NUMERICAL SOLUTION OF OPTIMAL CONTROL AND INVERSE PROBLEMS IN NON-REFLEXIVE BANACH SPACES

CUMULATIVE HABILITATION THESIS

Christian Clason

December 2012

Institute for Mathematics and Scientific Computing
Karl-Franzens-Universität Graz

CONTENTS

PREFACE v

LIST OF PUBLICATIONS viii

I SUMMARY

1 BACKGROUND 2
- 1.1 Measure spaces 2
- 1.2 Convex analysis 7
- 1.3 Semismooth Newton methods 12

2 OPTIMAL CONTROL WITH MEASURES 18
- 2.1 Elliptic problems with Radon measures 19
- 2.2 Parabolic problems with Radon measures 27
- 2.3 Elliptic problems with functions of bounded variation 31

3 OPTIMAL CONTROL WITH L^∞ FUNCTIONALS 34
- 3.1 L^∞ tracking 35
- 3.2 L^∞ control cost 38

4 INVERSE PROBLEMS WITH NON-GAUSSIAN NOISE 41
- 4.1 L^1 data fitting 43
- 4.2 L^∞ data fitting 49

5 APPLICATIONS IN BIOMEDICAL IMAGING 52
- 5.1 Diffuse optical imaging 52
- 5.2 Parallel magnetic resonance imaging 54

6 OUTLOOK 58

II OPTIMAL CONTROL WITH MEASURES

7 A DUALITY-BASED APPROACH TO ELLIPTIC CONTROL PROBLEMS IN NON-REFLEXIVE BANACH SPACES 60
7.1 Introduction 60
7.2 Existence and optimality conditions 63
7.3 Solution of the optimality systems 72
7.4 Numerical results 79
7.5 Conclusion 87
7.A Convergence of Moreau–Yosida regularization 87

8 A MEASURE SPACE APPROACH TO OPTIMAL SOURCE PLACEMENT 89
8.1 Introduction 89
8.2 Problem formulation and optimality system 90
8.3 Regularization 94
8.4 Semismooth Newton method 98
8.5 Numerical examples 100
8.6 Conclusion 105

9 APPROXIMATION OF ELLIPTIC CONTROL PROBLEMS IN MEASURE SPACES WITH SPARSE SOLUTIONS 106
9.1 Introduction 106
9.2 Optimality conditions 107
9.3 Approximation framework 110
9.4 Error estimates 115
9.5 A Neumann control problem 118
9.6 Computational results 121
9.7 Conclusion 125

10 PARABOLIC CONTROL PROBLEMS IN MEASURE SPACES WITH SPARSE SOLUTIONS 129
10.1 Introduction 129
10.2 Function spaces and well-posedness of the state equation 131
10.3 Analysis of the control problem 135
10.4 Approximation of the control problem 139
10.5 Error estimates 155
10.6 Numerical solution 157
10.7 Numerical examples 162
10.8 Conclusion 163
10.A Continuous optimality system 168

III OPTIMAL CONTROL WITH L^∞ FUNCTIONALS

11 MINIMAL INVASION: AN OPTIMAL L^∞ STATE CONSTRAINT PROBLEM 172
11.1 Introduction 172
11.2 Existence and regularization 174
11.3 Optimality system 177
11.4 Semismooth Newton method 182
11.5 Numerical results 188

12 A MINIMUM EFFORT OPTIMAL CONTROL PROBLEM FOR ELLIPTIC PDES 193
12.1 Introduction 193
12.2 Existence, uniqueness, and optimality system 195
12.3 Regularized problem 196
12.4 Solution of optimality system 200
12.5 Numerical examples 205
12.6 Conclusion 210
12.A Proof of Proposition 12.3.1 210
12.B Comparison of regularizations 212

IV INVERSE PROBLEMS WITH NON-GAUSSIAN NOISE

13 A SEMISMOOTH NEWTON METHOD FOR L^1 DATA FITTING WITH AUTOMATIC CHOICE OF REGULARIZATION PARAMETERS AND NOISE CALIBRATION 215
13.1 Introduction 215
13.2 Properties of minimizers 218
13.3 Solution by semismooth Newton method 222
13.4 Adaptive choice of regularization parameters 228
13.5 Numerical examples 237
13.6 Conclusion 244
13.A Convergence of smoothing for penalized box constraints 246
13.B Proof of Lemma 13.4.2 248
13.C Benchmark algorithms 250

14 A SEMISMOOTH NEWTON METHOD FOR NONLINEAR PARAMETER IDENTIFICATION PROBLEMS WITH IMPULSIVE NOISE 252
14.1 Introduction 252
14.2 L^1 fitting for nonlinear inverse problems 258
14.3 Solution by semismooth Newton method 263
14.4 Numerical examples 269
14.5 Conclusion 278
14.A Verification of properties for model problems 279

14.B Tables 285

15 L^∞ FITTING FOR INVERSE PROBLEMS WITH UNIFORM NOISE 288
15.1 Introduction 288
15.2 Well-posedness and regularization properties 290
15.3 Parameter choice 293
15.4 Numerical solution 295
15.5 Numerical examples 302
15.6 Conclusion 308

V APPLICATIONS IN BIOMEDICAL IMAGING

16 A DETERMINISTIC APPROACH TO THE ADAPTED OPTODE PLACEMENT FOR ILLUMINATION OF HIGHLY SCATTERING TISSUE 310
16.1 Introduction 310
16.2 Theory 312
16.3 Materials and methods 316
16.4 Results 317
16.5 Discussion 323

17 PARALLEL IMAGING WITH NONLINEAR RECONSTRUCTION USING VARIATIONAL PENALTIES 325
17.1 Introduction 325
17.2 Theory 326
17.3 Materials and methods 329
17.4 Results 332
17.5 Discussion 333
17.6 Conclusions 335

PREFACE

Historically, variational problems such as those arising in optimal control and inverse problems were predominantly posed in Hilbert spaces. Although this is indeed the correct setting for many physical models (e.g., those involving energy terms), it is just as often simply due to convenience and numerical tractability, and a Banach space setting would be more natural. In addition, interest in total variation minimization, sparsity constraints and bang-bang control have lead to significant progress in the analysis of such problems over the last decade. Numerical approaches, on the other hand, still tend to focus on either finite-dimensional (e.g., discretized) problems or those set in reflexive Banach spaces such as L^p, $1 < p < \infty$, due to their better differentiability properties. Hence, the motivation for this work, and its main contribution, is the development of efficient numerical algorithms for optimization problems in non-reflexive Banach spaces such as L^1 and L^∞. The main difficulty to overcome, apart from the non-standard functional-analytic setting, is the non-differentiability inherent in their formulation.

The problems treated here can be grouped as follows.

- *Optimal control problems in measure spaces*. These arise from control problems with sparsity constraints, which in finite dimensions can be enforced by ℓ^1 penalties. The corresponding infinite-dimensional problem, however, is not well-posed in L^1 due to the lack of weak compactness, and needs to be considered in spaces of Radon measures. Included here are also control problems in the space of functions of bounded variation (i.e., whose distributional gradient is a Radon measure), which can be used to promote piecewise constant controls.
- *Optimal control problems with* L^∞ *functionals*. The works in this group are concerned with problems with either tracking terms in L^∞, which correspond to minimizing the worst-case deviation from the target, or control costs in L^∞, which lead to bang-bang controls.
- *Inverse problems with nonsmooth discrepancy terms*. The standard L^2 data fitting term is statistically motivated by the assumption of Gaussian noise. For non-Gaussian noise, however, other data fitting terms turn out to be more appropriate. For impulsive noise (e.g., salt-and-pepper noise in digital imaging), L^1 fitting is more robust. Uniform noise (e.g., quantization errors) leads to L^∞ fitting.

- *Applications in biomedical imaging*. This group contains two examples from interdisciplinary cooperations where the non-reflexive Banach spaces considered above occur in applications. The first example demonstrates that the Radon measure space setting can be used to solve the problem of optimal placement of discrete light sources for the homogeneous illumination of tissue in optical tomography. The second example addresses an inverse problem in image reconstruction in magnetic resonance imaging using penalities of total variation-type.

For the numerical solution, Newton-type methods in function space are preferred due to their superlinear convergence and mesh independence. To apply these in spite of the lack of differentiability of the problems, a common approach is followed:

1. Using convex analysis (in particular, Fenchel duality) or a relaxation technique (or both), the original problem is transformed into a differentiable problem subject to pointwise constraints. Standard techniques (e.g, Maurer–Zowe-type conditions) then allow derivation of first order optimality conditions.
2. Due to the nonsmoothness of the original problem, these optimality systems are typically not sufficiently smooth to be solved by a Newton-type method. We therefore introduce a family of approximations that are amenable to such methods while avoiding unnecessary smoothing.
3. The resulting regularized optimality conditions lead to semismooth operator equations in function spaces, which can be solved using a semismooth Newton method. To deal with the local convergence of Newton-type methods, the Newton method is combined with a continuation strategy in the regularization parameter. In practice, this results in a globalization effect.

This thesis is organized as follows. Part I contains a summary of the submitted papers. It begins with a chapter collecting common background on partial differential equations with measure data, convex analysis, and semismooth Newton methods; the following chapters then summarize in turn the precise setting and the main results for each of the above groups. The purpose, besides introducing a consistent notation and terminology, is to motivate the appearing concepts and to illustrate their connections. Hence, some derivations and calculations are sketched, while formal statements of theorems and proofs are omitted; the reader is instead referred to the cited literature and to the full discussions in the corresponding chapters of the remaining parts. It should be pointed out that achieving a consistent notation and terminology in this part requires deviating, at times significantly, from that used in the original papers which make up Parts II–V of the thesis.

ACKNOWLEDGMENTS

This work was carried out as part of the SFB "Mathematical Optimization and Applications in Biomedical Sciences", and the financial support by the Austrian Science Fund (FWF) under grant SFB F32 is gratefully acknowledged. More importantly, the SFB fostered several exciting cooperations with the Institute of Medical Engineering of the TU Graz; here I want to thank in particular Manuel Freiberger and Florian Knoll for the enjoyable and fruitful collaboration.

It is also a pleasure to thank my colleagues – current and former – at the Institute of Mathematics, not only for many stimulating discussions, but also for making it such a pleasant place.

Finally, I wish to express my sincere gratitude to Prof. Karl Kunisch for giving me the opportunity to pursue my research in Graz and for his constant support. This work would not have been possible anywhere else.

Graz, September 2012

LIST OF PUBLICATIONS

This thesis consists of the following publications (in chronological order of submission), which have been retypeset from the original sources. Besides unifying the layout and the bibliography and correcting typos, no changes have been made.

C. Clason and K. Kunisch (2011). *A duality-based approach to elliptic control problems in non-reflexive Banach spaces*. ESAIM: Control, Optimisation and Calculus of Variations 17.1, pp. 243–266. DOI: 10.1051/cocv/2010003.

C. Clason, B. Jin, and K. Kunisch (2010). *A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration*. SIAM Journal on Imaging Sciences 3.2, pp. 199–231. DOI: 10.1137/090758003.

C. Clason, K. Ito, and K. Kunisch (2010). *Minimal invasion: An optimal L^∞ state constraint problem*. ESAIM: Math. Model. Numer. Anal. 45.3, pp. 505–522. DOI: 10.1051/m2an/2010064.

C. Clason and B. Jin (2012). *A semismooth Newton method for nonlinear parameter identification problems with impulsive noise*. SIAM Journal on Imaging Sciences 5, pp. 505–536. DOI: 10.1137/110826187.

F. Knoll, C. Clason, K. Bredies, M. Uecker, and R. Stollberger (2012). *Parallel imaging with nonlinear reconstruction using variational penalties*. Magnetic Resonance in Medicine 67.1, pp. 34–41. DOI: 10.1002/mrm.22964.

C. Clason, K. Ito, and K. Kunisch (2012). *A minimum effort optimal control problem for elliptic PDEs*. ESAIM: Mathematical Modelling and Numerical Analysis 46.4, pp. 911–927. DOI: 10.1051/m2an/2011074.

C. Clason and K. Kunisch (2012). *A measure space approach to optimal source placement*. Computational Optimization and Applications 53.1, pp. 155–171. DOI: 10.1007/s10589-011-9444-9.

E. Casas, C. Clason, and K. Kunisch (2012b). *Approximation of elliptic control problems in measure spaces with sparse solutions*. SIAM Journal on Control and Optimization 50.4, pp. 1735–1752. DOI: 10.1137/110843216.

C. Clason (2012). *L^∞ fitting for inverse problems with uniform noise*. Inverse Problems 28, p. 104007. DOI: 10.1088/0266-5611/28/10/104007.

E. Casas, C. Clason, and K. Kunisch (2012a). *Parabolic control problems in measure spaces with sparse solutions*. SIAM Journal on Control and Optimization. To appear.

P. Brunner, C. Clason, M. Freiberger, and H. Scharfetter (2012). *A deterministic approach to the adapted optode placement for illumination of highly scattering tissue*. Biomedical Optics Express 3.7, pp. 1732–1743. DOI: 10.1364/BOE.3.001732.

In addition, the following publications were written after the completion of the author's PhD degree in 2006.

C. Clason, B. Kaltenbacher, and S. Veljović (2009). *Boundary optimal control of the Westervelt and the Kuznetsov equation*. Journal of Mathematical Analysis and Applications 356.2, pp. 738–751. DOI: 10.1016/j.jmaa.2009.03.043.

S. Keeling, C. Clason, M. Hintermüller, F. Knoll, A. Laurain, and G. von Winckel (2012). *An image space approach to Cartesian based parallel MR imaging with total variation regularization*. Medical Image Analysis 16.1, pp. 189–200. DOI: 10.1016/j.media.2011.07.002.

F. Bauer and C. Clason (2011). *On theoretical limits in parallel magnetic resonance imaging*. International Journal of Tomography & Statistics 18.F11, pp. 10–23.

C. Clason and P. Hepperger (2009). *A forward approach to numerical data assimilation*. SIAM Journal on Scientific Computing 31.4, pp. 3090–3115. DOI: 10.1137/090746240.

F. Knoll, C. Clason, C. Diwoky, and R. Stollberger (2011). *Adapted random sampling patterns for accelerated MRI*. Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA) 24.1, pp. 43–50. DOI: 10.1007/s10334-010-0234-7.

M. Freiberger, C. Clason, and H. Scharfetter (2010a). *Adaptation and focusing of optode configurations for fluorescence optical tomography by experimental design methods*. Journal of Biomedical Optics 15.1, 016024, p. 016024. DOI: 10.1117/1.3316405.

F. Knoll, M. Unger, C. Clason, C. Diwoky, T. Pock, and R. Stollberger (2010). *Fast reduction of undersampling artifacts in radial MR angiography with 3D total variation on graphics hardware*. Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA) 23.2, pp. 103–114. DOI: 10.1007/s10334-010-0207-x.

C. Clason and G. von Winckel (2010). *On a bilinear optimization problem in parallel magnetic resonance imaging*. Applied Mathematics and Computation 216.4, pp. 1443–1452. DOI: 10.1016/j.amc.2010.02.047.

C. Clason, B. Jin, and K. Kunisch (2010). *A duality-based splitting method for ℓ^1–TV image restoration with automatic regularization parameter choice*. SIAM Journal on Scientific Computing 32.3, pp. 1484–1505. DOI: 10.1137/090768217.

M. Freiberger, C. Clason, and H. Scharfetter (2010b). *Total variation regularization for nonlinear fluorescence tomography with an augmented Lagrangian splitting approach*. Applied Optics 49.19. Selected for Virtual Journal for Biomedical Optics, Volume 5, Issue 11, pp. 3741–3747. DOI: 10.1364/AO.49.003741.

C. Clason and G. von Winckel (2012). *A general spectral method for the numerical simulation of one-dimensional interacting fermions*. Computer Physics Communications 183.2. Code published in *CPC Program Library* as AEKO_v1_1, pp. 405–417. DOI: 10.1016/j.cpc.2011.10.005.

C. Clason and B. Kaltenbacher (2013). *On the use of state constraints in optimal control of singular PDEs*. Systems & Control Letters 62.1, pp. 48–54. DOI: 10.1016/j.sysconle.2012.10.006.

A complete and up-to-date list of publications, including preprints and – where applicable – Matlab and Python code, can be found online at

http://www.uni-graz.at/~clason/publications.html.

Part I

SUMMARY

BACKGROUND

1

The purpose of this chapter is to collect the definitions and results on measure spaces, convex analysis and semismooth Newton methods that form the common basis for the results described in the remaining chapters, and to motivate the approach followed there.

1.1 MEASURE SPACES

We begin by giving some elementary definitions of dual spaces and operators, which serve to fix a common notation. In particular, we define spaces of Radon measures as dual spaces of continuous functions and discuss the well-posedness of partial differential equations with measures on the right hand side.

1.1.1 WEAK TOPOLOGIES

For a normed vector space V, we denote by V^* the topological dual of V. Note that this definition depends on the choice of the topology, specified via the *duality pairing* $\langle \cdot, \cdot \rangle_{V,V^*}$ between V and V^* (i.e., $(V, V^*, \langle \cdot, \cdot \rangle_{V,V^*})$ is a *dual pair*; see, e.g., [Werner 2011, Chapter VIII.3]). This fact that will play an important role in this work. The topological dual V^* is always a Banach space if equipped with the norm

$$\|v^*\|_{V^*} = \sup\{\langle v, v^* \rangle_{V,V^*} : v \in V,\ \|v\|_V \leqslant 1\}.$$

For non-reflexive spaces, two different topologies are of particular relevance.

(i) The *weak topology* corresponds to the duality pairing between V and V^* defined by

$$\langle v, v^* \rangle_{V,V^*} := v^*(v)$$

for all $v \in V$ and $v^* \in V^*$. In this case, V^* can be identified via the Hahn–Banach theorem with the space of all continuous linear forms on V, and the topological dual

coincides with the standard definitions. For example, the weak topological dual of $L^1(\Omega)$ can be identified with $L^\infty(\Omega)$, with the duality pairing reducing to

$$\langle v, v^* \rangle_{L^1, L^\infty} = \int_\Omega v(x) v^*(x)\, dx,$$

see, e.g., [Brezis 2010, Theorem 4.14]. If not specified otherwise, the topological dual is to be understood with respect to the weak topology.

(ii) If V^* is the weak topological dual of V, the duality pairing between V^* and V is defined by

$$\langle v^*, v \rangle_{V^*, V} := v^*(v)$$

for all $v^* \in V^*$ and $v \in V$. This allows identifying the *weak-$\star$ topological dual* (or *predual*) of V^* with V (i.e, the weak-$\star$ dual of $L^\infty(\Omega)$ is $L^1(\Omega)$).

(For reflexive Banach spaces, of course, both notions coincide.)

For a linear operator $A : X \to Y$ between the normed vector spaces X and Y, we call $A^* : Y^* \to X^*$ the *adjoint operator* to A if

$$\langle x, A^* y^* \rangle_{X, X^*} = \langle Ax, y^* \rangle_{Y, Y^*}$$

for all $x \in X$ and $y^* \in Y^*$. If the duality is taken with respect to the weak topology, this coincides again with the standard definition. On the other hand, if there exists $B : Y \to X$ such that $B^* = A$ with respect to the weak topology, we can identify the *weak-$\star$ adjoint* (or *preadjoint*) A^* of an operator $A : X^* \to Y^*$ with B, since

$$\langle x^*, By \rangle_{X^*, X} = \langle Ax^*, y \rangle_{Y^*, Y}$$

for all $x^* \in X^*$ and $y \in Y$.

1.1.2 SPACE OF RADON MEASURES

Let $\mathcal{M}(X)$ denote the vector space of all bounded Borel measures on $X \subset \mathbb{R}^n$, that is of all bounded σ-additive set functions $\mu : \mathcal{B}(X) \to \mathbb{R}$ defined on the Borel algebra $\mathcal{B}(X)$ satisfying $\mu(\emptyset) = 0$. The *total variation* of $\mu \in \mathcal{M}(X)$ is defined for all $B \in \mathcal{B}(\Omega)$ as

$$|\mu|(B) := \sup \left\{ \sum_{i=0}^{\infty} |\mu(B_i)| : \bigcup_{i=0}^{\infty} B_i = B \right\},$$

where the supremum is taken over all partitions of B. We recall that every Radon measure μ has a unique *Jordan decomposition* $\mu = \mu^+ - \mu^-$ into two positive measures (i.e., $\mu^+(B), \mu^-(B) \geqslant 0$ for all Borel sets B). The *support* $\operatorname{supp}(\mu)$ of a Radon measure μ is defined as the complement of the union of all open null sets with respect to μ.

By the Riesz representation theorem, $\mathcal{M}(X)$ can be identified with the dual of spaces of continuous functions on X, endowed with the norm $\|\nu\|_C = \sup_{x\in X} |\nu(x)|$. Based on the boundary behavior of the continuous functions, we discern three cases.

(i) Let $C_0(\Omega)$ be the completion of the space of all continuous functions with compact support in the simply connected domain $\Omega \subset \mathbb{R}^n$ with respect to the norm $\|\nu\|_C$, i.e, the space of all functions vanishing on the boundary $\partial\Omega$ (or at infinity if Ω is unbounded). In this case, we set $X = \Omega$ and have from [Elstrodt 2005, Satz VIII.2.26] that the weak topological dual of $C_0(\Omega)$ can be isometrically identified with $\mathcal{M}(\Omega)$.

(ii) If Ω is bounded, $\overline{\Omega}$ is compact, and we can identify $\mathcal{M}(\overline{\Omega})$ with the weak topological dual of the space $C(\overline{\Omega})$ of continuous functions that can be continuously extended to the boundary of Ω; see, e.g., [Dunford and Schwartz 1988, Theorem IV.6.3].

(iii) If the functions vanish only on an open part Γ of the boundary, we set $X = \overline{\Omega} \setminus \Gamma$. Then, X is compact, and we can apply the same argument as in case (ii) to deduce that the weak topological dual of $C_\Gamma(\overline{\Omega}) := \{\nu \in C(\overline{\Omega}) : \nu|_\Gamma = 0\}$ can be identified with the space $\mathcal{M}_\Gamma(\overline{\Omega}) := \{\mu \in \mathcal{M}(\overline{\Omega}) : \mu(\Gamma) = 0\}$.

For the sake of presentation, we will restrict the discussion to the first case; however, all results hold for the other two cases as well (with obvious modifications regarding the boundary behavior). The Riesz representation theorem leads to the equivalent characterization

$$\|\mu\|_{\mathcal{M}} = \sup\left\{\int_\Omega \nu \,\mathrm{d}\mu : \nu \in C_0(\Omega),\ \|\nu\|_{C_0} \leqslant 1\right\}. \tag{1.1.1}$$

In particular, this makes $\mathcal{M}(\Omega)$ a Banach space. For the purposes of dual pairings, we will always equip $\mathcal{M}(\Omega)$ with the weak-$\star$ topology, with respect to which $\|\cdot\|_{\mathcal{M}}$ is lower semicontinuous.

With the formalism of section 1.1.1, this allows identifying the weak-$\star$ dual of $\mathcal{M}(\Omega)$ with $C_0(\Omega)$, corresponding to the duality pairing

$$\langle \mu, \nu\rangle_{\mathcal{M},C} = \int_\Omega \nu \,\mathrm{d}\mu.$$

Note that (1.1.1) allows us to isometrically identify $L^1(\Omega)$ with a subspace of $\mathcal{M}(\Omega)$, such that $\|u\|_{\mathcal{M}} = \|u\|_{L^1}$ for all $u \in L^1(\Omega)$; see, e.g., [Brezis 2010, Chapter 4.5.3]. In addition, the Rellich–Kondrachov theorem yields that $W_0^{1,q}(\Omega) \hookrightarrow C_0(\Omega)$ for $q > n$, and this embedding is dense and compact. Hence we have the dense and compact embedding

$$\mathcal{M}(\Omega) \hookrightarrow W_0^{1,q}(\Omega)^* = W^{-1,q'}(\Omega)$$

for $1 < q' < \frac{n}{n-1}$. We will make use of this embedding to show well-posedness of partial differential equations with measure right hand sides.

For time-dependent measure-valued functions, the situation is slightly more delicate. Associated to the interval $(0,T)$ we define the space $L^2(0,T;C_0(\Omega))$ of measurable functions $z:(0,T)\to C_0(\Omega)$ for which the associated norm given by

$$\|z\|_{L^2(C_0)} := \left(\int_0^T \|z(t)\|_{C_0}^2 \,\mathrm{d}t\right)^{1/2}$$

is finite. Due to the fact that $C_0(\Omega)$ is a separable Banach space, $L^2(0,T;C_0(\Omega))$ is also a separable Banach space; see, e.g., [Warga 1972, Theorem I.5.18]. Let $L^2(0,T;\mathcal{M}(\Omega))$ denote the space of weakly measurable functions $u:[0,T]\to\mathcal{M}(\Omega)$ for which the norm

$$\|u\|_{L^2(\mathcal{M})} = \left(\int_0^T \|u(t)\|_{\mathcal{M}}^2 \,\mathrm{d}t\right)^{1/2}$$

is finite. This choice makes $L^2(0,T;\mathcal{M}(\Omega))$ a Banach space and guarantees that it can be identified with the weak topological dual of $L^2(0,T;C_0(\Omega))$, where the duality relation is given by

$$\langle z,u\rangle_{L^2(C_0),L^2(\mathcal{M})} = \int_0^T \langle z(t),u(t)\rangle_{C_0,\mathcal{M}}\,\mathrm{d}t,$$

with $\langle\cdot,\cdot\rangle_{C_0,\mathcal{M}}$ denoting the duality pairing between $C_0(\Omega)$ and $\mathcal{M}(\Omega)$; see [Edwards 1965, Theorem 8.20.3]. Vice versa, $L^2(0,T;C_0(\Omega))$ can be seen as the weak-$\star$ dual of $L^2(0,T;\mathcal{M}(\Omega))$.

Finally, we recall that $BV(\Omega)$, the space of functions of bounded variation, consists of all $u\in L^1(\Omega)$ for which the distributional gradient Du belongs to $(\mathcal{M}(\Omega))^n$. Furthermore, the mapping $u\mapsto\|u\|_{BV}$,

$$\|u\|_{BV} := \int_\Omega |Du|\,\mathrm{d}x = \sup\left\{\int_\Omega u(-\operatorname{div} v)\,\mathrm{d}x : v\in(C_0^\infty(\Omega))^n,\ \|v\|_{(C_0)^n}\leqslant 1\right\} \tag{1.1.2}$$

(which can be infinite) is lower semicontinuous in the topology of $L^1(\Omega)$, and $u\in L^1(\Omega)$ is in $BV(\Omega)$ if and only if $\|u\|_{BV}$ is finite. In this case $\|\cdot\|_{BV}$ is referred to as the *total variation seminorm*. (If $v\in H^1(\Omega)$, then $\|u\|_{BV}=\int_\Omega|\nabla u|\,\mathrm{d}x$.) Endowed with the norm $\|\cdot\|_{L^1}+\|\cdot\|_{BV}$, $BV(\Omega)$ is a (non-reflexive) Banach space; see, e.g., [Attouch, Buttazzo, and Michaille 2006, Chapter 10.1].

1.1.3 PARTIAL DIFFERENTIAL EQUATIONS WITH MEASURE DATA

Measure-valued right hand sides or boundary conditions in partial differential equations have attracted recent interest due to their role in the adjoint equation for optimal control problems with pointwise state constraints (see, e.g., [Casas 1986; Alibert and Raymond 1997]), although measure-valued right hand sides have already been treated in [Stampacchia 1965] in the context of the Green's function of the Dirichlet problem for an elliptic operator with

discontinuous coefficients. Correspondingly, several different solution concepts have been introduced in [Stampacchia 1965; Boccardo and Gallouët 1989; Casas 1986; Alibert and Raymond 1997]. All of these are fundamentally based on a duality technique, and have been shown to coincide; see [Meyer, Panizzi, and Schiela 2011]. Here, we follow [Casas 1986].

We first discuss elliptic problems. Consider the operator

$$Ay = -\sum_{j,k=1}^{n} \partial_j(a_{jk}(x)\partial_k y + d_j(x)y) + \sum_{j=1}^{n} b_j(x)\partial_j y + d(x)y,$$

and for $\mu \in \mathcal{M}(\Omega)$ the abstract Dirichlet problem

$$\begin{cases} Ay = \mu, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega. \end{cases} \tag{1.1.3}$$

We call $y \in L^1(\Omega)$ a *very weak solution* of (1.1.3) if

$$\int_\Omega y A^* z \, dx = \int_\Omega z \, d\mu \qquad \text{for all } z \in H^2(\Omega) \cap H_0^1(\Omega),$$

where A^* is the (weak) adjoint of A. Here, we shall for simplicity assume that A^* has maximal regularity as an operator from $W_0^{1,q}(\Omega)$ to $W^{-1,q}(\Omega)$ for $q > n$, which is the case if $a_{jk}, b_j \in C^{0,\delta}(\overline{\Omega})$ for some $\delta \in (0,1)$; A is uniformly elliptic; $d_j, d \in L^\infty(\Omega)$; the lower order coefficients are small enough (see, e.g., [Gilbarg and Trudinger 2001, Th. 8.3]); and $\partial\Omega$ is of class $C^{1,1}$, or Ω is a parallelepiped; see, e.g., [Ladyzhenskaya and Ural'tseva 1968, pp. 169–189] and [Troianiello 1987, Th. 2.24]. If not otherwise specified, any elliptic operator A mentioned in the following is assumed to satisfy these requirements. Under these conditions, A^* is an isomorphism from $W_0^{1,q}(\Omega)$ to $W^{-1,q}(\Omega)$, and the closed range theorem together with reflexivity of these spaces implies that A is an isomorphism from $W_0^{1,q'}(\Omega)$ to $W^{-1,q'}(\Omega)$ for $q' < \frac{n}{n-1}$. Hence by the continuous embedding $\mathcal{M}(\Omega) \hookrightarrow W^{-1,q'}(\Omega)$, problem (1.1.3) admits a unique solution $y \in W_0^{1,q'}(\Omega)$ satisfying

$$\|y\|_{W^{1,q'}} \leqslant C \, \|\mu\|_{\mathcal{M}}$$

for a constant C independent of μ. In this case, y also solves (1.1.3) in the usual weak sense. Note that this approach ensures the existence of a weak-$\star$ adjoint of A, which can be identified with A^*; and similarly for A^{-1} and $(A^*)^{-1}$. Furthermore, the compactness of the embedding $\mathcal{M}(\Omega) \hookrightarrow W^{-1,q'}(\Omega)$ yields that for any sequence μ_k converging weakly-$\star$ in $\mathcal{M}(\Omega)$ to μ, the sequence of corresponding solutions y_k converges strongly in $W_0^{1,q'}(\Omega)$ to y.

If A^* does not enjoy maximal regularity, we still have existence of a solution $y \in W^{1,q'}(\Omega)$ to (1.1.3), but uniqueness in $W^{1,q'}(\Omega)$ requires (one of several equivalent) additional assumptions (such as y being the limit of a sequence of regularized problems or satisfying an integration by parts formula). We refer to [Meyer, Panizzi, and Schiela 2011] for details.

The case of measure-valued boundary data or parabolic equations can be treated in an analogous fashion; see, e.g., [Casas 1993] and [Casas 1997], respectively. Finally, we note that by the chain of continuous embeddings $BV(\Omega) \hookrightarrow L^1(\Omega) \hookrightarrow \mathcal{M}(\Omega)$, we can apply the above results to $\mu \in BV(\Omega)$ as well.

1.2 CONVEX ANALYSIS

The task of finding a minimizer u of a Fréchet differentiable functional J can often be reduced to solving the first order necessary optimality conditions $J'(u) = 0$, which is sometimes referred to as *Fermat's principle*. If J is non-differentiable but convex, as is mostly the case in this work, the convex subdifferential replaces the nonexistant Fréchet derivative, as it satisfies Fermat's principle and allows for a rich calculus – in particular Fenchel duality – that can be used to obtain explicit optimality conditions. The classical reference in the context of this work is [Ekeland and Témam 1999], while [Attouch, Buttazzo, and Michaille 2006, Chapter 9] contains a readable and complete overview. A rigorous and extensive treatment can be found in the excellent textbook [Schirotzek 2007], which we follow here.

1.2.1 CONVEX CONJUGATES

Here and below, let V again be a normed vector space. Recall that a function $f : V \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is called *convex* if

$$f(\lambda v_1 + (1-\lambda)v_2) \leqslant \lambda f(v_1) + (1-\lambda) f(v_2)$$

for all $v_1, v_2 \in V$ and $\lambda \in (0,1)$, and *proper* if f is not identically equal to $+\infty$. For example, the *indicator function* δ_C of a nonempty, convex set $C \subset V$, defined by

$$\delta_C(v) := \begin{cases} 0 & \text{if } v \in C, \\ \infty & \text{otherwise,} \end{cases}$$

is convex and proper. This function will appear frequently in the following.

As we will see, one reason for the usefulness of convex subdifferentials in our context is their connection with the Legendre–Fenchel transform. For a function $f : V \to \bar{\mathbb{R}}$, the *Fenchel conjugate* (or *convex conjugate*) is defined as

$$f^* : V^* \to \bar{\mathbb{R}}, \qquad f^*(v^*) = \sup_{v \in V} \langle v, v^* \rangle_{V,V^*} - f(v). \tag{1.2.1}$$

The convex conjugate is always convex and lower semicontinuous. If f is convex and proper, then f^* is proper as well; see, e.g., [Schirotzek 2007, Proposition 2.2.3]. We also introduce the *biconjugate* of f, defined as

$$f^{**} : V \to \bar{\mathbb{R}}, \qquad f^{**}(v) = \sup_{v^* \in V^*} \langle v^*, v \rangle_{V^*,V} - f^*(v^*)$$

(i.e., if V^* is the weak dual of V, we take V as the weak-$\star$ dual of V^* (or vice versa) and set $f^{**} = (f^*)^*$). If f is proper, the Fenchel–Moreau–Rockafellar theorem states that $f^{**} = f$ if and only if f is convex and lower semicontinuous; see, e.g., [Schirotzek 2007, Theorem 2.2.4].

We give a few relevant examples; see [Schirotzek 2007, Examples 2.2.2, 2.2.5, and 2.2.6].

(i) Let $V = L^2(\Omega)$ and $f(v) = \frac{1}{2}\|v\|_{L^2}^2$. We identify V^* with V (i.e., the duality pairing is the inner product in $L^2(\Omega)$). Then, the function to be maximized in (1.2.1) is strictly concave and differentiable, so that the supremum is attained if and only if $v^* = f'(v) = v$. Inserting this into the definition and simplifying, we obtain

$$f^* : L^2(\Omega) \to \mathbb{R}, \qquad f^*(v^*) = \frac{1}{2}\|v^*\|_{L^2}^2 .$$

(ii) Let V be a normed vector space and $f(v) = \delta_{B_V}(v)$, where B_V is the unit ball with respect to the norm $\|\cdot\|_V$. We take V^* as the weak (or weak-$\star$) dual of V and compute $f^*(v^*)$ for $v^* \in V^*$:

$$\delta_{B_V}^*(v^*) = \sup_{v\in V} \langle v, v^*\rangle_{V,V^*} - \delta_{B_V}(v) = \sup_{\|v\|_V \leqslant 1} \langle v, v^*\rangle_{V,V^*} = \|v^*\|_{V^*} .$$

(iii) Let V be as above, V^* its weak topological dual and $f(v) = \|v\|_v$. We compute $f^*(v^*)$ for given $v^* \in V^*$ by discerning two cases:

a) $\|v^*\|_{V^*} \leqslant 1$. In this case, $\langle v, v^*\rangle_{V,V^*} \leqslant \|v\|_V \|v^*\|_{V^*} \leqslant \|v\|_V$ for all $v \in V$ and $\langle 0, v^*\rangle_{V,V^*} = 0 = \|0\|_V$. Hence,

$$f^*(v^*) = \sup_{v\in V} \langle v, v^*\rangle_{V,V^*} - \|v\|_V = 0.$$

b) $\|v^*\|_{V^*} > 1$. Then by the definition of the dual norm, there exists a $v_0 \in V$ with $\langle v_0, v^*\rangle_{V,V^*} > \|v_0\|_V$. Taking $\rho \to \infty$ in

$$0 < \rho\left(\langle v_0, v^*\rangle_{V,V^*} - \|v_0\|_V\right) = \langle \rho v_0, v^*\rangle_{V,V^*} - \|\rho v_0\|_V \leqslant f^*(v^*)$$

yields $f^*(v^*) = +\infty$.

We conclude that $f^* = \delta_{B_{V^*}}$.

If we take the dual with respect to the weak-$\star$ topology between V^* and V, this result also follows directly from the Fenchel–Moreau–Rockafellar theorem and example (ii) by noting that

$$\begin{aligned} \delta_{B_V}(v) = \delta_{B_V}^{**}(v) &= \sup_{v^*\in V^*} \langle v, v^*\rangle_{V,V^*} - \delta_{B_V}^*(v^*) \\ &= \sup_{v^*\in V^*} \langle v^*, v\rangle_{V^*,V} - \|v^*\|_{V^*} = f^*(v) \end{aligned}$$

for all $v \in V$.

Furthermore, straightforward calculation yields the following useful transformation rules; see, e.g., [Ekeland and Témam 1999, page 17]. For $f : V \to \bar{\mathbb{R}}$, we have for all $\alpha \in \mathbb{R}$ and $a \in V$ that

$$(\alpha f(\cdot))^*(v^*) = \alpha f^*(\alpha^{-1} v^*), \tag{1.2.2}$$

$$f(\cdot - a)^*(v^*) = f^*(v^*) + \langle a, v^* \rangle_{V,V^*}. \tag{1.2.3}$$

In particular, the above yields for every $\alpha > 0$

$$(\alpha \|\cdot\|_{L^1})^*(v^*) = \delta_{B_{L^\infty}}(\alpha^{-1} v^*) = \begin{cases} 0 & \text{if } |v^*(x)| \leqslant \alpha \text{ for almost all } x \in \Omega, \\ \infty & \text{otherwise.} \end{cases} \tag{1.2.4}$$

On the other hand, applying the same to $V^* = \mathcal{M}(\Omega)$ and its weak-$\star$ dual $V = C_0(\Omega)$, we obtain

$$(\alpha \|\cdot\|_{\mathcal{M}})^*(v) = \delta_{B_{C_0}}(\alpha^{-1} v) = \begin{cases} 0 & \text{if } |v(x)| \leqslant \alpha \text{ for all } x \in \Omega, \\ \infty & \text{otherwise.} \end{cases} \tag{1.2.5}$$

We will make use of this duality to pass from problems involving these nonsmooth norms to smooth problems with pointwise constraints.

1.2.2 CONVEX SUBDIFFERENTIALS

Let $f : V \to \bar{\mathbb{R}}$ be convex and proper, and let $\bar{v} \in V$ with $f(\bar{v}) < \infty$. The set

$$\partial f(\bar{v}) := \{v^* \in V^* : \langle v - \bar{v}, v^* \rangle_{V,V^*} \leqslant f(v) - f(\bar{v}) \text{ for all } v \in V\} \tag{1.2.6}$$

is called *subdifferential* of f at $\bar{v}$. Every $v^* \in \partial f(\bar{v})$ is called *subgradient* of f at $\bar{v}$. From the definition (1.2.6), we immediately obtain Fermat's principle for convex functions: The point $\bar{v}$ is a minimizer of f if and only if $f(\bar{v}) \leqslant f(v)$ for all $v \in V$, which is equivalent to $0 \in \partial f(\bar{v})$.

The convex subdifferential satisfies the following sum rule. Let $f_1, f_2 : V \to \bar{\mathbb{R}}$ be convex and proper. If there exists a point $\bar{v} \in V$ such that $f_1(\bar{v}), f_2(\bar{v}) < \infty$ and f_2 is continuous at $\bar{v}$, then

$$\partial(f_1 + f_2)(v) = \partial f_1(v) + \partial f_2(v) \tag{1.2.7}$$

for all $v \in V$ for which f_1 and f_2 are finite; see, e.g, [Schirotzek 2007, Proposition 4.5.1]. Further calculus rules can be obtained by relating the convex subdifferential to other derivatives. If f is convex, proper, and Gâteaux differentiable, then $\partial f(v) = \{f'(v)\}$; see, e.g., [Schirotzek 2007, Proposition 4.1.8]. On the other hand, any convex and proper function that is bounded from above is locally Lipschitz, and in this case the convex subdifferential coincides with the generalized gradient of Clarke; see, e.g., [Schirotzek 2007, Proposition 7.3.9]. In particular, we can apply the sum and chain rules for the generalized gradient; see [Clarke 1990, Theorems 2.3.3 and 2.3.10].

The usefulness of the convex subdifferential now lies in the fact that it can often be characterized explicitly. To give an example, we return to the indicator function of a convex set C. For $\bar{v} \in C$, we have

$$\begin{aligned} v^* \in \partial\delta_C(\bar{v}) &\Leftrightarrow \langle v - \bar{v}, v^* \rangle_{V,V^*} \leqslant \delta_C(v) && \text{for all } v \in V \\ &\Leftrightarrow \langle v - \bar{v}, v^* \rangle_{V,V^*} \leqslant 0 && \text{for all } v \in C, \end{aligned}$$

since the condition is trivially satisfied for all $v \notin C$. In other words, the subdifferential of the indicator function of a convex set is its normal cone. Of particular importance for us will be the case when the set C_α for $\alpha > 0$ is given by pointwise constraints,

$$C_\alpha = \{v \in C_0(\Omega) : -\alpha \leqslant v(x) \leqslant \alpha \quad \text{for all } x \in \Omega\},$$

where we can give a pointwise characterization of the subdifferential. By separate pointwise inspection of the

- *positive active set*: $x \in \mathcal{A}^+ := \{x \in \Omega : \bar{v}(x) = \alpha\}$,
- *negative active set*: $x \in \mathcal{A}^- := \{x \in \Omega : \bar{v}(x) = -\alpha\}$,
- *inactive set*: $x \in \mathcal{I} := \{x \in \Omega : |\bar{v}(x)| < \alpha\}$,

we obtain the equivalent *complementarity conditions* for $v^* \in \partial\delta_{C_\alpha}(\bar{v}) \subset \mathcal{M}(\Omega)$:

$$v^*(\mathcal{A}^+) \leqslant 0, \qquad v^*(\mathcal{A}^-) \geqslant 0, \qquad v^*(\mathcal{I}) = 0.$$

If v^* is sufficiently regular (e.g., $v^* \in L^2(\Omega)$ or V is finite-dimensional), the complementarity conditions can equivalently be expressed for any $\gamma > 0$ as

$$v^* + \max(0, -v^* + \gamma(\bar{v} - \alpha)) + \min(0, -v^* + \gamma(\bar{v} + \alpha)) = 0, \tag{1.2.8}$$

where max and min are taken pointwise almost everywhere in Ω (or componentwise in finite dimensions); this can again be seen by pointwise inspection. Optimality systems involving equation (1.2.8) can then be solved by Newton-type methods, and the regularity requirement is one reason why we will need to introduce approximations.

Another relevant example is the subdifferential of the norm of V. It is straightforward to verify using the definition of the subdifferential and the dual norm that

$$\partial(\|\cdot\|_V)(v) = \begin{cases} \{v^* \in V^* : \langle v, v^* \rangle_{V,V^*} = \|v\|_V \text{ and } \|v^*\|_{V^*} = 1\} & \text{if } v \neq 0, \\ B_{V^*} & \text{if } v = 0, \end{cases}$$

see [Schirotzek 2007, Proposition 4.6.2]. For $V = L^1(\Omega)$, we can use pointwise inspection to explicitly compute $v^* \in L^\infty(\Omega)$ for given v to obtain

$$\partial(\|\cdot\|_{L^1})(v) = \operatorname{sign}(v) := \begin{cases} 1 & \text{if } v(x) > 0, \\ -1 & \text{if } v(x) < 0, \\ t \in [-1, 1] & \text{if } v(x) = 0 \end{cases} \tag{1.2.9}$$

Since the multi-valued sign is not differentiable even in a generalized sense, we again need to consider an approximation before we can apply a Newton-type method.

1.2.3 FENCHEL DUALITY

We now discuss the relation between the Fenchel conjugate and the subdifferential of convex functions. Let f be a proper and convex function f. Then we immediately obtain from the definitions of the conjugate and the subdifferential that for all $v \in V$ with $f(v) < \infty$ and all $v^* \in V^*$ the *Fenchel–Young inequality*

$$\langle v, v^* \rangle_{V,V^*} \leqslant f(v) + f^*(v^*), \tag{1.2.10}$$

is satisfied, where equality holds (and thus the supremum in (1.2.1) is attained) if and only if $v^* \in \partial f(v)$. Hence, inserting in turn arbitrary $w^* \in V^*$ and $v^* \in \partial f(v)$ into (1.2.10) and subtracting yields

$$\langle v, w^* - v^* \rangle_{V,V^*} \leqslant (f(v) + f^*(w^*)) - (f(v) + f^*(v^*)) = f^*(w^*) - f^*(v^*)$$

for every $w^* \in V^*$, i.e., $v \in \partial f^*(v^*)$. If f is in addition lower semicontinuous, we can apply the Fenchel–Moreau–Rockafellar theorem to also obtain the converse, and thus

$$v^* \in \partial f(v) \quad \Leftrightarrow \quad v \in \partial f^*(v^*), \tag{1.2.11}$$

see [Schirotzek 2007, Proposition 4.4.4]. When combined with (1.2.4) or (1.2.5) and the characterization (1.2.8) or (1.2.9), this relation is the key in deriving useful optimality conditions for problems involving L^1 or measure space norms.

The *Fenchel duality theorem* combines in a particularly elegant way the relation (1.2.11), the sum rule (1.2.7), and a chain rule to obtain existence of and optimality conditions for a solution to a convex optimization problem. Let V and Y be Banach spaces, $\mathcal{F} : V \to \bar{\mathbb{R}}$, $\mathcal{G} : Y \to \bar{\mathbb{R}}$ be convex, proper, lower semicontinuous functions and $\Lambda : V \to Y$ be a continuous linear operator. If there exists a $v_0 \in V$ such that $\mathcal{F}(v_0) < \infty$, $\mathcal{G}(\Lambda v_0) < \infty$, and $\mathcal{G}$ is continuous at Λv_0, then

$$\inf_{v \in V} \mathcal{F}(v) + \mathcal{G}(\Lambda v) = \sup_{q \in Y^*} -\mathcal{F}^*(\Lambda^* q) - \mathcal{G}^*(-q), \tag{1.2.12}$$

and the optimization problem on the right hand side (referred to as the *dual problem*) has at least one solution; see, e.g., [Ekeland and Témam 1999, Theorem III.4.1]. (Existence of a solution to the problem on the left hand side – the *primal problem* – follows directly from the assumptions on $\mathcal{F}$, $\mathcal{G}$, and Λ by standard arguments.) Furthermore, the equality in (1.2.12) is attained at $(\bar{v}, \bar{q})$ if and only if the *extremality relations*

$$\begin{cases} \Lambda^* \bar{q} \in \partial \mathcal{F}(\bar{v}), \\ -\bar{q} \in \partial \mathcal{G}(\Lambda \bar{v}), \end{cases} \tag{1.2.13}$$

hold; see, e.g., [Ekeland and Témam 1999, Proposition III.4.1]. Depending on the context, one or both of these relations can be reformulated in terms of $\mathcal{F}^*$ and $\mathcal{G}^*$ using the equivalence (1.2.11). The conditions and consequences of the Fenchel duality theorem should be compared with classical regular point conditions (e.g., [Maurer and Zowe 1979; Ito and Kunisch 2008]) for the existence of Lagrange multipliers in constrained optimization.

1.3 SEMISMOOTH NEWTON METHODS

It remains to formulate a numerical method that can solve nonsmooth equations of the form (1.2.8) in an efficient manner. Just as the convex subdifferential proved to be suitable replacement for the Fréchet derivative in the context of optimality conditions, we need to consider a generalized derivative that can replace the Fréchet derivative in a Newton-type method and still allow superlinear convergence. In addition, it needs to provide a sufficiently rich calculus and the possibility for explicit characterization to be implementable in a numerical algorithm. These requirements lead to semismooth Newton methods. This section gives a brief overview of the theory in finite and infinite dimensions; for details and proofs, the reader is referred to the expositions in [Ito and Kunisch 2008; Ulbrich 2011; Schiela 2008].

To motivate the definitions, it will be instructive to first consider the convergence of an abstract generalized Newton method. Let Banach spaces X, Y, a mapping $F : X \to Y$, and $x^* \in X$ with $F(x^*) = 0$ be given. A generalized Newton method to compute an approximation of x^* can be described as follows:

1: Choose $x^0 \in X$
2: **for** $k = 0, 1, \ldots$ **do**
3: Choose an invertible linear operator $M_k \in \mathcal{L}(X, Y)$
4: Set $x^{k+1} = x^k - M_k^{-1} F(x^k)$
5: **end for**

We can now ask ourselves when convergence of the iterates $x^k \to x^*$ holds, and in particular when it is *superlinear*, i.e.,

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|_X}{\|x^k - x^*\|_X} = 0. \tag{1.3.1}$$

Set $M(x^k) := M_k$ and $d^k = x^k - x^*$. Then we can use the definition of the Newton step and the fact that $F(x^*) = 0$ to obtain

$$\begin{aligned}
\|x^{k+1} - x^*\|_X &= \|x^k - M(x_k)^{-1} F(x^k) - x^*\|_X \\
&= \|M(x_k)^{-1} \left[F(x^k) - F(x^*) - M(x_k)(x^k - x^*)\right]\|_X \\
&= \|M(x_k)^{-1} \left[F(x^k) - F(x^*) - M(x_k) d^k\right]\|_X \\
&\leqslant \|M(x_k)^{-1}\|_{\mathcal{L}(Y,X)} \|F(x^* + d^k) - F(x^*) - M(x^k) d^k\|_Y
\end{aligned}$$

Hence, (1.3.1) holds if both a

- *uniform regularity condition*: there exists a $C > 0$ such that

$$\left\|M(x_k)^{-1}\right\|_{\mathcal{L}(Y,X)} \leqslant C$$

  for all k, and an

- *approximation condition*:

$$\lim_{\|d^k\|_X \to 0} \frac{\left\|F(x^* + d^k) - F(x^*) - M(x^* + d^k)d^k\right\|_Y}{\|d^k\|_X} = 0,$$

hold. In this case, there exists a neighborhood $N(x^*)$ of x^* such that

$$\left\|x^{k+1} - x^*\right\|_X < \frac{1}{2}\left\|x^k - x^*\right\|_X$$

for an $x^k \in N(x^*)$, which by induction implies $d^k \to 0$ and hence the desired (local) superlinear convergence.

If F is continuously Fréchet differentiable, the approximation condition holds by definition for the Fréchet derivative $M_k = F'(x^k)$, and we arrive at the classical Newton method. For nonsmooth F, we simply take a linear operator which satisfies the uniform regularity and approximation conditions. Naturally, the choice $M_k \in \partial F(x^k)$ for an appropriate subdifferential suggests itself.

1.3.1 SEMISMOOTH NEWTON METHODS IN FINITE DIMENSIONS

If X and Y are finite-dimensional, an appropriate choice is the Clarke subdifferential. Recall that by Rademacher's theorem, every Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable almost everywhere; see, e.g. [Ziemer 1989, Theorem 2.2.1]. We can then define the *Clarke subdifferential* at $x \in \mathbb{R}^n$ as

$$\partial_C f(x) = \operatorname{co}\left\{\lim_{n\to\infty} f'(x_n) : \{x_n\}_{n\in\mathbb{N}} \text{ with } x_n \to x,\ f \text{ differentiable at } x_n\right\},$$

where co denotes the convex hull. To use an element of the Clarke subdifferential as linear operator in our Newton method, we need to ensure in particular that the approximation condition holds. In fact, we will require a slightly stronger condition. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *semismooth* at $x \in \mathbb{R}^n$ if

(i) f is Lipschitz continuous near x,

(ii) f is directionally differentiable at x,

(iii) $\lim_{\|h\|\to 0} \sup_{M\in\partial_C f(x+h)} \frac{\|F(x+h) - F(x) - Mh\|}{\|h\|} = 0.$

Note that we take the subgradient not in the linearization point but in a neighborhood, so we avoid evaluating $\partial_C f$ at the points where f is not differentiable. This definition is equivalent to the original one of [Mifflin 1977] (for real-valued functions) and [Qi and Sun 1993] (for vector-valued functions); see [Ulbrich 2011, Proposition 2.7].

For a locally Lipschitz continuous function, this leads to the *semismooth Newton method*

1: Choose $x^0 \in X$
2: **for** $k = 0, 1, \dots$ **do**
3: Choose $M_k \in \partial_C f(x^k)$
4: Set $x^{k+1} = x^k - M_k^{-1} f(x^k)$
5: **end for**

If f is semismooth at x^* with $f(x^*) = 0$ and all M_k satisfy the uniform regularity condition, this iteration converges (locally) superlinearly to x^*; see, e.g., [Ulbrich 2011, Proposition 2.12]. (In fact, condition (iii) of the definition is sufficient.) A similar abstract framework for the superlinear convergence of Newton methods was proposed in [Kummer 1988].

We close this section with some relevant examples. Clearly, if f is continuously differentiable at x, then f is semismooth at x with $\partial_C f(x) = \{f'(x)\}$. This can be extended to continuous piecewise differentiable functions. Let $f_1, \dots, f_N \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ be given. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *piecewise differentiable* if

$$f(x) \in \{f_1(x), \dots, f_N(x)\} \quad \text{for all } x \in \mathbb{R}^n.$$

Then, f is semismooth, and

$$\partial_C f(x) = \operatorname{co}\{f_i'(x) : f(x) = f_i(x) \text{ and } x \in \operatorname{cl}\operatorname{int}\{y : f(y) = f_i(y)\}\};$$

see, e.g., [Ulbrich 2011, Proposition 2.26]. This means that we can differentiate piecewise, and where pieces overlap, take the convex hull of all possible values at x excluding those that are only attained on a null set containing x. As a concrete example, the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \max(0, x)$ is semismooth, and

$$\partial_C f(x) = \begin{cases} \{0\} & \text{if } x < 0, \\ \{1\} & \text{if } x > 0, \\ [0, 1] & \text{if } x = 0. \end{cases}$$

Finally, a vector-valued function is semismooth if and only if all its component functions are semismooth; see [Ulbrich 2011, Proposition 2.10]. This implies semismoothness of (1.2.8) in finite dimensions.

1.3.2 SEMISMOOTH NEWTON METHODS IN INFINITE DIMENSIONS

In infinite dimensions, Rademacher's theorem is not available, and thus the construction above cannot be carried out. Instead of starting from Lipschitz continuous functions, we

directly demand the approximation condition to hold. We call $F : X \to Y$ *Newton differentiable* at $u \in X$ if there exists a neighborhood $N(u)$ and a mapping $G : N(u) \to \mathcal{L}(X, Y)$ with

$$\lim_{\|h\|_X \to 0} \frac{\|F(u+h) - F(u) - G(u+h)h\|_Y}{\|h\|_X} = 0. \tag{1.3.2}$$

Any $D_N F \in \{G(s) : s \in N(u)\}$ is then a *Newton derivative* at u. Note that Newton derivatives are in general not unique, and need not be elements of any generalized subdifferential. If F is Newton differentiable at u and

$$\lim_{t \to 0^+} G(u + th)h$$

exists uniformly in $\|h\|_X = 1$, then F is called *semismooth* at u. This approach to semismoothness in Banach spaces was proposed in [Hintermüller, Ito, and Kunisch 2002], based on the similar (but stronger) notion of slant differentiability introduced in [Chen, Nashed, and Qi 2000]. Related approaches to nonsmooth Newton methods in Banach spaces based on set-valued generalized derivatives were treated in [Kummer 2000] and [Ulbrich 2002]. The exposition here is adapted from [Ito and Kunisch 2008].

For Newton differentiable F, this definition leads to the semismooth Newton method

1: Choose $u^0 \in X$
2: **for** $k = 0, 1, \ldots$ **do**
3: Choose Newton derivative $D_N F(u^k)$
4: Set $u^{k+1} = u^k - D_N F(u^k)^{-1} F(u^k)$
5: **end for**

If F is Newton differentiable (in particular, if F is semismooth) at u^* with $F(u^*) = 0$ and all $D_N F(u) \in \{G(u) : u \in N(u^*)\}$ satisfy the uniform regularity condition $\|D_N F(u)\|_{\mathcal{L}(Y,X)} \leqslant C$, this iteration converges (locally) superlinearly to u^*; see, e.g., [Ito and Kunisch 2008, Theorem 8.16].

If we wish to apply a semismooth Newton method to a concrete function F such as the one in (1.2.8), we need to decide whether it is semismooth and give an explicit and computable Newton derivative. Clearly, if F is continuously Fréchet differentiable near u, then F is semismooth at u, and its Fréchet derivative $F'(u)$ is a Newton derivative (albeit not the only one). However, this cannot be extended directly to "piecewise differentiable" functions such as the pointwise max operator acting on functions in $L^p(\Omega)$. It is instructive to consider a concrete example. Take $F : L^p(\Omega) \to L^p(\Omega)$, $F(u) = \max(0, u)$. A candidate for its Newton derivative is defined by its action on $h \in L^p(\Omega)$ as

$$[G(u)h](x) = \begin{cases} 0 & u(x) < 0 \\ h(x) & u(x) > 0 \\ \delta\, h(x) & u(x) = 0 \end{cases}$$

for almost all $x \in \Omega$ and arbitrary $\delta \in \mathbb{R}$. (Since the Newton derivative coincides with the Fréchet derivative where F is continuously differentiable, we only have the freedom to choose

its value where $u(x) = 0$.) To show that the approximation condition (1.3.2) is violated at $u(x) = -|x|$ on $\Omega = (-1, 1)$ for any $1 \leqslant p < \infty$, we take the sequence

$$h_n(x) = \begin{cases} \frac{1}{n} & \text{if } |x| < \frac{1}{n}, \\ 0 & \text{otherwise,} \end{cases}$$

with $\|h_n\|_{L^p}^p = \frac{2}{n^{p+1}}$. Then, since $[F(u)](x) = \max(0, -|x|) = 0$ almost everywhere, we have

$$[F(u + h_n) - F(u) - G(u + h_n)h_n](x) = \begin{cases} -|x| & \text{if } |x| < \frac{1}{n}, \\ 0 & \text{if } |x| > \frac{1}{n}, \\ -\frac{\delta}{n} & \text{if } |x| = \frac{1}{n} \end{cases}$$

and thus

$$\|F(u + h_n) - F(u) - G(u + h_n)h_n\|_{L^p}^p = \int_{-\frac{1}{n}}^{\frac{1}{n}} |x|^p \, dx = \frac{2}{p+1} \left(\frac{1}{n}\right)^{p+1}.$$

This implies

$$\lim_{n\to\infty} \frac{\|F(u + h_n) - F(u) - G(u + h_n)h_n\|_{L^p}}{\|h_n\|_{L^p}} = \left(\frac{1}{p+1}\right)^{\frac{1}{p}} \neq 0$$

and hence that F is not semismooth from $L^p(\Omega)$ to $L^p(\Omega)$. A similar example can be constructed for $p = \infty$; see, e.g., [Ito and Kunisch 2008, Example 8.14].

On the other hand, if we consider $F : L^q(\Omega) \to L^p(\Omega)$ with $q > p$, the terms involving n^{-1} do not cancel and the approximation condition holds (at least for this choice of h_n). In fact, for arbitrary $h \in L^q(\Omega)$ one can use Hölder's inequality to create a term involving the Lebesgue measure of the support of the set where the "wrong" linearization is taken (i.e., where $\max(u(x) + h(x)) \neq \max(u(x)) + G(u(x) + h(x))h(x)$), which can be shown to go to zero as $h \to 0$; see [Hintermüller, Ito, and Kunisch 2002, Proposition 4.1]. Semismoothness in function spaces hence fundamentally requires a *norm gap*, which is another reason why approximation may be necessary to apply a semismooth Newton method to equations of type (1.2.8).

The above holds for any pointwise defined operator. If $\psi : \mathbb{R} \to \mathbb{R}$ is semismooth, the corresponding *Nemytskii operator* $\Psi : L^q(\Omega) \to L^p(\Omega)$, defined pointwise almost everywhere as

$$[\Psi(u)](x) := \psi(u(x)),$$

is semismooth if and only if $1 \leqslant p < q \leqslant \infty$, and a Newton derivative of Ψ at x, acting on h, can be taken as

$$[D_N(\Psi(u))h](x) \in \partial_C(\psi(u(x)))h(x).$$

This connection was first investigated systematically in [Ulbrich 2002]; an alternative approach which parallels the theory of Fréchet differentiability is followed in [Schiela 2008]. In particular, $F(u) = \max(0, u)$ is semismooth from $L^q(\Omega)$ to $L^p(\Omega)$ for any $q > p$, with Newton derivative

$$[D_N F(u)h](x) = \begin{cases} 0 & u(x) \leqslant 0 \\ h(x) & u(x) > 0. \end{cases}$$

This can be conveniently expressed with the help of the *characteristic function* $\chi_{\mathcal{A}}$ of the *active set* $\mathcal{A} := \{x \in \Omega : u(x) > 0\}$ (i.e., the function taking the value 1 at $x \in \mathcal{A}$ and 0 otherwise) as $D_N F(u) = \chi_{\mathcal{A}}$.

There is a useful calculus for Newton derivatives. It is straightforward to verify that the sum of two semismooth functions F_1 and F_2 is semismooth, and

$$D_N(F_1 + F_2)(u) := D_N F_1(u) + D_N F_2(u)$$

is a Newton derivative for any choice of Newton derivatives $D_N F_1$ and $D_N F_2$. We also have a chain rule: If $F : X \to Y$ is continuously Fréchet differentiable at $u \in X$ and $G : Y \to Z$ is Newton differentiable at $F(u)$, then $H := G \circ F$ is Newton differentiable at u with Newton derivative

$$D_N H(u + h) = D_N G(F(u + h))F'(u + h)$$

for any $h \in X$ sufficiently small; see [Ito and Kunisch 2008, Lemma 8.15].

A final remark. Although numerical computation almost always involves finite-dimensional problems, there is a practical reason for studying Newton methods in function spaces (besides the uniform framework and the frequently tidier notation this allows): If semismoothness and the uniform regularity condition can be verified for an infinite-dimensional problem, the respective property holds uniformly for any (conforming) discretization. In practice, this is reflected in the observation that the number of Newton iterations required to achieve a given tolerance does not increase with the fineness of the discretization. This property, called *mesh independence*, has been verified for semismooth Newton methods in [Hintermüller and Ulbrich 2004].

OPTIMAL CONTROL WITH MEASURES

2

This chapter is concerned with optimal control problems for elliptic and parabolic equations, where the controls are sought in spaces of Radon measures instead of the usual Lebesgue or Sobolev spaces. This setting is not a generalization for its own sake, but rather motivated by applications: In finite dimensional optimization, it has frequently been observed that minimizing ℓ^1-norms promotes solutions that are *sparser* than their ℓ^2-norm counterparts, i.e., that have fewer non-zero entries. This would also be desirable in the context of optimal control of partial differential equations, e.g., for the optimal placement of discrete actuators. These could be modeled as a distributed "control field", where a L^1 penalty would favor *sparse*, i.e., strongly localized controls, denoting both location and strength of the actuators; see [Stadler 2009]. Penalties of L^1 type would also be relevant in settings where the control cost is a linear function of its magnitude, e.g., representing fuel costs; see [Vossen and Maurer 2006]. However, optimal control problems in $L^1(\Omega)$ are not well-posed, since boundedness in $L^1(\Omega)$ is not sufficient for the existence of a weakly convergent subsequence. One possibility is to add additional L^2 penalties or L^∞ bounds on the control, in which case the existence of minimizers can be deduced from the Dunford–Pettis theorem; see, e.g., [Edwards 1965, Theorem 4.21.2]. This approach is followed in [Stadler 2009; Wachsmuth and Wachsmuth 2011a; Wachsmuth and Wachsmuth 2011b; Casas, Herzog, and Wachsmuth 2012]. On the other hand, we can identify $L^1(\Omega)$ with a subspace of $\mathcal{M}(\Omega)$ to obtain existence of a weak-$\star$ convergent subsequence in the latter. In this sense, the space $\mathcal{M}(\Omega)$ of Radon measures is the proper analogue of ℓ^1 for infinite-dimensional optimal control problems with sparsity constraints. A framework for the numerical solution of such problems is presented in section 2.1 for elliptic problems; its extension to parabolic problems is the topic of section 2.2.

In a similar fashion, total variation penalties favor piecewise constant controls and for that reason have attracted great interest in signal and image processing. In the context of optimal control problems, this would be relevant when the cost is proportional to changes in the control. Here, the proper setting in infinite dimensions is the space $BV(\Omega)$ of functions of bounded variation. The corresponding approach for elliptic problems is discussed in section 2.3.

2.1 ELLIPTIC PROBLEMS WITH RADON MEASURES

The challenge in the numerical solution of optimal control problems with measures arises from the non-reflexivity of the space $\mathcal{M}(\Omega)$ and the non-differentiability of its norm. However, a combination of Fenchel duality and Moreau–Yosida regularization allows approximating the optimal measure-space controls by a family of more regular controls that can be computed using a semismooth Newton method. The next section introduces this framework. Section 2.1.2 discusses the modifications necessary for restricted control and observation. An alternative to regularization is to consider a conforming discretization of the measure space, which is presented in section 2.1.3.

2.1.1 DUALITY-BASED FRAMEWORK

We consider the optimal control problem

$$\begin{cases} \min\limits_{u\in\mathcal{M}(\Omega)} \dfrac{1}{2}\|y-z\|_{L^2}^2 + \alpha\|u\|_{\mathcal{M}} \\ \text{s.t.} \quad Ay = u. \end{cases} \tag{2.1.1}$$

where $\Omega \subset \mathbb{R}^n$, $n \in \{2,3\}$, is a simply connected bounded domain with Lipschitz boundary $\partial\Omega$, and $\alpha > 0$ and $z \in L^2(\Omega)$ are given. Here, A is a linear second order elliptic differential operator taking homogeneous Dirichlet boundary conditions such that $\|A\cdot\|_{L^2}$ and $\|A^*\cdot\|_{L^2}$ are equivalent norms on

$$\mathcal{W} := H^2(\Omega) \cap H_0^1(\Omega) \hookrightarrow C_0(\Omega).$$

This is a slightly more restrictive assumption than maximal regularity for $p > n$, which can be relaxed; see sections 1.1.3 and 2.1.2. The main motivation for this restriction is to work with a standard Hilbert space for the dual problem; this will be particularly convenient when applying the framework to controls of bounded variation; see section 2.3.

Under this assumption, the equality constraint in (2.1.1) is well-posed, and the existence of a unique solution $\bar{u} \in \mathcal{M}(\Omega)$ follows from standard arguments; see Theorem 7.2.2.

To apply Fenchel duality, we take $C_0(\Omega)$ as the weak-$\star$ dual of $\mathcal{M}(\Omega)$ and set

$$\begin{aligned} &\mathcal{F} : \mathcal{M}(\Omega) \to \mathbb{R}, && \mathcal{F}(\nu) = \alpha\|\nu\|_{\mathcal{M}}, \\ &\mathcal{G} : \mathcal{W}^* \to \mathbb{R}, && \mathcal{G}(\nu) = \frac{1}{2}\left\|A^{-1}\nu - z\right\|_{L^2}^2, \\ &\Lambda : \mathcal{M}(\Omega) \to \mathcal{W}^*, && \Lambda\nu = \nu, \end{aligned}$$

i.e., Λ is the injection corresponding to the embedding $\mathcal{M}(\Omega) \hookrightarrow \mathcal{W}^*$. The conjugate of $\mathcal{G}$ can be directly calculated due to its Fréchet differentiability and the bijectivity of A; the conjugate

of $\mathcal{F}$ is given by (1.2.5). The adjoint Λ^* is the injection corresponding to the embedding $\mathcal{W} \hookrightarrow C_0(\Omega)$. Since $\mathcal{F}$ and $\mathcal{G}$ are convex, proper, and lower semicontinuous, and $\mathcal{G}$ is continuous at, e.g., $v = 0 = \Lambda v$, we can apply the Fenchel duality theorem to deduce that the dual problem

$$\begin{cases} \min\limits_{p \in \mathcal{W}} \dfrac{1}{2} \|A^* p + z\|_{L^2}^2 - \dfrac{1}{2} \|z\|_{L^2}^2 \\ \text{s.t.} \quad \|p\|_{C_0} \leqslant \alpha, \end{cases}$$

has a solution $\bar{p} \in \mathcal{W}$ which is unique by the assumption on A^*. Applying the equivalence (1.2.11) to both extremality relations in (1.2.13) then yields first order (necessary and sufficient) optimality conditions for $\bar{p}$: There exists $\bar{\lambda} := -\bar{u} \in \mathcal{M}(\Omega) \subset \mathcal{W}^*$ such that

$$\begin{cases} AA^* \bar{p} + Az + \bar{\lambda} = 0, \\ \langle \bar{\lambda}, p - \bar{p} \rangle_{\mathcal{M}, C_0} \leqslant 0, \end{cases} \tag{2.1.2}$$

holds for all $p \in \mathcal{W}$ with $\|p\|_{C_0} \leqslant \alpha$, where the first equation should be interpreted in the weak sense; see Corollary 7.2.5. From (2.1.2), we can deduce the following structural information for the Jordan decomposition $\bar{u} = \bar{u}^+ - \bar{u}^-$ of the optimal control:

$$\begin{aligned} \operatorname{supp}(\bar{u}^+) &\subset \{x \in \Omega : \bar{p}(x) = -\alpha\}, \\ \operatorname{supp}(\bar{u}^-) &\subset \{x \in \Omega : \bar{p}(x) = \alpha\}. \end{aligned}$$

This can be interpreted as a sparsity property: The optimal control $\bar{u}$ will be nonzero only on sets where the constraint on the dual variable $\bar{p}$ is active, which are typically small; and the larger the penalty α, the smaller the support of the control.

Due to the low regularity $\bar{\lambda} \in \mathcal{W}^*$, we cannot apply a semismooth Newton method directly. We therefore consider for $\gamma > 0$ the family of regularized problems

$$\begin{cases} \min\limits_{u \in L^2(\Omega)} \dfrac{1}{2} \|y - z\|_{L^2}^2 + \alpha \|u\|_{L^1} + \dfrac{1}{2\gamma} \|u\|_{L^2}^2 \\ \text{s.t.} \quad Ay = u, \end{cases} \tag{2.1.3}$$

which is strictly convex and thus has a unique solution $u_\gamma \in L^2(\Omega)$. Proceeding as above, we now set

$$\mathcal{F}_\gamma : \mathcal{M}(\Omega) \to \bar{\mathbb{R}}, \qquad \mathcal{F}_\gamma(v) = \alpha \|v\|_{\mathcal{M}} + \frac{1}{2\gamma} \|v\|_{L^2}^2,$$

which is finite if and only if $v \in L^2(\Omega)$. Direct calculation verifies that the weak-$\star$ Fenchel conjugate $\mathcal{F}_\gamma^* : C_0(\Omega) \to \bar{\mathbb{R}}$ is given by

$$\mathcal{F}_\gamma^*(v^*) = \frac{\gamma}{2} \|\max(0, v^* - \alpha)\|_{L^2}^2 + \frac{\gamma}{2} \|\min(0, v^* + \alpha)\|_{L^2}^2,$$

see Remark 7.3.2. The Fenchel duality theorem then yields the existence of a (unique) solution $p_\gamma \in \mathcal{W}$ of the dual problem

$$\min_{p \in \mathcal{W}} \frac{1}{2} \|A^* p + z\|_{L^2}^2 - \frac{1}{2} \|z\|_{L^2}^2 + \frac{\gamma}{2} \|\max(0, p - \alpha)\|_{L^2}^2 + \frac{\gamma}{2} \|\min(0, p + \alpha)\|_{L^2}^2,$$

as well as the optimality system

$$\begin{cases} AA^*p_\gamma + Az + \lambda_\gamma = 0, \\ \lambda_\gamma = \gamma \max(0, p_\gamma - \alpha)) + \gamma \min(0, p_\gamma + \alpha), \end{cases} \tag{2.1.4}$$

where $\lambda_\gamma = -u_\gamma \in L^2(\Omega)$. (The last equation should be compared with (1.2.8); the connection with the Fenchel dual of (2.1.3) justifies calling (2.1.4) a *Moreau–Yosida regularization* of (2.1.2).) As $\gamma \to \infty$, the solutions p_γ converge strongly in $\mathcal{W}$ to $\bar{p}$, while the λ_γ converge weakly-$\star$ in $\mathcal{W}^*$ to $\bar{\lambda}$; see Theorem 7.3.1.

We now consider (2.1.4) as a nonlinear equation $F(p) = 0$ for $F : \mathcal{W} \to \mathcal{W}^*$,

$$F(p) := AA^*p + Az + \gamma \max(0, p - \alpha) + \gamma \min(0, p + \alpha), \tag{2.1.5}$$

understood in the weak sense. Since $\mathcal{W} \hookrightarrow L^p(\Omega)$ for any $p > 2$, this equation is semismooth with Newton derivative

$$D_N F(p)h = AA^*h + \gamma \chi_{\{x : |p(x)| > \alpha\}} h.$$

By the assumption on A and A^*, the operator AA^* is an isometry from $\mathcal{W}$ to $\mathcal{W}^*$, which implies uniform invertibility of $D_N F(p)$ independently of p. The semismooth Newton method applied to F thus converges locally superlinearly to the solution of (2.1.5). The corresponding control $u_\gamma = -\lambda_\gamma$ can then be obtained from the second equation of (2.1.4). In practice, the basin of convergence shrinks with increasing γ; this can be remedied by computing a sequence of solutions, starting with $\gamma_0 = 1$, and using the solution u_{γ_k} as starting point for the computation of $u_{\gamma_{k+1}}$ with $\gamma_{k+1} > \gamma_k$. We shall refer to this procedure as a *continuation strategy*.

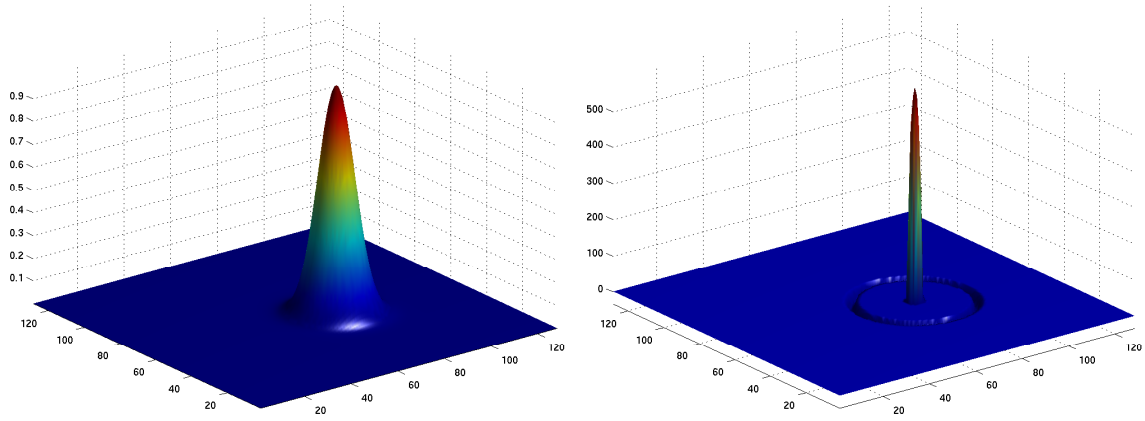
Figure 2.1 shows an example target and the corresponding optimal control u_γ for $A = -\Delta$, $\alpha = 10^{-3}$ and $\gamma = 10^7$, demonstrating the sparsity of the controls. More examples are given in section 7.4.

2.1.2 RESTRICTED CONTROL AND OBSERVATION

The optimal controls obtained from the above approach are strongly localized, and can be used as indicators for the optimal placement of point sources. In practical applications, it is often not possible to place sources in the whole computational domain; similarly, the state may need to be controlled only in a part of the domain. In case of restricted observation, however, the control-to-restricted-state mapping is no longer an isometry, and the above pure (pre)dual approach is no longer applicable. Nevertheless, useful optimality conditions of primal-dual type can still be obtained using Fenchel duality.

We thus consider the problem

$$\begin{cases} \min\limits_{u \in \mathcal{M}_\Gamma(\overline{\omega}_c)} \dfrac{1}{2} \|y|_{\omega_o} - z\|^2_{L^2(\omega_o)} + \alpha \|u\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} \\ \quad \text{s. t.} \quad Ay = \chi_{\omega_c} u, \end{cases} \tag{2.1.6}$$



(a) z (b) u_γ

Figure 2.1: Target z and corresponding optimal control u_γ for $\alpha = 10^{-3}$, $\gamma = 10^7$

where ω_o and ω_c represent the observation and control subdomains of the bounded domain $\Omega \subset \mathbb{R}^n$ with characteristic function χ_{ω_o} and χ_{ω_c}, respectively, and $z \in L^2(\omega_o)$ is given. Furthermore, $\mathcal{M}_\Gamma(\overline{\omega}_c)$ is the topological dual of $C_\Gamma(\overline{\omega}_c) := \{v \in C(\overline{\omega}_c) : v|_{\partial\omega_c \cap \Gamma} = 0\}$, where $\Gamma = \partial\Omega$ and the constraint $v|_{\partial\omega_c\cap\Gamma} = 0$ is dropped if $\partial\omega_c \cap \Gamma = \emptyset$; see section 1.1.2. Under the assumptions of section 1.1.3, the state equation is well-posed and problem (2.1.6) has a solution by standard arguments.

To define the control-to-observation mapping S_ω, we introduce for $q > n$ the canonical restriction operators

$$R_{\omega_o} : W_0^{1,q'}(\Omega) \to W^{1,q'}(\omega_o), \qquad R_{\omega_c} : W_0^{1,q}(\Omega) \to W^{1,q}(\omega_c)$$

and the injections

$$\mathcal{J}_{\omega_o} : W^{1,q'}(\omega_o) \to L^2(\omega_o), \qquad \mathcal{J}_{\omega_c} : W^{1,q}(\omega_c) \to C_\Gamma(\overline{\omega}_c)$$

and set

$$S_\omega : \mathcal{M}_\Gamma(\overline{\omega}_c) \to L^2(\omega_o), \qquad S_\omega(u) = \mathcal{J}_{\omega_o} R_{\omega_o} A^{-1} R^*_{\omega_c} \mathcal{J}^*_{\omega_c} u.$$

By construction, S_ω has the weak-$\star$ adjoint

$$S^*_\omega : L^2(\omega_o) \to C_\Gamma(\overline{\omega}_c), \qquad S^*_\omega(\varphi)) =_{\omega_c} R_{\omega_c}(A^*)^{-1} R^*_{\omega_o} \mathcal{J}^*_{\omega_o} \varphi.$$

We now apply Fenchel duality, this time setting

$$\begin{aligned} &\mathcal{F} : \mathcal{M}_\Gamma(\overline{\omega}_c) \to \mathbb{R}, && \mathcal{F}(v) = \alpha \|v\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)}, \\ &\mathcal{G} : L^2(\omega_o) \to \mathbb{R}, && \mathcal{G}(v) = \frac{1}{2}\|v - z\|^2_{L^2(\omega_o)}, \\ &\Lambda : \mathcal{M}_\Gamma(\overline{\omega}_c) \to L^2(\omega_o), && \Lambda v = S_\omega v. \end{aligned}$$

The Fenchel duality theorem now yields the existence of $\bar{q} \in L^2(\omega_o)$ satisfying

$$\begin{cases} -\bar{q} = S_\omega \bar{u} - z, \\ \bar{u} \in \partial I_{\{\|q\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha\}}(S_\omega^* \bar{q}), \end{cases}$$

where we have applied the equivalence (1.2.11) to the second relation only. Setting $\bar{p} = -S_\omega^* \bar{q} = S_\omega^*(S_\omega \bar{u} - z) \in C_\Gamma(\overline{\omega}_c)$ (i.e., introducing the adjoint state), we obtain the primal-dual optimality system for $(\bar{u}, \bar{p}) \in \mathcal{M}_\Gamma(\overline{\omega}_c) \times C_\Gamma(\overline{\omega}_c)$

$$\begin{cases} S_\omega^*(S_\omega \bar{u} - z) = \bar{p}, \\ \langle \bar{u}, \bar{p} - p \rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0, \end{cases} \tag{2.1.7}$$

for all $p \in C_\Gamma(\overline{\omega}_c)$ with $\|p\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$; see Theorem 8.2.3. Note that since S_ω is no longer bijective, we cannot solve the first equation for $\bar{u}$ as in (2.1.2).

Again, due to the low regularity of $\bar{u}$, we introduce a Moreau–Yosida regularization of (2.1.7):

$$\begin{cases} p_\gamma = S_\omega^*(S_\omega u_\gamma - z), \\ -u_\gamma = \gamma \max(0, p_\gamma - \alpha) + \gamma \min(0, p_\gamma + \alpha), \end{cases} \tag{2.1.8}$$

where S_ω is considered as an operator from $L^2(\omega_c) \to L^2(\omega_o)$. As in section 2.1.1, we deduce the existence of a unique solution $(u_\gamma, p_\gamma) \in L^2(\omega_c) \times W^{1,q}(\omega_c)$; see Theorem 8.3.1. For $\gamma \to \infty$, the family $\{u_\gamma\}_{\gamma>0}$ has a subsequence weakly-$\star$ converging to $\bar{u}$ in $\mathcal{M}_\Gamma(\overline{\omega}_c)$, and $\{p_\gamma\}_{\gamma>0}$ has a subsequence strongly converging to $\bar{p}$ in $W^{1,q}(\omega_c)$ and hence in $C_\Gamma(\overline{\omega}_c)$; see Theorem 8.3.2.

The regularized optimality system (2.1.8) can be written as an operator equation $F(u_\gamma) = 0$ for $F : L^2(\omega_c) \to L^2(\omega_c)$,

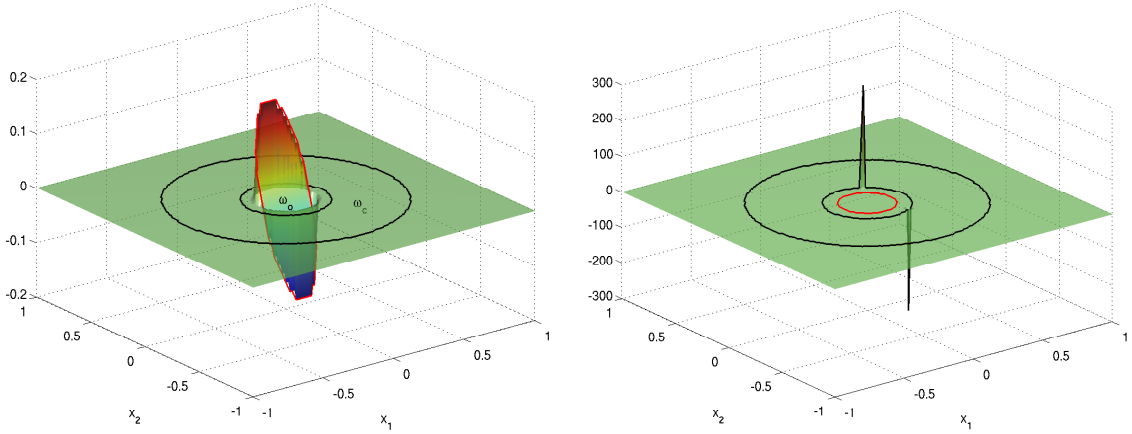
$$F(u) = u + \gamma \max(0, S_\omega^*(S_\omega u - z) - \alpha) + \gamma \min(0, S_\omega^*(S_\omega u - z) + \alpha).$$

Due to the smoothing properties of the adjoint solution operator S_ω^*, this equation is semismooth, with Newton derivative given by

$$D_N F(u)h = h + \gamma \chi_{\{x : |S_\omega^*(S_\omega u - z)(x)| > \alpha\}}(S_\omega^* S_\omega h).$$

Due to the presence of the first term and the continuity of S_ω and S_ω^*, the Newton derivatives have uniformly bounded inverses, and the semismooth Newton method converges locally superlinearly; see Theorem 8.4.1. The solution of the Newton step $D_F(u^k)\delta u = -F(u^k)$ can be computed using a matrix-free Krylov method such as GMRES, where the action of the Newton derivative on a given δu is computed by first solving the state equation $A\delta y = \chi_{\omega_c}\delta u$ followed by the adjoint equation $A^*\delta p = \chi_{\omega_o} y - z$ and setting $S_\omega^* S_\omega \delta u = \chi_{\omega_c}\delta p$. In practice, the Newton method is combined with the continuation strategy described above.

Figure 2.2 shows an example target and the corresponding optimal control u_γ for $A = -\nu\Delta - b \cdot \nabla$ with $\nu = 0.1$ and $b = (1, 0)^T$, $\alpha = 10^{-3}$, and $\gamma = 10^{12}$. It can be observed that



(a) z (b) u_γ

Figure 2.2: Target z and optimal control u_γ for $\gamma = 10^{12}$ and $\alpha = 10^{-3}$ (control domain ω_c and observation domain ω_o are shown in black and red, respectively)

the controls are concentrated on the boundary of the control domain ω_c (outlined in black) closest to the observation domain ω_o (in red). More examples can be found in section 8.5.

For some applications, it is important to ensure non-negativity of the controls; this is the case, e.g., if the controls represent light sources. This restriction can be incorporated by replacing $\mathcal{F}$ in the above framework with

$$\mathcal{F}_+ : \mathcal{M}_\Gamma(\overline{\omega}_c) \to \bar{\mathbb{R}}, \qquad \mathcal{F}_+(\nu) = \alpha \|\nu\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} + \delta_{\{\mu \in \mathcal{M}_\Gamma(\overline{\omega}_c) : \mu \geqslant 0\}}(\nu).$$

It follows from its definition that the Fenchel conjugate $\mathcal{F}_+^*$ is finite in $q \in C_\Gamma(\overline{\omega}_c)$ if and only if $q \leqslant \alpha$ everywhere, i.e.,

$$\mathcal{F}_+^* : C_\Gamma(\overline{\omega}_c) \to \bar{\mathbb{R}}, \qquad \mathcal{F}_+^*(q) = \delta_{\{\nu \in C_\Gamma(\overline{\omega}_c) : \nu \leqslant \alpha\}}(q).$$

This leads to the optimality system (recalling that $\bar{p} = -S_\omega^* \bar{q}$)

$$\begin{cases} S_\omega^*(S_\omega \bar{u} - z) = \bar{p}, \\ \langle \bar{u}, \bar{p} - p \rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0, \end{cases}$$

for all $p \in C_\Gamma(\overline{\omega}_c)$ with $p \geqslant -\alpha$, whose Moreau–Yosida regularization is

$$\begin{cases} p_\gamma = S_\omega^*(S_\omega u_\gamma - z), \\ -u_\gamma = \gamma \min(0, p_\gamma + \alpha). \end{cases}$$

The semismooth Newton method discussed above can be applied after straightforward modifications; see Remark 8.2.5 *ff.*

2.1.3 CONFORMING APPROXIMATION FRAMEWORK

In the previous sections, we have introduced a regularization in order to apply semismooth Newton methods in function spaces. Since for every $\gamma > 0$ the regularized controls u_γ are in $L^2(\Omega)$, the Newton steps can be discretized in this case using a standard finite difference or finite element method (*optimize-then-discretize*). On the other hand, if we apply a conforming discretization to problem (2.1.1), the resulting finite-dimensional optimality system will be semismooth without additional regularization. This will allow the numerical solution of the (semi-discretized) problem in measure space.

The crucial idea here is to construct a finite-dimensional subspace of $\mathcal{M}(\Omega)$ by considering a conforming discretization of $C_0(\Omega)$ and then mirroring the duality of $C_0(\Omega)$ and $\mathcal{M}(\Omega)$ on a discrete level. We start from the standard finite element approximation of continuous functions. Let $\{\mathcal{T}_h\}_{h>0}$ be a family of shape regular triangulations of $\overline{\omega}$ and let $\{x_j\}_{j=1}^{N_h}$ denote the interior nodes of the triangulation $\mathcal{T}_h$; see section 9.3 for the precise definitions. Associated to these nodes we consider the nodal basis formed by the continuous piecewise linear functions $\{e_j\}_{j=1}^{N_h}$ such that $e_j(x_i) = \delta_{ij}$ for every $1 \leqslant i, j \leqslant N_h$. We now define

$$Y_h = \left\{ y_h \in C_0(\Omega) : y_h = \sum_{j=1}^{N_h} y_j e_j, \text{ where } \{y_j\}_{j=1}^{N_h} \subset \mathbb{R} \right\}$$

endowed with the supremum norm. Since any function $y_h \in Y_h$ attains its maximum and minimum at one of the nodes, we have

$$\|y_h\|_{C_0} = \max_{1 \leqslant j \leqslant N_h} |y_j| = |\vec{y}_h|_\infty,$$

where we have identified y_h with the vector $\vec{y}_h = (y_1, \dots, y_{N_h})^T \in \mathbb{R}^{N_h}$ of its expansion coefficients, and $|\cdot|_p$ denotes the usual p-norm in $\mathbb{R}^{N_h}$. Similarly, we define

$$U_h = \left\{ u_h \in \mathcal{M}(\Omega) : u_h = \sum_{j=1}^{N_h} u_j \delta_{x_j}, \text{ where } \{u_j\}_{j=1}^{N_h} \subset \mathbb{R} \right\},$$

where δ_{x_j} is the Dirac measure corresponding to the node x_j, i.e., $\langle \delta_{x_j}, v \rangle_{\mathcal{M}, C_0} = v(x_j)$ for all $v \in C_0(\Omega)$. For $u_h \in U_h$, we have

$$\|u_h\|_{\mathcal{M}} = \sup_{\|v\|_C = 1} \sum_{j=1}^{N_h} u_j \langle \delta_{x_j}, v \rangle = \sum_{j=1}^{N_h} |u_j| = |\vec{u}_h|_1.$$

Hence endowed with these norms, U_h is the topological dual of Y_h with respect to the duality pairing

$$\langle u_h, y_h \rangle_{\mathcal{M}, C_0} = \sum_{j=1}^{N_h} u_j y_j = \vec{u}_h^T \vec{y}_h. \tag{2.1.9}$$

The natural conforming discretization of $\mathcal{M}(\Omega)$ is thus by a linear combination of Dirac measures.

To analyze the discretization of the optimal control problem, it will be useful to define the linear operators $\Pi_h : C_0(\Omega) \to Y_h$ and $\Lambda_h : \mathcal{M}(\Omega) \to U_h$ by

$$\Pi_h y = \sum_{j=1}^{N_h} y(x_j) e_j \quad \text{and} \quad \Lambda_h u = \sum_{j=1}^{N_h} \langle u, e_j \rangle \delta_{x_j}.$$

It is straightforward to verify using (2.1.9) that Λ_h is the weak-$\star$ adjoint of Π_h and that

$$\langle u, y_h \rangle_{\mathcal{M}, C_0} = \langle \Lambda_h u, y_h \rangle_{\mathcal{M}, C_0} \tag{2.1.10}$$

for all $u \in \mathcal{M}(\Omega)$ and $y_h \in Y_h$. Furthermore, $\Lambda_h u$ converges weakly-$\star$ in $\mathcal{M}(\Omega)$ to u as $h \to 0$ and $\|\Lambda_h\|_{\mathcal{L}(\mathcal{M}(\Omega), U_h)} \leqslant 1$; see Theorem 9.3.1.

We now consider the semi-discrete optimal control problem

$$\begin{cases} \min\limits_{u \in \mathcal{M}(\Omega)} \dfrac{1}{2} \|y_h - z\|^2_{L^2(\Omega_h)} + \alpha \|u\|_{\mathcal{M}(\Omega)}, \\ \text{s.t.} \quad a(y_h, v_h) = \langle u, v_h \rangle_{\mathcal{M}, C_0} \quad \text{for all } v_h \in Y_h, \end{cases} \tag{2.1.11}$$

where $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is the bilinear form associated with the operator A. Note that we have only discretized the state, but not the control; in this sense, this approach is related to the variational discretization method introduced in [Hinze 2005]. As before, we obtain the existence of an optimal control $\bar{u} \in \mathcal{M}(\Omega)$; however, since the mapping $u \mapsto y_h$ is not injective due to (2.1.10), the control is not unique. Nevertheless, by the same argument, there exists a unique $\bar{u}_h \in U_h$ such that every solution $\bar{u} \in \mathcal{M}(\Omega)$ satisfies $\Lambda_h \bar{u} = \bar{u}_h$; see Theorem 9.3.2. This means that we even if we restrict the control space to U_h, the computed control will be optimal for (2.1.11) as well.

Using the properties of Λ_h, one can show weak-$\star$ convergence of $\bar{u}_h$ to solutions of (2.1.1) in $\mathcal{M}(\Omega)$ and strong convergence of the corresponding states $\bar{y}_h$ to $\bar{y}$ in $L^2(\Omega)$ as $h \to 0$; see Theorem 9.3.5. If z is sufficiently smooth, we also obtain a rate for the latter:

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \leqslant C h^{\frac{\kappa}{2}},$$

where $\kappa = 1$ if $n = 2$ and $\kappa = 1/2$ if $n = 3$; see Theorem 9.4.2.

To compute the optimal control $\bar{u}_h$, we formulate problem (2.1.11) in terms of the coefficient vectors $\vec{u}_h$ and $\vec{y}_h$. Introducing the stiffness matrix A_h corresponding to A, we have

$$\begin{cases} \min\limits_{\vec{u}_h \in \mathbb{R}^{N_h}} \dfrac{1}{2} |\vec{y}_h - \vec{z}_h|_2^2 + \alpha |\vec{u}_h|_1, \\ \text{s.t.} \quad A_h \vec{y}_h = \vec{u}_h. \end{cases}$$

(Note that the "mass matrix" corresponding to $\langle u_h, v_h \rangle_{\mathcal{M},C_0}$ is the identity.) Applying Fenchel duality as above and introducing the optimal state vector $\bar{y}_h \in \mathbb{R}^{N_h}$, we obtain for the vectors $\bar{u}_h, \bar{p}_h \in \mathbb{R}^{N_h}$ the optimality conditions

$$\begin{cases} A_h \bar{y}_h = \bar{u}_h, \\ A_h^T \bar{p}_h = M_h(\bar{y}_h - y_{d,h}), \\ -\bar{u}_h = \max(0, -\bar{u}_h + \gamma(\bar{p}_h - \alpha)) + \min(0, -\bar{u}_h + \gamma(\bar{p}_h + \alpha)) \end{cases}$$

for any $\gamma > 0$, where M_h is the mass matrix corresponding to Y_h, and max and min should be understood componentwise in $\mathbb{R}^{N_h}$. Since we are in finite dimensions, this system can be solved using a semismooth Newton method. In practice, a continuation strategy based on a Moreau–Yosida regularization (obtained by dropping the terms $-\bar{u}_h$ on the right hand side of the last equation) is useful to compute a good starting point for the Newton iteration.

This framework can also be applied to the case of Neumann boundary controls in the space $\mathcal{M}(\Gamma)$; see section 9.5.

2.2 PARABOLIC PROBLEMS WITH RADON MEASURES

When applying the above framework to control problems involving parabolic partial differential equations, the situation is more difficult due to the low regularity of the states. For right hand sides in the space $\mathcal{M}(\Omega_T)$, where $\Omega_T := (0,T) \times \Omega$, the solution to the heat equation is only in $L^r(0,T;L^2(\Omega))$ for $r < 2$. (Using the duality technique, $r = 2$ would require $C(\Omega_T)$ regularity for solutions to the adjoint equation with right hand sides in $L^2(\Omega_T)$, which does not hold.) If we want to consider distributed L^2 tracking, we need to use controls that are more regular in time. This leads to the space $L^2(0,T;\mathcal{M}(\Omega))$ defined in section 1.1.2. The resulting controls are smooth in time, but exhibit sparsity in space; such controls can be used to model moving point sources. The spatio-temporal coupling of the corresponding control cost, however, presents a challenge for deriving numerically useful optimality conditions.

We thus consider the optimal control problem

$$\begin{cases} \min\limits_{u \in L^2(I,\mathcal{M}(\Omega))} \dfrac{1}{2}\|y - z\|^2_{L^2(\Omega_T)} + \alpha\|u\|_{L^2(\mathcal{M})}, \\ \quad \text{s.t.} \quad \partial_t y + Ay = u \quad \text{in } \Omega_T, \\ \qquad\qquad y(x,0) = y_0 \quad \text{in } \Omega \end{cases} \tag{2.2.1}$$

for given $y_0 \in L^2(\Omega)$. If A (and A^*) enjoys maximal parabolic regularity, the state equation is well-posed in $L^2(0,T;W_0^{1,q'}(\Omega))$ for all $q' \in [1, \frac{n}{n-1})$; see Theorem 10.2.2 for the case $A = -\Delta$. The control problem (2.2.1) then has a unique solution $\bar{u} \in L^2(0,T;\mathcal{M}(\Omega))$; see Theorem 10.3.2. Although the derivation of optimality conditions is deferred to later, let us

note that we can again deduce sparsity properties of the optimal control from them: For almost every $t \in [0,T]$,

$$\operatorname{supp}(\bar{u}^+(t)) \subset \{x \in \Omega : \bar{p}(x,t) = -\|\bar{p}(t)\|_{C_0}\},$$
$$\operatorname{supp}(\bar{u}^-(t)) \subset \{x \in \Omega : \bar{p}(x,t) = +\|\bar{p}(t)\|_{C_0}\},$$

where $\bar{p}$ denotes the adjoint state; see Theorem 10.3.3. This implies that the control is active where the adjoint state attains its maximum or minimum over Ω independently at each time t, and hence a purely spatial sparsity structure for the controls.

The approximation framework for $L^2(0,T;\mathcal{M}(\Omega))$ is again based on applying discrete duality to a conforming discretization of $L^2(0,T;C_0(\Omega))$. For the spatial discretization, we take the framework introduced in section 2.1.3; the temporal discretization uses piecewise constant functions. This leads to a dG(0)cG(1) discontinuous Galerkin approximation of the state equation; see, e.g., [Thomée 2006]. Specifically, we introduce a temporal grid $0 = t_0 < t_1 < \ldots < t_{N_\tau} = T$ with $\tau_k = t_k - t_{k-1}$ and set $\tau = \max_{1\leqslant k\leqslant N_\tau} \tau_k$. For every $\sigma = (\tau, h)$ we now define the discrete spaces

$$\mathcal{Y}_\sigma = \{y_\sigma \in L^2(0,T;C_0(\Omega)) : y_\sigma|_{(t_{k-1},t_k]} \in Y_h,\ 1 \leqslant k \leqslant N_\tau\},$$
$$\mathcal{U}_\sigma = \{u_\sigma \in L^2(0,T;\mathcal{M}(\Omega)) : u_\sigma|_{(t_{k-1},t_k]} \in U_h,\ 1 \leqslant k \leqslant N_\tau\}.$$

The elements $u_\sigma \in \mathcal{U}_\sigma$ and $y_\sigma \in \mathcal{Y}_\sigma$ can be represented in the form

$$u_\sigma = \sum_{k=1}^{N_\tau} u_{k,h}\chi_k \quad \text{and} \quad y_\sigma = \sum_{k=1}^{N_\tau} y_{k,h}\chi_k,$$

where χ_k is the characteristic function of the interval $(t_{k-1},t_k]$, $u_{k,h} \in U_h$, and $y_{k,h} \in Y_h$. Identifying again u_σ with the vector $\vec{u}_\sigma$ of expansion coefficients u_{kj}, we have for all $u_\sigma \in \mathcal{U}_\sigma$ that

$$\|u_\sigma\|^2_{L^2(\mathcal{M})} = \int_0^T \Big\| \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} u_{kj}\chi_k\delta_{x_j} \Big\|^2_{\mathcal{M}} dt = \sum_{k=1}^{N_\tau} \tau_k \left(\sum_{j=1}^{N_h} |u_{kj}|\right)^2 = \sum_{k=1}^{N_\tau} \tau_k |\vec{u}_k|_1^2$$

for $\vec{u}_k = (u_{k1},\ldots,u_{kN_h})^T$, and similarly for all $y_\sigma \in \mathcal{Y}_\sigma$ that

$$\|y_\sigma\|^2_{L^2(C_0)} = \sum_{k=1}^{N_\tau} \tau_k \left(\max_{1\leqslant j\leqslant N_h} |y_{kj}|\right)^2 = \sum_{k=1}^{N_\tau} \tau_k |\vec{y}_k|_\infty^2.$$

It is thus straightforward to verify that endowed with these norms, $\mathcal{U}_\sigma$ is the topological dual of $\mathcal{Y}_\sigma$ with respect to the duality pairing

$$\langle u_\sigma, y_\sigma\rangle_{L^2(\mathcal{M}),L^2(C_0)} = \sum_{k=1}^{N_\tau} \tau_k \sum_{j=1}^{N_h} u_{kj}y_{kj} = \sum_{k=1}^{N_\tau} \tau_k (\vec{u}_k^T \vec{y}_k). \tag{2.2.2}$$

As in the elliptic case, we now introduce the linear operators

$$\Phi_\sigma : L^2(0,T;\mathcal{M}(\Omega)) \to \mathcal{U}_\sigma, \qquad \Psi_\sigma : L^2(0,T;C_0(\Omega)) \to \mathcal{Y}_\sigma$$

by

$$\Phi_\sigma u = \sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \int_{I_k} \Lambda_h u(t)\,dt\,\chi_k, \qquad \Psi_\sigma y = \sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \int_{I_k} \Pi_h y(t)\,dt\,\chi_k,$$

which satisfy

$$\langle u, y_\sigma \rangle_{L^2(\mathcal{M}),L^2(C_0)} = \langle \Phi_\sigma u, y_\sigma \rangle_{L^2(\mathcal{M}),L^2(C_0)}$$

for all $u \in L^2(0,T;\mathcal{M}(\Omega))$ and $y_\sigma \in \mathcal{Y}_\sigma$. Furthermore, $\Phi_\sigma u$ converges weakly-$\star$ to u in $L^2(0,T;\mathcal{M}(\Omega))$ as $\sigma \to 0$ and $\|\Phi_\sigma\|_{\mathcal{L}(L^2(\mathcal{M}),\mathcal{U}_\sigma)} \leqslant 1$; see Theorem 10.4.2.

Since the dG(0)cg(1) discontinuous Galerkin approximation can be formulated as a variant of the implicit Euler method, the semi-discrete optimal control problem can be written as

(2.2.3)
$$\begin{cases} \min\limits_{u\in L^2(0,T;\mathcal{M}(\Omega))} \dfrac{1}{2}\|y_\sigma - z\|^2_{L^2(\Omega_T)} + \alpha\|u\|_{L^2(\mathcal{M})}, \\ \qquad \text{s.t.} \quad \left(\dfrac{y_{k,h} - y_{k-1,h}}{\tau_k}, v_h\right) + a(y_{k,h}, v_h) = \dfrac{1}{\tau_k}\displaystyle\int_{\tau_{k-1}}^{\tau_k} \langle u(t), v_h\rangle_{\mathcal{M},C_0}\,dt, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad y_{0,h} = y_0, \end{cases}$$

Again, since only the state is discretized, the solution $\bar{u}$ is not unique in $L^2(0,T;\mathcal{M}(\Omega))$, but there exists a unique $\bar{u}_\sigma \in \mathcal{U}_\sigma$ such that every solution $\bar{u} \in L^2(0,T;\mathcal{M}(\Omega))$ satisfies $\Phi_\sigma \bar{u} = \bar{u}_\sigma$. Convergence as $\sigma \to 0$, including rates, can be obtained in a similar fashion as in the elliptic case; see sections 10.4 and 10.5.

For the computation of the optimal control $\bar{u}_\sigma$, we formulate (2.2.3) in terms of the expansion coefficients $u_{k,h}$ and $y_{k,h}$. Let $N_\sigma = N_\tau \times N_h$ and identify as above $u_\sigma \in \mathcal{U}_\sigma$ with the vector $\vec{u}_\sigma = (u_{11},\dots,u_{1N_h},\dots,u_{N_\tau N_h})^T \in \mathbb{R}^{N_\sigma}$ of coefficients, and similarly $y_\sigma \in \mathcal{Y}_\sigma$ with $\vec{y}_\sigma$; see section 10.4.1. To keep the notation simple, we will omit the vector arrows from here on and fix $y_0 = 0$. Then the discrete state equation can be expressed as $L_\sigma y_\sigma = u_\sigma$ with

$$L_\sigma = \begin{pmatrix} \tau_1^{-1}M_h + A_h & 0 & 0 \\ -\tau_1^{-1}M_h & \tau_2^{-1}M_h + A_h & 0 \\ 0 & \ddots & \ddots \end{pmatrix} \in \mathbb{R}^{N_\sigma \times N_\sigma}.$$

Introducing for $v_\sigma \in \mathbb{R}^{N_\sigma}$ the vectors $v_k = (v_{k1},\dots,v_{kN_h})^T \in \mathbb{R}^{N_h}$, $1 \leqslant k \leqslant N_\tau$, the discrete optimal control problem (2.2.3) can be stated in reduced form as

$$\min_{u_\sigma\in\mathbb{R}^{N_\sigma}} \frac{1}{2}\sum_{k=1}^{N_\tau} \tau_k [L_\sigma^{-1}u_\sigma - z_\sigma]_k^T M_h [L_\sigma^{-1}u_\sigma - z_\sigma]_k + \alpha\left(\sum_{k=1}^{N_\tau} \tau_k |u_k|_1^2\right)^{1/2}.$$

We now set

$$\begin{aligned}
&\mathcal{F} : \mathbb{R}^{N_\sigma} \to \mathbb{R}, && \mathcal{F}(v) = \alpha \left(\sum_{k=1}^{N_\tau} \tau_k |v_k|_1^2 \right)^{1/2}, \\
&\mathcal{G} : \mathbb{R}^{N_\sigma} \to \mathbb{R}, && \mathcal{G}(v) = \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k (v_k - z_k)^T M_h (v_k - z_k), \\
&\Lambda : \mathbb{R}^{N_\sigma} \to \mathbb{R}^{N_\sigma}, && \Lambda v = L_\sigma^{-1} v,
\end{aligned}$$

and calculate the Fenchel conjugates with respect to the topology induced by the duality pairing (2.2.2). For $\mathcal{G}$, direct calculation yields that

$$\mathcal{G}^*(q) = \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k \left((q_k + M_h z_k)^T M_h^{-1} (q_k + M_h z_k) - z_k^T M_h z_k \right)$$

For $\mathcal{F}$, we have by example (iii) in section 1.2.1 that

$$\mathcal{F}^*(q) = \delta_{B_\alpha}(q) = \begin{cases} 0 & \text{if } \left(\sum_{k=1}^{N_\tau} \tau_k |q_k|_\infty^2 \right)^{1/2} \leqslant \alpha, \\ \infty & \text{otherwise.} \end{cases}$$

This leads to the dual problem

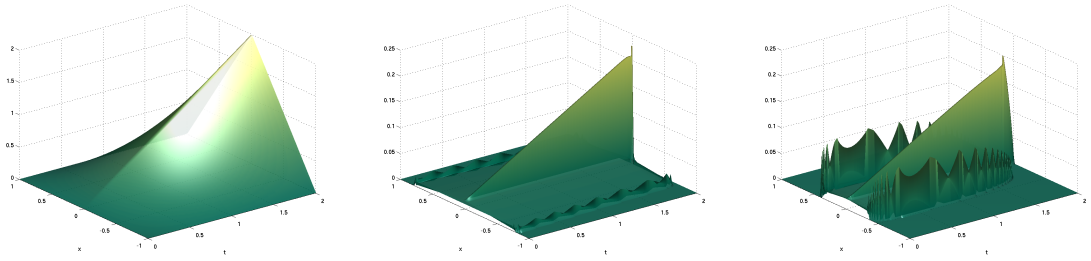
$$\min_{p_\sigma \in \mathbb{R}^{N_\sigma}} \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k ([L_\sigma^T p_\sigma]_k - M_h z_k)^T M_h^{-1} ([L_\sigma^T p_\sigma]_k - M_h z_k) + \delta_{B_\alpha}(p_\sigma).$$

Here, we cannot make direct use of the extremality relations since we have no pointwise characterization of the subdifferential of $\mathcal{F}^*$. We thus consider the following equivalent reformulation

$$\begin{cases} \min\limits_{p_\sigma \in \mathbb{R}^{N_\sigma}, c_\sigma \in \mathbb{R}^{N_\tau}} \dfrac{1}{2} \displaystyle\sum_{k=1}^{N_\tau} \tau_k ([L_\sigma^T p_\sigma]_k - M_h z_k)^T M_h^{-1} ([L_\sigma^T p_\sigma]_k - M_h z_k) \\ \qquad \text{s.t.} \quad |p_k|_\infty \leqslant c_k \text{ for all } 1 \leqslant k \leqslant N_\tau \quad \text{and} \quad \displaystyle\sum_{k=1}^{N_\tau} \tau_k c_k^2 = \alpha^2, \end{cases}$$

where $c_\sigma = (c_1, \dots, c_{N_\tau})^T \in \mathbb{R}^{N_\tau}$. Since the constraints satisfy a Maurer–Zowe regular point condition (that the feasible set contains an interior point), we obtain first order optimality conditions which can be reformulated as

$$\begin{cases} L_\sigma \bar{y}_\sigma - \bar{u}_\sigma = 0, \\ L_\sigma^T \bar{p}_\sigma - M_\sigma (\bar{y}_\sigma - z_\sigma) = 0, \\ \bar{u}_k + \max(0, -\bar{u}_k + \gamma(\bar{p}_k - \bar{c}_k)) + \min(0, -\bar{u}_k + \gamma(\bar{p}_k + \bar{c}_k)) = 0, \\ \displaystyle\sum_{j=1}^{N_h} [-\max(0, -\bar{u}_k + \gamma(\bar{p}_k - \bar{c}_k)) + \min(0, -\bar{u}_k + \gamma(\bar{p}_k + \bar{c}_k))]_j + 2\lambda \bar{c}_k = 0, \\ \displaystyle\sum_{k=1}^{N_\tau} \tau_k \bar{c}_k^2 - \alpha^2 = 0, \end{cases}$$



(a) z (b) $\bar{u}_\sigma$ ($\alpha = 10^{-3}$) (c) $\bar{u}_\sigma$ ($\alpha = 10^{-1}$)

Figure 2.3: Target z and corresponding measure space optimal controls $\bar{u}_\sigma$ for $\alpha = 10^{-3}$ and $\alpha = 10^{-1}$

where the third and fourth equations hold for all $1 \leqslant k \leqslant N_\tau$. This system can be solved by a semismooth Newton method; a good starting point again can be computed using a continuation strategy based on a Moreau–Yosida regularization of the complementarity conditions; see section 10.6.

Figure 2.3 shows an example target and the corresponding measure space optimal control $\bar{u}_\sigma$ for two different values of α. The results demonstrate the expected sparsity structure: For larger α, the controls are sparser in space, but smoother in time. More examples can be found in section 10.7.

2.3 ELLIPTIC PROBLEMS WITH FUNCTIONS OF BOUNDED VARIATION

To treat controls in the space $BV(\Omega)$, we follow the approach of section 2.1.1. We consider the problem

$$\begin{cases} \min\limits_{u \in BV(\Omega)} \dfrac{1}{2} \|y - z\|_{L^2}^2 + \alpha \|u\|_{BV} \\ \text{s.t.} \quad Ay = u \end{cases}$$

under the same assumptions as in section 2.1.1. Due to the embedding $BV(\Omega) \hookrightarrow \mathcal{M}(\Omega)$, the state equation is well-posed and existence of a unique minimizer follows again from standard arguments.

Here, we make use of the dense embedding $(C_0^\infty(\Omega))^n \hookrightarrow H^2_{\mathrm{div}}(\Omega)$ to apply Fenchel duality in a Hilbert space setting. In the following, Lebesgue spaces of vector valued functions are denoted by a blackboard bold letter corresponding to their scalar equivalent, e.g., $\mathbb{L}^2(\Omega) := (L^2(\Omega))^n$. Now let

$$H^2_{\mathrm{div}}(\Omega) := \left\{ v \in \mathbb{L}^2(\Omega) : \operatorname{div} v \in \mathcal{W},\ v \cdot \nu = 0 \text{ on } \partial\Omega \right\},$$

endowed with the norm $\|\nu\|^2_{H^2_{div}} := \|\nu\|^2_{\mathbb{L}^2} + \|\operatorname{div}\nu\|^2_{\mathcal{W}}$. We set

$$\begin{aligned} &\mathcal{F} : H^2_{div}(\Omega)^* \to \mathbb{R}, && \mathcal{F}(u) = \frac{1}{2}\left\|A^{-1}u - z\right\|^2_{L^2}, \\ &\mathcal{G} : H^2_{div}(\Omega)^* \to \mathbb{R}, && \mathcal{G}(\nu) = \alpha\|\nu\|_{\mathcal{M}^n}, \\ &\Lambda : \mathcal{W}^* \to H^2_{div}(\Omega)^*, && \Lambda\nu = D\nu, \end{aligned}$$

where D is the distributional gradient, and deduce from the Fenchel duality theorem that the dual problem

$$\begin{cases} \min\limits_{p\in H^2_{div}(\Omega)} \dfrac{1}{2}\|A^*\operatorname{div} p + z\|^2_{L^2} - \dfrac{1}{2}\|z\|^2_{L^2} \\ \quad \text{s. t.} \quad \|p\|_{(C_0)^n} \leqslant \alpha \end{cases}$$

has a solution (which however may not be unique); see Theorem 7.2.11. From the extremality relations, we obtain first order optimality conditions: There exists $\bar\lambda := D\bar u \in H^2_{div}(\Omega)^*$ such that

$$\begin{cases} \langle A^*\operatorname{div}\bar p + z, A^*\operatorname{div}\nu\rangle_{L^2} + \langle \lambda^*, \nu\rangle_{H^{2\,*}_{div}, H^2_{div}} = 0, \\ \langle \lambda^*, p - \bar p\rangle_{H^{2\,*}_{div}, H^2_{div}} \leqslant 0, \end{cases}$$

for all $\nu, p \in H^2_{div}(\Omega)$ with $\|p\|_{(C_0)^n} \leqslant \alpha$; see Corollary 7.2.12. These conditions also imply that for any $p \in H^2_{div}(\Omega)$, $p \geqslant 0$,

$$\langle D\bar u, p\rangle_{H^{2\,*}_{div}, H^2_{div}} = 0 \quad \text{if} \quad \operatorname{supp}(p) \subset \{x : |\bar p(x)|_\infty < \alpha\}.$$

This can be interpreted as a sparsity condition on the gradient of the control: The optimal control $\bar u$ will be constant on sets where the constraints on the dual variable $\bar p$ are inactive.

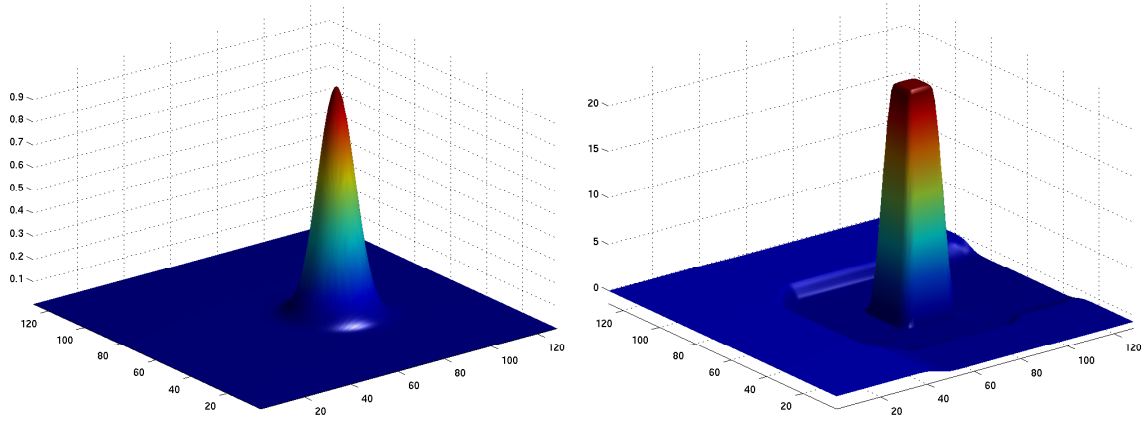
Since $\|A^*\operatorname{div} p\|_{L^2}$ is only a seminorm on $H^2_{div}(\Omega)$, we need to add additional regularization to ensure a unique solution. Since furthermore $H^2_{div}(\Omega)$ does not embed into $\mathbb{L}^q$ for $q > 2$ – which is necessary to apply a semismooth Newton method, – we set $\mathcal{H} := H^2_{div}(\Omega) \cap \mathcal{W}^n$ and consider the regularization

$$\begin{aligned} \min_{p\in\mathcal{H}} \frac{1}{2}\|A^*\operatorname{div} p + z\|^2_{L^2} + \frac{\beta}{2}\|\Delta p\|^2_{\mathbb{L}^2} - \frac{1}{2}\|z\|^2_{L^2} \\ + \frac{\gamma}{2}\|\max(0, p-\alpha)\|^2_{\mathbb{L}^2} + \frac{\gamma}{2}\|\min(0, p+\alpha)\|^2_{\mathbb{L}^2} \end{aligned}$$

with the corresponding optimality system

$$\begin{cases} \langle A^*\operatorname{div} p_\gamma + z, A^*\operatorname{div}\nu\rangle_{L^2} + \beta\langle \Delta p_\gamma, \Delta\nu\rangle_{\mathbb{L}^2} + \langle \lambda_\gamma, \nu\rangle_{H^{2\,*}_{div}, H^2_{div}} = 0, \\ \lambda_\gamma = \gamma\max(0, p_\gamma - \alpha) + \min(0, p_\gamma + \alpha), \end{cases}$$

for all $\nu \in H^2_{div}(\Omega)$, where Δ denotes the componentwise Laplacian with homogeneous Dirichlet boundary conditions, and the max, min are understood to act componentwise.



(a) z (b) u_γ

Figure 2.4: Target z and corresponding optimal control u_γ for $\alpha = 10^{-4}$, $\beta = 10^{-1}$, $\gamma = 10^7$

(Here and below, α stands for the vector $(\alpha, \dots, \alpha) \in \mathbb{R}^n$.) This system can be written as a semismooth operator equation $F(p) = 0$ for $F : \mathcal{H} \to \mathcal{H}^*$,

$$\langle F(p), v\rangle_{\mathcal{H}^*,\mathcal{H}} := \langle A^* \operatorname{div} p + z, A^* \operatorname{div} v\rangle_{L^2} + \beta \langle \Delta p, \Delta v\rangle_{\mathbb{L}^2} + \gamma \langle \max(0, p - \alpha) + \min(0, p + \alpha), v\rangle_{L^2}$$

for all $v \in \mathcal{H}$. Its Newton derivative $D_N F$ is given by its action on h as

$$\langle D_N F(p)h, v\rangle_{\mathcal{H}^*,\mathcal{H}} = \langle A^* \operatorname{div} h, A^* \operatorname{div} v\rangle_{L^2} + \beta \langle \Delta h, \Delta v\rangle_{L^2} + \gamma \left\langle \chi_{\{x:|p(x)|>\alpha\}} h, v\right\rangle_{L^2}, \tag{2.3.1}$$

where the last term is evaluated componentwise, i.e.,

$$(\chi_{\{x:|p(x)|>\alpha\}} h)_i(x) = \begin{cases} h_i(x) & \text{if } |p_i(x)| > \alpha, \\ 0, & \text{if } |p_i(x)| \leqslant \alpha, \end{cases}$$

for $i = 1, \dots, n$. Since the weak form of the Newton derivative (2.3.1) by construction defines an inner product on $\mathcal{H}$, its inverse is uniformly bounded and the semismooth Newton method converges locally superlinearly; see Theorem 7.3.5.

Figure 2.4 shows an example target and the corresponding optimal control u_γ for $A = -\Delta$, $\alpha = 10^{-4}$, $\beta = 10^{-1}$ and $\gamma = 10^7$. It can be seen that the optimal controls tend to be piecewise constant. Note that although the target possesses rotational invariance, the optimal control does not; this is due to the anisotropy of the vector norm $|\cdot|_\infty$ used in the definition of the total variation. More examples can be found in section 7.4.

OPTIMAL CONTROL WITH L^∞ FUNCTIONALS

3

This chapter treats optimal control problems where the functional to be minimized includes an L^∞ norm. We can separate such problems into two classes, depending on the role of the norm:

- Problems with *tracking terms* in L^∞ appear if the deviation from the target needs to be bounded uniformly everywhere in the domain; this amounts to a worst-case (in space) optimization problem.
- Problems with *control costs* in L^∞ lead to optimal controls of bang-bang type (i.e., the control attains its upper or lower bound almost everywhere). This is relevant in cases where the control action is virtually cost-free, but the cost of constructing the control apparatus depends on the possible range of the control.

Note that such problems are related to but different from problems with pointwise constraints on the state or the control, since the constraints themselves are subject to optimization. Compared to problems with pointwise constraints, both types of L^∞ optimal control problems have been studied relatively little in the context of partial differential equations.

The difficulty in their numerical solution arises from the fact that the subdifferential of the L^∞ norm is difficult to characterize. This can be circumvented by a reformulation: The problem

$$\min_u \|f(u)\|_{L^\infty}^2$$

can equivalently be expressed as

$$\min_{c,v,u} c^2 \quad \text{s.t.} \quad \|v\|_{L^\infty} \leqslant c,\ f(u) = v,$$

see, e.g., [Ruszczyński 2006, Example 3.39], [Grund and Rösch 2001], and [Prüfert and Schiela 2009]. In this way, optimality conditions can be obtained under standard regular point conditions. The corresponding Lagrange multipliers are only in $(L^\infty(\Omega))^*$, but semismooth Newton methods can be applied after introducing a Moreau–Yosida regularization. The squared L^∞ norm is considered in order to obtain positive definiteness of the Newton steps; note that this does not change the structural features of the problem, only the trade-off between minimizing the tracking term and the control cost for a fixed penalty parameter.

3.1 L^∞ TRACKING

We treat a slightly generalized problem

$$\begin{cases} \min\limits_{c\in\mathbb{R},u\in\mathbb{R}^m} \dfrac{c^2}{2} + \dfrac{\alpha}{2}|u|_2^2 \\ \quad \text{s.t.} \quad Ay = f + \sum\limits_{i=1}^{m} u_i \chi_{\omega_i}, \\ \qquad\quad -\beta_2 c + \psi_2 \leqslant y|_{\omega_0} \leqslant \beta_1 c + \psi_1, \end{cases} \tag{3.1.1}$$

where $\omega_i \subset \Omega$, $i = 0, \dots, m$ are open and connected sets in Ω and $f \in L^q(\Omega)$ for some $q < \max(2, n)$. Further

$$\beta_1, \beta_2 \in \mathbb{R} \text{ with } \beta_1, \beta_2 \geqslant 0 \quad \text{and} \quad \psi_1 \in L^\infty(\omega_0),\ \psi_2 \in L^\infty(\omega_0),$$

and we assume that $\beta_1 + \beta_2 > 0$ as well as $\max \psi_2 \leqslant \min \psi_1$. To simplify notation, we introduce the control operator $B : \mathbb{R}^m \to L^\infty(\Omega)$, $Bu = \sum_{i=1}^m u_i \chi_{\omega_i}$.

This problem can be given the following interpretation: A pollutant f enters the groundwater and is (diffusively and/or convectively) transported throughout the domain Ω. To minimize the concentration y of a pollutant in a town ω_0, wells $\omega_1, \dots, \omega_m$ are placed in Ω, through which a counter-agent u_i can be introduced. The problem is therefore to minimize the upper bound c in the formulation $y|_{\omega_0} \leqslant c$.

The case $\beta_1 = \beta_2 = 1$ and $\psi_1 = \psi_2 = 0$ corresponds to a problem with L^∞ tracking:

$$\begin{cases} \min\limits_{u\in\mathbb{R}^m} \dfrac{1}{2}\|y\|^2_{L^\infty(\omega_0)} + \dfrac{\alpha}{2}|u|_2^2 \\ \text{s.t.} \quad Ay = f + Bu. \end{cases} \tag{3.1.2}$$

Since the functional in (3.1.1) is continuous and radially unbounded, the problem admits a unique solution $(\bar{u}, \bar{c})$; see Proposition 11.2.3. Optimality conditions follow from a Maurer–Zowe regular point condition; see Theorem 11.3.1. However, this leads to Lagrange multipliers that are only in $(L^\infty(\Omega))^*$, which is not amenable to numerical realization.

We thus consider the Moreau–Yosida regularization of (3.1.1),

$$\begin{cases} \min\limits_{c\in\mathbb{R},u\in\mathbb{R}^m} \dfrac{c^2}{2} + \dfrac{\alpha}{2}|u|_2^2 + \dfrac{\gamma}{2}\|\max(0, y|_{\omega_0} - \beta_1 c - \psi_1)\|^2_{L^2} \\ \qquad\qquad + \dfrac{\gamma}{2}\|\min(0, y|_{\omega_0} + \beta_2 c - \psi_2)\|^2_{L^2}, \\ \quad \text{s.t.} \quad Ay = f + Bu. \end{cases}$$

This is a smooth, strictly convex optimization problem with equality constraints satisfying a Slater condition (that the linearized constraint $(y, u) \mapsto Ay - Bu$ is surjective), and hence the necessary and sufficient optimality conditions are

$$\begin{cases} \alpha u_{\gamma,i} - \langle p_\gamma, \chi_{\omega_i} \rangle = 0, \quad i = 1, \dots, m \\ c_\gamma - \langle \lambda_{\gamma,1}, \beta_1 \rangle + \langle \lambda_{\gamma,2}, \beta_2 \rangle = 0, \\ A^* p_\gamma + \tilde{\lambda}_\gamma = 0, \\ Ay_\gamma - f - Bu_\gamma = 0 \end{cases}$$

where

$$\begin{aligned} \lambda_{\gamma,1} &= \gamma \max(0, y_\gamma|_{\omega_0} - \beta_1 c - \psi_1), \\ \lambda_{\gamma,2} &= \gamma \min(0, y_\gamma|_{\omega_0} + \beta_2 c - \psi_2), \\ \lambda_\gamma &= \lambda_{\gamma,1} + \lambda_{\gamma,2}, \end{aligned}$$

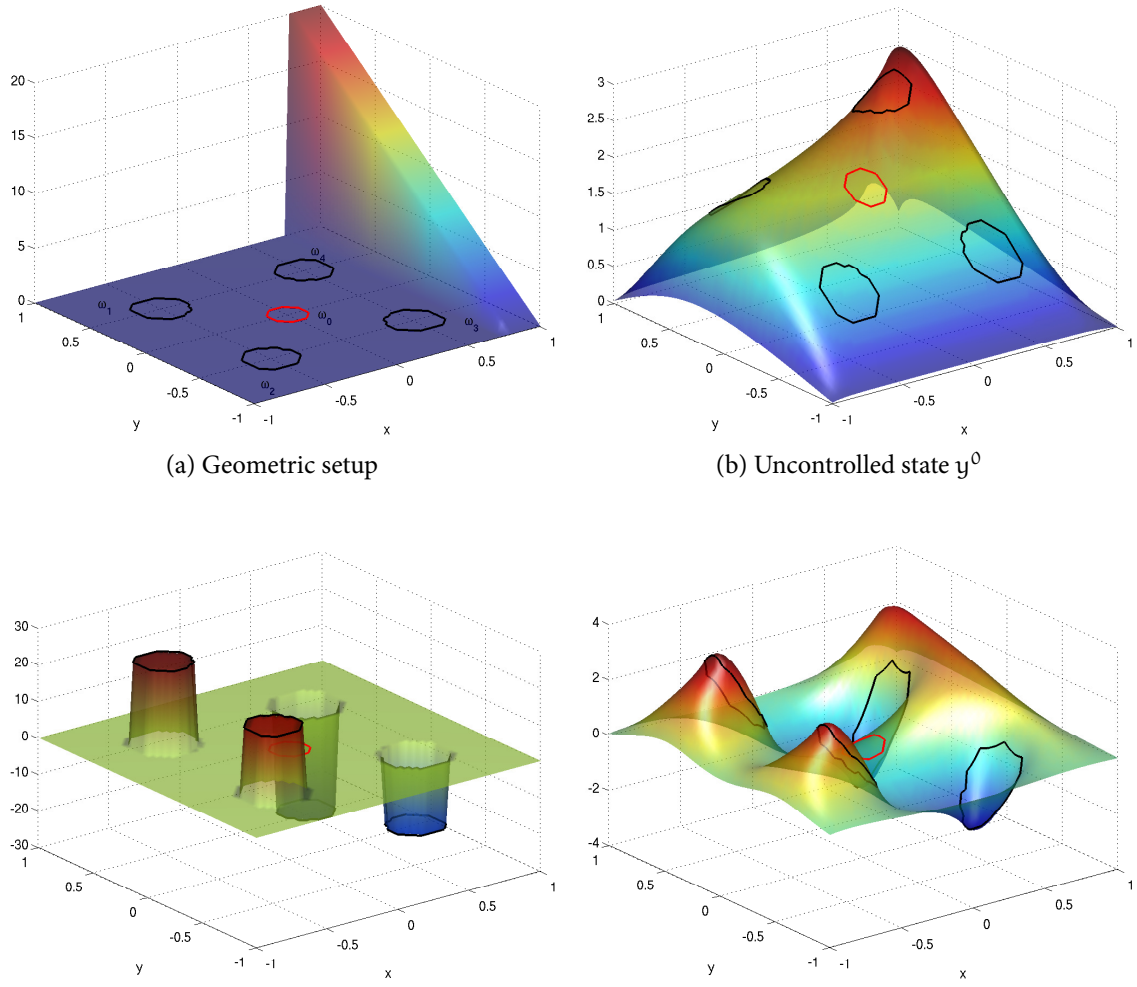
and $\tilde{\lambda}_\gamma$ denotes the extension by zero to $\Omega \setminus \omega_0$ of λ_γ. As $\gamma \to \infty$, we have convergence of $(c_\gamma, u_\gamma, y_\gamma)$ to (c^*, u^*, y^*) in $\mathbb{R} \times \mathbb{R}^m \times W^{1,q}(\Omega)$; see Proposition 11.2.3. Furthermore, we have the rate

$$\frac{1}{2}|c_\gamma - c^*|^2 + \frac{\alpha}{2}|u_\gamma - u^*|_2^2 = \mathcal{O}\left(\gamma^{-\frac{1-\theta}{1+\theta}}\right),$$

where $\theta = \frac{nq}{nq+2(q-n)}$; see Proposition 11.3.3.

Due to the regularity of the state equation and the embedding $\mathbb{R} \hookrightarrow L^\infty(\Omega)$, the optimality system, seen as an operator equation from $\mathbb{R}^m \times \mathbb{R} \times W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega)$ to $\mathbb{R}^m \times \mathbb{R} \times W^{-1,q'}(\Omega) \times W^{-1,q}(\Omega)$, is semismooth with respect to y and c. The Newton derivative can be calculated in the usual fashion; see section 11.4 (due to the necessary additional notation, it is not given here). It can be shown that the Newton derivative has a uniformly bounded inverse, implying local superlinear convergence of the semismooth Newton method; see Proposition 11.4.1. Again, the problem of local convergence can be remedied with a continuation strategy in γ.

Figure 3.1 shows a model example for problem (3.1.2): The circular observation domain ω_0 (the "town") is situated in the center of the unit square $[-1, 1]^2$. On one side, a contaminant given by the function $f = 100(1 + y)\chi_{\{x:\, x>.75\}}$ enters the computational domain. Around the town, $m = 4$ control domains ("wells") are spaced equally. We consider convective-diffusive transport, which is described by the operator $Ay = -\nu\Delta y + b \cdot \nabla y$ with $\nu = 0.1$ and $b = (-1, 0)^T$. Compared to the uncontrolled state y^0, the optimal state $\bar{y}$ is uniformly reduced within the observation domain ω_0. Since the state is bilaterally bounded, the controls opposite the support of the source are positive to avoid decreasing the lower bound. More examples are given in section 11.5.



(a) Geometric setup

(b) Uncontrolled state y^0

(c) Optimal control $\bar{u}$

(d) Optimal state $\bar{y}$

Figure 3.1: L^∞ tracking problem. The upper left plot shows the pollutant f, while the circles give the observation domain ω_0 (red) and the control domains $\omega_1, \ldots, \omega_4$ (black). The upper right plot shows the uncontrolled state $y^0 = A^{-1}f$. The lower plots show the optimal control and state, respectively, for $\alpha = 10^{-6}$.

3.2 L^∞ CONTROL COST

We consider the optimal control problem

$$\begin{cases} \min\limits_{u\in L^\infty(\Omega)} \dfrac{1}{2}\|y-z\|_{L^2}^2 + \dfrac{\alpha}{2}\|u\|_{L^\infty}^2 \\ \qquad \text{s.t.} \quad Ay = u. \end{cases}$$

Such problems are called *minimum effort problems* and have been studied in the context of ordinary partial differential equations; see, e.g., [Neustadt 1962]. They have received little attention for partial differential equations; see, e.g., [Zuazua 2007] and [Gugat and Leugering 2008] in the context of approximate and exact controllability of heat and wave equations. This may be related to the obvious difficulty arising from the non-differentiability appearing in the problem formulation.

Our approach rests on the equivalent formulation

$$\begin{cases} \min\limits_{c\in\mathbb{R},u\in L^\infty(\Omega)} \dfrac{1}{2}\|y-z\|_{L^2}^2 + \dfrac{\alpha}{2}c^2 + \delta_{B_{L^\infty}}(u) \\ \qquad \text{s.t.} \quad Ay = cu, \end{cases} \tag{3.2.1}$$

which is strictly convex due to the square of the optimal L^∞ bound c. Here and in the following, we exclude the trivial case $\bar{c} = 0$. This problem admits a unique solution $(\bar{u}, \bar{c}) \in L^\infty(\Omega) \times \mathbb{R}$, which satisfies the first order optimality conditions

$$\begin{cases} \langle -\bar{p}, u-\bar{u}\rangle_{L^2} \geqslant 0 \quad \text{for all } u \text{ with } \|u\|_{L^\infty} \leqslant 1, \\ \alpha\bar{c} - \langle \bar{u}, \bar{p}\rangle_{L^2} = 0, \\ \bar{y} - z + A^*\bar{p} = 0, \\ A\bar{y} - \bar{c}\bar{u} = 0, \end{cases}$$

with the optimal state $\bar{y} \in H_0^1(\Omega)$ and the Lagrange multiplier $\bar{p} \in H_0^1(\Omega)$. Identifying $L^1(\Omega)$ with the weak-$\star$ dual of $L^\infty(\Omega)$ and applying the equivalence (1.2.11) to the first relation (which is the explicit form of $\bar{p} \in \partial\delta_{B_{L^\infty}}(\bar{u})$), we obtain

$$\bar{u} \in \partial(\|\cdot\|_{L^1})(\bar{p}) = \operatorname{sign}(\bar{p}).$$

Note that this relation directly implies the bang-bang nature of the optimal controls. Inserting this into the remaining equations and eliminating $\bar{y}$ yields the reduced optimality system

$$\begin{cases} AA^*\bar{p} + \bar{c}\operatorname{sign}(\bar{p}) = Az, \\ \qquad \alpha\bar{c} - \|\bar{p}\|_{L^1} = 0. \end{cases}$$

Since the multi-valued sign is not differentiable even in a generalized sense, we introduce for $\beta > 0$ a Huber-type smoothing of the L^1 norm and its derivative:

$$\begin{cases} AA^*p_\beta + c_\beta \operatorname{sign}_\beta(p_\beta) = Az, \\ \qquad \alpha c_\beta - \|p_\beta\|_{L^1_\beta} = 0, \end{cases} \tag{3.2.2}$$

where we have defined

$$\|p\|_{L^1_\beta} := \int_\Omega |p(x)|_\beta \, dx, \qquad |p(x)|_\beta := \begin{cases} p(x) - \frac{\beta}{2} & \text{if } p(x) > \beta, \\ -p(x) - \frac{\beta}{2} & \text{if } p(x) < -\beta, \\ \frac{1}{2\beta} p(x)^2 & \text{if } |p(x)| \leqslant \beta, \end{cases} \tag{3.2.3}$$

and

$$\operatorname{sign}_\beta(p)(x) := \begin{cases} 1 & \text{if } p(x) > \beta, \\ -1 & \text{if } p(x) < -\beta, \\ \frac{1}{\beta} p(x) & \text{if } |p(x)| \leqslant \beta. \end{cases}$$

The optimality system (3.2.2) can also be obtained by adding the penalty $\frac{c}{2}\|u\|^2_{L^2}$ to (3.2.1) which allows deducing existence and uniqueness of solutions (c_β, p_β) to (3.2.2); see Proposition 12.3.1. The optimality condition $u_\beta = \operatorname{sign}_\beta(p_\beta)$ – corresponding to the first relation of (3.2.2) – implies that the regularized controls are of bang-zero-bang type: either $u_\beta(x) = \pm 1$ or $|u_\beta(x)| \ll 1$. As $\beta \to 0$, the family of regularized solutions (u_β, c_β) contains a subsequence that converges strongly in $L^q(\Omega) \times \mathbb{R}_+$ for any $q \in [1, \infty)$ to $(\bar u, \bar c)$; see Proposition 12.3.5. It also holds that $\beta \mapsto c_\beta \|u_\beta\|^2_{L^2}$ is monotonically decreasing; see Lemma 12.3.4.

Due to the local quadratic smoothing, $\|\cdot\|_{L^1_\beta}$ is Fréchet differentiable with derivative $\operatorname{sign}_\beta$. Furthermore, the mapping $\psi : \mathbb{R} \to \mathbb{R}$, $t \mapsto \operatorname{sign}_\beta(t)$ is continuous and piecewise differentiable and hence defines a semismooth Nemytskii operator from $L^q(\Omega)$ to $L^p(\Omega)$ for every $q > p$ with Newton derivative

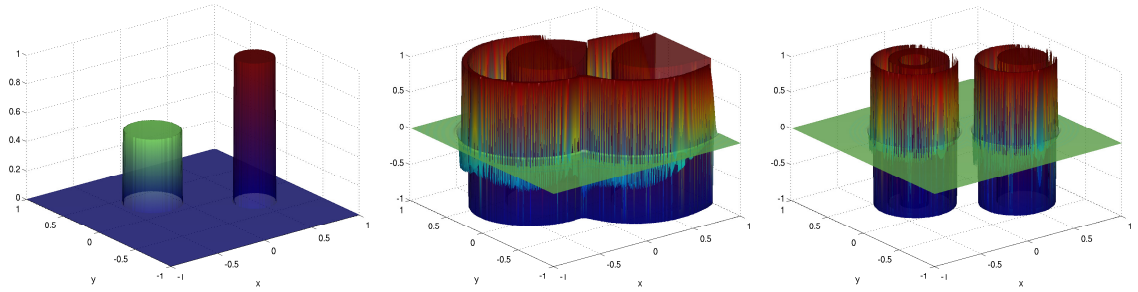
$$D_N \operatorname{sign}_\beta(v)h = \frac{1}{\beta}\chi_{\{x:\,|v(x)|\leqslant\beta\}}h.$$

for every $h \in L^q(\Omega)$. The regularity of $\bar p$ thus implies that (3.2.2) defines a semismooth operator equation $T(p, c) = 0$ for $T : \mathcal{W} \times \mathbb{R}_+ \to \mathcal{W}^* \times \mathbb{R}$. Uniform boundedness of the Newton step

$$\begin{cases} AA^*\delta p + c^k \frac{1}{\beta}\chi_{\{x:\,|v(x)|\leqslant\beta\}}\delta p + \operatorname{sign}_\beta(p^k)\delta c = -(AA^*p^k + c\operatorname{sign}_\beta(p^k) - Az), \\ \alpha\delta c - \langle \operatorname{sign}_\beta(p^k), \delta p\rangle = -(\alpha c^k - \|p^k\|_{L^1_\beta}) \end{cases}$$

once more follows from the fact that by assumption AA^* is an isometry from $\mathcal{W}$ to $\mathcal{W}^*$; see Proposition 12.4.1. This implies local superlinear convergence of the semismooth Newton method.

To combine this with a practical continuation in β requires adapting the stepsizes: If the Newton iteration did not converge for $\beta_n = \sigma\beta_{n-1}$ with $\sigma < 1$ after a given number of iterations (as monitored by the change in active sets), the result is discarded and the Newton iteration is restarted for a new $\beta_n = \sigma'\beta_{n-1}$ for $\sigma' > \sigma$. This requires an appropriate stopping rule to prevent stagnation. Here, a model function approach based on the function



(a) target z (b) control u_β ($\alpha = 5 \cdot 10^{-3}$) (c) control u_β ($\alpha = 5 \cdot 10^{-5}$)

Figure 3.2: Minimum effort problem; shown are the target z and the corresponding optimal controls u_β for $\alpha = 5 \cdot 10^{-3}$ and $\alpha = 5 \cdot 10^{-5}$

$\beta \mapsto c_\beta \|u_\beta\|_{L^2}^2$ is followed: Using the current and the previous iterates, one constructs a two-parameter interpolant $m(\beta)$ and takes $m(0)$ as an estimate of $\bar{c} \|\bar{u}\|_{L^2}^2$. If $c_{\beta_n} \|u_{\beta_n}\|_{L^2}^2 > \mu m(0)$ for a given efficiency index $\mu < 1$, the continuation is terminated; see section 12.4.2 for details.

Figure 3.2 shows results for the convection-diffusion problem from Figure 3.1 for two different values of α. The continuation strategy terminated in both cases at $\beta \approx 2 \cdot 10^{-7}$. The bang-zero-bang nature of the regularized controls can be observed clearly. Comparing the optimal L^∞ bounds $c_\beta = 0.8788$ and $c_\beta = 6.8161$, respectively, with the unscaled controls u_β demonstrates the tradeoff between magnitude and support inherent in minimum effort problems. More examples are given in section 12.5.

INVERSE PROBLEMS WITH NON-GAUSSIAN NOISE

4

In this chapter, we consider the inverse problem

$$S(u) = y^\delta$$

for a compact (possibly nonlinear) operator S, where we are interested in recovering an unknown true solution $u^\dagger$ from measurements $y^\delta = S(u^\dagger) + \xi^\delta$, where ξ^δ is some (random or deterministic) observation error of magnitude δ (often called *noise level*). Since S is compact, this problem is ill-posed even if y^δ is in the range of S and S is invertible, in the sense that the solution u^δ does not depend continuously on the data. For this reason, one usually computes an approximation u_α of u^δ by minimizing the *Tikhonov functional*

$$\mathcal{F}(S(u), y^\delta) + \alpha \mathcal{R}(u) \tag{4.0.1}$$

for an appropriate *discrepancy term* $\mathcal{F}$ and *regularization term* $\mathcal{R}$. If this problem has a unique solution u_α which converges to $u^\dagger$ as δ and α go to zero, this approach is called a *(Tikhonov) regularization* of the original inverse problem; the classical reference is [Engl, Hanke, and Neubauer 1996]. The choice of discrepancy and regularization term is crucial to achieve this, and the correct choice is intimately tied to a priori information about the problem. Specifically, the regularization term is often the (semi-)norm of an appropriate function space containing the true solution, and serves to enforce the desired structure of the solution. For example, higher order Sobolev or Lebesgue space norms yield smoother solutions, L^1-type norms promote sparsity, and total variation terms lead to piecewise constant solutions. This aspect, especially the last two examples, has attracted great interest in the last years. The discrepancy term has a similar connection with the observation error ξ^δ, and for random noise can often be deduced from statistical considerations. If ξ^δ is normally distributed, the appropriate discrepancy term is $\mathcal{F}(S(u), y^\delta) = \frac{1}{2}\left\|S(u) - y^\delta\right\|_{L^2}^2$, and this L^2 *data fitting term* is used in the vast majority of applications even if the Gaussian assumption is not justified. This is possibly due to the fact that the discrepancy terms for non-Gaussian noise may be nonsmooth, making the numerical solution challenging.

Here we will consider two such examples:

- *Impulsive noise* is characterized by significant outliers, i.e., large deviations which occur with much greater frequency than in Gaussian noise. On the other hand, not all data points are corrupted, i.e., there exist $x \in \Omega$ where $\xi^\delta(x) = 0$. Such noise frequently occurs in digital image acquisition and processing due to, e.g., malfunctioning pixels in camera sensors, faulty memory locations in hardware, or transmission in noisy channels. The assumption that there exist uncorrupted data points amounts to *sparsity* of the noise; this suggests choosing the discrepancy term

$$\mathcal{F}(S(u), y^\delta) = \left\| S(u) - y^\delta \right\|_{L^1}.$$

- *Uniform noise* can take any value between, say, $-\delta$ and δ with equal probability. Noise distributions of this type appear as statistical models of quantization errors and are therefore of relevance in any inverse problem where digital acquisition and processing of measured data plays a significant role, e.g., in the context of wireless sensor networks. Statistical considerations suggest that the choice

$$\mathcal{F}(S(u), y^\delta) = \left\| S(u) - y^\delta \right\|_{L^\infty}$$

  is appropriate in this case.

Although the choice of discrepancy terms has received less attention than the choice of regularization terms, there has been considerable recent progress in the general theory of inverse problems in Banach spaces which covers the above cases; see, e.g., [Burger and Osher 2004; Resmerita 2005; Pöschl 2009; Scherzer et al. 2009]. Efficient methods for their numerical solution, however, are less well studied.

One issue that distinguishes minimizing the Tikhonov functional (4.0.1) from optimal control problems with a similar structure is the role played by the parameter α, which governs the trade-off between attaining the data (or target) and enforcing the desired structural properties of the minimizer. In optimal control, this trade-off is usually part of the model and thus fixed in advance. For inverse problems, on the other hand, there exists an optimal choice for α, namely the one that yields a minimizer u_α that is as close as possible to $u^\dagger$. The parameter choice thus depends on y^δ and is part of the problem.

Of course, without knowing the true solution $u^\dagger$, this optimal choice is not possible. There are two classes of practical choice rules: The rules in the first class are based on the noise level δ and allow proving error estimates for u_α in terms of the noise level. One popular rule is the *Morozov discrepancy principle*, where α is chosen such that the discrepancy term is on the order of the noise level δ. On the other hand, *heuristic rules* such as the quasi-optimality principle do not require knowledge of the noise level. Although one can construct for any such rule a worst-case example for which convergence does not hold (known as the "Bakushinskii veto"; see [Bakushinskii 1984]), they are desirable in practice since the noise level may not be available. This is especially the case for non-Gaussian noise.

Here, a heuristic choice rule is proposed that involves auto-calibration of the noise level for non-Gaussian noise. Specifically, α is chosen such that the *balancing principle*

$$\sigma \mathcal{F}(S(u_\alpha), y^\delta) = \alpha \mathcal{R}(u_\alpha) \tag{4.0.2}$$

is satisfied. The parameter σ is a proportionality constant which depends on S and $\mathcal{R}$, but not on δ. The motivation is that if S is compact, $S(u)$ is smooth for any "reasonable" u, while non-Gaussian noise in general is not. If the discrepancy term is chosen appropriately, $\mathcal{F}(S(u), y^\delta)$ will therefore be a good estimate of the noise level δ for u reasonably close to $u^\dagger$. A similar assumption on the structural difference of noise and data allows proving (average-case) convergence rates for minimization-based heuristic choice rules; see [Kindermann 2011]. Of course, this is not a rigorous justification; but the rule performs quite well in practice and can be implemented using a simple fixed point iteration: For α_0 chosen sufficiently large, the iterates

$$\alpha_{k+1} := \sigma \frac{\mathcal{F}(S(u_{\alpha_k}), y^\delta)}{\mathcal{R}(u_{\alpha_k})} \tag{4.0.3}$$

define a monotonically decreasing sequence that converges to a solution of (4.0.2). This follows from the fact that by the minimizing property of u_α, the mappings $\alpha \mapsto \mathcal{F}(S(u_\alpha), y^\delta)$ and $\alpha \mapsto \mathcal{R}(u_\alpha)$ are monotonically decreasing and monotonically increasing, respectively, as $\alpha \to 0$; see [Clason, Jin, and Kunisch 2010b] and section 15.3. In practice, convergence is achieved within a few iterations.

4.1 L^1 DATA FITTING

We first consider inverse problems with data corrupted by impulsive noise. Specifically, we assume that y^δ is defined pointwise as

$$y^\delta(x) = \begin{cases} S(u^\dagger)(x) & \text{with probability } 1-d, \\ S(u^\dagger)(x) + \xi(x) & \text{with probability } d, \end{cases}$$

where $\xi(x)$ is a random variable, e.g., normally distributed with mean zero and typically large variance. The parameter $d \in (0,1)$ represents the percentage of corrupted data points. As discussed above, this implies that $y^\delta - S(u^\dagger)$ is sparse, suggesting use of the L^1 norm as discrepancy term. To avoid additional complications, we further assume in the following that $u^\dagger$ is an element of a Hilbert space $\mathcal{X}$ and correspondingly fix $\mathcal{R}(u) = \frac{1}{2}\|u\|_{\mathcal{X}}^2$.

4.1.1 LINEAR INVERSE PROBLEMS

We begin with linear inverse problems, i.e., $S(u) = Ku$ for a bounded linear operator $K : L^2(\Omega) \to L^2(\Omega)$, and consider

$$\min_{u \in L^2(\Omega)} \|Ku - y^\delta\|_{L^1(\Omega)} + \frac{\alpha}{2}\|u\|_{L^2}^2 .$$

Since $u \in L^2(\Omega)$, standard results ensure the well-posedness of this problem: There exists a unique solution u_α which depends continuously on the data y^δ, and if $\alpha \to 0$ and $\delta/\alpha \to 0$, the minimizers u_α converge to $u^\dagger$. Furthermore, under a so-called source condition (that $u^\dagger$ lies in the range of K^*, which implies additional regularity of the true solution) one obtains rates for this convergence; see section 13.2.1.

We now apply Fenchel duality. Setting

$$\begin{aligned} F &: L^2(\Omega) \to \mathbb{R}, & F(v) &= \frac{\alpha}{2}\|v\|_{L^2}^2, \\ G &: L^2(\Omega) \to \mathbb{R}, & G(v) &= \|v - y^\delta\|_{L^1}, \\ \Lambda &: L^2(\Omega) \to L^2(\Omega), & \Lambda v &= Kv, \end{aligned}$$

and computing the Fenchel conjugates (with respect to the weak duality between $L^1(\Omega)$ and $L^\infty(\Omega)$ in case of G), we obtain the dual problem

$$\min_{p \in L^2(\Omega)} \frac{1}{2\alpha}\|K^*p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2} + \delta_{B_{L^\infty}}(p), \tag{4.1.1}$$

where we have used (1.2.4) together with the transformation rules (1.2.2) (for F) and (1.2.3) (for G). The Fenchel duality theorem then yields existence of at least one minimizer $p_\alpha \in L^2(\Omega)$, which satisfies the extremality relations

$$\begin{cases} K^*p_\alpha = \alpha u_\alpha, \\ \langle Ku_\alpha - y^\delta, p_\alpha - p\rangle_{L^2} \leqslant 0, \end{cases} \tag{4.1.2}$$

for all $p \in L^2(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$, where we have again used the equivalence (1.2.11) to obtain the second relation; see Theorem 13.2.5. From the latter, we immediately deduce the following structural information:

$$\begin{aligned} \operatorname{supp}((Ku_\alpha - y^\delta)^+) &\subset \{x : p_\alpha(x) = 1\}, \\ \operatorname{supp}((Ku_\alpha - y^\delta)^-) &\subset \{x : p_\alpha(x) = -1\}. \end{aligned}$$

This can be interpreted as follows: the data is attained exactly at points where the bound constraint on p_α is inactive, and the sign of p_α is determined by the sign of the noise. The dual variable thus serves as a noise indicator.

If the inversion of K is ill-posed, solving the optimality system (4.1.2) is ill-posed as well. In addition, the low regularity $p_\alpha \in L^2(\Omega)$ prohibits application of semismooth Newton methods. We thus add a H^1 regularization term for p_α: For $\beta > 0$, we consider

$$\min_{p \in H^1(\Omega)} \frac{1}{2\alpha}\|K^*p\|_{L^2}^2 + \frac{\beta}{2}\|\nabla p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2} + \delta_{B_{L^\infty}}(p). \tag{4.1.3}$$

Under the assumption that $\ker K^* \cap \ker \nabla = \{0\}$, i.e., constant functions do not belong to the kernel of K^*, probem (4.1.3) is strictly convex and hence admits a unique solution. The family

of minimizers $\{p_\beta\}_{\beta>0}$ contains a subsequence converging weakly in $L^2(\Omega)$ to a minimizer p_α of (4.1.1); see Theorem 13.3.2. Since a regular point condition is satisfied for the box constraint, we obtain the optimality system

$$\begin{cases} \frac{1}{\alpha} KK^* p_\beta - \beta \Delta p_\beta - y^\delta + \lambda_\beta = 0, \\ \langle \lambda_\beta, p - p_\beta \rangle_{L^2} \leqslant 0, \end{cases}$$

for all $p \in H^1(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$ and a $\lambda_\beta \in (H^1(\Omega))^*$. Again, the low regularity of the Lagrange multiplier λ_β prevents a semismooth complementarity formulation, and we introduce for $\gamma > 0$ the Moreau–Yosida regularization

$$\begin{cases} \frac{1}{\alpha} KK^* p_\gamma - \beta \Delta p_\gamma - y^\delta + \lambda_\gamma = 0, \\ \lambda_\gamma = \gamma \max(0, p_\gamma - 1)) + \gamma \min(0, p_\gamma + 1). \end{cases} \tag{4.1.4}$$

One can show convergence for $(p_\gamma, \lambda_\gamma)$ as $\gamma \to \infty$ for fixed $\beta \geqslant 0$; see Theorem 13.3.1 and Theorem 13.A.1.

Due to the regularity of $p_\beta \in H^1(\Omega)$, the optimality system (4.1.4) defines a semismooth nonlinear equation $F(p) = 0$ with $F : H^1(\Omega) \to (H^1(\Omega))^*$,

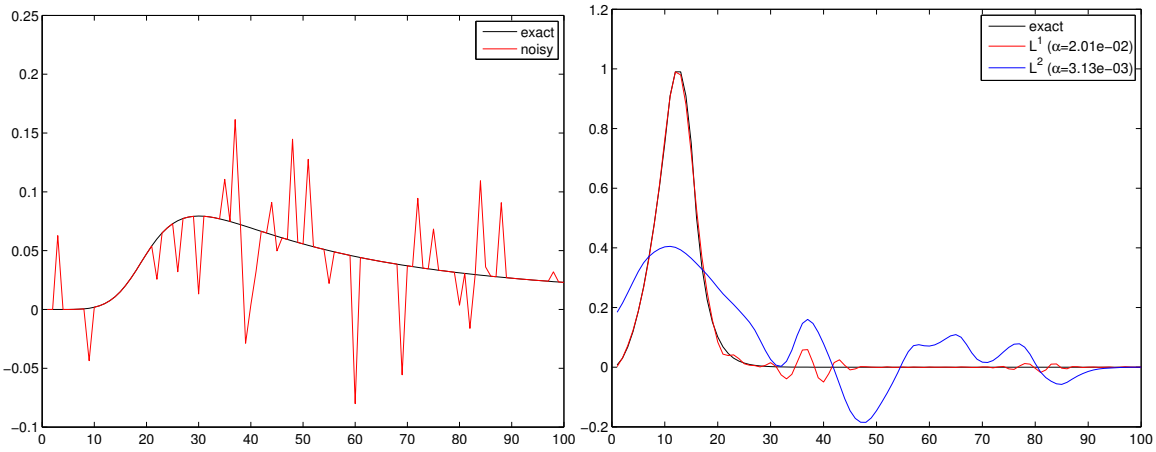
$$F(p) = \frac{1}{\alpha} KK^* p - \beta \Delta p + \gamma(\max(0, p - 1) + \min(0, p + 1)) - y^\delta,$$

which has the Newton derivative

$$D_N F(p)h = \frac{1}{\alpha} KK^* h - \beta \Delta h + \gamma \chi_{\{x : |p(x)| > 1\}} h.$$

Since by assumption the inner product $\beta \langle \nabla \cdot, \nabla \cdot \rangle_{L^2} + \frac{1}{\alpha} \langle K^* \cdot, K^* \cdot \rangle_{L^2}$ induces an equivalent norm on $H^1(\Omega)$ for any $\beta > 0$, the Lax–Milgram theorem implies uniform invertibility of $D_N F(p)$ for fixed $\beta, \gamma > 0$ and hence local superlinear convergence of the semismooth Newton method. Since β should be chosen as small as possible without making $D_N F(p)$ numerically singular, one can apply a continuation strategy which is terminated as soon as the computed p_β becomes infeasible, i.e., $\|p_\beta\|_{L^\infty} \gg 1$. In practice, the continuation strategy for β is sufficient to deal with the local convergence of the Newton method, and the parameter γ can be fixed at a large value, e.g., $\gamma = 10^9$.

Figure 4.1 shows a typical realization of noisy data for an inverse heat conduction problem with $d = 0.3$ and compares the performance of L^1 fitting with L^2 fitting, demonstrating the increased robustness of the former. For comparison, in both cases the parameter was chosen from a range of 100 logarithmically spaced values to give the lowest L^2 reconstruction error. The reconstruction with the parameter $\alpha_b = 2.239 \times 10^{-2}$ chosen according to the balancing principle (4.0.2) is very close to the optimal reconstruction with $\alpha_{opt} = 2.009 \times 10^{-2}$. Details and further one- and two-dimensional examples can be found in section 13.5.



(a) exact and noisy data $Ku^\dagger$, y^δ (b) true solution $u^\dagger$ and reconstructions u_α

Figure 4.1: Comparison of linear $L^1(\Omega)$ and $L^2(\Omega)$ fitting for inverse heat conduction problem ($d = 0.3$)

4.1.2 NONLINEAR INVERSE PROBLEMS

We now consider L^1 fitting for nonlinear inverse problems, in particular, for parameter identification for partial differential equations. Let $S : \mathcal{X} \to \mathcal{Y}$ denote the parameter-to-observation mapping, where $\mathcal{X}$ is a Hilbert space and the space $\mathcal{Y}$ compactly embeds into L^q for some $q > 2$. We also assume that y^δ is bounded almost everywhere, which is the case for data subject to impulsive noise. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are defined on the bounded domains $\omega \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$, respectively. To apply our approach, we assume that S is uniformly bounded in $U \subset \mathcal{X}$, completely continuous, and twice Fréchet differentiable with bounded first and second derivatives. These are generic assumptions in the context of parameter identification problems for partial differential equations, and are satisfied, for example, in the following situations.

- *Inverse potential problems* consist in recovering the potential u defined on $\omega = \Omega$ from noisy observational data y^δ in the domain $D = \Omega$, i.e., S maps $u \in U \subset \mathcal{X} = L^2(\Omega)$ to the solution $y \in \mathcal{Y} = H^1(\Omega)$ to

$$\begin{cases} -\Delta y + uy = f & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} = 0 & \text{on } \Gamma. \end{cases}$$

  Such problems arise in heat transfer, e.g., damping design and identifying heat radiative coefficients.

- *Inverse Robin coefficient problems* consist in recovering the Robin coefficient u defined on the boundary part $\omega = \Gamma_i$ from noisy observational data y^δ on the boundary part

$D = \Gamma_c$, i.e., S maps $u \in U \subset \mathcal{X} = L^2(\Gamma_i)$ to $y|_{\Gamma_c} \in \mathcal{Y} = H^{\frac{1}{2}}(\Gamma_c)$, where $v \mapsto v|_{\Gamma_c}$ denotes the Dirichlet trace operator and y is the solution to

$$\begin{cases} -\Delta y = 0 & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} = f & \text{on } \Gamma_c, \\ \dfrac{\partial y}{\partial n} + uy = 0 & \text{on } \Gamma_i. \end{cases}$$

This class of problems arises in corrosion detection and thermal analysis of quenching processes.

- *Inverse diffusion coefficient problems* consist in recovering the diffusion coefficient u defined on $\omega = \Omega$ from the noisy observational data y^δ in the domain $D = \Omega$, i.e., S maps $u \in U \subset \mathcal{X} = H^1(\Omega)$ to the solution $y \in \mathcal{Y} = W_0^{1,q}(\Omega)$, $q > 2$, to

$$\begin{cases} -\nabla \cdot (u \nabla y) = f & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma. \end{cases}$$

Such problems arise in estimating the permeability of underground flow and the conductivity of heat transfer.

Under the above assumptions, the Tikhonov functional

$$\min_{u \in U} \|S(u) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u\|_{\mathcal{X}}^2 \tag{4.1.5}$$

is well-posed by standard arguments, and convergence rates can be obtained under the usual source conditions on $u^\dagger$; see section 14.2.1.

Since S is nonlinear, we cannot apply the Fenchel duality theorem. We therefore proceed as in section 2.1.2. Due to the differentiability assumptions, S is strictly differentiable, and hence the Tikhonov functional is Lipschitz continuous. We can therefore use Clarke's calculus for generalized gradients to obtain for any local minimizer u_α of (4.1.5) the optimality system

$$\begin{cases} S'(u_\alpha)^* p_\alpha + \alpha j(u_\alpha - u_0) = 0, \\ \langle S(u_\alpha) - y^\delta, p - p_\alpha \rangle_{L^1, L^\infty} \leqslant 0 \quad \text{for all } \|p\|_{L^\infty} \leqslant 1, \end{cases}$$

with a $p_\alpha \in L^\infty(D)$ with $\|p_\alpha\|_{L^\infty} \leqslant 1$. Here $S'(u)^*$ denotes the adjoint of $S'(u)$ with respect to $L^2(D)$, and $j : \mathcal{X} \to \mathcal{X}^*$ is the (linear) duality mapping, i.e., $j(u) = \partial(\frac{1}{2}\|\cdot\|_{\mathcal{X}}^2)(u)$; see Theorem 14.2.7.

As in section 3.2, we now apply the equivalence (1.2.11) to the second relation (which is the explicit form of $S(u_\alpha) - y^\delta \in \partial \delta_{B_{L^\infty}}(p_\alpha)$) to obtain

$$p_\alpha \in \partial(\|\cdot\|_{L^1})(S(u_\alpha) - y^\delta) = \operatorname{sign}(S(u_\alpha) - y^\delta).$$

Inserting this into the first relation yields the necessary optimality condition

$$0 \in \alpha j(u_\alpha) + S'(u_\alpha)^*(\operatorname{sign}(S(u_\alpha) - y^\delta)).$$

Again, the non-differentiability of the multi-valued sign prevents application of Newton-type methods. We therefore consider for $\beta > 0$ the smoothed problem

$$\min_{u \in U} \|S(u) - y^\delta\|_{L^1_\beta} + \frac{\alpha}{2}\|u\|_X^2,$$

with $\|\cdot\|_{L^1_\beta}$ defined as in (3.2.3). As $\beta \to 0$, the family of minimizers $\{u_\beta\}_{\beta>0}$ contains a subsequence converging strongly in X to u_α; see Theorem 14.3.2.

Differentiability of $\|\cdot\|_{L^1_\beta}$ yields the necessary optimality condition

$$\alpha j(u_\beta) + S'(u_\beta)^*(\operatorname{sign}_\beta(S(u_\beta) - y^\delta)) = 0,$$

which defines a semismooth equation $F(u) = 0$ from X to X^* due to the linearity of the duality mapping in Hilbert spaces and the smoothing properties of S. By the chain rule for Newton derivatives we find the action of the Newton derivative on $\delta u \in X$ as

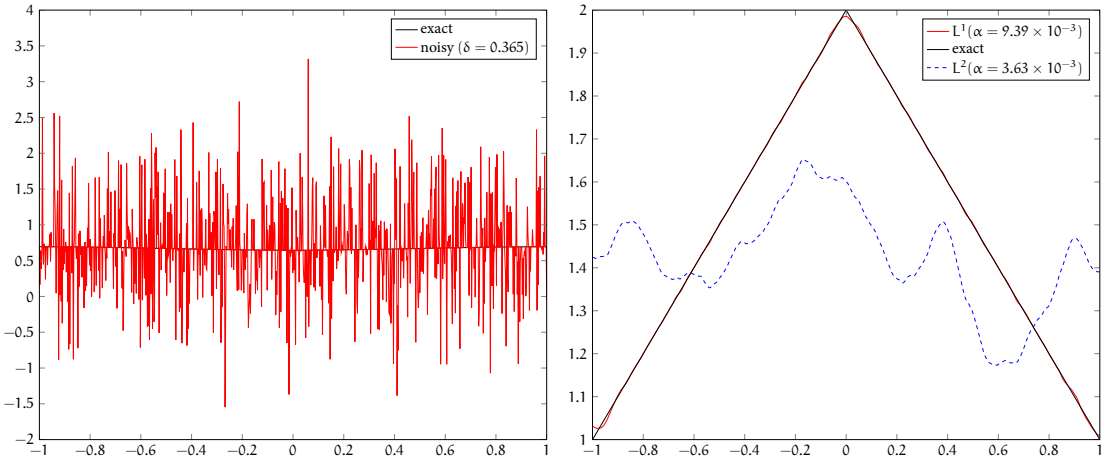
$$D_N F(u)\delta u = \alpha j'(u^k)\delta u + (S''(u^k)\delta u)^* \operatorname{sign}_\beta(S(u^k) - y^\delta)) + \beta^{-1} S'(u^k)^*(\chi_{\mathcal{J}^k} S'(u^k)\delta u).$$

Given a way to compute the action of the derivatives $S'(u)h$, $S'(u)^*h$ and $[S''(u)h]^*p$ for given u, p and h, the Newton system can be solved using a matrix-free Krylov method. In the context of parameter identification for partial differential equations, this involves solving linearized forward and adjoint equations; see section 14.A for the explicit form of these derivatives for the model problems listed above.

To deduce superlinear convergence of the semismooth Newton method, it remains to show uniform invertibility of $D_N F(u)$. Since the operator S is nonlinear and the functional is thus in general non-convex, we need to assume a local quadratic growth condition at a minimizer u_β: There exists a constant $\gamma > 0$ such that

$$\langle S''(u_\beta)(h,h), \operatorname{sign}_\beta(S(u_\beta) - y^\delta)\rangle_{L^2} + \alpha\|h\|_X^2 \geqslant \gamma\|h\|_X^2$$

holds for all $h \in X$. This is related to standard second-order sufficient optimality conditions in PDE-constrained optimization; see , e.g., [Tröltzsch 2010, Chapter 4.10]. The condition is satisfied for either large α or small noise (in the sense that $S(u_\beta) - y^\delta$ is sparse), which is a reasonable assumption for parameter fitting problems with impulsive noise. Under this condition, the inverse of $D_N F$ is uniformly bounded and thus local superlinear convergence holds; see Proposition 14.3.3. The Newton method is again combined with a continuation strategy in β, which is terminated if the semismooth Newton method failed to converge (as indicated by the change in active sets and the norm of the residual) after a given number of iterations.



(a) exact and noisy data $S(u^\dagger), y^\delta$ (b) true solution $u^\dagger$ and reconstructions u_α

Figure 4.2: Comparison of nonlinear L^1 and L^2 fitting for inverse potential problem ($d = 0.6$)

Figure 4.2 shows a typical realization of noisy data for the inverse potential problem with $d = 0.6$, and compares the performance of L^1 fitting with L^2 fitting. For L^1 fitting, the regularization parameter α was chosen according to the balancing principle; the fixed point iteration (4.0.3) converged after 4 iterations. For L^2 fitting, the parameter is chosen from a range to give the smallest reconstruction error. Again, L^1 fitting is much more robust than L^2 fitting. More results for the model problems in one and two dimensions are given in section 14.4.

4.2 L^∞ DATA FITTING

We now consider linear inverse problems with data corrupted by uniformly distributed noise. Specifically, we assume that $S(u) = Ku$ for a bounded linear operator $K : \mathcal{X} \to L^\infty(\Omega)$, and $y^\delta \in L^\infty(\Omega)$ is defined pointwise as

$$y^\delta(x) = Ku^\dagger(x) + \xi(x),$$

where $\xi(x)$ is a uniformly distributed random value in the range $[-d\,y_{\max}, d\,y_{\max}]$ for a noise parameter $d > 0$ and $y_{\max} = \|Ku^\dagger\|_\infty$. The main assumption on K is that

$$u_n \rightharpoonup u^\dagger \text{ in } \mathcal{X} \quad \text{implies} \quad Ku_n \to Ku^\dagger \text{ in } L^\infty(\Omega).$$

This holds if K is a compact operator or maps into a space compactly embedded into $L^\infty(\Omega)$ (as is commonly the case if K is the solution operator for a partial differential equation). We then consider for $p \in [1, \infty)$ the Tikhonov regularization

$$\min_{u \in \mathcal{X}} \frac{1}{p} \left\| Ku - y^\delta \right\|^p_{L^\infty(\Omega)} + \frac{\alpha}{2} \|u\|^2_{\mathcal{X}}. \tag{4.2.1}$$

Similar to the L^1 fitting case, well-posedness and convergence rates follow from standard results; see section 15.2. The reason for allowing $p > 1$ is to obtain positive definiteness of the Newton matrix; the value of p only influences the trade-off between minimizing the L^∞ norm of the residual and minimizing the norm of x, but not the relevant structural properties of the functional (in particular the geometry of the unit ball with respect to $\|\cdot\|_{L^\infty}^p$).

For the numerical solution of (4.2.1), we follow the approach presented in section 3.1. Fixing $p = 2$, we introduce the equivalent reformulation

$$\begin{cases} \min\limits_{(u,c)\in X\times\mathbb{R}} \dfrac{c^2}{2} + \dfrac{\alpha}{2}\|u\|_X^2 \\ \quad \text{s.t.} \quad \left\|Ku - y^\delta\right\|_{L^\infty(\Omega)} \leqslant c. \end{cases}$$

Since a regular point condition is satisfied for the bound constraint, there exist Lagrange multipliers $\lambda_1, \lambda_2 \in L^\infty(\Omega)^*$ with

$$\langle \lambda_1, \varphi\rangle_{L^\infty(\Omega)^*, L^\infty(\Omega)} \leqslant 0, \qquad \langle \lambda_2, \varphi\rangle_{L^\infty(\Omega)^*, L^\infty(\Omega)} \geqslant 0$$

for all $\varphi \in L^\infty(\Omega)$ with $\varphi \geqslant 0$ such that the minimizer (u_α, c_α) satisfies the optimality conditions

$$\begin{cases} \alpha j(u_\alpha) = K^*(\lambda_1 + \lambda_2), \\ \quad c_\alpha = \langle \lambda_1 - \lambda_2, -1\rangle_{L^\infty(\Omega)^*, L^\infty(\Omega)}, \\ \quad 0 = \left\langle \lambda_1, Ku_\alpha - y^\delta - c_\alpha\right\rangle_{L^\infty(\Omega)^*, L^\infty(\Omega)}, \\ \quad 0 = \left\langle \lambda_2, Ku_\alpha - y^\delta + c_\alpha\right\rangle_{L^\infty(\Omega)^*, L^\infty(\Omega)}, \end{cases}$$

see Theorem 15.4.1. Low regularity of the Lagrange multipliers once more prevents a semismooth complementarity formulation, and we introduce the Moreau–Yosida regularization

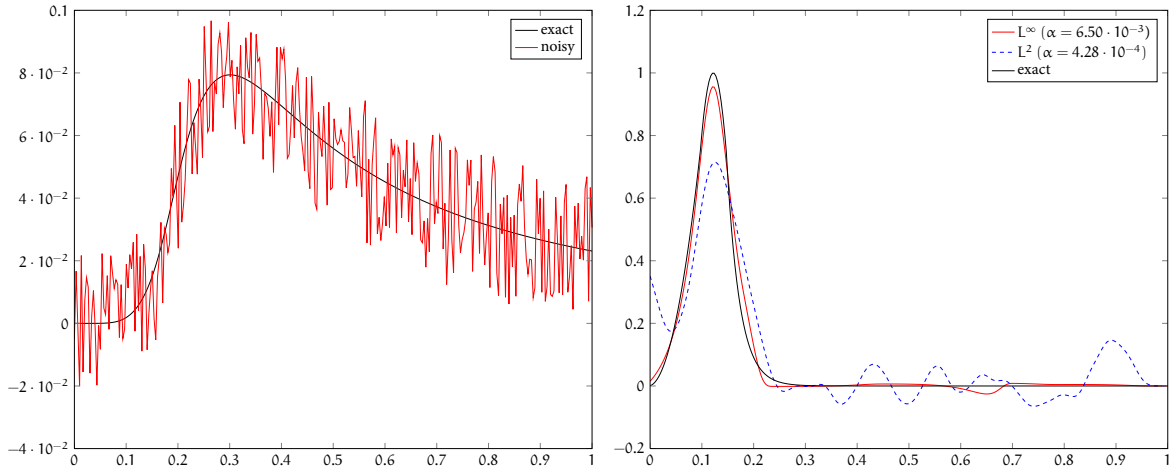
$$\min_{(u,c)\in X\times\mathbb{R}} \frac{c^2}{2} + \frac{\alpha}{2}\|u\|_X^2 + \frac{\gamma}{2}\left\|\max(0, Ku - y^\delta - c)\right\|_{L^2}^2 + \frac{\gamma}{2}\left\|\min(0, Ku - y^\delta + c)\right\|_{L^2}^2$$

which admits a unique minimizer $(u_\gamma, c_\gamma) \in X \times \mathbb{R}$. For $\gamma \to \infty$, the sequence of minimizers converges strongly to (u_α, c_α); see Theorem 15.4.2. Straightforward computation yields the (necessary and sufficient) optimality conditions

$$\begin{cases} \alpha j(u_\gamma) + \gamma K^*\left(\max(0, Ku_\gamma - y^\delta - c_\gamma) + \min(0, Ku_\gamma - y^\delta + c_\gamma)\right) = 0, \\ c_\gamma + \gamma\left\langle -\max(0, Ku_\gamma - y^\delta - c_\gamma) + \min(0, Ku_\gamma - y^\delta + c_\gamma), 1\right\rangle_{L^2(\Omega)} = 0. \end{cases}$$

This defines a semismooth equation $F(u, c) = 0$ from $X \times \mathbb{R}$ to $X^* \times \mathbb{R}$ due to the mapping properties of $K : X \to L^\infty(\Omega)$ and the embedding that maps $c \in \mathbb{R}$ to the constant function $x \mapsto c \in L^\infty(\Omega)$. The Newton derivative of F is defined by its action on $(\delta u, \delta c)$ as

$$D_N F(u, c)(\delta u, \delta c) = \begin{pmatrix} \alpha j'(u)\delta u + \gamma K^*((\chi_{\mathcal{A}_1} + \chi_{\mathcal{A}_2})K\delta u) + \gamma \delta c K^*(-\chi_{\mathcal{A}_1} + \chi_{\mathcal{A}_2}) \\ \gamma\left\langle -\chi_{\mathcal{A}_1} + \chi_{\mathcal{A}_2}, K\delta u\right\rangle_{L^2(\Omega)} + \left(1 + \gamma\left\langle \chi_{\mathcal{A}_1} + \chi_{\mathcal{A}_2}, 1\right\rangle_{L^2(\Omega)}\right)\delta c \end{pmatrix},$$



(a) exact and noisy data $Ku^\dagger, y^\delta$

(b) true solution $u^\dagger$ and reconstructions u_α

Figure 4.3: Comparison of linear L^∞ and L^2 fitting for inverse heat conduction problem ($d = 0.3$)

with

$$\mathcal{A}_1 = \{x : (Ku - y^\delta)(x) > c\}, \qquad \mathcal{A}_2 = \{x : (Ku - y^\delta)(x) < -c\}.$$

Since $\mathcal{A}_1$ and $\mathcal{A}_2$ are disjoint sets, straightforward calculation verifies that $D_N F$ is a positive definite operator independent of (u, c) and thus has a uniformly bounded inverse. This implies local superlinear convergence of the Newton method; see Theorem 15.4.4. Again, this is combined with a continuation strategy in γ.

Figure 4.3 shows a typical realization of noisy data for the inverse heat conduction problem with $d = 0.3$ and compares the performance of L^∞ fitting with L^2 fitting, demonstrating the increased robustness of the former. For L^∞ fitting, the regularization parameter α was chosen according to the balancing principle; the fixed point iteration (4.0.3) converged after 6 iterations. For L^2 fitting, the parameter is chosen from a range to give the smallest reconstruction error. More details and two-dimensional examples for deterministic quantization errors can be found in section 15.5.

APPLICATIONS IN BIOMEDICAL IMAGING

5

The final chapter of this part illustrates the relevance of non-reflexive Banach spaces in real-world applications with two examples from biomedical imaging: Optimal light source placement in *diffuse optical imaging*, which can be formulated as an optimal control problem in the space of Radon measures, and image reconstruction in *parallel magnetic resonance imaging*, which amounts to a bilinear inverse problem where regularization terms of total variation type can lead to improved reconstruction quality. These works arose from cooperations with the Institute of Medical Engineering of the TU Graz in the framework of the SFB "Mathematical Optimization and Applications in Biomedical Sciences".

5.1 DIFFUSE OPTICAL IMAGING

Fluorescent diffuse optical tomography is an imaging methodology where a biological sample is illuminated by near-infrared light emitted from point-like light sources (so-called *optodes*) such as optical fibers or lasers after a fluorescent marker has been introduced which selectively binds to inclusions to be detected, e.g., cancer cells. The light then diffuses through the tissue while being scattered and absorbed by inhomogeneities including the markers. The photons absorbed by the latter are reemitted at a different wavelength and are then transported back to the surface, where they are captured and used to reconstruct a tomographic image of the marked tissue. However, the diffusive nature of the photon transport makes this task challenging. The reconstruction would be facilitated if the photon density of the illuminating light can be made homogeneous within a region of interest, so that any variation in contrast must be due to the marker distribution. A similar problem occurs in photodynamic cancer therapy, where instead of a fluorescent marker, a photo-activable cytotoxin is used to selectively destroy cancer cells, and homogeneous illumination is critical to avoid local under- an overdoses.

Due to the complex surface shape of biological samples, the configuration of optodes required to achieve this is far from obvious. Previously published methods were based on a discrete approach, where a (large) set of possible locations was specified beforehand, from which the best locations are chosen such that a certain performance criterion is minimized; this

amounts to a combinatorial problem with exponential complexity. The corresponding optimal source magnitudes would then be computed in a second step. In contrast, the measure space optimal control approach described in section 2.1 yields both location and magnitude of point sources in a single step, without requiring an initial feasible configuration or specification of the desired number of optodes.

The model of the steady state of light propagation in a scattering medium is based on the diffusion approximation of the radiative transfer equation. This leads to a stationary elliptic partial differential equation for the photon distribution $\varphi \in H^1(\Omega)$,

$$\begin{cases} -\nabla \cdot (\kappa(x)\nabla\varphi(x)) + \mu_a(x)\varphi(x) = q(x) & \text{in } \Omega, \\ \kappa(x)\, \nu(x) \cdot \nabla\varphi(x) + \rho\varphi(x) = 0 & \text{on } \Gamma. \end{cases} \tag{5.1.1}$$

The geometry of the sample is given by the domain $\Omega \subset \mathbb{R}^n$, $n \in \{2, 3\}$, with boundary Γ whose outward normal vector is denoted by ν. The medium is characterized by the absorption coefficient μ_a, the reduced scattering coefficient μ_s', and the diffusion coefficient $\kappa = n\,[(\mu_a + \mu_s')]^{-1}$. The coefficient ρ models the reflection of a part of the photons at the boundary due to a mismatch in the index of refraction. Finally, the source term q models the light emission from the optodes.

The objective is to minimize the deviation from a constant illumination z in an observation region $\omega_o \subset\subset \Omega$. Due to the linearity of the forward problem, we can take $z = 1$ without loss of generality. In addition, we restrict the possible light source locations to a control region $\omega_c \subset\subset \Omega$ and enforce non-negativity of the source term q (which represents the optodes). This leads to the optimization problem

$$\begin{cases} \min\limits_{q \in \mathcal{M}(\overline{\omega}_c)} \dfrac{1}{2}\|\varphi|_{\omega_o} - z\|^2_{L^2(\omega_o)} + \alpha\|q\|_{\mathcal{M}(\overline{\omega}_c)} + \delta_{\{\mu:\, \mu \geqslant 0\}}(q) \\ \text{s. t.} \quad (5.1.1). \end{cases}$$

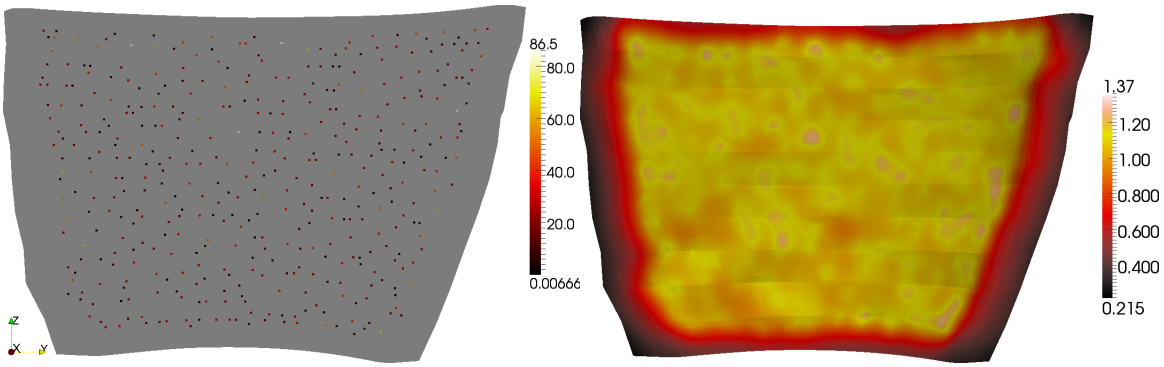
Applying the approach of section 2.1.2 yields for $\gamma > 0$ the family of optimality conditions

$$q_\gamma + \gamma \min(0, p_\gamma + \alpha) = 0,$$

where p_γ is again the adjoint state, i.e., the solution of the (selfadjoint) differential equation (5.1.1) with right hand side $\varphi_\gamma - z$. Instead of a matrix-free Krylov method, we first apply the discretization described in section 2.1.3 to obtain the discrete optimality system

$$\begin{cases} A_h\varphi_h - q_h = 0, \\ -M_o\varphi_h + A_h^T p_h = -M_o z, \\ q_h + \gamma \min(0, p_h|_{\omega_c} + \alpha) = 0, \end{cases}$$

where A_h denotes the stiffness matrix corresponding to (5.1.1), M_o the restricted mass matrix with entries $\langle e_i, e_j\rangle_{L^2(\omega_o)}$, and the last equation is to be understood componentwise in $\mathbb{R}^{N_h}$. Due to the discretization chosen for q_γ, we can eliminate the control using the first



(a) optodes (b) photon density

Figure 5.1: Optode positions and magnitudes (left) and photon densities (right, normalized to unit mean) for $\alpha = 0.8$

equation and apply a semismooth Newton method. In each Newton step, we have to solve for (φ^{k+1}, p^{k+1}) the block system

$$\begin{pmatrix} A_h & D_k \\ -M_o & A_h^T \end{pmatrix} \begin{pmatrix} \varphi^{k+1} \\ p^{k+1} \end{pmatrix} = \begin{pmatrix} -\alpha d^k \\ -M_o z \end{pmatrix},$$

where D_k is a diagonal matrix with the entries of the vector d^k,

$$d_j^k = \begin{cases} \gamma & \text{if } (p^k|_{\omega_c})_j < -\alpha, \\ 0 & \text{else,} \end{cases}$$

on the diagonal. This is once more combined with a continuation strategy in γ.

The feasibility of this approach is demonstrated for the practically relevant problem of designing light applicators in photodynamic therapy for mesotheliomas in the intrathoracic cavity, i.e., light-activated destruction of cancer cells on the surface of the lung. The goal is to achieve homogeneous illumination of the target region using a flexible diffuser with embedded light sources, taking into account the curvature of the surface. This is illustrated using a realistic three-dimensional diffuser model constructed from a CT scan of a human thorax, where the observation region ω_o is defined as the outer and inner surface of the model and ω_c is an interior manifold equidistant from both. Figure 5.1 shows the computed optimal light sources q_γ and resulting photon density φ_γ for $\alpha = 0.8$. The resulting deviation from a homogeneous illumination (with a coefficient of variation of 0.204) would be acceptable in practice. More results and quantitative evaluations are given in section 16.4.

5.2 PARALLEL MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging (MRI) is a medical imaging method that employs radio pulse echoes to measure the density of hydrogen atoms in a sample, which allows the discrimination

of different types of tissue. The spatial information is encoded, using a combination of spatially varying magnetic fields, in the phase and frequency of the time-dependent echo, which is then measured by coils surrounding the patient. A Fourier transform of the recorded signal will therefore yield an image of the sample. Mathematically, MRI can thus be thought of as direct measurement of the Fourier coefficients of the image. One of the major drawbacks of MRI in current practice is the speed of the image acquisition, since in principle each (discrete) Fourier coefficient $k(i,j)$ has to be acquired separately: Each coordinate pair (i,j) needs to be encoded by the gradient fields and measured by a separate radio pulse. By employing a phase encoding instead of a frequency encoding for one coordinate, a "line" of Fourier coefficients $k(i,\cdot)$ can be read out in parallel for every radio excitation. The standard approach for further speeding up the process acquires only a subset of these lines (e.g., every second, $k(2i,\cdot)$, or every fourth, $k(4i,\cdot)$); other strategies (so-called *trajectories*) such as sampling along radial and spiral "lines" are possible.

This, however, leads to aliasing, as the signal is now sampled below the Nyquist frequency, resulting in visible image corruption. As a remedy, parallel magnetic resonance imaging (PMRI) measures the radio echo using multiple independent coils, which are usually placed in a circle around the patient; in this way one hopes to compensate for the lost information. Since these coils have only limited aperture compared to a single coil, the resulting measurements are non-uniformly modulated. It is, therefore, necessary to recover both the missing Fourier coefficients and the unknown modulations (the so-called *sensitivities*) from a set of modulated and aliased coil images.

Mathematically, PMRI can be formulated as a nonlinear inverse problem where the sampling operator $\mathcal{F}_S$ (e.g., Fourier transform followed by multiplication with a binary mask) and the correspondingly acquired Fourier coefficients $g = (g_1, \dots, g_N)^T$ from N receiver coils are given, and the spin density u and the unknown (or not perfectly known) set of coil sensitivities $c = (c_1, \dots, c_N)^T$ have to be found such that

$$F(u,c) := (\mathcal{F}_S(u \cdot c_1), \dots, \mathcal{F}_S(u \cdot c_N))^T = g$$

holds. Since this problem is bilinear, Newton-type methods are not applicable. A standard approach for nonlinear inverse problems, the *iteratively regularized Gauß–Newton* (IRGN) method, consists in solving a sequence of quadratic problems obtained by linearization. Specifically, one computes in each step k for given $x^k := (u^k, c^k)$ the solution $\delta x := (\delta u, \delta c)$ to the minimization problem

$$\min_{\delta x} \frac{1}{2}\|F'(x^k)\delta x + F(x^k) - g\|^2 + \frac{\alpha_k}{2}\mathcal{W}(c^k + \delta c) + \beta_k \mathcal{R}(u^k + \delta u)$$

for given $\alpha_k, \beta_k > 0$, and then sets $x^{k+1} := x^k + \delta x$, $\alpha_{k+1} := q_\alpha \alpha_k$ and $\beta_{k+1} := q_\beta \beta_k$ with $0 < q_\alpha, q_\beta < 1$. Here, the term $\mathcal{W}(c) = \|w \cdot \mathcal{F}c\|^2$ is a penalty on the high-frequency Fourier coefficients of the sensitivities (enforcing smoothness of the modulations) and $\mathcal{R}$ is a regularization term for the image. In this work total variation-type penalties are used to prevent noise amplification as $\alpha_k, \beta_k \to 0$. For the solution of the quadratic subproblems, a

first order method is chosen which requires only application of Fourier transforms and pointwise operations and hence can be implemented efficiently on modern multi-core hardware such as graphics processing units. Inserting the characterization (1.1.2) of the total variation seminorm yields the nonsmooth convex-concave saddle-point problem

$$\min_{\delta x} \max_{p} \frac{1}{2}\|F'(x^k)\delta x + F(x^k) - g\|^2 + \frac{\alpha_k}{2}\mathcal{W}(c^k + \delta c) + \langle u^k + \delta u, -\operatorname{div} p\rangle + \delta_{C_{\beta_k}}(p)$$

with

$$C_\beta = \left\{p \in L^2(\Omega;\mathbb{C}^2) : \operatorname{div} p \in L^2(\Omega;\mathbb{C}),\ |p(x)|_2 \leqslant \beta \text{ for almost all } x \in \Omega\right\},$$

which is solved using a projected primal-dual extra-gradient method adapted from [Chambolle and Pock 2010]; see Algorithm 17.1. A similar approach can be applied when $\mathcal{R}$ is the *total generalized variation*, a higher order total variation-type penalty that promotes piecewise affine solutions; see [Bredies, Kunisch, and Pock 2010] and Algorithm 17.2.

Figure 5.2 shows reconstructions of real-time images of a beating heart using radial sampling with 25, 12 and 19 acquired lines (corresponding to undersampling of approximately 8.0, 9.6 and 10.6 times below the Nyquist limit, respectively), comparing L^2 regularization (IRGN) with total variation regularization (IRGN TV). The ability of the latter to prevent the noise amplification occuring in the former is evident. Additional results for different sampling strategies and total generalized variation are given in section 17.4.

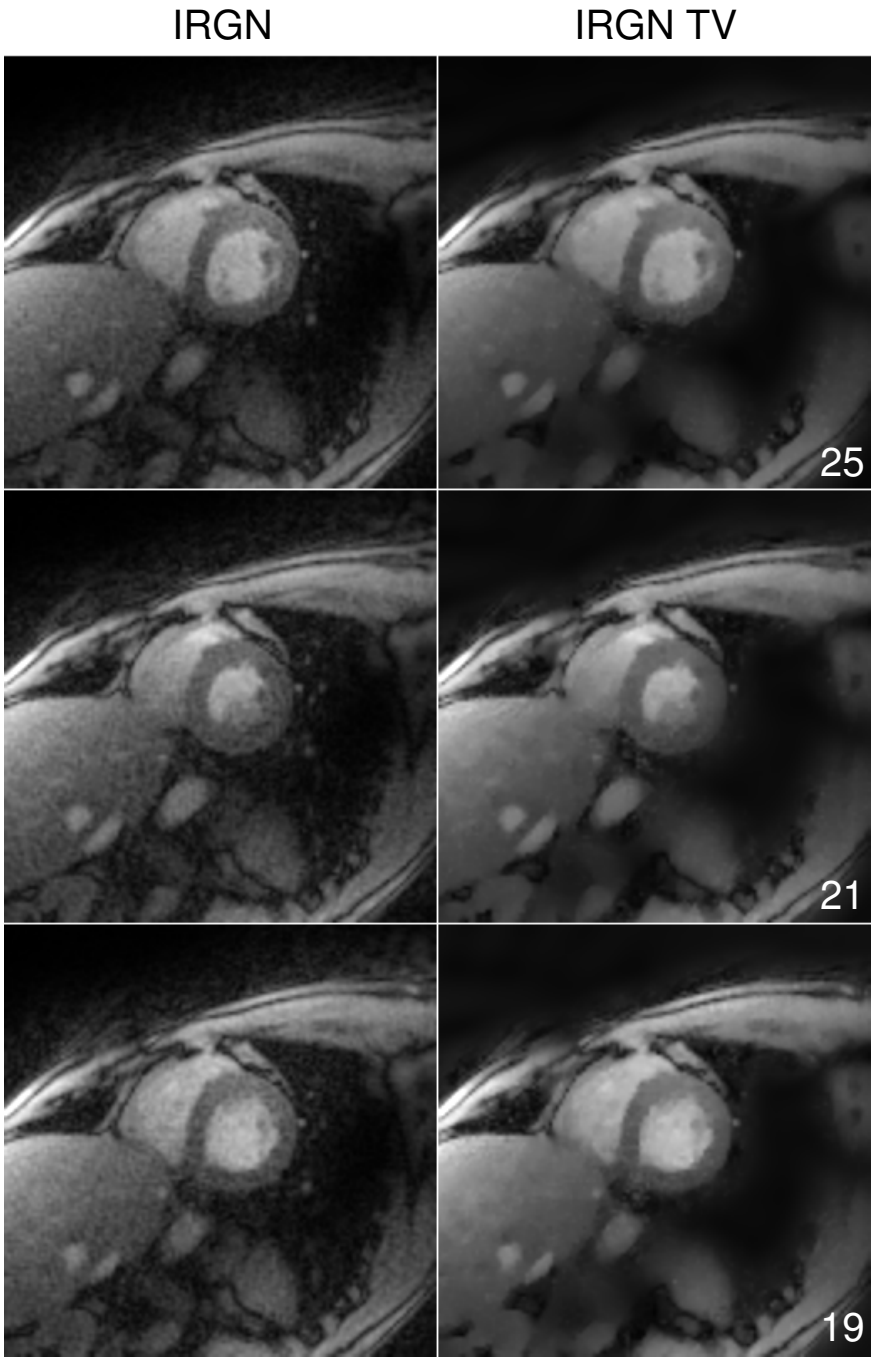


Figure 5.2: Reconstructions of real-time images of a beating heart from 25, 12 and 19 acquired radial lines with L^2 regularization (IRGN) and with total variation regularization (IRGN TV)

OUTLOOK

6

The results described in this thesis can be extended in many directions. Clearly, nonlinear optimal control problems, both in measure spaces and with L^∞ functionals, can be considered. Of interest would be nonlinear partial differential equations with controls on the right hand side as well as control of linear equations by lower order coefficients, e.g., potential terms. Also of practical relevance are controls with sparsity properties not in physical space but in a transform space, e.g., optimal control of the wave equation using controls that are a superposition of few frequencies. This problem can be treated using the approach proposed in section 2.3, with the Fourier transform in place of the distributional gradient. Tracking terms of L^1 and L^∞ type also appear as appropriate distance functionals in the context of optimal control of probability density functions (there known as *Kantorovich* and *Kolmogorov* distances, respectively).

For inverse problems, it would be of interest to extend the L^1 data fitting to the case of mixed noise (e.g., combined Gaussian–impulsive or Cauchy-distributed noise). For such noise, the regularization of the L^1 norm introduced in section 4.1.2 turns out to be the statistically appropriate discrepancy term (known as the *Huber norm*); it remains to replace the continuation strategy in β with a suitable parameter choice method. Still missing is a rigorous stochastic framework for impulsive noise models in function spaces, which would lead to inverse problems with data fitting in measure spaces. Besides applying the L^∞ fitting approach to nonlinear parameter identification problems, it would be possible to combine it with the approach for L^1-type regularization terms to obtain a numerical method for the so-called *Dantzig selector*. Of course, any method for solving inverse problems has to be tested with real-world data; hence applying the developed methods to practical inverse problems is of great interest.

A long-term goal is to combine the measure space approach for source placement with sensitivity maximization techniques to solve optimal light source and sensor placement problems in optical tomography. In parallel magnetic resonance imaging, formulating the image reconstruction problem as an inverse problem allows the inclusion of physiological parameters, either as additional penalties or as unknown parameters to be reconstructed.

These extensions are the topic of future research.

Part II

OPTIMAL CONTROL WITH MEASURES

A DUALITY-BASED APPROACH TO ELLIPTIC CONTROL PROBLEMS IN NON-REFLEXIVE BANACH SPACES

7

ABSTRACT

Convex duality is a powerful framework for solving nonsmooth optimal control problems. However, for problems set in non-reflexive Banach spaces such as $L^1(\Omega)$ or $BV(\Omega)$, the dual problem is formulated in a space which has difficult measure theoretic structure. The predual problem, on the other hand, can be formulated in a Hilbert space and entails the minimization of a smooth functional with box constraints, for which efficient numerical methods exist. In this work, elliptic control problems with measures and functions of bounded variation as controls are considered. Existence and uniqueness of the corresponding predual problems are discussed, as is the solution of the optimality systems by a semismooth Newton method. Numerical examples illustrate the structural differences in the optimal controls in these Banach spaces, compared to those obtained in corresponding Hilbert space settings.

7.1 INTRODUCTION

This work is concerned with the study of the optimal control problem

$$(\mathcal{P}) \qquad \begin{cases} \min\limits_{u \in \mathcal{X}} \dfrac{1}{2} \|y - z\|^2_{L^2(\Omega)} + \alpha \|u\|_{\mathcal{X}} \\ \text{s.t.} \quad Ay = u \end{cases}$$

where $\Omega \subset \mathbb{R}^n$, $n \in \{2, 3\}$, is a simply connected bounded domain with Lipschitz boundary $\partial\Omega$, $\mathcal{X}$ is a non-reflexive Banach space, and $\alpha > 0$ and $z \in L^2(\Omega)$ are given. Furthermore, A is a linear second order elliptic differential operator, taking appropriate boundary conditions. Throughout, we assume that:

(A) $\quad \|A\cdot\|_{L^2}$ and $\|A^*\cdot\|_{L^2}$ are equivalent norms on $H^2(\Omega) \cap H^1_0(\Omega)$,

where A^* denotes the adjoint of A with respect to the inner product in $L^2(\Omega)$.

If $\mathcal{X} = L^1(\Omega)$, this setting applies to optimal control problems where the cost of the control is a linear function of its magnitude (cf. [Vossen and Maurer 2006]); the case $\mathcal{X} = BV(\Omega)$ corresponds to settings where the cost is proportional to changes in the control. Of particular interest is how the structure of optimal controls in such Banach spaces differs from that of controls obtained in Hilbert spaces such as $L^2(\Omega)$ or $H_0^1(\Omega)$. For example, it is known that $L^1(\Omega)$-type costs promote sparsity, whereas $BV(\Omega)$-type penalties favor piecewise constant functions (cf. [Stadler 2009; Ring 2000], respectively). Note that for $\mathcal{X} = L^1(\Omega)$, problem ($\mathcal{P}$) is not well-posed: It need not have a minimizer in $L^1(\Omega)$, since the conditions of the Dunford-Pettis theorem are not satisfied (boundedness in $L^1(\Omega)$ is not a sufficient condition for the existence of a weakly converging subsequence). The natural functional-analytic framework for problems of this type is the space of bounded measures (cf. Remark 7.2.8). In the current paper, we will focus on optimal control problems in the space of measures and in the space of functions of bounded variation, and on their numerical treatment.

For the direct solution of ($\mathcal{P}$), one would need to address the problem of the discretization of measures. We therefore propose an alternative approach that, roughly speaking, consists in interpreting the optimality conditions for problem ($\mathcal{P}$) as an optimality system for the Lagrange multiplier associated with the equality constraint, which by a density argument can then be taken in an appropriate Hilbert space. This can be justified rigorously using Fenchel duality. This, together with the numerical results for simple model problems which highlight the significant difference of the controls in dependence of the chosen norm, constitutes the main contribution of the current paper.

This work is organized as follows: In the remainder of this section, we fix notations and recall some necessary background. In section 7.2, we derive predual formulations for the optimal control problem ($\mathcal{P}$) with measures and functions of bounded variation as controls (in § 7.2.1 and § 7.2.2, respectively), discuss the existence and uniqueness of their solutions, and derive optimality systems. Section 7.3 is concerned with the solution of the optimality systems by a semismooth Newton method, for which it is necessary to consider a regularization of the problem (§ 7.3.1). We can then show superlinear convergence of the method (§ 7.3.2). Here, we focus first on the case of measures, and discuss the corresponding issues for functions of bounded variation in § 7.3.3. Finally, we present numerical examples in section 7.4.

7.1.1 NOTATIONS AND BACKGROUND

For the reader's convenience, we give here the definitions and results on measure theory, functions of bounded variation, and convex duality relevant to this work. For more details and proofs, we refer to, e.g., [Ambrosio, Fusco, and Pallara 2000; Attouch, Buttazzo, and Michaille 2006] (our notation follows the latter). In the following, Lebesgue spaces of vector valued functions are denoted by a blackboard bold letter corresponding to their scalar equivalent, e.g., $\mathbb{L}^2(\Omega) := (L^2(\Omega))^n$.

MEASURE THEORY Let $\mathcal{M}(\Omega)$ denote the vector space of all bounded Borel measures on Ω, that is of all bounded σ-additive set functions $\mu : \mathcal{B}(\Omega) \to \mathbb{R}$ defined on the Borel algebra $\mathcal{B}(\Omega)$ satisfying $\mu(\emptyset) = 0$. The total variation of $\mu \in \mathcal{M}(\Omega)$ is defined for all $B \in \mathcal{B}(\Omega)$ by

$$|\mu|(B) := \sup \left\{ \sum_{i=0}^{\infty} |\mu(B_i)| : \bigcup_{i=0}^{\infty} B_i = B \right\},$$

where the supremum is taken over all partitions of B. Endowed with the norm $\|\mu\|_{\mathcal{M}} = |\mu|(\Omega)$, $\mathcal{M}(\Omega)$ is a Banach space. By the Riesz representation theorem, $\mathcal{M}(\Omega)$ can be isometrically identified with the topological dual of $C_0(\Omega)$, the space of all continuous functions with compact support in Ω, endowed with the norm $\|\nu\|_{C_0} = \sup_{x\in\Omega} |\nu(x)|_\infty$. This leads to the following equivalent characterization of the norm on $\mathcal{M}(\Omega)$:

$$\|\mu\|_{\mathcal{M}} = \sup_{\substack{\varphi\in C_0(\Omega),\\ \|\varphi\|_{C_0}\leqslant 1}} \int_\Omega \varphi \, d\mu.$$

FUNCTIONS OF BOUNDED VARIATION We recall that $BV(\Omega)$, the space of functions of bounded variation, consists of all $u \in L^1(\Omega)$ for which the distributional gradient Du belongs to $(\mathcal{M}(\Omega))^n$. Furthermore, the mapping $u \mapsto \|u\|_{BV}$,

$$\|u\|_{BV} := \int_\Omega |Du| \, dx = \sup \left\{ \int_\Omega u \operatorname{div} v \, dx : v \in (C_0^\infty(\Omega))^n, \ \|v\|_{(C_0)^n} \leqslant 1 \right\}$$

(which can be infinite) is lower semicontinuous in the topology of $L^1(\Omega)$, and $u \in L^1(\Omega)$ is in $BV(\Omega)$ if and only if $\|u\|_{BV}$ is finite. (If $v \in H^1(\Omega)$, then $\|u\|_{BV} = \int_\Omega |\nabla u| \, dx$.) Endowed with the norm $\|\cdot\|_{L^1} + \|\cdot\|_{BV}$, $BV(\Omega)$ is a non-reflexive Banach space.

One of the main features of $BV(\Omega)$ in comparison to Sobolev spaces is that it includes characteristic functions of sufficiently regular sets and piecewise smooth functions. In image reconstruction, $BV(\Omega)$-regularization is known to preserve edges better than regularization with $\|\nabla u\|_{\mathbb{L}^2}^2$.

FENCHEL DUALITY IN CONVEX OPTIMIZATION A complete discussion can be found in [Ekeland and Témam 1999]. Let V and Y be Banach spaces with topological duals V^* and Y^*, respectively, and let $\Lambda : V \to Y$ be a continuous linear operator. Setting $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, let $\mathcal{F} : V \to \bar{\mathbb{R}}$, $\mathcal{G} : Y \to \bar{\mathbb{R}}$ be convex lower semicontinuous functionals which are not identically equal ∞ and for which there exists a $v_0 \in V$ such that $\mathcal{F}(v_0) < \infty$, $\mathcal{G}(\Lambda v_0) < \infty$, and $\mathcal{G}$ is continuous at Λv_0. Let $\mathcal{F}^* : V^* \to \bar{\mathbb{R}}$ denote the Fenchel conjugate of $\mathcal{F}$ defined by

$$\mathcal{F}^*(q) = \sup_{v\in V} \langle q, v\rangle_{V^*,V} - \mathcal{F}(v),$$

which we will frequently calculate using the fact that

$$\mathcal{F}^*(q) = \langle q, v\rangle_{V^*,V} - \mathcal{F}(v) \quad \text{if and only if} \quad q \in \partial\mathcal{F}(v). \tag{7.1.1}$$

Here, $\partial\mathcal{F}$ denotes the subdifferential of the convex function $\mathcal{F}$, which reduces to the Gâteaux-derivative if it exists.

The Fenchel duality theorem states that under the assumptions given above,

$$\inf_{v\in V} \mathcal{F}(v) + \mathcal{G}(\Lambda v) = \sup_{q\in Y^*} -\mathcal{F}^*(\Lambda^* q) - \mathcal{G}^*(-q)\,, \tag{7.1.2}$$

holds, and that the right hand side of (7.1.2) has at least one solution. Furthermore, the equality in (7.1.2) is attained at (v^*, q^*) if and only if

$$\begin{cases} \Lambda^* q^* \in \partial\mathcal{F}(v^*), \\ -q^* \in \partial\mathcal{G}(\Lambda v^*). \end{cases} \tag{7.1.3}$$

7.2 EXISTENCE AND OPTIMALITY CONDITIONS

This section is concerned with the predual problem of ($\mathcal{P}$), discussing the existence and uniqueness of its solution and deriving the first order optimality system. We first consider the case $X = \mathcal{M}(\Omega)$, and then treat the case $X = BV(\Omega)$ in § 7.2.2.

7.2.1 CONTROLS IN $\mathcal{M}(\Omega)$

We consider the optimal control problem

$$\begin{cases} \min\limits_{u\in\mathcal{M}(\Omega)} \dfrac{1}{2}\|y - z\|_{L^2}^2 + \alpha\|u\|_{\mathcal{M}} \\ \text{s.t.} \quad Ay = u. \end{cases} \tag{$\mathcal{P}_{\mathcal{M}}$}$$

First, we have to address the well-posedness of the constraint for measure-valued data. We call $y \in L^1(\Omega)$ a very weak solution of $Ay = u \in \mathcal{M}(\Omega)$ if

$$\int_\Omega y A^*\varphi \,dx = \int_\Omega \varphi \,du \tag{7.2.1}$$

holds for all $\varphi \in C_0(\Omega)$ such that $A^*\varphi \in C_0(\Omega)$. Then, we have the following result [Stampacchia 1965, Th. 9.1]:

Proposition 7.2.1. *For $u \in \mathcal{M}(\Omega)$, the equation $Ay = u$ has a unique weak solution which satisfies $y \in W_0^{1,p}(\Omega)$ for all $1 \leqslant p < \frac{n}{n-1}$. Furthermore, there exists a constant $C > 0$ such that*

$$\|y\|_{W_0^{1,p}} \leqslant C\|u\|_{\mathcal{M}}$$

holds.

Since $W_0^{1,p}(\Omega)$ is compactly embedded in $L^2(\Omega)$, $(\mathcal{P}_\mathcal{M})$ is well-defined, and we can show existence of a minimizer:

Proposition 7.2.2. *Problem $(\mathcal{P}_\mathcal{M})$ has a unique solution $(y^*, u^*) \in L^2(\Omega) \times \mathcal{M}(\Omega)$.*

Proof. We consider a minimizing sequence $(u_n)_{n\in\mathbb{N}} \subset \mathcal{M}(\Omega)$ of $(\mathcal{P}_\mathcal{M})$. Since the pair $(y, u) = (0, 0)$ is feasible, $\frac{1}{2\alpha}\|z\|_{L^2}^2$ is an upper bound for $\|u_n\|_\mathcal{M}$. We can therefore extract a subsequence converging in the weak topology $\sigma(\mathcal{M}(\Omega), C_0(\Omega))$ to a $u^* \in \mathcal{M}(\Omega)$.

Setting $y_n := y(u_n) \in W_0^{1,p}(\Omega)$, i.e., the solution of (7.2.1) with $\mu = u_n$, we therefore can pass to the limit and obtain (by the density of $C_0(\Omega)$ in $L^2(\Omega)$) a $y^* \in W_0^{1,p}(\Omega) \subset L^2(\Omega)$ solving (7.2.1) for u^*. From the weak lower semicontinuity of the norms in $L^2(\Omega)$ and $\mathcal{M}(\Omega)$, we conclude that (y^*, u^*) is a minimizer of $(\mathcal{P}_\mathcal{M})$.

Finally, uniqueness of a minimizer follows directly from strict convexity of the norms and the assumption on A. □

We set $\mathcal{W} := H^2(\Omega) \cap H_0^1(\Omega)$ and consider the problem

$$(\mathcal{P}_\mathcal{M}^*) \qquad \begin{cases} \min\limits_{p\in\mathcal{W}} \dfrac{1}{2}\|A^*p + z\|_{L^2}^2 - \dfrac{1}{2}\|z\|_{L^2}^2 \\ \text{s. t.} \quad \|p\|_{C_0} \leqslant \alpha, \end{cases}$$

which we will show below to be the predual of $(\mathcal{P}_\mathcal{M})$. Due to the embedding of $\mathcal{W}$ in $C_0(\Omega)$, the constraint is well-defined, and problem $(\mathcal{P}_\mathcal{M}^*)$ has a unique solution:

Theorem 7.2.3. *Problem $(\mathcal{P}_\mathcal{M}^*)$ has a unique solution $p^* \in \mathcal{W}$.*

Proof. Let again $\{p_n\}_{n\in\mathbb{N}}$ be a minimizing sequence which is bounded in $\mathcal{W}$ by $2\|z\|_{L^2}^2$. We can thus extract a subsequence $\{p_{n_k}\}_{k\in\mathbb{N}}$ of feasible functions weakly converging to a $p^* \in \mathcal{W}$, which is again feasible. By the weak lower semicontinuity of the norm, we deduce

$$\liminf_{k\to\infty} \frac{1}{2}\|A^*p_{n_k} + z\|_{L^2}^2 \geqslant \frac{1}{2}\|A^*p^* + z\|_{L^2}^2 .$$

Hence, p^* is a minimizer of $(\mathcal{P}_\mathcal{M}^*)$.

By the assumption on A^*, the mapping $p \mapsto \frac{1}{2}\|A^*p + z\|_{L^2}^2$ is strictly convex, and the minimizer is unique. □

Theorem 7.2.4. *The dual of $(\mathcal{P}_\mathcal{M}^*)$ is $(\mathcal{P}_\mathcal{M})$, and the solutions $u^* \in \mathcal{M}(\Omega)$ of $(\mathcal{P}_\mathcal{M})$ and $p^* \in \mathcal{W}$ of $(\mathcal{P}_\mathcal{M}^*)$ are related by*

$$(7.2.2) \qquad \begin{cases} \langle u^*, v\rangle_{\mathcal{W}^*,\mathcal{W}} = \langle A^*p^* + z, A^*v\rangle_{L^2}, \\ 0 \geqslant \langle -u^*, p - p^*\rangle_{\mathcal{M},C_0}, \end{cases}$$

for all $p \in \mathcal{W}$ with $\|p\|_{C_0} \leqslant \alpha$.

Proof. We apply Fenchel duality for problem ($\mathcal{P}^*_{\mathcal{M}}$). Set

$$\mathcal{F} : \mathcal{W} \to \bar{\mathbb{R}}, \qquad \mathcal{F}(q) = \frac{1}{2}\|A^*q + z\|^2_{L^2} - \frac{1}{2}\|z\|^2_{L^2},$$
$$\mathcal{G} : C_0(\Omega) \to \bar{\mathbb{R}}, \qquad \mathcal{G}(q) = I_{\{\|q\|_{C_0} \leqslant \alpha\}} := \begin{cases} 0 & \text{if } \|q\|_{C_0} \leqslant \alpha, \\ \infty & \text{if } \|q\|_{C_0} > \alpha, \end{cases}$$

and $\Lambda : \mathcal{W} \to C_0(\Omega)$ the injection given by the continuous embedding. The Fenchel conjugate of $\mathcal{F}$ is given by

$$\mathcal{F}^* : \mathcal{W}^* \to \bar{\mathbb{R}}, \qquad \mathcal{F}^*(u) = \frac{1}{2}\left\|A^{-1}u - z\right\|^2_{L^2}.$$

The conjugate of $\mathcal{G}$ can be calculated as

$$\begin{aligned} \mathcal{G}^*(u) &= \sup_{q \in C_0(\Omega)} \langle u, q\rangle_{\mathcal{M}, C_0} - I_{\{\|q\|_{C_0} \leqslant \alpha\}} = \sup_{\substack{q \in C_0(\Omega), \\ \|q\|_{C_0} \leqslant \alpha}} \langle u, q\rangle_{\mathcal{M}, C_0} \\ &= \alpha \sup_{\substack{q \in C_0(\Omega), \\ \|q\|_{C_0} \leqslant 1}} \langle u, q\rangle_{\mathcal{M}, C_0} = \alpha \|u\|_{\mathcal{M}}, \end{aligned} \tag{7.2.3}$$

and $\Lambda^* : \mathcal{M}(\Omega) \to \mathcal{W}^*$ is again the injection from the dual of $C_0(\Omega)$ in $\mathcal{W}^*$.

It remains to verify the conditions of the Fenchel duality theorem. Since the norms in $L^2(\Omega)$ and $C_0(\Omega)$ are convex and lower semicontinuous, so are $\mathcal{F}$ and $\mathcal{G}$ (as indicator function of a convex set), which are also proper (e.g., for $q = 0$, at which point $\mathcal{G}$ is continuous). In addition, Λ is a continuous linear operator, and so we have that

$$\min_{\substack{p \in \mathcal{W}, \\ \|p\|_{C_0} \leqslant \alpha}} \frac{1}{2}\|A^*p + z\|^2_{L^2} - \frac{1}{2}\|z\|^2_{L^2} = \min_{u \in \mathcal{M}(\Omega)} \frac{1}{2}\left\|A^{-1}u - z\right\|^2_{L^2} + \alpha\|-u\|_{\mathcal{M}}.$$

Introducing $y \in W^{1,p}_0(\Omega)$ as the solution of $Ay = u \in \mathcal{M}(\Omega)$, we recover problem ($\mathcal{P}_{\mathcal{M}}$), and the relation (7.2.2) follows from the extremality relation (7.1.3). □

From this, we can derive the first order optimality conditions for problem ($\mathcal{P}^*_{\mathcal{M}}$):

Corollary 7.2.5. *Let $p^* \in \mathcal{W}$ be a solution of ($\mathcal{P}^*_{\mathcal{M}}$). Then there exists $\lambda^* \in \mathcal{M}(\Omega) \subset \mathcal{W}^*$ such that*

$$\begin{cases} \langle A^*p^* + z, A^*v\rangle_{L^2} + \langle \lambda^*, v\rangle_{\mathcal{M}, C_0} = 0, \\ \langle \lambda^*, p - p^*\rangle_{\mathcal{M}, C_0} \leqslant 0, \end{cases} \tag{7.2.4}$$

holds for all $v, p \in \mathcal{W}$ with $\|p\|_{C_0} \leqslant \alpha$. Moreover, the solution (p^*, λ^*) of (7.2.4) is unique.

Proof. By setting $\lambda^* = -u^*$ in the extremality relations (7.2.2), we immediately obtain (7.2.4) and existence of the Lagrange multiplier. Let (p_1, λ_1) and (p_2, λ_2) be two solutions of (7.2.4), and set $\delta p = p_1 - p_2$, $\delta\lambda = \lambda_1 - \lambda_2$. Then we have

$$\begin{aligned} \langle A^*\delta p, A^*v\rangle_{L^2} + \langle \delta\lambda, v\rangle_{\mathcal{M},C_0} &= 0, \\ \langle \delta\lambda, \delta p\rangle_{\mathcal{M},C_0} &\geqslant 0. \end{aligned}$$

The choice $v = \delta p$ implies that

$$\|A^*\delta p\|_{L^2}^2 = -\langle \delta\lambda, \delta p\rangle_{\mathcal{M},C_0} \leqslant 0,$$

and hence $\delta p = 0$. By the assumptions on A^*, it follows that $\delta\lambda = 0$. □

Splitting the first equation of (7.2.2) by introducing $y^* = A^*p + z$ yields

$$\text{(7.2.5)} \qquad \begin{cases} Ay^* = u^*, \\ A^*p^* = y^* - z, \\ \quad 0 \geqslant \langle u^*, p^* - p\rangle_{\mathcal{M},C_0} \text{ for all } p \in \mathcal{W}, \|p\|_{C_0} \leqslant \alpha. \end{cases}$$

Formally, this is the optimality system for $(\mathcal{P}_\mathcal{M})$, where $-p$ is the Lagrange multiplier for the constraint $Ay = u$, except that the non-reflexive Banach spaces $\mathcal{M}(\Omega)$ and $\mathcal{M}(\Omega)^*$ have been replaced by the Hilbert spaces $\mathcal{W}^*$ and $\mathcal{W}$, respectively.

From (7.2.5), we can deduce extra regularity for the Lagrange multiplier if the data is sufficiently smooth:

Corollary 7.2.6. *Let $1 \leqslant p < \frac{n}{n-1}$. If $z \in W^{1,p}(\Omega)$, $\partial\Omega \in C^{2,1}$ and the coefficients of A are in $W^{1,p}(\Omega)$ and $C^{2,1}$ for the principial part, then the optimal Lagrange multiplier for the state constraint in $(\mathcal{P}_\mathcal{M})$ satisfies $p^* \in \mathcal{W} \cap W^{3,p}(\Omega)$.*

Proof. From Proposition 7.2.1 and $u^* \in \mathcal{M}(\Omega)$, we obtain that the optimal state satisfies $y^* \in W_0^{1,p}(\Omega)$ for all $1 \leqslant p < \frac{n}{n-1}$. Hence, $A^*p^* = y^* - z$ yields $p^* \in W^{3,p}(\Omega)$. □

We can also give the following structural information for the solution of problem $(\mathcal{P}_\mathcal{M})$:

Corollary 7.2.7. *Let u^* be the minimizer of $(\mathcal{P}_\mathcal{M})$ and p^* the minimizer of $(\mathcal{P}^*_\mathcal{M})$. Then, $u^* = u^*_+ - u^*_-$, where u^*_+ and u^*_- are positive measures with support:*

$$\begin{aligned} \operatorname{supp}(u^*_+) &\subset \{x \in \Omega : p^*(x) = -\alpha\}, \\ \operatorname{supp}(u^*_-) &\subset \{x \in \Omega : p^*(x) = \alpha\}. \end{aligned}$$

This can be interpreted as a sparsity property: The optimal control u^* will be nonzero only on sets where the constraint on the dual variable p^* is active; hence the larger the penalty α, the smaller the support of the control.

Remark 7.2.8. If in addition, the minimizer satisfies $u^* \in L^1(\Omega)$, it is the solution of the problem

$$(\mathcal{P}_{L^1}) \qquad \begin{cases} \min\limits_{u \in L^1(\Omega)} \dfrac{1}{2} \|y - z\|_{L^2}^2 + \alpha \|u\|_{L^1} \\ \text{s.t.} \quad Ay = u \end{cases}$$

This follows from the embedding of $L^1(\Omega)$ into $\mathcal{M}(\Omega)$ and the fact that $\|\nu\|_{\mathcal{M}} = \|\nu\|_{L^1}$ for $\nu \in L^1(\Omega)$ (cf. [Brezis 1983, Ch. IV]).

REGULARIZATION OF $(\mathcal{P}_{\mathcal{M}})$ If we wish to avoid measures, we have to look for the minimizer in a stronger space than $L^1(\Omega)$, e.g., $L^2(\Omega)$. Consider then the following regularized problem:

$$(\mathcal{P}_{L^1,L^2}) \qquad \begin{cases} \min\limits_{u \in L^2(\Omega)} \dfrac{1}{2} \|y - z\|_{L^2}^2 + \alpha \|u\|_{L^1} + \dfrac{\beta}{2} \|u\|_{L^2}^2 \\ \text{s.t.} \quad Ay = u \end{cases}$$

Existence and uniqueness of a minimizer follows from standard arguments. Since $L^2(\Omega)$ is reflexive, we can directly calculate the dual problem: Set

$$\begin{aligned} &\mathcal{F} : L^2(\Omega) \to \bar{\mathbb{R}}, && \mathcal{F}(u) = \alpha \|u\|_{L^1}, \\ &\mathcal{G} : L^2(\Omega) \times L^2(\Omega) \to \bar{\mathbb{R}}, && \mathcal{G}(u, y) = \frac{1}{2} \|y - z\|_{L^2}^2 + \frac{\beta}{2} \|u\|_{L^2}^2, \\ &\Lambda : L^2(\Omega) \to L^2(\Omega) \times L^2(\Omega), && \Lambda u = (u, A^{-1}u). \end{aligned}$$

The Fenchel conjugate of $\mathcal{F}$ is again given by

$$\mathcal{F}^* : L^2(\Omega) \to \bar{\mathbb{R}}, \qquad \mathcal{F}(\bar{u}) = I_{\{\|\bar{u}\|_{L^\infty} \leqslant \alpha\}},$$

and we can calculate

$$\begin{cases} \mathcal{G}^* : L^2(\Omega) \times L^2(\Omega) \to \bar{\mathbb{R}}, \\ \mathcal{G}^*(\bar{u}, \bar{y}) = \dfrac{1}{2} \|\bar{y} + z\|_{L^2}^2 - \dfrac{1}{2} \|z\|_{L^2}^2 + \dfrac{1}{2\beta} \|\bar{u}\|_{L^2}^2, \end{cases}$$

as well as

$$\Lambda^* : L^2(\Omega) \times L^2(\Omega) \to L^2(\Omega), \quad \Lambda^*(\bar{u}, \bar{y}) = \bar{u} + A^{-*}\bar{y}.$$

Since $\mathcal{F}$ and $\mathcal{G}$ are convex and continuous operators (due to the continuous embedding of $L^2(\Omega)$ into $L^1(\Omega)$), and Λ is a continuous linear operator due to the well-posedness of

the equality constraint, the Fenchel duality theorem yields the existence of a minimizer $(\bar{u}^*, \bar{y}^*) \in L^2(\Omega) \times L^2(\Omega)$ of the dual problem

$$\begin{cases} \min\limits_{(\bar{u},\bar{y})\in L^2\times L^2} \dfrac{1}{2}\|-\bar{y}+z\|_{L^2}^2 - \dfrac{1}{2}\|z\|_{L^2}^2 + \dfrac{1}{2\beta}\|-\bar{u}\|_{L^2}^2 \\ \text{s.t.} \quad \|\bar{u} + A^{-*}\bar{y}\|_{L^\infty} \leqslant \alpha, \end{cases}$$

where we can substitute $p = -A^{-*}\bar{y} \in \mathcal{W}$ and $q = \bar{u} - p \in L^2(\Omega)$ to arrive at

$$(\mathcal{P}^*_{L^1,L^2}) \qquad \begin{cases} \min\limits_{(p,q)\in\mathcal{W}\times L^2} \dfrac{1}{2}\|A^*p+z\|_{L^2}^2 - \dfrac{1}{2}\|z\|_{L^2}^2 + \dfrac{1}{2\beta}\|p+q\|_{L^2}^2 \\ \text{s.t.} \quad \|q\|_{L^\infty} \leqslant \alpha. \end{cases}$$

In terms of p, the extremality relations linking the primal and dual minimizers u^* and (p^*, q^*) can be given as

$$\begin{cases} \langle u^*, v_1\rangle_{L^2} = \langle A^*p^* + z, A^*v_1\rangle_{L^2}, \\ \langle u^*, v_2\rangle_{L^2} = \dfrac{1}{\beta}\langle q^* + p^*, v_2\rangle_{L^2}, \\ 0 \geqslant \langle u^*, q^* - q\rangle_{L^2}, \end{cases}$$

for all $v_1 \in \mathcal{W}, v_2, q \in L^2(\Omega)$ with $\|q\|_{L^\infty} \leqslant \alpha$.

Note that now the Lagrange multiplier corresponding to the box constraint is in $L^2(\Omega)$ (as opposed to $\mathcal{W}^*$). Problem $(\mathcal{P}^*_{L^1,L^2})$ therefore can be seen as a regularization of $(\mathcal{P}^*_{\mathcal{M}})$ by introducing a new variable $q = -p$ and treating this equality constraint by penalization (cf. also [Stadler 2009]). In section 7.3 we will directly regularize the dual problem $(\mathcal{P}^*_{\mathcal{M}})$ by a Moreau-Yosida penalization with a parameter c, obtaining a regularized problem $(\mathcal{P}^*_{\mathcal{M},c})$. We will see that the Fenchel dual of $(\mathcal{P}^*_{\mathcal{M},c})$ is $(\mathcal{P}_{L^1,L^2})$ for an appropriate choice of the parameter c (see Remark 7.3.2).

7.2.2 CONTROLS IN $BV(\Omega)$

We now consider the optimal control problem

$$(\mathcal{P}_{BV}) \qquad \begin{cases} \min\limits_{u\in BV(\Omega)} \dfrac{1}{2}\|y-z\|_{L^2}^2 + \alpha\|u\|_{BV} \\ \text{s.t.} \quad Ay = u \end{cases}$$

Since the linear operator A^{-1} is injective on $BV(\Omega) \subset L^1(\Omega)$ by Proposition 7.2.1, the existence of a minimizer follows directly from [Chavent and Kunisch 1997, Th. 2.1]:

Proposition 7.2.9. *If $n = 2$, problem $(\mathcal{P}_{BV})$ has a unique solution $(y^*, u^*) \in L^2(\Omega) \times BV(\Omega)$.*

Now let

$$H^2_{\mathrm{div}}(\Omega) := \left\{ v \in \mathbb{L}^2(\Omega) : \operatorname{div} v \in \mathcal{W},\ v \cdot \nu = 0 \text{ on } \partial\Omega \right\},$$

endowed with the norm $\|v\|^2_{H^2_{\mathrm{div}}} := \|v\|^2_{\mathbb{L}^2} + \|\operatorname{div} v\|^2_{\mathcal{W}}$, and consider

$$(\mathcal{P}^*_{BV}) \qquad \begin{cases} \min\limits_{p \in H^2_{\mathrm{div}}(\Omega)} \dfrac{1}{2} \|A^* \operatorname{div} p + z\|^2_{L^2} - \dfrac{1}{2} \|z\|^2_{L^2} \\ \text{s. t.} \quad \|p\|_{\mathbb{L}^\infty} \leqslant \alpha. \end{cases}$$

This problem has a solution, which, however, may not be unique. Set

$$H^2_{\mathrm{div},0}(\Omega) = \left\{ v \in H^2_{\mathrm{div}}(\Omega) : \operatorname{div} v = 0 \right\}$$

and let $H^2_{\mathrm{div},0}(\Omega)^\perp$ the orthogonal complement in $H^2_{\mathrm{div}}(\Omega)$. Then we can show the following:

Theorem 7.2.10. *Problem $(\mathcal{P}^*_{BV})$ has a solution $p^* \in H^2_{\mathrm{div}}(\Omega)$. Moreover, there exists a unique $q^* \in H^2_{\mathrm{div},0}(\Omega)^\perp$ such that all such solutions satisfy $p^* \in \{q^*\} + H^2_{\mathrm{div},0}(\Omega)$.*

Proof. Consider again a minimizing sequence $\{p_n\}_{n\in\mathbb{N}} \subset H^2_{\mathrm{div}}(\Omega)$. The $\mathbb{L}^2(\Omega)$-norm of p_n is bounded via the box constraints, and the data fit term gives a bound on the $\mathcal{W}$-norm of $(\operatorname{div} p_n)$; together, this yields that the $H^2_{\mathrm{div}}(\Omega)$-norm of p_n is bounded. We can therefore extract a subsequence weakly converging in $H^2_{\mathrm{div}}(\Omega)$, and existence of the minimizer follows from the same arguments as in the proof of Theorem 7.2.3.

Since $H^2_{\mathrm{div},0}(\Omega)$ is a closed subspace of $H^2_{\mathrm{div}}(\Omega)$, it holds that

$$H^2_{\mathrm{div}}(\Omega) = H^2_{\mathrm{div},0}(\Omega)^\perp \oplus H^2_{\mathrm{div},0}(\Omega),$$

and that div is injective on $H^2_{\mathrm{div},0}(\Omega)^\perp$ by construction. Therefore, $\frac{1}{2}\|A^* \operatorname{div} p + z\|^2_{L^2}$ is strictly convex on $H^2_{\mathrm{div},0}(\Omega)^\perp$, so that $(\mathcal{P}^*_{BV})$ has a unique minimizer $q^* \in H^2_{\mathrm{div},0}(\Omega)^\perp$ there. On the other hand, given $p \in H^2_{\mathrm{div},0}(\Omega)$ with $\|q^* + p\|_{\mathbb{L}^\infty} \leqslant \alpha$, we find that $q^* + p \in H^2_{\mathrm{div}}(\Omega)$ is also a minimizer. □

Theorem 7.2.11. *The dual of $(\mathcal{P}^*_{BV})$ is $(\mathcal{P}_{BV})$, and the solutions $u^* \in BV(\Omega)$ of $(\mathcal{P}_{BV})$ and $p^* \in H^2_{\mathrm{div}}(\Omega)$ of $(\mathcal{P}^*_{BV})$ are related by*

$$(7.2.6) \qquad \begin{cases} \langle -u^*, v \rangle_{{H^2_{\mathrm{div}}}^*, H^2_{\mathrm{div}}} = \langle A^*(-\operatorname{div}) p^* + z, A^* v \rangle_{L^{2*}, L^2}, \\ 0 \geqslant \langle (-\operatorname{div})^* u^*, p - p^* \rangle_{{H^2_{\mathrm{div}}}^*, H^2_{\mathrm{div}}}, \end{cases}$$

for all $v \in \mathcal{W}$, $p \in H^2_{\mathrm{div}}(\Omega)$ with $\|p\|_{\mathbb{L}^\infty} \leqslant \alpha$.

Proof. We apply Fenchel duality. Setting

$$\begin{aligned} &\mathcal{F} : H^2_{\mathrm{div}}(\Omega) \to \bar{\mathbb{R}}, \quad && \mathcal{F}(q) = I_{\{\|q\|_{\mathbb{L}^\infty} \leqslant \alpha\}}, \\ &\mathcal{G} : \mathcal{W} \to \bar{\mathbb{R}}, && \mathcal{G}(q) = \frac{1}{2} \|A^* q - z\|^2_{L^2} - \frac{1}{2} \|z\|^2_{L^2}, \\ &\Lambda : H^2_{\mathrm{div}}(\Omega) \to \mathcal{W}, && \Lambda q = -\operatorname{div} q, \end{aligned}$$

problem $(\mathcal{P}^*_{BV})$ can be written as

$$\min_{p \in H^2_{\mathrm{div}}(\Omega)} \mathcal{F}(p) + \mathcal{G}(\Lambda p).$$

The Fenchel conjugate of $\mathcal{G}$ is given by

$$\mathcal{G}^* : H^2_{\mathrm{div}}(\Omega)^* \to \bar{\mathbb{R}}, \qquad \mathcal{G}^*(v) = \frac{1}{2} \left\|A^{-1} v + z\right\|^2_{L^2},$$

and the adjoint of Λ is

$$\Lambda^* : \mathcal{W}^* \to H^2_{\mathrm{div}}(\Omega)^*, \quad \Lambda^* v = (-\operatorname{div})^* v.$$

It remains to calculate $\mathcal{F}^* : H^2_{\mathrm{div}}(\Omega)^* \to \bar{\mathbb{R}}$. We have, as in (7.2.3), that

$$\mathcal{F}^*(v) = \alpha \sup_{\substack{q \in H^2_{\mathrm{div}}(\Omega), \\ \|q\|_{\mathbb{L}^\infty} \leqslant 1}} \langle v, q \rangle_{H^{2\,*}_{\mathrm{div}}, H^2_{\mathrm{div}}}.$$

By standard arguments, we can show that $(C^\infty_0(\Omega))^n$ is dense in $H^2_{\mathrm{div}}(\Omega)$ (cf. [Témam 2001, p. 26], [Amrouche, Ciarlet, and Ciarlet 2007]), so that we can equivalently write

$$\mathcal{F}^*(v) = \alpha \sup_{\substack{q \in (C^\infty_0(\Omega))^n, \\ \|q\|_{(C_0)^n} \leqslant 1}} \langle v, q \rangle_{H^{2\,*}_{\mathrm{div}}, H^2_{\mathrm{div}}},$$

and thus

$$\mathcal{F}^*((-\operatorname{div})^* u) = \alpha \sup_{\substack{q \in (C^\infty_0(\Omega))^n, \\ \|q\|_{(C_0)^n} \leqslant 1}} \langle u, -\operatorname{div} q \rangle_{\mathcal{W}^*, \mathcal{W}} = \alpha \|u\|_{BV},$$

which is finite if and only if $u \in BV(\Omega)$.

Again, $\mathcal{F}$ is convex and lower semicontinuous, $\mathcal{G}$ is convex and continuous on $\mathcal{W}$, and Λ is a continuous linear operator. The Fenchel duality theorem thus yields the duality of $(\mathcal{P}^*_{BV})$ and

$$\max_{u \in \mathcal{W}^*} -\mathcal{F}^*(\Lambda^* u) - \mathcal{G}^*(-u) = -\inf_{u \in BV(\Omega)} \alpha \|u\|_{BV} + \frac{1}{2} \left\|A^{-1} u - z\right\|^2_{L^2},$$

by which we recover $(\mathcal{P}_{BV})$. The relations (7.2.6) are once more the explicit form of the extremality relations (7.1.3). □

From this, we also obtain the first order necessary optimality conditions:

Corollary 7.2.12. *Let* $p^* \in H^2_{\mathrm{div}}(\Omega)$ *be a solution of* $(\mathcal{P}^*_{BV})$. *Then there exists* $\lambda^* \in H^2_{\mathrm{div}}(\Omega)^*$ *such that*

$$\begin{cases} \langle A^*(-\operatorname{div})p^* + z, A^*(-\operatorname{div})w\rangle_{L^2} + \langle \lambda^*, w\rangle_{H^2_{\mathrm{div}}{}^*, H^2_{\mathrm{div}}} = 0, \\ \langle \lambda^*, p - p^*\rangle_{H^2_{\mathrm{div}}{}^*, H^2_{\mathrm{div}}} \leqslant 0, \end{cases} \tag{7.2.7}$$

holds for all $w, p \in H^2_{\mathrm{div}}(\Omega)$ *with* $\|p\|_{\mathbb{L}^\infty} \leqslant \alpha$. *Moreover, the solution* (p^*, λ^*) *of* (7.2.4) *is unique in* $H^2_{\mathrm{div},0}(\Omega)^\perp \times H^2_{\mathrm{div}}(\Omega)^*$.

Proof. For $w \in H^2_{\mathrm{div}}(\Omega)$, we insert $v = -\operatorname{div} w \in \mathcal{W}$ in the first relation of (7.2.6), and set $\lambda^* := (-\operatorname{div})^* u^* \in H^2_{\mathrm{div}}(\Omega)^*$. Proceeding as in the proof of Corollary 7.2.5, we deduce that λ^* solving (7.2.7) is unique. Since the operator $\langle A^*(-\operatorname{div})\cdot, A^*(-\operatorname{div})\cdot\rangle_{L^2}$ is an inner product on $H^2_{\mathrm{div},0}(\Omega)^\perp$ by construction, we obtain a unique $q^* \in H^2_{\mathrm{div},0}(\Omega)^\perp$ such that all solutions p^* of (7.2.6) satisfy $p^* \in \{q^*\} + H^2_{\mathrm{div},0}(\Omega)$. □

Note that this implies that any solution p^* of problem $(\mathcal{P}^*_{BV})$ will yield the same (unique) solution u^* of $(\mathcal{P}_{BV})$ when calculated via the extremality relation (7.2.6).

Remark 7.2.13. In order to obtain a unique solution for problem $(\mathcal{P}^*_{BV})$, we can consider the following regularized problem for $\beta > 0$:

$$(\mathcal{P}^*_{BV,L^2}) \qquad \begin{cases} \min\limits_{p \in H^2_{\mathrm{div}}(\Omega)} \dfrac{1}{2}\|A^* \operatorname{div} p + z\|^2_{L^2} - \dfrac{1}{2}\|z\|^2_{L^2} + \dfrac{\beta}{2}\|p\|^2_{\mathbb{L}^2} \\ \text{s. t.} \quad \|p\|_{\mathbb{L}^\infty} \leqslant \alpha. \end{cases}$$

This can be expressed as a regularization of the primal problem $(\mathcal{P}_{BV})$ as well: Setting

$$\mathcal{F}(q) = I_{\{\|q\|_{\mathbb{L}^\infty} \leqslant \alpha\}} + \frac{\beta}{2}\|q\|^2_{\mathbb{L}^2}$$

and Λ as in Theorem 7.2.11, we find that

$$\mathcal{F}^*(\Lambda^* u) = \int_\Omega \varphi(\nabla u)\,dx,$$

where

$$\varphi(\vec{v})(x) = \begin{cases} \frac{1}{2\beta}|\vec{v}(x)|^2 & \text{if } |\vec{v}(x)| < \alpha\beta, \\ \alpha|\vec{v}(x)| - \frac{\alpha\beta}{2} & \text{if } |\vec{v}(x)| \geqslant \alpha\beta. \end{cases}$$

In the primal problem, the additional predual $\mathbb{L}^2(\Omega)$-regularization essentially results in locally replacing the $BV(\Omega)$-term with a $H^1(\Omega)$-penalty in a small neighborhood of the origin (cf. [Hintermüller and Stadler 2006]).

We can finally give some structural information on the optimal control in $BV(\Omega)$:

Corollary 7.2.14. *Let u^* be a minimizer of ($\mathcal{P}_{BV}$). Then the following holds for any $p \in H^2_{\mathrm{div}}(\Omega)$, $p \geqslant 0$:*

$$\langle (-\operatorname{div})^* u^*, p\rangle_{{H^2_{\mathrm{div}}}^*, H^2_{\mathrm{div}}} = 0 \ \textit{if} \ \operatorname{supp} p \subset \bigcap_{i=1}^{n} \{x : |p_i^*(x)| < \alpha\},$$

$$\langle (-\operatorname{div})^* u^*, p\rangle_{{H^2_{\mathrm{div}}}^*, H^2_{\mathrm{div}}} \geqslant 0 \ \textit{if} \ \operatorname{supp} p \subset \bigcup_{i=1}^{n} \{x : p_i^*(x) = \alpha\},$$

$$\langle (-\operatorname{div})^* u^*, p\rangle_{{H^2_{\mathrm{div}}}^*, H^2_{\mathrm{div}}} \leqslant 0 \ \textit{if} \ \operatorname{supp} p \subset \bigcup_{i=1}^{n} \{x : p_i^*(x) = -\alpha\}.$$

Again, this can be interpreted as a sparsity condition on the gradient of the control: The optimal control u^* will be piecewise constant on sets where the constraints on the dual variable p^* are inactive.

7.3 SOLUTION OF THE OPTIMALITY SYSTEMS

This section, we present a method for the numerical solution of the optimality systems (7.2.4) and (7.2.7). For this, we need to deal with the fact that the Lagrange multiplier corresponding to the box constraints is in general only in $\mathcal{W}^*$ (and $H^2_{\mathrm{div}}(\Omega)^*$, respectively). We introduce therefore regularized versions of (7.2.4) and (7.2.7) which can be solved using a semismooth Newton method having superlinear convergence.

Again, we first treat the solution of (7.2.4), and discuss the corresponding results and algorithm for problem (7.2.7) more briefly in § 7.3.3.

7.3.1 REGULARIZATION OF BOX CONSTRAINTS

In order to obtain Lagrange multipliers in $L^2(\Omega)$, we introduce the Moreau–Yosida regularization of problem ($\mathcal{P}^*_{\mathcal{M}}$) for $c > 0$:

$$(\mathcal{P}^*_{\mathcal{M},c}) \quad \min_{p\in\mathcal{W}} \frac{1}{2}\|A^*p + z\|^2_{L^2} - \frac{1}{2}\|z\|^2_{L^2} + \frac{1}{2c}\|\max(0, c(p-\alpha))\|^2_{L^2} + \frac{1}{2c}\|\min(0, c(p+\alpha))\|^2_{L^2},$$

where the max and min are taken pointwise in Ω. This is equivalent to the regularization in problem ($\mathcal{P}^*_{L^1,L^2}$) (cf. Remark 7.3.2) but more amenable to solution by a semismooth Newton method. Existence and uniqueness of a minimizer is directly deduced from lower

semicontinuity and strict convexity of the functional. The corresponding optimality system is given by

$$\begin{cases} \langle A^* p_c, A^* v\rangle_{L^2} + \langle z, A^* v\rangle_{L^2} + \langle \lambda_c, v\rangle_{\mathcal{W}^*,\mathcal{W}} = 0, \\ \lambda_c = \max(0, c(p_c - \alpha)) + \min(0, c(p_c + \alpha)), \end{cases} \tag{7.3.2}$$

where the Lagrange multiplier satisfies $\lambda_c \in W^{1,\infty}(\Omega)$.

First, we address the convergence of the solutions of (7.3.2) as c tends to infinity:

Theorem 7.3.1. *Let $(p_c, \lambda_c) \in \mathcal{W} \times \mathcal{W}^*$ be the solution of (7.3.2) for given $c > 0$, and $(p^*, \lambda^*) \in \mathcal{W} \times \mathcal{W}^*$ be the unique solution of (7.2.4). Then we have as $c \to \infty$:*

$$\begin{aligned} p_c &\to p^* \quad \text{in } \mathcal{W}, \\ \lambda_c &\rightharpoonup \lambda^* \quad \text{in } \mathcal{W}^*. \end{aligned}$$

Proof. From the optimality conditions (7.3.2), we have that pointwise in $x \in \Omega$

$$\lambda_c p_c = \max(0, c(p_c - \alpha))p_c + \min(0, c(p_c + \alpha))p_c = \begin{cases} c(p_c - \alpha)p_c, & p_c \geqslant \alpha, \\ 0, & |p_c| < \alpha, \\ c(p_c + \alpha)p_c, & p_c \leqslant -\alpha, \end{cases}$$

and hence that

$$\langle \lambda_c, p_c\rangle_{L^2} \geqslant \frac{1}{c} \|\lambda_c\|_{L^2}^2 . \tag{7.3.3}$$

Inserting p_c in (7.3.2), yields

$$\|A^* p_c\|_{L^2}^2 + \frac{1}{c} \|\lambda_c\|_{L^2}^2 \leqslant \|A^* p_c\|_{L^2} \|z\|_{L^2} , \tag{7.3.4}$$

and we deduce that $\|A^* p_c\|_{L^2} \leqslant \|z\|_{L^2}$, as well as

$$\begin{aligned} \|\lambda_c\|_{\mathcal{W}^*} &= \sup_{\substack{v \in \mathcal{W}, \\ \|v\|_{\mathcal{W}} \leqslant 1}} \langle \lambda_c, v\rangle_{\mathcal{W}^*,\mathcal{W}} \leqslant \sup_{\substack{v \in \mathcal{W}, \\ \|v\|_{\mathcal{W}} \leqslant 1}} [\langle A^* p_c, A^* v\rangle_{L^2} + \langle z, A^* v\rangle_{L^2}] \\ &\leqslant 2 \sup_{\substack{v \in \mathcal{W}, \\ \|v\|_{\mathcal{W}} \leqslant 1}} \|A^* v\|_{L^2} \|z\|_{L^2} =: K < \infty. \end{aligned}$$

Thus, (p_c, λ_c) is uniformly bounded in $\mathcal{W} \times \mathcal{W}^*$, so that we can deduce the existence of a $(\tilde{p}, \tilde{\lambda}) \in \mathcal{W} \times \mathcal{W}^*$ such that

$$(p_c, \lambda_c) \rightharpoonup (\tilde{p}, \tilde{\lambda}) \quad \text{in } \mathcal{W} \times \mathcal{W}^*.$$

Passing to the limit in (7.3.2), we obtain

$$\langle A^* \tilde{p}, A^* v\rangle_{L^2} + \langle z, A^* v\rangle_{L^2} + \langle \tilde{\lambda}, v\rangle_{\mathcal{W}^*,\mathcal{W}} = 0 \qquad \text{for all } v \in \mathcal{W}.$$

We next verify the feasibility of $\tilde{p}$. By pointwise inspection similar to (7.3.3), we obtain that

$$\frac{1}{c}\|\lambda_c\|_{L^2}^2 = c\,\|\max(0, p_c - \alpha)\|_{L^2}^2 + c\,\|\min(0, p_c + \alpha)\|_{L^2}^2 .$$

From (7.3.4), we have that $\frac{1}{c}\|\lambda_c\|_{L^2}^2 \leqslant \|z\|_{L^2}^2$, so that

$$\|\max(0, p_c - \alpha)\|_{L^2}^2 \leqslant \frac{1}{c}\|z\|_{L^2}^2 \to 0,$$
$$\|\min(0, p_c + \alpha)\|_{L^2}^2 \leqslant \frac{1}{c}\|z\|_{L^2}^2 \to 0$$

holds for $c \to \infty$. Since $p_c \to \tilde{p}$ strongly in $L^2(\Omega)$, this implies that

$$-\alpha \leqslant \tilde{p}(x) \leqslant \alpha \quad \text{for all } x \in \Omega.$$

It remains to pass to the limit in the second equation of (7.2.4). First, optimality of p_c yields that

$$\frac{1}{2}\|A^* p_c + z\|_{L^2}^2 \leqslant \frac{1}{2}\|A^* p + z\|_{L^2}^2$$

holds for all feasible $p \in \mathcal{W}$. Therefore, we have that

$$\limsup_{c\to\infty} \frac{1}{2}\|A^* p_c + z\|_{L^2}^2 \leqslant \frac{1}{2}\|A^* \tilde{p} + z\|_{L^2}^2$$

and thus $p_c \to \tilde{p}$ strongly in $\mathcal{W}$. Now observe that

$$\langle \lambda_c, p - p_c\rangle_{\mathcal{W}^*,\mathcal{W}} = \langle \max(0, c(p_c - \alpha)), p - p_c\rangle_{\mathcal{W}^*,\mathcal{W}} + \langle \min(0, c(p_c + \alpha)), p - p_c\rangle_{\mathcal{W}^*,\mathcal{W}} \leqslant 0$$

holds for all $p \in \mathcal{W}$ with $\|p\|_{C_0} \leqslant \alpha$, and thus

$$\langle \tilde{\lambda}, p - \tilde{p}\rangle_{\mathcal{W}^*,\mathcal{W}} \leqslant 0$$

is satisfied for all $p \in \mathcal{W}$ with $\|p\|_{C_0} \leqslant \alpha$. Therefore, $(\tilde{p}, \tilde{\lambda}) \in \mathcal{W} \times \mathcal{W}^*$ satisfies (7.2.4), and since the solution of (7.2.4) is unique, $\tilde{p} = p^*$ and $\tilde{\lambda} = \lambda^*$ follows. □

Remark 7.3.2. We can also relate the Moreau–Yosida regularization to the regularized problem ($\mathcal{P}^*_{L^1,L^2}$) via the primal problem: We proceed by calculating the dual of ($\mathcal{P}^*_{\mathcal{M},c}$) as in the proof of Theorem 7.2.4, defining $\mathcal{G} : C_0(\Omega) \to \bar{\mathbb{R}}$,

$$\mathcal{G}(q) = \frac{1}{2c}\|\max(0, c(q - \alpha))\|_{L^2}^2 + \frac{1}{2c}\|\min(0, c(q + \alpha))\|_{L^2}^2 .$$

To calculate the Fenchel conjugate, we use (7.1.1), which in this case implies that

$$u = \max(0, c(q - \alpha)) + \min(0, c(q + \alpha))$$

has to hold for the primal variable $u \in M$. If $u(x) > 0$, the right hand side has to be positive as well, which implies $u(x) = c(q(x) - \alpha)$ and hence $q(x) = \frac{1}{c}u(x) + \alpha$. Similarly, $u(x) < 0$ yields $q(x) = \frac{u}{c}(x) - \alpha$. For $u(x) = 0$, $-\alpha < p(x) < \alpha$ holds. Substituting in the definition of $\mathcal{G}^*$, we have that

$$\begin{aligned}\mathcal{G}^*(u) &= \int_{\{u>0\}} u(x)\left(\frac{1}{c}u(x) + \alpha\right) - \frac{1}{2c}\max(0, u(x))^2 \,dx \\ &\quad + \int_{\{u<0\}} u(x)\left(\frac{1}{c}u(x) - \alpha\right) - \frac{1}{2c}\min(0, u(x))^2 \,dx \\ &= \frac{1}{2c}\|u\|_{L^2}^2 + \alpha\|u\|_{L^1},\end{aligned}$$

which is finite if and only if $u \in L^2(\Omega)$. Setting $\beta := \frac{1}{c}$, we arrive at problem $(\mathcal{P}_{L^1,L^2})$. Since both regularized problems are posed in reflexive Hilbert spaces, they are equivalent.

7.3.2 SEMISMOOTH NEWTON METHOD

The regularized optimality system (7.3.2) can be solved efficiently using a semismooth Newton method (cf. [Hintermüller, Ito, and Kunisch 2002; Ulbrich 2002]), which is superlinearly convergent. For this purpose, we consider (7.3.2) as a nonlinear equation $F(p) = 0$ for $F : \mathcal{W} \to \mathcal{W}^*$, defined for $\nu \in \mathcal{W}$ by

$$\langle F(p), \nu\rangle_{\mathcal{W}^*,\mathcal{W}} := \langle A^*p + z, A^*\nu\rangle_{L^2} + \langle \max(0, c(p-\alpha)) + \min(0, c(p+\alpha)), \nu\rangle_{L^2}.$$

It is known (cf., e.g., [Ito and Kunisch 2008, Ex. 8.14]) that the projection operator

$$P_\alpha(p) := \max(0, (p-\alpha)) + \min(0, (p+\alpha))$$

is semismooth from $L^q(\Omega)$ to $L^p(\Omega)$, if and only if $q > p$, and has as Newton derivative

$$\partial_N P_\alpha(p)h := h\chi_{\{|p|>\alpha\}} = \begin{cases} h(x) & \text{if } |p(x)| > \alpha, \\ 0 & \text{if } |p(x)| \leqslant \alpha. \end{cases}$$

Since Frechét-differentiable functions and sums of semismooth functions are semismooth (with canonical Newton derivatives), we find that F is semismooth, and that its Newton derivative is

$$\langle \partial_N F(p)h, \nu\rangle_{\mathcal{W}^*,\mathcal{W}} = \langle A^*h, A^*\nu\rangle_{L^2} + c\left\langle h\chi_{\{|p|>\alpha\}}, \nu\right\rangle_{L^2},$$

for all $\nu \in \mathcal{W}$.

A semismooth Newton step consists in solving for p^{k+1} the equation

$$\partial_N F(p^k)(p^{k+1} - p^k) = -F(p^k). \tag{7.3.5}$$

Algorithm 7.1 Semismooth Newton method for (7.3.2)

1: Set $k = 0$, Choose $p^0 \in \mathcal{W}$
2: **repeat**
3: Set

$$\begin{aligned} \mathcal{A}_k^+ &= \left\{x : p^k(x) > \alpha\right\}, \\ \mathcal{A}_k^- &= \left\{x : p^k(x) < -\alpha\right\}, \\ \mathcal{A}_k &= \mathcal{A}_k^+ \cup \mathcal{A}_k^- \end{aligned}$$

4: Solve for $p^{k+1} \in \mathcal{W}$:

$$\left\langle A^* p^{k+1}, A^* v\right\rangle_{L^2} + c\left\langle p^{k+1}\chi_{\mathcal{A}_k}, v\right\rangle_{L^2} = -\left\langle z, A^* v\right\rangle_{L^2} + c\alpha\left\langle \chi_{\mathcal{A}_k^+} - \chi_{\mathcal{A}_k^-}, v\right\rangle_{L^2}$$

for all $v \in \mathcal{W}$
5: Set $k = k + 1$
6: **until** $(\mathcal{A}_k^+ = \mathcal{A}_{k-1}^+)$ and $(\mathcal{A}_k^- = \mathcal{A}_{k-1}^-)$

Defining the active and inactive sets

$$\mathcal{A}_k^+ := \left\{x : p^k(x) > \alpha\right\}, \quad \mathcal{A}_k^- := \left\{x : p^k(x) < -\alpha\right\}, \quad \mathcal{A}_k := \mathcal{A}_k^+ \cup \mathcal{A}_k^-,$$

step (7.3.5) can be written explicitly as finding $p^{k+1} \in \mathcal{W}$ such that

$$\left\langle A^* p^{k+1}, A^* v\right\rangle_{L^2} + c\left\langle p^{k+1}\chi_{\mathcal{A}_k}, v\right\rangle_{L^2} = -\left\langle z, A^* v\right\rangle_{L^2} + c\alpha\left\langle \chi_{\mathcal{A}_k^+} - \chi_{\mathcal{A}_k^-}, v\right\rangle_{L^2} \tag{7.3.6}$$

for all $v \in \mathcal{W}$. The resulting semismooth Newton method is given as Algorithm 7.1.

Theorem 7.3.3. *If $\left\|p_c - p^0\right\|_{\mathcal{W}}$ is sufficiently small, the iterates p^k of Algorithm 7.1 converge superlinearly in $\mathcal{W}$ to the solution p_c of (7.3.2) as $k \to \infty$.*

Proof. Since F is semismooth, it suffices to show that $(\partial_N F)^{-1}$ is uniformly bounded. Let $g \in \mathcal{W}^*$ be given. Due to the assumptions on A^*, the Riesz representation theorem ensures the existence of a unique $\varphi \in \mathcal{W}$ such that

$$\left\langle A^*\varphi, A^* v\right\rangle_{L^2} + c\left\langle \chi_{\mathcal{A}}\varphi, v\right\rangle_{L^2} = \left\langle g, v\right\rangle_{\mathcal{W}^*,\mathcal{W}}$$

holds for all $v \in \mathcal{W}$, independent of $\mathcal{A}$. Furthermore, φ satisfies

$$\|\varphi\|_{\mathcal{W}}^2 \leqslant C\|g\|_{\mathcal{W}^*}^2,$$

with a constant C depending only on A and Ω, giving the desired uniform bound. The superlinear convergence now follows from standard results (e.g., [Ito and Kunisch 2008, Th. 8.16]). □

The termination criterion in Algorithm 7.1, step 6, can be justified as follows:

Proposition 7.3.4. *If $\mathcal{A}^+_{k+1} = \mathcal{A}^+_k$ and $\mathcal{A}^-_{k+1} = \mathcal{A}^-_k$ holds, then p^{k+1} satisfies $F(p^{k+1}) = 0$.*

Proof. Since the solution of (7.3.6) is unique for fixed $\mathcal{A}^+_k, \mathcal{A}^-_k$, we have that $p^{k+1} = p^k$. Thus, setting $\mathcal{A}^+_{k+1} = \mathcal{A}^+_k$ and $\mathcal{A}^-_{k+1} = \mathcal{A}^-_k$ in (7.3.6) and noting that

$$c\chi_{\mathcal{A}^+_{k+1}} p^{k+1} - c\alpha\chi_{\mathcal{A}^+_{k+1}} = \max(0, c(p^{k+1} - \alpha)),$$

we see that (7.3.6) is equivalent to $F(p^{k+1}) = 0$. It follows that p^{k+1} is a solution of (7.3.2). □

7.3.3 CONTROLS IN $BV(\Omega)$

The arguments above rely on the fact that the term $\|A^* p\|_{L^2}$ in the functional is an equivalent norm on $\mathcal{W}$. For the problem in $BV(\Omega)$, the corresponding term $\|A^* \operatorname{div} p\|_{L^2}$ is only a seminorm on $H^2_{\mathrm{div}}(\Omega)$, and we need to add additional regularization. Since furthermore $H^2_{\mathrm{div}}(\Omega)$ does not embed into $\mathbb{L}^q$ for $q > 2$, we set $\mathcal{H} := H^2_{\mathrm{div}}(\Omega) \cap \mathcal{W}^n$ and consider

$$(\mathcal{P}^*_{BV,c}) \quad \min_{p\in\mathcal{H}} \frac{1}{2}\|A^* \operatorname{div} p + z\|^2_{L^2} + \frac{\beta}{2}\|p\|^2_{\mathcal{W}^n} - \frac{1}{2}\|z\|^2_{L^2} + \frac{1}{2c}\|\max(0, c(p-\alpha))\|^2_{\mathbb{L}^2} + \frac{1}{2c}\|\min(0, c(p+\alpha))\|^2_{\mathbb{L}^2}$$

with the corresponding optimality system

$$(7.3.8) \qquad \begin{cases} \langle A^*(-\operatorname{div})p_c + z, A^*(-\operatorname{div})v\rangle_{L^2} + \beta\langle \Delta p_c, \Delta v\rangle_{\mathbb{L}^2} + \langle \lambda_c, v\rangle_{H^2_{\mathrm{div}}{}^*, H^2_{\mathrm{div}}} = 0, \\ \lambda_c = \max(0, c(p_c - \alpha)) + \min(0, c(p_c + \alpha)), \end{cases}$$

for all $v \in H^2_{\mathrm{div}}(\Omega)$, where Δ denotes the componentwise Laplacian with homogeneous Dirichlet boundary conditions, and the max, min are understood to act componentwise. (Here and below, α stands for the vector $(\alpha, \dots, \alpha) \in \mathbb{R}^n$.)

The convergence for $\beta \to 0$ is impeded by the fact that no unique candidate for the limit exists. We can, however, show convergence for the corresponding regularization of problem ($\mathcal{P}^*_{BV,L^2}$), and the proof is given in the appendix. In the following, we will consider only the solution of problem ($\mathcal{P}^*_{BV,c}$).

We once more formulate the optimality system (7.3.8) as a semismooth operator equation $G(p) = 0$ for $G : \mathcal{H} \to \mathcal{H}^*$,

$$\langle G(p), v\rangle_{\mathcal{H}^*,\mathcal{H}} := \langle A^*(-\operatorname{div})p + z, A^*(-\operatorname{div})v\rangle_{L^2} + \beta\langle \Delta p, \Delta v\rangle_{\mathbb{L}^2} + \langle \max(0, c(p-\alpha)) + \min(0, c(p+\alpha)), v\rangle_{L^2}$$

Algorithm 7.2 Semismooth Newton method for (7.3.8)

1: Set $k = 0$, Choose $p^0 \in \mathcal{H}$
2: **repeat**
3: Set for $i = 1, \dots, n$

$$\mathcal{A}^+_{i,k} = \left\{x : p^k_i(x) > \alpha\right\}, \ \mathcal{A}^-_{i,k} = \left\{x : p^k_i(x) < -\alpha\right\}, \ \mathcal{A}_{i,k} = \mathcal{A}^+_{i,k} \cup \mathcal{A}^-_{i,k}$$

4: Solve for $p^{k+1} \in \mathcal{H}$:

$$\left\langle A^* \operatorname{div} p^{k+1}, A^* \operatorname{div} v\right\rangle_{L^2} + \beta \sum_{i=1}^n \left\langle \Delta p^{k+1}_i, \Delta v_i\right\rangle_{L^2} + c \sum_{i=1}^n \left\langle p^{k+1}_i \chi_{\mathcal{A}_{k,i}}, v_i\right\rangle_{L^2} = \left\langle z, A^* \operatorname{div} v\right\rangle_{L^2} + c\alpha \sum_{i=1}^n \left\langle \chi_{\mathcal{A}^+_{i,k}} - \chi_{\mathcal{A}^-_{i,k}}, v_i\right\rangle_{L^2}$$

for all $v \in \mathcal{H}$
5: Set $k = k + 1$
6: **until** $(\mathcal{A}^+_{i,k} = \mathcal{A}^+_{i,k-1})$ and $(\mathcal{A}^-_{i,k} = \mathcal{A}^-_{i,k-1})$ for all $i = 1, \dots, n$

for all $v \in \mathcal{H}$ with Newton derivative

$$\left\langle \partial_N G(p)h, v\right\rangle_{\mathcal{H}^*,\mathcal{H}} = \left\langle A^*(-\operatorname{div})h, A^*(-\operatorname{div})v\right\rangle_{L^2} + \beta \left\langle \Delta h, \Delta v\right\rangle_{L^2} + c \left\langle h\chi_{\{|p|>\alpha\}}, v\right\rangle_{L^2} \tag{7.3.9}$$

for all $v \in \mathcal{H}$, where the term $h\chi_{\{|p|>\alpha\}}$ is evaluated componentwise, i.e.,

$$(h\chi_{\{|p|>\alpha\}})_i = \begin{cases} h_i(x) & \text{if } |p_i(x)| > \alpha, \\ 0 & \text{if } |p_i(x)| \leqslant \alpha. \end{cases} \tag{7.3.10}$$

for $i = 1, \dots, n$. One step of the semismooth Newton method consists thus in finding $p^{k+1} \in \mathcal{H}$ such that

$$\left\langle A^* \operatorname{div} p^{k+1}, A^* \operatorname{div} v\right\rangle_{L^2} + \beta \left\langle \Delta p^{k+1}, \Delta v\right\rangle_{\mathbb{L}^2} + c \left\langle p^{k+1} \chi_{\mathcal{A}_k}, v\right\rangle_{\mathbb{L}^2} = \left\langle z, A^* \operatorname{div} v\right\rangle_{L^2} + c \left\langle \alpha\chi_{\mathcal{A}^+_k} - \alpha\chi_{\mathcal{A}^-_k}, v\right\rangle_{\mathbb{L}^2}$$

holds for all $v \in \mathcal{H}$. The active sets $\mathcal{A}$ and their characteristic functions are defined componentwise as in (7.3.10). The full Newton method is given as Algorithm 7.2.

Since the weak form of the Newton derivative (7.3.9) by construction defines an inner product on $\mathcal{H}$, the same argument used in Theorem 7.3.3 yields uniform boundedness of $(\partial_N G)^{-1}$. Hence Algorithm 7.2 converges locally superlinearly and terminates if the active sets coincide:

Theorem 7.3.5. *If* $\|p_c - p^0\|_{\mathcal{H}}$ *is sufficiently small, the iterates* p^k *of Algorithm 7.2 converge superlinearly in* $\mathcal{H}$ *to the solution* p_c *of (7.3.8) as* $k \to \infty$. *Additionally, if* $\mathcal{A}^+_{i,k+1} = \mathcal{A}^+_{i,k}$ *and* $\mathcal{A}^-_{i,k+1} = \mathcal{A}^-_{i,k}$ *holds for all* $i = 1, \dots, N$, *then* $p^{k+1} = p_c$.

Finally, arguing as in Remark 7.3.2, we obtain that the Moreau–Yosida regularization of the predual problem is equivalent to adding the penalty term $\frac{1}{2c}\|\nabla u\|^2_{L^2}$ to the primal problem ($\mathcal{P}_{BV}$) and minimizing over $u \in H^1(\Omega)$.

7.4 NUMERICAL RESULTS

Traditionally, optimal control problems for partial differential equations are formulated with quadratic control costs. When the cost is proportional to the control (or its gradient), it is of interest how this change affects the structure of the optimal controls. The numerical results for a simple model problem presented in this section allow a comparison. Specifically, we consider $A = -\Delta$, the Laplacian with homogeneous Dirichlet conditions on the domain $\Omega = [-1, 1]^2 \subset \mathbb{R}^2$. The differential operators were discretized using standard finite differences on a 128 by 128 grid. To ensure symmetry of the system matrices, the adjoints of A and $-\operatorname{div}$ were taken as the transpose of the corresponding discretization. The implementation was done in Matlab. We consider the following two targets, shown in Figure 7.1:

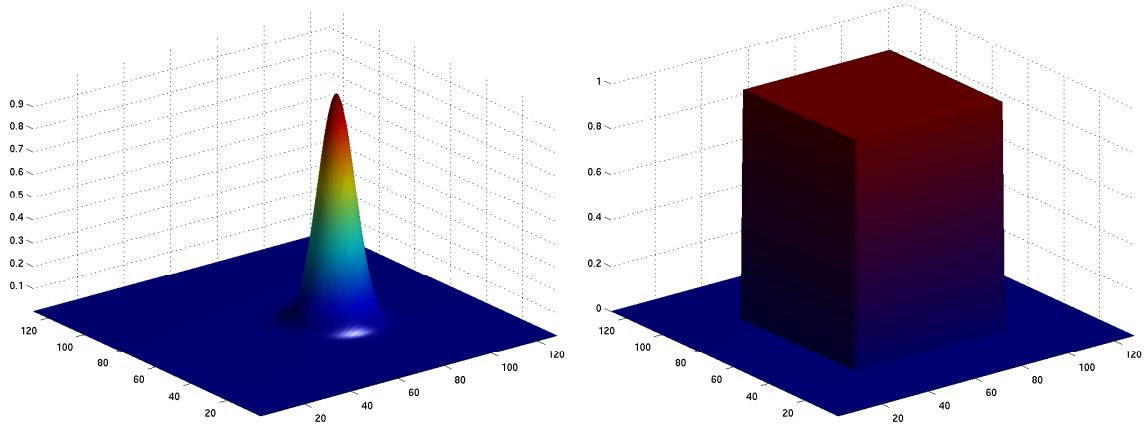
$$z_a(x, y) = e^{-50[(x-0.2)^2+(y+0.1)^2)]},$$
$$z_b(x, y) = \chi_{\{|x|<\frac{1}{2}, |y|<\frac{1}{2}\}}.$$

The solutions of the predual problem for these targets will be denoted by p^*_a and p^*_b respectively, and similarly for the resulting optimal controls u^*_a, u^*_b and corresponding states y^*_a, y^*_b.

CONTROLS IN $\mathcal{M}(\Omega)$ We set $\alpha = 10^{-3}$ and $c = 10^7$ and compute the solution of (7.3.2) using Algorithm 7.1. The solution of problem ($\mathcal{P}_{\mathcal{M}}$) is then obtained using the first relation of (7.2.2). The resulting optimal control and corresponding optimal state are shown in Figure 7.3. The sparsity of the optimal control can be seen (cf. Figure 7.2, and also Corollary 7.2.7): The control is zero wherever the dual variable is inactive (i.e., $-\alpha < p^* < \alpha$), negative on the set where the upper bound is active (i.e., $p^* \geqslant \alpha$), and positive where the lower bound is active (i.e., $p^* \leqslant -\alpha$). Note that the solution p^* of the regularized problem is allowed to be infeasible, although this can be controlled with larger c.

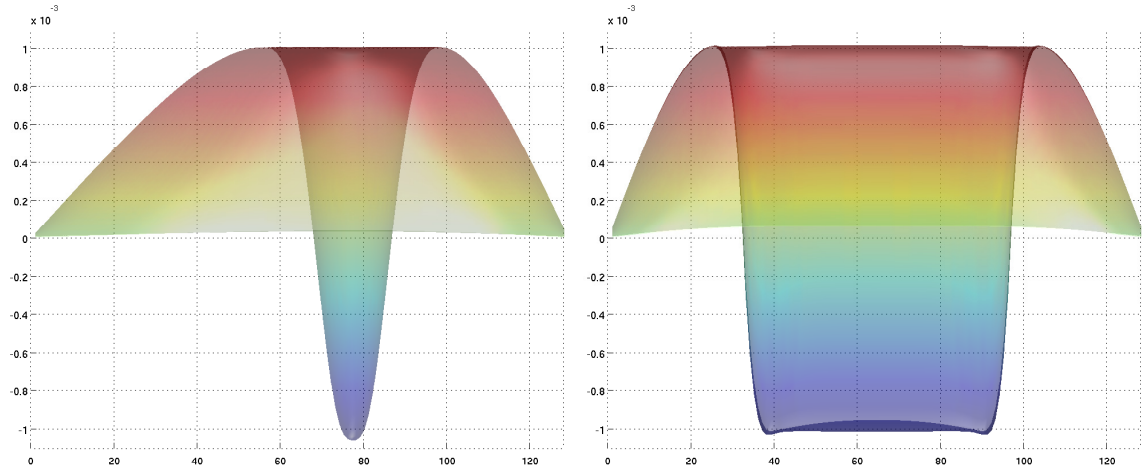
We compare the controls obtained in $\mathcal{M}(\Omega)$ with the solution of the control problem in $L^2(\Omega)$:

$$(\mathcal{P}_{L^2}) \qquad \begin{cases} \min\limits_{u \in L^2(\Omega)} \dfrac{1}{2}\|y - z\|^2_{L^2} + \dfrac{\alpha}{2}\|u\|^2_{L^2} \\ \text{s.t.} \quad Ay = u. \end{cases}$$



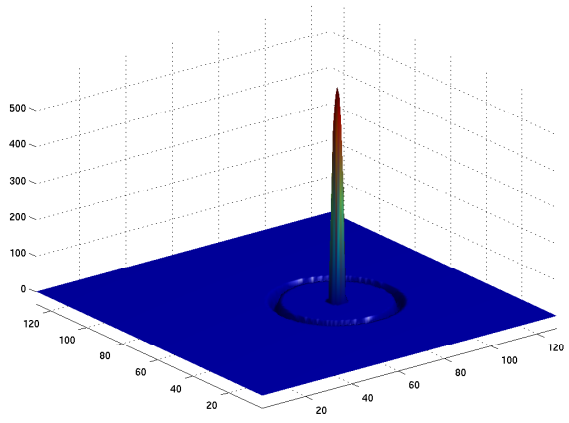
(a) z_a (b) z_b

Figure 7.1: Test targets.

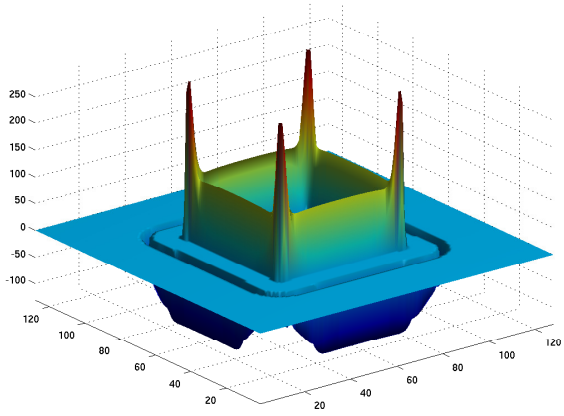


(a) p_a^* (b) p_b^*

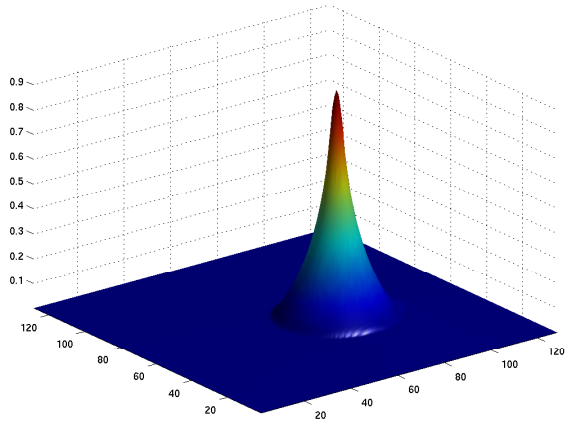
Figure 7.2: Solutions $p^* \in \mathcal{W}$ of $(\mathcal{P}^*_{\mathcal{M},c})$ for $\alpha = 10^{-3}$, $c = 10^7$. Shown is the projection along x_2.



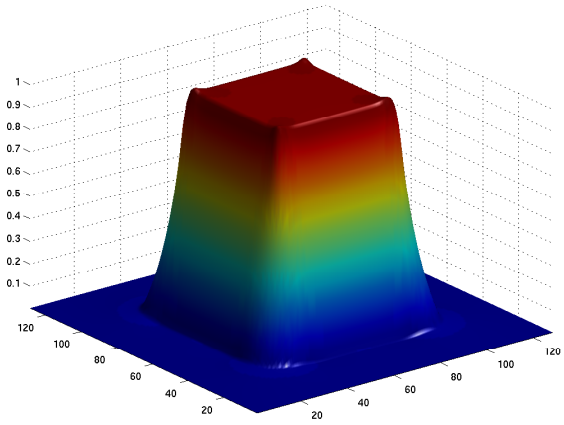
(a) Optimal control $u_a^* \in \mathcal{M}(\Omega)$



(b) Optimal control $u_b^* \in \mathcal{M}(\Omega)$



(c) Optimal state y_a^*



(d) Optimal state y_b^*

Figure 7.3: Solutions of problem ($\mathcal{P}_{\mathcal{M}}$) calculated via predual problem ($\alpha = 10^{-3}$, $c = 10^7$).

Table 7.1: Convergence of iterates p^k in semismooth Newton method for problem ($\mathcal{P}^*_{\mathcal{M},c}$). Given is the error quotient e_k defined by (7.4) for the final iterates.

k	21	22	23	24	25	26	27
e_k	0.9633	0.6766	0.6164	0.5099	0.3467	0.1048	0

The solution is computed from the optimality system, obtained using standard Lagrangian techniques, which can be written in reduced form as $u + \alpha AA^*u = Az$, if $z \in \mathcal{W}$. The corresponding optimal control and state for the same value of $\alpha = 10^{-3}$ are given in Figure 7.4. We point out that a better approximation of the target is possible with controls in $\mathcal{M}(\Omega)$ (compare the height of the peaks for target z_a). Note also that the optimal control in $L^2(\Omega)$ is nonzero almost everywhere, while $u^* \in \mathcal{M}(\Omega)$ is sparse.

We illustrate the superlinear convergence of the Newton method for the case of target z_a with the same parameters as given above. Table 7.1 gives the error quotients

$$e_k := \frac{\left\|p^{k+1} - p^*\right\|_{\mathcal{W}}}{\left\|p^k - p^*\right\|_{\mathcal{W}}}$$

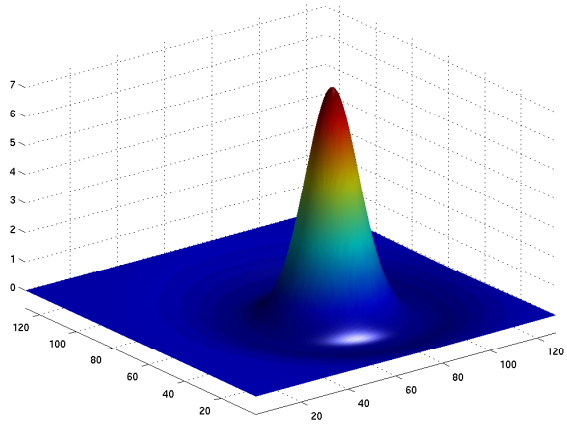
for the final iterates p^k in the semismooth Newton algorithm 7.1. The quotients decrease monotonically, verifying the local superlinear convergence shown in Theorem 7.3.3.

CONTROLS IN $BV(\Omega)$ We repeat the computations for problem ($\mathcal{P}^*_{BV,c}$), setting $\alpha = 10^{-4}$, $\beta = 10^{-1}$ and $c = 10^7$. Here, we compare with the optimal control in $H^1_0(\Omega)$:

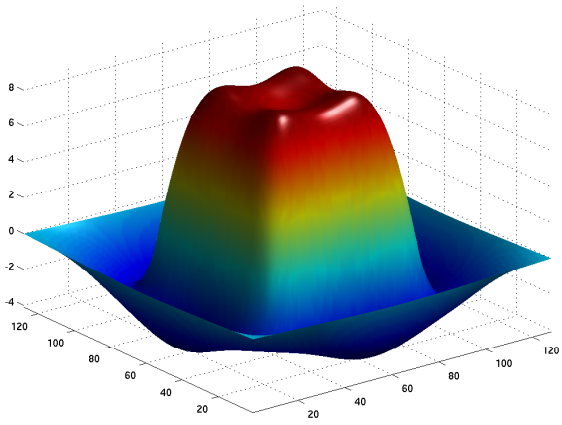
$$(\mathcal{P}_{H^1}) \qquad \begin{cases} \min\limits_{u \in H^1_0(\Omega)} \dfrac{1}{2}\|y - z\|^2_{L^2} + \dfrac{\alpha}{2}\|u\|^2_{H^1(\Omega)} \\ \text{s.t.} \quad Ay = u, \end{cases}$$

where the solution is computed from $u - \alpha AA^*\Delta u = Az$, if $z \in \mathcal{W}$. The resulting controls in $BV(\Omega)$ and $H^1_0(\Omega)$ and corresponding states are given in Figure 7.5 and 7.6, respectively. It can be clearly seen that the $BV(\Omega)$ cost favors piecewise constant controls; this phenomenon is well-known in the mathematical imaging community as staircasing (cf. [Ring 2000]). Again, the target is better attained by controls in $BV(\Omega)$ compared with $H^1_0(\Omega)$.

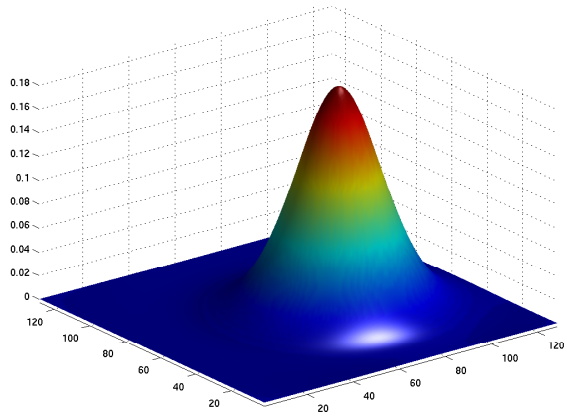
To show the effect of the control cost in the case of $BV(\Omega)$, we increase α from 10^{-4} to 10^{-3} for the target z_b. The optimal control and corresponding state are shown in Figure 7.7). The control is now constant over a much larger region, at the cost of being further from the target. For this example, we can easily illustrate Corollary 7.2.14: The derivative with respect to x_1 or x_2 of the control u^* is zero in regions where the box constraint for the corresponding component of the predual solution $p^* = (p^*_1, p^*_2)$ is inactive (cf. Figure 7.8).



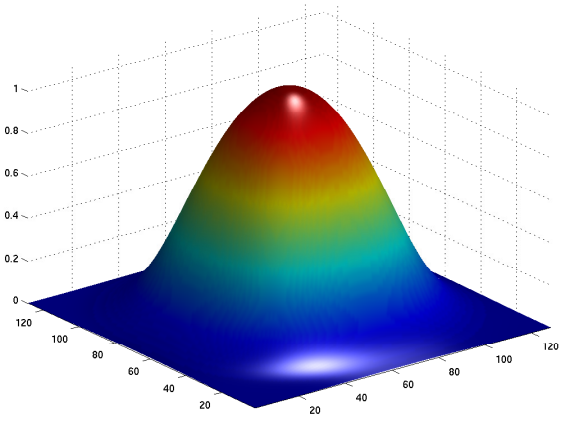
(a) Optimal control $u_a^* \in L^2(\Omega)$



(b) Optimal control $u_b^* \in L^2(\Omega)$

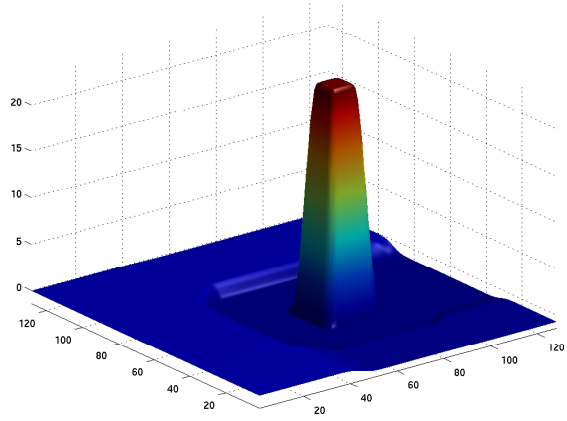


(c) Optimal state y_a^*

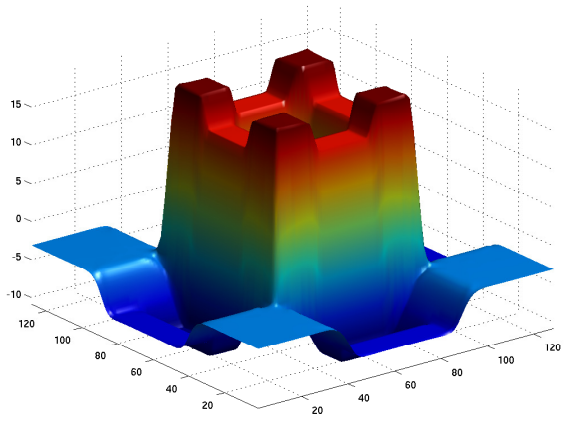


(d) Optimal state y_b^*

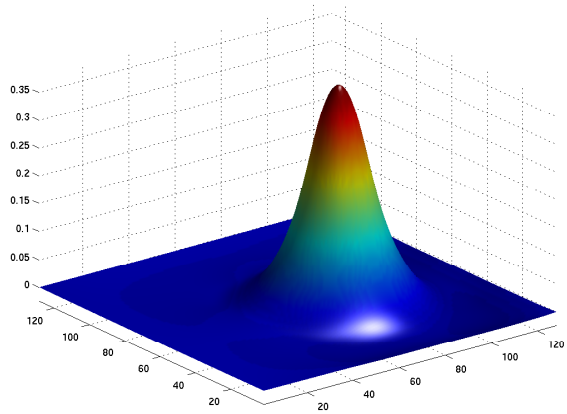
Figure 7.4: Solutions of problem ($\mathcal{P}_{L^2}$) for $\alpha = 10^{-3}$.



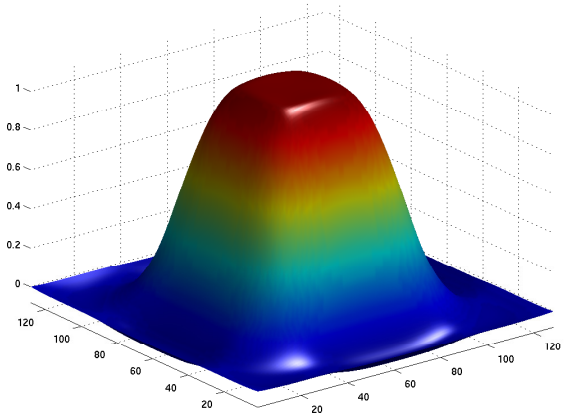
(a) Optimal control $u_a^* \in BV(\Omega)$



(b) Optimal control $u_b^* \in BV(\Omega)$

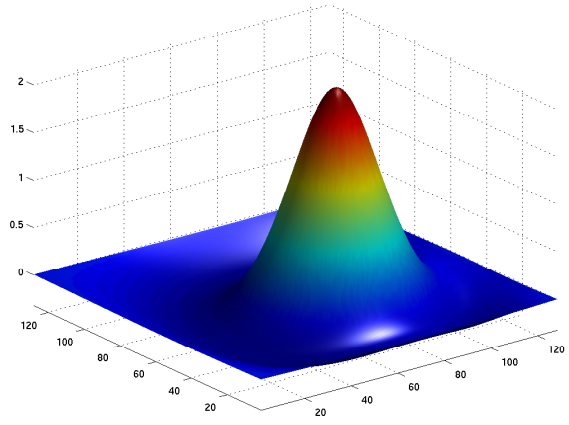


(c) Optimal state y_a^*

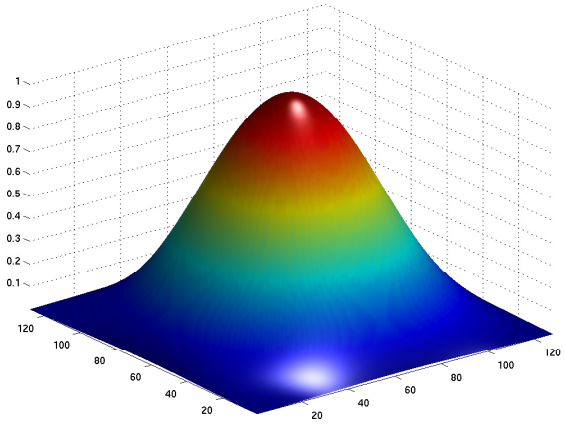


(d) Optimal state y_b^*

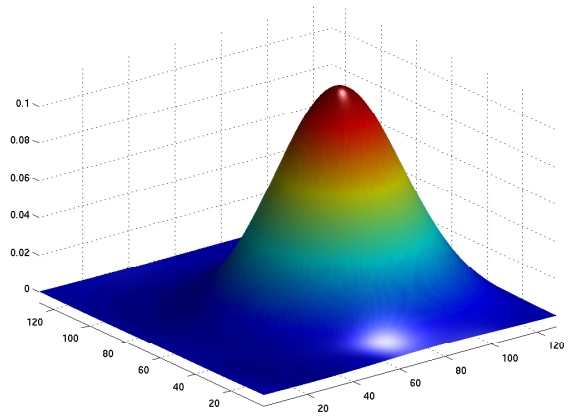
Figure 7.5: Solutions of problem ($\mathcal{P}_{BV}$) calculated via predual problem ($\alpha = 10^{-4}$, $\beta = 10^{-1}$, $c = 10^7$).



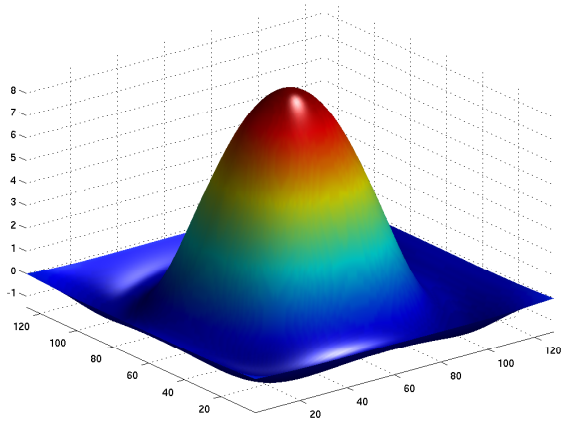
(a) Optimal control $u_a^* \in H_0^1(\Omega)$



(b) Optimal control $u_b^* \in H_0^1(\Omega)$

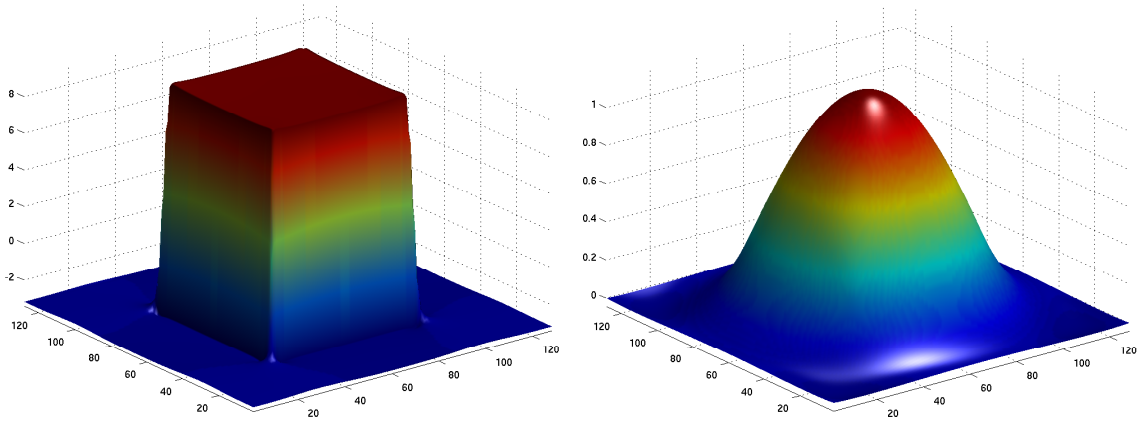


(c) Optimal state y_a^*



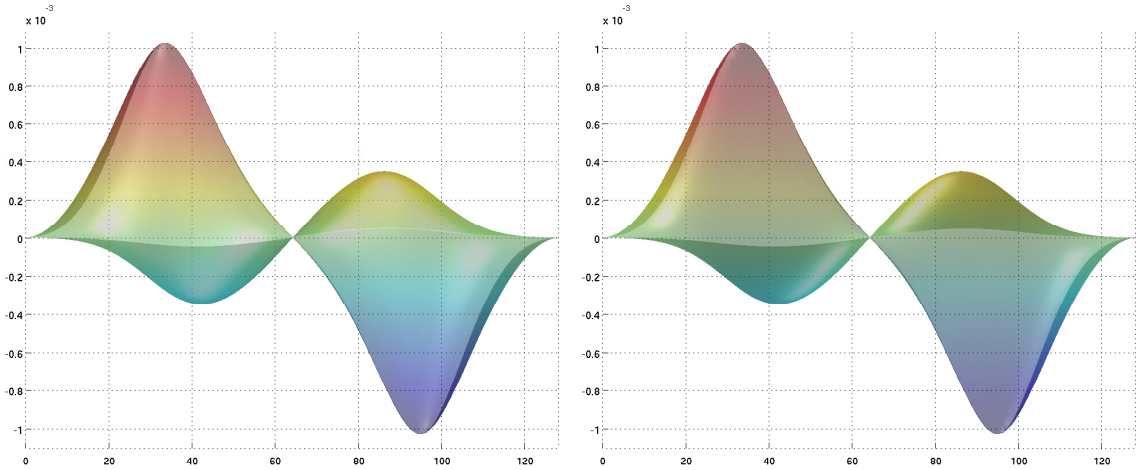
(d) Optimal state y_b^*

Figure 7.6: Solutions of problem ($\mathcal{P}_{H^1}$) for $\alpha = 10^{-4}$.



(a) Optimal control $u_b^* \in BV(\Omega)$ (b) Optimal state y_b^*

Figure 7.7: Solution of problem ($\mathcal{P}_{BV}$) for target z_b with $\alpha = 10^{-3}$.



(a) Projection of p_1^* along x_2 (b) Projection of p_2^* along x_1

Figure 7.8: Solution $p_b^* = (p_1^*, p_2^*) \in H^2_{div}(\Omega)$ of ($\mathcal{P}^*_{BV,c}$) with $\alpha = 10^{-3}$, $c = 10^7$.

7.5 CONCLUSION

We have presented a framework for the efficient solution of elliptic optimal control problems in non-reflexive Banach spaces such as the space of bounded Radon measures or of functions of bounded variation. Specifically, the Fenchel duality theorem allows the reformulation of these (non-differentiable) problems as smooth box constrained problems in a Hilbert space. The corresponding optimality systems can be solved with a semismooth Newton method, which converges superlinearly after regularization. We also demonstrated the structural differences between the optimal controls for these nonsmooth problems and the solutions of the corresponding quadratic formulations (i.e., $L^2(\Omega)$ and $H_0^1(\Omega)$).

The proposed approach can be extended to time-dependent (e.g., parabolic) problems as well, which will be investigated in a subsequent work.

7.A CONVERGENCE OF MOREAU–YOSIDA REGULARIZATION

In this appendix we consider the regularization of ($\mathcal{P}^*_{BV,L^2}$):

$$\min_{p\in\mathcal{H}} \frac{1}{2}\|A^*\operatorname{div} p + z\|_{L^2}^2 + \frac{\beta}{2}\|p\|_{\mathbb{L}^2}^2 + \frac{1}{2c}\|p\|_{\mathcal{W}^n}^2 - \frac{1}{2}\|z\|_{L^2}^2 + \frac{1}{2c}\|\max(0, c(p-\alpha))\|_{\mathbb{L}^2}^2 + \frac{1}{2c}\|\min(0, c(p+\alpha))\|_{\mathbb{L}^2}^2,$$

$\beta > 0$ fixed. For the sake of presentation, we have used the same parameter for the Moreau–Yosida regularization and the $\mathcal{W}^n$-smoothing term, but the same result holds if we replace $\frac{1}{2c}\|p\|_{\mathcal{W}^n}^2$ by $\frac{\gamma}{2}\|p\|_{\mathcal{W}^n}^2$ and take the limit as $c \to \infty, \gamma \to 0$.

The corresponding optimality system is

$$\text{(7.A.1)} \qquad \begin{cases} \langle A^*\operatorname{div} p_c, A^*\operatorname{div} v\rangle_{L^2} + \beta\langle p_c, v\rangle_{\mathbb{L}^2} + \frac{1}{c}\langle \Delta p_c, \Delta v\rangle_{\mathbb{L}^2} \\ \qquad\qquad - \langle z, A^*\operatorname{div} v\rangle_{L^2} + \langle \lambda_c, v\rangle_{\mathcal{H}^*,\mathcal{H}} = 0, \\ \lambda_c = \max(0, c(p_c-\alpha)) + \min(0, c(p_c+\alpha)), \end{cases}$$

for all $v \in \mathcal{H}$. This equation has a unique solution $(\lambda_c, p_c) \in \mathcal{H}^* \times \mathcal{H}$. We wish to show convergence as $c \to \infty$ to the unique solution $(\lambda^*, p^*) \in H^2_{\mathrm{div}}(\Omega)^* \times H^2_{\mathrm{div}}(\Omega)$ of the optimality system for ($\mathcal{P}^*_{BV,L^2}$):

$$\text{(7.A.2)} \qquad \begin{cases} \langle A^*\operatorname{div} p^*, A^*\operatorname{div} v\rangle_{L^2} + \beta\langle p^*, v\rangle_{\mathbb{L}^2} - \langle z, A^*\operatorname{div} v\rangle_{L^2} + \langle \lambda^*, v\rangle_{H^{2\,*}_{\mathrm{div}},H^2_{\mathrm{div}}} = 0, \\ \langle \lambda^*, p^* - p\rangle_{H^{2\,*}_{\mathrm{div}},H^2_{\mathrm{div}}} \leqslant 0, \end{cases}$$

for all $p \in H^2_{\mathrm{div}}(\Omega)$ with $\|p\|_{\mathbb{L}^\infty} \leqslant \alpha$.

The proof is similar to that of Theorem 7.3.1. Arguing as above, we have that

$$\langle \lambda_c, p_c \rangle_{L^2} \geqslant \frac{1}{c} \|\lambda_c\|_{L^2}^2 .$$

and hence from (7.A.1) that

$$\|A^* \operatorname{div} p_c\|_{L^2}^2 + \frac{1}{c} \|p_c\|_{\mathcal{W}^n}^2 + \beta \|p_c\|_{\mathbb{L}^2}^2 + \frac{1}{c} \|\lambda_c\|_{L^2}^2 \leqslant \|A^* \operatorname{div} p_c\|_{L^2} \|z\|_{L^2} .$$

This implies that the $H^2_{\mathrm{div}}(\Omega)$-norm of p_c is bounded uniformly in c by $\|z\|_{L^2}$, and that

$$\|\lambda_c\|_{\mathcal{H}^*} \leqslant 2 \sup_{\substack{\nu \in \mathcal{H}, \\ \|\nu\|_{\mathcal{H}} \leqslant 1}} \|A^* \operatorname{div} \nu\|_{L^2} \|z\|_{L^2} < \infty.$$

It follows that (λ_c, p_c) converges weakly subsequentially in $\mathcal{H}^* \times H^2_{\mathrm{div}}(\Omega)$ to a $(\tilde{\lambda}, \tilde{p}) \in \mathcal{H}^* \times H^2_{\mathrm{div}}(\Omega)$, which satisfies

$$\langle A^* \operatorname{div} \tilde{p}, A^* \operatorname{div} \nu \rangle_{L^2} + \beta \langle \tilde{p}, \nu \rangle_{\mathbb{L}^2} - \langle z, A^* \operatorname{div} \nu \rangle_{L^2} + \langle \tilde{\lambda}, \nu \rangle_{\mathcal{H}^*, \mathcal{H}} = 0 \tag{7.A.3}$$

for all $\nu \in \mathcal{H}$. By the density of $\mathcal{H}$ in $H^2_{\mathrm{div}}(\Omega)$ (via $(C_0^\infty(\Omega))^n \subset \mathcal{H}$), equation (7.A.3) holds for all $\nu \in H^2_{\mathrm{div}}(\Omega)$, and we can identify $\tilde{\lambda}$ with an element in $H^2_{\mathrm{div}}(\Omega)^*$ (replacing the duality pairing with the one in $H^2_{\mathrm{div}}(\Omega)$).

The feasibility of $\tilde{p}$ follows exactly as in the proof of Theorem 7.3.1. Similarly, we deduce from the optimality of p_c and the density of $\mathcal{H}$ in $H^2_{\mathrm{div}}(\Omega)$ that

$$\limsup_{c \to \infty} \frac{1}{2} \|A^* \operatorname{div} p_c + z\|_{L^2}^2 + \frac{\beta}{2} \|p_c\|_{\mathbb{L}^2}^2 \leqslant \frac{1}{2} \|A^* \operatorname{div} \tilde{p} + z\|_{L^2}^2 + \frac{\beta}{2} \|\tilde{p}\|_{\mathbb{L}^2}^2$$

holds. The convergence of p_c to $\tilde{p}$ is therefore strong in $H^2_{\mathrm{div}}(\Omega)$, and we can pass to the limit in

$$\langle \lambda_c, p - p_c \rangle_{\mathcal{H}^*, \mathcal{H}} \leqslant 0 \quad \text{for all feasible } p \in \mathcal{H},$$

to obtain

$$\langle \tilde{\lambda}, p - \tilde{p} \rangle_{\mathcal{H}^*, \mathcal{H}} \leqslant 0 \quad \text{for all feasible } p \in \mathcal{H}.$$

Again, the density of $\mathcal{H}$ in $H^2_{\mathrm{div}}(\Omega)$ and the fact that $\tilde{\lambda} \in H^2_{\mathrm{div}}(\Omega)^*$ allows us to conclude that

$$\langle \tilde{\lambda}, p - \tilde{p} \rangle_{{H^2_{\mathrm{div}}}^*, H^2_{\mathrm{div}}} \leqslant 0$$

holds for all feasible $p \in H^2_{\mathrm{div}}(\Omega)$. Therefore, $(\tilde{\lambda}, \tilde{p}) \in H^2_{\mathrm{div}}(\Omega)^* \times H^2_{\mathrm{div}}(\Omega)$ satisfies (7.A.2), and $\tilde{p} = p^*$ and $\tilde{\lambda} = \lambda^*$ follows from the uniqueness of its solution.

A MEASURE SPACE APPROACH TO OPTIMAL SOURCE PLACEMENT

8

ABSTRACT

The problem of optimal placement of point sources is formulated as a distributed optimal control problem with sparsity constraints. For practical relevance, partial observations as well as partial and non-negative controls need to be considered. Although well-posedness of this problem requires a non-reflexive Banach space setting, a primal-predual formulation of the optimality system can be approximated well by a family of semismooth equations, which can be solved by a superlinearly convergent semismooth Newton method. Numerical examples indicate the feasibility for optimal light source placement problems in diffusive photochemotherapy.

8.1 INTRODUCTION

This work is concerned with the (formal) optimal control problem

$$(\mathcal{P}) \qquad \begin{cases} \min\limits_{y,u} \dfrac{1}{2} \|y|_{\omega_o} - z\|^2_{L^2(\omega_o)} + \alpha \|u\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} \\ \text{subject to } Ay = \chi_{\omega_c} u, \quad y|_{\partial\Omega} = 0, \end{cases}$$

where A is a linear second-order elliptic operator, ω_o and ω_c represent the observation and control subdomains of the bounded domain $\Omega \subset \mathbb{R}^n$ with characteristic function χ_{ω_o} and χ_{ω_c}, respectively, and $z \in L^2(\omega_o)$ is given. For convenience, we abbreviate $\Gamma = \partial\Omega$. Furthermore $\mathcal{M}_\Gamma(\overline{\omega}_c)$ denotes the topological dual of $C_\Gamma(\overline{\omega}_c) := \{v \in C(\overline{\omega}_c) : v|_{\partial\omega_c \cap \Gamma} = 0\}$, where the constraint $v|_{\partial\omega_c\cap\Gamma} = 0$ is dropped if $\partial\omega_c \cap \Gamma = \emptyset$. The norm on $\mathcal{M}_\Gamma(\overline{\omega}_c)$ is given by

$$\|u\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} = \sup_{\substack{\varphi \in C_\Gamma(\overline{\omega}_c) \\ \|\varphi\|_{C_\Gamma(\overline{\omega}_c)} \leqslant 1}} \int_{\omega_c} \varphi \, du, \tag{8.1.1}$$

which coincides with $\|u\|_{L^1(\omega_c)}$ if $u \in L^1(\omega_c)$ (identified with a subspace of $\mathcal{M}_\Gamma(\overline{\omega}_c)$) holds. Since $\overline{\omega}_c \setminus \Gamma$ is a locally compact Hausdorff space, the Riesz representation theorem allows identifying elements of $\mathcal{M}_\Gamma(\overline{\omega}_c)$ with Radon measures that have compact support in $\overline{\omega}_c \setminus \Gamma$ (cf. [Elstrodt 2005, Th. VIII.2.19]).

The problem is motivated by the question of optimal source placement, e. g., in diffusive optical tomography, since the L^1 norm is known to promote sparsity in optimization. The connection between L^1 control costs and source placement was first discussed in [Stadler 2009]. However, problem ($\mathcal{P}$) is not well-posed in L^1, since L^1 lacks the necessary weak compactness properties. Problems with L^1 control cost and L^∞ control constraints were considered in [Stadler 2009], [Wachsmuth and Wachsmuth 2011a], [Wachsmuth and Wachsmuth 2011b] and [Casas, Herzog, and Wachsmuth 2012], while a measure space setting was first investigated in [Clason and Kunisch 2011].

In this work, we address the feasibility of optimal source placement by optimal control in measure spaces by including partial observation, control on subdomains and non-negativity of the controls, which was not considered in the previously cited works. The Fenchel predual framework as utilized in [Clason and Kunisch 2011] is not applicable in this situation, so we consider a primal-predual setting. This framework can be modified to allow for nonlinear control-to-state mappings, which also do not fit into the earlier Fenchel duality framework.

This paper is organized as follows. In section 8.2, we discuss the well-posedness of the optimal control problem for measure source terms defined on subdomains and derive the optimality system. Section 8.3 is devoted to the regularization of the optimality system and addresses the convergence of the regularized solutions to those of the original problem. The numerical solution using a semismooth Newton method is discussed in section 8.4. Finally, in section 8.5 we give numerical examples to indicate the feasibility of the proposed approach for a problem of optimal light source placement in photochemotherapy.

Throughout, we take as $W_0^{1,r}(\Omega)$ the closure of $\{v \in C^\infty(\Omega) : v|_{\partial\Omega} = 0\}$ in the $W^{1,r}(\Omega)$ norm, $r \in (1,\infty)$. We denote by $W^{-1,r'}(\Omega) = (W_0^{1,r}(\Omega))^*$ the topological dual of $W_0^{1,r}(\Omega)$. Moreover, for $\omega \subset \Omega$ we set $W^{1,r}(\omega) = \{\varphi|_\omega : \varphi \in W_0^{1,r}(\Omega)\}$ with dual denoted by $(W^{1,r}(\omega))^*$.

8.2 PROBLEM FORMULATION AND OPTIMALITY SYSTEM

We first address the well-posedness of the state equation. Let $\mathcal{M}(\Omega)$ denote the topological dual of $C_0(\overline{\omega})$ endowed with the operator norm, cf. (8.1.1). By the Riesz representation theorem (e.g., [Elstrodt 2005, Th. VIII.2.10]), $\mathcal{M}(\Omega)$ can be identified with the Banach space of finite Radon measures. We further choose $q \in (1, \frac{n}{n-1})$ and set $q' = \frac{q-1}{q} \in (n,\infty)$. For this choice, we have $W_0^{1,q'}(\Omega) \hookrightarrow C_0(\overline{\omega})$, and this embedding is compact.

We consider the operator

$$Ay = -\sum_{j,k=1}^{n} \partial_j(a_{jk}(x)\partial_k y + d_j(x)y) + \sum_{j=1}^{n} b_j(x)\partial_j y + d(x)y,$$

and for $\mu \in \mathcal{M}(\Omega)$ the abstract Dirichlet problem

$$\begin{cases} Ay = \mu, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \tag{8.2.1}$$

which is to be interpreted in variational form, i.e., y satisfies

$$-\sum_{j,k=1}^{n} \langle a_{jk}\partial_j y, \partial_k v\rangle_{L^2} + \sum_{j=1}^{n} \langle b_j \partial_j y, v\rangle_{L^2} + \sum_{k=1}^{n} \langle y, d_k \partial_k v\rangle_{L^2} + \langle dy, v\rangle_{L^2} = \int_\Omega v \, d\mu \tag{8.2.2}$$

for all $v \in W_0^{1,q'}(\Omega)$. Here, Ω is a bounded domain in $\mathbb{R}^n$ with $C^{1,\delta}$ boundary $\partial\Omega$, $a_{jk}, b_j \in C^{0,\delta}(\overline{\omega})$ for some $\delta \in (0,1)$, $d_j, d \in L^\infty(\Omega)$, and it is assumed that 0 is not an eigenvalue of A (e.g., A is uniformly elliptic and the lower order coefficients are small enough, cf. [Gilbarg and Trudinger 2001, Th. 8.3]). These assumptions imply that the adjoint A^* of A is an isomorphism from $W_0^{1,q'}(\Omega)$ to $W^{-1,q'}(\Omega)$, see, e.g., [Troianiello 1987, Th. 3.16], [Gallouët and Monier 1999]. Consequently, A is an isomorphism from $W_0^{1,q}(\Omega)$ to $W^{-1,q}(\Omega)$. In particular, (8.2.1) admits a unique solution satisfying

$$\|y\|_{W_0^{1,q}(\Omega)} \leqslant C \, \|\mu\|_{\mathcal{M}(\Omega)}$$

for a constant C independent of μ, by the fact that $\mathcal{M}(\Omega)$ embeds continuously into $W^{-1,q}(\Omega)$ (see, e. g., [Stampacchia 1965, Th. 9.1] and [Meyer, Panizzi, and Schiela 2011, Th. 4.1]). We refer to [Meyer, Panizzi, and Schiela 2011] for a discussion of the various (equivalent) characterizations of solutions to (8.2.2) and their uniqueness if A^* is not surjective on $W^{-1,q'}(\Omega)$.

We now define the control-to-state mapping associated to ($\mathcal{P}$). For this purpose, let

$$R_{\omega_o} : W_0^{1,q}(\Omega) \to W^{1,q}(\omega_o), \qquad R_{\omega_c} : W_0^{1,q'}(\Omega) \to W^{1,q'}(\omega_c)$$

denote the canonical restriction operators from Ω to ω_c and ω_o, respectively, with adjoints

$$R_{\omega_o}^* : (W^{1,q}(\omega_o))^* \to W^{-1,q'}(\Omega), \qquad R_{\omega_c}^* : (W^{1,q'}(\omega_c))^* \to W^{-1,q}(\Omega).$$

Further we shall employ the injections

$$\mathcal{J}_{\omega_o} : W^{1,q}(\omega_o) \to L^2(\omega_o), \qquad \mathcal{J}_{\omega_c} : W^{1,q'}(\omega_c) \to C_\Gamma(\overline{\omega}_c)$$

with adjoints

$$\mathcal{J}_{\omega_o}^* : L^2(\omega_o) \to (W^{1,q}(\omega_o))^*, \qquad \mathcal{J}_{\omega_c}^* : \mathcal{M}_\Gamma(\overline{\omega}_c) \to (W^{1,q'}(\omega_c))^*.$$

Then we set

$$S_\omega : \mathcal{M}_\Gamma(\overline{\omega}_c) \to L^2(\omega_o), \qquad u \mapsto \mathcal{J}_{\omega_o} R_{\omega_o} A^{-1} R^*_{\omega_c} \mathcal{J}^*_{\omega_c} u,$$

and note that S_ω is a bounded linear operator. Since R_{ω_o}, R_{ω_c}, $\mathcal{J}_{\omega_o}$ and $\mathcal{J}_{\omega_c}$ have dense ranges, their adjoints are injections. To argue that $\mathcal{J}_{\omega_c}$ has dense range, let $\varphi \in C_\Gamma(\overline{\omega}_c)$. By Tietze's extension theorem (e. g., [DiBenedetto 2002, Th. 3.1]), there exists a $\widetilde{\varphi} \in C_\Gamma(\overline{\omega}_c)$ with $\widetilde{\varphi}|_{\partial\Omega} = 0$ and $\widetilde{\varphi}|_{\overline{\omega}_c} = \varphi$ (see also [Troianiello 1987, Th. 1.N]). Moreover $\widetilde{\varphi}$ can be approximated by $\widetilde{\varphi}_n \in W_0^{1,q'}(\Omega)$ in the $W_0^{1,q'}(\Omega)$-norm, and hence $\widetilde{\varphi}_n|_{\overline{\omega}_c} \in W^{1,q'}(\omega_c)$ approximates $\varphi \in C_\Gamma(\overline{\omega}_c)$.

We will also need the following continuity property of the control-to-state mapping.

Proposition 8.2.1. *For any sequence* $\{u_k\} \subset \mathcal{M}_\Gamma(\overline{\omega}_c)$ *converging weakly-⋆ in* $\mathcal{M}_\Gamma(\overline{\omega}_c)$*, the sequence* $S_\omega(u_k)$ *converges strongly to* $S_\omega(u)$ *in* $L^2(\omega_o)$.

Proof. Since $q' > n$, the embedding $W^{1,q'}(\omega_c) \hookrightarrow C_\Gamma(\overline{\omega}_c)$ is compact. Therefore, the adjoint embedding $\mathcal{M}_\Gamma(\overline{\omega}_c) \hookrightarrow (W^{1,q'}(\omega_c))^*$ is compact as well. Weak-⋆ convergence of u_k in $\mathcal{M}_\Gamma(\overline{\omega}_c)$ thus implies strong convergence of $\mathcal{J}^*_{\omega_c} u_k$. The claim then follows from the continuity of $\mathcal{J}_{\omega_o} R_{\omega_o} A^{-1} R^*_{\omega_c}$. □

The reduced problem corresponding to ($\mathcal{P}$) can then be formulated as

$$\min_{u \in \mathcal{M}_\Gamma(\overline{\omega}_c)} \frac{1}{2} \|S_\omega u - z\|^2_{L^2(\omega_o)} + \alpha \|u\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)}. \tag{8.2.3}$$

Existence of a minimizer u^* follows from the fact that bounded sequences in $\mathcal{M}_\Gamma(\overline{\omega}_c)$ contain a weakly-⋆ convergent subsequence, and that $u \mapsto \|u\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)}$ is weak-⋆ lower semicontinuous.

Remark 8.2.2. If a minimizer u^* satisfies $u^* \in L^1(\omega_c)$, it is also a solution of the problem

$$\min_{u \in L^1(\omega_o)} \frac{1}{2} \|S_\omega u - z\|^2_{L^2(\omega_o)} + \alpha \|u\|_{L^1(\omega_c)}.$$

This follows from the embedding of $L^1(\omega_c)$ into $\mathcal{M}_\Gamma(\overline{\omega}_c)$ and the fact that $\|\nu\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} = \|\nu\|_{L^1(\omega_c)}$ for $\nu \in L^1(\omega_c)$ (cf. [Brezis 2010, Ch. IV]).

We wish to employ a Fenchel duality argument for the derivation of a necessary optimality condition for (8.2.3). To avoid dealing with $(\mathcal{M}_\Gamma(\overline{\omega}_c))^*$, we shall consider a predual of (8.2.3) rather than a dual problem. Such a procedure was previously used in [Bredies and Pikkarainen 2012; Clason and Kunisch 2011; Hintermüller and Kunisch 2004], for example. For this purpose, we introduce

$${}^*S_\omega : L^2(\omega_o) \to C_\Gamma(\overline{\omega}_c), \qquad \varphi \mapsto \mathcal{J}_{\omega_c} R_{\omega_c} (A^*)^{-1} R^*_{\omega_o} \mathcal{J}^*_{\omega_o} \varphi,$$

noting that

$$({}^*S_\omega)^* = S_\omega,$$

i. e., ${}^*S_\omega$ is the "preadjoint" to S_ω.

Theorem 8.2.3. *Let $u^* \in \mathcal{M}_\Gamma(\overline{\omega}_c)$ be a solution to (8.2.3). Then there exists a $p^* \in C_\Gamma(\overline{\omega}_c)$ satisfying*

$$\text{(OS)} \qquad \begin{cases} {}^*S_\omega(S_\omega u^* - z) = p^*, \\ \langle u^*, p^* - p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0, \qquad \|p^*\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha, \end{cases}$$

for all $p \in C_\Gamma(\overline{\omega}_c)$ with $\|p\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$.

Proof. We consider the following problem in $L^2(\omega_o)$, which will be shown to be the predual of (8.2.3):

$$\min_{q\in L^2(\omega_o)} \frac{1}{2}\|q+z\|^2_{L^2(\omega_o)} - \frac{1}{2}\|z\|^2_{L^2(\omega_o)} + I_{\{\|q\|_{C_\Gamma(\overline{\omega}_c)}\leqslant\alpha\}}({}^*S_\omega q) =: \min_{q\in L^2(\omega_o)} \mathcal{F}(q) + \mathcal{G}({}^*S_\omega q), \tag{8.2.4}$$

where $\mathcal{F} : L^2(\omega_o) \to \mathbb{R}$ and $\mathcal{G} : C_\Gamma(\overline{\omega}_c) \to \mathbb{R}\cup\{\infty\}$. A short computation shows that the Fenchel conjugates $\mathcal{F}^* : L^2(\omega_o) \to \mathbb{R}$ and $\mathcal{G}^* : \mathcal{M}_\Gamma(\overline{\omega}_c) \to \mathbb{R}$ are given by

$$\mathcal{F}^*(v) = \frac{1}{2}\|v - z\|^2_{L^2(\omega_o)}, \qquad \mathcal{G}^*(v) = \alpha\|v\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)}.$$

Since $q \mapsto \mathcal{F}(q)+\mathcal{G}({}^*S_\omega q)$ is continuous at 0, the Fenchel duality theorem (see, e. g., [Ekeland and Témam 1999, Th. 4.1]) is applicable and implies that

$$\begin{aligned} \min_{q\in L^2(\omega_o)} \mathcal{F}(q) + \mathcal{G}({}^*S_\omega q) &= \min_{u\in\mathcal{M}_\Gamma(\overline{\omega}_c)} \mathcal{F}^*(S_\omega u) + \mathcal{G}^*(-u) \\ &= \min_{u\in\mathcal{M}_\Gamma(\overline{\omega}_c)} \frac{1}{2}\|S_\omega u - z\|^2_{L^2(\omega_o)} + \alpha\|u\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)}, \end{aligned} \tag{8.2.5}$$

where we utilize $({}^*S_\omega)^* = S_\omega$. Moreover (cf. [Ekeland and Témam 1999, Prop. 4.1]), to every minimizer $q^* \in L^2(\omega_o)$ of the left hand side of (8.2.5) corresponds a minimizer $u^* \in \mathcal{M}_\Gamma(\overline{\omega}_c)$ of the right hand side satisfying the relationship

$$\begin{cases} S_\omega u^* = q^* + z, \\ -u^* \in \partial I_{\{\|q\|_{C_\Gamma(\overline{\omega}_c)}\leqslant\alpha\}}({}^*S_\omega q^*). \end{cases}$$

From the second relation, we have $\|{}^*S_\omega q^*\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$ and

$$\langle -u^*, p - {}^*S_\omega q^*\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0 \qquad \text{for all } \|p\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha. \tag{8.2.6}$$

Setting $p^* = {}^*S_\omega q^* = {}^*S_\omega(S_\omega u^* - z)$ we find that

$$\langle u^*, p^* - p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0 \qquad \text{for all } \|p\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$$

and $\|p^*\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$. □

We note that by construction $p^* \in W^{1,q'}(\omega_c)$ holds. From the second relation of (OS), we can also obtain the following structural information on an optimal control u^*.

Corollary 8.2.4. *Let* (u^*, p^*) *be a solution to* (OS). *Then for any* $p \in C_\Gamma(\overline{\omega}_c)$ *with* $p \geqslant 0$,

$$\begin{aligned}
\langle u^*, p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} &= 0 \qquad \text{if } \operatorname{supp}(p) \subset \{x \in \overline{\omega}_c : |p^*(x)| < \alpha\},\\
\langle u^*, p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} &\leqslant 0 \qquad \text{if } \operatorname{supp}(p) \subset \{x \in \overline{\omega}_c : p^*(x) = \alpha\},\\
\langle u^*, p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} &\geqslant 0 \qquad \text{if } \operatorname{supp}(p) \subset \{x \in \overline{\omega}_c : p^*(x) = -\alpha\}.
\end{aligned}$$

This can be interpreted as a sparsity property: An optimal control u^* will be non-zero only on sets where the constraint on p^* is active; hence the larger the penalty α, the smaller the support of the control.

Remark 8.2.5 (Non-negative controls). If in ($\mathcal{P}$) only non-negative controls are admitted, we replace $\mathcal{G}^*(v)$ by

$$\mathcal{G}_+^* : \mathcal{M}_\Gamma(\overline{\omega}_c) \to \mathbb{R} \cup \{\infty\}, \qquad v \mapsto I_{\{f \leqslant 0\}}(v) + \alpha \|v\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)}$$

(noting that the dual problem involves the term $\mathcal{G}_+^*(-u^*)$). This is the Fenchel dual of

$$\mathcal{G}_+ : C_\Gamma(\overline{\omega}_c) \to \mathbb{R} \cup \{\infty\}, \qquad q \mapsto I_{\{f \geqslant -\alpha\}}(q),$$

and (8.2.6) must be replaced by

$$\langle -u^*, p - {}^*S_\omega q^*\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0 \qquad \text{for all } p \geqslant -\alpha.$$

The optimality conditions for the case of non-negative controls become

$$(\mathrm{OS}_+) \qquad \begin{cases} {}^*S_\omega(S_\omega u^* - z) = p^*, \\ \langle u^*, p^* - p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0, \qquad p^* \geqslant -\alpha \end{cases}$$

for all $p \in C_\Gamma(\overline{\omega}_c)$ with $p \geqslant -\alpha$.

8.3 REGULARIZATION

The numerical solution of the optimality system (OS) is based on a Moreau–Yosida regularization of (OS). For given $c > 0$, we search for $(u_c, p_c) \in L^2(\omega_c) \times W^{1,q'}(\omega_c)$ which satisfy

$$(\mathrm{OS}_c) \qquad \begin{cases} p_c = S_\omega^*(S_\omega u_c - z), \\ -u_c = c \max(0, p_c - \alpha) + c \min(0, p_c + \alpha), \end{cases}$$

where the max and min are taken pointwise in $\overline{\omega}_c$. Here, S_ω is considered as an operator from $L^2(\omega_c) \to L^2(\omega_o)$. The action of its adjoint $S_\omega^* : L^2(\omega_o) \to L^2(\omega_c)$ coincides with that of ${}^*S_\omega$. Moreover, the range of S_ω^* is contained in $W^{1,q'}(\omega_c)$.

This regularization can be interpreted as a quadratic penalization of the box constraints in (8.2.4).

Theorem 8.3.1. *There exists a unique solution* $(u_c, p_c) \in L^2(\omega_c) \times W^{1,q'}(\omega_c)$ *of* (OS_c).

Proof. The claim will follow from the fact that (OS_c) are the necessary optimality conditions of the problem

$$(\mathcal{P}_c) \qquad \min_{u \in L^2(\omega_c)} \frac{1}{2} \|S_\omega u - z\|^2_{L^2(\omega_o)} + \alpha \|u\|_{L^1(\omega_c)} + \frac{1}{2c} \|u\|^2_{L^2(\omega_c)}.$$

The cost function in ($\mathcal{P}_c$) is continuous, bounded from below and strictly convex due to the presence of the $L^2(\omega_c)$ term, hence ($\mathcal{P}_c$) admits a unique minimizer $u_c \in L^2(\omega_c)$. To express ($\mathcal{P}_c$) abstractly, we introduce

$$\begin{aligned} \mathcal{F}_c^* : L^2(\omega_c) \to \mathbb{R}, \qquad & u \mapsto \frac{1}{2}\|S_\omega u - z\|^2_{L^2(\omega_o)}, \\ \mathcal{G}_c^* : L^2(\omega_c) \to \mathbb{R}, \qquad & u \mapsto \alpha \|u\|_{L^1(\omega_c)} + \frac{1}{2c} \|u\|^2_{L^2(\omega_c)}. \end{aligned}$$

The optimality condition for ($\mathcal{P}_c$) is given by

$$0 \in S_\omega^*(S_\omega u_c - z) + \partial \mathcal{G}_c^*(u_c)$$

or equivalently,

$$(8.3.1) \qquad \begin{cases} p_c = S_\omega^*(S_\omega u_c - z), \\ -p_c \in \partial \mathcal{G}_c^*(u_c), \end{cases}$$

where the first equation implies $p_c \in W^{1,q'}(\omega_c) \hookrightarrow C_\Gamma(\overline{\omega}_c)$.

We claim that $\mathcal{G}_c^*$ is the Fenchel conjugate of

$$\mathcal{G}_c : L^2(\omega_c) \to \mathbb{R}, \quad p \mapsto \frac{c}{2} \|\max(0, p - \alpha)\|^2_{L^2(\omega_c)} + \frac{c}{2} \|\min(0, p + \alpha)\|^2_{L^2(\omega_c)}.$$

To show this, we compute the Fenchel conjugate of $\mathcal{G}_c$ at $u \in L^2(\omega_c)$, which is defined as

$$\mathcal{G}_c^*(u) = \sup_{q \in L^2(\omega_c)} \langle u, q \rangle_{L^2(\omega_c)} - \mathcal{G}_c(q).$$

The supremum is attained at $p \in L^2(\omega_c)$ if and only if

$$u = \partial \mathcal{G}_c(p) = c \max(0, p - \alpha) + c \min(0, p + \alpha)$$

holds almost everywhere in ω_c. If $u(x) > 0$, the right hand side has to be positive as well, which implies that $u(x) = c(p(x) - \alpha)$ and hence $p(x) = \frac{1}{c}u(x) + \alpha$. Similarly, $u(x) < 0$

yields $p(x) = \frac{1}{c}u(x) - \alpha$. For $u(x) = 0$, we deduce that $-\alpha \leqslant p(x) \leqslant \alpha$ holds. Substituting in the definition of $\mathcal{G}^*$, we have that

$$\begin{aligned}\mathcal{G}_c^*(u) &= \int_{\{u>0\}} u(x)\Big(\frac{1}{c}u(x) + \alpha\Big) - \frac{1}{2c}\max(0, u(x))^2\,dx \\ &\quad + \int_{\{u<0\}} u(x)\Big(\frac{1}{c}u(x) - \alpha\Big) - \frac{1}{2c}\min(0, u(x))^2\,dx \\ &= \frac{1}{2c}\|u\|^2_{L^2(\omega_c)} + \alpha\|u\|_{L^1(\omega_c)}.\end{aligned}$$

Since $\mathcal{G}_c$ is Lipschitz continuous, the second condition in (8.3.1) can be expressed as (cf., e. g., [Attouch, Buttazzo, and Michaille 2006, Th. 9.5.1])

$$u_c \in \partial\mathcal{G}_c(-p_c) = \{c(\max(0, -p_c - \alpha) + \min(0, -p_c + \alpha))\}.$$

Noting that $\max(0, -p) = -\min(0, p)$, the optimality conditions (OS_c) follow.

Turning to uniqueness, let (u_c, p_c) and $(\bar{u}_c, \bar{p}_c)$ be two solutions to (OS_c) and set $(\delta u, \delta p) = (u_c - \bar{u}_c, p_c - \bar{p}_c)$. Then, subtracting the corresponding optimality conditions and taking the inner product with $(\delta u, \delta p)$ implies that

$$\begin{aligned}0 = \|S_\omega \delta u\|^2_{L^2(\omega_o)} &+ c\,\langle \max(0, p_c - \alpha) - \max(0, \bar{p}_c - \alpha), p_c - \bar{p}_c\rangle_{L^2(\omega_c)} \\ &+ c\,\langle \min(0, p_c + \alpha) - \min(0, \bar{p}_c + \alpha), p_c - \bar{p}_c\rangle_{L^2(\omega_c)}\end{aligned} \tag{8.3.2}$$

Since the mappings $p \mapsto \max(0, p)$ and $p \mapsto \min(0, p)$ are monotone, we obtain that the inner products in (8.3.2) are non-negative and thus that $S_\omega \delta u = 0$. Since $\delta p = S_\omega^*(S_\omega \delta u) = 0$ by linearity of state and adjoint equation, we deduce $p_c = \bar{p}_c$ and hence $u_c = \bar{u}_c$ from the second equation of (OS_c). □

Next, we address the convergence of solutions of (OS_c) as $c \to \infty$.

Theorem 8.3.2. *Let $(u_c, p_c) \in L^2(\omega_c) \times W^{1,q'}(\omega_c)$ be solutions of (OS_c) for $c > 0$. Then the family (u_c, p_c) contains a subsequence, denoted by the same symbol, such that*

$$\begin{aligned}u_c &\rightharpoonup^\star u^* \quad \text{in } \mathcal{M}_\Gamma(\overline{\omega}_c), \\ p_c &\to p^* \quad \text{in } W^{1,q'}(\omega_c) \text{ and hence in } C_\Gamma(\overline{\omega}_c),\end{aligned}$$

and (u^*, p^*) is a solution of (OS).

Proof. Since $u_c = 0$ is an admissible control, we have

$$\begin{aligned}\alpha\|u_c\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} &\leqslant \frac{1}{2}\|S_\omega u_c - z\|^2_{L^2(\omega_o)} + \alpha\|u_c\|_{\mathcal{M}_\Gamma(\overline{\omega}_c)} + \frac{1}{2c}\|u_c\|^2_{L^2(\omega_c)} \\ &\leqslant \frac{1}{2}\|z\|^2_{L^2(\omega_o)}.\end{aligned} \tag{8.3.3}$$

The family of minimizers $\{u_c\}_{c>0}$ is thus bounded in $\mathcal{M}_\Gamma(\overline{\omega}_c)$, and hence there exists a subsequence (also denoted by $\{u_c\}$) which converges weakly-$\star$ in $\mathcal{M}_\Gamma(\overline{\omega}_c)$ to a $\tilde{u} \in \mathcal{M}_\Gamma(\overline{\omega}_c)$. Since $S_\omega(u_c) \to S_\omega(\tilde{u})$ strongly in $L^2(\omega_o)$ by Proposition 8.2.1, we deduce from the continuity of S_ω^* that

$$p_c \to \tilde{p} := S_\omega^*(S_\omega \tilde{u} - z)$$

strongly in $W^{1,q'}(\omega_c)$ and hence in $C_\Gamma(\overline{\omega}_c)$.

We next verify the feasibility of $\tilde{p}$. By squaring the second relation of (OS_c) and inspecting pointwise, we obtain that

$$\frac{1}{c}\,\|u_c\|^2_{L^2(\omega_c)} = c\,\|\max(0, p_c - \alpha)\|^2_{L^2(\omega_c)} + c\,\|\min(0, p_c + \alpha)\|^2_{L^2(\omega_c)}\,.$$

From (8.3.3), we have that $\frac{1}{c}\,\|u_c\|^2_{L^2(\omega_c)} \leqslant \|z\|^2_{L^2(\omega_o)}$, so that

$$\|\max(0, p_c - \alpha)\|^2_{L^2(\omega_c)} \leqslant \frac{1}{c}\,\|z\|^2_{L^2(\omega_o)} \to 0,$$
$$\|\min(0, p_c + \alpha)\|^2_{L^2(\omega_c)} \leqslant \frac{1}{c}\,\|z\|^2_{L^2(\omega_o)} \to 0,$$

hold for $c \to \infty$. Since $p_c \to \tilde{p}$ strongly in $C_\Gamma(\overline{\omega}_c)$, this implies that

$$-\alpha \leqslant \tilde{p}(x) \leqslant \alpha \quad \text{for all } x \in \overline{\omega}_c.$$

It remains to pass to the limit in the second equation of (OS_c). Observe that

$$\begin{aligned}\langle -u_c, p - p_c\rangle_{L^2(\omega_c)} &= c\,\langle \max(0, p_c - \alpha), p - p_c\rangle_{L^2(\omega_c)} \\ &\quad + c\,\langle \min(0, p_c + \alpha), p - p_c\rangle_{L^2(\omega_c)} \leqslant 0\end{aligned}$$

holds for all $p \in C_\Gamma(\overline{\omega}_c)$ with $\|p\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$, and thus that

$$\langle \tilde{u}, \tilde{p} - p\rangle_{\mathcal{M}_\Gamma(\overline{\omega}_c), C_\Gamma(\overline{\omega}_c)} \leqslant 0$$

is satisfied for all $p \in C_\Gamma(\overline{\omega}_c)$ with $\|p\|_{C_\Gamma(\overline{\omega}_c)} \leqslant \alpha$. Therefore, $(\tilde{u}, \tilde{p}) \in \mathcal{M}_\Gamma(\overline{\omega}_c) \times C_\Gamma(\overline{\omega}_c)$ satisfies (OS). □

Remark 8.3.3 (Non-negative controls). By a similar argument as in the proof of Theorem 8.3.1, it can be shown that

$$\mathcal{G}^*_{+,c} : L^2(\omega_c) \to \bar{\mathbb{R}}, \qquad \nu \mapsto I_{\{f \leqslant 0\}}(\nu) + \frac{1}{2c}\,\|\nu\|^2_{L^2(\omega_c)} + \alpha\,\|\nu\|_{L^1(\omega_c)}\,,$$

is the Fenchel conjugate of

$$\mathcal{G}_{+,c} : L^2(\omega_c) \to \bar{\mathbb{R}}, \qquad q \mapsto \frac{c}{2}\,\|\min(0, q + \alpha)\|^2_{L^2(\omega_c)}\,,$$

and thus that the corresponding regularized optimality system is

$$(OS_{+,c}) \qquad \begin{cases} p_c = S_\omega^*(S_\omega u_c - z), \\ -u_c = c\,\min(0, p_c + \alpha). \end{cases}$$

The convergence result for $c \to \infty$ as in Theorem 8.3.2 remains valid.

8.4 SEMISMOOTH NEWTON METHOD

For the numerical solution, we consider (OS_c) as the operator equation $F(u_c) = 0$ for $F : L^2(\omega_c) \to L^2(\omega_c)$, given by

$$F(u) = u + c\max(0, S_\omega^*(S_\omega u - z) - \alpha) + c\min(0, S_\omega^*(S_\omega u - z) + \alpha).$$

It is known (e.g., from [Ito and Kunisch 2008, Ex. 8.14]), that the function $v \mapsto \max(0, v - \alpha)$ is Newton differentiable from L^p to L^q for any $p > q \geqslant 1$ with Newton derivative in direction h given pointwise almost everywhere by

$$[D_N \max(0, v - \alpha)]\, h = \chi_{\{v>\alpha\}} h = \begin{cases} h(x), & \text{if } v(x) > \alpha, \\ 0, & \text{if } v(x) \leqslant \alpha. \end{cases}$$

An analogous statement holds for the pointwise min function. Since $S_\omega^* v \in W^{1,q'}(\omega_c)$ holds for all $v \in L^2(\omega_o)$, F is Newton differentiable and the chain rule for Newton derivatives (e.g., [Ito and Kunisch 2008, Lemma 8.15]) yields that the action of the Newton derivative of

$$G_+(u) := \max(0, S_\omega^*(S_\omega u - z) - \alpha)$$

in the direction h is given by

$$\begin{aligned} D_N G_+(u)h &= \chi_{\{S_\omega^*(S_\omega u - z)>\alpha\}}(S_\omega^* S_\omega h) \\ &= \begin{cases} (S_\omega^* S_\omega h)(x) & \text{if } (S_\omega^*(S_\omega u - z))(x) > \alpha, \\ 0 & \text{if } (S_\omega^*(S_\omega u - z))(x) \leqslant \alpha, \end{cases} \end{aligned}$$

and a similar claim holds for the min term. A semismooth Newton step thus consists in solving for δu in the equation

$$\begin{aligned} \delta u + c\chi_{\{|S_\omega^*(S_\omega u^k - z)|>\alpha\}}(S_\omega^* S_\omega \delta u) = -u^k - c\max(0, S_\omega^*(S_\omega u^k - z) - \alpha) \\ - c\min(0, S_\omega^*(S_\omega u^k - z) + \alpha) \end{aligned} \tag{8.4.1}$$

and setting $u^{k+1} = u^k + \delta u$. The semismooth Newton step (8.4.1) can be solved using an iterative Krylov solver (e.g., GMRES), where the action of the Newton derivative on given δu is computed by first solving the linearized state equation for the state differential δy followed by the adjoint equation for the adjoint differential δp. The full procedure is given as Algorithm 8.1.

It remains to verify the well-posedness and uniform boundedness of the Newton step (8.4.1).

Proposition 8.4.1. *For fixed $\alpha, c > 0$ and for any $u \in L^2(\omega_c)$, the mapping $D_N F(u) \in \mathcal{L}(L^2(\omega_c), L^2(\omega_c))$ is invertible, and there exists a constant $C > 0$ independent of u such that*

$$\|D_N F(u)^{-1}\|_{\mathcal{L}(L^2(\omega_c), L^2(\omega_c))} \leqslant C$$

holds.

Algorithm 8.1 Semismooth Newton method

Choose u^0, set $k = 0$
repeat
  Solve for $\widetilde{y}^k$ in $Ay = R^*_{\omega_c} u^k$, set $y^k = R_{\omega_o}\widetilde{y}^k$
  Solve for $\widetilde{p}^k$ in $A^*p = R^*_{\omega_o}(y^k - z)$, set $p^k = R_{\omega_c}\widetilde{p}^k$
  Compute active sets

$$\begin{aligned}\mathcal{A}^k_+ &= \left\{x \in \omega_c : p^k(x) > \alpha\right\}\\ \mathcal{A}^k_- &= \left\{x \in \omega_c : p^k(x) < -\alpha\right\},\\ \mathcal{A}^k &= \mathcal{A}^k_+ \cup \mathcal{A}^k_-\end{aligned}$$

  Set $F(u^k) = -u^k - c\chi_{\mathcal{A}^k_+}(p^k - \alpha) - c\chi_{\mathcal{A}^k_-}(p^k + \alpha)$
  Compute δu by solving $D_N F(u^k)\delta u = F(u^k)$ using ApplyNewtonMatrix in Krylov method
  Set $u^{k+1} = u^k + \delta u$, $k \leftarrow k + 1$
until ($\mathcal{A}^k_+ = \mathcal{A}^{k-1}_+$ and $\mathcal{A}^k_- = \mathcal{A}^{k-1}_-$) or $k = k^*$
function ApplyNewtonMatrix($\delta u, \mathcal{A}^k$)
  Solve for $\widetilde{\delta y}$ in $A\delta y = R^*_{\omega_c}\delta u$, set $\delta y = R_{\omega_o}\widetilde{\delta y}$
  Solve for $\widetilde{\delta p}$ in $A^*\delta p = R^*_{\omega_o}\delta y$, set $\delta p = R_{\omega_c}\widetilde{\delta p}$
  return $\delta u + c\chi_{\mathcal{A}^k}\delta p$
end function

Proof. Let $u \in L^2(\omega_c)$ be given and set

$$\mathcal{A} = \{|S^*_\omega(S_\omega u - z)| > \alpha\}$$

as well as $\mathcal{I} = \overline{\omega}_c \setminus \mathcal{A}$.

For arbitrary $w \in L^2(\omega_c)$, we need to find $\delta u \in L^2(\omega_c)$ satisfying

$$\delta u + c\chi_{\mathcal{A}}(S^*_\omega S_\omega \delta u) = w. \tag{8.4.2}$$

From this, we have that $\delta u = w$ almost everywhere in $\mathcal{I}$. By the linearity of S_ω and S^*_ω, we can thus write

$$c\chi_{\mathcal{A}}(S^*_\omega S_\omega \delta u) = c\chi_{\mathcal{A}}(S^*_\omega S_\omega(\chi_{\mathcal{A}}\delta u)) + c\chi_{\mathcal{A}}(S^*_\omega S_\omega(\chi_{\mathcal{I}} w)).$$

Inserting this identity into (8.4.2) and testing with δu, we obtain

$$\begin{aligned}\|\delta u\|^2_{L^2(\omega_c)} + c\,\|S_\omega(\chi_{\mathcal{A}}\delta u)\|^2_{L^2(\omega_o)} &= \langle w, \delta u\rangle_{L^2(\omega_c)} - c\,\langle S_\omega(\chi_{\mathcal{I}} w), S_\omega(\chi_{\mathcal{A}}\delta u)\rangle_{L^2(\omega_o)}\\ &\leqslant \|w\|_{L^2(\omega_c)}\,\|\delta u\|_{L^2(\omega_c)} + C\,\|w\|_{L^2(\omega_c)}\,\|\delta u\|_{L^2(\omega_c)}\\ &\leqslant C\,\|w\|_{L^2(\omega_c)}\,\|\delta u\|_{L^2(\omega_c)}\end{aligned}$$

by the continuity of S_ω. Together this implies

$$\|\delta u\|_{L^2(\omega_c)} \leqslant C \|w\|_{L^2(\omega_c)}$$

with a constant $C > 0$ depending on c but independent of $\mathcal{A}$ and therefore of u, which yields the claimed uniform boundedness. □

From this, the superlinear convergence of the semismooth Newton method follows from standard arguments (e.g., [Ito and Kunisch 2008, Th. 8.16]). The termination criterion in Algorithm 8.1, step 9, can be justified as follows: If $\mathcal{A}_\pm^{k+1} = \mathcal{A}_\pm^k$ holds, then u^{k+1} satisfies $F(u^{k+1}) = 0$ (cf. [Ito and Kunisch 2008, Rem. 7.1.1]).

For the numerical implementation, we use a continuation strategy: Solve for u_{c_k}, set $c_{k+1} = qc_k$ with $q > 1$, and use u_{c_k} as initial guess for the computation of $u_{c_{k+1}}$.

Remark 8.4.2 (Non-negative controls). By setting $\mathcal{A}_+^k = \emptyset$, Algorithm 8.1 can be applied to the numerical solution of ($OS_{+,c}$). The superlinear convergence holds in this case as well.

8.5 NUMERICAL EXAMPLES

We illustrate the proposed approach with a simple convection-diffusion equation, described by the operator $Ay = -\nu\Delta y - b \cdot \nabla y$ with $\nu = 0.1$ and $b = (1, 0)^T$ and homogeneous Dirichlet conditions on the unit square $[-1, 1]^2$. The control and observation domains are given by

$$\omega_c = \left\{x \in \Omega : \tfrac{1}{16} \leqslant |x|^2 \leqslant \tfrac{1}{2}\right\},$$
$$\omega_o = \left\{x \in \Omega : |x|^2 \leqslant \tfrac{1}{32}\right\},$$

and the target is $z = \chi_{\omega_o} x_2$ (see Figure 8.1). The differential operators are discretized using standard finite differences with $N = 128$ nodes in each direction, and Algorithm 8.1 is implemented in Matlab.

The parameters for this example are set as follows. In the continuation scheme for the penalty parameter c, the initial value is $c_0 = 1$, the incrementation factor is set to $q = 10$, and the continuation is terminated at $c^* = 10^{12}$. The semismooth Newton method is terminated if the active sets coincide or $k^* = 10$ iterations are reached. For the solution of the linear systems arising in the semismooth Newton step, we use Matlab's built-in GMRES with a relative tolerance of 10^{-9} and a maximum number of iterations of 100. The Matlab code of our implementation can be downloaded from `http://www.uni-graz.at/~clason/codes/measurecontrol.m`.

The discrete optimal controls u_α and corresponding states y_α for different values of α are shown in Figure 8.2. As α is decreased, the state becomes closer to the target, while the control becomes less sparse. Note that the loss in sparsity is manifested by an increasing number of

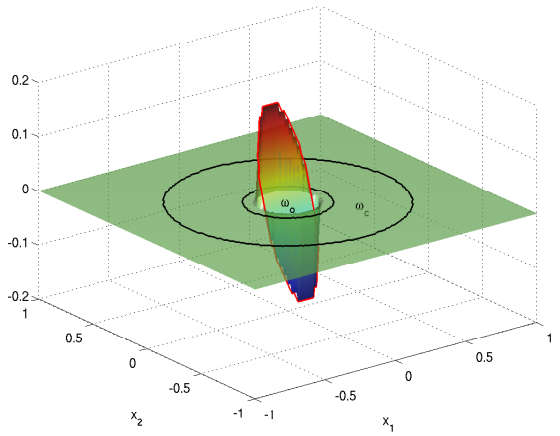


Figure 8.1: Target z, control domain ω_c and observation domain ω_o.

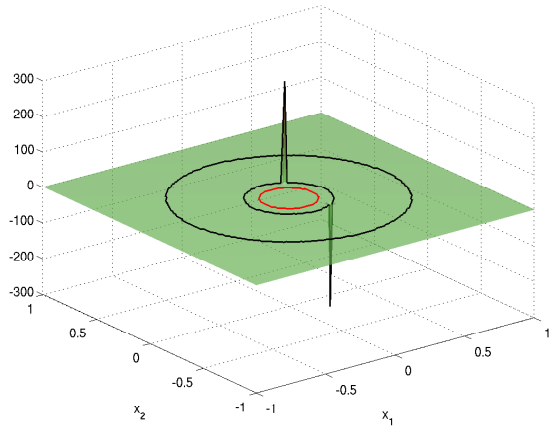
point sources, but the support of the control remains localized. This is due to the structural properties of the optimality system: the control is allowed to be active only where the dual variable p^* is active and must be identically zero everywhere else (cf. Corollary 8.2.4). Also, the controls are placed asymmetrically due to the directionality of the convection term. We point out that the placement of the corresponding sources is not obvious.

We indicate the superlinear convergence of the semismooth Newton method by fixing $c = 10^5$ and $\alpha = 10^{-4}$ and starting from the initial guess $u_0 = 0$. Table 8.1 shows the norm of the residual $\|F(u^k)\|_{L^2}$ and the change in active sets $\delta\mathcal{A}^k$ for each iteration in the semismooth Newton method, verifying the locally superlinear convergence.

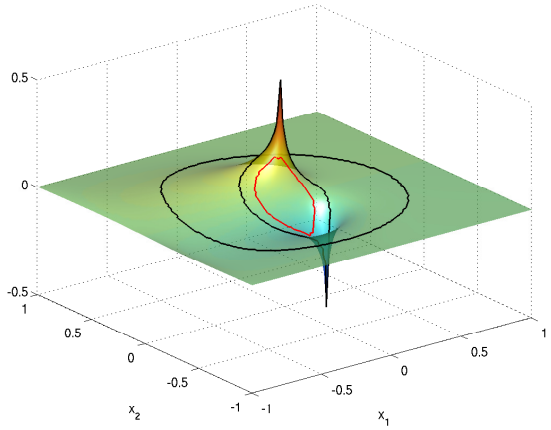
The feasibility of our approach for the optimal placement of sources is illustrated with an example that is motivated by an application in photochemotherapy. Here, ω_o denotes a region where a photosensitive chemotherapeutic agent is locally activated by laser light from multiple strategically placed fiber-optic light sources [Baas et al. 1997]. The latter can be focused inside a small boundary layer, which corresponds to the control domain ω_c. This example further demonstrates the dependence of the locations of the optimal controls on the geometry of the problem, here determined by an irregular domain (see Figure 8.3).

Table 8.1: Convergence of semismooth Newton method. Shown are the norm of the residual of (8.4.1) and the change in active sets $\delta\mathcal{A}^k$ in each iteration.

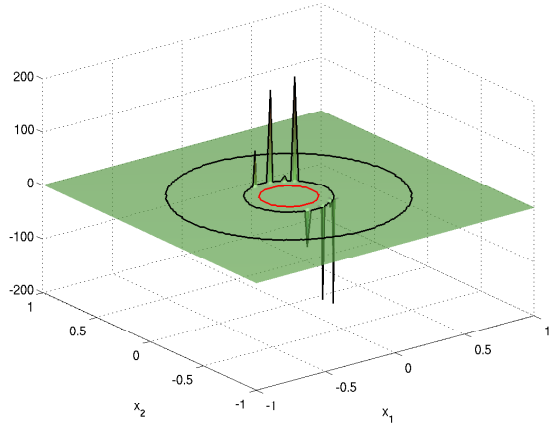
k	1	2	3	4	5
$\|F(u^k)\|$	$3.84 \cdot 10^2$	$3.78 \cdot 10^1$	$6.12 \cdot 10^0$	$6.16.78 \cdot 10^{-1}$	$1.08 \cdot 10^{-10}$
$\delta\mathcal{A}^k$	3678	532	106	8	0



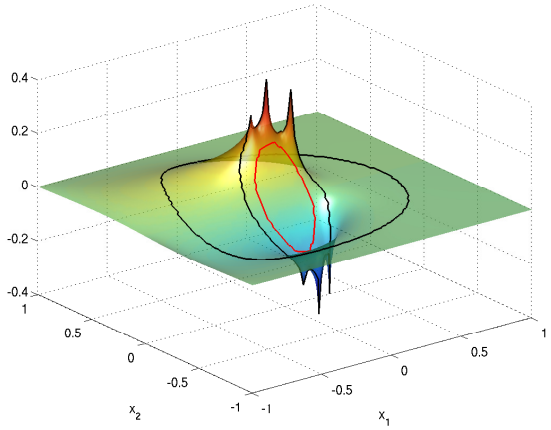
(a) u_α, $\alpha = 10^{-3}$



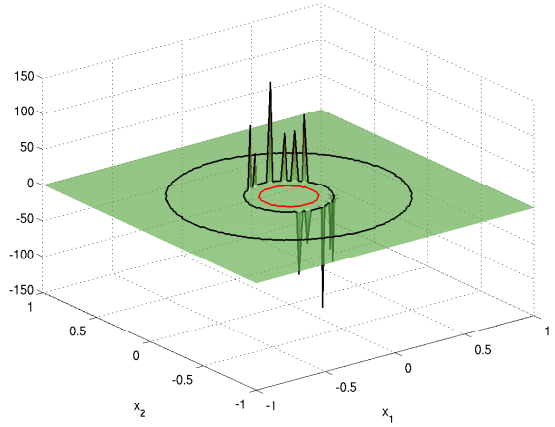
(b) y_α, $\alpha = 10^{-3}$



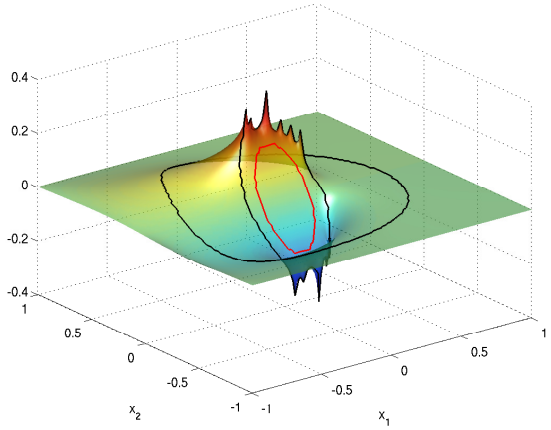
(c) u_α, $\alpha = 10^{-4}$



(d) y_α, $\alpha = 10^{-4}$

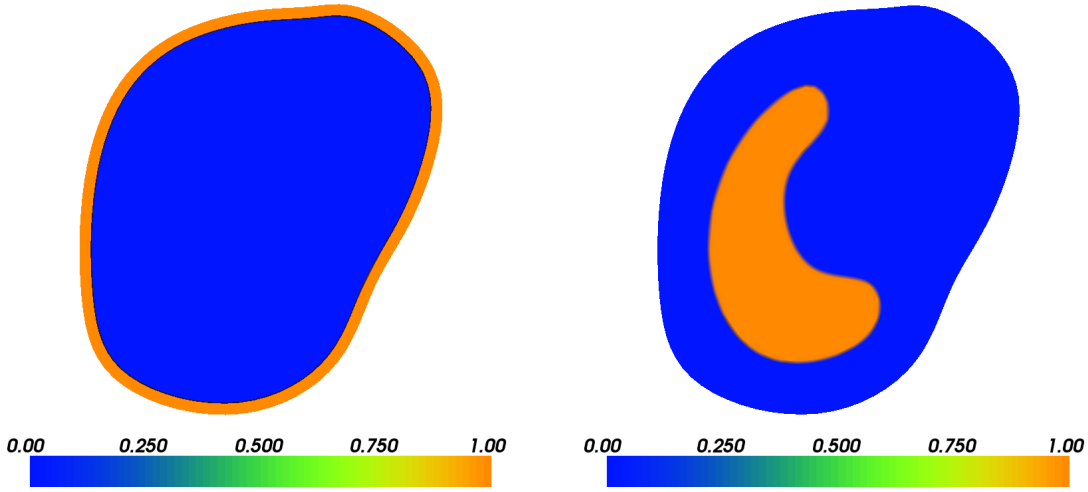


(e) u_α, $\alpha = 10^{-5}$



(f) y_α, $\alpha = 10^{-5}$

Figure 8.2: Optimal controls u_α and states y_α for different values of α.



(a) control domain ω_c (b) observation domain ω_o

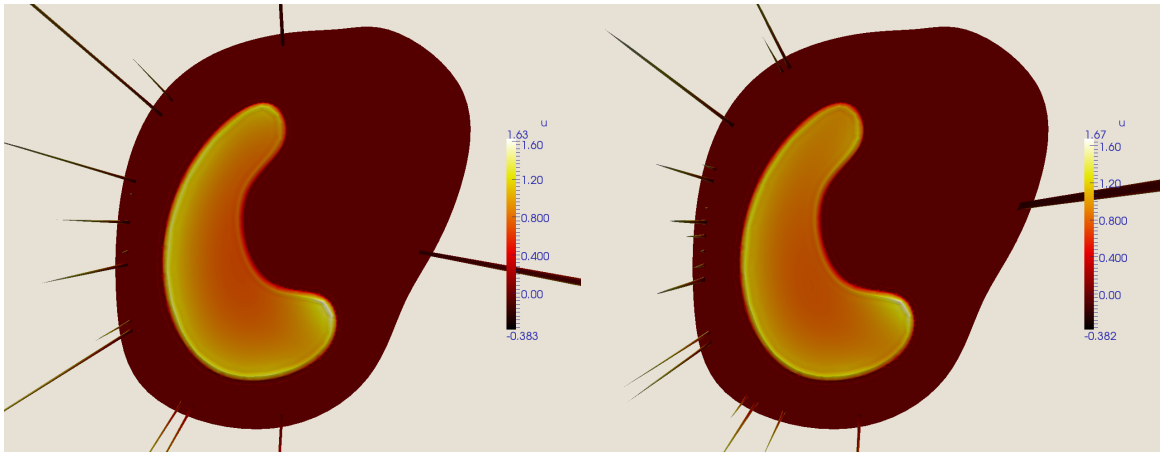
Figure 8.3: Geometry for model problem. Shown are the indicator functions of control and observation domain.

The corresponding state equation is

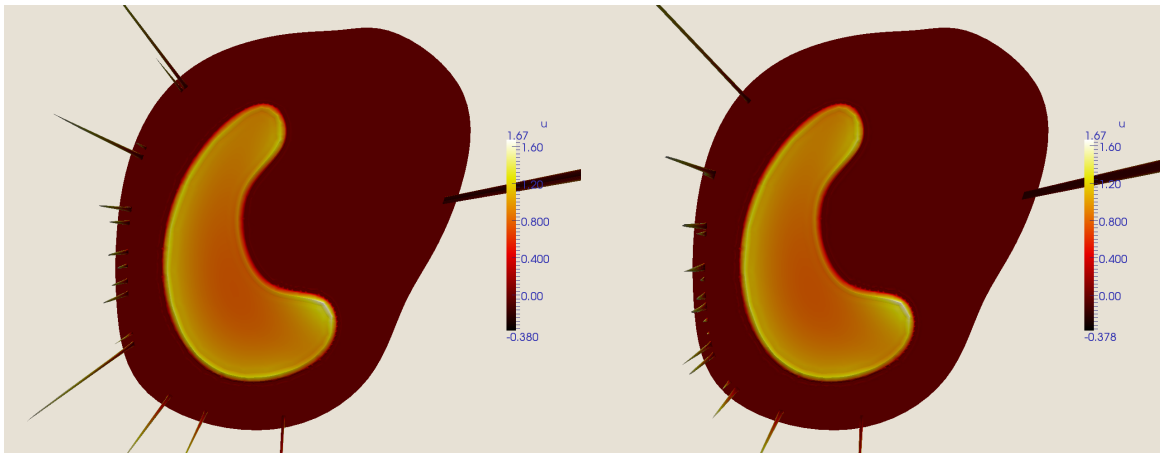
$$\begin{cases} -\nabla \cdot \left(\frac{1}{2(\mu_a+\mu_s)} \nabla y \right) + \mu_a y = \chi_{\omega_c} u & \text{on } \Omega, \\ \frac{1}{2(\mu_a+\mu_s)} \partial_\nu y + \rho y = 0 & \text{on } \partial\Omega, \end{cases}$$

which describes diffusive photon transport in tissue. Here, μ_a is the tissue's absorption coefficient, μ_s is the scattering coefficient, and ρ is the reflection coefficient at the boundary $\partial\Omega$. In our tests, we set $\mu_a = 0.03$, $\mu_s = 0.275$ and $\rho = 0.1992$ to model a small piece of lung tissue. The objective is then to achieve a homogeneous illumination of the region of interest ω_o for the optimal activation of the chemo-sensitive agent. Due to the linearity of the equation, we set without loss of generality $z \equiv 1$. As noted in Remark 8.4.2, non-negativity of the controls is enforced by setting $\mathcal{A}_+ \equiv \emptyset$.

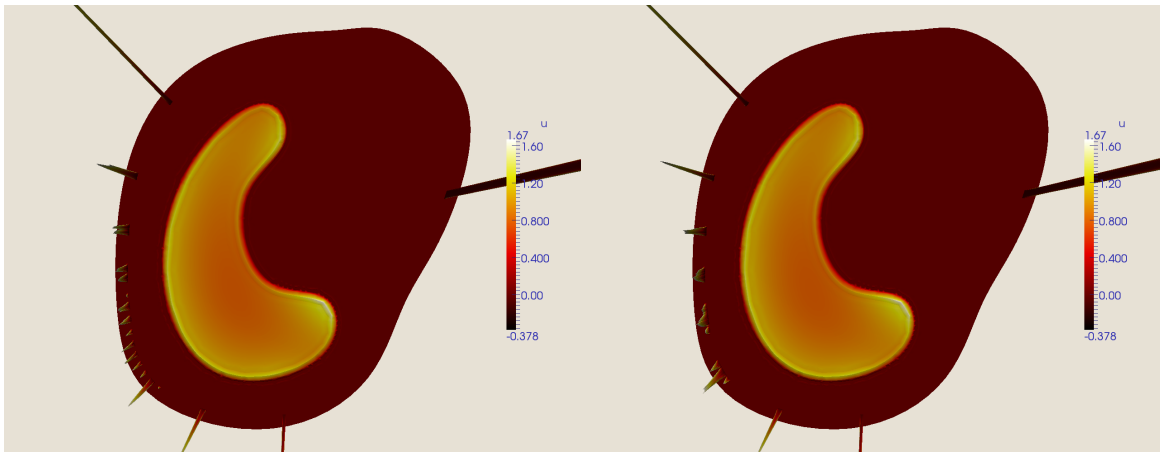
Due to the irregularity of the domain, we now use a standard finite element discretization of state and adjoint equation in weak form (cf. (8.2.2)) with triangular elements, where the discretized control is taken as piecewise constant and the discretized state and adjoint as piecewise linear. The implementation is based on the open source FEniCS project [Logg and Wells 2010]. Here, the final penalty parameter is set to $c^* = 10^9$, the remaining parameters being unchanged. The results for different values of α are shown in Figure 8.4. The influence of α on sparsity of the controls and homogeneity of the illumination can be observed clearly. Note that the placement of the optimal point sources is again not obvious, and depends on α in a non-intuitive manner (compare the location of the major point source on the right hand side of the domain between Figs. 8.4a and 8.4b).



(a) $\alpha = 5 \cdot 10^{-2}$ (b) $\alpha = 1 \cdot 10^{-2}$



(c) $\alpha = 5 \cdot 10^{-3}$ (d) $\alpha = 1 \cdot 10^{-3}$



(e) $\alpha = 5 \cdot 10^{-4}$ (f) $\alpha = 1 \cdot 10^{-4}$

Figure 8.4: Optimal controls u and state y for different values of α. Shown is a superposition of u on ω_c (as height plot) and y on ω_o (as color plot).

8.6 CONCLUSION

The problem of optimal placement of point sources was formulated in a non-reflexive Banach space setting. The optimality system for this nonsmooth optimization problem was derived and a family of regularized problems, which can be approximated efficiently by semismooth Newton methods, was analyzed. The numerical examples demonstrate the effectivity for optimal light source placement problems in diffusive photochemotherapy. Current work is concerned with the application of the proposed approach to patient-specific geometries. Formally, the primal-dual framework considered here can be extended to nonlinear control-to-state mappings, although the proper functional analytic treatment of the linearization remains challenging. Finally, it would be of interest to consider parabolic state equations.

ACKNOWLEDGMENTS

The authors wish to thank Patricia Brunner, Manuel Freiberger and Hermann Scharfetter of the Institute of Medical Engineering, Graz University of Technology, for their help on the photochemotherapy example.

APPROXIMATION OF ELLIPTIC CONTROL PROBLEMS IN MEASURE SPACES WITH SPARSE SOLUTIONS

9

ABSTRACT

Optimal control problems in measure spaces governed by elliptic equations are considered for distributed and Neumann boundary control, which are known to promote sparse solutions. Optimality conditions are derived and some of the structural properties of their solutions, in particular sparsity, are discussed. A framework for their approximation is proposed which is efficient for numerical computations and for which we prove convergence and provide error estimates.

9.1 INTRODUCTION

This paper is dedicated to the approximation of the optimal control problem

$$\min_{u\in\mathcal{M}(\Omega)} J(u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \alpha\|u\|_{\mathcal{M}(\Omega)}, \tag{P}$$

where y is the unique solution to the Dirichlet problem

$$\begin{cases} -\Delta y + c_0 y &= u \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \Gamma, \end{cases} \tag{9.1.1}$$

with $c_0 \in L^\infty(\Omega)$ and $c_0 \geqslant 0$. We assume that $\alpha > 0$, $y_d \in L^2(\Omega)$ and Ω is a bounded domain in $\mathbb{R}^n$, $n = 2$ or 3, which is supposed to either be convex or have a $C^{1,1}$ boundary Γ. The controls are taken in the space of regular Borel measures $\mathcal{M}(\Omega)$. As usual, $\mathcal{M}(\Omega)$ is

identified by the Riesz theorem with the dual space of $C_0(\Omega)$ – consisting of the continuous functions in $\overline{\omega}$ vanishing on Γ – endowed with the norm

$$\|u\|_{\mathcal{M}(\Omega)} = \sup_{\|z\|_{C_0(\Omega)} \leqslant 1} \langle u, z \rangle = \sup_{\|z\|_{C_0(\Omega)} \leqslant 1} \int_\Omega z(x)\, du,$$

which is equivalent to the total variation of u.

It has been observed that the use of measures leads to optimal controls which are sparse. This is relevant for many applications in distributed parameter control; see [Clason and Kunisch 2011]. Moreover, the support of the optimal control provides information on the optimal placements of control actuators. Formally, the same features can be achieved by using $L^1(\Omega)$ control cost. In this case, however, the optimal control problem is not well-posed in the sense of a possible lack of existence of a minimizer due to the fact that $L^1(\Omega)$ does not allow an appropriate topology for compactness arguments. Other techniques have been used to overcome this difficulty, including the use of regularization techniques or the introduction of control constraints; see, for instance, [Casas, Herzog, and Wachsmuth 2012], [Stadler 2009], [Wachsmuth and Wachsmuth 2011a].

The focus of this paper is to give an approximation framework which, in spite of the difficulties due to the presence of measures, leads to implementable schemes for which a priori error estimates can be provided. We show that the optimal control measure can be approximated efficiently by a linear combination of Dirac measures. This is important for practical applications because it provides a way of controlling a distributed system by finitely many point actuators, giving information on where they have to be placed. A similar framework in the context of inverse problems was considered in [Bredies and Pikkarainen 2012].

The plan of the paper is as follows. In the next section we provide optimality conditions for (P) and derive some properties of the solution, in particular sparsity and actuator location. In § 9.3, we introduce the approximation framework and prove convergence of the discretized problems to the continuous one. Rate of convergence results are provided in § 9.4. In § 9.5 we show that analogous results can also be obtained for Neumann control problems. Finally, the last section is devoted to numerical test problems.

9.2 OPTIMALITY CONDITIONS

Before establishing the optimality conditions for problem (P) and deducing some consequences from them, let us observe some important facts. First, given a measure $u \in \mathcal{M}(\Omega)$, we say that y is a solution to (9.1.1) if

$$\int_\Omega y A z \, dx = \int_\Omega z \, du \qquad \text{for all } z \in H^2(\Omega) \cap H_0^1(\Omega), \tag{9.2.1}$$

where $A = -\Delta + c_0 I$. It is well known, see for instance [Casas 1986], that there exists a unique solution to (9.1.1) in the sense of (9.2.1). Moreover, $y \in W_0^{1,p}(\Omega)$ for every $1 \leqslant p < \frac{n}{n-1}$ and

$$\|y\|_{W_0^{1,p}(\Omega)} \leqslant C_p \|u\|_{\mathcal{M}(\Omega)}. \tag{9.2.2}$$

Since $W_0^{1,p}(\Omega) \subset L^2(\Omega)$ for every $\frac{2n}{n+2} \leqslant p < \frac{n}{n-1}$, the cost functional is well defined on $\mathcal{M}(\Omega)$. Furthermore, the control-to-state mapping is injective, and therefore the cost functional J is strictly convex. Then, it can be obtained by the standard approach that (P) has a unique solution; see [Clason and Kunisch 2011] for details. Hereafter, this optimal solution will be denoted by $\bar{u}$ with an associated state $\bar{y}$. By using subdifferential calculus of convex functions and introducing the adjoint state we get the following results (see also [Clason and Kunisch 2011; Clason and Kunisch 2012]).

Theorem 9.2.1. *There exists a unique element $\bar{\varphi} \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying*

$$\begin{cases} -\Delta\bar{\varphi} + c_0\bar{\varphi} &= \bar{y} - y_d \quad \text{in } \Omega, \\ \bar{\varphi} &= 0 \quad \text{on } \Gamma, \end{cases}$$

such that

$$\alpha\|\bar{u}\|_{\mathcal{M}(\Omega)} + \int_\Omega \bar{\varphi}\, d\bar{u} = 0, \tag{9.2.3}$$

$$\|\bar{\varphi}\|_{C_0(\Omega)} \begin{cases} = \alpha & \text{if } \bar{u} \neq 0, \\ \leqslant \alpha & \text{if } \bar{u} = 0. \end{cases} \tag{9.2.4}$$

Proof. By standard arguments from Lagrange multiplier theory and the Sobolev embedding theorem, we deduce the existence of a $\lambda \in C_0(\Omega)$ with

$$\lambda \in \partial\|\cdot\|_{\mathcal{M}(\Omega)}(\bar{u}) \quad \text{and} \quad \alpha\lambda = -\bar{\varphi}. \tag{9.2.5}$$

By the definition of the convex subdifferential, the first inclusion is equivalent to

$$\langle \lambda, u - \bar{u}\rangle + \|\bar{u}\|_{\mathcal{M}(\Omega)} \leqslant \|u\|_{\mathcal{M}(\Omega)} \tag{9.2.6}$$

for all $u \in \mathcal{M}(\Omega)$. Taking $u = 2\bar{u}$ and $u = 0$, respectively, we obtain the two inequalities

$$\langle \lambda, \bar{u}\rangle \leqslant \|\bar{u}\|_{\mathcal{M}(\Omega)} \leqslant \langle \lambda, \bar{u}\rangle$$

and hence (9.2.3) by the second relation of (9.2.5). Inserting (9.2.3) and $\lambda = -\frac{1}{\alpha}\bar{\varphi}$ into (9.2.6) yields

$$\langle \bar{\varphi}, u\rangle \leqslant \alpha\|u\|_{\mathcal{M}(\Omega)},$$

which implies (9.2.4). □

As pointed out in [Clason and Kunisch 2011], if we consider the Jordan decomposition of $\bar{u} = \bar{u}^+ - \bar{u}^-$, then we deduce from (9.2.3) and (9.2.4) that

$$\begin{cases} \operatorname{supp}(\bar{u}^+) \subset \{x \in \Omega : \bar{\varphi}(x) = -\alpha\}, \\ \operatorname{supp}(\bar{u}^-) \subset \{x \in \Omega : \bar{\varphi}(x) = +\alpha\}. \end{cases} \tag{9.2.7}$$

From (9.2.7) we note that $\bar{u} \equiv 0$ on the set $\{x \in \Omega : |\bar{\varphi}(x)| < \alpha\}$. As the numerical results will show, the set $\{x \in \Omega : |\bar{\varphi}(x)| = \alpha\}$ is small, which yields the sparsity of $\bar{u}$. Moreover, we have the following property for the penalty parameter.

Proposition 9.2.2. *There exists $\tilde{\alpha} > 0$ such that $\bar{u} = 0$ for every $\alpha > \tilde{\alpha}$.*

Proof. Let us denote by J_α the cost functional associated to the parameter α. Similarly, let $(u_\alpha, y_\alpha, \varphi_\alpha)$ denote the solution to the corresponding optimality system. For each $\alpha > 0$ the following inequalities hold

$$\frac{1}{2}\|y_\alpha - y_d\|^2_{L^2(\Omega)} \leqslant J_\alpha(u_\alpha) \leqslant J_\alpha(0) = \frac{1}{2}\|y_d\|^2_{L^2(\Omega)}.$$

Consequently,

$$\|y_\alpha - y_d\|_{L^2(\Omega)} \leqslant \|y_d\|_{L^2(\Omega)} \quad \forall \alpha > 0.$$

From the adjoint state equation and the embedding of $H^2(\Omega) \cap H^1_0(\Omega) \hookrightarrow C_0(\Omega)$, we deduce the existence of a constant $C > 0$ such that

$$\|\varphi_\alpha\|_{C_0(\Omega)} \leqslant C\|y_\alpha - y_d\|_{L^2(\Omega)} \leqslant C\|y_d\|_{L^2(\Omega)}.$$

Setting $\tilde{\alpha} = C\|y_d\|_{L^2(\Omega)}$, we obtain from the above inequality and (9.2.4) that $u_\alpha = 0$ for every $\alpha > \tilde{\alpha}$. □

In the case where we consider the observation of the state only in a subset $\omega_y \subset \Omega$, then we have the following property of the support of the optimal control.

Proposition 9.2.3. *Let ω_y be an open subset of Ω such that $\Omega \setminus \omega_y$ is connected and consider the functional*

$$J_{\omega_y}(u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\omega_y)} + \alpha\|u\|_{\mathcal{M}(\Omega)}.$$

Then the associated optimal control $\bar{u}$ satisfies $\operatorname{supp}(\bar{u}) \subset \overline{\omega}_y$.

Proof. For the functional under consideration, the adjoint state equation is given by

$$\begin{cases} -\Delta\bar{\varphi} + c_0\bar{\varphi} &= (\bar{y} - y_d)\chi_{\omega_y} & \text{in } \Omega, \\ \bar{\varphi} &= 0 & \text{on } \Gamma, \end{cases}$$

where χ_{ω_y} is the characteristic function of ω_y. Applying the maximum principle to the problem

$$\begin{cases} -\Delta\bar{\varphi} + c_0\bar{\varphi} &= 0 \quad \text{in } \Omega \setminus \overline{\omega}_y, \\ \bar{\varphi} &= 0 \quad \text{on } \Gamma, \end{cases}$$

we deduce that $\bar{\varphi}$ is identically zero in $\Omega \setminus \overline{\omega}_y$ or

$$\min_{x' \in \partial\omega_y} \bar{\varphi}(x') < \bar{\varphi}(x) < \max_{x' \in \partial\omega_y} \bar{\varphi}(x') \ \ \forall x \in \Omega \setminus \overline{\omega}_y.$$

In both cases the equality (9.2.4) can only be achieved in $\overline{\omega}_y$, therefore (9.2.7) implies the claim of the proposition. □

Let us close this section by pointing out that the results of our paper can also be adapted to the situation where the control domain is a priori restricted to a strict subdomain ω_u of Ω, and the controls are restricted to be non-negative (cf. [Clason and Kunisch 2012]).

9.3 APPROXIMATION FRAMEWORK

In this section Ω will be assumed to be convex. We consider a nodal basis finite element approximation of (P). Associated with a parameter h we consider a family of triangulations $\{\mathcal{T}_h\}_{h>0}$ of $\overline{\omega}$. To every element $T \in \mathcal{T}_h$ we assign two parameters $\rho(T)$ and $\sigma(T)$, where $\rho(T)$ denotes the diameter of T and $\sigma(T)$ is the diameter of the biggest ball contained in T. The size of the grid is given by $h = \max_{T\in\mathcal{T}_h} \rho(T)$. The following usual regularity assumptions on the triangulation are assumed.

(i) There exist two positive constants ρ and σ such that

$$\frac{\rho(T)}{\sigma(T)} \leqslant \sigma \quad \text{and} \quad \frac{h}{\rho(T)} \leqslant \rho$$

hold for every $T \in \mathcal{T}_h$ and all $h > 0$.

(ii) Let us set $\overline{\Omega}_h = \bigcup_{T\in\mathcal{T}_h} T$ with Ω_h and Γ_h its interior and boundary respectively. We assume that the vertices of $\mathcal{T}_h$ placed on the boundary Γ_h are also points of Γ. From [Raviart and Thomas 1983, inequality (5.2.19)] we know

$$|\Omega \setminus \Omega_h| \leqslant Ch^2, \tag{9.3.1}$$

where $|\cdot|$ denotes the Lebesgue measure.

Associated to these triangulations we define the space

$$Y_h = \left\{ y_h \in C_0(\Omega) : y_{h|T} \in \mathcal{P}_1 \text{ for every } T \in \mathcal{T}_h, \text{ and } y_h = 0 \text{ in } \overline{\omega} \setminus \Omega_h \right\},$$

where $\mathcal{P}_1$ is the space formed by the polynomials of degree less than or equal to one. For every $u \in \mathcal{M}(\Omega)$, we denote by y_h the unique element of Y_h satisfying

$$a(y_h, z_h) = \int_{\Omega_h} z_h \, du \quad \forall z_h \in Y_h, \tag{9.3.2}$$

where $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is the bilinear form associated to the operator A, i.e.,

$$a(y, z) = \int_\Omega [\nabla y(x) \nabla z(x) + c_0(x) y(x) z(x)] \, dx.$$

The approximation of the optimal control problem (P) is defined as

$$\min_{u \in \mathcal{M}(\Omega)} J_h(u_h) = \frac{1}{2} \|y_h - y_d\|^2_{L^2(\Omega_h)} + \alpha \|u\|_{\mathcal{M}(\Omega)}, \tag{P_h}$$

where y_h is the solution to (9.3.2).

Since we have not discretized the control space, this approach is related to the variational discretization method introduced in [Hinze 2005]. Below we will show that among all the solutions to (P_h) there is a unique one which is a finite linear combination of Dirac measures concentrated in the interior vertices of the triangulation, leading to a simple numerical implementation.

Before any discussion of the solutions to problem (P_h), let us introduce some additional notation. Hereafter we will denote by $\{x_j\}_{j=1}^{N(h)}$ the interior nodes of the triangulation $\mathcal{T}_h$. Associated to these nodes we consider the nodal basis of Y_h given by the functions $\{e_j\}_{j=1}^{N(h)}$ such that $e_j(x_i) = \delta_{ij}$ for every $1 \leqslant i, j \leqslant N(h)$. Then every element y_h of Y_h can be written in the form

$$y_h = \sum_{j=1}^{N(h)} y_j e_j, \qquad \text{where} \quad y_j = y_h(x_j), \quad 1 \leqslant j \leqslant N(h).$$

We also consider the space

$$D_h = \left\{ u_h \in \mathcal{M}(\Omega) : u_h = \sum_{j=1}^{N(h)} \lambda_j \delta_{x_j}, \text{ where } \{\lambda_j\}_{j=1}^{N(h)} \subset \mathbb{R} \right\}.$$

Above δ_{x_j} denotes the Dirac measure centered at the node x_j. It is obvious that D_h can be identified with the dual of Y_h through the duality relation

$$\langle u_h, y_h \rangle = \sum_{j=1}^{N(h)} \lambda_j y_j.$$

Now, we define the linear operators $\Pi_h : C_0(\Omega) \to Y_h$ and $\Lambda_h : \mathcal{M}(\Omega) \to D_h$ by

$$\Pi_h y = \sum_{j=1}^{N(h)} y(x_j) e_j \quad \text{and} \quad \Lambda_h u = \sum_{j=1}^{N(h)} \langle u, e_j \rangle \delta_{x_j}.$$

The operator Π_h is the nodal interpolation operator for Y_h, and we have the following result concerning the operator Λ_h.

Theorem 9.3.1. *The following properties hold.*

(i) For every $u \in \mathcal{M}(\Omega)$ and every $z \in C_0(\Omega)$ and $z_h \in Y_h$ we have

$$\langle u, z_h \rangle = \langle \Lambda_h u, z_h \rangle, \tag{9.3.3}$$

$$\langle u, \Pi_h z \rangle = \langle \Lambda_h u, z \rangle. \tag{9.3.4}$$

(ii) For every $u \in \mathcal{M}(\Omega)$ we have

$$\|\Lambda_h u\|_{\mathcal{M}(\Omega)} \leqslant \|u\|_{\mathcal{M}(\Omega)}, \tag{9.3.5}$$

$$\Lambda_h u \overset{*}{\rightharpoonup} u \text{ in } \mathcal{M}(\Omega) \text{ and } \|\Lambda_h u\|_{\mathcal{M}(\Omega)} \to \|u\|_{\mathcal{M}(\Omega)}. \tag{9.3.6}$$

(iii) There exist a constant $C > 0$ such that for every $u \in \mathcal{M}(\Omega)$

$$\|u - \Lambda_h u\|_{W^{-1,p}(\Omega)} \leqslant C h^{1-n/p'} \|u\|_{\mathcal{M}(\Omega)}, \quad 1 < p < \frac{n}{n-1}, \tag{9.3.7}$$

$$\|u - \Lambda_h u\|_{(W_0^{1,\infty}(\Omega))^*} \leqslant C h \|u\|_{\mathcal{M}(\Omega)}, \tag{9.3.8}$$

where p' is the conjugate of p.

(iv) Given $u \in \mathcal{M}(\Omega)$, let y_h and $\tilde{y}_h$ be the solutions to (9.3.2) associated to the controls u and $\Lambda_h u$, respectively. Then the equality $y_h = \tilde{y}_h$ holds.

Proof. For $z_h = \sum_{j=1}^{N(h)} z_j e_j$ we have

$$\langle u, z_h \rangle = \sum_{j=1}^{N(h)} z_j \langle u, e_j \rangle = \sum_{j=1}^{N(h)} \langle u, e_j \rangle \langle \delta_{x_j}, z_h \rangle = \langle \Lambda_h u, z_h \rangle,$$

which proves (9.3.3). For (9.3.4) we proceed as follows

$$\langle u, \Pi_h z \rangle = \sum_{j=1}^{N(h)} z(x_j) \langle u, e_j \rangle = \sum_{j=1}^{N(h)} \langle u, e_j \rangle \langle \delta_{x_j}, z \rangle = \langle \Lambda_h u, z \rangle.$$

To verify (9.3.5) we introduce the function $s_h \in Y_h$ by

$$s_h = \sum_{j=1}^{N(h)} s_j e_j, \quad \text{with} \quad s_j = \begin{cases} +1 & \text{if } \langle u, e_j \rangle > 0, \\ -1 & \text{if } \langle u, e_j \rangle < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$\|\Lambda_h u\|_{\mathcal{M}(\Omega)} = \sum_{j=1}^{N(h)} |\langle u, e_j\rangle| = \sum_{j=1}^{N(h)} s_j \langle u, e_j\rangle = \langle u, s_h\rangle \leqslant \|u\|_{\mathcal{M}(\Omega)} \|s_h\|_{C_0(\Omega)} = \|u\|_{\mathcal{M}(\Omega)}.$$

Let us prove (9.3.6). Since $\{\Lambda_h u\}_{h>0}$ is bounded in $\mathcal{M}(\Omega)$ there exists a subsequence, denoted in the same way, such that $\Lambda_h u \overset{*}{\rightharpoonup} v$ in $\mathcal{M}(\Omega)$. From (9.3.3) we get that

$$\langle v, e_j\rangle = \lim_{h\to 0} \langle \Lambda_h u, e_j\rangle = \langle u, e_j\rangle \ \ \forall\, 1 \leqslant j \leqslant N(h),$$

which implies that $\langle v, z_h\rangle = \langle u, z_h\rangle$ for every $z_h \in Y_h$. Hence, for every $z \in C_0(\Omega)$

$$\langle v, z\rangle = \lim_{h\to 0} \langle v, \Pi_h z\rangle = \lim_{h\to 0} \langle u, \Pi_h z\rangle = \langle u, z\rangle,$$

therefore $u = v$. Since any subsequence converges to u, the whole sequence converges to u weakly* in $\mathcal{M}(\Omega)$. From this convergence and (9.3.5) we obtain

$$\|u\|_{\mathcal{M}(\Omega)} \leqslant \liminf_{h\to 0} \|\Lambda_h u\|_{\mathcal{M}(\Omega)} \leqslant \limsup_{h\to 0} \|\Lambda_h u\|_{\mathcal{M}(\Omega)} \leqslant \|u\|_{\mathcal{M}(\Omega)},$$

consequently (9.3.6) holds.

To prove (9.3.7) we take an arbitrary element $z \in W_0^{1,p'}(\Omega)$, with $1 \leqslant p < \frac{n}{n-1}$. Using (9.3.4) and the well known interpolation error estimates in Sobolev spaces (see, for instance, [Ciarlet 1978, Chapter 3]) we obtain

$$\begin{aligned}\langle u - \Lambda_h u, z\rangle = \langle u, z - \Pi_h z\rangle &\leqslant \|u\|_{\mathcal{M}(\Omega)} \|z - \Pi_h z\|_{C_0(\Omega)} \\ &\leqslant Ch^{1-n/p'} \|u\|_{\mathcal{M}(\Omega)} \|z\|_{W_0^{1,p'}(\Omega)}.\end{aligned}$$

Since $W^{-1,p}(\Omega)$ is the dual of $W_0^{1,p'}(\Omega)$ for $1 < p < \frac{n}{n-1}$, (9.3.7) follows from the above inequalities. For $p = 1$, we have $p' = \infty$ and the above inequality can be expressed as

$$\langle u - \Lambda_h u, z\rangle \leqslant Ch \|u\|_{\mathcal{M}(\Omega)} \|z\|_{W_0^{1,\infty}(\Omega)}, \quad \forall z \in W_0^{1,\infty}(\Omega).$$

Since $W^{-1,1}(\Omega)$ is not the dual space of $W_0^{1,\infty}(\Omega)$, from this inequalities we only get (9.3.8).

The last statement of the theorem is an immediate consequence of (9.3.3). □

Now, we turn to the study of (P_h). First, we observe that analogously to J, the functional J_h is convex. However, it is not strictly convex. This is a consequence of the non-injectivity of the control-to-discrete-state mapping and the non-strict convexity of the norm of $\mathcal{M}(\Omega)$. Although the existence of a solution can be proved in the same way as for the problem (P), we cannot claim its uniqueness. Nevertheless, if $\bar{u}_h$ is a solution to (P_h) and we take $\bar{u}_h = \Lambda_h \bar{u}_h$, then the statement (iv) of Theorem 9.3.1 and the inequality (9.3.5) imply that

$J_h(\bar{u}_h) \leqslant J_h(\tilde{u}_h)$, hence $\bar{u}_h$ is also a solution to (P_h). Since for $u_h \in D_h$, the mapping $u_h \mapsto y_h(u_h)$, the solution to (9.3.2) for $u = u_h$, is linear, injective and $\dim D_h = \dim Y_h$, this mapping is bijective. Therefore, the cost functional J_h is strictly convex on D_h, hence (P_h) has a unique solution in D_h, which will be denoted by $\bar{u}_h$ hereafter. We summarize this discussion in the following theorem.

Theorem 9.3.2. *Problem* (P_h) *admits at least one solution. Among them there exists a unique one* $\bar{u}_h$ *belonging to* D_h. *Moreover, any other solution* $\tilde{u}_h \in \mathcal{M}(\Omega)$ *of* (P_h) *satisfies that* $\Lambda_h \tilde{u}_h = \bar{u}_h$.

Remark 9.3.3. The fact that (P_h) has exactly one solution in D_h is of practical interest. Indeed, recall that, as an element of D_h, $\bar{u}_h$ has a unique representation of the form

$$\bar{u}_h = \sum_{j=1}^{N(h)} \bar{\lambda}_j \delta_{x_j}.$$

Then, the numerical computation of $\bar{u}_h$ is reduced to the computation of the coefficients $\{\bar{\lambda}_j\}_{j=1}^{N(h)}$.

Remark 9.3.4. All results remain valid for Lagrange elements of arbitrary degree, where the x_j should be taken as the nodes associated with the degrees of freedom (which no longer need to correspond to vertices of the triangulation, e.g., vertices and edge midpoints for quadratic elements).

We finish this section by proving the convergence of the solutions in D_h to problems (P_h) to the solution to (P).

Theorem 9.3.5. *For every* $h > 0$, *let* $\bar{u}_h$ *be the unique solution to* (P_h) *belonging to* D_h *and let* $\bar{u}$ *be the solution to* (P). *Then the following convergence properties hold for* $h \to 0$:

$$\bar{u}_h \overset{*}{\rightharpoonup} \bar{u} \text{ in } \mathcal{M}(\Omega), \tag{9.3.9}$$

$$\|\bar{u}_h\|_{\mathcal{M}(\Omega)} \to \|\bar{u}\|_{\mathcal{M}(\Omega)}, \tag{9.3.10}$$

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \to 0, \tag{9.3.11}$$

$$J_h(\bar{u}_h) \to J(\bar{u}), \tag{9.3.12}$$

where $\bar{y}$ *and* $\bar{y}_h$ *are the continuous and discrete states associated to* $\bar{u}$ *and* $\bar{u}_h$, *respectively.*

Proof. First of all, let us verify that

$$u_h \overset{*}{\rightharpoonup} u \text{ in } \mathcal{M}(\Omega) \quad \text{implies} \quad \|y_h(u_h) - y_u\|_{L^2(\Omega)} \to 0, \tag{9.3.13}$$

where $y_h(u_h)$ and y_u are the discrete and continuous states associated to the controls u_h and u, respectively. From the compact embedding $\mathcal{M}(\Omega) \hookrightarrow W^{-1,p}(\Omega)$ for every $1 \leqslant p < \frac{n}{n-1}$, we deduce the strong converge $u_h \to u$ in $W^{-1,p}(\Omega)$. Let us denote by y_{u_h} the continuous state associated to u_h. From [Jerison and Kenig 1995] we obtain the strong convergence $y_{u_h} \to y_u$

in $W^{1,p}(\Omega)$, where we have used that the boundary Γ is Lipschitz continuous as a consequence of the convexity of Ω. Moreover, from [Casas 1985] we have that $\|y_h(u_h) - y_{u_h}\|_{L^2(\Omega)} \to 0$. Finally, by the triangular inequality we obtain the desired convergence.

Turning to the verification of (9.3.9), we observe that

$$\alpha\|\bar{u}_h\|_{\mathcal{M}(\Omega)} \leqslant J_h(\bar{u}_h) \leqslant J_h(0) = \frac{1}{2}\|y_d\|^2_{L^2(\Omega_h)} \leqslant \frac{1}{2}\|y_d\|^2_{L^2(\Omega)},$$

which implies the boundedness of $\{\bar{u}_h\}_{h>0}$ in $\mathcal{M}(\Omega)$. By taking a subsequence, we have that $\bar{u}_h \overset{*}{\rightharpoonup} \nu$ in $\mathcal{M}(\Omega)$. Then using (9.3.1), (9.3.13), the lower semicontinuity of the norm $\|\cdot\|_{\mathcal{M}(\Omega)}$ and (9.3.6) we get

$$J(\nu) \leqslant \liminf_{h\to 0} J_h(\bar{u}_h) \leqslant \limsup_{h\to 0} J_h(\bar{u}_h) \leqslant \limsup_{h\to 0} J_h(\Lambda_h\bar{u}) = J(\bar{u}).$$

Hence $\nu = \bar{u}$ by the uniqueness of the solution to (P), and the whole sequence $\{\bar{u}_h\}_{h>0}$ converges weakly* to $\bar{u}$. Also, from the above inequality we get (9.3.12). Using again (9.3.13), we deduce (9.3.11). Finally, (9.3.10) follows immediately from (9.3.11) and (9.3.12). □

9.4 ERROR ESTIMATES

This section is devoted to the proof of error estimates for the optimal costs as well as for the optimal states. We still require Ω to be convex and in addition we assume

$$y_d \in L^r(\Omega) \text{ with } r = \begin{cases} 4 & \text{if } n = 2, \\ \frac{8}{3} & \text{if } n = 3. \end{cases} \tag{9.4.1}$$

As in the previous sections, we denote by $\bar{y}$ and $\bar{y}_h$ the continuous and discrete states associated to the optimal controls $\bar{u}$ and $\bar{u}_h$, respectively.

Theorem 9.4.1. *There exists a constant $C > 0$ independent of h such that*

$$|J(\bar{u}) - J_h(\bar{u}_h)| \leqslant Ch^\kappa, \tag{9.4.2}$$

where $\kappa = 1$ if $n = 2$ and $\kappa = 1/2$ if $n = 3$.

Proof. We establish some preliminary estimates. Given $u \in \mathcal{M}(\Omega)$, with associated continuous and discrete states y and y_h, we know from [Casas 1985] that

$$\|y - y_h\|_{L^2(\Omega_h)} \leqslant Ch^\kappa\|u\|_{\mathcal{M}(\Omega)}, \tag{9.4.3}$$

with κ defined as in the statement of the theorem.

Taking r as in (9.4.1) and using Hölder's inequality and (9.3.1), we deduce that for all $\varphi \in L^r(\Omega)$,

$$\|\varphi\|_{L^2(\Omega\setminus\Omega_h)} \leqslant \|\varphi\|_{L^r(\Omega\setminus\Omega_h)}|\Omega \setminus \Omega_h|^{\frac{r-2}{2r}} \leqslant C\|\varphi\|_{L^r(\Omega\setminus\Omega_h)}h^{\frac{\kappa}{2}} \tag{9.4.4}$$

holds. As a consequence of (9.4.3) and (9.4.4), with $\varphi = y - y_d$, we get

$$\begin{aligned}(9.4.5)\quad \left|\|y-y_d\|^2_{L^2(\Omega)} - \|y_h-y_d\|^2_{L^2(\Omega_h)}\right| &\leqslant \|y-y_d\|^2_{L^2(\Omega\setminus\Omega_h)} \\ &\quad + \left(\|y-y_d\|_{L^2(\Omega_h)} + \|y_h-y_d\|_{L^2(\Omega_h)}\right)\|y-y_h\|_{L^2(\Omega_h)} \\ &\leqslant C\left(\|y-y_d\|^2_{L^r(\Omega\setminus\Omega_h)} + [\|y-y_d\|_{L^2(\Omega_h)} + \|y_h-y_d\|_{L^2(\Omega_h)}]\|u\|_{\mathcal{M}(\Omega)}\right)h^\kappa.\end{aligned}$$

Now, by the optimality of $\bar u$ and $\bar u_h$ we have

$$J(\bar u) - J_h(\bar u) \leqslant J(\bar u) - J_h(\bar u_h) \leqslant J(\bar u_h) - J_h(\bar u_h),$$

hence

$$(9.4.6)\qquad |J(\bar u) - J_h(\bar u_h)| \leqslant \max\{|J(\bar u) - J_h(\bar u)|, |J(\bar u_h) - J_h(\bar u_h)|\}.$$

From (9.3.10) we deduce that $\{\bar u_h\}_{h>0}$ is bounded in $\mathcal{M}(\Omega)$. Therefore, (9.2.2) implies that the continuous associated states $\{y_{\bar u_h}\}_{h>0}$ are bounded in $W^{1,p}_0(\Omega)$ for every $1 \leqslant p < \frac{n}{n-1}$ and therefore also in $L^r(\Omega)$. We apply (9.4.5) with $u = \bar u_h$ and $u = \bar u$, respectively. Together with (9.4.6) this establishes (9.4.2). □

In the following theorem we establish a rate of convergence for the states.

Theorem 9.4.2. *There exists a constant $C > 0$ independent of h such that*

$$(9.4.7)\qquad \|\bar y - \bar y_h\|_{L^2(\Omega)} \leqslant Ch^{\frac{\kappa}{2}},$$

with κ as defined in Theorem 9.4.1.

Proof. Let $S : \mathcal{M}(\Omega) \to L^2(\Omega)$ and $S_h : \mathcal{M}(\Omega) \to L^2(\Omega)$ be the solution operators associated to the equations (9.1.1) and (9.3.2), respectively. From (9.4.3) it follows that

$$(9.4.8)\qquad \|Su - S_h u\|_{L^2(\Omega_h)} \leqslant Ch^\kappa \|u\|_{\mathcal{M}(\Omega)}.$$

By the optimality of $\bar u$ we have for all $u \in \mathcal{M}(\Omega)$, that

$$(S\bar u - y_d, Su - S\bar u) + \alpha[\|u\|_{\mathcal{M}(\Omega)} - \|\bar u\|_{\mathcal{M}(\Omega)}] \geqslant 0,$$

where $(\cdot,\cdot)$ denotes the scalar product in $L^2(\Omega)$. In particular, taking $u = \bar u_h$, we get

$$(9.4.9)\qquad (S\bar u - y_d, S\bar u_h - S\bar u) + \alpha[\|\bar u_h\|_{\mathcal{M}(\Omega)} - \|\bar u\|_{\mathcal{M}(\Omega)}] \geqslant 0.$$

Analogously, the optimality of $\bar u_h$ implies that

$$(9.4.10)\qquad (S_h\bar u_h - y_d, S_h\bar u - S_h\bar u_h) + \alpha[\|\bar u\|_{\mathcal{M}(\Omega)} - \|\bar u_h\|_{\mathcal{M}(\Omega)}] \geqslant 0.$$

We point out that by definition of Y_h, we have $S_h u = 0$ in $\Omega \setminus \Omega_h$. Then, the scalar product above in $L^2(\Omega)$ coincides with that in $L^2(\Omega_h)$. Now, we rearrange terms in (9.4.10) as follows:

$$
\begin{aligned}
(9.4.11)\qquad (S\bar{u}_h - y_d, S\bar{u} - S\bar{u}_h) + (S_h\bar{u}_h - S\bar{u}_h, S_h\bar{u} - S_h\bar{u}_h) \\
+ (y_d, S\bar{u} - S_h\bar{u} + S_h\bar{u}_h - S\bar{u}_h) + (S\bar{u}_h, S_h\bar{u} - S\bar{u} + S\bar{u}_h - S_h\bar{u}_h) \\
+ \alpha[\|\bar{u}\|_{\mathcal{M}(\Omega)} - \|\bar{u}_h\|_{\mathcal{M}(\Omega)}] \geqslant 0.
\end{aligned}
$$

Now, adding (9.4.9) and (9.4.11) we obtain

$$
\begin{aligned}
(9.4.12)\qquad \|S\bar{u} - S_h\bar{u}_h\|^2_{L^2(\Omega)} &= (S\bar{u} - S_h\bar{u}_h, S\bar{u} - S_h\bar{u}_h) \\
&\leqslant (S_h\bar{u}_h - S\bar{u}_h, S_h\bar{u} - S_h\bar{u}_h) \\
&\quad + (y_d - S\bar{u}_h, S\bar{u} - S_h\bar{u} + S_h\bar{u}_h - S\bar{u}_h).
\end{aligned}
$$

Let us estimate the right hand terms. For the first one we apply the Cauchy–Schwarz inequality, exploit the fact $S_h\bar{u} - S_h\bar{u}_h = 0$ in $\Omega \setminus \Omega_h$ and use (9.4.8), to deduce

$$
(9.4.13)\quad (S_h\bar{u}_h - S\bar{u}_h, S_h\bar{u} - S_h\bar{u}_h) \leqslant \|S_h\bar{u}_h - S\bar{u}_h\|_{L^2(\Omega_h)}\|S_h\bar{u} - S_h\bar{u}_h\|_{L^2(\Omega_h)} \leqslant Ch^\kappa,
$$

where we have used that $\{\bar{u}_h\}_{h>0}$, $\{S_h\bar{u}\}_{h>0}$ and $\{S_h\bar{u}_h\}_{h>0}$ are bounded due to (9.3.10), (9.3.11) and (9.4.3), respectively. For the second term we use (9.4.4) and once again (9.4.8) as well as the fact that $S_h u = 0$ in $\Omega \setminus \Omega_h$, to obtain

(9.4.14)

$$
\begin{aligned}
(y_d - S\bar{u}_h, S\bar{u} - S_h\bar{u} + S_h\bar{u}_h - S\bar{u}_h) &\leqslant \|y_d - S\bar{u}_h\|_{L^2(\Omega\setminus\Omega_h)}\|S(\bar{u} - \bar{u}_h)\|_{L^2(\Omega\setminus\Omega_h)} \\
&\quad + \|y_d - S\bar{u}_h\|_{L^2(\Omega_h)}\|(S - S_h)(\bar{u} - \bar{u}_h)\|_{L^2(\Omega_h)} \\
&\leqslant C\left(\|y_d - S\bar{u}_h\|_{L^r(\Omega\setminus\Omega_h)}\|S(\bar{u} - \bar{u}_h)\|_{L^r(\Omega\setminus\Omega_h)} + \|\bar{u} - \bar{u}_h\|_{\mathcal{M}(\Omega)}\right) h^\kappa \leqslant Ch^\kappa,
\end{aligned}
$$

where we have also used that $y_d \in L^r(\Omega)$ and (9.2.2). Finally, (9.4.12), (9.4.13) and (9.4.14) prove (9.4.7). □

Remark 9.4.3. Let us observe that (9.4.2) and (9.4.7) imply that

$$
\left|\|\bar{u}\|_{\mathcal{M}(\Omega)} - \|\bar{u}_h\|_{\mathcal{M}(\Omega)}\right| \leqslant Ch^{\frac{\kappa}{2}}
$$

for some constant $C > 0$ independent of h.

Remark 9.4.4. All the previous results remain correct for a general elliptic operator

$$
Ay = -\sum_{i,j=1}^{n} \partial_{x_j}[a_{ij}\partial_{x_i} y] + a_0 y,
$$

provided the coefficients a_{ij} are Lipschitz continuous functions in $\overline{\omega}$ and $a_0 \geqslant 0$ is in $L^\infty(\Omega)$.

9.5 A NEUMANN CONTROL PROBLEM

In this section, we assume that the system is controlled on the boundary. The control problem is formulated as follows

$$\min_{u\in\mathcal{M}(\Gamma)} J_\Gamma(u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \alpha\|u\|_{\mathcal{M}(\Gamma)}, \tag{P_Γ}$$

where y is the unique solution to the Neumann problem

$$\begin{cases} -\Delta y + c_0 y &= f \quad \text{in } \Omega, \\ \partial_\nu y &= u \quad \text{on } \Gamma, \end{cases} \tag{9.5.1}$$

for $c_0 \in L^\infty(\Omega)$, $c_0 \geqslant 0$ and $c_0 \not\equiv 0$, and given $f \in L^1(\Omega)$. Here we will assume $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3, to be convex and polyhedral. Again by the Riesz representation theorem $\mathcal{M}(\Gamma)$ is identified with the dual space of $C(\Gamma)$; see, for instance, [Rudin 1970, Chapter 6]. Concerning the state equation (9.5.1), analogously to the Dirichlet problem (9.1.1), we say that an element $y \in W^{1,p}(\Omega)$, $p < \frac{n}{n-1}$, is a solution to (9.5.1) if

$$\int_\Omega y A z \, dx + \int_\Gamma y \partial_\nu z \, d\sigma = \int_\Omega f z \, dx + \int_\Gamma z \, du \qquad \text{for all } z \in H^2(\Omega).$$

We have the following theorem.

Theorem 9.5.1. *The problem* (9.5.1) *has a unique solution belonging to* $W^{1,p}(\Omega)$ *for every* $1 \leqslant p < \frac{n}{n-1}$, *and there exists a constant* $C_p > 0$ *such that*

$$\|y\|_{W^{1,p}(\Omega)} \leqslant C_p \left(\|f\|_{L^1(\Omega)} + \|u\|_{\mathcal{M}(\Gamma)} \right).$$

As a consequence of this theorem, we have that the functional $J_\Gamma : \mathcal{M}(\Gamma) \to \mathbb{R}$ is well defined. Moreover, it is continuous and strictly convex. Therefore, it has a unique minimizer that hereafter will be denoted by $\bar{u}$, with associated optimal state $\bar{y}$. Analogously to Theorem 9.2.1, if we denote the adjoint state associated to $\bar{u}$ by $\bar{\varphi}$,

$$\begin{cases} -\Delta \bar{\varphi} + c_0 \bar{\varphi} &= \bar{y} - y_d \quad \text{in } \Omega, \\ \partial_\nu \bar{\varphi} &= 0 \quad \text{on } \Gamma, \end{cases}$$

then the following identities hold

$$\alpha\|\bar{u}\|_{\mathcal{M}(\Gamma)} + \int_\Gamma \bar{\varphi} \, d\bar{u} = 0, \tag{9.5.2}$$

$$\|\bar{\varphi}\|_{C(\Gamma)} \begin{cases} = \alpha & \text{if } \bar{u} \neq 0, \\ \leqslant \alpha & \text{if } \bar{u} = 0. \end{cases} \tag{9.5.3}$$

Then, (9.5.2) and (9.5.3) imply a sparsity structure of $\bar{u}$ analogous to (9.2.7).

To carry out the numerical analysis of problem (P), we consider the same triangulation as in § 9.3. On this triangulation we define the space of discrete states by

$$Y_h = \left\{ y_h \in C(\overline{\omega}) : y_{h|T} \in \mathcal{P}_1 \text{ for every } T \in \mathcal{T}_h \right\},$$

and the discrete state equation

$$a(y_h, z_h) = \int_\Omega f z_h \, dx + \int_\Gamma z_h \, du \qquad \text{for all } z_h \in Y_h. \tag{9.5.4}$$

The approximation of the Neumann control problem results in

$$\min_{u \in \mathcal{M}(\Gamma)} J_h(u_h) = \frac{1}{2} \|y_h - y_d\|^2_{L^2(\Omega)} + \alpha \|u\|_{\mathcal{M}(\Gamma)}, \tag{$P_{\Gamma,h}$}$$

where y_h is the solution to (9.5.4). Before analyzing this problem, let us prove the following error estimates concerning the discretization of the state equation.

Theorem 9.5.2. *Given $u \in \mathcal{M}(\Gamma)$, let y and y_h be the solutions to (9.5.1) and (9.5.4). Then, there exists a constant $C > 0$ independent of h, f and u such that*

$$\|y - y_h\|_{L^2(\Omega)} \leqslant C h^\kappa \left(\|f\|_{L^1(\Omega)} + \|u\|_{\mathcal{M}(\Gamma)} \right), \tag{9.5.5}$$

with κ as in Theorem 9.4.1.

Proof. Here we follow the lines of the proof [Casas 1985, Theorem 3]. For any function $g \in L^2(\Omega)$, let $z \in H^2(\Omega)$ be the solution to

$$\begin{cases} -\Delta z + c_0 z &= g & \text{in } \Omega, \\ \partial_\nu z &= 0 & \text{on } \Gamma, \end{cases}$$

and $z_h \in Y_h$ the solution to

$$a(z_h, \varphi_h) = \int_\Omega g \varphi_h \, dx \qquad \text{for all } \varphi_h \in Y_h.$$

Using Green's formula, we obtain

$$\begin{aligned} \int_\Omega g(y - y_h) \, dx = a(y - y_h, z) = a(y, z) - a(y_h, z) = a(y, z) - a(y_h, z_h) \\ &= \int_\Omega f(z - z_h) \, dx + \int_\Gamma (z - z_h) \, du \\ &\leqslant \left(\|f\|_{L^1(\Omega)} + \|u\|_{\mathcal{M}(\Gamma)} \right) \|z - z_h\|_\infty \\ &\leqslant C \left(\|f\|_{L^1(\Omega)} + \|u\|_{\mathcal{M}(\Gamma)} \right) h^\kappa \|z\|_{H^2(\Omega)} \\ &\leqslant C \left(\|f\|_{L^1(\Omega)} + \|u\|_{\mathcal{M}(\Gamma)} \right) h^\kappa \|g\|_{L^2(\Omega)}, \end{aligned}$$

where we have used the classical finite element error estimate; see, for instance, [Ciarlet 1978, Chapter 3]. Since $g \in L^2(\Omega)$ is arbitrary, this gives the desired estimate. □

Analogously to § 9.3, we will denote by $\{x_j\}_{j=1}^{M(h)}$ the boundary nodes of the triangulation $\mathcal{T}_h$. Associated to these nodes we consider the space

$$Y_h^\Gamma = \left\{ y_h \in C(\Gamma) : y_{h|T\cap\Gamma} \in \mathcal{P}_1(T \cap \Gamma) \text{ for every } T \in \mathcal{T}_h^\Gamma \right\},$$

where $\{\mathcal{T}_h^\Gamma\}_{h>0}$ is the family of boundary triangles. A nodal basis of Y_h^Γ is given by the functions $\{e_j\}_{j=1}^{M(h)}$ such that $e_j(x_i) = \delta_{ij}$ for every $1 \leqslant i, j \leqslant M(h)$. Then, every element y_h of Y_h^Γ can be written in the form

$$y_h = \sum_{j=1}^{M(h)} y_j e_j, \quad \text{where} \quad y_j = y_h(x_j),\ 1 \leqslant j \leqslant M(h).$$

We also consider the space

$$D_h^\Gamma = \left\{ u_h \in \mathcal{M}(\Gamma) : u_h = \sum_{j=1}^{M(h)} \lambda_j \delta_{x_j},\ \text{where } \{\lambda_j\}_{j=1}^{M(h)} \subset \mathbb{R} \right\}.$$

Above, δ_{x_j} denotes the Dirac measure centered at the node x_j. It is obvious that D_h^Γ can be identified with the dual of Y_h^Γ through the duality relation

$$\langle u_h, y_h \rangle = \sum_{j=1}^{M(h)} \lambda_j y_j.$$

Now, we define the linear operators $\Pi_h : C(\Gamma) \to Y_h^\Gamma$ and $\Lambda_h : \mathcal{M}(\Gamma) \to D_h^\Gamma$ by

$$\Pi_h y = \sum_{j=1}^{M(h)} y(x_j) e_j \quad \text{and} \quad \Lambda_h u = \sum_{j=1}^{M(h)} \langle u, e_j \rangle \delta_{x_j}.$$

With the above notation, the identities (9.3.3) and (9.3.4) remain valid and (9.3.5) and (9.3.6) hold with Ω replaced by Γ. Also, the statement (iv) of Theorem 9.3.1 remains correct for $u \in \mathcal{M}(\Gamma)$. This, in particular, implies that Theorem 9.3.2 remains valid for the case of Neumann boundary control.

The analogous inequalities to (9.3.7) and (9.3.8) are

$$\|u - \Lambda_h u\|_{W^{-\frac{1}{p},p}(\Gamma)} \leqslant C h^{1-n/p'} \|u\|_{\mathcal{M}(\Gamma)}, \quad 1 < p < \frac{n}{n-1},$$
$$\|u - \Lambda_h u\|_{W^{1,\infty}(\Gamma)^*} \leqslant C h \|u\|_{\mathcal{M}(\Gamma)}.$$

To prove these inequalities let us consider an arbitrary element $z \in W^{\frac{1}{p},p'}(\Gamma)$. It is well known that $W^{\frac{1}{p},p'}(\Gamma)$ is the trace space of $W^{1,p'}(\Omega) \subset C(\overline{\omega})$; see [Nečas 1967]. Given $w \in W^{1,p'}(\Omega)$,

let us denote by w_h its nodal interpolation on the triangulation of $\overline{\omega}$. Then, arguing as in § 9.3, we obtain

$$\begin{aligned}\langle u - \Lambda_h u, z\rangle = \langle u, z - \Pi_h z\rangle &\leqslant \|u\|_{\mathcal{M}(\Gamma)}\|z - \Pi_h z\|_{C(\Gamma)}\\ &\leqslant \|u\|_{\mathcal{M}(\Gamma)} \inf_{w\in W^{1,p'}(\Omega),\gamma(w)=z} \|w - w_h\|_{C(\overline{\omega})}\\ &\leqslant Ch^{1-n/p'}\|u\|_{\mathcal{M}(\Omega)} \inf_{w\in W^{1,p'}(\Omega),\gamma(w)=z} \|w\|_{W_0^{1,p'}(\Omega)}\\ &= Ch^{1-n/p'}\|u\|_{\mathcal{M}(\Omega)}\|z\|_{W^{\frac{1}{p},p'}(\Gamma)}.\end{aligned}$$

Since $W^{\frac{1}{p},p'}(\Gamma)^* = W^{-\frac{1}{p},p}(\Gamma)$, the inequality (9.3.10) follows from the above inequality. The inequality (9.3.11) is proved analogously.

Hereafter, $\bar{u}_h$ will denote the unique solution to (P_h) in the space D_h with the associated discrete state $\bar{y}_h$. Then, as a consequence of Theorem 9.5.2 and the previous observations, we get that Theorem 9.3.2 remains true with Ω replaced by Γ.

Finally, error estimates analogous to (9.4.2) and (9.4.7) can be obtained following the same arguments, replacing (9.4.3) by (9.5.5) and taking into account that $\Omega = \Omega_h$, which obviously simplifies the proofs.

9.6 COMPUTATIONAL RESULTS

We illustrate the theoretical results of the previous sections with numerical examples in two dimensions. For our computational domain, we take the square $\Omega_h = \Omega = [-1, 1]^2$, which is discretized using the standard uniform triangulation arising from $N \times N$ equidistributed nodes. Unless stated otherwise, we fix $N = 128$, which corresponds to $h \approx 0.0157$, $c_0 = 0$, and $\alpha = 10^{-2}$.

The numerical solution of the discrete optimality system is based on an equivalent formulation of the optimality conditions (9.2.3) and (9.2.4). Returning to the characterization (9.2.5) of the subgradient, we have that the adjoint state $\bar{\varphi} \in C_0(\Omega)$ satisfies

$$-\bar{\varphi} \in \alpha\partial\|\cdot\|_{\mathcal{M}(\Omega)}(\bar{u}).$$

By the definition of the convex subdifferential, this is equivalent to

$$\bar{u} \in \partial I_{\{z\in C_0(\Omega):\|z\|_{C_0(\Omega)}\leqslant\alpha\}}(-\bar{\varphi}),$$

since the Fenchel conjugate of the indicator function of the (scaled) unit ball in $C_0(\Omega)$ is the (scaled) norm in $\mathcal{M}(\Omega)$. The subdifferential of the indicator function is then given by the normal cone, which can be characterized by the variational inequality

$$\langle \bar{u}, \bar{\varphi} - \varphi\rangle \leqslant 0 \quad \text{for all } \|\varphi\|_{C_0(\Omega)} \leqslant \alpha.$$

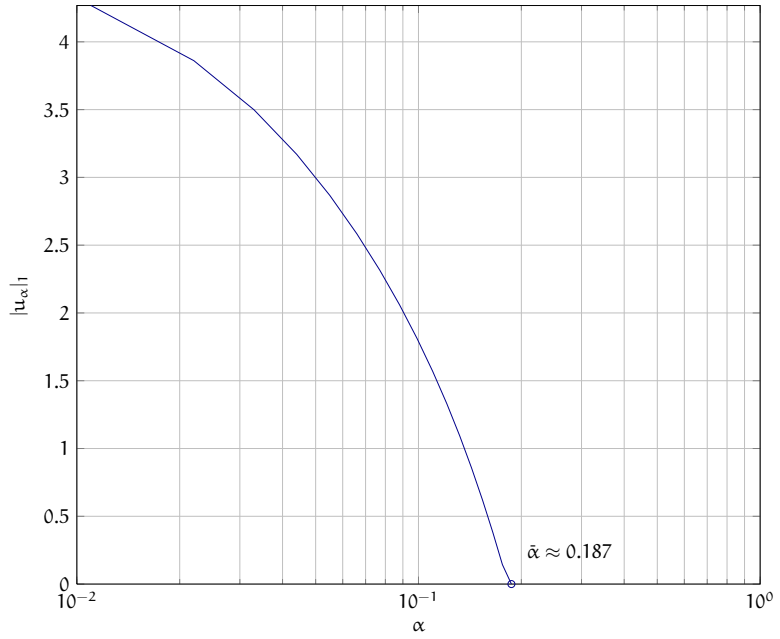


Figure 9.1: Dependence of norm of optimal control u_h on penalty parameter α.

We now pass to the discrete setting by replacing the continuous control $\bar{u}$ with its discretization $\bar{u}_h$ and introducing the discrete adjoint state $\bar{\varphi}_h = \sum_{j=1}^{N(h)} \varphi_j e_j \in Y_h$. The above variational inequality can then be reformulated using a complementarity function as

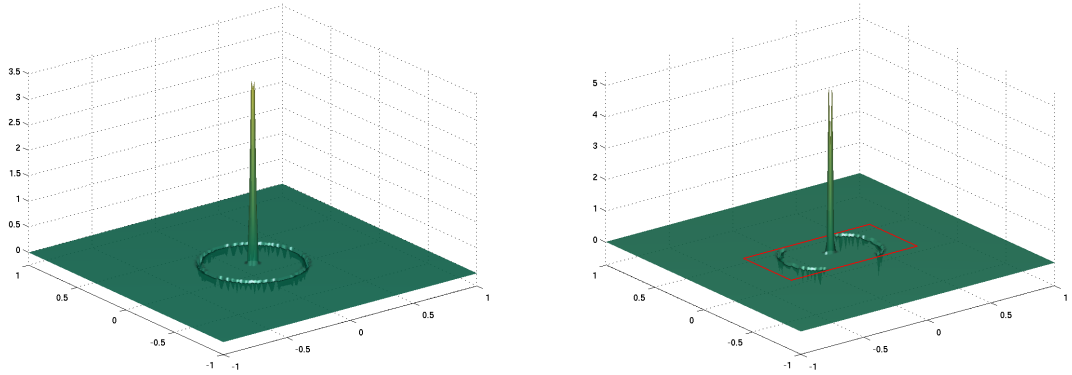
$$\bar{u}_h + \max(0, -\bar{u}_h + \bar{\varphi}_h - \alpha) + \min(0, -\bar{u}_h + \bar{\varphi}_h + \alpha) = 0,$$

which should be understood component-wise in terms of the vector of expansion coefficients $(\lambda_1, \dots, \lambda_{N(h)})$ and $(\varphi_1, \dots, \varphi_{N(h)})$. This is a locally Lipschitz mapping from $\mathbb{R}^{N(h)} \times \mathbb{R}^{N(h)} \to \mathbb{R}^{N(h)}$ and thus the reformulated discrete optimality system can be solved by a locally superlinearly convergent semismooth Newton method [Kummer 1992; Qi and Sun 1993]. The corresponding algorithm was implemented in Matlab (R2011a).

We first illustrate the structural properties of the optimal controls. Figure 9.1 shows the norm of the optimal control u_α as a function of the penalty parameter α. As verified in Proposition 9.2.2, there exists an $\bar{\alpha}$ (≈ 0.187), such that $u_\alpha \equiv 0$ for $\alpha > \bar{\alpha}$.

The statement of Proposition 9.2.3 is illustrated in Figure 9.2, where the optimal controls for the target $y_d = 10 \exp(-50\|x\|^2)$ and different observation domains ω_y are compared. As a reference, Figure 9.2a shows the control for $\omega_y = \Omega$ (in the form of its expansion coefficients λ_j at each grid point, with linear interpolation for better visibility). In contrast, the control for $\omega_y = \chi_{\{|x_1|<1/2\}}\chi_{\{|x_2|<1/4\}} \subsetneq \Omega$ vanishes outside of ω_y, see Figure 9.2b.

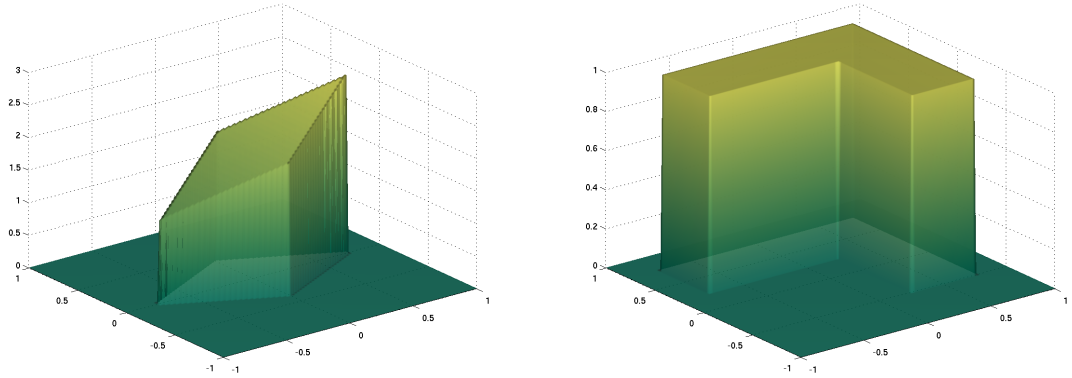
We now investigate the convergence behavior as $h \to 0$. In the absence of a known exact solution, we take as reference solution the computed optimal discrete control and optimal discrete state on the finest grid with $N^* = 2^{10}$, corresponding to $h^* = 2 \cdot 10^{-3}$. We first consider distributed control, with the target $y_{d,1}$ given in Figure 9.3a. Figure 9.4a shows the difference $|J_h - J_{h^*}|$ for a series of successively refined, nested grids for $N = 2^3, \dots, 2^9$.



(a) u_h, observation on $\omega_y = \Omega$

(b) u_h, observation on $\omega_y \subsetneq \Omega$ (in red)

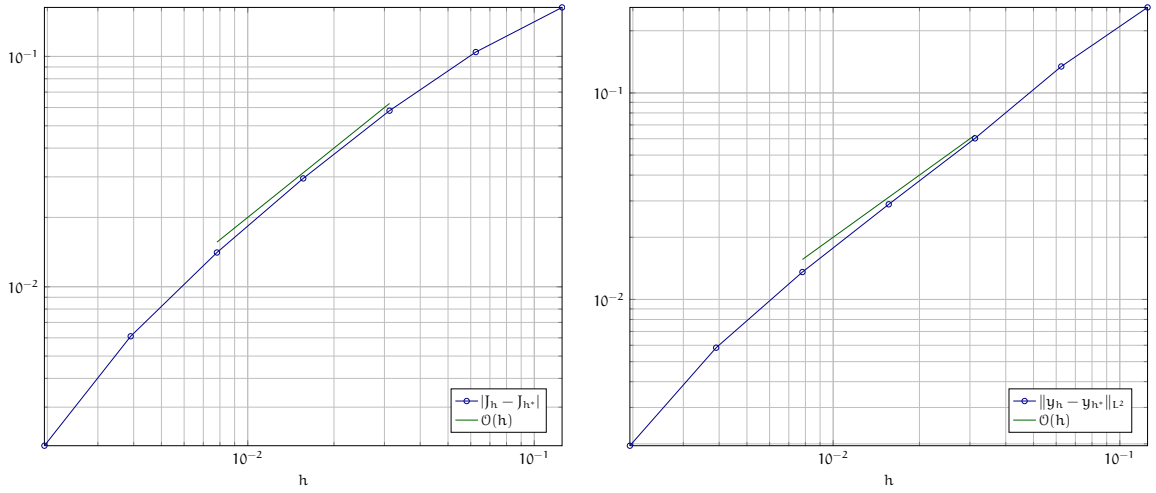
Figure 9.2: Comparison of optimal controls u_h for full observation ($\omega_y = \Omega$) and partial observation ($\omega_y \subsetneq \Omega$, marked in red).



(a) target $y_{d,1}$

(b) target $y_{d,2}$

Figure 9.3: Target states for convergence rate examples



(a) functional value J_h

(b) L^2 norm of state y_h

Figure 9.4: Illustration of convergence order for distributed control

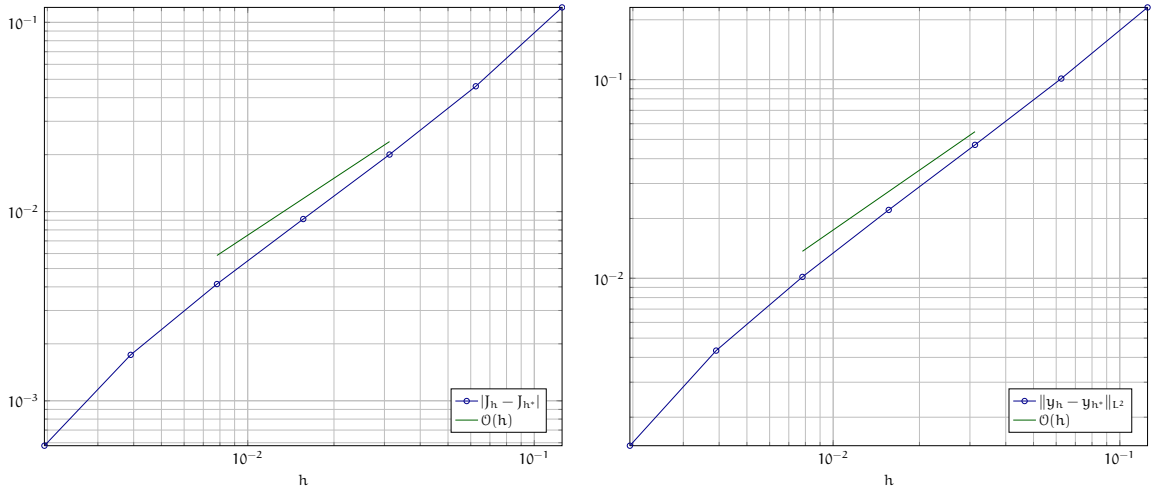
The observed linear convergence rate agrees well with the rate obtained in Theorem 9.4.1. The corresponding L^2 error $\|y_h - y_{h^*}\|_{L^2}$ of the discrete states also decays with a linear rate, which is faster than predicted by Theorem 9.4.2.

For the case of Neumann control, we set $\alpha = 5 \cdot 10^{-2}$ and $c_0 = 10^{-2}$ and consider the target $y_{d,2}$ shown in Figure 9.3b. Again, both the error in the functional value (Figure 9.5a) and in the state (Figure 9.5b) follow an approximately linear convergence rate.

To illustrate the sparsity properties of Neumann boundary controls, Figure 9.6 shows the optimal control $u_{h,\alpha}$ (again, in the form of its linearly interpolated coefficients λ_j) for $\alpha = 10^{-3}$, 10^{-2} and 10^{-1}, plotted along boundary sections as indicated.

Finally, we address a control theoretic issue.[1] In problem (P), the penalty parameter controls both the sparsity and the magnitude – and hence the effect – of the optimal controls. This may not be desirable; in fact, the measure theoretic formulation may serve primarily the purpose of optimal actuator placement, and this is different from the optimal control problem. We therefore carried out experiments where we identified control locations based on the optimal measure space control, and then solved a standard quadratic optimal problem. Furthermore, true point controls may not be realizable in practice, so one would define the control to act as piecewise constants on patches centered on the optimal Dirac measures. To illustrate these points, we report on an experiment for the target shown in Figure 9.3b with $N = 65$ nodes in each direction, where we solve problem (P) for a value of α chosen to give a strongly localized optimal control ($\alpha = 2 \cdot 10^{-2}$, cf. Figure 9.7a). We then select the four Dirac measures δ_{ξ_i}, $i = 1, \dots, 4$, of largest magnitude and define control patches ω_i by combining all triangles adjacent to the node ξ_i of each Dirac measure (cf. Figure 9.7b, in green). For comparison, we

[1] This example was not part of the published version of the work this chapter is based on.



(a) functional value J_h

(b) L^2 norm of state y_h

Figure 9.5: Illustration of convergence order for Neumann control

heuristically select four control patches of the same area around the vertices of a rectangle containing the support of the target (cf. Figure 9.7b, in red). Note that the "optimal" locations are placed asymmetrically with respect to the target (outlined in black).

Then, we solve for each set of patches the optimal control problem

$$\min_{u\in\mathbb{R}^4} \frac{1}{2}\|y_h - y_d\|^2_{L^2(\Omega)} + \frac{\beta}{2}|u|_2^2,$$

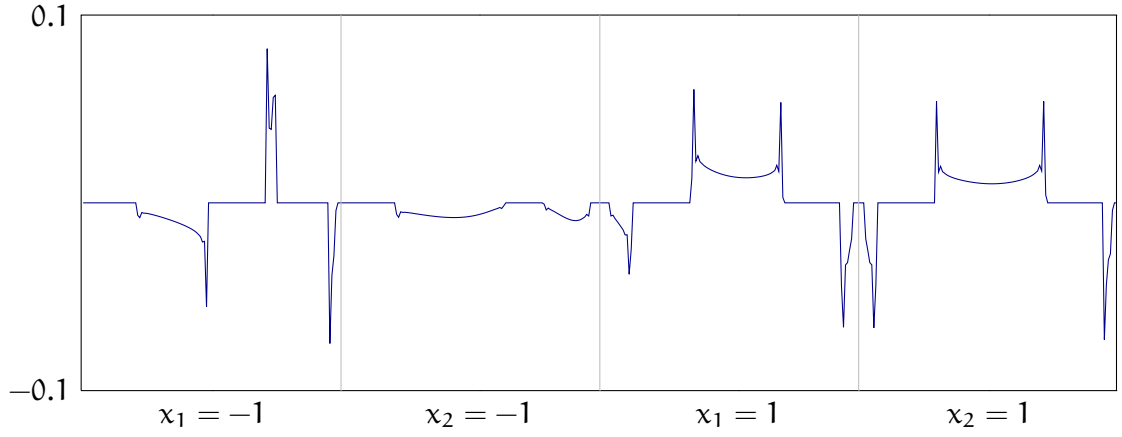
where $y_h \in Y_h$ is the finite element solution of the Dirichlet problem

$$a(y_h, z_h) = \sum_{i=1}^{4} (u\chi_{\omega_i}, z_h) \qquad \text{for all } z_h \in Y_h,$$

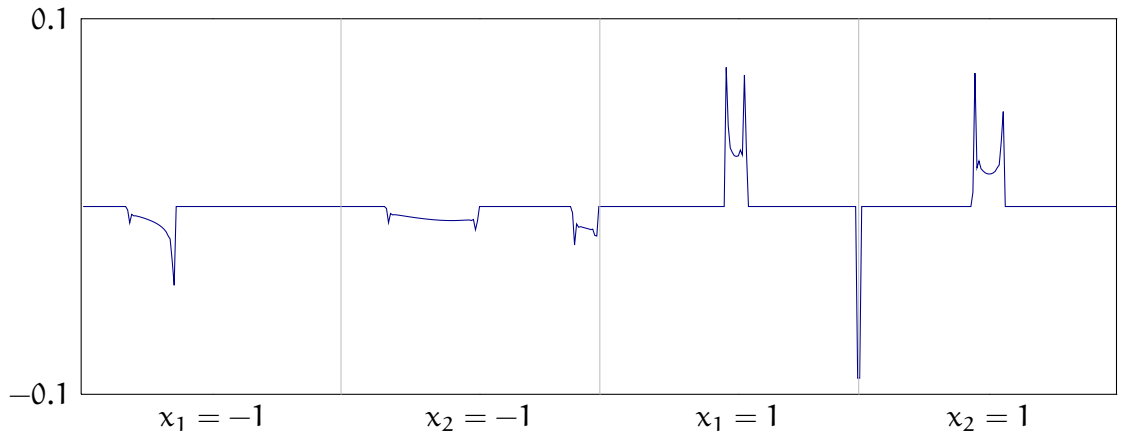
for all $z_h \in Y_h$, with the characteristic function χ_{ω_i} taken as piecewise constant on each element. Here, β is chosen in each case to achieve an optimal control with given energy $|u^*|_2 \approx 1000$. For the patches chosen according to the optimal measure space control, this was $\beta = 1.95 \cdot 10^{-7}$, and for the heuristic choice we set $\beta = 1.15 \cdot 10^{-7}$. The resulting controls are shown in Figure 9.7c and 9.7d, respectively. Although the controls are visually similar, the tracking error $\|y_h^* - y_d\|^2_{L^2(\Omega)}$ for the heuristic locations is almost twice as large as that for the locations based on the optimal measure space control (cf. Table 9.1).

9.7 CONCLUSION

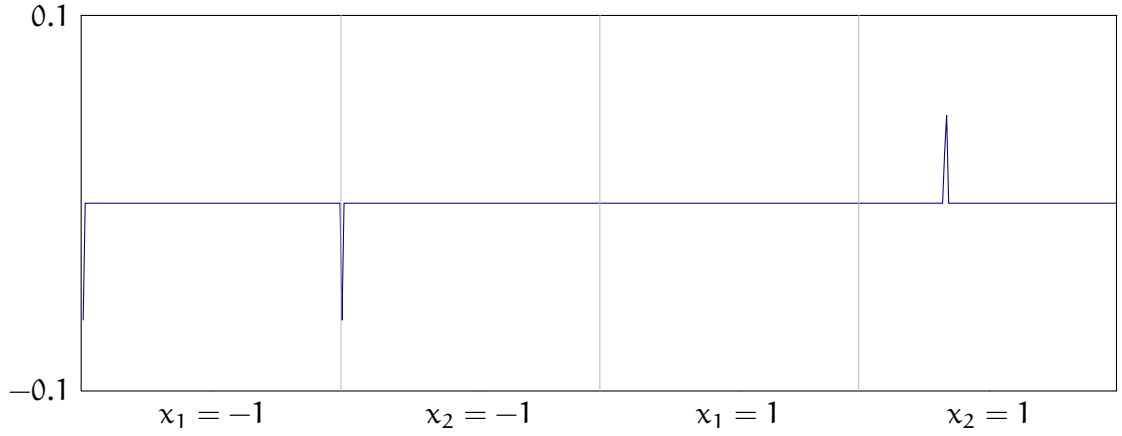
By considering optimal control problems in spaces of measures, controls with strong sparsity properties can be obtained. Although the non-reflexive Banach space setting complicates the



(a) $\alpha = 10^{-3}$



(b) $\alpha = 10^{-2}$

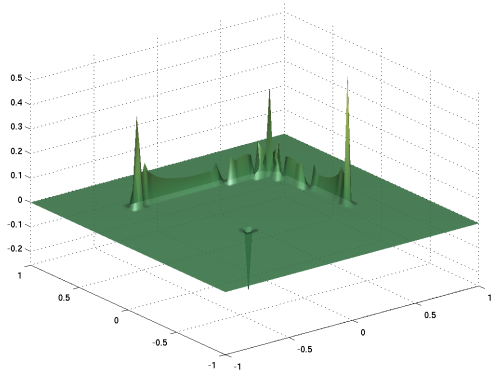


(c) $\alpha = 10^{-1}$

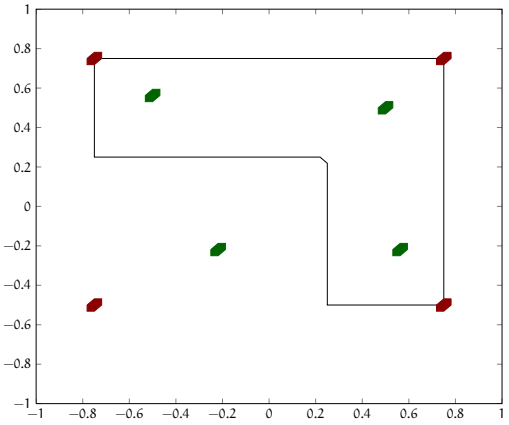
Figure 9.6: Optimal Neumann control $u_{h,\alpha}$ for increasing values of α

Table 9.1: Results of comparison of optimal and heuristic placement of controls.

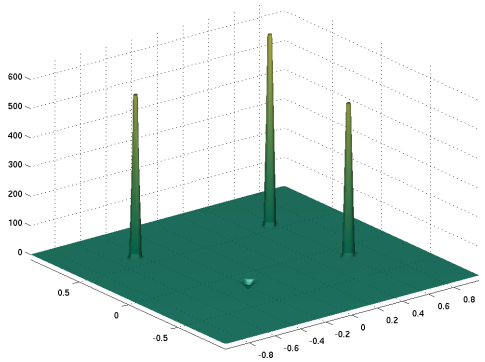
placement	tracking error	control energy	control area
optimal	0.44843	999.7636	0.011719
heuristic	0.86210	1001.3401	0.011719



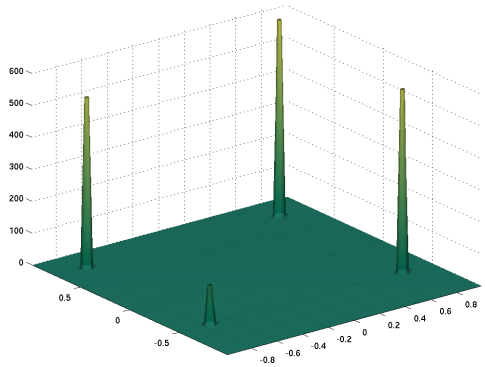
(a) measure control ($\alpha = 2 \cdot 10^{-2}$)



(b) control patches (green: based on Fig. 9.7a, red: heuristic)



(c) optimal control based on measure control



(d) optimal control based on heuristic placement

Figure 9.7: Comparison of optimal and heuristic placement of control patches.

analysis, a straightforward numerical approximation that retains the structural properties of the measure norm is possible. In a sense, the results of this paper justify the "intuitive" discretization of regular Borel measures by Dirac measures on a set of nodes.

ACKNOWLEDGMENTS

The first author was supported by Spanish Ministerio de Ciencia e Innovación under projects MTM2008-04206 and "Ingenio Mathematica (i-MATH)" No. CSD2006-00032 (Consolider Ingenio 2010). The second and third authors were supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences").

PARABOLIC CONTROL PROBLEMS IN MEASURE SPACES WITH SPARSE SOLUTIONS

10

ABSTRACT

Optimal control problems in measure spaces lead to controls that have small support, which is desirable, e.g., in the context of optimal actuator placement. For problems governed by parabolic partial differential equations, well-posedness is guaranteed in the space of square-integrable measure-valued functions, which leads to controls with a spatial sparsity structure. A conforming approximation framework allows deriving numerically accessible optimality conditions as well as convergence rates. In particular, although the state is discretized, the control problem can still be formulated and solved in the measure space. Numerical examples illustrate the structural features of the optimal controls.

10.1 INTRODUCTION

This paper is concerned with the analysis and approximation of the optimal control problem

$$\text{(P)} \qquad \min_{u \in L^2(I,\mathcal{M}(\Omega))} J(u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_T)} + \alpha\|u\|_{L^2(\mathcal{M})},$$

where $I = [0, T]$ and y is the unique solution to the initial–boundary value problem

$$\text{(10.1.1)} \qquad \begin{cases} \partial_t y - \Delta y &= u \quad \text{in } \Omega_T = \Omega \times (0,T), \\ y &= 0 \quad \text{on } \Sigma_T = \Gamma \times (0,T), \\ y(x,0) &= y_0 \quad \text{in } \Omega \end{cases}$$

for given $y_0 \in L^2(\Omega)$. We assume that $\alpha > 0$, $y_d \in L^2(\Omega_T)$ and Ω is a bounded domain in $\mathbb{R}^n$, $1 \leqslant n \leqslant 3$, which is supposed to either be convex or have a $C^{1,1}$ boundary Γ. Hereafter $\mathcal{M}(\Omega)$ denotes the space of regular Borel measures in Ω and $\|u\|_{L^2(\mathcal{M})}$ denotes the norm of u in the space $L^2(I, \mathcal{M}(\Omega))$; see section 10.2 below for details.

Formulating the control problem in a measure space is motivated by the observation that the resulting optimal controls possess sparsity properties (i.e., have small support), which is desirable in many applications such as optimal sensor or actuator placement; see [Clason and Kunisch 2012; Casas, Clason, and Kunisch 2012] in the context of elliptic equations. Although similar features can be achieved using L^1 control costs, the corresponding control problem in general does not admit a solution in the absence of further regularization because L^1 spaces lack the necessary compactness properties. For parabolic problems, the situation is even more delicate since (10.1.1) is not well-posed for right hand sides in $\mathcal{M}(\Omega_T)$ (which would require $C(\Omega_T)$ regularity for the adjoint equation; see Definition 10.2.1 below). This leads to considering controls in $L^2(I, \mathcal{M}(\Omega))$. The associated norm $\|u\|_{L^2(\mathcal{M})}$ for the control is a natural one from the point of view of well-posedness of the state equation (10.1.1) and allows for sparsity in space. The numerical results will illustrate precisely this property of our formulation. The spatio-temporal coupling of the corresponding control cost, however, presents a challenge for deriving numerically useful optimality conditions.

Besides the analysis of the control problem (P), the main focus of this paper consists in providing an approximation framework which, in spite of the difficulties due to the measure space setting, leads to implementable schemes for which a priori error estimates can be provided. We show that the optimal measure controls can be approximated efficiently by linear combinations of Dirac measures in space which are piecewise constant in time. We point out that even after discretization, the control problem is formulated and solved in the measure space.

Let us mention some related works. A similar approximation framework for elliptic control problems in measure spaces was proposed in [Casas, Clason, and Kunisch 2012]. Differently from the elliptic case, parabolic control problems with sparsity-promoting constraints have received very little attention. In [Casas and Zuazua 2012], the approximate control of $y(T)$ by measures $u \in \mathcal{M}([t_0, t_1] \times \Omega)$ with $0 < t_0 < t_1 < T$ is discussed (using the smoothing property of the heat equation to ensure $y(T) \in L^2(\Omega)$); finite-dimensional approximation and numerical solution are not addressed. Although not specifically concerned with parabolic equations, the approach of [Herzog, Stadler, and Wachsmuth 2012] covers control problems with $L^1(\Omega, L^2([0, T]))$ control costs (together with additional pointwise control constraints). The resulting optimal controls have *directional sparsity*, i.e., their support is constant in time. In contrast, we will show that solutions to (P) have a non-separable sparsity structure.

This paper is organized as follows. In the next section, we discuss the functional analytic setting of the control problem and analyze well-posedness of the state equation. Section 10.3 is concerned with existence of and optimality conditions for solutions to (P), the latter implying a sparsity property of the optimal controls. The proposed approximation framework is the subject of section 10.4, where we introduce the discretization (§ 10.4.1) and show convergence of solutions to the discretized state equation (§ 10.4.2) and to the discrete optimal control problem (§ 10.4.3). Convergence rates are derived in section 10.5. Section 10.6 addresses the numerical solution of the discrete control problem, for which we derive a reformulated optimality system that is amenable to solution by a semismooth Newton method. (The

continuous counterpart of this optimality system is sketched in Appendix 10.A.) Finally, section 10.7 illustrates the structure of the optimal controls with some numerical examples.

10.2 FUNCTION SPACES AND WELL-POSEDNESS OF THE STATE EQUATION

In this section we first define the control space and give some of its properties. Then, we turn to the analysis of the state equation.

10.2.1 CONTROL SPACE

We denote by $C_0(\Omega)$ the space of continuous functions in $\overline{\omega}$ vanishing on $\Gamma = \partial\Omega$, endowed with the supremum norm $\|\cdot\|_\infty$. Its topological dual is identified with the space of regular Borel measures in Ω, denoted by $\mathcal{M}(\Omega)$. Moreover, we have

$$\|u\|_{\mathcal{M}} = \sup\left\{\int_\Omega z\,du : z \in C_0(\Omega) \text{ and } \|z\|_\infty \leqslant 1\right\} = |u|(\Omega),$$

where $|u|$ denotes the total variation measure.

Associated to the interval $I = [0,T]$ we define the spaces $L^2(I, C_0(\Omega))$ and $L^2(I, \mathcal{M}(\Omega))$, where $L^2(I, C_0(\Omega))$ is the space of measurable functions $z : [0,T] \to C_0(\Omega)$ for which the associated norm given by

$$\|z\|_{L^2(C_0)} = \left(\int_0^T \|z(t)\|_\infty^2\,dt\right)^{1/2}$$

is finite. Due to the fact that $C_0(\Omega)$ is a separable Banach space, $L^2(I, C_0(\Omega))$ is also a separable Banach space; see e.g. [Warga 1972, Theorem I.5.18].

As a consequence of the non-separability of $\mathcal{M}(\Omega)$, the definition of the space $L^2(I, \mathcal{M}(\Omega))$ is more delicate. Indeed, we need to distinguish between weakly and strongly measurable functions $u : [0,T] \to \mathcal{M}(\Omega)$. Hereafter we denote by $L^2(I, \mathcal{M}(\Omega))$ the space of weakly measurable functions u for which the norm

$$\|u\|_{L^2(\mathcal{M})} = \left(\int_0^T \|u(t)\|_{\mathcal{M}}^2\,dt\right)^{1/2}$$

is finite. This choice makes $L^2(I, \mathcal{M}(\Omega))$ a Banach space and guarantees that it can be identified with the dual of $L^2(I, C_0(\Omega))$, where the duality relation is given by

$$\langle u, z\rangle_{L^2(\mathcal{M}),L^2(C_0)} = \int_0^T \langle u(t), z(t)\rangle\,dt,$$

with $\langle\cdot,\cdot\rangle$ denoting the duality between $\mathcal{M}(\Omega)$ and $C_0(\Omega)$. The reader is referred to [Edwards 1965, section 8.14.1 and Proposition 8.15.3] for the different notions of measurability and [Edwards 1965, Theorem 8.20.3] for the duality identification. (The distinction between weak and strong measurability is not required for the space $L^2(I, C_0(\Omega))$ because $C_0(\Omega)$ is separable and hence both notions are equivalent; see [Edwards 1965, Theorem 8.15.2].)

10.2.2 ANALYSIS OF THE STATE EQUATION

Given $1 < p < \infty$, we denote by $W_0^{1,p}(\Omega)$ the Sobolev space of functions of $L^p(\Omega)$ with distributional derivatives in $L^p(\Omega)$ and having a zero trace on Γ and we set $W^{-1,p'}(\Omega)$ to be the dual of $W_0^{1,p}(\Omega)$, where $1/p' + 1/p = 1$. These spaces are reflexive and separable, and hence the spaces $L^2(I, W_0^{1,p}(\Omega))$ formed by the measurable functions $y : [0,T] \to W_0^{1,p}(\Omega)$ for which the norm

$$\|y\|_{L^2(W_0^{1,p})} = \left(\int_0^T \|y(t)\|^2_{W_0^{1,p}}\,dt\right)^{1/2}$$

is finite, are separable and reflexive Banach spaces whose dual is identified with $L^2(I, W^{-1,p'}(\Omega))$; see [Edwards 1965, Theorem 8.25.5].

The notion of solution to the state equation makes use of the following space of test functions

$$\mathcal{Z} = \{z \in H^{2,1}(\Omega_T) : z = 0 \text{ on } \Sigma_T \text{ and } z(T) = 0 \text{ in } \Omega\},$$

where

$$H^{2,1}(\Omega_T) = \left\{z \in L^2(\Omega_T) : \partial_t z, \frac{\partial^{|\beta|} z}{\partial x^\beta} \in L^2(\Omega_T), \text{ with } \beta \in \mathbb{N}^n,\ |\beta| \leqslant 2\right\}$$

is endowed with the graph norm. By the Rellich–Kondrachov theorem, $\mathcal{Z}$ embeds compactly into $L^2(I, C_0(\Omega))$.

Definition 10.2.1. *We say that $y \in L^2(\Omega_T)$ is a solution to equation (10.1.1) if*

$$\int_{\Omega_T} y(-\partial_t z - \Delta z)\,dx\,dt = \int_0^T \langle u(t), z(t)\rangle\,dt + \int_\Omega y_0(x) z(x,0)\,dx, \quad \forall z \in \mathcal{Z}. \tag{10.2.1}$$

Theorem 10.2.2. *For all $(u, y_0) \in L^2(I, \mathcal{M}(\Omega)) \times L^2(\Omega)$ the equation (10.1.1) has a unique solution y. Moreover, $y \in L^2(I, W_0^{1,p}(\Omega))$ for every $p \in [1, \frac{n}{n-1})$ and there exist constants C_p such that*

$$\|y\|_{L^2(W_0^{1,p})} \leqslant C_p\left(\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}\right). \tag{10.2.2}$$

Proof. We adapt the proof of [Casas 1997]. Let $\{u_k\}_k$ be a sequence in $C(\overline{\omega}_T)$ satisfying

$$u_k \overset{*}{\rightharpoonup} u \text{ in } L^2(I, \mathcal{M}(\Omega)) \text{ and } \|u_k\|_{L^2(L^1)} \leqslant \|u\|_{L^2(\mathcal{M})}. \tag{10.2.3}$$

Let $y_k \in L^2(I, H_0^1(\Omega))$ denote the variational solution to

$$\begin{cases} \partial_t y_k - \Delta y_k &= u_k & \text{in } \Omega_T, \\ y_k &= 0 & \text{on } \Sigma_T, \\ y_k(x,0) &= y_0(x) & \text{in } \Omega. \end{cases} \tag{10.2.4}$$

For $\psi = (\psi_0, \dots, \psi_n) \in \mathcal{D}(\Omega_T)^{n+1}$ we denote by $z \in \mathcal{Z}$ the solution to

$$\begin{cases} -\partial_t z - \Delta z &= \psi_0 - \sum_{j=1}^n \partial_{x_j} \psi_j & \text{in } \Omega_T, \\ z &= 0 & \text{on } \Sigma_T, \\ z(x,T) &= 0 & \text{in } \Omega. \end{cases} \tag{10.2.5}$$

From the last two equations we get for any $1 < p < \frac{n}{n-1}$

$$\begin{aligned} \int_{\Omega_T} (\psi_0 y_k + \sum_{j=1}^n \psi_j \partial_{x_j} y_k) \, dx \, dt &= \int_{\Omega_T} u_k z \, dx \, dt + \int_\Omega y_0(x) z(x,0) \, dx \\ &\leqslant \|u_k\|_{L^2(L^1)} \|z\|_{L^2(W_0^{1,p'})} + \|y_0\|_{L^2(\Omega)} \|z(0)\|_{L^2(\Omega)}. \end{aligned}$$

In the following estimate we use maximal regularity of the heat equation in an essential way. If Ω is convex, its boundary is of Lipschitz class, and hence there exists a $\hat{p}$ with $\hat{p} > 4$ if $n = 2$ and $\hat{p} > 3$ when $n = 3$ such that $\Delta : W_0^{1,p}(\Omega) \to W^{-1,p}(\Omega)$ is an isomorphism for each $\hat{p}' < p < \hat{p}$, where $1/\hat{p}' + 1/\hat{p} = 1$; see [Jerison and Kenig 1995]. (If $n = 1$ or if Ω has a $C^{1,1}$ boundary, $\Delta : W_0^{1,p}(\Omega) \to W^{-1,p}(\Omega)$ is an isomorphism for every $1 < p < +\infty$.) In particular, combining [Haller–Dintelmann and Rehberg 2009, Theorem 5.4] and (10.2.3), we obtain for every $\hat{p}' < p < \frac{n}{n-1} < \hat{p}$ the existence of a constant $\hat{C}_p$ such that

$$\begin{aligned} \int_{\Omega_T} y_k \left(\psi_0 - \sum_{j=1}^n \partial_{x_j} \psi_j \right) dx \, dt &= \int_{\Omega_T} (\psi_0 y_k + \sum_{j=1}^n \psi_j \partial_{x_j} y_k) \, dx \, dt \\ &\leqslant \hat{C}_p \left(\|y_0\|_{L^2(\Omega)} + \|u\|_{L^2(\mathcal{M})} \right) \sum_{j=0}^n \|\psi_j\|_{L^2(L^{p'})}. \end{aligned}$$

From the density of $\{\psi_0 - \sum_{j=1}^n \partial_{x_j} \psi_j : \psi \in \mathcal{D}(\Omega_T)^{n+1}\}$ in $L^2(I, W^{-1,p'}(\Omega))$ and the duality identification $L^2(I, W_0^{1,p}(\Omega))^* = L^2(I, W^{-1,p'}(\Omega))$, we deduce the boundedness of $\{y_k\}_{k=1}^\infty$ in $L^2(I, W_0^{1,p}(\Omega))$ and the existence of a constant C_p such that

$$\|y_k\|_{L^2(W_0^{1,p})} \leqslant C_p \left(\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)} \right). \tag{10.2.6}$$

Using the reflexivity of $L^2(I, W_0^{1,p}(\Omega))$, we can obtain a subsequence, denoted in the same way, and an element $y \in L^2(I, W_0^{1,p}(\Omega))$ such that $y_k \rightharpoonup y$ in $L^2(I, W_0^{1,p}(\Omega))$.

For $\psi_0 \in L^2(\Omega_T)$ arbitrary and $z \in \mathcal{Z}$ solution to (10.2.5) for $\psi_i = 0$, $1 \leqslant j \leqslant n$, it follows from (10.2.4) and (10.2.5) that

$$\int_{\Omega_T} y_k(-\partial_t z - \Delta z)\,dx\,dt = \int_{\Omega_T} y_k \psi_0\,dx\,dt = \int_{\Omega_T} u_k z\,dx\,dt + \int_\Omega y_0(x) z(x,0)\,dx.$$

Passing to the limit in this identity and in (10.2.6), we obtain (10.2.1) and (10.2.2). Using the fact that $\partial_t + \Delta$ is an isomorphism from $\mathcal{Z}$ to $L^2(\Omega_T)$ and (10.2.1), we conclude the uniqueness of $y \in W_0^{1,p}(\Omega)$.

Finally, independence of y with respect to p follows from the existence of a solution y in $L^2(I, W_0^{1,p}(\Omega))$ for every $\hat{p}' < p < \frac{n}{n-1}$ and its uniqueness in $L^2(\Omega_T)$, since $W_0^{1,p_1}(\Omega) \subset W_0^{1,p_2}(\Omega)$ for $p_1 > p_2$. □

Remark 10.2.3.

(i) The solution to (10.1.1) belongs to $L^2(I, W_0^{1,p}(\Omega))$ for every $\hat{p} \leqslant p < \frac{n}{n-1}$, and from the equation (10.1.1) we know that $\partial_t y \in L^2(I, W^{-1,p}(\Omega))$. Observe that $W_0^{1,p}(\Omega) \subset L^2(\Omega)$ for $p \geqslant p_0 := \max\{\hat{p}', \frac{2n}{n+2}\}$, with $\hat{p}$ as in the proof of Theorem 10.2.2, and hence $y \in L^2(\Omega_T)$. As a consequence, we deduce that $y \in C(I, L^2(\Omega))$; see [Showalter 1997, Proposition III.1.2].

(ii) Under our regularity conditions, an equivalent definition for the solution to equation (10.1.1) is the following. A function $y \in L^2(I, W_0^{1,p}(\Omega))$ with $p_0 < p < \frac{n}{n-1}$ is called a solution to (10.1.1) if

$$-\int_0^T \langle y(t), \partial_t z(t)\rangle_{W_0^{1,p}, W^{-1,p'}}\,dt + \int_{\Omega_T} \nabla y \nabla z\,dx\,dt = \int_0^T \langle u(t), z(t)\rangle\,dt + \int_\Omega y_0(x) z(x,0)\,dx$$

for all $z \in L^2(I, W_0^{1,p'}(\Omega))$ such that $\partial_t z \in L^2(I, W^{-1,p'}(\Omega))$ (which implies $z(\cdot,0) \in L^2(\Omega)$; see (i)) and $z(T) = 0$. This follows from (10.2.1) and the density of $\mathcal{Z}$ in this new space of test functions. Theorem 10.2.2 remains valid with this definition if we only assume for Ω to have a Lipschitz boundary. This is the regularity of Ω required to have the maximal parabolic regularity; see [Haller–Dintelmann and Rehberg 2009]. We have chosen the above definition because it is more convenient for the numerical analysis to be developed later in this paper.

(iii) The preceding theorem as well as the rest of the results given in this paper are valid if we replace the heat operator in (10.1.1) by a more general parabolic operator $\partial_t + A$ that enjoys maximal parabolic regularity.

We finish this section by proving a continuity result of the states with respect to the controls.

Theorem 10.2.4. *Let $\{u_k\}_{k=1}^\infty \subset L^2(I, \mathcal{M}(\Omega))$ be a sequence such that $u_k \overset{*}{\rightharpoonup} u$ in $L^2(I, \mathcal{M}(\Omega))$. If y_k and y denote the states associated to u_k and u, respectively, then $\|y_k - y\|_{L^2(\Omega_T)} \to 0$.*

Proof. For every k, let $z_k \in \mathcal{Z}$ satisfy

$$\begin{cases} -\partial_t z_k - \Delta z_k &= y - y_k & \text{in } \Omega_T, \\ z_k &= 0 & \text{on } \Sigma_T, \\ z_k(x,T) &= 0 & \text{in } \Omega. \end{cases}$$

Then, from Definition 10.2.1 and using the boundedness of $\{u_k\}_{k=1}^\infty$ in $L^2(I,\mathcal{M}(\Omega))$, we have

$$\begin{aligned} \|y - y_k\|^2_{L^2(\Omega_T)} &= \int_{\Omega_T} (y-y_k)(-\partial_t z_k - \Delta z_k)\,dx\,dt = \int_0^T \langle u(t) - u_k(t), z_k(t)\rangle\,dt \\ &\leqslant \|u - u_k\|_{L^2(\mathcal{M})}\|z_k\|_{L^2(C_0)} \leqslant C\|z_k\|_{L^2(C_0)}. \end{aligned}$$

From Theorem 10.2.2, we know that $y_k \rightharpoonup y$ in $L^2(\Omega_T)$, therefore $z_k \rightharpoonup 0$ in $H^{2,1}(\Omega_T)$. Since the embedding $H^{2,1}(\Omega_T) \subset L^2(I, C_0(\Omega))$ is compact, we get that $\|z_k\|_{L^2(C_0)} \to 0$. This convergence and the above inequality conclude the proof. □

10.3 ANALYSIS OF THE CONTROL PROBLEM

In this section we establish existence of an optimal control and derive the optimality conditions.

Proposition 10.3.1. *The control problem* (P) *has a unique solution* $\bar{u}$.

Proof. Let $\{u_k\}_{k=1}^\infty$ be a minimizing sequence, which is thus bounded in the space $L^2(I,\mathcal{M}(\Omega))$. Since the predual $L^2(I, C_0(\Omega))$ is separable, there exists a subsequence, denoted in the same way, converging weakly-$\star$ to some $\bar{u} \in L^2(I,\mathcal{M}(\Omega))$. From Theorem 10.2.4 we get that $y(u_k) \to y(\bar{u})$ strongly in $L^2(\Omega_T)$. Hence, the weakly-$\star$ lower semicontinuity of the norm $\|\cdot\|_{L^2(\mathcal{M})}$ implies that $\bar{u}$ is a solution. The uniqueness is a consequence of the strict convexity of J, which follows from the injectivity of the control-to-state mapping. □

Hereafter $\bar{u}$ will denote the solution to (P) and $\bar{y}$ the associated state. Now, we give the first order optimality conditions, which are necessary and sufficient due to the convexity of (P).

Theorem 10.3.2. *There exists a unique element* $\bar{\varphi} \in H^{2,1}(\Omega_T)$ *satisfying*

$$\begin{cases} -\partial_t \bar{\varphi} - \Delta \bar{\varphi} &= \bar{y} - y_d & \text{in } \Omega_T, \\ \bar{\varphi} &= 0 & \text{on } \Sigma_T, \\ \bar{\varphi}(x,T) &= 0 & \text{in } \Omega, \end{cases} \tag{10.3.1}$$

such that

$$\int_0^T \langle \bar{u}(t), \bar{\varphi}(t)\rangle\,dt + \alpha\|\bar{u}\|_{L^2(\mathcal{M})} = 0, \tag{10.3.2}$$

$$\|\bar{\varphi}\|_{L^2(C_0)} \begin{cases} = \alpha & \text{if } \bar{u} \neq 0, \\ \leqslant \alpha & \text{if } \bar{u} = 0. \end{cases} \tag{10.3.3}$$

Proof. Let us introduce $j(u) = \|u\|_{L^2(\mathcal{M})}$ and $F(u) = \frac{1}{2}\|y(u) - y_d\|^2_{L^2(\Omega_T)}$, so that $J(u) = F(u) + \alpha j(u)$. By the differentiability of F and the convexity of j we obtain

$$F'(\bar{u})(u - \bar{u}) + \alpha j(u) - \alpha j(\bar{u}) \geqslant 0 \quad \forall u \in L^2(I, \mathcal{M}(\Omega)),$$

and hence

$$\int_{\Omega_T} (\bar{y} - y_d)(y(u) - \bar{y})\, dx\, dt + \alpha j(u) - \alpha j(\bar{u}) \geqslant 0.$$

Utilizing the adjoint equation (10.3.1) and the state equation (10.2.1), we deduce from the above inequality

$$\int_0^T \langle u(t) - \bar{u}(t), \bar{\varphi}(t)\rangle\, dt + \alpha j(u) - \alpha j(\bar{u}) \geqslant 0 \quad \forall u \in L^2(I, \mathcal{M}(\Omega)). \tag{10.3.4}$$

Taking $u = 2\bar{u}$ and $u = \frac{1}{2}\bar{u}$, respectively, in (10.3.4) we obtain (10.3.2). On the other hand, setting $u = \bar{u} - v$ in (10.3.4), it follows that

$$\int_0^T \langle v(t), \bar{\varphi}(t)\rangle\, dt \leqslant \alpha(j(\bar{u} - v) - j(\bar{u})) \leqslant \alpha\|v\|_{L^2(\mathcal{M})} \quad \forall v \in L^2(I, \mathcal{M}(\Omega)). \tag{10.3.5}$$

By the duality $L^2(I, \mathcal{M}(\Omega)) = L^2(I, C_0(\Omega))^*$ we have that

$$\|\bar{\varphi}\|_{L^2(C_0)} = \max_{\|v\|_{L^2(\mathcal{M})} \leqslant 1} \int_0^T \langle v(t), \bar{\varphi}(t)\rangle\, dt \leqslant \alpha. \tag{10.3.6}$$

Then, (10.3.3) is an immediate consequence of (10.3.2) and (10.3.6). □

From now on, we will assume that the optimal control $\bar{u} \neq 0$. By using (10.3.2) and (10.3.3) we can prove some sparsity property for $\bar{u}$. Let us consider the Jordan decomposition $\bar{u}(t) = \bar{u}^+(t) - \bar{u}^-(t)$ for almost every $t \in I$. Then we have the following theorem.

Theorem 10.3.3. *For almost every* $t \in I$ *the following embeddings hold*

$$\operatorname{supp}(\bar{u}^+(t)) \subset \{x \in \Omega : \bar{\varphi}(x, t) = -\|\bar{\varphi}(t)\|_\infty\}, \tag{10.3.7}$$

$$\operatorname{supp}(\bar{u}^-(t)) \subset \{x \in \Omega : \bar{\varphi}(x, t) = +\|\bar{\varphi}(t)\|_\infty\}. \tag{10.3.8}$$

Proof. Since $\bar{\varphi} : I \times \overline{\omega} \to \mathbb{R}$ is a Caratheodory function, there exists a measurable selection $t \in I \mapsto x_t \in \overline{\omega}$ such that $\bar{\varphi}(x_t, t) = \|\bar{\varphi}(t)\|_\infty$; see [Ekeland and Témam 1999, Chapter 8, Theorem 1.2]. Now, we define the element $v \in L^2(I, \mathcal{M}(\Omega))$ by $v(t) = \operatorname{sign}(\bar{\varphi}(x_t))\|u(t)\|_{\mathcal{M}}\delta_{x_t}$. We have to check that $v : I \to \mathcal{M}(\Omega)$ is weakly measurable. To this end the only delicate point is the weak measurability of $t \in I \mapsto \delta_{x_t} \in \mathcal{M}(\Omega)$. This follows from the measurability of the

mapping $t \mapsto x_t$ and the continuity of $x \in \overline{\omega} \mapsto \delta_x \in \mathcal{M}(\Omega)$ when $\mathcal{M}(\Omega)$ is endowed with the weak-$\star$ topology. By definition of ν we get

$$\langle \nu(t), \bar{\varphi}(t) \rangle = \|\bar{u}(t)\|_{\mathcal{M}} \|\bar{\varphi}(t)\|_\infty \geqslant -\langle \bar{u}(t), \bar{\varphi}(t) \rangle \tag{10.3.9}$$

and

$$\begin{aligned} \|\nu\|_{L^2(\mathcal{M})} &= \left(\int_0^T \|\bar{u}(t)\|_{\mathcal{M}}^2 \|\delta_{x_t}\|_{\mathcal{M}}^2 \, dt \right)^{1/2} = \left(\int_0^T \|\bar{u}(t)\|_{\mathcal{M}}^2 \, dt \right)^{1/2} \\ &= \|\bar{u}\|_{L^2(\mathcal{M})}. \end{aligned} \tag{10.3.10}$$

From (10.3.2), (10.3.9), (10.3.5) and (10.3.10) we obtain

$$\alpha \|\bar{u}\|_{L^2(\mathcal{M})} = -\int_0^T \langle \bar{u}(t), \bar{\varphi}(t) \rangle \, dt \leqslant \int_0^T \langle \nu(t), \bar{\varphi}(t) \rangle \, dt \leqslant \alpha \|\nu\|_{L^2(\mathcal{M})} = \alpha \|\bar{u}\|_{L^2(\mathcal{M})}.$$

As a consequence of these inequalities and (10.3.9) we conclude that

$$\|\bar{u}(t)\|_{\mathcal{M}} \|\bar{\varphi}(t)\|_\infty = -\langle \bar{u}(t), \bar{\varphi}(t) \rangle \quad \text{for a. e. } t \in I. \tag{10.3.11}$$

Finally, (10.3.7) and (10.3.8) follow from (10.3.11) and Lemma 10.3.4 below applied to $\mu = -\bar{u}(t)$. □

Lemma 10.3.4. *Let $\mu \in \mathcal{M}(\Omega)$ and $z \in C_0(\Omega)$, both of them not zero, be such that*

$$\langle \mu, z \rangle = \|\mu\|_{\mathcal{M}} \|z\|_\infty, \tag{10.3.12}$$

and let $\mu = \mu^+ - \mu^-$ be the Jordan decomposition of μ. Then we have

$$\operatorname{supp}(\mu^+) \subset \Omega_+ = \{x \in \Omega : z(x) = +\|z\|_\infty\}, \tag{10.3.13}$$

$$\operatorname{supp}(\mu^-) \subset \Omega_- = \{x \in \Omega : z(x) = -\|z\|_\infty\}. \tag{10.3.14}$$

Proof. We will prove (10.3.13), the proof of (10.3.14) being analogous. First we observe that due to (10.3.12) we obtain for all measures $\nu \in \mathcal{M}(\Omega)$ with $\|\nu\|_{\mathcal{M}} \leqslant \|\mu\|_{\mathcal{M}}$ that

$$\langle \nu, z \rangle \leqslant \|\nu\|_{\mathcal{M}} \|z\|_\infty \leqslant \|\mu\|_{\mathcal{M}} \|z\|_\infty = \langle \mu, z \rangle. \tag{10.3.15}$$

We have as well that

$$\langle \mu, z \rangle = \langle \mu^+, z^+ \rangle + \langle \mu^-, z^- \rangle - \langle \mu^+, z^- \rangle - \langle \mu^-, z^+ \rangle \leqslant \langle \mu^+, z^+ \rangle + \langle \mu^-, z^- \rangle.$$

Moreover, the inequality is strict unless μ^+ and μ^- are concentrated at the set of points $x \in \Omega$ where $z(x) \geqslant 0$ and $z(x) \leqslant 0$, respectively. Let us define the sets

$$A_+ = \{x \in \Omega : z(x) \geqslant 0\} \text{ and } A_- = \{x \in \Omega : z(x) \leqslant 0\}$$

and the measures $\nu^+ = \mu^+ \mid_{A_+}$, $\nu^- = \mu^- \mid_{A_-}$ and $\nu = \nu^+ - \nu^-$. Then we have that $\|\nu\|_{\mathcal{M}} \leqslant \|\mu\|_{\mathcal{M}}$ and $\langle \nu, z\rangle > \langle \mu, z\rangle$ if $\operatorname{supp}(\mu^+) \not\subset A_+$ or $\operatorname{supp}(\mu^-) \not\subset A_-$. Because of (10.3.15) we conclude that $\operatorname{supp}(\mu^+) \subset A_+$ and $\operatorname{supp}(\mu^-) \subset A_-$. Now we distinguish two cases in the proof of (10.3.13) depending on whether the norm bound is attained from above.

Case 1: $\max_{x\in\overline{\omega}} z(x) < \|z\|_\infty$. In this case we prove that $\mu^+ = 0$. Indeed, let $x_0 \in \Omega$ such that $z(x_0) = -\|z\|_\infty$ and define $\nu = -\mu^+(\Omega)\delta_{x_0} - \mu^-$. Then it is obvious that $\|\nu\|_{\mathcal{M}} = \|\mu\|_{\mathcal{M}}$. If $\mu^+ \neq 0$, since the support of μ^+ is in A_+ and $\max_{x\in\overline{\omega}} z(x) < \|z\|_\infty$, we have that

$$\langle \nu, z\rangle = \|z\|_\infty \mu^+(\Omega) - \langle \mu^-, z\rangle > \langle \mu^+, z\rangle - \langle \mu^-, z\rangle = \langle \mu, z\rangle,$$

which contradicts (10.3.15). Then, (10.3.13) holds.

Case 2: $\max_{x\in\overline{\omega}} z(x) = \|z\|_\infty$. Let $x_0 \in \Omega$ be such that $z(x_0) = \|z\|_\infty$. We argue by contradiction and assume that $\mu^+(S) > 0$ where

$$S = \{x \in \Omega : 0 \leqslant z(x) < \|z\|_\infty\}.$$

We take $\nu = \mu^+(\Omega)\delta_{x_0} - \mu^-$ and once again

$$\|\nu\|_{\mathcal{M}} = \|\mu\|_{\mathcal{M}} \text{ and } \langle \nu, z\rangle = \mu^+(\Omega)\|z\|_\infty - \langle \mu^-, z\rangle > \langle \mu, z\rangle,$$

since $\mu^+(S) > 0$. Again this contradicts (10.3.15). Therefore, $\mu^+(S) = 0$ and hence (10.3.13) follows from the inclusion $\operatorname{supp}(\mu^+) \subset A_+$. □

Corollary 10.3.5. *There exists* $\bar\alpha > 0$ *such that* $\bar u = 0$ *for every* $\alpha > \bar\alpha$.

Proof. Let us denote by J_α the cost functional associated to the parameter α. Similarly, let $(u_\alpha, y_\alpha, \varphi_\alpha)$ denote the solution to the corresponding optimality system. For each $\alpha > 0$ we have the inequalities

$$\frac{1}{2}\|y_\alpha - y_d\|^2_{L^2(\Omega_T)} \leqslant J_\alpha(u_\alpha) \leqslant J_\alpha(0) = \frac{1}{2}\|\hat y_0 - y_d\|^2_{L^2(\Omega_T)},$$

where $\hat y_0$ denotes the uncontrolled state, i.e., the solution to (10.1.1) with $u = 0$. Consequently, $\|y_\alpha - y_d\|_{L^2(\Omega_T)} \leqslant \|\hat y_0 - y_d\|_{L^2(\Omega_T)}$ holds for every $\alpha > 0$. From the adjoint state equation (10.3.1) and the embedding of $H^{2,1}(\Omega_T) \hookrightarrow L^2(I, C(\overline{\omega}))$, we deduce the existence of a constant $C > 0$ such that

$$\|\varphi_\alpha\|_{L^2(C_0)} \leqslant C'\|\bar\varphi\|_{H^{2,1}} \leqslant C\|y_\alpha - y_d\|_{L^2(\Omega_T)} \leqslant C\|\hat y_0 - y_d\|_{L^2(\Omega_T)}.$$

Setting $\bar\alpha = C\|\hat y_0 - y_d\|_{L^2(\Omega_T)}$, we obtain from the above inequality and (10.3.3) that $u_\alpha = 0$ for every $\alpha > \bar\alpha$. □

10.4 APPROXIMATION OF THE CONTROL PROBLEM

We consider a dG(0)cG(1) discontinuous Galerkin approximation of the state equation (10.1.1) (i.e., piecewise constant in time and linear nodal basis finite elements in space; see, e.g., [Thomée 2006]). Associated with a parameter h we consider a family of triangulations $\{\mathcal{K}_h\}_{h>0}$ of $\overline{\omega}$. To every element $K \in \mathcal{K}_h$ we assign two parameters $\rho(K)$ and $\vartheta(K)$, where $\rho(K)$ denotes the diameter of K and $\vartheta(K)$ is the diameter of the biggest ball contained in K. The size of the grid is given by $h = \max_{K\in\mathcal{K}_h} \rho(K)$. We will denote by $\{x_j\}_{j=1}^{N_h}$ the interior nodes of the triangulation $\mathcal{K}_h$. In this section Ω will be assumed to be convex. In addition, the following usual regularity assumptions on the triangulation are assumed.

(i) There exist two positive constants ρ_Ω and ϑ_Ω such that

$$\frac{h}{\rho(K)} \leqslant \rho_\Omega \quad \text{and} \quad \frac{\rho(K)}{\vartheta(K)} \leqslant \vartheta_\Omega$$

hold for every $K \in \mathcal{K}_h$ and all $h > 0$.

(ii) Let us set $\overline{\Omega}_h = \bigcup_{K\in\mathcal{K}_h} K$ with Ω_h and Γ_h being its interior and boundary, respectively. We assume that the vertices of $\mathcal{K}_h$ placed on the boundary Γ_h are also points of Γ and there exists a constant $C_\Gamma > 0$ such that $\operatorname{dist}(x, \Gamma) \leqslant C_\Gamma h^2$ for every $x \in \Gamma_h$. This always holds if Γ is a C^2 boundary. In the case of polygonal or polyhedral domains, it is reasonable to assume that the triangulation satisfies that $\Gamma_h = \Gamma$. From this assumption we know [Raviart and Thomas 1983, section 5.2] that

$$|\Omega \setminus \Omega_h| \leqslant Ch^2, \tag{10.4.1}$$

where $|\cdot|$ denotes the Lebesgue measure.

We also introduce a temporal grid $0 = t_0 < t_1 < \ldots < t_{N_\tau} = T$ with $\tau_k = t_k - t_{k-1}$ and set $\tau = \max_{1\leqslant k\leqslant N_\tau} \tau_k$. We assume that there exist $\rho_T > 0$, $C_{\Omega,T} > 0$ and $c_{\Omega,T} > 0$ independent of h and τ such that

$$\tau \leqslant \rho_T \tau_k, \text{ for } 1 \leqslant k \leqslant N_\tau \quad \text{and} \quad c_{\Omega,T} h^{\max\{n,2\}} \leqslant \tau \leqslant C_{\Omega,T} h^{\max\{n,2\}}. \tag{10.4.2}$$

We will use the notation $\sigma = (\tau, h)$ and $\Omega_{hT} = \Omega_h \times (0, T)$.

10.4.1 DISCRETIZATION OF THE CONTROLS AND STATES

We first discuss the spatial discretization, which follows [Casas, Clason, and Kunisch 2012]. Associated to the interior nodes $\{x_j\}_{j=1}^{N_h}$ of $\mathcal{K}_h$ we consider the spaces

$$U_h = \left\{ u_h \in \mathcal{M}(\Omega) : u_h = \sum_{j=1}^{N_h} u_j \delta_{x_j}, \text{ where } \{u_j\}_{j=1}^{N_h} \subset \mathbb{R} \right\}$$

and

$$Y_h = \left\{ y_h \in C_0(\Omega) : y_h = \sum_{j=1}^{N_h} y_j e_j, \text{ where } \{y_j\}_{j=1}^{N_h} \subset \mathbb{R} \right\},$$

where $\{e_j\}_{j=1}^{N_h}$ is the nodal basis formed by the continuous piecewise linear functions such that $e_j(x_i) = \delta_{ij}$ for every $1 \leqslant i, j \leqslant N_h$. Such functions attain their maximum and minimum at one of the nodes, and thus for all $y_h \in Y_h$,

$$\|y_h\|_\infty = \max_{1 \leqslant j \leqslant N_h} |y_j| = |\vec{y}_h|_\infty,$$

where we have identified y_h with the vector $\vec{y}_h = (y_1, \dots, y_{N_h})^T \in \mathbb{R}^{N_h}$ of its expansion coefficients, and $|\cdot|_p$ denotes the usual p-norm in $\mathbb{R}^{N_h}$. Similarly, we have for all $u_h \in U_h$ that

$$\|u_h\|_{\mathcal{M}} = \sup_{\|v\|_\infty = 1} \sum_{j=1}^{N_h} u_j \langle \delta_{x_j}, v \rangle = \sum_{j=1}^{N_h} |u_j| = |\vec{u}_h|_1 \quad \text{for all } u_h \in U_h.$$

Hence endowed with these norms, U_h is the topological dual of Y_h with respect to the duality pairing

$$\langle u_h, y_h \rangle = \sum_{j=1}^{N_h} u_j y_j = \vec{u}_h^T \vec{y}_h.$$

For every σ we define the space of discrete controls and states by

$$\mathcal{U}_\sigma = \{u_\sigma \in L^2(I, U_h) : u_\sigma|_{I_k} \in U_h,\ 1 \leqslant k \leqslant N_\tau\}$$

and

$$\mathcal{Y}_\sigma = \{y_\sigma \in L^2(I, Y_h) : y_\sigma|_{I_k} \in Y_h,\ 1 \leqslant k \leqslant N_\tau\},$$

where $I_k = (t_{k-1}, t_k]$. The elements $u_\sigma \in \mathcal{U}_\sigma$ and $y_\sigma \in \mathcal{Y}_\sigma$ can be represented in the form

$$u_\sigma = \sum_{k=1}^{N_\tau} u_{k,h} \chi_k \quad \text{and} \quad y_\sigma = \sum_{k=1}^{N_\tau} y_{k,h} \chi_k,$$

where χ_k is the characteristic function of I_k, $u_{k,h} \in U_h$ and $y_{k,h} \in Y_h$. Moreover, by definition of U_h and Y_h, we can write

$$u_\sigma = \sum_{k=1}^{N_\tau} \sum_{j=1}^{N_h} u_{kj} \chi_k \delta_{x_j} \quad \text{and} \quad y_\sigma = \sum_{k=1}^{N_\tau} \sum_{j=1}^{N_h} y_{kj} \chi_k e_j.$$

Thus $\mathcal{U}_\sigma$ and $\mathcal{Y}_\sigma$ are finite dimensional spaces of dimension $N_\tau \times N_h$, and bases are given by $\{\chi_k \delta_{x_j}\}_{k,j}$ and $\{\chi_k e_j\}_{k,j}$. Identifying again u_σ with the vector $\vec{u}_\sigma$ of expansion coefficients

u_{kj}, we have for all $u_\sigma \in \mathcal{U}_\sigma$ that

$$\begin{aligned}\|u_\sigma\|^2_{L^2(\mathcal{M})} &= \int_0^T \Big\| \sum_{k=1}^{N_\tau} \sum_{j=1}^{N_h} u_{kj}\chi_k\delta_{x_j} \Big\|^2_{\mathcal{M}} \,dt = \sum_{k=1}^{N_\tau} \int_{I_k} \Big\| \sum_{j=1}^{N_h} u_{kj}\delta_{x_j} \Big\|^2_{\mathcal{M}} \,dt \\ &= \sum_{k=1}^{N_\tau} \tau_k \left(\sum_{j=1}^{N_h} |u_{kj}| \right)^2 \\ &= \sum_{k=1}^{N_\tau} \tau_k |\vec{u}_k|_1^2\end{aligned}$$

for $\vec{u}_k = (u_{k1}, \dots, u_{kN_h})^T$, and similarly for all $y_\sigma \in \mathcal{Y}_\sigma$ that

$$\|y_\sigma\|^2_{L^2(C_0)} = \sum_{k=1}^{N_\tau} \tau_k \left(\max_{1\leqslant j\leqslant N_h} |y_{kj}| \right)^2 = \sum_{k=1}^{N_\tau} \tau_k |\vec{y}_k|^2_\infty.$$

It is thus straightforward to verify that endowed with these norms, $\mathcal{U}_\sigma$ is the topological dual of $\mathcal{Y}_\sigma$ with respect to the duality pairing

$$\langle u_\sigma, y_\sigma \rangle = \sum_{k=1}^{N_\tau} \tau_k \sum_{j=1}^{N_h} u_{kj} y_{kj} = \sum_{k=1}^{N_\tau} \tau_k (\vec{u}_k^T \vec{y}_k). \tag{10.4.3}$$

Next we define the linear operators $\Lambda_h : \mathcal{M}(\Omega) \to U_h \subset \mathcal{M}(\Omega)$ and $\Pi_h : C_0(\Omega) \to Y_h \subset C_0(\Omega)$ by

$$\Lambda_h u = \sum_{j=1}^{N_h} \langle u, e_j \rangle \delta_{x_j} \quad \text{and} \quad \Pi_h y = \sum_{j=1}^{N_h} y(x_j) e_j.$$

The operator Π_h is the nodal interpolation operator for Y_h. Concerning the operator Λ_h we have the following result.

Theorem 10.4.1 ([Casas, Clason, and Kunisch 2012, Theorem 3.1]). *The following properties hold.*

(i) For every $u \in \mathcal{M}(\Omega)$ and every $y \in C_0(\Omega)$ and $y_h \in Y_h$ we have

$$\begin{aligned}\langle u, y_h \rangle &= \langle \Lambda_h u, y_h \rangle, \\ \langle u, \Pi_h y \rangle &= \langle \Lambda_h u, y \rangle.\end{aligned}$$

(ii) For every $u \in \mathcal{M}(\Omega)$ we have

$$\begin{aligned}&\|\Lambda_h u\|_{\mathcal{M}} \leqslant \|u\|_{\mathcal{M}}, \\ &\Lambda_h u \overset{*}{\rightharpoonup} u \text{ in } \mathcal{M}(\Omega) \text{ and } \|\Lambda_h u\|_{\mathcal{M}} \to \|u\|_{\mathcal{M}} \text{ as } h \to 0.\end{aligned}$$

(iii) There exists a constant $C > 0$ *such that for every* $u \in \mathcal{M}(\Omega)$ *we have*

$$\|u - \Lambda_h u\|_{W^{-1,p}(\Omega)} \leqslant C h^{1-n/p'} \|u\|_{\mathcal{M}}, \quad 1 < p < \frac{n}{n-1},$$
$$\|u - \Lambda_h u\|_{(W_0^{1,\infty}(\Omega))^*} \leqslant C h \|u\|_{\mathcal{M}},$$

with $1/p' + 1/p = 1$.

Similarly to Λ_h and Π_h we define the linear operators

$$\Phi_\sigma : L^2(I, \mathcal{M}(\Omega)) \to \mathcal{U}_\sigma \subset L^2(I, \mathcal{M}(\Omega))$$

and

$$\Psi_\sigma : L^2(I, C_0(\Omega)) \to \mathcal{Y}_\sigma \subset L^2(I, C_0(\Omega))$$

by

$$\Phi_\sigma u = \sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \int_{I_k} \Lambda_h(u(t))\, dt\, \chi_k = \sum_{k=1}^{N_\tau} \sum_{j=1}^{N_h} \frac{1}{\tau_k} \int_{I_k} \langle u(t), e_j \rangle\, dt\, \chi_k \delta_{x_j},$$

$$\Psi_\sigma y = \sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \int_{I_k} \Pi_h(y(t))\, dt\, \chi_k = \sum_{k=1}^{N_\tau} \sum_{j=1}^{N_h} \frac{1}{\tau_k} \int_{I_k} y(x_j, t)\, dt\, \chi_k e_j.$$

Analogously to Theorem 10.4.1 we obtain the following result concerning Φ_σ and Ψ_σ.

Theorem 10.4.2. *The following properties hold.*

(i) For every $u_\sigma \in \mathcal{U}_\sigma$ *and every* $y_\sigma \in \mathcal{Y}_\sigma$ *we have*

$$\Phi_\sigma u_\sigma = u_\sigma \text{ and } \Psi_\sigma y_\sigma = y_\sigma. \tag{10.4.4}$$

(ii) For every $u \in L^2(I, \mathcal{M}(\Omega))$ *and every* $y \in L^2(I, C_0(\Omega))$ *and* $y_\sigma \in \mathcal{Y}_\sigma$ *we have*

$$\langle u, y_\sigma \rangle = \langle \Phi_\sigma u, y_\sigma \rangle, \tag{10.4.5}$$
$$\langle u, \Psi_\sigma y \rangle = \langle \Phi_\sigma u, y \rangle. \tag{10.4.6}$$

(iii) For every $u \in L^2(I, \mathcal{M}(\Omega))$ *and* $y \in L^2(I, C_0(\Omega))$ *we have*

$$\|\Phi_\sigma u\|_{L^2(\mathcal{M})} \leqslant \|u\|_{L^2(\mathcal{M})}, \tag{10.4.7}$$
$$\|\Psi_\sigma y\|_{L^2(C_0)} \leqslant \|y\|_{L^2(C_0)}. \tag{10.4.8}$$

(iv) For every $u \in L^2(I, \mathcal{M}(\Omega))$ *and* $y \in L^2(I, C_0(\Omega))$ *we have*

$$\Phi_\sigma u \overset{*}{\rightharpoonup} u \text{ in } L^2(I, \mathcal{M}(\Omega)) \text{ and } \|\Phi_\sigma u\|_{L^2(\mathcal{M})} \to \|u\|_{L^2(\mathcal{M})}, \tag{10.4.9}$$
$$\Psi_\sigma y \to y \text{ in } L^2(I, C_0(\Omega)). \tag{10.4.10}$$

Proof. The formulas of (10.4.4) follow from the linearity of the operators and the identities $\Phi_\sigma(\chi_l\delta_{x_i}) = \chi_l\delta_{x_i}$ and $\Psi_\sigma(\chi_l e_i) = \chi_l e_i$ for all $1 \leqslant l \leqslant N_\tau$ and $1 \leqslant i \leqslant N_h$.

Identity (10.4.5) is a consequence of (10.4.4) and (10.4.6). Let us prove the latter. First we observe that

$$\Phi_\sigma u = \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} u_{kj}\chi_k\delta_{x_j}, \quad \text{with } u_{kj} = \frac{1}{\tau_k}\int_{I_k}\langle u(t), e_j\rangle\,dt, \tag{10.4.11}$$

$$\Psi_\sigma y = \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} y_{kj}\chi_k e_j, \quad \text{with } y_{kj} = \frac{1}{\tau_k}\int_{I_k} y(x_j, t)\,dt. \tag{10.4.12}$$

From (10.4.11) and (10.4.12) we have

$$\begin{aligned}\langle \Phi_\sigma u, y\rangle &= \int_0^T \langle (\Phi_\sigma u)(t), y(t)\rangle\,dt = \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} u_{kj}\int_0^T \langle \chi_k\delta_{x_j}, y(t)\rangle\,dt\\ &= \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} u_{kj}\int_{I_k} y(x_j, t)\,dt\\ &= \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} \tau_k u_{kj} y_{kj}.\end{aligned}$$

Analogously we get

$$\begin{aligned}\langle u, \Psi_\sigma y\rangle &= \int_0^T \langle u(t), (\Psi_\sigma y)(t)\rangle\,dt = \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} y_{kj}\int_0^T \langle u(t), \chi_k e_j\rangle\,dt\\ &= \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} y_{kj}\int_{I_k} \langle u(t), e_j\rangle\,dt\\ &= \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} \tau_k u_{kj} y_{kj},\end{aligned}$$

as desired.

We turn to (10.4.7). First we recall that the norm of $\Phi_\sigma u$ is given by

$$\|\Phi_\sigma u\|_{L^2(\mathcal{M})} = \left(\sum_{k=1}^{N_\tau}\tau_k\left(\sum_{j=1}^{N_h}|u_{kj}|\right)^2\right)^{1/2}.$$

Next we define $y_\sigma \in \mathcal{Y}_\sigma$ by

$$y_{kj} = \left(\sum_{i=1}^{N_h}|u_{ki}|\right)\operatorname{sign}(u_{kj}),$$

where we set $\operatorname{sign}(0) = 0$. For y_σ we compute the expressions

$$\begin{aligned}(10.4.13)\qquad \langle u, y_\sigma\rangle &= \int_0^T \langle u(t), y_\sigma(t)\rangle \,dt = \sum_{k=1}^{N_\tau} \int_{I_k} \sum_{j=1}^{N_h} y_{kj}\langle u(t), e_j\rangle \,dt \\ &= \sum_{k=1}^{N_\tau} \tau_k \sum_{j=1}^{N_h} y_{kj} u_{kj} = \sum_{k=1}^{N_\tau} \tau_k \left(\sum_{j=1}^{N_h} |u_{kj}|\right)^2 \\ &= \|\Phi_\sigma u\|^2_{L^2(\mathcal{M})}\end{aligned}$$

and

$$\begin{aligned}(10.4.14)\qquad \|y_\sigma\|_{L^2(C_0)} &= \left(\int_0^T \|y_\sigma(t)\|^2_\infty \,dt\right)^{1/2} = \left(\sum_{k=1}^{N_\tau} \int_{I_k} \Big\| \sum_{j=1}^{N_h} y_{kj} e_j \Big\|^2_\infty \,dt\right)^{1/2} \\ &= \left(\sum_{k=1}^{N_\tau} \tau_k \left(\sum_{j=1}^{N_h} |u_{kj}|\right)^2\right)^{1/2} \\ &= \|\Phi_\sigma u\|_{L^2(\mathcal{M})}.\end{aligned}$$

From (10.4.13) and (10.4.14) we deduce

$$\|\Phi_\sigma u\|^2_{L^2(\mathcal{M})} = \langle u, y_\sigma\rangle \leqslant \|u\|_{L^2(\mathcal{M})}\|y_\sigma\|_{L^2(C_0)} = \|u\|_{L^2(\mathcal{M})}\|\Phi_\sigma u\|_{L^2(\mathcal{M})},$$

which implies (10.4.7).

To establish (10.4.8) we choose $y \in L^2(I, C_0(\Omega))$ and estimate

$$\begin{aligned}\|\Psi_\sigma y\|_{L^2(C_0)} &= \left(\sum_{k=1}^{N_\tau} \int_{I_k} \|(\Psi_\sigma y)(t)\|^2_\infty \,dt\right)^{1/2} \\ &= \left(\sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \Big\| \sum_{j=1}^{N_h} \left(\int_{I_k} y(x_j, t)\,dt\right) e_j \Big\|^2_\infty\right)^{1/2} \\ &\leqslant \left(\sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \left(\int_{I_k} \|y(t)\|_\infty \,dt\right)^2\right)^{1/2} \leqslant \left(\sum_{k=1}^{N_\tau} \int_{I_k} \|y(t)\|^2_\infty \,dt\right)^{1/2} \\ &= \|y\|_{L^2(C_0)}.\end{aligned}$$

Before proving (10.4.9), we will consider (10.4.10). It is well known that (10.4.10) holds for functions in $C^\infty(\overline{\omega}_T)$ vanishing on Σ_T. From the density of these functions in $L^2(I, C_0(\Omega))$ and from inequality (10.4.8) we deduce (10.4.10).

Finally, we prove (10.4.9). From (10.4.7) we know that $\{\Phi_\sigma u\}_\sigma$ is bounded in the space $L^2(I, \mathcal{M}(\Omega))$. Then, there exists a subsequence, denoted in the same way, and an element $\tilde{u} \in L^2(I, \mathcal{M}(\Omega))$ such that $\Phi_\sigma u \overset{*}{\rightharpoonup} \tilde{u}$ in $L^2(I, \mathcal{M}(\Omega))$. Then, for every $y \in L^2(I, C_0(\Omega))$ it holds that

$$\lim_{\sigma\to 0} \int_0^T \langle (\Phi_\sigma u)(t), y(t)\rangle \, dt = \int_0^T \langle \tilde{u}(t), y(t)\rangle \, dt.$$

Using (10.4.6) and (10.4.10) we find

$$\lim_{\sigma\to 0} \int_0^T \langle (\Phi_\sigma u)(t), y(t)\rangle \, dt = \lim_{\sigma\to 0} \int_0^T \langle u(t), (\Psi_\sigma y)(t)\rangle \, dt = \int_0^T \langle u(t), y(t)\rangle \, dt.$$

Combining these two equalities we have that

$$\int_0^T \langle \tilde{u}(t), y(t)\rangle \, dt = \int_0^T \langle u(t), y(t)\rangle \, dt \quad \forall y \in L^2(I, C_0(\Omega)),$$

therefore $u = \tilde{u}$ and the whole sequence $\{\Phi_\sigma u\}_\sigma$ converges weakly-$\star$ to u.

By the convergence $\Phi_\sigma u \overset{*}{\rightharpoonup} u$ and (10.4.7) we obtain

$$\|u\|_{L^2(\mathcal{M})} \leqslant \liminf_{\sigma\to 0} \|\Phi_\sigma u\|_{L^2(\mathcal{M})} \leqslant \limsup_{\sigma\to 0} \|\Phi_\sigma u\|_{L^2(\mathcal{M})} \leqslant \|u\|_{L^2(\mathcal{M})},$$

which concludes the proof of (10.4.9). □

We finish this section by proving the following approximation result.

Theorem 10.4.3. *Let y and y^σ be the solutions to (10.1.1) corresponding to u and $\Phi_\sigma u$, respectively. Then there exists a constant $C > 0$ independent of u and σ such that*

$$\|y - y^\sigma\|_{L^2(\Omega_T)} \leqslant C h^{2-\frac{n}{2}} \|u\|_{L^2(\mathcal{M})} \quad \forall u \in L^2(I, \mathcal{M}(\Omega)). \tag{10.4.15}$$

Proof. Let $f \in L^2(\Omega_T)$ be arbitrary and take $z \in \mathcal{Z}$ satisfying

$$\begin{cases} -\partial_t z - \Delta z &= f \quad \text{in } \Omega_T, \\ z &= 0 \quad \text{on } \Sigma_T, \\ z(x, T) &= 0 \quad \text{in } \Omega. \end{cases} \tag{10.4.16}$$

Due to the convexity of Ω, there exists a constant $\tilde{C}$ independent of f such that $\|z\|_{H^{2,1}(\Omega_T)} \leqslant \tilde{C}\|f\|_{L^2(\Omega_T)}$. By (10.2.1) and (10.4.6) we get

$$\begin{aligned} \int_{\Omega_T} (y - y^\sigma) f \, dx \, dt &= \int_0^T \langle u(t) - (\Phi_\sigma u)(t), z(t)\rangle \, dt \\ &= \int_0^T \langle u(t), z(t) - (\Psi_\sigma z)(t)\rangle \, dt \\ &\leqslant \|u\|_{L^2(\mathcal{M})} \|z - \Psi_\sigma z\|_{L^2(C_0)}. \end{aligned} \tag{10.4.17}$$

Now, we will prove that

$$\|z - \Psi_\sigma z\|_{L^2(C_0)} \leqslant Ch^{2-\frac{n}{2}} \|z\|_{H^{2,1}(\Omega_T)}. \tag{10.4.18}$$

From the error estimates of the interpolation in Sobolev spaces [Ciarlet 1978, Chapter 3] we get

$$\begin{aligned} \|z - \Pi_h z\|_{L^2(C_0)} &= \left(\int_0^T \|z(t) - \Pi_h z(t)\|_\infty^2 \, dt \right)^{1/2} \\ &\leqslant Ch^{2-\frac{n}{2}} \left(\int_0^T \|z(t)\|_{H^2(\Omega)}^2 \, dt \right)^{1/2} \\ &\leqslant Ch^{2-\frac{n}{2}} \|z\|_{H^{2,1}(\Omega_T)}. \end{aligned} \tag{10.4.19}$$

Here and below C denotes a constant independent of σ. By an inverse inequality (see [Ciarlet 1978, Theorem 17.2]) and using (10.4.2) for the last inequality in the following estimate we obtain

(10.4.20)

$$\begin{aligned} \|\Pi_h z - \Psi_\sigma z\|_{L^2(C_0)} &= \left(\sum_{k=1}^{N_\tau} \int_{I_k} \left\| \Pi_h z(t) - \frac{1}{\tau_k} \int_{I_k} \Pi_h z(s) \, ds \right\|_\infty^2 dt \right)^{1/2} \\ &\leqslant \left(\sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \int_{I_k} \int_{I_k} \|\Pi_h z(t) - \Pi_h z(s)\|_\infty^2 \, ds \, dt \right)^{1/2} \\ &\leqslant \frac{C}{h^{n/2}\sqrt{\tau}} \left(\sum_{k=1}^{N_\tau} \int_{I_k} \int_{I_k} \|\Pi_h z(t) - \Pi_h z(s)\|_{L^2(\Omega)}^2 \, ds \, dt \right)^{1/2} \\ &\leqslant \frac{C}{h^{n/2}\sqrt{\tau}} \left(\sum_{k=1}^{N_\tau} \int_{I_k} \int_{I_k} \|\Pi_h z(t) - z(t)\|_{L^2(\Omega)}^2 \, ds \, dt \right)^{1/2} \\ &\quad + \frac{C}{h^{n/2}\sqrt{\tau}} \left(\sum_{k=1}^{N_\tau} \int_{I_k} \int_{I_k} \|\Pi_h z(s) - z(s)\|_{L^2(\Omega)}^2 \, ds \, dt \right)^{1/2} \\ &\quad + \frac{C}{h^{n/2}\sqrt{\tau}} \left(\sum_{k=1}^{N_\tau} \int_{I_k} \int_{I_k} \|z(t) - z(s)\|_{L^2(\Omega)}^2 \, ds \, dt \right)^{1/2} \\ &\leqslant \frac{Ch^2}{h^{n/2}} \|z\|_{H^{2,1}(\Omega_T)} + \frac{C}{h^{n/2}\sqrt{\tau}} \left(\sum_{k=1}^{N_\tau} \int_{I_k} \int_{I_k} \| \int_{I_k} \partial_t z(\theta) \, d\theta \|_{L^2(\Omega)}^2 \, ds \, dt \right)^{1/2} \\ &\leqslant C \frac{h^2 + \tau}{h^{n/2}} \|z\|_{H^{2,1}(\Omega_T)} \leqslant Ch^{2-\frac{n}{2}} \|z\|_{H^{2,1}(\Omega_T)}. \end{aligned}$$

Inequality (10.4.18) follows from (10.4.19) and (10.4.20). Finally, (10.4.17) and (10.4.18) leads to

$$\int_{\Omega_T} (y - y^\sigma) f \, dx \, dt \leqslant Ch^{2-\frac{n}{2}} \|z\|_{H^{2,1}(\Omega_T)} \leqslant Ch^{2-\frac{n}{2}} \|f\|_{L^2(\Omega_T)} \quad \forall f \in L^2(\Omega_T),$$

which implies (10.4.15). □

10.4.2 DISCRETE STATE EQUATION

In this section we approximate the state equation and provide error estimates. We recall that I_k was defined as $(t_{k-1}, t_k]$ and consequently $y_{k,h} = y_\sigma(t_k) = y_\sigma|_{I_k}$, $1 \leqslant k \leqslant N_\tau$. To approximate the state equation in time we use a dG(0) discontinuous Galerkin method, which can be formulated as an implicit Euler time stepping scheme. Given a control $u \in L^2(I, \mathcal{M}(\Omega))$, for $k = 1, \dots, N_\tau$ and $z_h \in Y_h$ we set

$$\begin{cases} \left(\dfrac{y_{k,h} - y_{k-1,h}}{\tau_k}, z_h\right) + a(y_{k,h}, z_h) = \dfrac{1}{\tau_k}\displaystyle\int_{I_k} \langle u(t), z_h\rangle \, dt, \\ y_{0,h} = y_{0h}, \end{cases} \tag{10.4.21}$$

where $(\cdot,\cdot)$ denotes the scalar product in $L^2(\Omega)$, a is the bilinear form associated to the operator $-\Delta$, i.e.,

$$a(y, z) = \int_\Omega \nabla y \nabla z \, dx,$$

and y_{0h} is an element of Y_h satisfying for some $C_0 > 0$

$$\|y_0 - y_{0h}\|_{H^{-1}(\Omega)} \leqslant C_0 h \|y_0\|_{L^2(\Omega)}. \tag{10.4.22}$$

For instance we can choose for y_{0h} the projection $P_h y_0$ of y_0 on Y_h given by the variational equation

$$(P_h y_0, z_h) = (y_0, z_h) \ \ \forall z_h \in Y_h.$$

For any such choice of y_{0h}, the estimate (10.4.22) implies that there exists a constant $C_1 > 0$ independent of h such that

$$\|y_{0h}\|_{L^2(\Omega)} \leqslant C_1 \|y_0\|_{L^2(\Omega)}. \tag{10.4.23}$$

Indeed, by using an inverse inequality and the well known estimates for the projection operator $P_h : L^2(\Omega) \to Y_h$, we obtain

$$\begin{aligned} \|y_{0h}\|_{L^2(\Omega)} &\leqslant \|y_{0h} - P_h y_0\|_{L^2(\Omega)} + \|P_h y_0\|_{L^2(\Omega)} \leqslant \frac{C}{h}\|y_{0h} - P_h y_0\|_{H^{-1}(\Omega)} + \|y_0\|_{L^2(\Omega)} \\ &\leqslant \frac{C}{h}\left(\|y_{0h} - y_0\|_{H^{-1}(\Omega)} + \|y_0 - P_h y_0\|_{H^{-1}(\Omega)}\right) + \|y_0\|_{L^2(\Omega)} \\ &\leqslant (C+1)\|y_0\|_{L^2(\Omega)}. \end{aligned}$$

Obviously (10.4.21) defines a unique solution y_σ. Let us observe that from (10.4.5) we have the following important consequence.

Lemma 10.4.4. *Let y_σ and $\tilde{y}_\sigma$ denote the solutions to (10.4.21) associated to the controls u and $\Phi_\sigma u$, respectively. Then the identity $y_\sigma = \tilde{y}_\sigma$ holds.*

The rest of the section is devoted to the proof of the stability of the scheme (10.4.21) and to the derivation of error estimates for $\|y - y_\sigma\|_{L^2(\Omega_T)}$, where y and y_σ are the solutions to (10.1.1) and (10.4.21) associated to a given control $u \in L^2(I, \mathcal{M}(\Omega))$. To this end, we introduce some operators that will be used in the proof of the theorems. For every h we consider the Ritz projection $R_h : H_0^1(\Omega) \to Y_h$ given by

$$a(y_h, R_h z) = a(y_h, z) \quad \forall y_h \in Y_h.$$

From the theory of finite elements we know that for all $z \in H^2(\Omega) \cap H_0^1(\Omega)$,

$$\begin{cases} \|z - R_h z\|_{L^2(\Omega)} + h\|z - R_h z\|_{H^1(\Omega)} \leqslant Ch^2\|z\|_{H^2(\Omega)}, \\ \|z - R_h z\|_{L^\infty(\Omega)} \leqslant Ch^{2-\frac{n}{2}}\|z\|_{H^2(\Omega)}. \end{cases} \tag{10.4.24}$$

Now, for every $\sigma = (\tau, h)$ we define $\mathcal{R}_\sigma : L^2(I, H_0^1(\Omega)) \to \mathcal{Y}_\sigma$ by

$$\mathcal{R}_\sigma z = \sum_{k=1}^{N_\tau} \frac{1}{\tau_k} \int_{I_k} R_h z(t)\, dt \chi_k = \sum_{k=1}^{N_\tau} z_{k,h}\chi_k.$$

The operator $\mathcal{R}_\sigma$ enjoys for all $z \in L^2(I, H_0^1(\Omega))$ and $y_\sigma \in \mathcal{Y}_\sigma$ the property

$$\int_0^T a(y_\sigma(t), z(t) - \mathcal{R}_\sigma z(t))\, dt = \sum_{k=1}^{N_\tau} \int_{I_k} a(y_{k,h}, z(t) - z_{k,h})\, dt = 0. \tag{10.4.25}$$

Indeed, for every $k = 1, \dots, N_\tau$ we have

$$\begin{aligned} \int_{I_k} a(y_{k,h}, z(t))\, dt &= \int_{I_k} a(y_{k,h}, R_h z(t))\, dt \\ &= \tau_k a(y_{k,h}, \frac{1}{\tau_k} \int_{I_k} R_h z(t)\, dt) \\ &= \int_{I_k} a(y_{h,k}, z_{h,k})\, dt. \end{aligned}$$

Theorem 10.4.5. *Given a control $u \in L^2(I, \mathcal{M}(\Omega))$, let y_σ be the solution to (10.4.21) corresponding to u. Then, there exist constants $C_i > 0$, $i = 1, 2$, independent of u and σ such that*

$$\begin{aligned} \sum_{k=1}^{N_\tau} \|y_{k,h} - y_{k-1,h}\|_{L^2(\Omega)}^2 + \tau \max_{1 \leqslant k \leqslant N_\tau} \|\nabla y_{k,h}\|_{L^2(\Omega)}^2 \\ \leqslant C_1 \left(\|y_0\|_{L^2(\Omega)}^2 + \|u\|_{L^2(\mathcal{M})}^2 \right), \end{aligned} \tag{10.4.26}$$

$$\|y_\sigma\|_{L^2(\Omega_T)} \leqslant C_2 \left(\|y_0\|_{L^2(\Omega)} + \|u\|_{L^2(\mathcal{M})} \right). \tag{10.4.27}$$

Proof. Let us set $z_h = y_{k,h} - y_{k-1,h}$ in (10.4.21). Then we obtain for $1 \leqslant k \leqslant N_\tau$ that

$$\frac{1}{\tau_k}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} + a(y_{k,h}, y_{k,h} - y_{k-1,h}) = \frac{1}{\tau_k}\int_{I_k} \langle u(t), y_{k,h} - y_{k-1,h}\rangle \, dt.$$

From here we get with the aid of an inverse estimate [Ciarlet 1978, Theorem 17.2]

$$\begin{aligned}
\frac{1}{\tau}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} &+ \frac{1}{2}[a(y_{k,h}, y_{k,h}) - a(y_{k-1,h}, y_{k-1,h})] \\
&\leqslant \frac{1}{\tau_k}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} + \frac{1}{2}[a(y_{k,h}, y_{k,h}) - a(y_{k-1,h}, y_{k-1,h}) \\
&\quad + a(y_{k,h} - y_{k-1,h}, y_{k,h} - y_{k-1,h})] \\
&= \frac{1}{\tau_k}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} + a(y_{k,h}, y_{k,h} - y_{k-1,h}) \\
&= \frac{1}{\tau_k}\int_{I_k} \langle u(t), y_{k,h} - y_{k-1,h}\rangle \, dt \\
&\leqslant \frac{1}{\sqrt{\tau_k}}\|u\|_{L^2(I_k,\mathcal{M})}\|y_{k,h} - y_{k-1,h}\|_\infty \\
&\leqslant \frac{Ch^{-n/2}}{\sqrt{\tau_k}}\|u\|_{L^2(I_k,\mathcal{M})}\|y_{k,h} - y_{k-1,h}\|_{L^2(\Omega)} \\
&\leqslant \frac{C^2h^{-n}\tau}{2\tau_k}\|u\|^2_{L^2(I_k,\mathcal{M})} + \frac{1}{2\tau}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} \\
&\leqslant \frac{C^2\rho_T C_{\Omega,T}}{2\tau}\|u\|^2_{L^2(I_k,\mathcal{M})} + \frac{1}{2\tau}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)}.
\end{aligned}$$

In the last inequality we have used (10.4.2). Summing from $k = 1$ to m and using (10.4.2), it follows that

$$\frac{1}{\tau}\sum_{k=1}^{m}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} + a(y_{m,h}, y_{m,h}) - a(y_{0h}, y_{0h}) \leqslant \frac{C^2\rho_T C_{\Omega,T}}{\tau}\|u\|^2_{L^2(\mathcal{M})}.$$

Hence

$$\sum_{k=1}^{m}\|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega)} + \tau\|\nabla y_{m,h}\|^2_{L^2(\Omega)} \leqslant C(\|y_0\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\mathcal{M})}). \tag{10.4.28}$$

Here we have used an inverse inequality, (10.4.2), and (10.4.23) to get

$$\tau\|y_{0h}\|^2_{H^1(\Omega)} \leqslant \frac{C\tau}{h^2}\|y_{0h}\|^2_{L^2(\Omega)} \leqslant C\|y_0\|^2_{L^2(\Omega)}.$$

Finally, since $1 \leqslant m \leqslant N_\tau$ is arbitrary, (10.4.26) follows from (10.4.28).

Now we prove (10.4.27). Given $f \in L^2(\Omega_T)$, we take $z \in \mathcal{Z}$ satisfying (10.4.16). Integrating by

parts we get

$$\begin{aligned}\int_{\Omega_T} y_\sigma f\,dxdt &= \sum_{k=1}^{N_\tau}\int_{I_k}\int_\Omega y_{k,h}(x)f(x,t)\,dxdt\\ &= \sum_{k=1}^{N_\tau}\int_{I_k}\{-\partial_t(y_{k,h},z(t)) + a(y_{k,h},z(t))\}\,dt\\ &= \sum_{k=1}^{N_\tau}\left\{(y_{k,h},z(t_{k-1})-z(t_k)) + \int_{I_k} a(y_{k,h},z(t))\,dt\right\}\\ &= \sum_{k=1}^{N_\tau}\left\{(y_{k,h}-y_{k-1,h},z(t_{k-1})) + \int_{I_k} a(y_{k,h},z(t))\,dt\right\} + (y_{0h},z(0)).\end{aligned}$$

Taking $z_\sigma = \mathcal{R}_\sigma z$, we get from the above identity and (10.4.25) that

$$\begin{aligned}\int_{\Omega_T} y_\sigma f\,dxdt &= \sum_{k=1}^{N_\tau}\{(y_{k,h}-y_{k-1,h},z_{k,h}) + \tau_k a(y_{k,h},z_{k,h})\} + (y_{0h},z(0))\\ &\quad + \sum_{k=1}^{N_\tau}\left\{(y_{k,h}-y_{k-1,h},z(t_{k-1})-z_{k,h}) + \int_{I_k} a(y_{k,h},z(t)-z_{k,h})\,dt\right\}\\ &= \int_0^T \langle u(t),z_\sigma(t)\rangle\,dt + (y_{0h},z(0)) + \sum_{k=1}^{N_\tau}(y_{k,h}-y_{k-1,h},z(t_{k-1})-z_{k,h}).\end{aligned} \tag{10.4.29}$$

Let us estimate each of these terms. From the definition of z_σ and (10.4.23) we obtain

$$\begin{aligned}\int_0^T \langle u(t),z_\sigma(t)\rangle\,dt + (y_{0h},z(0)) &\leqslant \|u\|_{L^2(\mathcal{M})}\|z_\sigma\|_{L^2(C_0)} + \|y_{0h}\|_{L^2(\Omega)}\|z(0)\|_{L^2(\Omega)}\\ &\leqslant C\|z\|_{H^{2,1}(\Omega_T)}(\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}),\end{aligned} \tag{10.4.30}$$

where we have used that there exists a constant $C > 0$ independent of σ such that

$$\|\mathcal{R}_\sigma v\|_{L^2(C_0)} \leqslant C\|v\|_{H^{2,1}(\Omega_T)} \quad \forall v \in H^{2,1}(\Omega_T). \tag{10.4.31}$$

Indeed,

$$\begin{aligned}\|\mathcal{R}_\sigma v\|_{L^2(C_0)} &= \left(\sum_{k=1}^{N_\tau}\int_{I_k}\|\mathcal{R}_\sigma v(t)\|_\infty^2\,dt\right)^{1/2} = \left(\sum_{k=1}^{N_\tau}\int_{I_k}\Big\|\frac{1}{\tau_k}\int_{I_k} R_h v(s)\,ds\Big\|_\infty^2\,dt\right)^{1/2}\\ &\leqslant \left(\sum_{k=1}^{N_\tau}\int_{I_k}\|R_h v(s)\|_\infty^2\,ds\right)^{1/2}.\end{aligned}$$

Using (10.4.24) we deduce that

$$\|R_h w\|_\infty \leqslant \|R_h w - w\|_\infty + \|w\|_\infty \leqslant Ch^\kappa\|w\|_{H^2(\Omega)} + \|w\|_\infty \leqslant C\|w\|_{H^2(\Omega)}$$

for every $w \in H^2(\Omega) \cap H_0^1(\Omega)$, with $\kappa = 1$ if $n \leqslant 2$ and $\kappa = 1/2$ if $n = 3$. Then, (10.4.31) follows from the above inequalities.

Concerning the last term of (10.4.29), we will prove

$$\sum_{k=1}^{N_\tau} (y_{k,h} - y_{k-1,h}, z(t_{k-1}) - z_{k,h}) \leqslant Ch^\kappa \|z\|_{H^{2,1}(\Omega_T)} (\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}), \tag{10.4.32}$$

where κ is defined as above. First we observe that (10.4.26) implies

$$\begin{aligned}\sum_{k=1}^{N_\tau} |(y_{k,h} - y_{k-1,h}, z(t_{k-1}) - z_{k,h})| &\\ \leqslant \left(\sum_{k=1}^{N_\tau} \|y_{k,h} - y_{k-1,h}\|^2_{L^2(\Omega_h)}\right)^{1/2} &\left(\sum_{k=1}^{N_\tau} \|z(t_{k-1}) - z_{k,h}\|^2_{L^2(\Omega_h)}\right)^{1/2} \\ \leqslant C(\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}) &\left(\sum_{k=1}^{N_\tau} \|z(t_{k-1}) - z_{k,h}\|^2_{L^2(\Omega_h)}\right)^{1/2}.\end{aligned} \tag{10.4.33}$$

From the definition of z_σ and (10.4.24) we deduce

$$\begin{aligned}\|z(t_{k-1}) - z_{k,h}\|_{L^2(\Omega_h)} &= \left(\int_{\Omega_h} \left|\frac{1}{\tau_k}\int_{I_k} \{z(t_{k-1}) - R_h z(s)\}\,ds\right|^2 dx\right)^{1/2} \\ &\leqslant \left(\frac{1}{\tau_k}\int_{\Omega_h}\int_{I_k} |z(t_{k-1}) - R_h z(s)|^2\,ds dx\right)^{1/2} \\ &\leqslant \left(\frac{1}{\tau_k}\int_{\Omega_h}\int_{I_k} |z(t_{k-1}) - z(s)|^2\,ds dx\right)^{1/2} \\ &\quad + \left(\frac{1}{\tau_k}\int_{I_k} \|z(s) - R_h z(s)\|^2_{L^2(\Omega_h)}\,ds\right)^{1/2} \\ &\leqslant \left(\int_{\Omega_h}\int_{I_k}\int_{I_k} |\partial_t z(\theta)|^2\,d\theta\,ds\,dx\right)^{1/2} \\ &\quad + Ch^2\left(\frac{1}{\tau_k}\int_{I_k} \|z(s)\|^2_{H^2(\Omega)}\,ds\right)^{1/2} \\ &\leqslant \sqrt{\tau}\|\partial_t z\|_{L^2(I_k,L^2(\Omega))} + \frac{Ch^2\sqrt{\rho_T}}{\sqrt{\tau}}\|z\|_{L^2(I_k,H^2(\Omega))} \\ &\leqslant Ch^\kappa(\|\partial_t z\|_{L^2(I_k,L^2(\Omega))} + \|z\|_{L^2(I_k,H^2(\Omega))}).\end{aligned}$$

Inserting this estimate in (10.4.33) we infer (10.4.32). Finally, (10.4.29), (10.4.30) and (10.4.32) imply that

$$\int_{\Omega_T} y_\sigma f\,dxdt \leqslant C\|f\|_{L^2(\Omega_T)}(\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}) \quad \forall f \in L^2(\Omega_T),$$

which is equivalent to (10.4.27) □

In the next theorem we show error estimates for the discretization of the state equation.

Theorem 10.4.6. *Given* $u \in L^2(I, \mathcal{M}(\Omega))$, *let* y *and* y_σ *be the solutions to* (10.1.1) *and* (10.4.21). *Then, there exists a constant* C *independent of* $u \in L^2(I, \mathcal{M}(\Omega))$, $y_0 \in L^2(\Omega)$, *and* σ *such that*

$$\|y - y_\sigma\|_{L^2(\Omega_T)} \leqslant Ch^\kappa(\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}), \tag{10.4.34}$$

where $\kappa = 1$ *if* $n \leqslant 2$ *and* $\kappa = 1/2$ *if* $n = 3$.

Proof. As in the proof of Theorem 10.4.5, we take an arbitrary element $f \in L^2(\Omega_T)$, $z \in \mathcal{Z}$ solution to (10.4.16), and $z_\sigma = \mathcal{R}_\sigma z$. Then, from (10.2.1) we obtain

$$\begin{aligned}\int_{\Omega_T} (y - y_\sigma) f \,dx\,dt = \int_0^T \langle u(t), z(t)\rangle \,dt + \int_\Omega y_0(x) z(x,0)\,dx \\ - \sum_{k=1}^{N_\tau} \int_{I_k} \{-(y_{k,h}, \partial_t z(t)) + a(y_{k,h}, z(t))\}\,dt.\end{aligned} \tag{10.4.35}$$

Integrating by parts we get

$$\begin{aligned}\sum_{k=1}^{N_\tau} \int_{I_k} -(y_{k,h}, \partial_t z(t))\,dt &= \sum_{k=1}^{N_\tau} (y_{k,h}, z(t_{k-1}) - z(t_k)) \\ &= \sum_{k=1}^{N_\tau} (y_{k,h} - y_{k-1,h}, z(t_{k-1})) + (y_{0h}, z(0)).\end{aligned}$$

From this identity, (10.4.21), and (10.4.25) we deduce

$$\begin{aligned}&\sum_{k=1}^{N_\tau} \int_{I_k} \{-(y_{k,h}, \partial_t z(t)) + a(y_{k,h}, z(t))\}\,dt \\ &\qquad = \sum_{k=1}^{N_\tau} \int_{I_k} \left\{(y_{k,h} - y_{k-1,h}, z(t_{k-1})) + \int_{I_k} a(y_{k,h}, z(t))\right\} dt + (y_{0h}, z(0)) \\ &\qquad = \sum_{k=1}^{N_\tau} \int_{I_k} \left\{(y_{k,h} - y_{k-1,h}, z_{k,h}) + \int_{I_k} a(y_{k,h}, z_{k,h})\right\} dt \\ &\qquad\quad + \sum_{k=1}^{N_\tau} \int_{I_k} (y_{k,h} - y_{k-1,h}, z(t_{k-1}) - z_{k,h}) + (y_{0h}, z(0)) \\ &\qquad = \int_0^T \langle u(t), z_\sigma(t)\rangle \,dt + \int_\Omega y_{0h}(x) z(x,0)\,dx \\ &\qquad\quad + \sum_{k=1}^{N_\tau} \int_{I_k} (y_{k,h} - y_{k-1,h}, z(t_{k-1}) - z_{k,h}).\end{aligned}$$

Inserting this identity in (10.4.35) we infer

(10.4.36)
$$\int_{\Omega_T}(y-y_\sigma)f\,dx\,dt=\int_0^T\langle u(t),z(t)-\mathcal{R}_\sigma z(t)\rangle\,dt+\int_\Omega(y_0(x)-y_{0h}(x))z(x,0)\,dx \\ -\sum_{k=1}^{N_\tau}\int_{I_k}(y_{k,h}-y_{k-1,h},z(t_{k-1})-z_{k,h}).$$

Let us estimate each of these three terms. For the first term we observe that

$$\|z-\mathcal{R}_\sigma z\|_{L^2(C_0)}\leqslant Ch^\kappa\|z\|_{H^{2,1}(\Omega_T)}.$$

The proof of this inequality is the same than the one of (10.4.18); it is enough to replace Π_h by R_h and to use (10.4.24). Using this inequality we obtain the first estimate as follows:

(10.4.37)
$$\left|\int_0^T\langle u(t),z(t)-\mathcal{R}_\sigma z(t)\rangle\,dt\right|\leqslant\|u\|_{L^2(\mathcal{M})}\|z-\mathcal{R}_\sigma z\|_{L^2(C_0)} \\ \leqslant ch^\kappa\|u\|_{L^2(\mathcal{M})}\|z\|_{H^{2,1}(\Omega_T)}.$$

For the second term we proceed with the aid of (10.4.23):

(10.4.38)
$$\left|\int_\Omega(y_0(x)-y_{0h}(x))z(x,0)\,dx\right|\leqslant\|y_0-y_{0h}\|_{H^{-1}(\Omega)}\|z(0)\|_{H_0^1(\Omega)} \\ \leqslant Ch\|y_0\|_{L^2(\Omega)}\|z\|_{H^{2,1}(\Omega_T)}.$$

Finally, the third term of (10.4.36) was estimated in (10.4.32). Thus, using (10.4.37), (10.4.38), and (10.4.32) in (10.4.36) the inequality

$$\int_{\Omega_T}(y-y_\sigma)f\,dx\,dt\leqslant Ch^\kappa(\|u\|_{L^2(\mathcal{M})}+\|y_0\|_{L^2(\Omega)})\|z\|_{H^{2,1}(\Omega_T)} \\ \leqslant Ch^\kappa(\|u\|_{L^2(\mathcal{M})}+\|y_0\|_{L^2(\Omega)})\|f\|_{L^2(\Omega_T)}$$

is obtained, which leads to (10.4.34) □

10.4.3 DISCRETE OPTIMAL CONTROL PROBLEM

The approximation of the optimal control problem (P) is defined as

(P_σ)
$$\min_{u\in L^2(I,\mathcal{M}(\Omega))}J_\sigma(u)=\frac{1}{2}\|y_\sigma-y_d\|^2_{L^2(\Omega_{hT})}+\alpha\|u\|_{L^2(\mathcal{M})},$$

where y_σ is the discrete state associated to u, i.e., the solution to (10.4.21). Let us observe that analogously to J, the functional J_σ is convex. However, it is not strictly convex due to the

non-injectivity of the control-to-discrete-state mapping and the non-strict convexity of the norm of $L^2(I,\mathcal{M}(\Omega))$. Although the existence of a solution can be shown in the same way as for the problem (P), we therefore cannot deduce its uniqueness. On the other hand, if $\tilde{u}_\sigma$ is a solution to (P_σ) and if we take $\bar{u}_\sigma = \Phi_\sigma\tilde{u}_\sigma$, then Lemma 10.4.4 and the inequality (10.4.7) imply that $J_\sigma(\bar{u}_\sigma) \leqslant J_\sigma(\tilde{u}_\sigma)$, hence $\bar{u}_\sigma$ is also a solution to (P_σ). Since for $u_\sigma \in \mathcal{U}_\sigma$, the mapping $u_\sigma \mapsto y_\sigma(u_\sigma)$, with $y_\sigma(u_\sigma)$ the solution to (10.4.21) for $u = u_\sigma$, is linear, injective and $\dim \mathcal{U}_\sigma = \dim \mathcal{Y}_\sigma$, this mapping is bijective. Therefore, the cost functional J_σ is strictly convex on $\mathcal{U}_\sigma$, hence (P_σ) has a unique solution in $\mathcal{U}_\sigma$, which will be denoted by $\bar{u}_\sigma$ hereafter. We summarize this discussion in the following theorem.

Theorem 10.4.7. *Problem* (P_σ) *admits at least one solution. Among all solutions, there exists a unique solution* $\bar{u}_\sigma$ *belonging to* $\mathcal{U}_\sigma$. *Moreover, any other solution* $\tilde{u} \in L^2(I,\mathcal{M}(\Omega))$ *to* (P_σ) *satisfies* $\Phi_\sigma\tilde{u} = \bar{u}_\sigma$.

Remark 10.4.8. The fact that problem (P_σ) has exactly one solution in $\mathcal{U}_\sigma$ is of practical interest. Indeed, recall that $\bar{u}_\sigma$, as an element of $\mathcal{U}_\sigma$, can be uniquely represented as

$$\bar{u}_\sigma = \sum_{k=1}^{N_\tau}\sum_{j=1}^{N_h} \bar{u}_{kj}\chi_k\delta_{x_j}.$$

The numerical computation of $\bar{u}_\sigma$ therefore is equivalent to the computation of the coefficients $\{\bar{u}_{kj} : 1 \leqslant k \leqslant N_\tau,\ 1 \leqslant j \leqslant N_h\}$; see section 10.6.

We finish this section by analyzing the convergence of the solution in $\mathcal{U}_\sigma$ to (P_σ) to the solution to (P).

Theorem 10.4.9. *For every* σ, *let* $\bar{u}_\sigma$ *be the unique solution to problem* (P_σ) *belonging to* $\mathcal{U}_\sigma$ *and let* $\bar{u}$ *be the solution to problem* (P). *Then the following convergence properties hold for* $\sigma \to 0^+$:

$$\bar{u}_\sigma \overset{*}{\rightharpoonup} \bar{u} \ \text{ in } L^2(I,\mathcal{M}(\Omega)), \tag{10.4.39}$$

$$\|\bar{u}_\sigma\|_{L^2(\mathcal{M})} \to \|\bar{u}\|_{L^2(\mathcal{M})}, \tag{10.4.40}$$

$$\|\bar{y} - \bar{y}_\sigma\|_{L^2(\Omega_T)} \to 0, \tag{10.4.41}$$

$$J_\sigma(\bar{u}_\sigma) \to J(\bar{u}), \tag{10.4.42}$$

where $\bar{y}$ *and* $\bar{y}_\sigma$ *are the continuous and discrete states associated to* $\bar{u}$ *and* $\bar{u}_\sigma$, *respectively.*

Proof. First of all, let us show that

$$u_\sigma \overset{*}{\rightharpoonup} u \text{ in } L^2(I,\mathcal{M}(\Omega)) \quad \text{implies} \quad \|y_\sigma - y\|_{L^2(\Omega_T)} \to 0, \tag{10.4.43}$$

where y_σ and y are the discrete and continuous states associated to the controls u_σ and u, respectively. Indeed, let us write $y - y_\sigma = (y - y^\sigma) + (y^\sigma - y_\sigma)$, where y^σ is the continuous state associated to u_σ. Then by Theorems 10.2.4 and 10.4.6 we deduce (10.4.43).

Turning to the verification of (10.4.39), we observe that

$$\alpha\|\bar{u}_\sigma\|_{L^2(\mathcal{M})} \leqslant J_\sigma(\bar{u}_\sigma) \leqslant J_\sigma(0) = \frac{1}{2}\|\hat{y}_{\sigma 0} - y_d\|^2_{L^2(\Omega_{hT})} \leqslant \frac{1}{2}\|\hat{y}_{\sigma 0} - y_d\|^2_{L^2(\Omega_T)}$$

with $\hat{y}_{\sigma 0}$ denoting the uncontrolled discrete state, which implies the boundedness of $\{\bar{u}_\sigma\}_\sigma$ in $L^2(I, \mathcal{M}(\Omega))$. By taking a subsequence, we have that $\bar{u}_\sigma \overset{*}{\rightharpoonup} u$ in $L^2(I, \mathcal{M}(\Omega))$. Then using (10.4.1), (10.4.43), lower semicontinuity of the norm $\|\cdot\|_{L^2(\mathcal{M})}$ and (10.4.9) we obtain

$$J(u) \leqslant \liminf_{\sigma\to 0} J_\sigma(\bar{u}_\sigma) \leqslant \limsup_{\sigma\to 0} J_\sigma(\bar{u}_\sigma) \leqslant \limsup_{\sigma\to 0} J_\sigma(\Psi_\sigma \bar{u}) = J(\bar{u}).$$

Hence $u = \bar{u}$ by the uniqueness of the solution to (P), and the whole sequence $\{\bar{u}_\sigma\}_\sigma$ converges weakly-$\star$ to $\bar{u}$. In addition, the above inequality implies (10.4.42). Using again (10.4.43), we deduce (10.4.41). Finally, (10.4.40) follows immediately from (10.4.41) and (10.4.42). □

10.5 ERROR ESTIMATES

We now turn to the proof of error estimates for the optimal costs and for the optimal states. We still require Ω to be convex and assume in addition

$$y_d \in L^2(I, L^r(\Omega)) \text{ with } r = \begin{cases} 2 & \text{if } n = 1, \\ 4 & \text{if } n = 2, \\ \frac{8}{3} & \text{if } n = 3. \end{cases} \tag{10.5.1}$$

Recall that $\bar{y}$ and $\bar{y}_\sigma$ denote the continuous and discrete states associated to the optimal controls $\bar{u}$ and $\bar{u}_\sigma$, respectively.

Theorem 10.5.1. *There exists a constant $C > 0$ independent of σ such that*

$$|J(\bar{u}) - J_\sigma(\bar{u}_\sigma)| \leqslant Ch^\kappa, \tag{10.5.2}$$

where $\kappa = 1$ if $n \leqslant 2$ and $\kappa = 1/2$ if $n = 3$.

Proof. Taking r as in (10.5.1) and using Hölder's inequality and (10.4.1), we deduce that for all $\varphi \in L^2(I, L^r(\Omega))$ and $n = 2$ or 3,

$$\|\varphi\|_{L^2(I,L^2(\Omega\setminus\Omega_h))} \leqslant \|\varphi\|_{L^2(I,L^r(\Omega\setminus\Omega_h))}|\Omega \setminus \Omega_h|^{\frac{r-2}{2r}} \leqslant C\|\varphi\|_{L^2(I,L^r(\Omega\setminus\Omega_h))}h^{\frac{\kappa}{2}} \tag{10.5.3}$$

holds. Observe that $\Omega = \Omega_h$ for $n = 1$; consequently (10.5.3) holds with $C = 0$.

Let y and y_σ be the continuous and discrete states associated to a given control u. As a consequence of (10.4.34) and (10.5.3), with $\varphi = y - y_d$, we obtain

$$\begin{aligned} \left|\|y - y_d\|^2_{L^2(\Omega_T)} - \|y_\sigma - y_d\|^2_{L^2(\Omega_{hT})}\right| &\leqslant \|y - y_d\|^2_{L^2(I,L^2(\Omega\setminus\Omega_h))} \\ &\quad + \left(\|y - y_d\|_{L^2(\Omega_{hT})} + \|y_\sigma - y_d\|_{L^2(\Omega_{hT})}\right)\|y - y_\sigma\|_{L^2(\Omega_{hT})} \\ &\leqslant C\left(\|y - y_d\|^2_{L^2(I,L^r(\Omega\setminus\Omega_h))} + \|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}\right)h^\kappa. \end{aligned} \tag{10.5.4}$$

Now, by the optimality of $\bar{u}$ and $\bar{u}_\sigma$ we have

$$J(\bar{u}) - J_\sigma(\bar{u}) \leqslant J(\bar{u}) - J_\sigma(\bar{u}_\sigma) \leqslant J(\bar{u}_\sigma) - J_\sigma(\bar{u}_\sigma),$$

and hence

$$|J(\bar{u}) - J_\sigma(\bar{u}_\sigma)| \leqslant \max\{|J(\bar{u}) - J_\sigma(\bar{u})|, |J(\bar{u}_\sigma) - J_\sigma(\bar{u}_\sigma)|\}. \tag{10.5.5}$$

From (10.4.40) we deduce that $\{\bar{u}_\sigma\}_\sigma$ is bounded in $L^2(I, \mathcal{M}(\Omega))$. Therefore, (10.2.2) implies that the continuous associated states $\{y_{\bar{u}_\sigma}\}_\sigma$ are bounded in $L^2(I, W_0^{1,p}(\Omega))$ for every $1 \leqslant p < \frac{n}{n-1}$, and therefore in $L^2(I, L^r(\Omega))$ as well. We now apply (10.5.4) with $u = \bar{u}_\sigma$ and $u = \bar{u}$, respectively. Together with (10.5.5) this establishes (10.5.2). □

In the following theorem we establish a rate of convergence for the states.

Theorem 10.5.2. *There exists a constant $C > 0$ independent of h such that*

$$\|\bar{y} - \bar{y}_\sigma\|_{L^2(\Omega_T)} \leqslant C h^{\frac{\kappa}{2}}, \tag{10.5.6}$$

with κ as defined in Theorem 10.4.1.

Proof. Let $S : L^2(I, \mathcal{M}(\Omega)) \to L^2(\Omega_T)$ and $S_\sigma : L^2(I, \mathcal{M}(\Omega)) \to L^2(\Omega_T)$ be the solution operators associated to the equations (10.1.1) and (10.4.21), respectively. From (10.4.34) it follows that

$$\|Su - S_\sigma u\|_{L^2(\Omega_T)} \leqslant C h^\kappa (\|u\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)}). \tag{10.5.7}$$

By the optimality of $\bar{u}$ we have for all $u \in L^2(I, \mathcal{M}(\Omega))$ that

$$(S\bar{u} - y_d, Su - S\bar{u}) + \alpha[\|u\|_{L^2(\mathcal{M})} - \|\bar{u}\|_{L^2(\mathcal{M})}] \geqslant 0,$$

where $(\cdot, \cdot)$ now denotes the scalar product in $L^2(\Omega_T)$. In particular, taking $u = \bar{u}_\sigma$, we get

$$(S\bar{u} - y_d, S\bar{u}_\sigma - S\bar{u}) + \alpha[\|\bar{u}_\sigma\|_{L^2(\mathcal{M})} - \|\bar{u}\|_{L^2(\mathcal{M})}] \geqslant 0. \tag{10.5.8}$$

Analogously, the optimality of $\bar{u}_\sigma$ implies that

$$(S_\sigma\bar{u}_\sigma - y_d, S_\sigma\bar{u} - S_\sigma\bar{u}_\sigma) + \alpha[\|\bar{u}\|_{L^2(\mathcal{M})} - \|\bar{u}_\sigma\|_{L^2(\mathcal{M})}] \geqslant 0. \tag{10.5.9}$$

We point out that by definition of Y_h, we have $S_\sigma u = 0$ in $I \times (\Omega \setminus \Omega_h)$. Then, the scalar product above in $L^2(\Omega_T)$ coincides with that in $L^2(\Omega_{hT})$. Now, we rearrange terms in (10.5.9) as follows:

$$\begin{aligned}(S\bar{u}_\sigma - y_d, S\bar{u} - S\bar{u}_\sigma) + (S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma, S_\sigma\bar{u} - S_\sigma\bar{u}_\sigma) \\ + (y_d, S\bar{u} - S_\sigma\bar{u} + S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma) + (S\bar{u}_\sigma, S_\sigma\bar{u} - S\bar{u} + S\bar{u}_\sigma - S_\sigma\bar{u}_\sigma) \\ + \alpha[\|\bar{u}\|_{L^2(\mathcal{M})} - \|\bar{u}_\sigma\|_{L^2(\mathcal{M})}] \geqslant 0.\end{aligned} \tag{10.5.10}$$

Adding (10.5.8) and (10.5.10) we obtain

$$\begin{aligned}(10.5.11)\qquad \|S\bar{u}-S_\sigma\bar{u}_\sigma\|^2_{L^2(\Omega_T)} &= (S\bar{u}-S_\sigma\bar{u}_\sigma, S\bar{u}-S_\sigma\bar{u}_\sigma)\\ &\leqslant (S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma, S_\sigma\bar{u}-S_\sigma\bar{u}_\sigma)\\ &\quad + (y_d - S\bar{u}_\sigma, S\bar{u}-S_\sigma\bar{u}+S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma).\end{aligned}$$

Let us estimate the right hand terms. For the first one we apply the Cauchy–Schwarz inequality and use (10.5.7) to deduce

$$(10.5.12)\quad (S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma, S_\sigma\bar{u}-S_\sigma\bar{u}_\sigma) \leqslant \|S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma\|_{L^2(\Omega_T)}\|S_\sigma\bar{u}-S_\sigma\bar{u}_\sigma\|_{L^2(\Omega_T)} \leqslant Ch^\kappa,$$

where we have used that $\{\bar{u}_\sigma\}_\sigma, \{S_\sigma\bar{u}\}_\sigma$ and $\{S_\sigma\bar{u}_\sigma\}_\sigma$ are bounded due to (10.4.40) and (10.4.27). For the second term we use once again (10.5.7) to obtain

$$\begin{aligned}(10.5.13)\quad &(y_d - S\bar{u}_\sigma, S\bar{u}-S_\sigma\bar{u}+S_\sigma\bar{u}_\sigma - S\bar{u}_\sigma)\\ &\quad + \|y_d - S\bar{u}_\sigma\|_{L^2(\Omega_T)}\|(S-S_\sigma)(\bar{u}-\bar{u}_\sigma)\|_{L^2(\Omega_T)}\\ &\quad + C(\|\bar{u}-\bar{u}_\sigma\|_{L^2(\mathcal{M})} + \|y_0\|_{L^2(\Omega)})h^\kappa \leqslant Ch^\kappa,\end{aligned}$$

where we have also used that $y_d \in L^2(I, L^r(\Omega))$ and (10.2.2). Finally, (10.5.11), (10.5.12) and (10.5.13) prove (10.5.6). □

Remark 10.5.3. Let us observe that (10.5.2) and (10.5.6) imply that

$$\left|\|\bar{u}\|_{L^2(\mathcal{M})} - \|\bar{u}_\sigma\|_{L^2(\mathcal{M})}\right| \leqslant Ch^{\frac{\kappa}{2}}$$

for some constant $C > 0$ independent of σ.

10.6 NUMERICAL SOLUTION

We now address the computation of minimizers $\bar{u}_\sigma$ of problem (P_σ). First of all, we note that if we define $y_{d,\sigma}$ as the $L^2(\Omega_{hT})$ projection of y_d on $\mathcal{Y}_\sigma$, then

$$J_\sigma(u) = \frac{1}{2}\|y_\sigma - y_{d,\sigma}\|^2_{L^2(\Omega_{hT})} + \alpha\|u\|_{L^2(\mathcal{M})} + \frac{1}{2}\|y_d - y_{d,\sigma}\|^2_{L^2(\Omega_{hT})}.$$

Therefore, the problems (P_σ) and

$$(Q_\sigma)\qquad \min_{u\in L^2(I,\mathcal{M}(\Omega))} \tilde{J}_\sigma(u) = \frac{1}{2}\|y_\sigma - y_{d,\sigma}\|^2_{L^2(\Omega_{hT})} + \alpha\|u\|_{L^2(\mathcal{M})}$$

are equivalent. In this section we present a numerical algorithm to solve (Q_σ) as an alternative formulation to (P_σ).

Due to the spatio-temporal coupling of the norm in $L^2(I, \mathcal{M}(\Omega))$, its subdifferential is difficult to characterize. However, using Fenchel duality combined with an equivalent reformulation

that decouples the spatio-temporal structure, we can obtain optimality conditions that can be solved using a semismooth Newton method.

For the reader's convenience, we recall the Fenchel duality theory, e.g., from [Ekeland and Témam 1999, Chapter 4]. Let V and Y be Banach spaces with topological duals V^* and Y^*, respectively, and let $\Lambda : V \to Y$ be a continuous linear operator. Setting $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, let $\mathcal{F} : V \to \bar{\mathbb{R}}$, $\mathcal{G} : Y \to \bar{\mathbb{R}}$ be convex lower semi-continuous functionals which are not identically equal ∞ and for which there exists a $v_0 \in V$ such that $\mathcal{F}(v_0) < \infty$, $\mathcal{G}(\Lambda v_0) < \infty$, and $\mathcal{G}$ is continuous at Λv_0. Let $\mathcal{F}^* : V^* \to \bar{\mathbb{R}}$ denote the Fenchel conjugate of $\mathcal{F}$ defined by

$$\mathcal{F}^*(q) = \sup_{v \in V} \langle q, v \rangle_{V^*,V} - \mathcal{F}(v),$$

which we can calculate using the fact that

$$\mathcal{F}^*(q) = \langle q, v \rangle_{V^*,V} - \mathcal{F}(v) \quad \text{if and only if} \quad q \in \partial \mathcal{F}(v). \tag{10.6.1}$$

Here, $\partial \mathcal{F}$ denotes the subdifferential of the convex function $\mathcal{F}$, which reduces to the Gâteaux-derivative if it exists, and the left hand side arises from differentiating the duality pairing.

The Fenchel duality theorem states that under the assumptions given above,

$$\inf_{v \in V} \mathcal{F}(v) + \mathcal{G}(\Lambda v) = \sup_{q \in Y^*} -\mathcal{F}^*(\Lambda^* q) - \mathcal{G}^*(-q)\,, \tag{10.6.2}$$

holds, and that the right hand side of (10.6.2) has at least one solution. Furthermore, the equality in (10.6.2) is attained at $(\bar{v}, \bar{q})$ if and only if

$$\begin{cases} \Lambda^* \bar{q} \in \partial \mathcal{F}(\bar{v}), \\ -\bar{q} \in \partial \mathcal{G}(\Lambda \bar{v}), \end{cases} \tag{10.6.3}$$

where the derivative of the duality pairing again enters the left hand side.

We now apply the Fenchel duality theorem to (Q_σ), which we express in terms of the expansion coefficients $\bar{u}_{kj}$. Let $N_\sigma = N_\tau \times N_h$ and identify as above $u_\sigma \in \mathcal{U}_\sigma$ with the vector $\vec{u}_\sigma = (u_{11}, \dots, u_{1N_h}, \dots, u_{N_\tau N_h})^T \in \mathbb{R}^{N_\sigma}$ of coefficients, and similarly $y_{d,\sigma} \in \mathcal{Y}_\sigma$; see section 10.4.1. To keep the notation simple, we will omit the vector arrows from here on. Denote by $M_h = (\langle e_j, e_k \rangle)_{j,k=1}^{N_h}$ the mass matrix and by $A_h = (a(e_j, e_k))_{j,k=1}^{N_h}$ the stiffness matrix corresponding to Y_h. For the sake of presentation, we fix $y_0 = 0$. Then the discrete state equation (10.4.21) can be expressed as $L_\sigma y_\sigma = u_\sigma$ with

$$L_\sigma = \begin{pmatrix} \tau_1^{-1} M_h + A_h & 0 & 0 \\ -\tau_1^{-1} M_h & \tau_2^{-1} M_h + A_h & 0 \\ 0 & \ddots & \ddots \end{pmatrix} \in \mathbb{R}^{N_\sigma \times N_\sigma}.$$

(Note that the "mass matrix" corresponding to $(\langle \delta_{x_j}, e_k \rangle)_{j,k=1}^{N_h}$ is the identity.) Introducing for $v_\sigma \in \mathbb{R}^{N_\sigma}$ the vectors $v_k = (v_{k1}, \dots, v_{kN_h})^T \in \mathbb{R}^{N_h}$, $1 \leqslant k \leqslant N_\tau$, the discrete optimal control problem (Q_σ) can be stated in reduced form as

$$\min_{u_\sigma \in \mathbb{R}^{N_\sigma}} \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k [L_\sigma^{-1} u_\sigma - y_{d,\sigma}]_k^T M_h [L_\sigma^{-1} u_\sigma - y_{d,\sigma}]_k + \alpha \left(\sum_{k=1}^{N_\tau} \tau_k |u_k|_1^2 \right)^{1/2}.$$

We now set $\Lambda : \mathbb{R}^{N_\sigma} \to \mathbb{R}^{N_\sigma}$, $\Lambda v = L_\sigma^{-1} v$,

$$\begin{aligned} \mathcal{F} : \mathbb{R}^{N_\sigma} \to \mathbb{R}, \qquad & \mathcal{F}(v) = \alpha \left(\sum_{k=1}^{N_\tau} \tau_k |v_k|_1^2 \right)^{1/2}, \\ \mathcal{G} : \mathbb{R}^{N_\sigma} \to \mathbb{R}, \qquad & \mathcal{G}(v) = \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k (v_k - y_{d,k})^T M_h (v_k - y_{d,k}), \end{aligned}$$

and calculate the Fenchel conjugates with respect to the topology induced by the duality pairing (10.4.3). For $\mathcal{G}$, we have by direct calculation that

$$\begin{aligned} \mathcal{G}^*(q) &= \sup_{v \in \mathbb{R}^{N_\sigma}} \sum_{k=1}^{N_\tau} \tau_k q_k^T v_k - \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k (v_k - y_{d,k})^T M_h (v_k - y_{d,k}) \\ &= \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k \left((q_k + M_h y_{d,k})^T M_h^{-1} (q_k + M_h y_{d,k}) - y_{d,k}^T M_h y_{d,k} \right) \end{aligned}$$

since the supremum is attained if and only if $q_k = M_h(v_k - y_{d,k})$ for each $1 \leqslant k \leqslant N_\tau$ due to (10.6.1) and the definition of the duality pairing. For $\mathcal{F}$, we appeal to the fact that in any Banach space the Fenchel conjugate (with respect to the weak-$\star$ topology) of a norm is the indicator function of the unit ball with respect to the dual norm (see, e.g., [Schirotzek 2007, Example 2.2.6]), and to the duality between $\mathcal{U}_\sigma$ and $\mathcal{Y}_\sigma$, to obtain

$$\mathcal{F}^*(q) = \iota_\alpha(q) := \begin{cases} 0 & \text{if } \left(\sum_{k=1}^{N_\tau} \tau_k |q_k|_\infty^2 \right)^{1/2} \leqslant \alpha, \\ \infty & \text{otherwise.} \end{cases}$$

The adjoint $\Lambda^* : \mathbb{R}^{N_\sigma} \to \mathbb{R}^{N_\sigma}$ (with respect to the above duality pairing) is given by L_σ^{-T}. Dropping the constant term in $\mathcal{G}^*$ and substituting $p_\sigma = \Lambda^* q_\sigma$, i.e., $q_\sigma = L_\sigma^T p_\sigma$, we obtain the dual problem

$$\min_{p_\sigma \in \mathbb{R}^{N_\sigma}} \frac{1}{2} \sum_{k=1}^{N_\tau} \tau_k ([L_\sigma^T p_\sigma]_k - M_h y_{d,k})^T M_h^{-1} ([L_\sigma^T p_\sigma]_k - M_h y_{d,k}) + \iota_\alpha(p_\sigma). \tag{10.6.4}$$

Since $v_0 = 0 = \Lambda v_0$ satisfies the regular point condition, the Fenchel duality theorem is applicable, implying the existence of a solution $\bar{p}_\sigma$ which is unique due to the strict convexity in (10.6.4).

While the second relation of (10.6.3),

$$\tau_k(L_\sigma^T \bar{p}_\sigma)_k = \tau_k M_h (L_\sigma^{-1} \bar{u}_\sigma - y_{d,\sigma})_k \quad \text{for all} \quad 1 \leqslant k \leqslant N_\tau, \tag{10.6.5}$$

can in principle be used to obtain $\bar{u}_\sigma$ from $\bar{p}_\sigma$, the first relation remains impractical for numerical computation. We thus consider the following equivalent reformulation of (10.6.4), which decouples the spatio-temporal constraint given by the term $\iota_\alpha(p_\sigma)$:

$$\begin{cases} \min\limits_{p_\sigma \in \mathbb{R}^{N_\sigma}, c_\sigma \in \mathbb{R}^{N_\tau}} \dfrac{1}{2} \displaystyle\sum_{k=1}^{N_\tau} \tau_k ([L_\sigma^T p_\sigma]_k - M_h y_{d,k})^T M_h^{-1} ([L_\sigma^T p_\sigma]_k - M_h y_{d,k}) \\ \text{s.t.} \quad |p_k|_\infty \leqslant c_k \text{ for all } 1 \leqslant k \leqslant N_\tau \quad \text{and} \quad \displaystyle\sum_{k=1}^{N_\tau} \tau_k c_k^2 = \alpha^2, \end{cases}$$

where $c_\sigma = (c_1, \dots, c_{N_\tau})^T \in \mathbb{R}^{N_\tau}$. Since the constraints satisfy a Slater condition (take $p_\sigma = 0$ and $c_k = T^{-1/2}\alpha$, $1 \leqslant k \leqslant N_\tau$), we obtain (e.g., from [Maurer and Zowe 1979]) existence of Lagrange multipliers $\mu_k^1, \mu_k^2 \in \mathbb{R}^{N_h}$, $1 \leqslant k \leqslant N_\tau$, and $\lambda \in \mathbb{R}$ such that the (unique) solution $(\bar{p}_\sigma, \bar{c}_\sigma)$ satisfies the optimality conditions

$$\begin{cases} \tau_k [L_\sigma M_\sigma^{-1} (L_\sigma^T \bar{p}_\sigma - M_\sigma y_{d,\sigma})]_k = \mu_k^1 + \mu_k^2, \quad 1 \leqslant k \leqslant N_\tau, \\ \displaystyle\sum_{j=1}^{N_h} (-\mu_{kj}^1 + \mu_{kj}^2) + 2\lambda \tau_k \bar{c}_k = 0, \quad 1 \leqslant k \leqslant N_\tau, \\ (\mu_k^1)^T (\bar{p}_k - \bar{c}_k) = 0, \quad (\mu_k^2)^T (\bar{p}_k + \bar{c}_k) = 0, \quad \mu_k^1 \leqslant 0, \ \mu_k^2 \geqslant 0, \quad 1 \leqslant k \leqslant N_\tau, \\ \displaystyle\sum_{k=1}^{N_\tau} \tau_k \bar{c}_k^2 - \alpha^2 = 0, \end{cases} \tag{10.6.6}$$

where $M_\sigma \in \mathbb{R}^{N_\sigma \times N_\sigma}$ is a block diagonal matrix containing N_τ copies of M_h.

We now rewrite the optimality system in a form amenable to the numerical solution using a semismooth Newton method. First, μ_k^1 and μ_k^2 are scaled by $\tau_k > 0$ to eliminate this factor from the first and second relation (which does not affect the complementarity conditions). Using the componentwise max and min functions, the complementarity conditions for μ_k^1, μ_k^2 and $\bar{p}_k$ can be expressed equivalently for any $\gamma > 0$ as

$$\mu_k^1 + \max(0, -\mu_k^1 + \gamma(\bar{p}_k - \bar{c}_k)) = 0, \qquad \mu_k^2 + \min(0, -\mu_k^2 + \gamma(\bar{p}_k + \bar{c}_k)) = 0.$$

Since $\mu_k^2 = 0$ if $\bar{p}_k > -\bar{c}_k$ and $\mu_k^1 = 0$ if $\bar{p}_k < \bar{c}_k$, we have by componentwise inspection

$$\max(0, -\mu_k^1 + \gamma(\bar{p}_k - \bar{c}_k)) = \max(0, -\mu_k^1 - \mu_k^2 + \gamma(\bar{p}_k - \bar{c}_k)).$$

We argue similarly for the min term. Furthermore, comparing the first relation of (10.6.6) with (10.6.5), we deduce that $\bar{u}_k = \mu_k^1 + \mu_k^2$ for all $1 \leqslant k \leqslant N_\tau$. Finally, to avoid having to form M_σ^{-1}, we introduce $\bar{y}_\sigma \in \mathbb{R}^{N_\sigma}$ satisfying

$$L_\sigma^T \bar{p}_\sigma = M_\sigma (\bar{y}_\sigma - y_{d,\sigma}).$$

Inserting these relations into (10.6.6), we obtain for every $\gamma > 0$ the optimality system (10.6.7)

$$\begin{cases} L_\sigma \bar{y}_\sigma - \bar{u}_\sigma = 0, \\ L_\sigma^T \bar{p}_\sigma - M_\sigma(\bar{y}_\sigma - y_{d,\sigma}) = 0, \\ \bar{u}_k + \max(0, -\bar{u}_k + \gamma(\bar{p}_k - \bar{c}_k)) + \min(0, -\bar{u}_k + \gamma(\bar{p}_k + \bar{c}_k)) = 0, \quad 1 \leqslant k \leqslant N_\tau \\ \sum_{j=1}^{N_h} [-\max(0, -\bar{u}_k + \gamma(\bar{p}_k - \bar{c}_k)) + \min(0, -\bar{u}_k + \gamma(\bar{p}_k + \bar{c}_k))]_j + 2\lambda \bar{c}_k = 0, \\ \qquad\qquad 1 \leqslant k \leqslant N_\tau, \\ \sum_{k=1}^{N_\tau} \tau_k \bar{c}_k^2 - \alpha^2 = 0. \end{cases}$$

Since the max and min functions are globally Lipschitz mappings in finite dimensions, this defines a semismooth equation which can be solved using a generalized Newton method; see, e.g., [Qi and Sun 1993; Kummer 1992]. Here we recall that the Newton derivative of $\max(0, v)$ with respect to v is given componentwise by

$$[D_N \max(0, v)h]_k = \begin{cases} h_k & \text{if } v_k \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

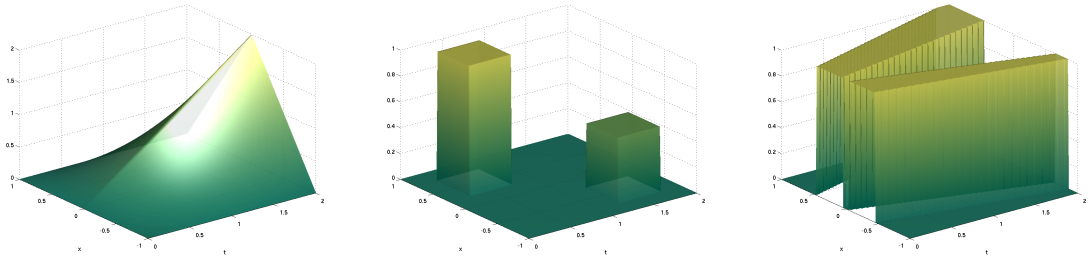
and similarly that of $\min(0, v)$. In practice, we have to account for the possibly local convergence of the Newton method. To compute a suitable starting point, as an initialization step we successively solve a sequence of approximating problems that are obtained from (10.6.7) by replacing the max and min terms with

$$\max(0, \gamma(\bar{p}_k - \bar{c}_k)) \quad \text{and} \quad \min(0, \gamma(\bar{p}_k + \bar{c}_k)),$$

respectively, and letting γ tend to infinity. (This can be interpreted as a Moreau–Yosida regularization of the complementarity conditions.) Since now u_k no longer appears in the argument of the max and min functions, it can be eliminated from the optimality system using the third equation (which also allows computing $\bar{u}_k$ given $(\bar{p}_k, \bar{c}_k)$), yielding (10.6.8)

$$\begin{cases} L_\sigma^T p_\gamma - M_\sigma(y_\gamma - y_{d,\sigma}) = 0, \\ L_\sigma y_\gamma + \gamma[\max(0, p_\gamma - c_\gamma) + \min(0, p_\gamma + c_\gamma)] = 0, \\ \sum_{j=1}^{N_h} \gamma\,[-\max(0, p_{\gamma,k} - c_{\gamma,k}) + \min(0, p_{\gamma,k} + c_{\gamma,k})]_j + 2\lambda_\gamma c_{\gamma,k} = 0, \quad 1 \leqslant k \leqslant N_\tau \\ \sum_{k=1}^{N_\tau} \tau_k c_{\gamma,k}^2 - \alpha^2 = 0. \end{cases}$$

Starting with $\gamma = 1$ and $p^0 = y^0 = 0$, $c^0 = T^{-1/2}\alpha$ and $\lambda^0 = 1$, we solve (10.6.8) using a semismooth Newton method, increase γ, and compute a new solution for increased γ with



(a) z_1 (b) z_2 (c) z_3

Figure 10.1: Targets for numerical experiments

the previous solution as starting point. Once a solution satisfies the constraints (or a stopping value γ^* is reached), we use it as a starting point for the solution of (10.6.7) with $\gamma = 1$.

Remark 10.6.1. By virtue of the chosen discretization (specifically, the adjoint consistency of discontinuous Galerkin methods and the discrete topology mirroring the continuous one), the discrete optimality system (10.6.8) coincides with the discretization of the continuous optimality system obtained by applying Fenchel duality, the relaxation approach and a Moreau–Yosida approximation to problem (P). Since the continuous optimality system may be of independent interest, the derivation is sketched in Appendix 10.A.

10.7 NUMERICAL EXAMPLES

We illustrate the structure of the optimal controls with some one-dimensional examples. For this purpose we set $\Omega = (-1, 1)$, $T = 2$, $\nu = 10^{-1}$ and consider the state equation

$$\begin{cases} y_t - \nu \Delta y = u, \\ y(0) = 0, \end{cases}$$

with homogeneous Dirichlet conditions. The spatial domain is discretized using $N_h = 128$ uniformly distributed nodes (which corresponds to $h \approx 0.0156$). Following (10.4.2), we take $N_\tau = 1024$ time steps (which corresponds to $\tau \approx 0.00195$). The targets are chosen as (see Figure 10.1)

$$z_1 = t(1 - |x|),$$

$$z_2 = \begin{cases} 1 & \text{if } 0.25 \leqslant t \leqslant 0.75 \text{ and } 0.25 \leqslant x \leqslant 0.75, \\ \frac{1}{2} & \text{if } 1.25 \leqslant t \leqslant 1.75 \text{ and } -0.25 \geqslant x \geqslant -0.75, \\ 0 & \text{otherwise,} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{if } |x - 0.25 - t/4| < (0.2 + t/20), \\ 1 & \text{if } |x + 0.25 + t/4| < (0.2 - t/20), \\ 0 & \text{otherwise.} \end{cases}$$

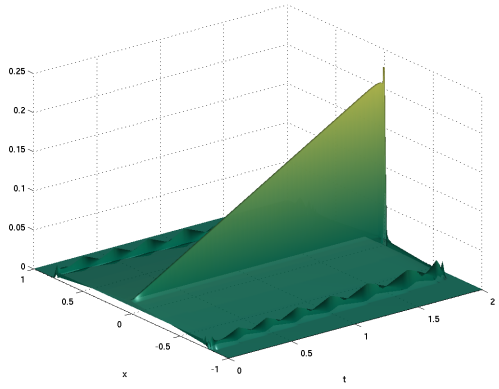
The semismooth Newton method for the solution of the optimality system (10.6.7) is implemented in MATLAB, where the initialization is calculated as discussed in section 10.6 with $\gamma_{k+1} = 10\gamma_k$ and $\gamma^* = 10^{12}$. For each target the optimal control is computed for $\alpha = 10^{-3}$ and $\alpha = 10^{-1}$. In every case, the discrete optimality system is solved to an accuracy below 10^{-12}, and the bounds on p_σ and on c_σ are attained within machine precision.

The respective optimal controls u_σ (in the form of linearly interpolated expansion coefficients u_{kj}), optimal states y_σ and bounds c_σ are shown in Figure 10.2–10.4. The predicted sparsity structure of the optimal controls can be observed clearly: The spatio-temporal coupling of the control cost predominantly promotes spatial sparsity; see Figure 10.3b in particular. The structural features of the norm $\|u\|_{L^2(\mathcal{M})}$ are further illustrated by the fact that larger values of α lead to both increased sparsity in space and increased smoothness in time. It is instructive to compare the optimal controls obtained with our $\|u\|_{L^2(\mathcal{M})}$ regularization to those obtained numerically using a (Moreau–Yosida approximation of a) $\mathcal{M}(\Omega_T)$ norm penalty term. Figure 10.5 shows the latter for all considered targets and values of α. While for $\alpha = 10^{-3}$ both types of control have comparable structure, for $\alpha = 10^{-1}$ the controls in $\mathcal{M}(\Omega_T)$ demonstrate strong temporal sparsity which is absent in the case of controls in $L^2(I, \mathcal{M}(\Omega))$.

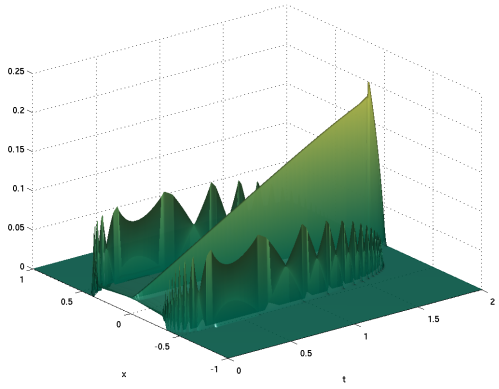
We now investigate the convergence behavior as $h \to 0$. In the absence of a known exact solution, we take as a reference solution the computed optimal discrete control and optimal discrete state on the finest grid with $N_{h^*} = 256$ and $N_{\tau^*} = 4096$, corresponding to $h^* \approx 0.00781$ and $\tau^* \approx 0.000488$. As a representative example, we consider the target z_1 and $\alpha = 0.1$. Figure 10.6a shows the difference $|J_h - J_{h^*}|$ for a series of successively refined grids with $N_h = 32, 40, \ldots, 128$ and $N_{\tau(h)} = \frac{1}{16}N_h^2$. The observed approximately linear convergence rate agrees with the rate obtained in Theorem 10.5.1. The corresponding L^2 error $\|y_h - y_{h^*}\|_{L^2}$ of the discrete states also decays with a linear rate, which is faster than predicted by Theorem 10.5.2. A similar behavior was observed in the elliptic case; see [Casas, Clason, and Kunisch 2012].

10.8 CONCLUSION

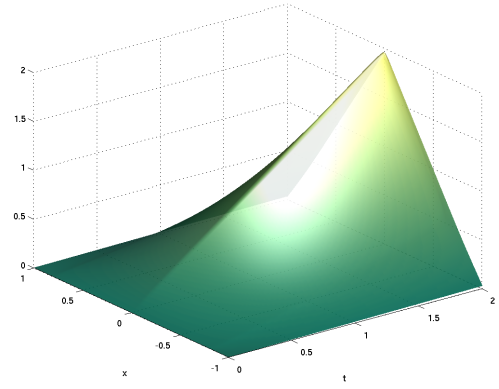
For the appropriate functional-analytic setting of parabolic optimal control problems in measure spaces, there exists a straightforward approximation framework that retains the structural properties of the norm in the measure-valued Banach space and allows deriving numerically accessible optimality conditions as well as convergence rates. In particular, although the state is discretized, the control problem is still formulated and solved in measure space. The numerical results demonstrate that the optimal controls exhibit the expected sparsity pattern.



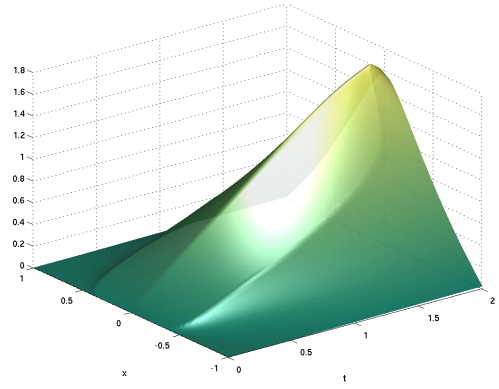
(a) u_σ



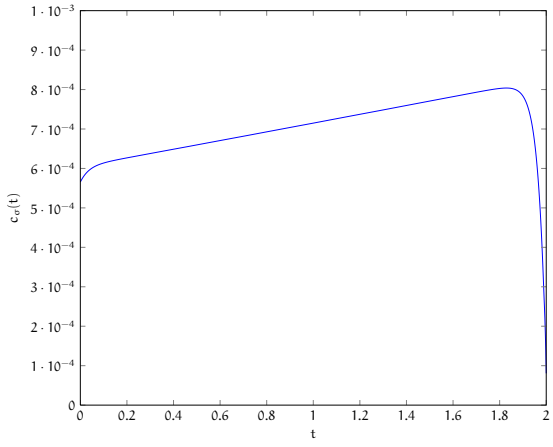
(b) u_σ



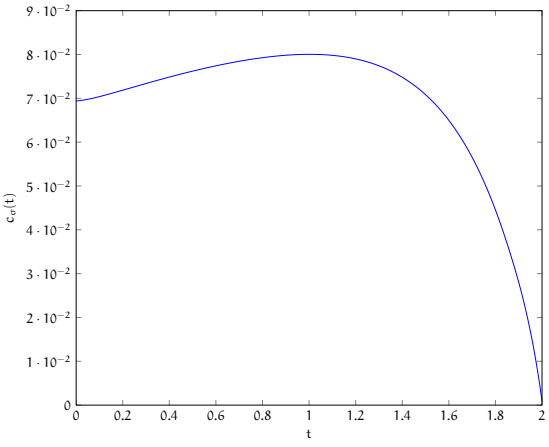
(c) y_σ



(d) y_σ

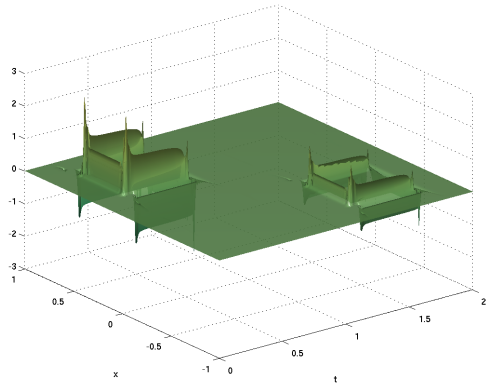


(e) c_σ

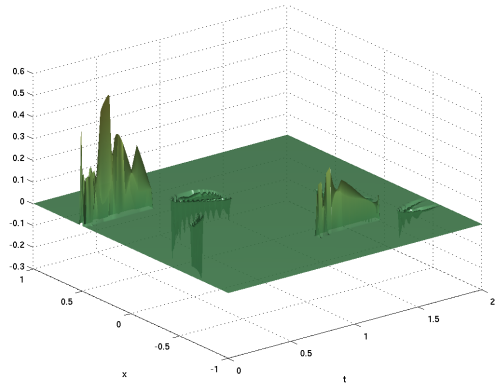


(f) c_σ

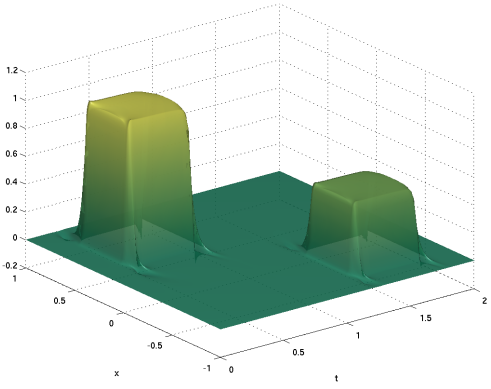
Figure 10.2: Optimal control u_σ, state y_σ and bound c_σ for target z_1 and $\alpha = 10^{-3}$ (left), $\alpha = 10^{-1}$ (right).



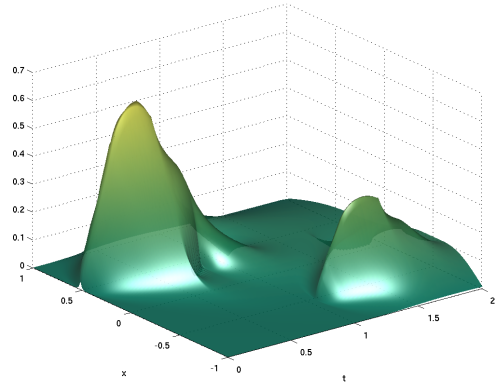
(a) u_σ



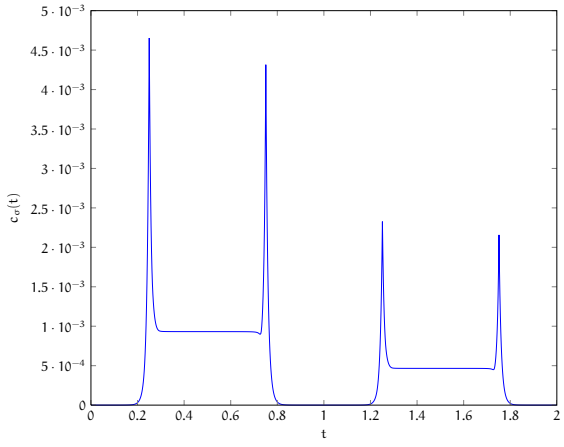
(b) u_σ



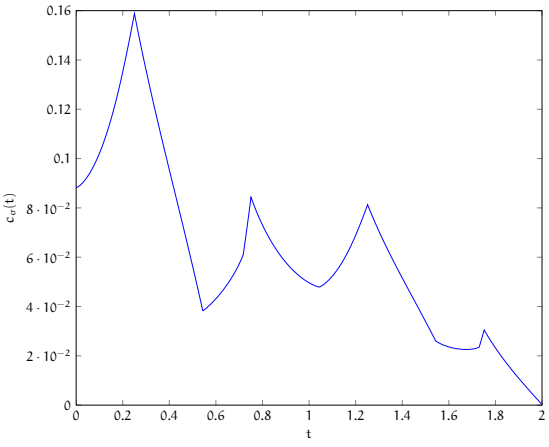
(c) y_σ



(d) y_σ

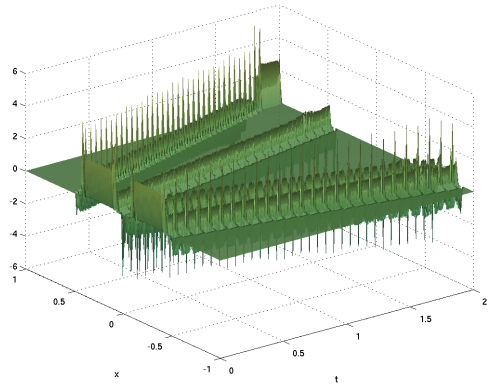


(e) c_σ

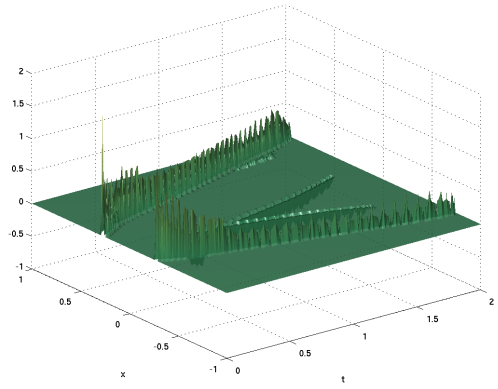


(f) c_σ

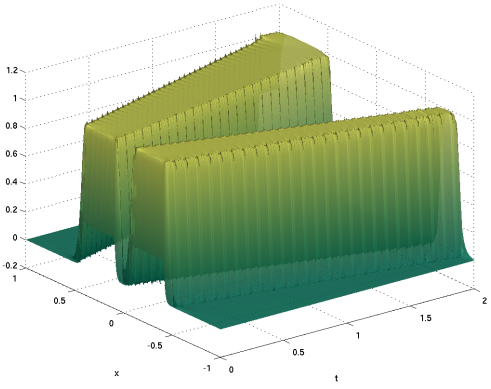
Figure 10.3: Optimal control u_σ, state y_σ and bound c_σ for target z_2 and $\alpha = 10^{-3}$ (left), $\alpha = 10^{-1}$ (right).



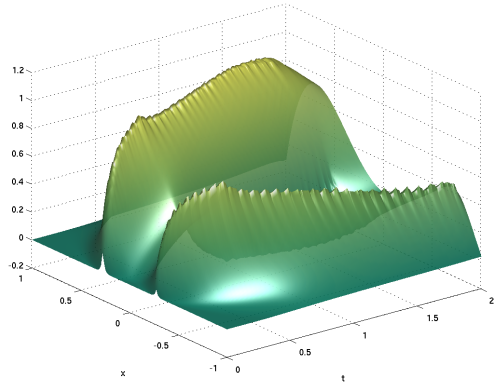
(a) u_σ



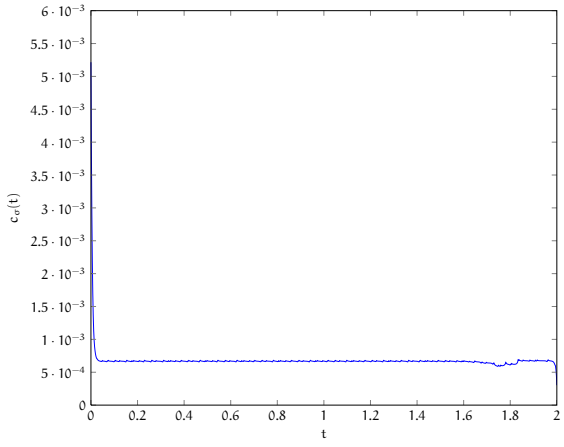
(b) u_σ



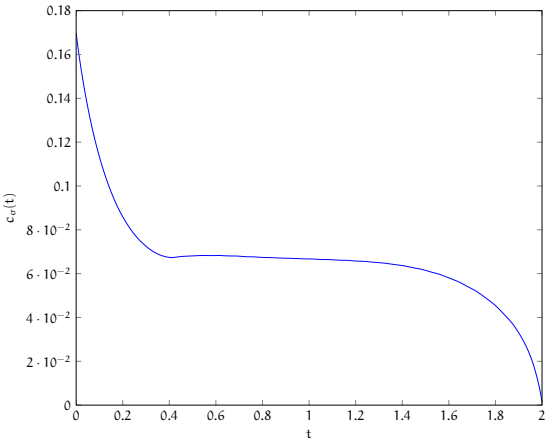
(c) y_σ



(d) y_σ

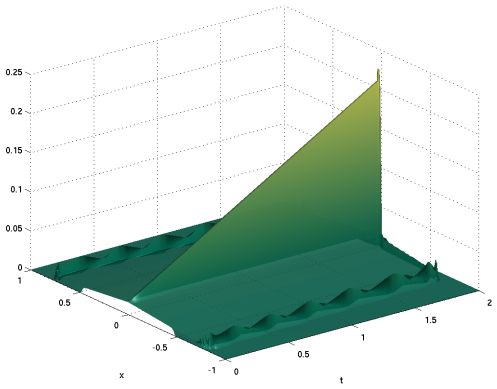


(e) c_σ

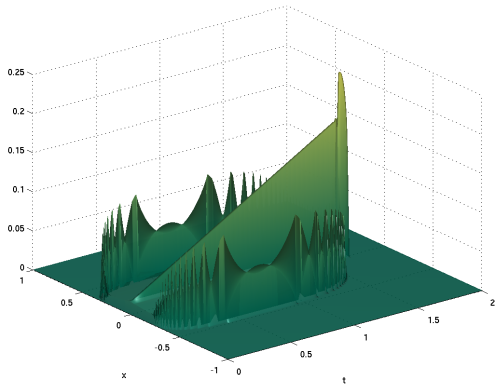


(f) c_σ

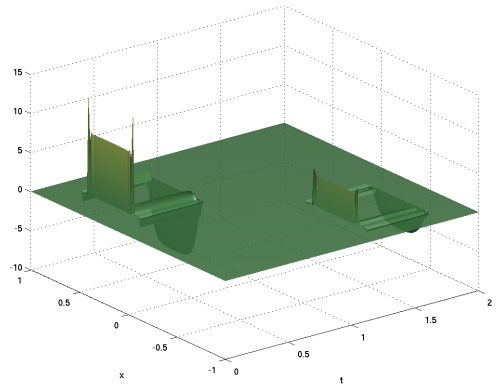
Figure 10.4: Optimal control u_σ, state y_σ and bound c_σ for target z_3 and $\alpha = 10^{-3}$ (left), $\alpha = 10^{-1}$ (right).



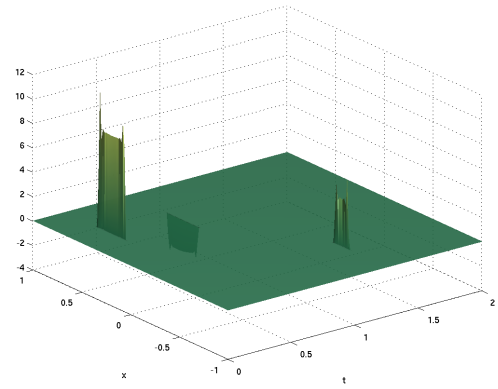
(a) target z_1, $\alpha = 10^{-3}$



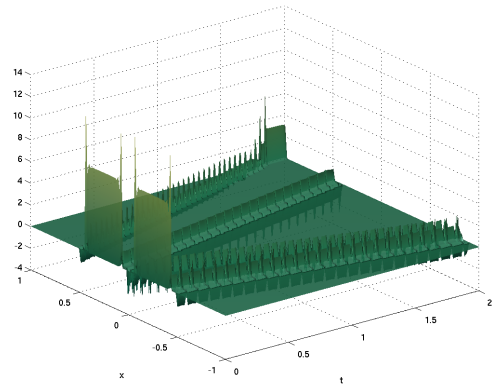
(b) target z_1, $\alpha = 10^{-1}$



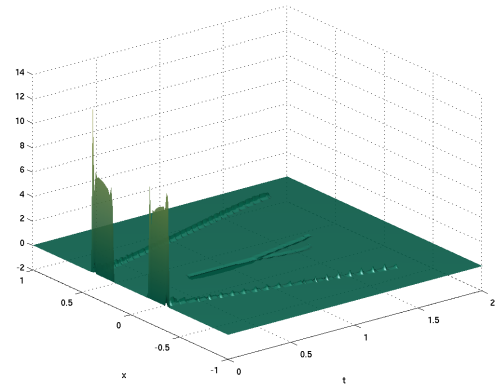
(c) target z_2, $\alpha = 10^{-3}$



(d) target z_2, $\alpha = 10^{-1}$

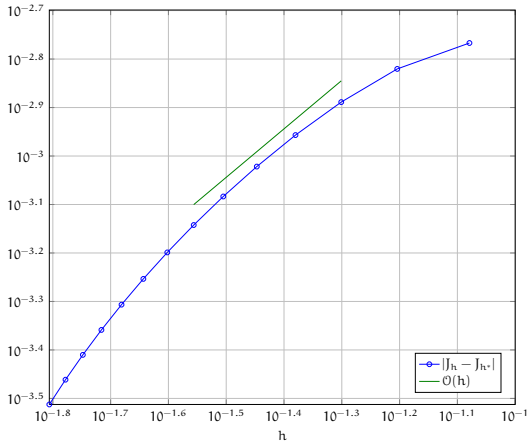


(e) target z_3, $\alpha = 10^{-3}$

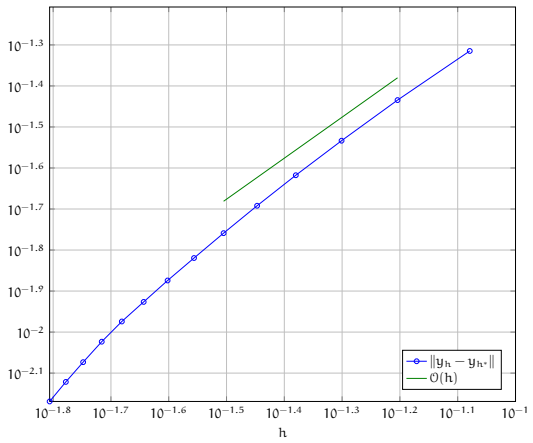


(f) target z_3, $\alpha = 10^{-1}$

Figure 10.5: Optimal controls with $\mathcal{M}(\Omega_T)$ penalty.



(a) Functional value J_h



(b) L^2 norm of state y_h

Figure 10.6: Illustration of convergence order for target z_1 and $\alpha = 0.1$.

ACKNOWLEDGMENTS

The first author was supported by the Spanish Ministerio de Ciencia e Innovación under projects MTM2011-22711. The remaining authors were supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences").

10.A CONTINUOUS OPTIMALITY SYSTEM

In this section we sketch the derivation of the continuous optimality system using Fenchel duality and the relaxation approach. Let $S : L^2(I, \mathcal{M}(\Omega)) \to L^2(\Omega_T)$ denote the solution operator corresponding to the state equation (10.1.1) with homogeneous initial conditions. It will be convenient to introduce the parabolic differential operator L such that the solution y to (10.1.1) satisfies $Ly = u$. Then we can express problem (P) in reduced form as

$$\min_{u \in L^2(I,\mathcal{M}(\Omega))} \frac{1}{2}\|Su - y_d\|^2_{L^2(\Omega_T)} + \alpha\|u\|_{L^2(I,\mathcal{M}(\Omega))}.$$

To apply Fenchel duality, we set

$$\begin{aligned} &\mathcal{F} : L^2(I, \mathcal{M}(\Omega)) \to \mathbb{R}, && \mathcal{F}(v) = \alpha\|v\|_{L^2(I,\mathcal{M}(\Omega))}, \\ &\mathcal{G} : L^2(\Omega_T) \to \mathbb{R}, && \mathcal{G}(v) = \frac{1}{2}\|v - y_d\|^2_{L^2(\Omega_T)}, \\ &\Lambda : L^2(I, \mathcal{M}(\Omega)) \to L^2(\Omega_T), && \Lambda u = Su. \end{aligned}$$

Similarly to the discrete case, the Fenchel conjugates (with respect to the weak-$\star$ topology) are given by

$$\begin{aligned} &\mathcal{F}^* : L^2(I, C_0(\Omega)) \to \mathbb{R}, && \mathcal{F}^*(q) = \iota_\alpha(q) \\ &\mathcal{G}^* : L^2(\Omega_T) \to \mathbb{R}, && \mathcal{G}^*(q) = \frac{1}{2}\|q + y_d\|^2_{L^2(\Omega_T)} - \frac{1}{2}\|y_d\|^2_{L^2(\Omega_T)}, \end{aligned}$$

where

$$\iota_\alpha(q) = \begin{cases} 0 & \text{if } \|q\|_{L^2(I,C_0(\Omega))} \leqslant \alpha, \\ \infty & \text{otherwise.} \end{cases}$$

Due to the definition of the solution to (10.1.1) via duality (see Definition 10.2.1), we obtain the existence of a weak-$\star$ adjoint operator $\Lambda^* := S^* : L^2(\Omega_T) \to L^2(I, C_0(\Omega))$ defined via the solution to (10.2.5). Furthermore, there exists a weak-$\star$ adjoint L^* of L such that for given $\psi_0 \in L^2(\Omega_T)$, the solution $z \in L^2(I, C_0(\Omega))$ of (10.2.5) satisfies $L^* z = \psi_0$. The dual problem is then found to be

$$\min_{q \in L^2(\Omega_T)} \frac{1}{2}\|q - y_d\|^2_{L^2(\Omega_T)} + \iota_\alpha(S^* q).$$

We again substitute $p = S^* q \in L^2(I, C_0(\Omega))$, i.e., $q = L^* p$, introduce $c \in L^2(I)$ by

$$c(t) := \|p(t)\|_\infty \qquad \text{for a. e. } 0 \leqslant t \leqslant T,$$

and consider

$$\text{(10.A.1)} \qquad \begin{cases} \min\limits_{p \in L^2(I, C_0(\Omega)), c \in L^2(I)} \dfrac{1}{2}\|L^* p - y_d\|^2_{L^2(\Omega_T)} \\ \text{s. t.} \quad \|p(t)\|_\infty \leqslant c(t) \text{ for a. e. } 0 \leqslant t \leqslant T \\ \text{and} \quad \displaystyle\int_0^T c(t)^2 \, dt = \alpha^2. \end{cases}$$

The Moreau–Yosida regularization of (10.A.1) is given by

$$\begin{cases} \min\limits_{p \in L^2(I, C_0(\Omega)), c \in L^2(I)} \dfrac{1}{2}\|L^* p - y_d\|^2_{L^2(\Omega_T)} + \dfrac{\gamma}{2}\Big[\|\max(0, p - c)\|^2_{L^2(\Omega_T)} \\ \qquad\qquad + \|\min(0, p + c)\|^2_{L^2(\Omega_T)}\Big] \\ \text{s. t.} \quad \displaystyle\int_0^T c(t)^2 \, dt = \alpha^2, \end{cases}$$

where the max and min functions should be understood pointwise in Ω for almost every $0 \leqslant t \leqslant T$. Its solution is denoted by $(p_\gamma, c_\gamma) \in L^2(I, C_0(\Omega)) \times L^2(I)$. Since the cost functional is Fréchet differentiable and a Slater condition is again satisfied for the constraint

on c (take $c = T^{-1/2}\alpha$), we obtain existence of a Lagrange multiplier $\lambda_\gamma \in \mathbb{R}$. Introducing once more y_γ satisfying $L^* p_\gamma = y_\gamma - y_d$, this yields the continuous optimality system

$$\begin{cases} L^* p_\gamma - (y_\gamma - y_d) = 0, \\ L y_\gamma + \gamma \max(0, p_\gamma - c_\gamma) + \gamma \min(0, p_\gamma + c_\gamma) = 0, \\ \gamma \displaystyle\int_\Omega - \max(0, p_\gamma - c_\gamma) + \min(0, p_\gamma + c_\gamma)\, dx + 2\lambda_\gamma c_\gamma = 0, \\ \displaystyle\int_0^T c_\gamma^2\, dt - \alpha^2 = 0. \end{cases}$$

By approximating p_γ and y_γ in $\mathcal{Y}_\sigma$, using the fact that for linear finite elements the pointwise maximum and minimum is attained at the nodes, and the adjoint consistency of discontinuous Galerkin methods (i.e., $(L^*)_\sigma = L_\sigma^T$), we recover (10.6.8).

Part III

OPTIMAL CONTROL WITH L^∞ FUNCTIONALS

MINIMAL INVASION: AN OPTIMAL L^∞ STATE CONSTRAINT PROBLEM

11

ABSTRACT

In this work, the least pointwise upper and/or lower bounds on the state variable on a specified subdomain of a control system under piecewise constant control action are sought. This results in a nonsmooth optimization problem in function spaces. Introducing a Moreau–Yosida regularization of the state constraints, the problem can be solved using a superlinearly convergent semismooth Newton method. Optimality conditions are derived, convergence of the Moreau–Yosida regularization is proved, and well-posedness and superlinear convergence of the Newton method is shown. Numerical examples illustrate the features of this problem and the proposed approach.

11.1 INTRODUCTION

We consider the following relaxed L^∞-type control problem:

$$(\mathcal{P}) \qquad \begin{cases} \min\limits_{c\in\mathbb{R},u\in\mathbb{R}^m} \dfrac{c^2}{2} + \dfrac{\alpha}{2}|u|_2^2 \\ \text{s. t. } Ay = f + \displaystyle\sum_{i=1}^{m} u_i \chi_{\omega_i} \qquad \text{in } \Omega, \\ -\beta_2 c + \psi_2 \leqslant y|_{\omega_0} \leqslant \beta_1 c + \psi_1 \quad \text{in } \omega_0. \end{cases}$$

Here $\alpha > 0$, Ω is a bounded domain in $\mathbb{R}^n$, A is a linear second order elliptic partial differential operator of convection-diffusion type carrying appropriate boundary conditions (to be made more explicit below), $\omega_i \subset \Omega$, $i = 0, \dots, m$ are subdomains and as such open and connected sets in Ω with characteristic functions

$$\chi_{\omega_i}(x) = \begin{cases} 1 & x \in \omega_i, \\ 0 & x \notin \omega_i, \end{cases}$$

and $f \in L^q(\Omega)$ for some $q < \max(2, n)$. Further

$$\beta_1, \beta_2 \in \mathbb{R} \text{ with } \beta_1, \beta_2 \geqslant 0 \quad \text{and} \quad \psi_1 \in L^\infty(\omega_0),\ \psi_2 \in L^\infty(\omega_0),$$

and we assume that $\beta_1 + \beta_2 > 0$ as well as $\max \psi_2 \leqslant \min \psi_1$. Furthermore, we assume that $\psi_1(\bar{x}) = \psi_2(\bar{x})$ for some $\bar{x} \in \overline{\omega}_0$, which can always be guaranteed by re-parametrization according to

$$\bar{\psi}_1 = \psi_1 + \beta_1 \bar{c}, \qquad \bar{\psi}_2 = \psi_2 - \beta_2 \bar{c},$$

where $\bar{c} = \frac{d}{\beta_1+\beta_2} \leqslant 0$ with $d = \max(\psi_2 - \psi_1) \leqslant 0$. Indeed, let $\bar{x} = \arg\max(\psi_2 - \psi_1)$. Then note that $\bar{\psi}_1 - \bar{\psi}_2 \leqslant 0$ and $\bar{\psi}_1(\bar{x}) - \bar{\psi}_2(\bar{x}) = 0$. Hence after re-parametrization it necessarily holds that $c \geqslant 0$.

To simplify notation, we introduce the control operator $B : \mathbb{R}^m \to L^\infty(\Omega)$,

$$Bu = \sum_{i=1}^{m} u_i \chi_{\omega_i}.$$

This problem can be given the following interpretation: A pollutant f enters the groundwater and is (diffusively and/or convectively) transported throughout the domain Ω. To minimize the concentration y of a pollutant in a city ω_0, wells $\omega_1, \ldots, \omega_m$ are placed in Ω, through which a counter-agent u_i can be introduced. The problem is therefore to minimize the upper bound c in the formulation $y|_{\omega_0} \leqslant c$, or, if the concentration is supposed to be non-negative, $0 \leqslant y|_{\omega_0} \leqslant c$. In general the concentration only satisfies inhomogeneous boundary conditions $\tilde{y} = g$ on $\partial\Omega$. To return to the formulation introduced above we transform to homogeneous boundary conditions by means of $y = \tilde{y} - g_{ext}$, where g_{ext} is a smooth extension of g into Ω. The resulting constraints on y are of the form $-g_{ext} \leqslant y \leqslant -g_{ext} + c$, and are a special case of the constrains considered above. The approach we present can readily be applied to a problem with a unilateral constraint on y.

In case $\beta_1 = \beta_2 = 1$ and $\psi_1 = \psi_2 = 0$ the inequality constraints in ($\mathcal{P}$) result in the norm constraint problem

$$\|y\|_{L^\infty(\omega_0)} \leqslant c.$$

which can equivalently be expressed as the following quadratic problem with affine constraints:

$$\begin{cases} \min\limits_{u \in \mathbb{R}^m} \dfrac{1}{2} \|y\|^2_{L^\infty(\omega_0)} + \dfrac{\alpha}{2} |u|_2^2 \\ \text{s. t. } Ay = f + Bu. \end{cases} \tag{11.1.1}$$

Clearly ($\mathcal{P}$) is related to state-constrained optimal control problems but it is different since it involves c as a free variable. To find the smallest c such that the constraints in ($\mathcal{P}$) admit

a feasible solution and such that the objective is minimized is the objective of this work. Note that for (11.1.1) with $f \neq 0$ it is required that $c > 0$ to guarantee that the constraint $\|y\|_{L^\infty(\omega_0)} \leqslant c$ is feasible.

Problem ($\mathcal{P}$) with $\omega_i = \Omega$ was treated in [Grund and Rösch 2001; Prüfert and Schiela 2009]. In [Grund and Rösch 2001] a discretize before optimize approach was pursued so that phenomena due to lack of L^2-regularity of the Lagrange multipliers are not apparent. The work in [Prüfert and Schiela 2009] rests on an interior point treatment of the state constraints. In addition to the different treatment that we follow in this work, in the numerical examples we also focus on effects and the interpretation which result from the choice of the control and observation domains as proper subsets of Ω.

This article is organized as follows. In a short section 11.2 we present the regularization that we employ, prove existence and uniqueness of a solution to ($\mathcal{P}$) and establish the asymptotic behavior of the solutions to the regularized problems. Section 11.3 is devoted to the optimality systems for the original and the regularized problems. The semismooth Newton method and its analysis are considered in section 11.4, and the final section 11.5 contains numerical results.

11.2 EXISTENCE AND REGULARIZATION

This section is devoted to specifying the regularization that we use, and to establish existence and uniqueness results. We first address well-posedness of the state equation. We consider the operator

$$Ay = -\sum_{j,k=1}^{n} \partial_j(a_{jk}(x)\partial_k y + d_j(x)y) + \sum_{j=1}^{n} b_j(x)\partial_j y + d(x)y,$$

where the coefficients satisfy $a_{jk} \in C^{0,\delta}(\overline{\omega})$ for some $\delta \in (0,1)$ and $b_j, d \in L^\infty(\Omega)$, and the corresponding Dirichlet problem

$$\begin{cases} Ay = g, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \tag{11.2.1}$$

where the domain Ω is open, bounded and of class $C^{1,\delta}$ and $g \in H^{-1}(\Omega)$ is given. If 0 is not an eigenvalue of A, this problem has a unique solution in $H_0^1(\Omega)$. A sufficient assumption for this is the existence of constants $\lambda, \nu > 0$ such that

$$\begin{cases} \lambda|\xi|_2^2 \leqslant a_{jk}\xi_j\xi_k \quad \text{for all } \xi \in \mathbb{R}^n, \qquad \displaystyle\sum_{j,k=1}^{n} |a_{j,k}|^2 \leqslant \lambda^2, \\ \lambda^{-2}\displaystyle\sum_{j=1}^{n} \left(|d_j|^2 + |b_j|^2\right) + \lambda^{-1}|d| \leqslant \nu^2, \qquad d - \partial_j d_j \geqslant 0, \quad \text{for all } 1 \leqslant j \leqslant n, \end{cases}$$

where the last inequality should be understood in the generalized sense (cf., e.g., [Gilbarg and Trudinger 2001, Th. 8.3]). Concerning the regularity of this solution, we have the following theorem [Troianiello 1987, Th. 3.16]:

Proposition 11.2.1. *For each $g \in W^{-1,q}(\Omega)$ with $2 < q < \infty$, the solution y of (11.2.1) satisfies $y \in W_0^{1,q}(\Omega)$. Moreover, there exists a constant $C > 0$ independent of g such that*

$$\|y\|_{W^{1,q}(\Omega)} \leqslant C \|g\|_{W^{-1,q}(\Omega)}.$$

holds.

In particular, since $f \in L^q(\Omega)$ with $q > n$, this implies the existence of a unique solution $y \in W_0^{1,q}(\Omega)$ of the state equation $Ay = f + Bu$ for any control vector $u \in \mathbb{R}^m$. This affine solution mapping will be denoted by

$$y : \mathbb{R}^m \to W_0^{1,q}(\Omega), \text{ with } y(u) = A^{-1}(f + Bu).$$

Recalling the continuous embedding $W^{1,q}(\Omega) \hookrightarrow C(\bar{\Omega})$ for any $q > n$ we have moreover that $y \in C(\bar{\Omega})$.

To apply a semismooth Newton method, we introduce the Moreau–Yosida regularization of ($\mathcal{P}$), i.e. for $\gamma > 0$ we consider:

$$(\mathcal{P}_\gamma) \qquad \begin{cases} \min\limits_{c\in\mathbb{R}, u\in\mathbb{R}^m} \dfrac{c^2}{2} + \dfrac{\alpha}{2}|u|_2^2 + \dfrac{\gamma}{2}\|\max(0, y|_{\omega_0} - (\beta_1 c + \psi_1))\|_{L^2}^2 \\ \qquad\qquad + \dfrac{\gamma}{2}\|\min(0, y|_{\omega_0} + \beta_2 c - \psi_2)\|_{L^2}^2, \\ \text{s. t. } Ay = f + Bu. \end{cases}$$

For the case $\beta_1 = \beta_2 = 1$ and $\psi_1 = \psi_2 = 0$ this can be expressed compactly as

$$\begin{cases} \min\limits_{c\in\mathbb{R}, u\in\mathbb{R}^m} \dfrac{c^2}{2} + \dfrac{\alpha}{2}|u|_2^2 + \dfrac{\gamma}{2}\|\max(0, |y|_{\omega_0}| - c)\|_{L^2}^2, \\ \text{s. t. } Ay = f + Bu. \end{cases}$$

Proposition 11.2.2. *Problem ($\mathcal{P}$) admits a unique solution (c^*, u^*). Moreover, for every $\gamma > 0$ there exists a unique solution (c_γ, u_γ) to ($\mathcal{P}_\gamma$). The associated states will be denoted by $y^* = y(u^*)$ and $y_\gamma = y(u_\gamma)$ respectively.*

Proof. Problem ($\mathcal{P}$) can equivalently be expressed as

$$\min_{u\in\mathbb{R}^m} J(u) = \min_{u\in\mathbb{R}^m} \frac{1}{2}\left[\operatorname*{ess\,sup}_{x\in\omega_0} \max\left(\frac{y - \psi_1}{\beta_1}, \frac{-y + \psi_2}{\beta_2}\right)\right]^2 + \frac{\alpha}{2}|u|_2^2,$$

where $J : \mathbb{R}^m \to \mathbb{R}$ is continuous and radially unbounded. The mapping

$$u \mapsto \operatorname*{ess\,sup}_{x\in\omega_0} \max\left(\frac{y(u)-\psi_1}{\beta_1}, \frac{-y(u)+\psi_2}{\beta_2}\right)$$

is convex and hence $u \mapsto J(u)$ is strictly convex. As a consequence ($\mathcal{P}$) has a unique solution with

$$0 \leqslant c^* = \operatorname*{ess\,sup}_{x\in\omega_0} \max\left(\frac{y(u^*)-\psi_1}{\beta_1}, \frac{-y(u^*)+\psi_2}{\beta_2}\right).$$

An analogous argument implies the existence of a unique solution to ($\mathcal{P}_\gamma$). □

We also define

$$\begin{cases} \lambda_\gamma = \lambda_{\gamma,1} + \lambda_{\gamma,2}, \\ \lambda_{\gamma,1} = \gamma \max(0, y_\gamma|_{\omega_0} - (\beta_1 c + \psi_1)), \quad \lambda_{\gamma,2} = \gamma \min(0, y_\gamma|_{\omega_0} + \beta_2 c - \psi_2), \end{cases}$$

which will turn out to be the regularized Lagrange multiplier associated to the inequality constraint on $y|_{\omega_0}$. Note that $\lambda_{\gamma,1} \geqslant 0$, $\lambda_{\gamma,2} \leqslant 0$ and that strict inequalities cannot hold simultaneously.

Proposition 11.2.3. *We have*

$$(c_\gamma, u_\gamma, y_\gamma) \to (c^*, u^*, y^*) \text{ in } \mathbb{R} \times \mathbb{R}^m \times W^{1,q}(\Omega),$$

and

$$\frac{1}{\sqrt{\gamma}} \|\lambda_\gamma\|_{L^2(\omega_0)} \to 0, \text{ as } \gamma \to \infty. \tag{11.2.2}$$

Proof. Due to the optimality of (c_γ, u_γ) we have

$$\frac{(c_\gamma)^2}{2} + \frac{\alpha}{2}|u_\gamma|_2^2 + \frac{1}{2\gamma}\|\lambda_\gamma\|^2_{L^2(\omega_0)} \leqslant \frac{(c^*)^2}{2} + \frac{\alpha}{2}|u^*|_2^2. \tag{11.2.3}$$

Consequently

$$\{c_\gamma\}_{\gamma>0},\ \{u_\gamma\}_{\gamma>0},\ \{\frac{1}{\gamma}\|\lambda_\gamma\|^2_{L^2(\omega_0)}\}_{\gamma>0},\ \{\|y_\gamma\|_{W^{1,q}}\}_{\gamma>0}$$

are bounded.

Thus there exists a sequence $\{\gamma_k\}$ and $(\hat{c}, \hat{u}, \hat{y}) \in \mathbb{R} \times \mathbb{R}^m \times W_0^{1,q}$ such that $(c_{\gamma_k}, u_{\gamma_k}, y_{\gamma_k})$ converges to $(\hat{c}, \hat{u}, \hat{y})$. Taking the limit in (11.2.3) and in $Ay_{\gamma_k} - f - Bu_{\gamma_k} = 0$ we find that $(\hat{c}, \hat{u}, \hat{y})$ coincides with the unique solution (c^*, u^*, y^*) of ($\mathcal{P}$). Due to uniqueness of (c^*, u^*, y^*) the whole family $(c_\gamma, u_\gamma, y_\gamma)$ converges in $\mathbb{R} \times \mathbb{R}^m \times W^{1,q}(\Omega)$ to (c^*, u^*, y^*). Taking the limit in (11.2.3) implies (11.2.2). □

11.3 OPTIMALITY SYSTEM

In this section we derive the optimality systems for ($\mathcal{P}$) and ($\mathcal{P}_\gamma$) and the relationship between them.

We introduce the Lagrangian for the regularized problem

$$L(u,c,y,p) = \frac{c^2}{2} + \frac{\alpha}{2}|u|_2^2 + \frac{\gamma}{2}\|\max(0, y|_{\omega_0} - (\beta_1 c + \psi_1))\|_{L^2}^2 + \frac{\gamma}{2}\|\min(0, y|_{\omega_0} + \beta_2 c - \psi_2)\|_{L^2}^2 + \langle p, Ay - f - Bu\rangle_{W_0^{1,q'}, W^{-1,q}},$$

where

$$L : \mathbb{R}^m \times \mathbb{R} \times W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega) \to \mathbb{R}, \qquad \frac{1}{q} + \frac{1}{q'} = 1.$$

Since the linearized equality constraint in ($\mathcal{P}$) given by

$$(\bar{u}, \bar{y}) \mapsto A\bar{y} - B\bar{u}$$

is surjective, the necessary and sufficient optimality system for ($\mathcal{P}_\gamma$) is found to be

$$\begin{cases} \alpha u_{\gamma,i} - \langle p_\gamma, \chi_{\omega_i}\rangle = 0, \quad i = 1, \dots, m \\ c_\gamma - \langle \lambda_{\gamma,1}, \beta_1\rangle + \langle \lambda_{\gamma,2}, \beta_2\rangle = 0, \\ A^* p_\gamma + \tilde{\lambda}_\gamma = 0, \\ Ay_\gamma - f - Bu_\gamma = 0 \end{cases} \tag{11.3.1}$$

where $\tilde{\lambda}_\gamma$ denotes the extension by zero to $\Omega \setminus \omega_0$ of λ_γ.

Theorem 11.3.1 (Optimality system for ($\mathcal{P}$)). *There exist* $\lambda_i \in L^\infty(\omega_0)^*$, $i = 1, 2$, *and* $p^* \in W_0^{1,q'}(\Omega)$ *such that*

$$\begin{cases} \alpha u_i^* - \langle p^*, \chi_{\omega_i}\rangle = 0, \ i = 1, \dots, m \\ c^* - \langle \lambda_1, \beta_1\rangle_{L^{\infty *}, L^\infty} + \langle \lambda_2, \beta_2\rangle_{L^{\infty *}, L^\infty} = 0, \\ \langle p^*, A\varphi\rangle + \langle \lambda_1 + \lambda_2, \varphi|_{\omega_0}\rangle_{L^{\infty *}, L^\infty} = 0, \text{ for all } \varphi \in W_0^{1,q}(\Omega) \\ Ay^* - f - Bu^* = 0, \\ \langle \lambda_1, y^*|_{\omega_0} - (\beta_1 c^* + \psi_1)\rangle_{L^{\infty *}, L^\infty} = 0, \ \langle \lambda_2, y^*|_{\omega_0} + (\beta_2 c^* - \psi_2\rangle_{L^{\infty *}, L^\infty} = 0, \\ \langle \lambda_1, \varphi\rangle_{L^{\infty *}, L^\infty} \geqslant 0, \ \langle \lambda_2, \varphi\rangle_{L^{\infty *}, L^\infty} \leqslant 0, \text{ for all } \varphi \in L^\infty(\Omega). \text{ with } \varphi \geqslant 0. \end{cases} \tag{11.3.2}$$

Moreover $\{p_\gamma, \lambda_\gamma\}_{\gamma>0}$ *is bounded in* $W_0^{1,q'}(\Omega) \times L^1(\omega_0)$ *and for every subsequence such that* p_{γ_k} *converges weakly in* $W_0^{1,q'}(\Omega)$ *and* λ_{γ_k} *converges weakly* in* $(L^\infty(\omega_0))^*$ *the subsequential limits satisfy* (11.3.2).

Proof. Let $G : \mathbb{R}^m \times \mathbb{R} \to L^\infty(\omega_0) \times L^\infty(\omega_0)$ be defined by

$$G(u, c) = \begin{pmatrix} y(u)|_{\omega_0} - \beta_1 c - \psi_1 \\ -y(u)|_{\omega_0} - \beta_2 c + \psi_2 \end{pmatrix},$$

and

$$K = \{k \in L^\infty(\omega_0) : k \leqslant 0\}.$$

Then ($\mathcal{P}$) can be expressed in abstract form as

$$\min_{u \in \mathbb{R}^m, c \in \mathbb{R}} J(u, c) = \frac{1}{2} c^2 + \frac{\alpha}{2} \|u\|^2 \qquad \text{subject to } G(u, c) \in K \times K. \tag{11.3.3}$$

The regular point condition for (11.3.3) (cf. [Maurer and Zowe 1979; Ito and Kunisch 2008]) is given by

$$0 \in \left\{ G'(u^*, c^*)(\mathbb{R}^m \times \mathbb{R}) + G(u^*, c^*) - (K \times K) \right\}. \tag{11.3.4}$$

To verify (11.3.4) we consider for arbitrary $(g_1, g_2) \in L^\infty(\omega_0) \times L^\infty(\omega_0)$

$$\begin{pmatrix} A^{-1}(Bu)|_{\omega_0} - \beta_1 c \\ -A^{-1}(Bu)|_{\omega_0} - \beta_2 c \end{pmatrix} + \begin{pmatrix} y^*|_{\omega_0} - \beta_1 c^* - \psi_1 \\ -y^*|_{\omega_0} - \beta_2 c^* + \psi_2 \end{pmatrix} - \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}. \tag{11.3.5}$$

Set $u = 0$ and

$$c = \max\left(\beta_1^{-1} \operatorname{ess\,sup}(-g_1), \beta_2^{-1} \operatorname{ess\,sup}(-g_2), 0\right).$$

Then the first coordinate in (11.3.5) is satisfied with

$$k_1 = -g_1 + y^*|_{\omega_0} - \beta_1 c^* - \psi_1 - \beta_1 c \leqslant 0.$$

Similarly for the second coordinate we have

$$k_2 = -g_2 - y^*|_{\omega_0} - \beta_2 c^* + \psi_2 - \beta_2 c \leqslant 0.$$

Hence there exist $(\lambda_1, \lambda_2) \in L^\infty(\omega_0)^* \times L^\infty(\omega_0)^*$ such that the last two lines in (11.3.2) hold.

For later reference we specify that the above argument implies that

(11.3.6)
$$\begin{cases} \text{for all } \rho > 0 \text{ there exists } M \geqslant 0 \text{ such that for all } g = (g_1, g_2) \in B_\rho \\ \text{there exists } (u, c, k_1, k_2) \text{ satisfying} \\ G'(u^*, c^*)(u, c) + G(u^*, c^*) - \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \text{ and} \\ |(u, c, k_1 - y^*|_{\omega_0} + \beta_1 c^* + \psi_1, k_2 + y^*|_{\omega_0} + \beta_2 c^* - \psi_2)|_{\mathbb{R}^m \times \mathbb{R} \times L^\infty \times L^\infty} \leqslant M \rho, \end{cases}$$

where $B_\rho = \{(g_1, g_2) \in L^\infty \times L^\infty : \|g_i\|_{L^\infty} \leqslant \rho\}$. In fact, $u = 0$ is possible.

We also have

$$J'(u^*, c^*) + \langle(\lambda_1, -\lambda_2), G'(u^*, c^*)\rangle_{(L^\infty \times L^\infty)^*, L^\infty \times L^\infty} = 0.$$

Exploiting this equality we find

$$c^* - \langle\lambda_1, \beta_1\rangle_{L^{\infty *}, L^\infty} + \langle\lambda_2, \beta_2\rangle_{L^{\infty *}, L^\infty} = 0,$$

which is the second equation in (11.3.2), and

$$\alpha u_i + \langle\lambda_1 + \lambda_2, A^{-1}(\chi_{\omega_i})|_{\omega_0}\rangle = 0, \text{ for all } i = 1, \dots, m = 0.$$

Introducing the adjoint state as the solution to

$$\langle p^*, A\varphi\rangle + \langle\chi_{\omega_0}, (\lambda_1 + \lambda_2)\varphi\rangle = 0, \text{ for all } \varphi \in C_0^\infty(\Omega)$$

provides the first and third equation in (11.3.2). Since $\lambda_1, \lambda_2 \in L^\infty(\omega_0)^*$ we have that $p^* \in W_0^{1,q'}(\Omega)$. This concludes the proof of the optimality system (11.3.2).

Next we argue that $\{\lambda_{\gamma,1}\}_{\gamma>0}$ and $\{\lambda_{\gamma,2}\}_{\gamma>0}$ are bounded in $L^1(\omega_0)$. For this purpose we set

$$\vec{\lambda}_\gamma = (\lambda_{\gamma,1}, -\lambda_{\gamma,2})^T = (\gamma \max(0, y_\gamma|_{\omega_0} - (\beta_1 c + \psi_1), -\gamma \min(0, y|_{\omega_0} + \beta_2 c - \psi_2))^T.$$

For $\rho > 0$ fixed choose $g = (g_1, g_2) \in B_\rho$ arbitrarily. Appealing to (11.3.6) for $-g$, there exists $(\tilde{u}, \tilde{c}, k)$, with $k = (k_1, k_2) \in K \times K$, such that for $(u, c) = (\tilde{u} + u^*, \tilde{c} + c^*)$,

$$-g = \big(G'(u^*, c^*)((u, c) - (u^*, c^*))\big) - (k - G(u^*, c^*))$$

holds. Taking the inner product with $\vec{\lambda}_\gamma$ we have

$$\begin{aligned}
-\langle g, \vec{\lambda}_\gamma\rangle &= \big\langle G'(u_\gamma, c_\gamma)((u, c) - (u^*, c^*)), \vec{\lambda}_\gamma\big\rangle - \big\langle k - G(u^*, c^*)), \vec{\lambda}_\gamma\big\rangle \\
&= \big\langle G'(u_\gamma, c_\gamma)((u, c) - (u_\gamma, c_\gamma)), \vec{\lambda}_\gamma\big\rangle + \big\langle G'(u_\gamma, c_\gamma)((u_\gamma, c_\gamma) - (u^*, c^*)), \vec{\lambda}_\gamma\big\rangle \\
&\quad - \big\langle k, \vec{\lambda}_\gamma\big\rangle + \big\langle G(u^*, c^*) - G(u_\gamma, c_\gamma), \vec{\lambda}_\gamma\big\rangle + \big\langle G(u_\gamma, c_\gamma), \vec{\lambda}_\gamma\big\rangle.
\end{aligned}$$

Note that $\langle k, \vec{\lambda}_\gamma\rangle \leqslant 0$, $\langle G(u_\gamma, c_\gamma), \vec{\lambda}_\gamma\rangle \leqslant 0$ and $J'(u_\gamma, c_\gamma) + G'(u_\gamma, \lambda_\gamma)^* \vec{\lambda}_\gamma = 0$. Therefore

$$\begin{aligned}
\langle g, \vec{\lambda}_\gamma\rangle \leqslant \big\langle J'(u_\gamma, c_\gamma), (u, c) - (u_\gamma, c_\gamma)\big\rangle \\
- \big\langle G'(u_\gamma, c_\gamma)((u_\gamma, c_\gamma) - (u^*, c^*)), \vec{\lambda}_\gamma\big\rangle - \big\langle G(u^*, c^*) - G(u_\gamma, c_\gamma), \vec{\lambda}_\gamma\big\rangle.
\end{aligned}$$

Hence by (11.2.3) and (11.3.6) there exists $\tilde{M}$ independent of γ and (u, c) such that

$$\sup_{g \in B_\rho} \langle g, \vec{\lambda}_\gamma\rangle \leqslant \tilde{M}\,(1 + \|\vec{\lambda}_\gamma\|_{L^1 \times L^1}\, \|(u^*, c^*) - (u_\gamma, c_\gamma)\|).$$

Using (11.2.3) once again there exists $\hat{M}$ independent of γ such that

$$\sup_{g\in B_\rho} \langle g, \vec{\lambda}_\gamma\rangle \leqslant \tilde{M}\,(1 + \|\vec{\lambda}_\gamma\|_{L^1\times L^1}).$$

This implies boundedness of $\{\|\lambda_{\gamma,1}\|_{L^1}\}_{\gamma>0}$ and $\{\|\lambda_{\gamma,2}\|_{L^1}\}_{\gamma>0}$. It follows that the sequence $\{\|p_\gamma\|_{W_0^{1,q'}}\}_{\gamma>0}$ is bounded as well. From Proposition 11.2.3 we recall that $(c_\gamma, u_\gamma, y_\gamma) \to (c^*, u^*, y^*)$ in $\mathbb{R}\times\mathbb{R}^m\times W^{1,q}(\Omega)$, and in particular $y_\gamma \to y^*$ in $C(\bar\Omega)$. Since $W_0^{1,q'}(\Omega)$ embeds compactly into $L^r(\Omega)\subset L^1(\Omega)$ for $r = \frac{\frac{nq}{q-1}}{n-\frac{q}{q-1}} = \frac{nq}{nq-(n+q)}$, there exist subsequences of $p_\gamma, \lambda_{\gamma_1}, \lambda_{\gamma_2}$, denoted by the same symbols, and $p^*, \lambda_1, \lambda_2$ such that $p_\gamma \to p^*$ weakly in $W^{1,q'}(\Omega)$ and strongly in $L^1(\Omega)$, and $\lambda_{\gamma,1}, \lambda_{\gamma,2}$ converge to λ_1, λ_2 weakly* in $L^\infty(\omega_0)^*$.

From (11.2.2) we have $0 = \lim_{\gamma\to\infty} \frac{1}{\gamma}\|\lambda_\gamma\|^2_{L^2(\omega_0)}$ and hence

$$0 = \lim_{\gamma\to\infty} \frac{1}{\gamma}\|\lambda_{\gamma,1}\|^2_{L^2(\omega_0)}, \qquad 0 = \lim_{\gamma\to\infty} \frac{1}{\gamma}\|\lambda_{\gamma,2}\|^2_{L^2(\omega_0)}.$$

Consequently

$$\begin{aligned} 0 &= \lim_{\gamma\to\infty} \langle \lambda_{\gamma,1}, \max(0, y_\gamma|_{\omega_0} - (\beta_1 c_\gamma + \psi))\rangle \\ &= \lim_{\gamma\to\infty} \langle \lambda_{\gamma,1}, y_\gamma|_{\omega_0} - (\beta_1 c_\gamma + \psi)\rangle = \langle \lambda_1, y^*|_{\omega_0} - (\beta_1 c^* + \psi_1)\rangle_{L^{\infty *}, L^\infty}. \end{aligned}$$

In a similar way one shows that $0 = \langle \lambda_2, y^*|_{\omega_0} + (\beta_2 c^* - \psi_2)\rangle_{L^{\infty *}, L^\infty}$. This gives the fifth line of the optimality system (11.3.2). The remaining properties of the optimality system (11.3.2) can easily be obtained by passing to the limit in (11.3.1). □

Remark 11.3.2. The regularity of the Lagrange multipliers $\lambda_1, \lambda_2 \in L^\infty(\omega_0)^*$ is associated to the fact that the obstacle functions ψ_1, ψ_2 are in $L^\infty(\omega_0)$. If ψ_1, ψ_2 are taken in $C(\overline{\omega}_0)$, the above proof can be adapted to show that λ_1, λ_2 are bounded Radon measures in $C(\overline{\omega}_0)^*$.

Since the problem ($\mathcal{P}$) is strictly convex, the system (11.3.2) provides a sufficient optimality condition. In particular, the $(\bar u, \bar c, \bar y)$ satisfying (11.3.2) are unique.

In the final result of this section we address the question of rate of convergence of the regularized solutions to the solution of the original problem as $\gamma\to\infty$.

Proposition 11.3.3. *We have*

$$\frac{1}{2}|c_\gamma - c^*|^2 + \frac{\alpha}{2}|u_\gamma - u^*|^2_2 = \mathcal{O}\left(\frac{1}{\gamma^{\frac{1-\theta}{1+\theta}}}\right),$$

where $\theta = \frac{nq}{nq+2(q-n)}$.

Proof. Let $z^1_\gamma = y_\gamma|_{\omega_0} - (\beta_1 c_\gamma + \psi_1)$ and $z^2_\gamma = y_\gamma|_{\omega_0} + (\beta_2 c_\gamma - \psi_2)$. Due to optimality of $(c_\gamma, u_\gamma, y_\gamma)$ we find

$$\frac{(c_\gamma)^2}{2} + \frac{\alpha}{2}|u_\gamma|^2_2 + \frac{\gamma}{2}\|\max(0, z^1_\gamma)\|^2_{L^2} + \frac{\gamma}{2}\|\min(0, z^2_\gamma)\|^2_{L^2} \leqslant \frac{(c^*)^2}{2} + \frac{\alpha}{2}|u^*|^2_2\,. \tag{11.3.7}$$

We shall use the relationships

$$\begin{aligned}\frac{(c_\gamma)^2}{2} - \frac{(c^*)^2}{2} + \frac{\alpha}{2}|u_\gamma|^2_2 - \frac{\alpha}{2}|u^*|^2_2 \\ = (c_\gamma - c^*)c^* + \frac{1}{2}|c_\gamma - c^*|^2 + \alpha\langle u_\gamma - u^*, u^*\rangle + \frac{\alpha}{2}|u_\gamma - u^*|^2_2,\end{aligned} \tag{11.3.8}$$

and

$$\langle A(y_\gamma - y^*), p^*\rangle = \langle u_\gamma - u^*, B^*p^*\rangle = \alpha\langle u_\gamma - u^*, u^*\rangle, \tag{11.3.9}$$

and, using (11.3.2),

$$\begin{aligned}\langle A(y_\gamma - y^*), p^*\rangle &= -\langle \lambda_1, y_\gamma|_{\omega_0} - y^*|_{\omega_0}\rangle - \langle \lambda_2, y_\gamma|_{\omega_0} - y^*|_{\omega_0}\rangle \\ &= -\langle \lambda_1, y_\gamma|_{\omega_0} - (c_\gamma\beta_1 + \psi_1)\rangle - \beta_1\langle \lambda_1, c_\gamma - c^*\rangle \\ &\quad + \langle \lambda_1, y^*|_{\omega_0} - (c^*\beta_1 + \psi_1)\rangle \\ &\quad - \langle \lambda_2, y_\gamma|_{\omega_0} + (c_\gamma\beta_2 - \psi_2)\rangle + \beta_2\langle \lambda_2, c_\gamma - c^*\rangle \\ &\quad + \langle \lambda_2, y^*|_{\omega_0} + (c^*\beta_2 - \psi_2)\rangle \\ &= -\langle \lambda_1, z^1_\gamma\rangle - \langle \lambda_2, z^2_\gamma\rangle - c^*(c_\gamma - c^*).\end{aligned} \tag{11.3.10}$$

Combining (11.3.7), (11.3.8), (11.3.9), and (11.3.10), we obtain

$$\begin{aligned}\frac{1}{2}|c_\gamma - c^*|^2 + \frac{\alpha}{2}|u_\gamma - u^*|^2 &\leqslant -(c_\gamma - c^*)c^* - \alpha\langle u_\gamma - u^*, u^*\rangle \\ &\quad - \frac{\gamma}{2}\|\max(0, z^1_\gamma)\|^2_{L^2} - \frac{\gamma}{2}\|\min(0, z^2_\gamma)\|^2_{L^2} \\ &= \langle \lambda_1, z^1_\gamma\rangle + \langle \lambda_2, z^2_\gamma\rangle \\ &\quad - \frac{\gamma}{2}\|\max(0, z^1_\gamma)\|^2_{L^2} - \frac{\gamma}{2}\|\min(0, z^2_\gamma)\|^2_{L^2} \\ &\leqslant \|\lambda_1\|_{(L^\infty)^*}\|\max(0, z^1_\gamma)\|_{L^\infty} - \frac{\gamma}{2}\|\max(0, z^1_\gamma)\|^2_{L^2} \\ &\quad + \|\lambda_2\|_{(L^\infty)^*}\|\min(0, z^2_\gamma)\|_{L^\infty} - \frac{\gamma}{2}\|\min(0, z^2_\gamma)\|^2_{L^2}\end{aligned} \tag{11.3.11}$$

In the following estimates K denotes a generic constant, which is independent of $\alpha > 0$ and $\gamma > 0$. We use the estimate

$$\|\hat{z}\|_{L^\infty} \leqslant K\|\hat{z}\|^\theta_{W^{1,q}}\,\|\hat{z}\|^{1-\theta}_{L^2},$$

for $\hat{z} = \max(0, z_\gamma^1)$, where $\theta = \frac{nq}{nq+2(q-n)}$, see e.g. [Adams and Fournier 2003, p. 141]. This further implies that

$$\|\hat{z}\|_{L^\infty} \leqslant K \left(\frac{\|\lambda_1\|_{(L^\infty)^*}}{\gamma} \right)^{\frac{1-\theta}{2}} \|\hat{z}\|_{W^{1,q}}^{\theta} \left(\frac{\gamma}{\|\lambda_1\|_{(L^\infty)^*}} \right)^{\frac{1-\theta}{2}} \|\hat{z}\|_{L^2}^{1-\theta},$$

and

$$\|\hat{z}\|_{L^\infty} \leqslant K \left(\frac{\|\lambda_1\|_{(L^\infty)^*}}{\gamma} \right)^{\frac{1-\theta}{1+\theta}} \|\hat{z}\|_{W^{1,q}}^{\frac{2\theta}{1+\theta}} + \frac{\gamma}{2\|\lambda_1\|_{(L^\infty)^*}} \|\hat{z}\|_{L^2}^2,$$

where we use that $\frac{1-\theta}{2} + \frac{1+\theta}{2} = 1$. Arguing similarly for $\hat{z} = \min(0, z_\gamma^2)$ and combining these estimates with (11.3.11) implies that

$$\frac{1}{2}|c_\gamma - c^*|^2 + \frac{\alpha}{2}|u_\gamma - u^*|_2^2 \leqslant K \frac{1}{\gamma^{\frac{1-\theta}{1+\theta}}} \left(\|\lambda_1\|_{(L^\infty)^*}^{\frac{2}{1+\theta}} \|z_\gamma^1\|_{W^{1,q}}^{\frac{2\theta}{1+\theta}} + \|\lambda_2\|_{(L^\infty)^*}^{\frac{2}{1+\theta}} \|z_\gamma^2\|_{W^{1,q}}^{\frac{2\theta}{1+\theta}} \right).$$

Since $\{y_\gamma\}_{\gamma \geqslant 1}$ is bounded in $W^{1,q}(\Omega)$ this implies the claim. □

In the case $n = 2$ we find that $\theta = \frac{q}{2q-4}$ and so we obtain the convergence rate $\mathcal{O}(\frac{1}{\gamma^{1/3-\epsilon}})$ for any $\epsilon > 0$, provided that q is sufficiently large. For $n = 2$, using solutions $y_\gamma \in H^2(\Omega)$, the above proof can also be adapted to obtain the rate $\mathcal{O}(\gamma^{-1/3})$.

11.4 SEMISMOOTH NEWTON METHOD

This section is devoted to the solution of the optimality system (11.3.1) by a semismooth Newton method. For this purpose we define

$$F : \mathbb{R}^m \times \mathbb{R} \times W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega) \to \mathbb{R}^m \times \mathbb{R} \times W^{-1,q'}(\Omega) \times W^{-1,q}(\Omega)$$

by

$$F(u, c, y, p) = \begin{pmatrix} \alpha u - \langle p, \vec{\chi}_\omega \rangle \\ c - \gamma \langle \beta_1, \max(0, y|_{\omega_0} - (\beta_1 c + \psi_1)) \rangle + \gamma \langle \beta_2, \min(0, y|_{\omega_0} + \beta_2 c - \psi_2) \rangle \\ \gamma \widetilde{\max}(0, y|_{\omega_0} - (\beta_1 c + \psi_1)) + \gamma \widetilde{\min}(0, y|_{\omega_0} + \beta_2 c - \psi_2) + A^* p \\ Ay - Bu \end{pmatrix}, \tag{11.4.1}$$

where we have set

$$\langle \vec{\chi}_\omega, \tilde{p} \rangle = (\langle \chi_{\omega_1}, \tilde{p} \rangle, \dots, \langle \chi_{\omega_m}, \tilde{p} \rangle)^T,$$

and $\widetilde{\max}$, $\widetilde{\min}$ denote extensions by zero from ω_0 to $\Omega \setminus \omega_0$. Recall from [Ito and Kunisch 2008] that $z \mapsto \max(0, z)$ is Newton differentiable from $L^{p_1}(\Omega) \to L^{p_2}(\Omega)$ provided that $1 \leqslant p_1 < p_2 \leqslant \infty$ with Newton derivative given in the a.e. sense by

$$D \max(0, z) = \begin{cases} 1, & \text{if } z(x) \geqslant 0 \\ 0, & \text{if } z(x) < 0. \end{cases}$$

An analogous statement holds for the min operation. It follows that

$$G_1(y, c) = \widetilde{\max}(0, y|_{\omega_0} - (\beta_1 c + \psi_1))$$

is Newton differentiable for fixed c from $W_0^{1,q}(\Omega) \to W^{-1,q'}(\Omega)$ with Newton derivative with respect to y given by

$$D_y G_1(y, c)\, \bar{y} = \bar{\chi}\, \bar{y},$$

where $\bar{\chi}$ is given in the a.e. sense by

$$\bar{\chi} = \bar{\chi}(y, c) = \begin{cases} 1, & \text{if } x \in \omega_0 \text{ and } y|_{\omega_0}(x) - (\beta_1 c + \psi_1(x)) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11.4.2}$$

Analogously

$$G_2(y, c) = \widetilde{\min}(0, y|_{\omega_0} + \beta_2 c - \psi_2)$$

is Newton differentiable for fixed c from $W_0^{1,q}(\Omega) \to W^{-1,q'}(\Omega)$ with Newton derivative with respect to y given by

$$D_y G_2(y, c)\, \bar{y} = \underline{\chi}\, \bar{y},$$

where $\underline{\chi}$ is given in the a.e. sense by

$$\underline{\chi} = \underline{\chi}(y, c) = \begin{cases} 1, & \text{if } x \in \omega_0 \text{ and } y|_{\omega_0}(x) + \beta_2 c - \psi_2(x) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11.4.3}$$

Let us also consider the mappings $H_1, H_2 : W_0^{1,q}(\Omega) \times \mathbb{R} \to \mathbb{R}$, defined by

$$\begin{aligned} H_1(y, c) &= \langle \beta_1, \max(0, y|_{\omega_0} - (\beta_1 c + \psi_1)) \rangle_{L^2(\omega_0)} \\ &= \langle \beta_1, \widetilde{\max}(0, y|_{\omega_0} - (\beta_1 c + \psi_1)) \rangle_{L^2(\Omega)} \end{aligned}$$

and

$$H_2(y, c) = \langle \beta_2, \widetilde{\min}(0, y|_{\omega_0} + (\beta_2 c - \psi_2)) \rangle_{L^2(\Omega)}.$$

Their Newton derivatives with respect to y are found to be

$$D_y H_1(y,c)\,\bar{y} = \beta_1\langle \bar{\chi}, \bar{y}\rangle, \qquad D_y H_2(y,c)\,\bar{y} = \beta_2\langle \underline{\chi}, \bar{y}\rangle.$$

Similarly, we obtain

$$D_c G_1(y,c)\bar{c} = \beta_1 \bar{\chi}\bar{c},$$

and

$$D_c H_1(y,c)\bar{c} = \beta_1^2\langle \bar{\chi}\rangle \bar{c},$$

where $\langle z\rangle = \int_\Omega z\,dx$. One proceeds analogously for $D_c G_2$ and $D_c H_2$.

All together the Newton derivative of F at an arbitrary point $(u,c,y,p) \in \mathbb{R}^{m+1} \times W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega)$ is given by

(11.4.4)
$$DF(u,c,y,p)(\bar{u},\bar{c},\bar{y},\bar{p}) = \begin{pmatrix} \alpha\bar{u} - \langle \bar{p}, \vec{\chi}_\omega\rangle \\ (1+\gamma\beta_1^2\langle\bar{\chi}\rangle + \gamma\beta_2^2\langle\underline{\chi}\rangle)\bar{c} - \gamma\beta_1\langle\bar{\chi},\bar{y}\rangle + \gamma\beta_2\langle\underline{\chi},\bar{y}\rangle \\ -\gamma\beta_1\bar{\chi}\bar{c} + \gamma\beta_2\underline{\chi}\bar{c} + \gamma\bar{\chi}\bar{y} + \gamma\underline{\chi}\bar{y} + A^*\bar{p} \\ A\bar{y} - B\bar{u} \end{pmatrix},$$

where we have assumed that $c > 0$ holds. A semismooth Newton step is given by

$$DF(u,c,y,p)(\bar{u},\bar{c},\bar{y},\bar{p}) = -F(u,c,y,p).$$

We next address its well-posedness.

Proposition 11.4.1. *For each $(u,c,y,p) \in \mathbb{R}^m \times \mathbb{R} \times W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega)$, the mapping $DF(u,c,y,p)$ is invertible, and there exists a constant $C > 0$ independent of (u,c,y,p) such that*

$$\|DF(u,c,y,p)^{-1}\|_{\mathcal{L}(\mathbb{R}^{m+1}\times W^{-1,q'}\times W^{-1,q},\,\mathbb{R}^{m+1}\times W_0^{1,q}\times W_0^{1,q'})} \leqslant C.$$

Proof. For $w = (w_1,\dots,w_4)^T \in \mathbb{R}^m \times \mathbb{R} \times W^{-1,q'}(\Omega) \times W^{-1,q}(\Omega)$ we consider the equation $DF(u,c,y,p)(\bar{u},\bar{c},\bar{y},\bar{p}) = w$, i.e.,

(11.4.5)
$$\begin{cases} \alpha\bar{u} - \langle \vec{\chi}_\omega, \bar{p}\rangle & = w_1 \\ (1+\gamma\beta_1^2\langle\bar{\chi}\rangle + \gamma\beta_2^2\langle\underline{\chi}\rangle)\bar{c} - \gamma\beta_1\langle\bar{\chi},\bar{y}\rangle + \gamma\beta_2\langle\underline{\chi},\bar{y}\rangle & = w_2 \\ -\gamma\beta_1\bar{\chi}\bar{c} + \gamma\beta_2\underline{\chi}\bar{c} + \gamma\bar{\chi}\bar{y} + \gamma\underline{\chi}\bar{y} + A^*\bar{p} & = w_3 \\ -B\bar{u} + A\bar{y} & = w_4. \end{cases}$$

Therefore from the third and fourth equation in (11.4.5), we obtain

$$\begin{aligned}\bar{y} &= A^{-1}(B\bar{u}) + A^{-1}w_4 = A^{-1}(B\bar{u}) + r_1,\\ \bar{p} &= A^{-*}\big(w_3 + \gamma\bar{c}(\beta_1\bar{\chi} - \beta_2\underline{\chi}) - \gamma\bar{y}(\bar{\chi}+\underline{\chi})\big)\\ &= A^{-*}\big(w_3 + \gamma\bar{c}(\beta_1\bar{\chi} - \beta_2\underline{\chi}) - \gamma A^{-1}(B\bar{u})(\bar{\chi}+\underline{\chi}) - \gamma(\bar{\chi}+\underline{\chi})\rangle A^{-1}w_4\\ &= \gamma\bar{c}A^{-*}(\beta_1\bar{\chi} - \beta_2\underline{\chi}) - \gamma A^{-*}((\bar{\chi}+\underline{\chi})A^{-1}(Bu)) + r_2,\end{aligned}$$

where

$$r_1 = A^{-1}w_4 \in W_0^{1,q}(\Omega) \qquad r_2 = A^{-*}(w_3 - \gamma(\bar{\chi}+\underline{\chi})A^{-1}w_4) \in W_0^{1,q'}(\Omega).$$

Using these representations for $\bar{y}, \bar{p}$ in the first two equations of (11.4.5) we obtain for $\bar{u}, \bar{c}$ and $i = 1, \dots, m$

$$\begin{cases} \alpha\bar{u}_i - \gamma\bar{c}\langle\chi_{\omega_i}, A^{-*}(\beta_1\bar{\chi} - \beta_2\underline{\chi})\rangle + \gamma\displaystyle\sum_{j=1}^{m}\bar{u}_j\langle(A^{-1}\chi_{\omega_i})(\bar{\chi}+\underline{\chi}), A^{-1}\chi_{\omega_j}\rangle - \langle\chi_{\omega_i}, r_2\rangle \\ \qquad\qquad = w_{1,i},\\ (1 + \gamma(\beta_1^2\langle\bar{\chi}\rangle + \beta_2^2\langle\underline{\chi}\rangle))\bar{c} - \gamma\displaystyle\sum_{j=1}^{m}\bar{u}_j\langle\beta_1\bar{\chi} - \beta_2\underline{\chi}, A^{-1}\chi_{\omega_j}\rangle - \gamma\langle\beta_1\bar{\chi} - \beta_2\underline{\chi}, r_1\rangle = w_2,\end{cases}$$

equivalently in matrix form

(11.4.6)
$$\begin{pmatrix} \alpha I + \gamma\langle\vec{\psi}, (\bar{\chi}+\underline{\chi})\vec{\psi}\rangle & -\gamma\langle\vec{\psi}, \beta_1\bar{\chi} - \beta_2\underline{\chi}\rangle \\ -\gamma\langle\beta_1\bar{\chi} - \beta_2\underline{\chi}, \vec{\psi}\rangle & 1 + \gamma(\beta_1^2\langle\bar{\chi}\rangle + \beta_2^2\langle\underline{\chi}\rangle)\end{pmatrix}\begin{pmatrix}\bar{u}\\ \bar{c}\end{pmatrix} = \begin{pmatrix} w_1 + \langle\vec{\chi}_\omega, r_2\rangle \\ \gamma\langle\beta_1\bar{\chi} - \beta_2\underline{\chi}, r_1\rangle + w_2\end{pmatrix},$$

where I is the $m \times m$ identity matrix,

$$\psi_i = A^{-1}\chi_{\omega_i}, \quad \vec{\psi} = (\psi_1, \dots, \psi_m)^T,$$

and

$$(\langle\vec{\psi}, (\bar{\chi}+\underline{\chi})\vec{\psi}\rangle)_{i,j} = \langle\vec{\psi}_i, (\bar{\chi}+\underline{\chi})\vec{\psi}_j\rangle.$$

The matrix on the left hand side of (11.4.6) can be expressed as

$$M = M_1 + \gamma M_2 + \gamma M_3,$$

where

$$M_1 = \begin{pmatrix}\alpha I & 0\\ 0 & 1\end{pmatrix},\ M_2 = \begin{pmatrix}\langle\vec{\psi}, \bar{\chi}\vec{\psi}\rangle & -\beta_1\langle\vec{\psi}, \bar{\chi}\rangle\\ -\beta_1\langle\vec{\psi}, \bar{\chi}\rangle & \beta_1^2\langle\bar{\chi}\rangle\end{pmatrix},\ M_3 = \begin{pmatrix}\langle\vec{\psi}, \underline{\chi}\vec{\psi}\rangle & \langle\beta_2\vec{\psi}, \underline{\chi}\rangle\\ \langle\beta_2\vec{\psi}, \underline{\chi}\rangle & \beta_2^2\langle\underline{\chi}\rangle\end{pmatrix}.$$

We next argue that the symmetric matrix M_2 is positive semi-definite. For $\langle\bar\chi\rangle = 0$, this is straight-forward. Henceforth assume that $\langle\bar\chi\rangle \neq 0$ and let $(\bar u, \bar c) \in \mathbb{R}^{m+1}$ be arbitrary.

Then we have, using that $2ab + b^2 \geqslant -a^2$,

$$\begin{aligned}(\bar u^T, \bar c) M_2 (\bar u^T, \bar c)^T &= \bar u^T \langle \vec\psi, \bar\chi\vec\psi\rangle\, \bar u - 2\beta_1 \bar c \langle \vec\psi, \bar\chi\rangle^T \bar u + \beta_1^2 \langle\bar\chi\rangle\, \bar c^2 \\ &\geqslant \bar u^T \langle \vec\psi, \bar\chi\vec\psi\rangle\, \bar u - \frac{1}{\langle\bar\chi\rangle} \bar u^T \langle\vec\psi, \bar\chi\rangle\, \langle\vec\psi, \bar\chi\rangle^T \bar u = \bar u^T \mathcal{M}_2 \bar u,\end{aligned}$$

where

$$(\mathcal{M}_2)_{i,j} = \langle \vec\psi_i, \bar\chi\vec\psi_j\rangle - \frac{1}{\langle\bar\chi\rangle}\langle\vec\psi_i, \bar\chi\rangle\langle\vec\psi_j, \bar\chi\rangle.$$

The diagonal elements of $\mathcal{M}_2$ are given by

$$(\mathcal{M}_2)_{i,i} = \langle\psi_i, \bar\chi\psi_i\rangle - \frac{1}{\langle\bar\chi\rangle}\langle\psi_i, \bar\chi\rangle^2 = \|(\psi_i - \frac{1}{\langle\bar\chi\rangle}\langle\psi_i, \bar\chi\rangle)\bar\chi\|_{L^2}^2,$$

and for the off-diagonal elements we find

$$(\mathcal{M}_2)_{i,j} = \langle\psi_i, \bar\chi\psi_j\rangle - \frac{1}{\langle\bar\chi\rangle}\langle\psi_i, \bar\chi\rangle\, \langle\psi_j, \bar\chi\rangle = \big\langle\, \psi_i - \frac{1}{\langle\bar\chi\rangle}\langle\psi_i, \bar\chi\rangle, (\psi_j - \frac{1}{\langle\bar\chi\rangle}\langle\psi_j, \bar\chi\rangle)\bar\chi\,\big\rangle.$$

Therefore we have

$$(\mathcal{M}_2)_{i,j} = \big\langle\, \psi_i - \frac{1}{\langle\bar\chi\rangle}\langle\psi_i, \bar\chi\rangle, (\psi_j - \frac{1}{\langle\bar\chi\rangle}\langle\psi_j, \bar\chi\rangle)\bar\chi\,\big\rangle.$$

Consequently $\mathcal{M}_2$ is a Gramian matrix and thus positive semi-definite, and we find that the same holds true for M_2. Analogously one argues that M_3 is positive semi-definite. All together we obtain

$$\|M^{-1}\|_{\mathbb{R}^{(m+1)\times(m+1)}} \leqslant \max(\frac{1}{\sqrt{\alpha}}, 1).$$

This estimate is independent of $\bar\chi, \underline{\chi}, \omega_i, \omega_0$.

Using (11.4.6) there exist constants C_1 and C_2 such that

$$|(\bar u, \bar c)|_{\mathbb{R}^{m+1}} \leqslant C_1 |w|_{\mathbb{R}^{m+1}\times W_0^{-1,q'}\times W_0^{-1,q}}$$

and

$$|(\bar y, \bar p)|_{W_0^{1,q}\times W_0^{1,q'}} \leqslant C_2 |w|_{\mathbb{R}^{m+1}\times W_0^{-1,q'}\times W_0^{-1,q}}$$

hold. This implies the claim. □

Algorithm 11.1 Semismooth Newton method

1: Choose x^0, γ^0, $\tau > 1$, $\varepsilon > 0$, k^*; set $j = 0$
2: **repeat**
3: Increment $j \leftarrow j + 1$
4: Set $x_0 = x^{j-1}$, $k = 0$
5: **repeat**
6: Increment $k \leftarrow k + 1$
7: Compute indicator function of active sets : $\bar{\chi}(y_{k-1}, c_{k-1})$, $\underline{\chi}(y_{k-1}, c_{k-1})$ from (11.4.2) and (11.4.3)
8: Solve for δx:

$$DF(x_{k-1})\delta x = -F(x_{x-1}),$$

where F and DF are given by (11.4.1) and (11.4.4), respectively
9: Update $x_k = x_{k-1} + \delta x$
10: **until** $\bar{\chi}(y_{k-1}, c_{k-1}) = \bar{\chi}(y_{k-2}, c_{k-2})$ and $\underline{\chi}(y_{k-1}, c_{k-1}) = \underline{\chi}(y_{k-2}, c_{k-2})$, or $k = k^*$
11: Set $x^j = x_k$
12: Set $\gamma^j = \tau\gamma^{j-1}$
13: **until** $\sup_{x\in\omega_0}(y - (\beta_1 c + \psi_1)) < \varepsilon$ and $\inf_{x\in\omega_0}(y + \beta_2 c - \psi_2) < \varepsilon$

Thus F is semismooth, and from standard results (e.g., [Ito and Kunisch 2008, Th. 8.16]) we deduce the following convergence result for the semismooth Newton method. For convenience we set

$$x = (u, c, y, p) \in \mathbb{R}^m \times \mathbb{R} \times W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega),$$

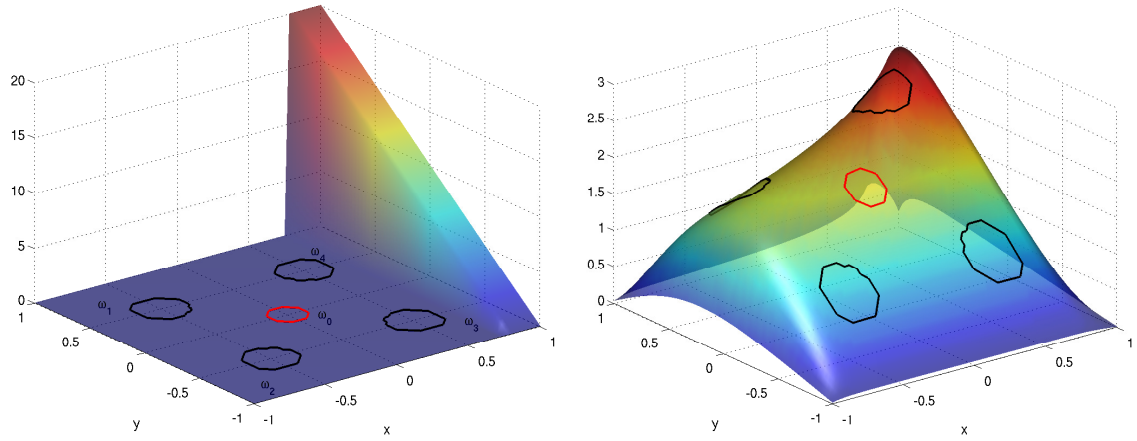
and similarly x_k, δx et cetera.

Theorem 11.4.2. *For each $\gamma > 0$ the iteration $DF(x_{k-1})(x_k - x_{k-1}) = -F(x_{k-1})$ converges superlinearly to $x_\gamma = (u_\gamma, c_\gamma, y_\gamma, p_\gamma)$ provided that x_0 is sufficiently close to x_γ.*

The full procedure for the solution of problem $(\mathcal{P}_\gamma)$ is given as Algorithm 11.1. Note that Algorithm 11.1 contains as inner loop the semismooth Newton method and as outer loop the increase of the penalty parameter γ. The convergence of these two processes was analyzed in Theorem 11.4.2 and Proposition 11.2.3 respectively. Here we choose a simple strategy for increasing γ. In related contexts [Hintermüller and Kunisch 2006] we proposed a path-following technique which could be adapted to the present situation. The termination criterion in step 10 is motivated by the following property of the semismooth Newton method.

Proposition 11.4.3. *If $\bar{\chi}(y_{k+1}, c_{k+1}) = \bar{\chi}(y_k, c_k)$ and $\underline{\chi}(y_{k+1}, c_{k+1}) = \underline{\chi}(y_k, c_k)$ holds, then x_{k+1} satisfies $F(x_{k+1}) = 0$.*

This can be verified by simple inspection, and is shown in [Ito and Kunisch 2008, Rem. 7.1.1].



(a) Geometric setup.

(b) Uncontrolled state y^0.

Figure 11.1: Model problem. The left plot shows the pollutant f, while the circles give the observation domain ω_0 (red) and the control domains $\omega_1, \ldots, \omega_4$ (black). The right plot shows the uncontrolled state $y^0 = A^{-1} f$.

11.5 NUMERICAL RESULTS

Here we give the results of some numerical tests for a model problem in two dimensions. The geometric situation is given in Figure 11.1a: The circular observation domain ω_0 (the "town") is situated in the center of the unit square $[-1, 1]^2$. On one side, a contaminant given by the function $f = 100(1 + y)\chi_{\{x>.75\}}$ enters the domain. Around the town, $m = 4$ control domains ("wells") are spaced equally. We consider convective-diffusive transport, which is described by the operator $Ay = -\nu\Delta y + b \cdot \nabla y$ with $\nu = 0.1$ and $b = (-1, 0)^T$ (i.e., transport parallel to the x-axis away from the source) with homogeneous Dirichlet conditions. The uncontrolled state y^0, which solves $Ay^0 = f$, is shown in Figure 11.1b.

The parameters were set to $x^0 = 0, \gamma^0 = 1, \tau = 0.1$, and $\varepsilon = 10^{-9}$. The penalty parameter was set to $\alpha = 10^{-6}$. The differential operators were discretized by finite differences with $N = 64$ grid points. We give results for $\psi_1 = \psi_2 = 0$. Since the convergence behavior is very similar in all test cases, we only show details for the motivating example of optimal L^∞-constraints, §11.5.3. A Matlab code implementing the algorithm for these examples can be downloaded from http://www.uni-graz.at/~clason/codes/mininvasion.zip.

11.5.1 UNILATERAL CONSTRAINT

We begin by considering the motivating example, which is the unilaterally constrained problem

$$\begin{cases} \min\limits_{c\in\mathbb{R},u\in\mathbb{R}^m} \dfrac{c^2}{2} + \dfrac{\alpha}{2}|u|_2^2 \\ \text{s.t. } Ay = f + Bu, \qquad y|_{\omega_0} \leqslant c. \end{cases}$$

The numerical solution can be obtained using the above algorithm by simply dropping the min terms and setting all corresponding active sets to zero. Algorithm 11.1 terminated at $\gamma^* = 10^6$, using at most 8 (for $\gamma = 1$) Newton iterations. The computed optimal control is $u^* = (-0.0418744, -0.037166, -25.3717, -29.2032)^T$ (shown in Fig. 11.2c), which results in a minimal norm bound $c^* = 8.33217 \cdot 10^{-4}$. Correspondingly, the maximum value of y^* on ω_0 is $8.33217 \cdot 10^{-4}$, while its minimum value is -0.565084 (cf. Figs. 11.2a, 11.2b). It can be seen that only the controls acting on the control domains located between the source and the observation domain are active. For completeness, we also show the Lagrange multiplier p^* for the pde constraint in Fig. 11.2d.

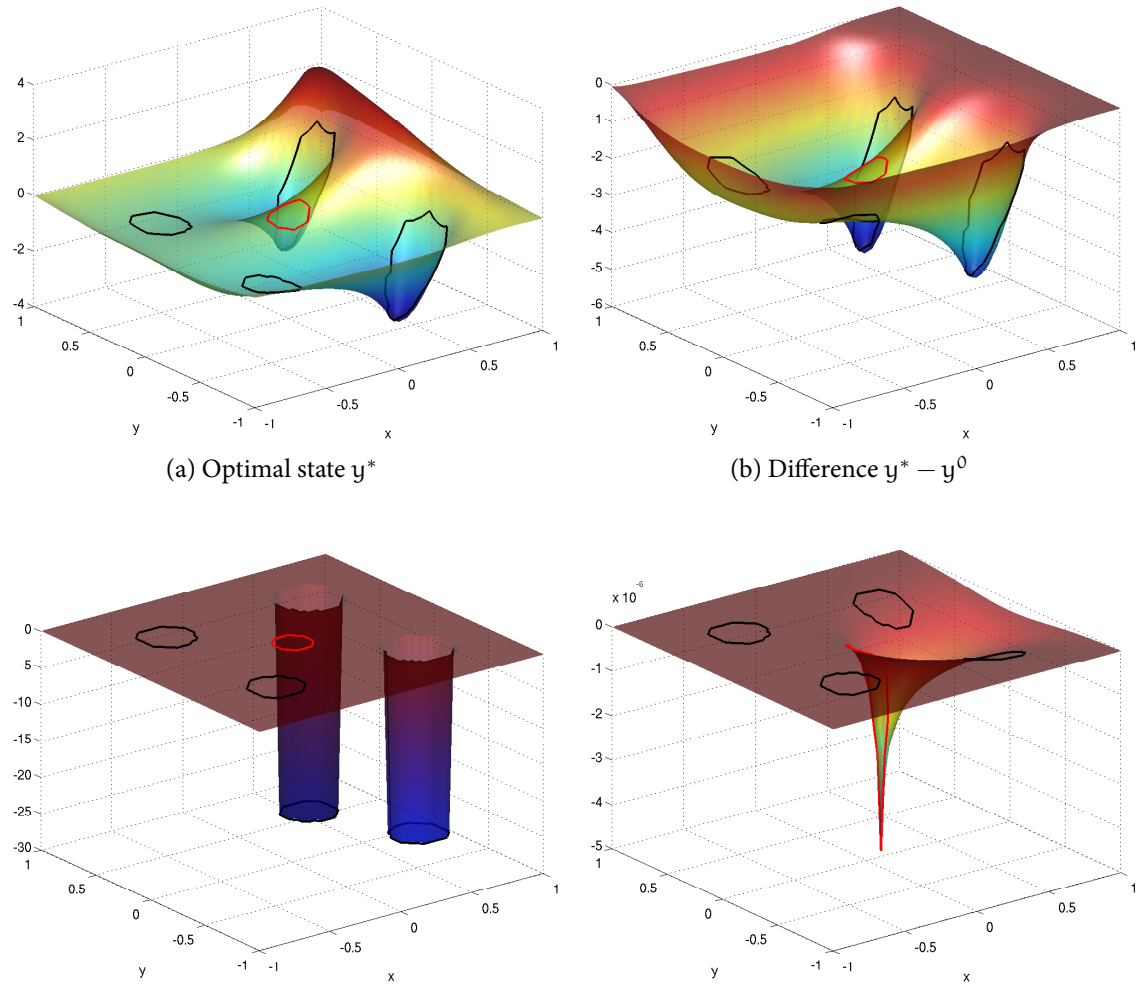
11.5.2 NON-NEGATIVITY CONSTRAINT

We next consider the case of minimizing an upper bound, while enforcing non-negativity of the state, i.e., we set $\beta_1 = 1$ and $\beta_2 = 0$. Again, Algorithm 11.1 terminated at $\gamma^* = 10^9$, after at most 7 Newton iterations. The computed optimal control is $u^* = (70.5931, 58.8345, -17.0403, -26.6347)^T$, which results in a minimal upper bound $c^* = 0.350723$ and identical maximal value of y^* on ω_0. The minimal value of y^* on ω_0 is $-1.67788 \cdot 10^{-10}$, within the prescribed tolerance of $\varepsilon = 10^{-9}$. Optimal state y^*, difference $y^* - y^0$, optimal control u^* and Lagrange multiplier p^* are shown in Figure 11.3.

We note that due to the non-negativity constraint, the controls near the contaminant inflow cannot act as strongly as in example 11.5.1, and that the optimal control is no longer uniformly negative. Thus the achievable upper bound is larger than in example 11.5.1.

11.5.3 L^∞ NORM CONSTRAINT

Finally, we consider the case $\beta_1 = \beta_2 = 1$, i.e., the L^∞ norm constraint problem (11.1.1). The iteration terminated at $\gamma^* = 10^9$, using at most 7 (for $c = 1$) Newton iterations. Table 11.1 shows the distance $d(\gamma) := \frac{1}{2}|c_\gamma - c_{\gamma^*}|^2 + \frac{\alpha}{2}|u_\gamma - u_{\gamma^*}|_2^2$, which indicates that the convergence rate proved in Proposition 11.3.3 is not optimal. For $\gamma = 1$, the norm of the residuals $|F(x_k)|_2$ in the semismooth Newton method is given in Table 11.2, verifying the locally superlinear convergence shown in Theorem 11.4.2. The computed optimal control is $u^* = (22.4538, 18.7443, -19.272, -28.7974)^T$ which results in a minimal norm bound



(a) Optimal state y^*

(b) Difference $y^* - y^0$

(c) Optimal control u^*

(d) Lagrange multiplier p^*

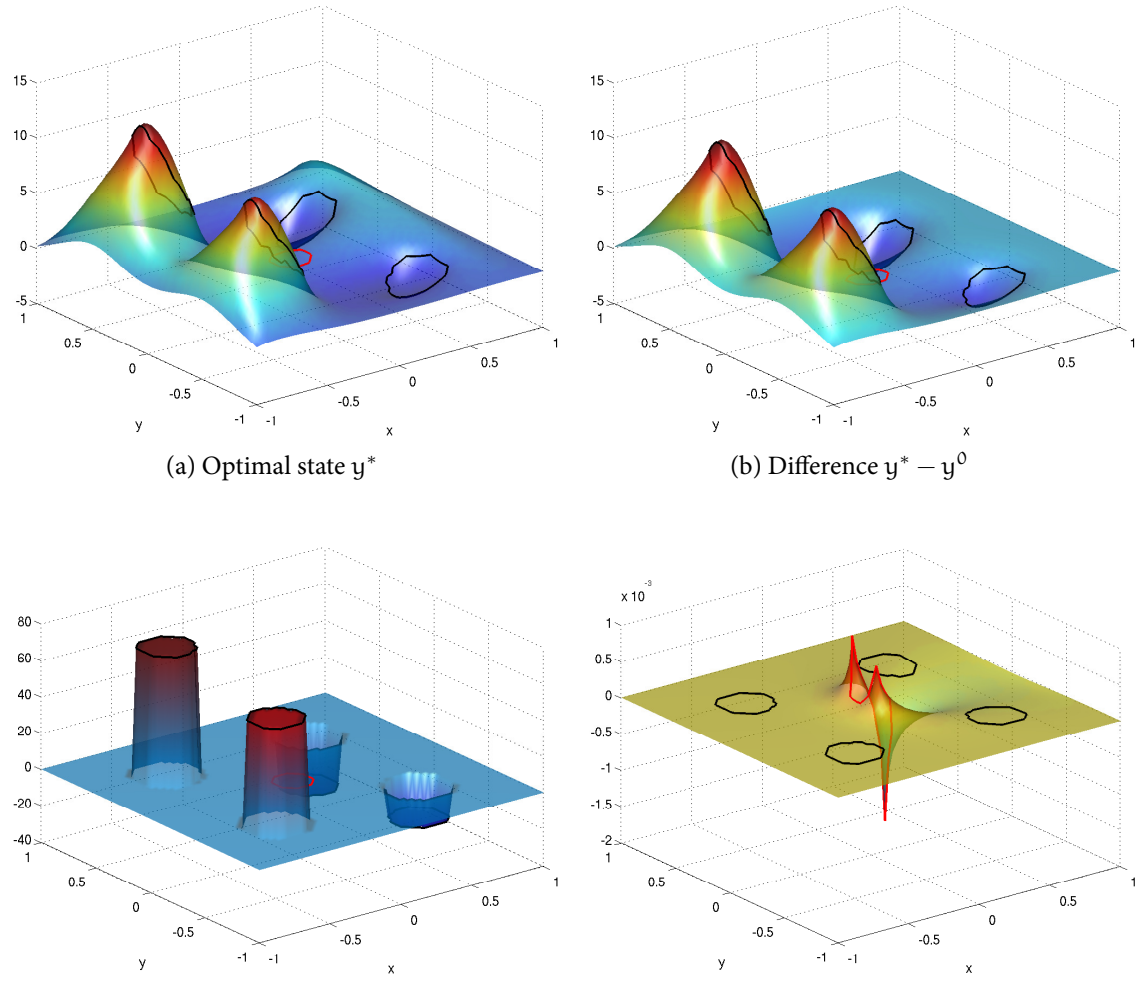
Figure 11.2: Results for unilateral constraint ($y|_{\omega_0} \leqslant c$).

Table 11.1: Convergence in γ. Shown are the distances $d(\gamma) := \frac{1}{2}|c_\gamma - c_{\gamma^*}|^2 + \frac{\alpha}{2}|u_\gamma - u_{\gamma^*}|^2_2$.

γ	1e0	1e1	1e2	1e3	1e4	1e5	1e6	1e7	1e8
$d(\gamma)$	5.11e-4	1.85e-5	3.34e-7	3.37e-9	3.37e-11	3.37e-13	3.37e-15	3.30e-17	2.73e-19

Table 11.2: Convergence of semismooth Newton method. Shown is the norm of the residual of (11.4.1) for the iterates x_k.

k	0	1	2	3	4	5	6
$\|F(x_k)\|_2$	2.62e+2	9.12e+1	1.34e+0	7.63e-1	3.26e-1	7.07e-2	2.85e-12



(a) Optimal state y^*

(b) Difference $y^* - y^0$

(c) Optimal control u^*

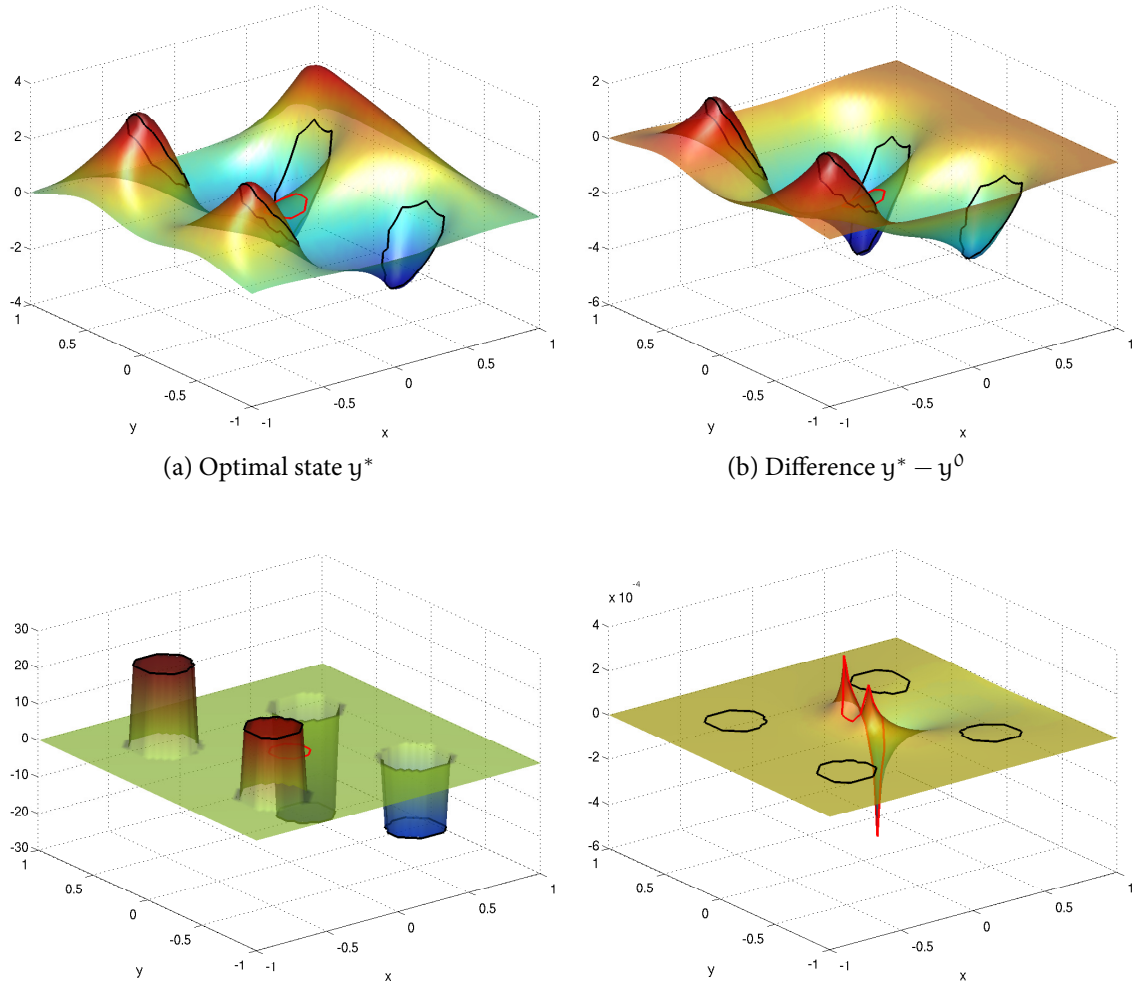
(d) Lagrange multiplier p^*

Figure 11.3: Results for upper bound minimization with non-negativity constraint ($\beta_1 = 1$, $\beta_2 = 0$).

$c^* = 0.202759$. The corresponding maximum and minimum value of y^* on ω_0 is 0.202759 (a difference of -1.70265 compared to the maximum of the uncontrolled state y^0) and -0.202759, respectively. Again, optimal state y^*, difference $y^* - y^0$, optimal control u^* and Lagrange multiplier p^* are shown in Figure 11.4.

ACKNOWLEDGMENTS

The research of the first and last named authors was supported in part by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in



(a) Optimal state y^*

(b) Difference $y^* - y^0$

(c) Optimal control u^*

(d) Lagrange multiplier p^*

Figure 11.4: Results for L^∞ norm minimization ($\beta_1 = \beta_2 = 1$).

Biomedical Sciences"). The research of the second named author was partially supported by the Army Research Office under DAAD19-02-1-0394.

A MINIMUM EFFORT OPTIMAL CONTROL PROBLEM FOR ELLIPTIC PDES

12

ABSTRACT

This work is concerned with a class of minimum effort problems for partial differential equations, where the control cost is of L^∞-type. Since this problem is non-differentiable, a regularized functional is introduced that can be minimized by a superlinearly convergent semismooth Newton method. Uniqueness and convergence for the solutions to the regularized problem are addressed, and a continuation strategy based on a model function is proposed. Numerical examples for a convection-diffusion equation illustrate the behavior of minimum effort controls.

12.1 INTRODUCTION

We investigate the optimal control problem

$$\begin{cases} \min\limits_{u\in L^\infty} \dfrac{1}{2}\|y-z\|_{L^2}^2 + \dfrac{\alpha}{2}\|u\|_{L^\infty}^2 \\ \text{s. t. } Ay = u \qquad \text{in } \Omega, \end{cases} \tag{12.1.1}$$

where $\alpha > 0$, Ω is a bounded domain in $\mathbb{R}^n$, A is a linear second order elliptic partial differential operator of convection-diffusion type carrying appropriate boundary conditions, and $z \in L^2(\Omega)$. Problem (12.1.1) expresses the fact that we wish to determine the best possible control u which steers the state y as close as possible to z, with minimum effort. We consider (12.1.1) as a simple reference problem. The techniques to be presented here can certainly be generalized in many aspects. In particular, the results are applicable if the controls act on subdomains ω strictly contained in Ω. We shall frequently consider an equivalent formulation

given by

$$\begin{cases} \min\limits_{c\in\mathbb{R},u\in L^\infty} \dfrac{1}{2}\|y-z\|_{L^2}^2 + \dfrac{\alpha}{2}c^2 \\ \text{s.t. } Ay = u \qquad \text{in } \Omega, \\ \|u\|_{L^\infty} \leqslant c, \end{cases} \tag{12.1.2}$$

where the nondifferentiability that appears in the cost of (12.1.1) is moved to the constraint set of (12.1.2). This problem resembles a bilaterally constrained optimal control problem, but it is different in that the bound on the control is itself a variable that is subject to minimization. Below, we shall consider yet another reformulation involving a scaling of the control according to $u \to cu$. This will have the advantage that the constraint is not parameter dependent but fixed, at the expense of a bilinear structure occurring in the transformed state-equation constraint.

Problems involving L^∞ control costs – so-called *minimum effort problems* – have received rather little attention in the mathematical literature so far despite their obvious practical relevance. This may be related to the obvious difficulty arising from the nondifferentiability appearing in the problem formulation. We shall demonstrate that semismooth Newton methods in a function space setting are an efficient method to overcome this difficulty. Published investigations of minimum effort problems focus on the case of – mostly linear – control systems in the context of ordinary differential equations. We particularly mention [Neustadt 1962], where sufficient conditions for the optimal controls to be bang-bang are given. In [Ben-Asher, Cliff, and Burns 1989], numerical approaches to solve minimum effort problems are discussed and applications to spacecraft maneuvers are given. The application of semismooth Newton methods to minimum effort problems is presented in [Ito and Kunisch 2011]. In contrast, the corresponding problem for partial differential equations has been studied less frequently (e.g., in [Zuazua 2007] and [Gugat and Leugering 2008] in the context of approximate and exact controllability of heat and wave equations). In passing we also point to a related but different class of problems, where instead of a bound on the controls, bounds on the state are minimized. This type of constraints can be interpreted as minimal invasion problems and was considered in [Prüfert and Schiela 2009; Grund and Rösch 2001] and [Clason, Ito, and Kunisch 2010].

In section 12.2 we discuss existence and uniqueness of a solution to (12.1.1), and present the first order optimality condition. Section 12.3 contains a regularization procedure that is the basis for the numerical treatment by a semismooth Newton method together with a continuation strategy based on a model function approach, all of which are investigated in section 12.4. Numerical examples are presented in section 12.5.

12.2 EXISTENCE, UNIQUENESS, AND OPTIMALITY SYSTEM

We first address well-posedness of the state equation. We consider the operator

$$Ay = -\sum_{j,k=1}^{n} \partial_j(a_{jk}(x)\partial_k y + d_j(x)y) + \sum_{j=1}^{n} b_j(x)\partial_j y + d(x)y,$$

where the coefficients satisfy $a_{jk} \in C^{0,\delta}(\overline{\omega})$ for some $\delta \in (0,1)$ and $b_j, d \in L^\infty(\Omega)$, and the corresponding Dirichlet problem

$$\begin{cases} Ay = g, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases}$$

where the domain $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, is open, bounded with at least Lipschitz continuous boundary $\partial\Omega$, and $g \in L^2(\Omega)$ is given. If 0 is not an eigenvalue of A, this problem has a unique solution in $H_0^1(\Omega)$. A sufficient assumption for this is the existence of constants $\lambda, \Lambda, \nu > 0$ such that

$$\begin{cases} \lambda|\xi|_2^2 \leqslant a_{jk}\xi_j\xi_k \quad \text{for all } \xi \in \mathbb{R}^n, \qquad \displaystyle\sum_{j,k=1}^{n} |a_{j,k}|^2 \leqslant \Lambda^2, \\ \lambda^{-2}\displaystyle\sum_{j=1}^{n} \left(|d_j|^2 + |b_j|^2\right) + \lambda^{-1}|d| \leqslant \nu^2, \qquad d - \partial_j d_j \geqslant 0, \quad \text{for all } 1 \leqslant j \leqslant n, \end{cases}$$

(cf., e.g., [Gilbarg and Trudinger 2001, Th. 8.3]). In particular, this implies the existence of a unique solution $y \in H_0^1(\Omega)$ of the state equation $Ay = u$ for any control $u \in L^\infty(\Omega)$. We further assume that the domain Ω is sufficiently regular (e.g., $\partial\Omega$ is of class $C^{1,1}$ or Ω is a parallelepiped [Ladyzhenskaya and Ural'tseva 1968, pp. 169–189], [Troianiello 1987, Th. 2.24]) that in addition $y \in H^2(\Omega)$ holds.

Consider now the minimum effort problem (12.1.2). Observe that (12.1.2) contains the implicit constraint $c \geqslant 0$. Except in the case $c^* = 0$, problem (12.1.2) can equivalently be expressed by rescaling the control u:

$$(\mathcal{P}) \qquad \begin{cases} \displaystyle\min_{c\in\mathbb{R}_+, u\in L^\infty} \left\{ J(y,c) \equiv \frac{1}{2}\|y - z\|_{L^2}^2 + \frac{\alpha}{2}c^2 \right\} \\ \text{s. t. } Ay = cu \quad \text{in } \Omega, \\ \|u\|_{L^\infty} \leqslant 1 \quad \text{in } \Omega. \end{cases}$$

By standard arguments, we obtain existence of a minimizer $(y^*, u^*, c^*) \in H_0^1(\Omega) \times L^\infty(\Omega) \times \mathbb{R}_+$ of $(\mathcal{P})$. For $c^* = 0$, any control u with $\|u\|_{L^\infty} \leqslant 1$ is a minimizer. The degenerate case $c^* = 0$ can be excluded if and only if $J(y^*, c^*) < \frac{1}{2}\|z\|_{L^2}^2$ with $Ay^* = c^*u^*$ and $\|u^*\|_{L^\infty} \leqslant 1$, which will henceforth be assumed.

For $c^* \neq 0$, this solution is unique. In fact, if (c_1, u_1) and (c_2, u_2) are two (possibly different) solutions to ($\mathcal{P}$) with $c_1 \neq 0$ and $c_2 \neq 0$, then they are also solutions to (12.1.1), where the cost can be expressed as $F(u) = \frac{1}{2}\|A^{-1}u - z\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^\infty}^2$. Since A^{-1} is injective, $u \mapsto \frac{1}{2}\|A^{-1}u - z\|_{L^2}^2$ is strictly convex. Furthermore, $u \mapsto \frac{\alpha}{2}\|u\|_{L^\infty}^2$ is convex. Consequently, $u \mapsto F(u)$ is strictly convex on $L^\infty(\Omega)$ and hence $u_1 = u_2$ and consequently $c_1 = c_2$ holds.

Using standard subdifferential calculus (cf., e.g., [Ekeland and Témam 1999]), we obtain the existence of a Lagrange multiplier $p^* \in H_0^1(\Omega)$ satisfying the necessary optimality conditions

$$\text{(OS)} \qquad \begin{cases} \langle -p^*, u - u^* \rangle_{L^2} \geqslant 0 \quad \text{for all } u \text{ with } \|u\|_{L^\infty} \leqslant 1, \\ \alpha c^* - \langle u^*, p^* \rangle_{L^2} = 0, \\ y^* - z + A^* p^* = 0, \\ Ay^* - c^* u^* = 0. \end{cases}$$

From the assumption on the regularity of Ω, we have in addition $p^* \in H^2(\Omega)$.

By pointwise inspection of the first relation of (OS) we deduce that

$$u(x) = \begin{cases} 1 & \text{if } p(x) > 0, \\ -1 & \text{if } p(x) < 0, \\ t \in [-1, 1] & \text{if } p(x) = 0 \end{cases}$$

holds, which can be equivalently expressed as $u = \operatorname{sign}(p)$. Inserting this into the second relation of (OS) and eliminating y and u from the last two relations, we obtain the reduced optimality system

$$\text{(OS')} \qquad \begin{cases} AA^* p^* + c^* \operatorname{sign}(p^*) = Az, \\ \alpha c^* - \|p^*\|_{L^1} = 0, \end{cases}$$

where the first equation should be interpreted in variational form, i.e., as

$$\langle A^* p^*, A^* v \rangle_{L^2} + c^* \langle \operatorname{sign}(p^*), v \rangle_{L^2} = \langle z, A^* v \rangle_{L^2}$$

for all $v \in H_0^1(\Omega) \cap H^2(\Omega)$.

12.3 REGULARIZED PROBLEM

From (OS'), it is clear that the optimality system is not differentiable even in a generalized sense. We therefore introduce the following regularization in Problem ($\mathcal{P}$), where we again only consider $c \geqslant 0$:

$$(\mathcal{P}_\beta) \qquad \begin{cases} \min\limits_{c \in \mathbb{R}_+, u \in L^\infty} \left\{ J_\beta(y, u, c) \equiv \dfrac{1}{2}\|y - z\|_{L^2}^2 + \dfrac{\beta c}{2}\|u\|_{L^2}^2 + \dfrac{\alpha}{2}c^2 \right\} \\ \text{s. t. } Ay = cu \quad \text{in } \Omega, \\ \|u\|_{L^\infty} \leqslant 1. \end{cases}$$

As before, existence of a minimizer $(y_\beta, u_\beta, c_\beta) \in H^1_0(\Omega) \times L^\infty(\Omega) \times \mathbb{R}_+$ follows from standard arguments. The case $c_\beta = 0$ is excluded by the assumption that $J(y^*, c^*) < \frac{1}{2}\|z\|^2_{L^2}$, where (y^*, u^*, c^*) is the solution to $(\mathcal{P})$. In fact, if the c_β-component of the solution to $(\mathcal{P}_\beta)$ is zero, then

$$\begin{aligned}\frac{1}{2}\|z\|^2_{L^2} &= \frac{1}{2}\|y_\beta - z\|^2_{L^2} + \frac{\beta c_\beta}{2}\|u_\beta\|^2_{L^2} + \frac{\alpha}{2}c_\beta^2 \\ &\leqslant \frac{1}{2}\|y^* - z\|^2_{L^2} + \frac{\beta c^*}{2}\|u^*\|^2_{L^2} + \frac{\alpha}{2}(c^*)^2 = J(y^*, c^*) < \frac{1}{2}\|z\|^2_{L^2},\end{aligned}$$

which gives a contradiction.

Due to the bilinear structure of the equality constraint in $(\mathcal{P}_\beta)$, uniqueness of the solution is not obvious. The (technical) proof of the following statement is given in Appendix 12.A.

Proposition 12.3.1. *If $\alpha > 0$ is sufficiently large, then the solution $(y_\beta, u_\beta, c_\beta)$ to $(\mathcal{P}_\beta)$ is unique for every $\beta > 0$. For any $\alpha > 0$ and given c_β, the corresponding components u_β, y_β are unique, and conversely c_β and hence y_β is uniquely determined by u_β.*

For $c_\beta > 0$, we obtain the existence of a $p_\beta \in H^1_0(\Omega)$ satisfying the necessary optimality conditions for $(\mathcal{P}_\beta)$:

$$(\mathrm{OS}_\beta) \qquad \begin{cases} \langle \beta u_\beta - p_\beta, u - u_\beta\rangle_{L^2} \geqslant 0 \quad \text{for all } u \text{ with } \|u\|_{L^\infty} \leqslant 1, \\ \alpha c_\beta + \dfrac{\beta}{2}\|u_\beta\|^2_{L^2} - \langle u_\beta, p_\beta\rangle_{L^2} = 0, \\ y_\beta - z + A^* p_\beta = 0, \\ A y_\beta - c_\beta u_\beta = 0. \end{cases}$$

We have again by pointwise inspection of the first relation that

$$u(x) = \operatorname{sign}_\beta(p)(x) := \begin{cases} 1 & \text{if } p(x) > \beta, \\ -1 & \text{if } p(x) < -\beta, \\ \frac{1}{\beta}p(x) & \text{if } |p(x)| \leqslant \beta \end{cases}$$

holds. Using again the higher regularity $p_\beta \in H^2(\Omega)$, we can insert this into the second relation of (OS_β) and eliminate y and u to obtain

$$(\mathrm{OS}'_\beta) \qquad \begin{cases} AA^* p_\beta + c_\beta \operatorname{sign}_\beta(p_\beta) = Az, \\ \alpha c_\beta - \|p_\beta\|_{L^1_\beta} = 0, \end{cases}$$

where we have defined

$$\|p\|_{L^1_\beta} := \int_\Omega |p(x)|_\beta \, dx, \qquad |p(x)|_\beta := \begin{cases} p(x) - \frac{\beta}{2} & \text{if } p(x) > \beta, \\ -p(x) - \frac{\beta}{2} & \text{if } p(x) < -\beta, \\ \frac{1}{2\beta}p(x)^2 & \text{if } |p(x)| \leqslant \beta. \end{cases}$$

Remark 12.3.2. The chosen regularization is motivated by the following consideration: We can write (12.1.2) in the form

$$\min_{c}\left(\min_{u}\frac{1}{2}\|A^{-1}u-z\|_{L^2}^2+I_{\{\|v\|_{L^\infty}\leqslant c\}}(u)\right)+\frac{\alpha}{2}c^2,$$

Formally applying Fenchel duality for the inner minimization problem (where we consider $c>0$ fixed), we obtain by noting that the Fenchel dual of the indicator function of the L^∞-ball is the scaled L^1 norm

$$(\mathcal{P}^*)\qquad \min_{c}\left(\sup_{p}-\frac{1}{2}\|A^*p+z\|_{L^2}^2-c\,\|p\|_{L^1}\right)+\frac{\alpha}{2}c^2.$$

Our regularization now amounts to replacing the non-differentiable L^1-norm by the quadratic approximation $\|p\|_{L^1_\beta}$, which has second (Newton-)derivatives and can be considered as a Huber-type smoothing of the L^1-norm. The optimality system for the regularized dual problem (after replacing p by $-p$) is then given by (OS'_β). In Appendix 12.B, we compare different regularization strategies, which will turn out to be less convenient.

Remark 12.3.3. The proposed approach can also be applied when the control acts on a proper subdomain $\omega\subset\Omega$. Introducing the extension operator E_ω from ω to Ω, we consider the regularized problem

$$\begin{cases}\min\limits_{c\in\mathbb{R}_+,u\in L^2(\omega)}\dfrac{1}{2}\|y-z\|_{L^2}^2+\dfrac{\beta c}{2}\|u\|_{L^2(\omega)}^2+\dfrac{\alpha}{2}c^2\\ \text{s.t. } Ay=cE_\omega u\quad\text{in }\Omega,\\ \|u\|_{L^\infty(\omega)}\leqslant 1,\end{cases}$$

with the necessary optimality conditions

$$\begin{cases} c\,\langle\beta u-E_\omega^*p,\tilde u-u\rangle_{L^2(\omega)}\geqslant 0 & \text{for all }\tilde u\text{ with }\|\tilde u\|_{L^\infty(\omega)}\leqslant 1,\\ \alpha c+\dfrac{\beta}{2}\|u\|_{L^2(\omega)}^2-\langle u,p\rangle_{L^2(\omega)}=0 & \text{if } c>0,\\ y-z+A^*p=0,\\ Ay-cE_\omega u=0,\end{cases}$$

where E_ω^* denotes the restriction operator to ω. By case discrimination and pointwise inspection we can again obtain the reduced optimality system

$$\begin{cases} AA^*p+c\,\operatorname{sign}_\beta(\chi_\omega p)=Az,\\ \alpha c-\|\chi_\omega p\|_{L^1_\beta}=0,\end{cases}$$

where $\chi_\omega=E_\omega E_\omega^*$ is the characteristic function of ω. The solution to this system can be computed using the semismooth Newton method described in section 12.4.1 after changing the definition of the active and inactive sets to $\mathcal{A}_+\cap\omega$, $\mathcal{A}_-\cap\omega$ and $\mathcal{I}\cap\omega$, respectively.

We next address the convergence of solutions to $(\mathcal{P}_\beta)$ as $\beta \to 0$. First, we show monotonicity properties of the solutions with respect to β.

Lemma 12.3.4. *For $\beta > 0$ let $(y_\beta, u_\beta, c_\beta)$ denote any solution to $(\mathcal{P}_\beta)$ and let (y^*, u^*, c^*) denote the solution to $(\mathcal{P})$. Then for any $\beta < \beta'$ we have that*

$$c_{\beta'}\|u_{\beta'}\|_{L^2}^2 \leqslant c_\beta \|u_\beta\|_{L^2}^2, \tag{12.3.1}$$

$$J(y_\beta, c_\beta) \leqslant J(y_{\beta'}, c_{\beta'}), \tag{12.3.2}$$

$$J(y_\beta, c_\beta) + \frac{\beta c_\beta}{2}\|u_\beta\|_{L^2}^2 \leqslant J(y^*, c^*) + \frac{\beta c^*}{2}\|u^*\|_{L^2}^2. \tag{12.3.3}$$

Proof. For $0 \leqslant \beta < \beta'$ (where $c_0 \equiv c^*$ etc.) we have that

$$J(y_\beta, c_\beta) + \frac{\beta c_\beta}{2}\|u_\beta\|_{L^2}^2 \leqslant J(y_{\beta'}, c_{\beta'}) + \frac{\beta c_{\beta'}}{2}\|u_{\beta'}\|_{L^2}^2$$

which implies

$$\begin{aligned} J(y_\beta, c_\beta) + \frac{\beta c_\beta}{2}\|u_\beta\|_{L^2}^2 + \frac{(\beta'-\beta)c_{\beta'}}{2}\|u_{\beta'}\|_{L^2}^2 &\leqslant J(y_{\beta'}, c_{\beta'}) + \frac{\beta' c_{\beta'}}{2}\|u_{\beta'}\|_{L^2}^2 \\ &\leqslant J(y_\beta, c_\beta) + \frac{\beta' c_\beta}{2}\|u_\beta\|_{L^2}^2. \end{aligned} \tag{12.3.4}$$

From the outer inequalities we deduce that

$$(\beta - \beta')\left(c_{\beta'}\|u_{\beta'}\|_{L^2}^2 - c_\beta\|u_\beta\|_{L^2}^2\right) \leqslant 0,$$

which implies relation (12.3.1).

From the first inequality of (12.3.4), we obtain

$$J(y_\beta, c_\beta) - J(y_{\beta'}, c_{\beta'}) \leqslant \beta\left(c_\beta\|u_\beta\|_{L^2}^2 - c_{\beta'}\|u_{\beta'}\|_{L^2}^2\right) \leqslant 0,$$

and relation (12.3.2) follows.

Finally, relation (12.3.3) is a consequence of the second inequality of (12.3.4) by setting $\beta = 0$ and $\beta' = \beta$. □

We can now show strong subsequential convergence of minimizers of $(\mathcal{P}_\beta)$.

Proposition 12.3.5. *Any selection of solutions $\{(y_\beta, u_\beta, c_\beta)\}_{\beta>0}$ of $(\mathcal{P}_\beta)$ is bounded in $H_0^1(\Omega) \times L^\infty(\Omega) \times \mathbb{R}_+$. For $\beta \to 0$, it converges weakly-$\star$ to the solution to $(\mathcal{P})$, and the convergence is strong in $H_0^1(\Omega) \times L^q(\Omega) \times \mathbb{R}_+$ for any $q \in [1, \infty)$.*

Proof. Since $(y, u, c) = (0, 0, 0)$ is feasible for the constraints in $(\mathcal{P}_\beta)$, we have

$$\|y_\beta - z\|_{L^2}^2 + \beta c_\beta \|u_\beta\|_{L^2}^2 + \alpha c_\beta^2 \leqslant \|z\|_{L^2}^2$$

and hence $\{c_\beta\}_{\beta>0}$ is bounded. The family $\{u_\beta\}_{\beta>0}$ is bounded by 1 in $L^\infty(\Omega)$ and consequently $\{y_\beta\}_{\beta>0}$ is bounded in $H_0^1(\Omega)$ for any $q \in [1, \infty)$.

Hence there exists $(\bar{y}, \bar{u}, \bar{c}) \in H_0^1(\Omega) \times L^\infty(\Omega) \times \mathbb{R}_+$ such that, on a subsequence denoted by the same symbols, $(y_\beta, u_\beta, c_\beta) \rightharpoonup^\star (\bar{y}, \bar{u}, \bar{c})$ holds in $H_0^1(\Omega) \times L^\infty(\Omega) \times \mathbb{R}$. Passing to the limit in the variational formulation of $Ay_\beta = c_\beta u_\beta$, we find that $A\bar{y} = \bar{c}\bar{u}$. By the weak lower semicontinuity of the L^∞-norm, we have that $\|\bar{u}\|_{L^\infty} \leqslant 1$. Weak lower semicontinuity of J_β from $L^2(\Omega) \times L^2(\Omega) \times \mathbb{R}_+ \to \mathbb{R}$ implies that $(\bar{y}, \bar{u}, \bar{c})$ is a solution to $(\mathcal{P})$. Since the solution to $(\mathcal{P})$ is unique, $(\bar{y}, \bar{u}, \bar{c})$ coincides with (y^*, u^*, c^*) and the whole family $\{(y_\beta, u_\beta, c_\beta)\}_{\beta>0}$ converges.

To show strong convergence, we insert the weak limit (u^*, c^*) in the inequality (12.3.1) (setting $\beta = 0$) to deduce from the lower semicontinuity of the norm that

$$c_\beta \|u_\beta\|_{L^2}^2 \leqslant c^* \|u^*\|_{L^2}^2 \leqslant \liminf_{\beta \to 0} c_\beta \|u_\beta\|_{L^2}^2$$

holds. From this, we deduce

$$\limsup_{\beta \to 0} \|u_\beta\|_{L^2}^2 \leqslant \|u^*\|_{L^2}^2 \leqslant \liminf_{\beta \to 0} \|u_\beta\|_{L^2}^2$$

and hence strong convergence in $L^2(\Omega)$ – and thus in $L^q(\Omega)$ for every $q \in [1, \infty)$ as well – of the subsequence u_β to u^*. This also implies the strong convergence of $y_\beta = A^{-1}(c_\beta u_\beta)$ in $H_0^1(\Omega)$. □

Inserting (y^*, u^*, c^*) into (12.3.3) (setting $\beta = 0$) yields

$$0 \leqslant J(y_\beta, c_\beta) - J(y^*, c^*) \leqslant \beta \left(c^* \|u^*\|_{L^2} - c_\beta \|u_\beta\|_{L^2}\right).$$

From the strong convergence of u_β, we therefore obtain the following convergence rate result.

Corollary 12.3.6. *As $\beta \to 0$, it holds that*

$$J(y_\beta, c_\beta) - J(y^*, c^*) = o(\beta).$$

12.4 SOLUTION OF OPTIMALITY SYSTEM

In this section, we discuss the computation of approximate minimizers of $(\mathcal{P})$. The first subsection is concerned with the solution for fixed $\beta > 0$ of the regularized optimality system (OS_β). We then propose a continuation strategy in β where the stopping criterion is based on a model function.

12.4.1 SEMISMOOTH NEWTON METHOD

For the numerical solution of the regularized problem ($\mathcal{P}_\beta$), we consider the reduced optimality system (OS'_β) as an operator equation $T(p,c) = (0,0)$ for

$$T : \mathcal{W} \times \mathbb{R}_+ \to H^{-2}(\Omega) \times \mathbb{R}, \quad (p,c) \mapsto \begin{pmatrix} AA^*p + c\,\mathrm{sign}_\beta(p) - Az \\ \alpha c - \|p\|_{L^1_\beta} \end{pmatrix},$$

where $\mathcal{W} := H^1_0(\Omega) \cap H^2(\Omega)$ and $H^{-2}(\Omega) := (\mathcal{W})^*$. Obviously, T is differentiable with respect to c. We next argue Newton differentiability of T with respect to p. First observe that

$$\mathrm{sign}_\beta(v) = \tfrac{1}{\beta}(v - \max(0, v-\beta) - \min(0, v+\beta))$$

and recall (e.g., from [Ito and Kunisch 2008, Th. 8.5]; see also [Schiela 2008]) that for any $\beta \in \mathbb{R}$, the function $z \mapsto \max(0, z-\beta)$ is Newton differentiable from $L^p(\Omega)$ to $L^q(\Omega)$ for any $p > q \geqslant 1$ with its Newton derivative in direction h given pointwise by

$$\big(D_N \max(0, v-\beta)h\big)(x) = \begin{cases} h(x), & \text{if } v(x) > \beta, \\ 0, & \text{if } v(x) \leqslant \beta. \end{cases}$$

An analogous statement holds for the min function. The function sign_β is thus Newton differentiable from $L^p(\Omega)$ to $L^q(\Omega)$ as well, where the Newton derivative of sign_β is given by

$$\big(D_N \,\mathrm{sign}_\beta(p)h\big)(x) = \begin{cases} 0, & \text{if } |p(x)| > \beta, \\ \frac{1}{\beta}h(x), & \text{if } |p(x)| \leqslant \beta. \end{cases}$$

Since the mapping $\psi : \mathbb{R} \to \mathbb{R}$, $t \mapsto |t|_\beta$, is differentiable with globally Lipschitz continuous derivative $t \mapsto \mathrm{sign}_\beta(t)$, ψ defines a differentiable Nemytskii operator from $L^p(\Omega)$ to $L^2(\Omega)$ for every $p \geqslant 4$ (see, e.g., [Tröltzsch 2010, Chap. 4.3] and the references therein). This yields the Newton differentiability of $\|\cdot\|_{L^1_\beta}$ from $L^p(\Omega)$, $p \geqslant 4$, to $\mathbb{R}$, with Newton derivative

$$D_N(\|p\|_{L^1_\beta})h = \langle \mathrm{sign}_\beta(p), h\rangle_{L^2}.$$

The Newton differentiability of T thus follows from the smoothing properties of AA^*.

Defining the active and inactive sets by

$$\begin{aligned} \mathcal{A}_+ &= \{x \in \Omega : p^k(x) > \beta\}, \\ \mathcal{A}_- &= \{x \in \Omega : p^k(x) < -\beta\}, \\ \mathcal{A} &= \mathcal{A}_+ \cup \mathcal{A}_-, \quad \mathcal{I} = \Omega \setminus \mathcal{A} \end{aligned}$$

with indicator functions $\chi_{\mathcal{A}_+}, \dots, \chi_{\mathcal{I}}$, a semismooth Newton step consists in finding $\delta p, \delta c$ for given p^k, c^k such that

$$\begin{cases} AA^*\delta p + c^k \frac{1}{\beta}\chi_{\mathcal{I}}\delta p + \operatorname{sign}_\beta(p^k)\delta c = -(AA^*p^k + c\operatorname{sign}_\beta(p^k) - Az), \\ \alpha\delta c - \langle \operatorname{sign}_\beta(p^k), \delta p\rangle = -(\alpha c^k - \|p^k\|_{L^1_\beta}) \end{cases} \tag{12.4.1}$$

holds.

We now show that the Newton system (12.4.1) is uniformly invertible for fixed $\beta > 0$.

Proposition 12.4.1. *For each $(p, c) \in \mathcal{W} \times \mathbb{R}_+$, the mapping $M : \mathcal{W} \times \mathbb{R} \to H^{-2}(\Omega) \times \mathbb{R}$,*

$$M(\delta p, \delta c) := \begin{pmatrix} AA^*\delta p + c\frac{1}{\beta}\chi_{\mathcal{I}}\delta p + \operatorname{sign}_\beta(p)\delta c \\ \alpha\delta c - \langle \operatorname{sign}_\beta(p), \delta p\rangle \end{pmatrix}$$

is invertible, and there exists a constant $C > 0$ independent of (p, c) such that there holds

$$\|M^{-1}\| \leqslant C.$$

Proof. Due to the regularity of Ω, we have that A^* acts as an isomorphism from $\mathcal{W}$ to a closed subspace of $L^2(\Omega)$. It thus suffices to observe that

$$\langle (\delta p, \delta c), M(\delta p, \delta c)\rangle = \|A^*\delta p\|^2_{L^2} + \frac{c}{\beta}\|\delta p\,\chi_{\mathcal{I}}\|^2_{L^2} + \alpha\delta c^2 > 0$$

for $(\delta p, \delta c) \neq 0$, independent of p and $c > 0$. □

Thus, system (12.4.1) is semismooth, and from standard results (e.g., [Ito and Kunisch 2008, Th. 8.5]) we deduce the following convergence result for the semismooth Newton method.

Theorem 12.4.2. *For every $\alpha, \beta > 0$, the Newton iteration (12.4.1) converges superlinearly to the solution (p_β, c_β) of (OS'_β), provided that (p^0, c^0) is sufficiently close to (p_β, c_β).*

The following finite termination property (e.g., [Ito and Kunisch 2008, Rem. 7.1.1]) will be useful in formulating a continuation scheme in β:

Proposition 12.4.3. *If $\mathcal{A}^{k+1}_+ = \mathcal{A}^k_+$ and $\mathcal{A}^{k+1}_- = \mathcal{A}^k_-$ holds, then $T(p^{k+1}, c^{k+1}) = 0$.*

12.4.2 CONTINUATION STRATEGY FOR β

While Theorem 12.4.2 guarantees locally superlinear convergence for every $\beta > 0$, in practice the region of convergence for the semismooth Newton method shrinks with decreasing β. In order to compute a good approximation of the original minimum effort problem ($\mathcal{P}$), we make use of a continuation approach: Starting with large β_n, we compute the minimizer $(p_{\beta_n}, c_{\beta_n})$, decrease β_n by a given factor q_m and compute the corresponding minimizers $(p_{\beta_{n+1}}, c_{\beta_{n+1}})$ starting from the initial guess $(p^0, c^0) = (p_{\beta_n}, c_{\beta_n})$. If the Newton iteration did not converge after a fixed number of iterations (as determined by the change in active sets), we increase the reduction factor by setting $q_{m+1} = (q_m)^t$ for a fixed $t < 1$ and restart the iteration with new $\beta_n = q_{m+1}\beta_{n-1}$.

Let us address the optimal stopping of the decrease of the regularization parameter. For very small values of β there is little change in the value of c_β and $c_\beta \|u_\beta\|_{L^2}^2$. This observation from numerical tests can be used to develop a stopping rule based on a model function. From Lemma 12.3.4 it is known that $\beta \mapsto c_\beta \|u_\beta\|_{L^2}^2$ is monotonically decreasing. Let $\mu > 0$ denote the desired efficiency level of the regularization term. Then the stopping parameter $\hat{\beta}$ is chosen such that

$$c_{\hat{\beta}} \|u_{\hat{\beta}}\|_{L^2}^2 > \mu c^* \|u^*\|_{L^2}^2.$$

Since c^* and u^* are unknown, we propose to introduce a model function $m(\beta)$ which approximates $\beta \mapsto c_\beta \|u_\beta\|_{L^2}^2$. The specific choice we make is

$$m(\beta) = \frac{K_1}{(K_2 + \beta)^2}. \tag{12.4.2}$$

Since $c^* \|u^*\|_{L^2}^2$ is finite, we can expect that $0 < K_1 < \infty$ and $0 < K_2 < \infty$. The constants K_1, K_2 can be determined by interpolation from evaluations with two successive solutions to ($\mathcal{P}_\beta$). The continuation is then stopped if

$$c_{\beta_n} \|u_{\beta_n}\|_{L^2}^2 > \mu m_n(0)$$

is satisfied, where m_n is constructed from the interpolation conditions at β_n and β_{n-1}.

The choice (12.4.2) for m is based partly on numerical experience and partly on the following heuristic considerations. The necessary optimality condition implies that

$$0 = cA^{-1}u - z + \beta A u,$$

where we ignore the inequality constraint on u. Considering A as a scalar variable, rather than as an operator, and denoting it by a henceforth, we have

$$(c\, a^{-2} + \beta)u - a^{-1}z = 0. \tag{12.4.3}$$

Here we may consider u as the value of $u(x)$ at some x where the constraint is not yet active. The range of interest for creating the model function covers small values of β, where

Algorithm 12.1 Path-following semismooth Newton method

Choose β_0, q_0, t, k_*, μ
Set $(c_0, p_0) = (0, 0)$, $n = 0$, $q = q_0$
repeat ▷ continuation in β
  Set $k = 0$, $p^0 = p_n$, $c^0 = c_n$
  repeat ▷ semismooth Newton method
    Compute active sets $\mathcal{A}_+^k$, $\mathcal{A}_-^k$
    Solve Newton system (12.4.1) for δp, δc and set

$$p^{k+1} = p^k + \delta p, \quad c^{k+1} = c^k + \delta c, \quad k \leftarrow k+1$$

  until $(\mathcal{A}_+^k = \mathcal{A}_+^{k-1}$ and $\mathcal{A}_-^k = \mathcal{A}_-^{k-1})$ or $k > k_*$
  if $(\mathcal{A}_+^k \neq \mathcal{A}_+^{k-1}$ or $\mathcal{A}_-^k \neq \mathcal{A}_-^{k-1})$ **then**
    Set $q \leftarrow (q)^t$, $\beta_n = q\beta_{n-1}$ ▷ increase β, restart Newton iteration
  else
    Set $\beta_{n+1} = q\beta_n$ ▷ decrease β
    Set $p_{n+1} = p^k$, $c_{n+1} = c^k$, $n \leftarrow n+1$ ▷ accept computed step
    Compute $u_n = \operatorname{sign}_{\beta_n}(p_n)$
    Determine $m_n(\beta)$ from $(\beta_n, c_n\|u_n\|_{L^2}^2)$, $(\beta_{n-1}, c_{n-1}\|u_{n-1}\|_{L^2}^2)$
  end if
until $c_n\|u_n\|_{L^2}^2 > \mu m_n(0)$

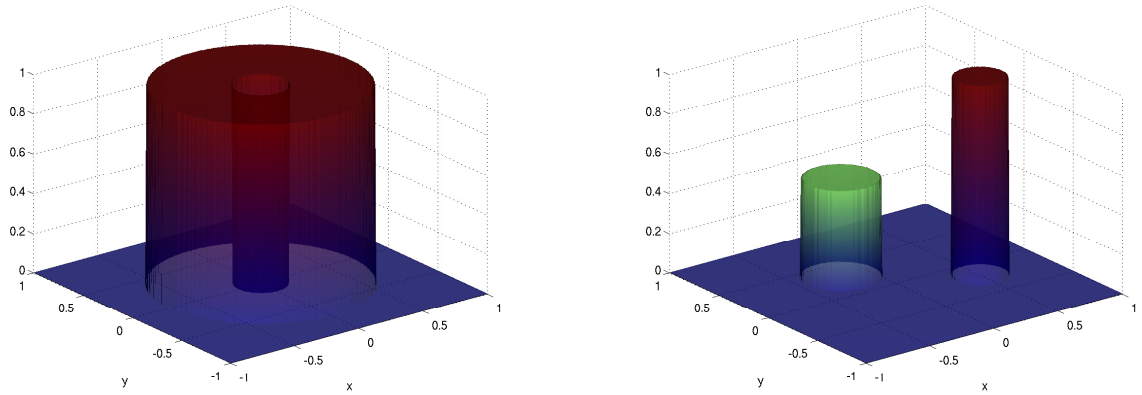
numerical results show little dependence of c_β on β. Assuming therefore that c is a constant, and differentiating (12.4.3) with respect to β, we obtain

$$(c\,\alpha^{-2} + \beta)\frac{d}{d\beta}u + u = 0.$$

The solution to this ordinary differential equation is given by $u = \frac{k_1}{c\,\alpha^{-2}+\beta}$. This suggests using $m(\beta)$ as a model function for $c_\beta\|u_\beta\|_{L^2}^2$.

The full procedure for the numerical approximation of the solution to ($\mathcal{P}$) is given as Algorithm 12.1.

Remark 12.4.4. The convergence of the path-following method can be accelerated by starting with a damped Newton iteration (cf. [Sun and Zeng 2010]), where we only take fractional Newton steps. In our experiments, a sequence of step sizes $\tau^k = \frac{k+1}{k+2}$ showed good results. While this modification is not necessary for the convergence of the method, it allows larger steps in the decrease of β. The benefit depends on α, from about 20% performance increase for $\alpha = 10^{-2}$ to about 80% for $\alpha = 10^{-5}$. Since the focus of this work is not on optimal performance of the numerical solution, the examples shown below do not make use of this damping strategy.



(a) z_1 (b) z_2

Figure 12.1: Target functions

12.5 NUMERICAL EXAMPLES

To illustrate the features of the optimal controls arising in the minimum effort problem, we consider convective-diffusive transport, which is described by the operator $Ay = -\nu\Delta y + b \cdot \nabla y$ with $\nu = 0.1$ and $b = (-1, 0)^T$ with homogeneous Dirichlet conditions on the unit square $[-1, 1]^2$. We show results for two target functions

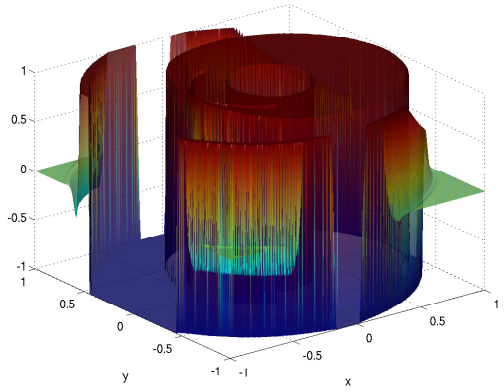
$$z_1(x, y) = \chi_{\{x^2+y^2<1/2\}}\chi_{\{x^2+y^2>1/32\}},$$
$$z_2(x, y) = \chi_{\{(x-0.5)^2+(y+0.2)^2<1/32\}} + \frac{1}{2}\chi_{\{(x+0.2)^2+(y-0.3)^2<1/16\}},$$

which are shown in Figure 12.1.

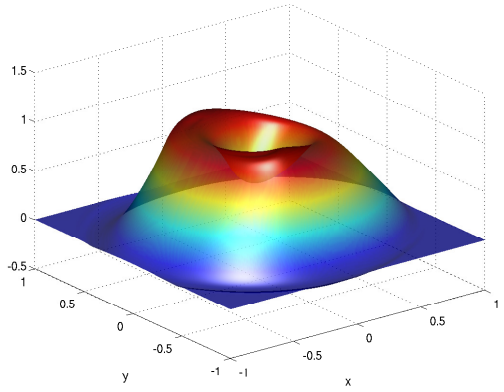
The parameters in Algorithm 12.1 were set to $\beta_0 = 1$, $q_0 = 10^{-1}$, $t = 0.5$, $k_* = 10$, and $\mu = 0.99$. The differential operators were discretized using finite differences on a uniform grid with $N = 256$ nodes in each direction. A Matlab function implementing Algorithm 12.1 can be downloaded from http://www.uni-graz.at/~clason/codes/mineffort.m.

We first compare the optimal (scaled) controls $(cu)_\alpha \equiv c^*u^*$ for different values of α. The controls and corresponding states y_α are shown in Figure 12.2 for the target z_1 and in Figure 12.3 for the target z_2. The bang-bang nature of the minimum effort control can be seen clearly. The optimal L^∞ bounds c_α are given in Table 12.1. In all cases, the optimal control u_α is feasible, i.e., $\max(u_\alpha) = -\min(u_\alpha) = 1$. According to the model function, the continuation was stopped around $2 \cdot 10^{-7}$ in all cases except for target z_1 with $\alpha = 5 \cdot 10^{-3}$, where the iteration was terminated at $2.4 \cdot 10^{-6}$.

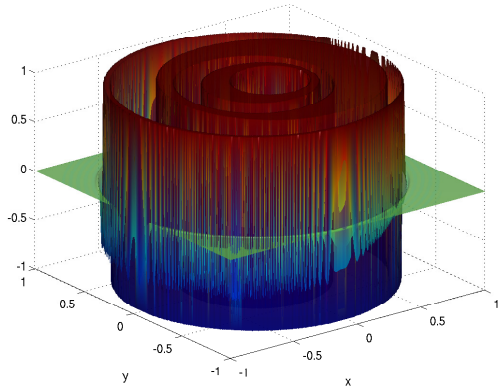
We illustrate the convergence behavior with respect to β exemplarily for the target z_2 and $\alpha = 5 \cdot 10^{-3}$ in Figure 12.4. Figure 12.4a shows the iteration history of c_β, where every circle represents a computed value. Figure 12.4b illustrates Corollary 12.3.6 by plotting the difference



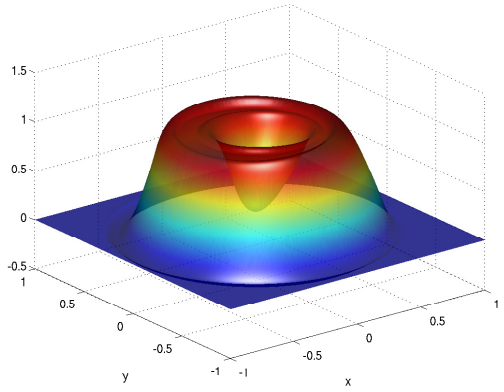
(a) u_α, $\alpha = 5 \cdot 10^{-3}$



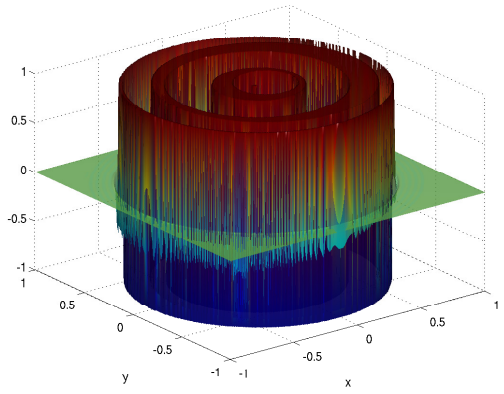
(b) y_α, $\alpha = 5 \cdot 10^{-3}$



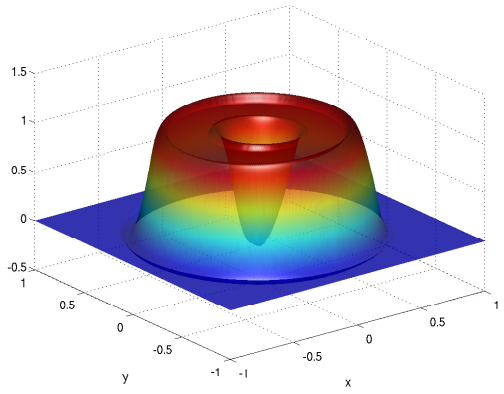
(c) u_α, $\alpha = 5 \cdot 10^{-4}$



(d) y_α, $\alpha = 5 \cdot 10^{-4}$

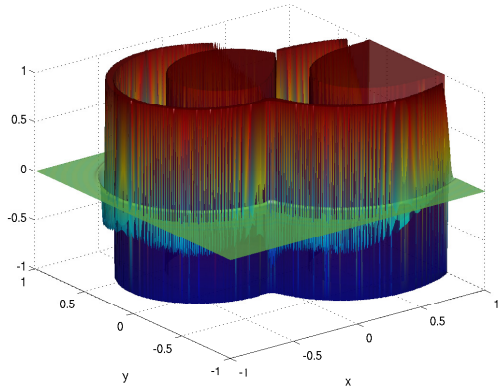


(e) u_α, $\alpha = 5 \cdot 10^{-5}$

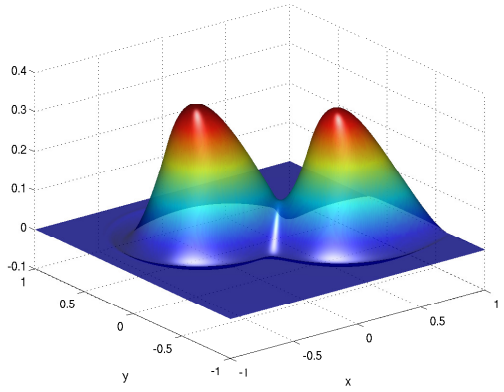


(f) y_α, $\alpha = 5 \cdot 10^{-5}$

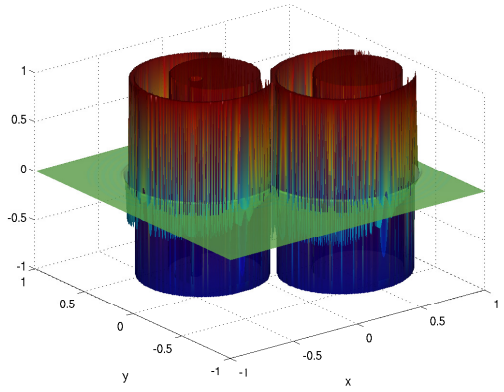
Figure 12.2: Optimal controls u_α and states y_α for the target z_1 and different α.



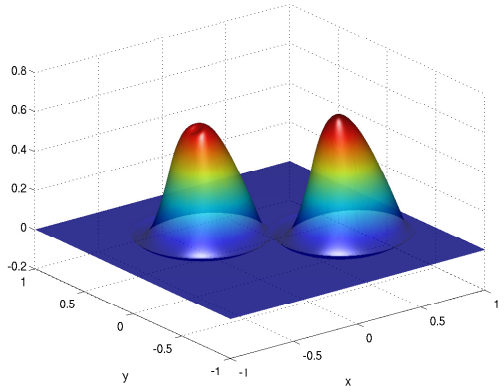
(a) u_α, $\alpha = 5 \cdot 10^{-3}$



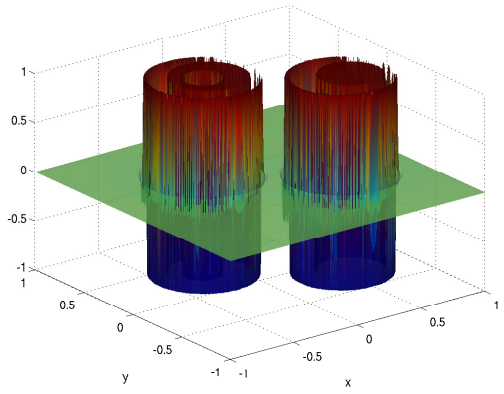
(b) y_α, $\alpha = 5 \cdot 10^{-3}$



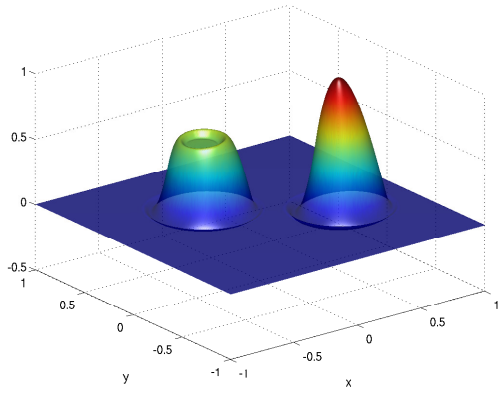
(c) u_α, $\alpha = 5 \cdot 10^{-4}$



(d) y_α, $\alpha = 5 \cdot 10^{-4}$

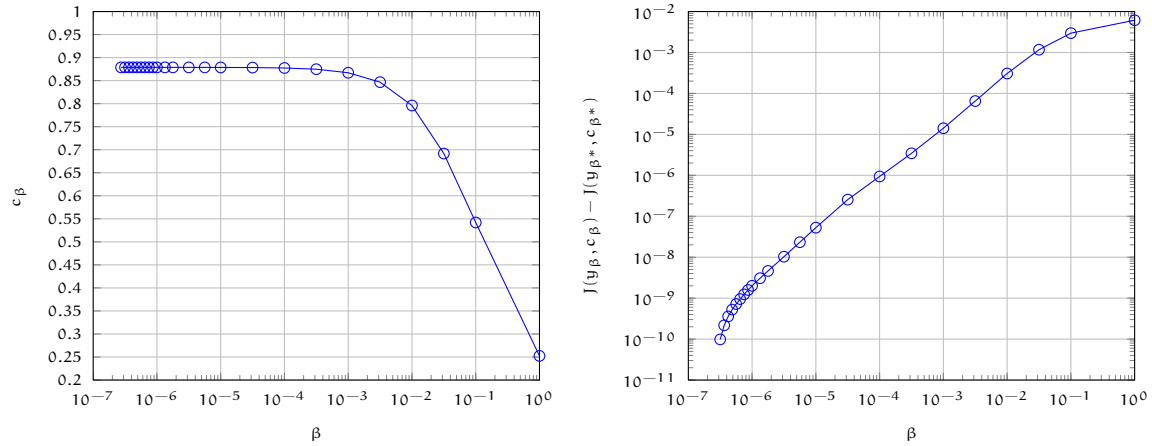


(e) u_α, $\alpha = 5 \cdot 10^{-5}$



(f) y_α, $\alpha = 5 \cdot 10^{-5}$

Figure 12.3: Optimal controls u_α and states y_α for the target z_2 and different α.



(a) optimal L^∞ bound c_β

(b) difference $J(y_\beta, c_\beta) - J(y_{\beta_{\min}}, c_{\beta_{\min}})$

Figure 12.4: Illustration of the convergence behavior with respect to β. Every circle corresponds to a computed step in the continuation method.

between the current functional value $J(\beta) \equiv J(y_\beta, c_\beta)$ and the final computed value of $J(\beta_n)$. We point out the asymptotic superlinear decay as $\beta \to 0$.

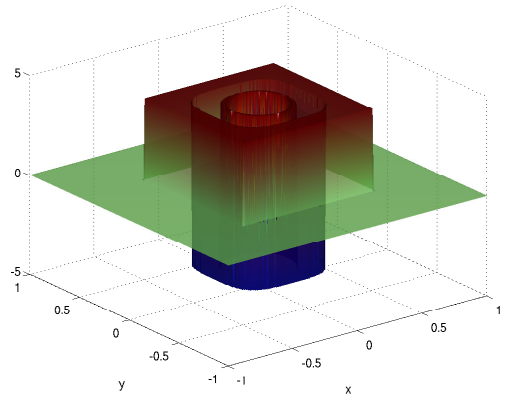
We indicate the superlinear convergence of the semismooth Newton method by fixing $\beta = 5 \cdot 10^{-2}$ and computing p_β, c_β from the starting guess $(p_0, c_0) = (0, 0)$ (again, for target z_2 and $\alpha = 5 \cdot 10^{-3}$). Table 12.2 shows the norm of the residual $\|T(p^k, c^k)\|_{L^2}$ in the semismooth Newton method, verifying the locally superlinear convergence shown in Theorem 12.4.2.

Finally, we consider the effect of the geometry of the control domain on the optimal control and state. For this, we choose the target $z_3(x, y) = 1$, set $b = (0, 0)$ and compare subdomains of equal area on which the control is allowed to act (cf. Remark 12.3.3): The control domain ω_n consists of 1, 4 or 9 uniformly distributed squares whose areas each sum to 1 (see Figure 12.5). The penalty parameter was fixed at $\alpha = 10^{-3}$, and to allow quantitative comparison, the continuation in β was terminated in each case when $\beta \leqslant 10^{-7}$ was satisfied. The resulting controls and states are shown in Figure 12.5, and the corresponding optimal L^∞-bound c^* are 4.1039, 4.3707 and 4.6298, respectively.

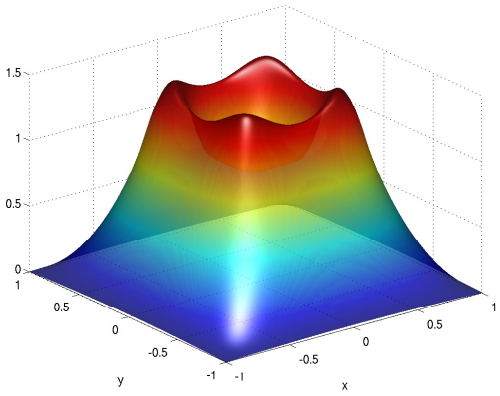
Table 12.1: Optimal L^∞-bounds c_α for targets z_1 (left) and z_2 (right) and different α.

α	$5 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-3}$
c_α	1.9622	4.4236	9.8518

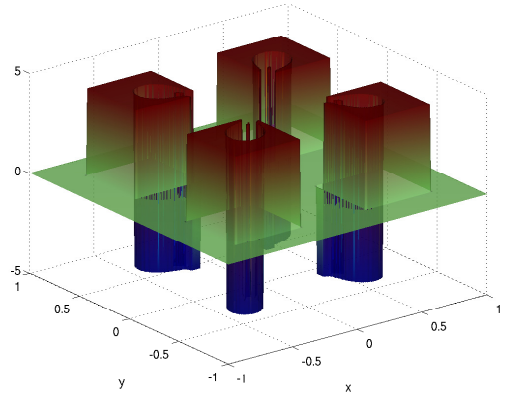
α	$5 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-3}$
c_α	0.8788	2.6066	6.8161



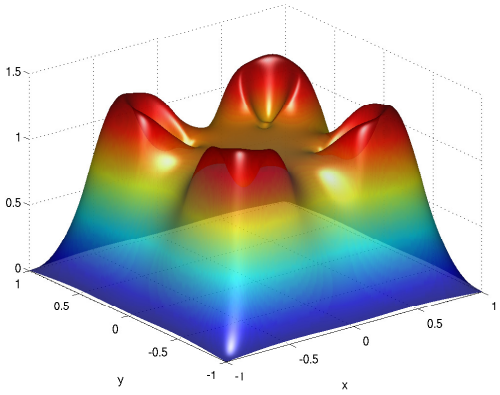
(a) optimal control (cu) for control domain ω_1



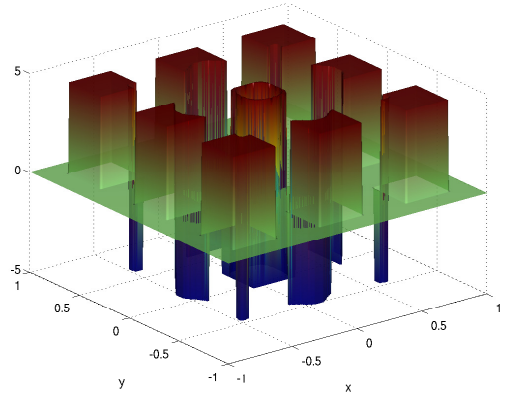
(b) optimal state y for control domain ω_1



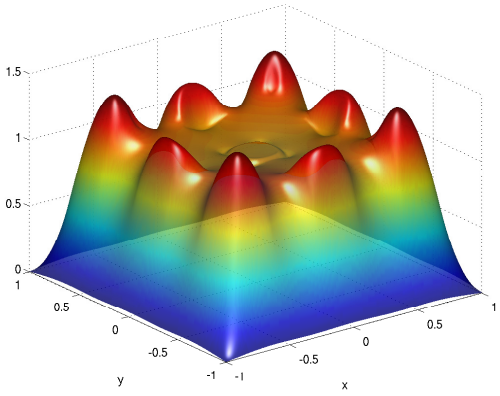
(c) optimal control (cu) for control domain ω_2



(d) optimal state y for control domain ω_2



(e) optimal control (cu) for control domain ω_3



(f) optimal state y for control domain ω_3

Figure 12.5: Optimal controls (cu) and states y for the target $z_3 = 1$ and different control domains ($\alpha = 10^{-3}$).

Table 12.2: Convergence of semismooth Newton method. Shown is the norm of the residual of (12.4.1) for the iterates (p^k, c^k).

k	0	1	2	3	4	5	6	7	8
$\|T(p^k, c^k)\|_{L^2}$	174	6.75	1.93	0.584	0.197	0.0224	$2.36 \cdot 10^{-4}$	$1.62 \cdot 10^{-7}$	$1.12 \cdot 10^{-10}$

12.6 CONCLUSION

A semismooth Newton technique based on an appropriate regularization was analyzed and investigated numerically for a class of minimum effort optimal control problems for elliptic equations. The numerical results show that while the unregularized minimum effort controls can be expected to be of bang-bang type, the regularized controls mostly assume values on the boundary of the admissible control set and are zero on open subsets of the control domain. This sparsity property could be of practical interest in itself, and can certainly be the focus of further investigations.

ACKNOWLEDGMENTS

The research of the first and last named authors was supported in part by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences"). The research of the second named author was partially supported by the Army Research Office under DAAD19-02-1-0394.

12.A PROOF OF PROPOSITION 12.3.1

It will be convenient to introduce the reduced cost

$$F(u, c) = \frac{1}{2}\|A^{-1}(cu) - z\|_{L^2}^2 + \frac{\beta c}{2}\|u\|_{L^2}^2 + \frac{\alpha}{2}c^2,$$

with the corresponding optimality conditions

$$\text{(12.A.1)} \quad \begin{cases} c_\beta \left\langle \beta u_\beta - A^{-*}z + c_\beta A^{-*}A^{-1}u_\beta, u - u_\beta \right\rangle_{L^2} \geqslant 0 \quad \text{for all } \|u\|_{L^\infty} \leqslant 1, \\ \alpha c_\beta + \dfrac{\beta}{2}\|u_\beta\|_{L^2}^2 - \langle u_\beta, A^{-*}z\rangle_{L^2} + c_\beta\|A^{-1}u_\beta\|_{L^2}^2 = 0. \end{cases}$$

To apply a Taylor expansion of F, we compute the partial derivatives of F at (u_β, c_β):

$$\begin{aligned}
F_{uu} &= c_\beta^2 A^{-*}A^{-1} + \beta c_\beta I,\\
F_{cc} &= \|A^{-1}u_\beta\|_{L^2} + \alpha,\\
F_{uc} &= 2c_\beta A^{-*}A^{-1}u_\beta - A^{-*}z + \beta u_\beta,\\
F_{uuc} &= 2c_\beta A^{-*}A^{-1} + \beta I,\\
F_{ccu} &= 2A^{-*}A^{-1}u_\beta,\\
F_{uucc} &= 2A^{-*}A^{-1},
\end{aligned}$$

with F_{ccc}, F_{uuu}, the remaining fourth-order and all higher-order derivatives being zero.

Now let (u, c) be any admissible pair and set

$$\hat{u} = u - u_\beta, \qquad \hat{c} = c - c_\beta.$$

Applying Taylor expansion of $F(u, c)$ at (u_β, c_β) and making use of (12.A.1), we find that (12.A.2)

$$\begin{aligned}
F(u, c) - F(u_\beta, c_\beta) &= c_\beta \left\langle \beta u_\beta - A^{-*}z + c_\beta A^{-*}A^{-1}u_\beta, \hat{u}\right\rangle_{L^2} + \frac{c_\beta^2}{2}\|A^{-1}\hat{u}\|_{L^2}^2\\
&\quad + \frac{\beta c_\beta}{2}\|\hat{u}\|_{L^2}^2 + \frac{1}{2}\|A^{-1}u_\beta\|_{L^2}^2 + \alpha\hat{c}^2\\
&\quad + \left\langle \beta u_\beta - A^{-*}z + c_\beta A^{-*}A^{-1}u_\beta, \hat{u}\right\rangle_{L^2}\hat{c}\\
&\quad + c_\beta \left\langle A^{-*}A^{-1}u_\beta, \hat{v}\right\rangle_{L^2}\hat{c} + c_\beta\|A^{-1}\hat{u}\|_{L^2}^2\hat{c}\\
&\quad + \left\langle A^{-1}u_\beta, A^{-1}\hat{u}\right\rangle_{L^2}\hat{c}^2 + \frac{1}{2}\|A^{-1}\hat{u}\|_{L^2}^2\hat{c}^2\\
&\geqslant \frac{c_\beta^2}{2}\|A^{-1}\hat{u}\|_{L^2}^2 + \frac{\beta}{2}(c_\beta + \hat{c})\|\hat{u}\|_{L^2}^2 + \frac{1}{2}\left(\|A^{-1}u_\beta\|_{L^2}^2 + \alpha\right)\hat{c}^2\\
&\quad + c_\beta \left\langle A^{-1}u_\beta, A^{-1}\hat{u}\right\rangle_{L^2}\hat{c} + c_\beta\|A^{-1}\hat{u}\|_{L^2}^2\hat{c}\\
&\quad + \left\langle A^{-1}u_\beta, A^{-1}\hat{u}\right\rangle_{L^2}\hat{c}^2 + \frac{1}{2}\|A^{-1}\hat{u}\|_{L^2}^2\hat{c}^2\\
&\geqslant \frac{c_\beta^2}{2}\left(1 - \frac{1}{\eta}\right)\|A^{-1}\hat{u}\|_{L^2}^2 + \frac{\beta}{2}(c_\beta + \hat{c})\|\hat{u}\|_{L^2}^2\\
&\quad + \frac{1}{2}\left(\alpha - \eta\|A^{-1}u_\beta\|_{L^2}^2\right)\hat{c}^2,
\end{aligned}$$

where we have used that

$$\left\langle A^{-1}u_\beta, A^{-1}\hat{u}\right\rangle_{L^2} = \left\langle \sqrt{\eta}\,A^{-1}u_\beta, (\sqrt{\eta})^{-1}A^{-1}\hat{u}\right\rangle_{L^2} \geqslant -\frac{\eta}{2}\|A^{-1}u_\beta\|_{L^2}^2 - \frac{1}{2\eta}\|A^{-1}\hat{u}\|_{L^2}^2$$

for every $\eta > 0$.

Let $K = \sup\left\{\|A^{-1}u\|_{L^2}^2 : \|u\|_{L^\infty} \leqslant 1\right\}$. Then the factors in front of $\|A^{-1}\hat{u}\|_{L^2}^2$ and $\hat{c}^2$ are non-negative if $\eta = 1$ and $\alpha > K^2$ holds. Under this condition, we have that

$$F(u, c) - F(u_\beta, c_\beta) \geqslant \frac{\beta}{2}(c_\beta + \hat{c})\|\hat{u}\|_{L^2}^2 + (\alpha - K^2)\hat{c}^2 + c_\beta\|A^{-1}\hat{u}\|_{L^2}^2\hat{c}. \tag{12.A.3}$$

Now for fixed $\beta > 0$, let (u_β, c_β) and (u'_β, c'_β) be two solutions to $(\mathcal{P}_\beta)$. Without loss of generality, we may assume that $c'_\beta \geqslant c_\beta$. Taking $(u, c) = (u'_\beta, c'_\beta)$, we deduce from (12.A.3) that $(u_\beta, c_\beta) = (u'_\beta, c'_\beta)$. Moreover, if for two possibly different solutions we have $c_\beta = c'_\beta$, then from (12.A.2) with any $\eta > 1$ we find $u_\beta = u'_\beta$. Conversely, if $u_\beta = u'_\beta$ holds, then by choosing $0 < \eta < \frac{\alpha}{K}$ we obtain $c_\beta = c'_\beta$.

12.B COMPARISON OF REGULARIZATIONS

In this section we compare the chosen regularization strategy, where the penalty term is scaled linearly with c, with two alternatives where the penalty term is constant or quadratic in c. We restrict the discussion to the case $c > 0$. First, we consider the regularization

$$\begin{cases} \min\limits_{c\in\mathbb{R}_+, u\in L^\infty} \dfrac{1}{2}\|y - z\|^2_{L^2} + \dfrac{\beta}{2}\|u\|^2_{L^2} + \dfrac{\alpha}{2}c^2 \\ \text{s.t. } Ay = cu \quad \text{in } \Omega, \\ \|u\|_{L^\infty} \leqslant 1. \end{cases}$$

The corresponding optimality system is

$$\begin{cases} 0 \in \beta u - cp + \partial I_{\{\|u\|_{L^\infty}\leqslant 1\}}, \\ 0 = \alpha c - \langle u, p\rangle, \\ 0 = y - z + A^* p, \\ 0 = Ay - cu. \end{cases}$$

Again, we can rewrite the first two relations by pointwise inspection as

$$u(x) = \operatorname{sign}_\beta(cp)(x) = \begin{cases} 1 & \text{if } cp(x) > \beta, \\ -1 & \text{if } cp(x) < -\beta, \\ \frac{1}{\beta}cp & \text{if } |cp(x)| \leqslant \beta \end{cases}$$

and

$$\alpha c = \frac{1}{c}\|cp\|_\beta := \int_{\{|cp|>\beta\}} |p(x)|\,dx + \frac{c}{\beta}\int_{\{|cp|\leqslant\beta\}} |p(x)|^2\,dx$$

to obtain the reduced system

$$\begin{cases} AA^* p + c\operatorname{sign}_\beta(cp) = Az, \\ \alpha c - \dfrac{1}{c}\|cp\|_\beta = 0, \end{cases}$$

Observe that now we have the product of c and p in the smoothed terms. If we formally compute the semismooth Newton step by fixing one variable and differentiating case by case,

we obtain (setting $\mathcal{A}_+ = \{x \in \Omega : cp(x) > \beta\}$ and so on) the system

$$\begin{cases} AA^*\delta p + c^2 \frac{1}{\beta}\chi_{\mathcal{I}}\delta p + (\chi_{\mathcal{A}_+} - \chi_{\mathcal{A}_-} + \frac{2}{\beta}c^k p^k)\delta c = -(AA^* p^k + c\,\mathrm{sign}_\beta(p^k) - Az), \\ \alpha\delta c - \langle(\chi_{\mathcal{A}_+} - \chi_{\mathcal{A}_-} + \frac{2}{\beta}c^k p^k), \delta p\rangle - \frac{1}{\beta}\|p^k\chi_{\mathcal{I}}\|_{L^2}\delta c = -(\alpha c - \|p^k\|_{L^1_\beta}). \end{cases}$$

To show that this defines a positive definite operator, we would now need to argue that the term $(\alpha - \frac{1}{\beta}\|p^k\chi_{\mathcal{I}}\|^2_{L^2})$ is positive.

On the other hand, we can regularize with the scaled control cu, which leads to

$$\begin{cases} \min\limits_{c\in\mathbb{R}_+, u\in L^\infty} \dfrac{1}{2}\|y - z\|^2_{L^2} + \dfrac{\beta}{2}\|cu\|^2_{L^2} + \dfrac{\alpha}{2}c^2 \\ \text{s. t. } Ay = cu \qquad \text{in } \Omega, \\ \|u\|_{L^\infty} \leqslant 1. \end{cases}$$

The corresponding optimality system is

$$\begin{cases} 0 \in \beta c^2 u - cp + \partial I_{\{\|u\|_{L^\infty}\leqslant 1\}}, \\ 0 = \alpha c + \beta c\|u\|^2_{L^2} - \langle u, p\rangle, \\ 0 = y - z + A^* p, \\ 0 = Ay - cu. \end{cases}$$

Pointwise inspection then allows expressing the first two relations as

$$u(x) = \mathrm{sign}_{\beta c}(p)(x) = \begin{cases} 1 & \text{if } p(x) > \beta c, \\ -1 & \text{if } p(x) < -\beta c, \\ \frac{1}{\beta c}p(x) & \text{if } |p(x)| \leqslant \beta c \end{cases}$$

and

$$\alpha c = \int_{\{|p|>\beta c\}} (|p(x)| - \beta c)\ dx.$$

The reduced optimality system is thus

$$\begin{cases} AA^* p + c\,\mathrm{sign}_{\beta c}(p) = Az, \\ \alpha c - \displaystyle\int_{\{|p|>\beta c\}} (|p(x)| - \beta c)\ dx = 0, \end{cases}$$

Here, we have c appearing in the definition of the smoothed functions.

Part IV

INVERSE PROBLEMS WITH NON-GAUSSIAN NOISE

A SEMISMOOTH NEWTON METHOD FOR L^1 DATA FITTING WITH AUTOMATIC CHOICE OF REGULARIZATION PARAMETERS AND NOISE CALIBRATION

13

ABSTRACT

This paper considers the numerical solution of inverse problems with an L^1 data fitting term, which is challenging due to the lack of differentiability of the objective functional. Utilizing convex duality, the problem is reformulated as minimizing a smooth functional with pointwise constraints, which can be efficiently solved using a semismooth Newton method. In order to achieve superlinear convergence, the dual problem requires additional regularization. For both the primal and the dual problem, the choice of the regularization parameters is crucial. We propose adaptive strategies for choosing these parameters. The regularization parameter in the primal formulation is chosen according to a balancing principle derived from the model function approach, whereas the one in the dual formulation is determined by a path-following strategy based on the structure of the optimality conditions. Several numerical experiments confirm the efficiency and robustness of the proposed method and adaptive strategy.

13.1 INTRODUCTION

This work is concerned with solving the inverse problem

$$Kx = y^\delta,$$

where $K : L^2(\Omega) \to L^2(\Omega)$ is a bounded linear operator, $\Omega \subset \mathbb{R}^n$ is a bounded domain, and $y^\delta \in L^2(\Omega)$ are noisy measurements with noise level $\|y^\dagger - y^\delta\|_{L^1} \leqslant \delta$ ($y^\dagger$ being the noise-free data). This problem is ill-posed in the sense of Hadamard, and in particular, the solution often fails to depend continuously on the data. The now standard approach is Tikhonov

regularization, which typically incorporates *a priori* information and amounts to solving a minimization problem of the form

$$\frac{1}{2}\|Kx - y^\delta\|^2_{L^2(\Omega)} + \alpha R(x)$$

where R is the regularization term, and α is a regularization parameter determining the relative weight of these two terms. The choice of the regularization term R is application dependent, and in the sequel, we shall focus on the choice $R(x) = \frac{1}{2}\|x\|^2_{L^2}$, which is suitable for smooth solutions.

The classical Tikhonov regularization uses a L^2 data fitting term, which statistically speaking is most appropriate for Gaussian noise. The success of this formulation relies crucially on the validity of the Gaussian assumption [Kärkkäinen, Kunisch, and Majava 2005] (no heavy tails and the noise distribution is symmetric); in some practical applications, however, the noise is non-Gaussian. For instance, the noise may follow a Laplace distribution as in certain inverse problems arising in signal processing [Alliney and Ruzinsky 1994]. Noise models of impulse type, e.g. salt-and-pepper or random-valued noise, arise in image processing because of malfunctioning pixels in camera sensors, faulty memory locations in hardware, or transmission in noisy channels [Bovik 2005], and call for the use of L^1 data fitting. The advantage of using the L^1 norm is given by the fact that the solution is more robust when compared to the L^2 norm [Huber 1981]. In particular, a small number of outliers has less influence on the solution, whereas the L^2 formulation needs some extra processing stage utilizing robust procedures to locate the outliers [Rousseeuw and Leroy 1987]. These considerations motivate the formulation

$$\mathcal{J}_\alpha(x) = \|Kx - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x\|^2_{L^2}.$$

Recently, minimization of cost functions involving L^1 data fitting have received growing interests in diverse disciplines, e.g. signal processing [Alliney and Ruzinsky 1994; Alliney 1997], image processing [Nikolova 2002; Nikolova 2004; Chan and Esedoğlu 2005; Kärkkäinen, Kunisch, and Majava 2005] and distributed parameter identification [Chaabane, Ferchichi, and Kunisch 2004]. Alliney [Alliney 1997] studied the properties of a discrete variational problem and established its relation with recursive median filters. Nikolova [Nikolova 2002] showed that in L^1 data fitting for discrete denoising problems, a certain number of data points can be attained exactly, and thus theoretically justified its superior performance over the standard model for certain type of noise. Chan and Esedoğlu [Chan and Esedoğlu 2005] investigated the analytical properties of minimizers and their implication for multiscale image decomposition and parameter selection in the context of total variation image denoising. These results were recently extended and refined by a number of authors [Yin, Goldfarb, and Osher 2007; Allard 2007/08; Duval, Aujol, and Gousseau 2009].

Numerical methods for the solution of L^1 data fitting problems have also received some attention, see for instance [Kärkkäinen, Kunisch, and Majava 2005] (an active set algorithm

for denoising), [Fu et al. 2006] (an interior point algorithm for image restoration problems), [Rodríguez and Wohlberg 2009] (a generalization of the classical iteratively reweighted least squares method), [Yang, Zhang, and Yin 2009] (an alternating minimization algorithm for color image restoration), and [Chan, Dong, and Hintermüller 2010] (a primal-dual algorithm for image restoration). Note that these studies focus on structured matrices, e.g. identity or (block) Toeplitz, instead of general (infinite-dimensional) operators. Except [Kärkkäinen, Kunisch, and Majava 2005], the above-mentioned works focus on total variation regularization because of their interests in image processing.

This paper focuses on the efficient numerical solution of the $L^1(\Omega)$ data fitting problem in infinite dimensions using a semismooth Newton method and on the automatic choice of the regularization parameter with a model function approach. Semismooth Newton methods were applied to inverse problems with L^1-type functionals in, e.g., [Griesse and Lorenz 2008] (for sparsity constrained ℓ^2-minimization) and [Dong, Hintermüller, and Neri 2009] (for ℓ^1-tv image restoration). One particular advantage of semismooth Newton methods in function spaces [Hintermüller, Ito, and Kunisch 2002; Ulbrich 2002; Ito and Kunisch 2008] is their mesh independence (i.e., the necessary number of iterations is independent on the problem size). The model function approach was originally proposed for efficiently solving Morozov's discrepancy equation [Ito and Kunisch 1992; Kunisch and Zou 1998; Xie and Zou 2002]. However, the discrepancy principle requires an accurate estimate of the noise level δ, which might be unavailable in practice. Therefore, it is useful to estimate the noise level and to develop heuristic parameter choice rules based on this estimate. In [Kunisch and Zou 1998] it was suggested to estimate the noise level using an iterative approach involving model functions. Numerically, it was observed that the method gives an excellent approximation of the noise level δ after two or three iterations. However, the method was formulated for least-squares data fitting problems, and also the mechanism of the iteration remained unexplored.

Our main contributions are as follows. Firstly, we propose and analyze an efficient semismooth Newton method for solving the L^1 data fitting problem. Secondly, we derive heuristic choice rules for the regularization parameters in the primal and dual problems based on the idea of balancing, which do not require knowledge of the noise level. Thirdly, the convergence property of a fixed point iteration for the automatic parameter choice is investigated.

This paper is organized as follows. In section 13.2, we treat the primal and a dual formulation of the problem. The regularizing properties, especially the convergence rate results for *a priori* and *a posteriori* parameter choice rules, are shown, and optimality conditions are established. Section 13.3 is devoted to the solution of the dual problem using a semismooth Newton method. The additional regularization guaranteeing superlinear convergence is discussed in § 13.3.1, while § 13.3.2 concerns the convergence of the semismooth Newton method. Our adaptive rules for choosing regularization parameters in the primal and dual problems are presented in section 13.4. We conclude with several numerical experiments involving test problems in one and two dimensions.

13.2 PROPERTIES OF MINIMIZERS

13.2.1 PRIMAL PROBLEM

In this section, we consider the properties of the primal problem

$$(\mathcal{P}) \qquad \min_{x\in L^2(\Omega)} \left\{ \mathcal{J}_\alpha(x) \equiv \|Kx - y^\delta\|_{L^1(\Omega)} + \frac{\alpha}{2}\|x\|_{L^2}^2 \right\}.$$

The functional $\mathcal{J}_\alpha$ is strictly convex, and thus has a unique minimizer x_α. The next result, whose proof is rather standard and is thus omitted [Engl, Hanke, and Neubauer 1996], summarizes the regularizing property of the functional $\mathcal{J}_\alpha$. For the next result as well as for Theorem 2.3 and Theorem 2.4, we shall assume that the noise free data $y^\dagger$ is attainable, i.e. there exists some $x^\dagger \in L^2(\Omega)$ such that $y^\dagger = Kx^\dagger$.

Theorem 13.2.1. *For each fixed α, there exists a unique minimizer x_α to the functional $\mathcal{J}_\alpha$ which depends continuously on the data y^δ. Moreover, if the regularization parameter α satisfies $\alpha \to 0$, and if $\lim_{\delta\to 0^+} \delta/\alpha = 0$, then x_α converges to $x^\dagger$, a minimum norm solution of the inverse problem, as $\delta \to 0$.*

We also need the following results on properties of the value function

$$F(\alpha) = \|Kx_\alpha - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2.$$

The proofs can be found in [Ito, Jin, and Zou 2011; Jin, Zhao, and Zou 2012].

Theorem 13.2.2. *The functions $\|Kx_\alpha - y^\delta\|_{L^1}$ and $\|x_\alpha\|_{L^2}^2$ are continuous, and, respectively, monotonically increasing and decreasing with respect to α in the sense that*

$$(\alpha_1 - \alpha_2)(\|Kx_{\alpha_1} - y^\delta\|_{L^1} - \|Kx_{\alpha_2} - y^\delta\|_{L^1}) \geqslant 0,$$
$$(\alpha_1 - \alpha_2)(\|x_{\alpha_1}\|_{L^2}^2 - \|x_{\alpha_2}\|_{L^2}^2) \leqslant 0.$$

The value function $F(\alpha)$ is continuous and increasing, and it is differentiable with derivative

$$F'(\alpha) = \frac{1}{2}\|x_\alpha\|_{L^2}^2.$$

The next result shows a convergence rate result for *a priori* parameter choice rules under certain source conditions. To explicitly indicate the dependence of the minimizer x_α on the data y^δ, we shall use the notation x_α^δ for the next two results. We will also assume that the following source condition holds: There exists some $w \in L^\infty(\Omega)$ such that the exact solution $x^\dagger$ satisfies

$$x^\dagger = K^* w. \tag{13.2.1}$$

Theorem 13.2.3. *Assume that the source condition (13.2.1) holds. Then for sufficiently small δ and $\alpha = \mathcal{O}(\delta^\varepsilon)$ with $\varepsilon \in (0,1)$, the minimizer x_α^δ of the functional $\mathcal{J}_\alpha$ satisfies*

$$\|x_\alpha^\delta - x^\dagger\|_{L^2} \leqslant c\delta^{\frac{1-\varepsilon}{2}}.$$

Proof. By the minimizing property of x_α^δ, we have

$$\|Kx_\alpha^\delta - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha^\delta\|_{L^2}^2 \leqslant \|Kx^\dagger - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x^\dagger\|_{L^2}^2 \leqslant \delta + \frac{\alpha}{2}\|x^\dagger\|_{L^2}^2. \tag{13.2.2}$$

Therefore, we have

$$\|Kx_\alpha^\delta - y^\delta\|_{L^1} + \frac{\alpha}{2}(\|x_\alpha^\delta\|_{L^2}^2 - \|x^\dagger\|_{L^2}^2 - 2\langle x^\dagger, x_\alpha^\delta - x^\dagger\rangle) \leqslant \delta - \alpha\langle x^\dagger, x_\alpha^\delta - x^\dagger\rangle,$$

i.e.,

$$\|Kx_\alpha^\delta - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_{L^2}^2 \leqslant \delta - \alpha\langle x^\dagger, x_\alpha^\delta - x^\dagger\rangle,$$

Now by the source condition (13.2.1) and the triangle inequality we have

$$\begin{aligned}\|Kx_\alpha^\delta - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_{L^2}^2 &\leqslant \delta - \alpha\langle K^* w, x_\alpha^\delta - x^\dagger\rangle \\ &= \delta - \alpha\langle w, Kx_\alpha^\delta - y^\dagger\rangle \\ &\leqslant \delta + \alpha\|w\|_{L^\infty}(\|Kx_\alpha^\delta - y^\delta\|_{L^1} + \|y^\delta - y^\dagger\|_{L^1}).\end{aligned}$$

Rearranging the terms gives

$$(1 - \alpha\|w\|_{L^\infty})\|Kx_\alpha^\delta - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_{L^2}^2 \leqslant (1 + \alpha\|w\|_{L^\infty})\delta.$$

The desired convergence rate now follows using $\alpha = \mathcal{O}(\delta^\varepsilon)$. □

Next we consider Morozov's classical discrepancy principle [Morozov 1966], which consists in choosing α as a solution of the following nonlinear equation

$$\|Kx_\alpha^\delta - y^\delta\|_{L^1} = c\delta, \tag{13.2.3}$$

for some $c \geqslant 1$. Under the condition that $\lim_{\alpha\to 0^+}\|Kx_\alpha - y^\delta\|_{L^1} < c\delta$ and that $\lim_{\alpha\to\infty}\|Kx_\alpha - y^\delta\|_{L^1} = \|y^\delta\|_{L^1} > c\delta$, there exists at least one positive solution to Morozov's equation (13.2.3), which follows from the continuity and monotonicity results, see Theorem 13.2.2.

In contrast to *a priori* choice rules, the discrepancy principle yields a concrete scheme for determining an appropriate regularization parameter, and is mathematically rigorous in that consistency and convergence rates can be established [Engl, Kunisch, and Neubauer 1989; Engl, Hanke, and Neubauer 1996]. Its applicability for $\mathcal{J}_\alpha$ follows directly from the results in [Jin, Zhao, and Zou 2012].

The next result shows a convergence rate result for the discrepancy principle. We point out that there have been few convergence rate results for discrepancy principle for Tikhonov functional other than the classical L^2-L^2 formulation.

Theorem 13.2.4. *Assume that the source condition (13.2.1) holds, and that the regularization parameter $\alpha = \alpha(\delta)$ is determined according to the discrepancy principle. Then the minimizer x_α^δ of the functional $\mathcal{J}_\alpha$ satisfies*

$$\|x_\alpha^\delta - x^\dagger\|_{L^2} \leqslant [2(c+1)\|w\|_{L^\infty}]^{\frac{1}{2}} \delta^{\frac{1}{2}}.$$

Proof. Relation (13.2.2) together with Morozov's equation (13.2.3) for $\alpha(\delta)$ gives

$$\|x_\alpha^\delta\|_{L^2}^2 \leqslant \|x^\dagger\|_{L^2}^2,$$

from which it follows that

$$\begin{aligned}
\|x_\alpha^\delta - x^\dagger\|_{L^2}^2 &\leqslant 2\langle x^\dagger, x^\dagger - x_\alpha^\delta\rangle = 2\langle K^* w, x^\dagger - x_\alpha^\delta\rangle \\
&\leqslant 2\|w\|_{L^\infty}\|Kx^\dagger - Kx_\alpha^\delta\|_{L^1} \\
&\leqslant 2\|w\|_{L^\infty}(\|Kx^\dagger - y^\delta\|_{L^1} + \|Kx_\alpha^\delta - y^\delta\|_{L^1}) \\
&\leqslant 2\|w\|_{L^\infty}(\delta + c\delta),
\end{aligned}$$

again by (13.2.3). This yields the desired estimate. □

13.2.2 DUAL PROBLEM

In this section, we consider the problem

$$(\mathcal{P}^*) \qquad \begin{cases} \min\limits_{p \in L^2(\Omega)} \dfrac{1}{2\alpha}\|K^* p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2} \\ \text{s. t.} \quad \|p\|_{L^\infty} \leqslant 1, \end{cases}$$

which we will show to be the dual problem of ($\mathcal{P}$).

Theorem 13.2.5. *The dual problem of ($\mathcal{P}$) is ($\mathcal{P}^*$), which has at least one solution, and the solutions $x_\alpha \in L^2(\Omega)$ of ($\mathcal{P}$) and $p_\alpha \in L^2(\Omega)$ of ($\mathcal{P}^*$) are related by*

$$(13.2.4) \qquad \begin{cases} K^* p_\alpha = \alpha x_\alpha, \\ \quad 0 \leqslant \langle Kx_\alpha - y^\delta, p - p_\alpha\rangle_{L^2}, \end{cases}$$

for all $p \in L^2(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$.

Proof. We apply Fenchel duality [Ekeland and Témam 1999], setting

$$\begin{aligned}
&\mathcal{F} : L^2(\Omega) \to \mathbb{R}, && \mathcal{F}(x) = \frac{\alpha}{2}\|x\|_{L^2}^2, \\
&\mathcal{G} : L^2(\Omega) \to \mathbb{R}, && \mathcal{G}(x) = \|x - y^\delta\|_{L^1}, \\
&\Lambda : L^2(\Omega) \to L^2(\Omega), && \Lambda x = Kx.
\end{aligned}$$

The Fenchel conjugates of $\mathcal{F}$ and $\mathcal{G}$ are given by

$$\begin{aligned} \mathcal{F}^* : L^2(\Omega) \to \mathbb{R}, \qquad & \mathcal{F}^*(q) = \frac{1}{2\alpha}\|q\|_{L^2}^2, \\ \mathcal{G}^* : L^2(\Omega) \to \mathbb{R} \cup \{\infty\}, \qquad & \mathcal{G}^*(q) = \begin{cases} \langle q, y^\delta \rangle_{L^2} & \text{if } \|q\|_{L^\infty} \leqslant 1, \\ \infty & \text{if } \|q\|_{L^\infty} > 1. \end{cases} \end{aligned}$$

Since the functionals $\mathcal{F}$ and $\mathcal{G}$ are convex lower semicontinuous, proper and continuous at $x_0 = 0 = Kx_0$, and K is a continuous linear operator, the Fenchel duality theorem states that

$$\inf_{x \in L^2(\Omega)} \mathcal{F}(x) + \mathcal{G}(\Lambda x) = \sup_{p \in L^2(\Omega)} -\mathcal{F}^*(\Lambda^* p) - \mathcal{G}^*(-p), \tag{13.2.5}$$

holds, and that the right-hand side of (13.2.5) has at least one solution.

Furthermore, the equality in (13.2.5) is attained at (x_α, p_α) if and only if

$$\begin{cases} \Lambda^* p_\alpha \in \partial \mathcal{F}(x_\alpha), \\ -p_\alpha \in \partial \mathcal{G}(\Lambda x_\alpha). \end{cases}$$

Since $\mathcal{F}$ is Fréchet differentiable, the first relation of (13.2.4) follows by direct calculation. Recall that by the definition of the subgradient

$$-p_\alpha \in \partial \mathcal{G}(\Lambda x_\alpha) \Leftrightarrow \Lambda x_\alpha \in \partial \mathcal{G}^*(-p_\alpha)$$

holds. Subdifferential calculus then yields

$$\Lambda x_\alpha - y^\delta \in \partial I_{\{\|-p_\alpha\|_{L^\infty} \leqslant 1\}},$$

where I_S denotes the indicator function of the set S, whose subdifferential coincides with the normal cone at S (cf., e.g., [Ito and Kunisch 2008, Ex. 4.21]). We thus obtain that

$$0 \geqslant \langle Kx_\alpha - y^\delta, p + p_\alpha \rangle_{L^2}$$

for all $p \in L^2(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$, from which the second relation follows. □

Remark 13.2.6. The solution of problem ($\mathcal{P}^*$) is no longer unique, rather any solution p_α can be written as $p_\alpha = \tilde{p}_\alpha + \hat{p}_\alpha$ with $\tilde{p}_\alpha \in \ker K^*$ and a unique $\hat{p}_\alpha \in (\ker K^*)^\perp$. Nevertheless, the corresponding primal solution x_α calculated using the first extremality relation (13.2.4) will be unique. The treatment of the non-uniqueness in the numerical solution of problem ($\mathcal{P}^*$) will be discussed in section 13.3.1.

Assisted with Theorem 13.2.5, we can now derive the first order optimality conditions for problem ($\mathcal{P}^*$).

Corollary 13.2.7. *Let $p_\alpha \in L^2(\Omega)$ be a solution of ($\mathcal{P}^*$). Then there exists $\lambda_\alpha \in L^2(\Omega)$ such that*

$$\begin{cases} \frac{1}{\alpha} K K^* p_\alpha - y^\delta + \lambda_\alpha = 0, \\ \qquad \langle \lambda_\alpha, p - p_\alpha \rangle_{L^2} \leqslant 0, \end{cases} \tag{13.2.6}$$

holds for all $p \in L^2(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$.

Proof. By applying $\frac{1}{\alpha} K$ to the first relation in (13.2.4) and setting $\lambda_\alpha = -(Kx_\alpha - y^\delta)$, we immediately obtain the existence of a Lagrange multiplier satisfying (13.2.6). □

The following structural information for the solution of problem ($\mathcal{P}$) is a direct consequence of (13.2.4).

Corollary 13.2.8. *Let x_α be the minimizer of ($\mathcal{P}$). Then the following holds for any $p \in L^2(\Omega)$, $p \geqslant 0$:*

$$\begin{aligned} \langle Kx_\alpha - y^\delta, p \rangle_{L^2} &= 0 \qquad \text{if } \operatorname{supp} p \subset \{x : |p_\alpha(x)| < 1\}, \\ \langle Kx_\alpha - y^\delta, p \rangle_{L^2} &\geqslant 0 \qquad \text{if } \operatorname{supp} p \subset \{x : p_\alpha(x) = 1\}, \\ \langle Kx_\alpha - y^\delta, p \rangle_{L^2} &\leqslant 0 \qquad \text{if } \operatorname{supp} p \subset \{x : p_\alpha(x) = -1\}. \end{aligned}$$

This can be interpreted as follows: the bound constraint on the dual solution p_α is active where the data is not attained by the primal solution x_α.

13.3 SOLUTION BY SEMISMOOTH NEWTON METHOD

13.3.1 REGULARIZATION

If the inversion of K is ill-posed, problem ($\mathcal{P}^*$) remains ill-posed in spite of the pointwise bounds on p. To counter this and to ensure superlinear convergence of the semismooth Newton method for solving the constrained optimization problem, we introduce the regularized problem

$$\begin{cases} \min\limits_{p \in H^1(\Omega)} \dfrac{1}{2\alpha} \|K^* p\|_{L^2}^2 + \dfrac{\beta}{2} \|\nabla p\|_{L^2}^2 - \langle p, y^\delta \rangle_{L^2} \\ \text{s.t.} \quad \|p\|_{L^\infty} \leqslant 1, \end{cases} \tag{$\mathcal{P}^*_\beta$}$$

for $\beta > 0$. The interplay between the pointwise bound on p and the semi-norm regularization term will enable an easy choice for the regularization parameter β, which will be explained in § 13.4.2. We assume that $\ker K^* \cap \ker \nabla = \{0\}$, i.e. constant functions do not belong to the kernel of K^*. Under this assumption the inner product $\frac{1}{\alpha}\langle K^* \cdot, K^* \cdot \rangle + \beta \langle \nabla \cdot, \nabla \cdot \rangle$ induces an

equivalent norm on $H^1(\Omega)$, and problem $(\mathcal{P}^*_\beta)$ admits a unique solution p_β. This assumption can be removed if the semi-norm regularization is replaced by the full $H^1(\Omega)$ norm.

To solve $(\mathcal{P}^*_\beta)$ numerically, we introduce a Moreau-Yosida regularization of the box constraints and consider

$$(\mathcal{P}^*_{\beta,c}) \quad \min_{p\in H^1} \frac{1}{2\alpha}\|K^*p\|^2_{L^2} + \frac{\beta}{2}\|\nabla p\|^2_{L^2} - \langle p, y^\delta\rangle_{L^2} + \frac{1}{2c}\|\max(0, c(p-1))\|^2_{L^2} + \frac{1}{2c}\|\min(0, c(p+1))\|^2_{L^2}.$$

for $c > 0$, where the max and min are taken pointwise. For fixed β and c, under the above assumption, $\frac{1}{2\alpha}\|K^*p\|^2_{L^2} + \frac{\beta}{2}\|\nabla p\|^2_{L^2}$ is strictly convex and hence problem $(\mathcal{P}^*_{\beta,c})$ admits a unique minimizer p_c. The optimality system for $(\mathcal{P}^*_{\beta,c})$ is given by

$$\begin{cases} \frac{1}{\alpha}KK^*p_c - \beta\Delta p_c - y^\delta + \lambda_c = 0, \\ \lambda_c = \max(0, c(p_c-1)) + \min(0, c(p_c+1)). \end{cases} \tag{13.3.2}$$

This yields higher regularity for the Lagrange multiplier λ_c: Since the max (and min) operator is continuous from $W^{1,\infty}(\Omega)$ to $W^{1,\infty}(\Omega)$ (cf., e.g., [Brezis 1983]) and $p_c \in H^1(\Omega)$ by construction, the second equation of (13.3.2) ensures that $\lambda_c \in H^1(\Omega)$ as well.

Now we address the convergence as $c \to \infty$ of the solutions of (13.3.2) to that of problem $(\mathcal{P}^*_\beta)$. To this end, we introduce the optimality system for problem $(\mathcal{P}^*_\beta)$:

$$\begin{cases} \frac{1}{\alpha}KK^*p_\beta - \beta\Delta p_\beta - y^\delta + \lambda_\beta = 0, \\ \langle \lambda_\beta, p - p_\beta\rangle_{L^2} \leqslant 0, \end{cases} \tag{13.3.3}$$

for all $p \in H^1(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$ and a $\lambda_\beta \in (H^1(\Omega))^*$, the dual space of $H^1(\Omega)$.

Theorem 13.3.1. *For $\beta > 0$ fixed, let $(p_c, \lambda_c) \in H^1(\Omega) \times (H^1(\Omega))^*$ be the solution of (13.3.2) for $c > 0$, and $(p_\beta, \lambda_\beta) \in H^1(\Omega) \times (H^1(\Omega))^*$ the solution of (13.3.3). Then we have as $c \to \infty$:*

$$\begin{aligned} p_c &\to p_\beta \quad \text{in } H^1(\Omega), \\ \lambda_c &\rightharpoonup \lambda_\beta \quad \text{in } (H^1(\Omega))^*. \end{aligned}$$

Proof. From the optimality system (13.3.2), we have that pointwise in $x \in \Omega$

$$\lambda_c p_c = \max(0, c(p_c-1))p_c + \min(0, c(p_c+1))p_c = \begin{cases} c(p_c-1)p_c, & p_c \geqslant 1, \\ 0, & |p_c| < 1, \\ c(p_c+1)p_c, & p_c \leqslant -1 \end{cases}$$

holds and thus, since $\lambda_c \in W^{1,\infty}(\Omega) \subset L^2(\Omega)$, that

$$\langle \lambda_c, p_c\rangle_{L^2} \geqslant \frac{1}{c}\|\lambda_c\|^2_{L^2}. \tag{13.3.4}$$

Inserting p_c as a test function in the variational form of (13.3.2),

$$\frac{1}{\alpha}\langle K^* p_c, K^* v\rangle_{L^2} + \beta\,\langle \nabla p_c, \nabla v\rangle_{L^2} - \langle y^\delta, v\rangle_{L^2} + \langle \lambda_c, v\rangle_{H^{1*},H^1} = 0, \tag{13.3.5}$$

for all $v \in H^1(\Omega)$, yields

$$\frac{1}{\alpha}\|K^* p_c\|_{L^2}^2 + \beta\|\nabla p_c\|_{L^2}^2 + \frac{1}{c}\|\lambda_c\|_{L^2}^2 \leqslant \|p_c\|_{L^2}\|y^\delta\|_{L^2}, \tag{13.3.6}$$

and by recalling that by assumption the first two terms define an equivalent norm on $H^1(\Omega)$, we deduce that $\|p_c\|_{H^1} \leqslant C\left\|y^\delta\right\|_{L^2}$ for some constant C. Moreover,

$$\begin{aligned}
\|\lambda_c\|_{(H^1(\Omega))^*} &= \sup_{\substack{v\in H^1(\Omega),\\ \|v\|_{H^1}\leqslant 1}} \langle \lambda_c, v\rangle_{H^{1*},H^1} \\
&= \sup_{\substack{v\in H^1(\Omega),\\ \|v\|_{H^1}\leqslant 1}} \left[-\frac{1}{\alpha}\langle K^* p_c, K^* v\rangle_{L^2} - \beta\,\langle \nabla p_c, \nabla v\rangle_{L^2} + \left\langle y^\delta, v\right\rangle_{L^2}\right] \\
&\leqslant \sup_{\substack{v\in H^1(\Omega),\\ \|v\|_{H^1}\leqslant 1}} \left[C_1\|p_c\|_{H^1}\|v\|_{H^1} + \|v\|_{L^2}\|y^\delta\|_{L^2}\right] \\
&\leqslant (CC_1+1) \sup_{\substack{v\in H^1(\Omega),\\ \|v\|_{H^1}\leqslant 1}} \|v\|_{H^1}\ \|y^\delta\|_{L^2} =: K < \infty,
\end{aligned}$$

where C_1 is another norm equivalence constant. Thus, (p_c, λ_c) is uniformly bounded in $H^1(\Omega) \times (H^1(\Omega))^*$, and there exists some $(\tilde{p}, \tilde{\lambda}) \in H^1(\Omega) \times (H^1(\Omega))^*$ such that

$$(p_c, \lambda_c) \rightharpoonup (\tilde{p}, \tilde{\lambda}) \quad \text{in } H^1(\Omega) \times (H^1(\Omega))^*.$$

Passing to the limit in (13.3.5), we obtain

$$\frac{1}{\alpha}\langle K^*\tilde{p}, K^* v\rangle_{L^2} + \beta\,\langle \nabla\tilde{p}, \nabla v\rangle_{L^2} - \langle y^\delta, v\rangle_{L^2} + \left\langle\tilde{\lambda}, v\right\rangle_{H^{1*},H^1} = 0$$

for all $v \in H^1(\Omega)$.

We next verify the feasibility of $\tilde{p}$. By pointwise inspection similar to (13.3.4), we obtain that

$$\frac{1}{c}\|\lambda_c\|_{L^2}^2 = c\,\|\max(0, p_c - 1)\|_{L^2}^2 + c\,\|\min(0, p_c + 1)\|_{L^2}^2.$$

From (13.3.6), we have that $\frac{1}{c}\|\lambda_c\|_{L^2}^2 \leqslant C\left\|y^\delta\right\|_{L^2}^2$, so that

$$\begin{aligned}
\|\max(0, p_c - 1)\|_{L^2}^2 &\leqslant \frac{1}{c} C\|y^\delta\|_{L^2}^2 \to 0, \\
\|\min(0, p_c + 1)\|_{L^2}^2 &\leqslant \frac{1}{c} C\|y^\delta\|_{L^2}^2 \to 0
\end{aligned}$$

as $c \to \infty$. Since $p_c \to \tilde{p}$ strongly in $L^2(\Omega)$, this implies that

$$-1 \leqslant \tilde{p}(x) \leqslant 1 \quad \text{for all } x \in \Omega.$$

It remains to show that the second equation of (13.3.3) holds. First, the minimizing property of p_c yields that

$$\frac{1}{2\alpha}\|K^* p_c\|_{L^2}^2 + \frac{\beta}{2}\|\nabla p_c\|_{L^2}^2 - \langle p_c, y^\delta\rangle_{L^2} \leqslant \frac{1}{2\alpha}\|K^* p\|_{L^2}^2 + \frac{\beta}{2}\|\nabla p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2}$$

holds for all feasible $p \in H^1(\Omega)$. Therefore, we have that

$$\limsup_{c\to\infty}\left[\frac{1}{2\alpha}\|K^* p_c\|_{L^2}^2 + \frac{\beta}{2}\|\nabla p_c\|_{L^2}^2 - \langle p_c, y^\delta\rangle_{L^2}\right] \leqslant \frac{1}{2\alpha}\|K^*\tilde{p}\|_{L^2}^2 + \frac{\beta}{2}\|\nabla\tilde{p}\|_{L^2}^2 - \langle\tilde{p}, y^\delta\rangle_{L^2}$$

and consequently $p_c \to \tilde{p}$ strongly in $H^1(\Omega)$. Now observe that

$$\langle\lambda_c, p - p_c\rangle_{H^{1*},H^1} = \langle\max(0, c(p_c - 1)), p - p_c\rangle_{L^2} + \langle\min(0, c(p_c + 1)), p - p_c\rangle_{L^2} \leqslant 0$$

holds for all $p \in H^1(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$. Thus

$$\langle\tilde{\lambda}, p - \tilde{p}\rangle_{H^{1*},H^1} \leqslant 0$$

is satisfied for all $p \in H^1(\Omega)$ with $\|p\|_{L^\infty} \leqslant 1$. Therefore, $(\tilde{p}, \tilde{\lambda}) \in H^1(\Omega) \times (H^1(\Omega))^*$ satisfies (13.3.3), and since the solution of (13.3.3) is unique, $\tilde{p} = p_\beta$ and $\tilde{\lambda} = \lambda_\beta$ follows. □

Next we address the convergence of the solution of $(\mathcal{P}^*_\beta)$ as $\beta \to 0$ to a solution of $(\mathcal{P}^*)$, which might be nonunique if the operator K is not injective. The functional in $(\mathcal{P}^*)$ is convex, as is the set of all its minimizers, and thus if problem $(\mathcal{P}^*)$ admits a solution in $H^1(\Omega)$, this set has an element with minimal $H^1(\Omega)$-semi-norm, denoted by $p^\dagger$.

Theorem 13.3.2. *Let $\{\beta_n\} \to 0$. Then the sequence of minimizers $\{p_{\beta_n}\}$ of $(\mathcal{P}^*_\beta)$ has a subsequence converging weakly to a minimizer of problem $(\mathcal{P}^*)$. If the operator K is injective or there exists a unique $p^\dagger$ as defined above, then the whole sequence converges weakly to $p^\dagger$.*

Proof. Since the minimizers $p_n := p_{\beta_n}$ of $(\mathcal{P}^*_\beta)$ satisfy $\|p_n\|_{L^\infty} \leqslant 1$, the sequence $\{p_n\}$ is uniformly bounded in $L^2(\Omega)$ independently of n. Therefore, there exists a subsequence, also denoted by $\{p_n\}$, converging weakly in $L^2(\Omega)$ to some $p^* \in L^2(\Omega)$. By the weak lower semi-continuity of norms, we have that

$$\|K^* p^*\|_{L^2}^2 \leqslant \liminf_{n\to\infty}\|K^* p_n\|_{L^2}^2, \qquad \langle p^*, y^\delta\rangle_{L^2} = \lim_{n\to\infty}\langle p_n, y^\delta\rangle_{L^2},$$

and

$$\|p^*\|_{L^\infty} \leqslant \liminf_{n\to\infty}\|p_n\|_{L^\infty} \leqslant 1.$$

Therefore, by the minimizing property of p_n, for any fixed $p \in H^1(\Omega)$ we have that

$$\begin{aligned}\frac{1}{2\alpha}\|K^*p^*\|_{L^2}^2 - \langle p^*, y^\delta\rangle_{L^2} &\leqslant \liminf_{n\to\infty}\left(\frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2\right)\\ &\leqslant \liminf_{n\to\infty}\left(\frac{1}{2\alpha}\|K^*p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p\|_{L^2}^2\right)\\ &= \frac{1}{2\alpha}\|K^*p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2}.\end{aligned}$$

Therefore, p^* is a minimizer of problem ($\mathcal{P}^*$) over $H^1(\Omega)$. Now the density of $H^1(\Omega)$ in $L^2(\Omega)$ shows that p^* is also a minimizer of problem ($\mathcal{P}^*$) over $L^2(\Omega)$.

Now by the minimizing properties of $p^\dagger$ and p_n, we have that

$$\frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2 \leqslant \frac{1}{2\alpha}\|K^*p^\dagger\|_{L^2}^2 - \langle p^\dagger, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p^\dagger\|_{L^2}^2$$

and

$$\frac{1}{2\alpha}\|K^*p^\dagger\|_{L^2}^2 - \langle p^\dagger, y^\delta\rangle_{L^2} \leqslant \frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2}.$$

Adding these two inequalities together, we deduce that

$$\|\nabla p_n\|_{L^2}^2 \leqslant \|\nabla p^\dagger\|_{L^2}^2,$$

which together with the weak lower-semicontinuity of the $H^1(\Omega)$-semi-norm yields

$$\|\nabla p^*\|_{L^2}^2 \leqslant \|\nabla p^\dagger\|_{L^2}^2,$$

i.e. p^* is a minimizer with minimal $H^1(\Omega)$-semi-norm. If K is injective or $p^\dagger$ is unique, then it follows that $p^* = p^\dagger$. Consequently, each subsequence has a subsequence converging weakly to $p^\dagger$, and the whole sequence converges weakly. □

Remark 13.3.3. For the numerical solution of the dual problem, we will let $\beta \to 0$ for some fixed $c > 0$ (cf. § 13.4.2). Thus it is useful to have the convergence of the solution $p_c = p_{\beta,c}$ of ($\mathcal{P}^*_{\beta,c}$) as $\beta \to 0$. The proof of this result is similar to that of Theorem 13.3.2, and is given in Appendix 13.A.

Remark 13.3.4. For completeness, we also state how the regularization introduced in this section affects the primal problem. Setting

$$\begin{aligned}&\mathcal{F} : L^2(\Omega) \to \mathbb{R}, && \mathcal{F}(p) = -\langle p, y^\delta\rangle_{L^2} + \frac{1}{2c}\|\max(0, c(p-1))\|_{L^2}^2 \\ &&& \qquad + \frac{1}{2c}\|\min(0, c(p+1))\|_{L^2}^2,\\ &\mathcal{G} : L^2(\Omega) \times (L^2(\Omega))^n \to \mathbb{R}, && \mathcal{G}(p, q) = \frac{1}{2\alpha}\|p\|_{L^2}^2 + \frac{\beta}{2}\|q\|_{L^2}^2,\\ &\Lambda : H^1(\Omega) \to L^2(\Omega) \times (L^2(\Omega))^n, && \Lambda p = (K^*p, \nabla p),\end{aligned}$$

and calculating the corresponding duals, we find that the dual of problem ($\mathcal{P}^*_{\beta,c}$) is

$$\min_{\substack{x\in L^2(\Omega),\\ z\in H(\mathrm{div})}} \|Kx + \mathrm{div}\, z - y^\delta\|_{L^1(\Omega)} + \frac{1}{2c}\|Kx + \mathrm{div}\, z - y^\delta\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|x\|_{L^2}^2 + \frac{1}{2\beta}\|z\|_{(L^2(\Omega))^n}^2.$$

13.3.2 SEMISMOOTH NEWTON METHOD

The regularized optimality system (13.3.2) can be solved efficiently using a semismooth Newton method (cf. [Hintermüller, Ito, and Kunisch 2002; Ulbrich 2002]), which is superlinearly convergent. For this purpose, we consider (13.3.2) as a nonlinear equation $F(p) = 0$ with $F : H^1(\Omega) \to (H^1(\Omega))^*$,

$$F(p) := \frac{1}{\alpha} KK^* p - \beta \Delta p + \max(0, c(p-1)) + \min(0, c(p+1)) - y^\delta.$$

It is known (cf., e.g., [Ito and Kunisch 2008, Ex. 8.14]) that the projection operator

$$P(p) := \max(0, (p-1)) + \min(0, (p+1))$$

is semismooth from $L^q(\Omega)$ to $L^p(\Omega)$, if and only if $q > p$, and has as Newton derivative

$$D_N P(p)h = h\chi_{\{|p|>1\}} := \begin{cases} h(x) & \text{if } |p(x)| > 1, \\ 0 & \text{if } |p(x)| \leqslant 1. \end{cases}$$

Since sums of Frechét differentiable functions and semismooth functions are semismooth (with canonical Newton derivatives), we find that F is semismooth, and that its Newton derivative is

$$D_N F(p)h = \frac{1}{\alpha} KK^* h - \beta \Delta h + ch\chi_{\{|p|>1\}}.$$

A semismooth Newton step consists in solving for $p^{k+1} \in H^1(\Omega)$ the equation

$$D_N F(p^k)(p^{k+1} - p^k) = -F(p^k). \tag{13.3.7}$$

Upon defining the active and inactive sets

$$\mathcal{A}_k^+ := \{x : p^k(x) > 1\}, \quad \mathcal{A}_k^- := \{x : p^k(x) < -1\}, \quad \mathcal{A}_k := \mathcal{A}_k^+ \cup \mathcal{A}_k^-,$$

the step (13.3.7) can be written explicitly as finding $p^{k+1} \in H^1(\Omega)$ such that

$$\frac{1}{\alpha} KK^* p^{k+1} - \beta \Delta p^{k+1} + c\chi_{\mathcal{A}_k} p^{k+1} = y^\delta + c(\chi_{\mathcal{A}_k^+} - \chi_{\mathcal{A}_k^-}). \tag{13.3.8}$$

The resulting semismooth Newton method is given as Algorithm 13.1.

Theorem 13.3.5. *If $\|p_c - p^0\|_{H^1}$ is sufficiently small, the sequence of iterates $\{p^k\}$ of Algorithm 13.1 converge superlinearly in $H^1(\Omega)$ to the solution p_c of (13.3.2) as $k \to \infty$.*

Algorithm 13.1 Semismooth Newton method for (13.3.2)

1: Set $k = 0$, choose $p^0 \in H^1(\Omega)$
2: **repeat**
3: Set

$$\begin{aligned} \mathcal{A}_k^+ &= \left\{x \in \Omega : p^k(x) > 1\right\}, \\ \mathcal{A}_k^- &= \left\{x \in \Omega : p^k(x) < -1\right\}, \\ \mathcal{A}_k &= \mathcal{A}_k^+ \cup \mathcal{A}_k^- \end{aligned}$$

4: Solve for $p^{k+1} \in H^1(\Omega)$:

$$\frac{1}{\alpha} K K^* p^{k+1} - \beta \Delta p^{k+1} + c\chi_{\mathcal{A}_k} p^{k+1} = y^\delta + c(\chi_{\mathcal{A}_k^+} - \chi_{\mathcal{A}_k^-}).$$

5: Set $k = k + 1$
6: **until** $(\mathcal{A}_k^+ = \mathcal{A}_{k-1}^+)$ and $(\mathcal{A}_k^- = \mathcal{A}_{k-1}^-)$, or $k = k_{max}$

Proof. Since F is semismooth, it suffices to show that $(D_N F)^{-1}$ is uniformly bounded. Let $g \in (H^1(\Omega))^*$ be given. By assumption, the inner product $\beta \langle \nabla \cdot, \nabla \cdot \rangle_{L^2} + \frac{1}{\alpha} \langle K^* \cdot, K^* \cdot \rangle_{L^2}$ induces an equivalent norm on $H^1(\Omega)$, and thus the Lax–Milgram theorem ensures the existence of a unique $\varphi \in H^1$ such that

$$\beta \langle \nabla \varphi, \nabla v \rangle_{L^2} + \frac{1}{\alpha} \langle K^* \varphi, K^* v \rangle_{L^2} + c \langle \chi_{\mathcal{A}} \varphi, v \rangle_{L^2} = \langle g, v \rangle_{H^{1*}, H^1}$$

holds for all $v \in H^1(\Omega)$, independently of $\mathcal{A}$. Furthermore, φ satisfies

$$\|\varphi\|_{H^1}^2 \leqslant C \|g\|_{H^{1*}}^2,$$

with a constant C depending only on α, β, K and Ω. This yields the desired uniform bound. The superlinear convergence now follows from standard results (e.g., [Ito and Kunisch 2008, Theorem 8.16]). □

The termination criterion in Algorithm 13.1, step 6, can be justified as follows:

Proposition 13.3.6. *If $\mathcal{A}_{k+1}^+ = \mathcal{A}_k^+$ and $\mathcal{A}_{k+1}^- = \mathcal{A}_k^-$ holds, then p^{k+1} satisfies $F(p^{k+1}) = 0$.*

This can be verified by simple inspection, and is shown in [Ito and Kunisch 2008, Remark 7.1.1].

13.4 ADAPTIVE CHOICE OF REGULARIZATION PARAMETERS

13.4.1 CHOICE OF α BY A MODEL FUNCTION APPROACH

In this section, we propose a fixed point algorithm for adaptively determining the regularization parameter α based on a balancing principle and the model function approach. Specifically,

we first state a balancing principle based on the value function introduced in section 13.2.1, and prove existence of a regularization parameter α^* satisfying it. Then we propose a fixed point iteration which computes α^* by making use of a model function, which is a rational interpolant of the value function. The last part is devoted to the proof of convergence of the fixed point iteration.

Balancing principle

The idea of our regularization parameter choice method is to balance the data fitting term $\varphi(\alpha) = \|Kx_\alpha - y^\delta\|_{L^1}$ and the penalty term $\alpha F'(\alpha) = \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2$. The balancing principle thus consists in choosing the regularization parameter α^* as the solution of the equation

$$(\sigma - 1)\varphi(\alpha^*) = \alpha^* F'(\alpha^*), \tag{13.4.1}$$

where $\sigma > 1$ controls the relative weight between these two terms. Similar balancing ideas underlie a number of heuristic parameter choice rules, e.g. the local minimum criterion [Engl, Hanke, and Neubauer 1996], the zero-crossing method [Johnston and Gulrajani 2002] and the L-curve criterion [Hansen 1998].

Theorem 13.4.1. *For σ sufficiently close to 1 and $y^\delta \neq 0$, there exists at least one positive solution to the balancing equation (13.4.1).*

For the proof of this result, we need the next lemma, which is proved in Appendix 13.B.

Lemma 13.4.2. *The following limits hold true*

$$\lim_{\alpha\to 0^+} \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 = \lim_{\alpha\to+\infty} \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 = 0.$$

In the following, we will repeatedly make use of the residual in (13.4.1):

$$r(\alpha) = \alpha F'(\alpha) - (\sigma - 1)\varphi(\alpha). \tag{13.4.2}$$

By Theorem 13.2.2, the functions $\varphi(\alpha)$ and $F'(\alpha)$ are continuous, and thus the function $r(\alpha)$ is continuous. We can now prove existence of a positive solution to the balancing equation (13.4.1).

Proof of Theorem 13.4.1. Lemma 13.4.2 shows that the following limits hold for $\sigma > 1$:

$$\lim_{\alpha\to 0^+} r(\alpha) = -(\sigma - 1) \lim_{\alpha\to 0^+} \|Kx_\alpha - y^\delta\|_{L^1} \leqslant 0,$$
$$\lim_{\alpha\to+\infty} r(\alpha) = -(\sigma - 1)\|y^\delta\|_{L^1} < 0.$$

However, $\|Kx_\alpha - y^\delta\|_{L^1} \leqslant \|y^\delta\|_{L^1}$, and $\sup_{\alpha\in(0,+\infty)} \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 > 0$. Consequently, we have

$$r(\alpha) = \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 - (\sigma - 1)\|Kx_\alpha - y^\delta\|_{L^1} \geqslant \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 - (\sigma - 1)\|y^\delta\|_{L^1}.$$

Therefore, there exists a $\sigma_0 > 1$ such that $\sup_{\alpha\in(0,+\infty)} r(\alpha) > 0$ for all $\sigma \in (1, \sigma_0)$, and the existence of a positive solution follows.

Model function and fixed point iteration

To find a solution of the balancing equation, we write equation (13.4.1) as

$$F(\alpha^*) = \sigma(F(\alpha^*) - \alpha^* F'(\alpha^*))$$

and consider a fixed point iteration, where α^{k+1} is chosen as the solution of

$$F(\alpha^{k+1}) = \sigma(F(\alpha^k) - \alpha^k F'(\alpha^k)). \tag{13.4.3}$$

To compute this solution, we make use of the model function approach, proposed in [Ito and Kunisch 1992] for determining regularization parameters, which locally approximates the value function $F(\alpha)$ by rational polynomials. In this paper, we consider a model function of the form

$$m(\alpha) = b + \frac{s}{t + \alpha}.$$

Noting that $x_\alpha \to 0$ for $\alpha \to \infty$, we fix $b = \|y^\delta\|_{L^1}$ to match the asymptotic behavior of $F(\alpha)$ (although larger values of b work as well for our purposes). The parameters s and t are determined by the interpolation conditions

$$m(\alpha) = F(\alpha), \qquad m'(\alpha) = F'(\alpha),$$

which together with the definition of $m(\alpha)$ gives

$$b + \frac{s}{t + \alpha} = F(\alpha), \qquad -\frac{s}{(t + \alpha)^2} = F'(\alpha).$$

The parameters s and t can be derived explicitly. We recall that by Theorem 13.2.2, we have $F'(\alpha) = \frac{1}{2}\|x_\alpha\|^2_{L^2}$, and this value can be calculated without any extra computational effort. If we replace the left hand side of (13.4.3) by the value $m_k(\alpha^{k+1})$ of the model function $m_k(\alpha)$ derived from the interpolation conditions at α^k, we can thus explicitly compute a new iterate α^{k+1}. The resulting iteration is given as Algorithm 13.2. One of its salient features lies in not requiring knowledge of the noise level. Indeed, since $F(0)$ represents a lower bound for the noise level, the quantity $m(0)$ may be taken as a valid estimate of the noise level if the model function $m(\alpha)$ approximates reasonably the value function $F(\alpha)$ in the neighborhood of $\alpha = 0$.

Having found a fixed point $\bar\alpha$ of Algorithm 13.2, we find – using the fact that $\hat m = m(\bar\alpha)$ satisfies the interpolation condition at $\bar\alpha$ – that $\bar\alpha$ is a solution of the balancing equation (13.4.1). We can show that under some very general assumptions, Algorithm 13.2 converges locally to such a fixed point. To this end, we call a solution α^* to equation (13.4.1) a regular attractor if there exists an $\varepsilon > 0$ such that $r(\alpha) < 0$ for $\alpha \in (\alpha^* - \varepsilon, \alpha^*)$ and $r(\alpha) > 0$ for $\alpha \in (\alpha^*, \alpha^* + \varepsilon)$.

Theorem 13.4.3. *Assume that α_0 satisfies $r(\alpha_0) < 0$ and that it is close to a regular attractor α^*. Then the sequence $\{\alpha_k\}$ generated by Algorithm 13.2 converges to α^*.*

The proof is given in the next subsection.

Algorithm 13.2 Fixed-point algorithm for balancing equation

1: Set $k = 0$, choose $\alpha_0 > 0$, $b \geqslant \|y\|_{L^1}$ and $\sigma > 1$
2: **repeat**
3: Compute x_{α_k} by path-following semismooth Newton method ▷ Alg. 13.3
4: Compute $F(\alpha_k)$ and $F'(\alpha_k)$
5: Construct the model function $m_k(\alpha) = b + \frac{s_k}{t_k + \alpha}$ from the interpolation condition at α_k by setting

$$s_k = -\frac{(b - F(\alpha_k))^2}{F'(\alpha_k)},$$
$$t_k = \frac{b - F(\alpha_k)}{F'(\alpha_k)} - \alpha_k.$$

6: Solve for α_{k+1} in $m_k(\alpha_{k+1}) = \sigma(F(\alpha_k) - \alpha_k F'(\alpha_k))$, i.e.

$$\hat{m}_k = F(\alpha_k) - \alpha_k F'(\alpha_k),$$
$$\alpha_{k+1} = \frac{s_k}{\sigma \hat{m}_k - b} - t_k.$$

7: Set $k = k + 1$
8: **until** $k = k_{max}$

Convergence of fixed point iteration

To analyze the convergence of the fixed point algorithm, we first observe that

$$\begin{aligned}\alpha_{k+1} &= \frac{s_k}{\sigma \hat{m} - b} - t_k \\ &= \frac{(F(\alpha_k) - b)^2 - (b - F(\alpha_k) - \alpha_k F'(\alpha_k))(b - \sigma\varphi(\alpha_k))}{F'(\alpha_k)(b - \sigma\varphi(\alpha_k))},\end{aligned}$$

where $\varphi(\alpha) = \|Kx_\alpha - y^\delta\|_{L^1}$ denotes again the norm of the residual. The numerator can be simplified as follows

$$\begin{aligned}(F(\alpha_k) - b)^2 - (b - F(\alpha_k) - \alpha_k F'(\alpha_k))(b - \sigma\varphi(\alpha_k)) \\ = (\alpha_k F'(\alpha_k))^2 + (\sigma - 1)\varphi(\alpha_k)[b - F(\alpha_k) - \alpha_k F'(\alpha_k)].\end{aligned}$$

Therefore, the fixed point iteration reads as follows

$$\alpha_{k+1} = \frac{(\alpha_k F'(\alpha_k))^2 + (\sigma - 1)\varphi(\alpha_k)[b - F(\alpha_k) - \alpha_k F'(\alpha_k)]}{F'(\alpha_k)(b - \sigma\varphi(\alpha_k))} =: \alpha_k \frac{N_k}{D_k}. \tag{13.4.4}$$

Under the assumption $b > \sigma\|y^\delta\|_{L^1}$, the denominator D_k is positive, and thus the iteration is well defined. Moreover, the following identity holds

$$N_k - D_k = [(\sigma - 1)\varphi(\alpha_k) - \alpha_k F'(\alpha_k)][b - F(\alpha_k)].$$

Therefore, it follows from (13.4.4) that if $r(\alpha_k) = \alpha_k F'(\alpha_k) - (\sigma - 1)\varphi(\alpha_k) > 0$, then $N_k < D_k$ and consequently $\alpha_{k+1} < \alpha_k$, otherwise $\alpha_{k+1} > \alpha_k$ holds. Next consider

$$\begin{aligned}\alpha_{k+1} - \alpha_k &= \alpha_k \frac{N_k}{D_k} - \alpha_k = \frac{N_k - D_k}{F'(\alpha_k)(b - \sigma\varphi(\alpha_k))} \\ &= \frac{[(\sigma-1)\varphi(\alpha_k) - \alpha_k F'(\alpha_k)][b - F(\alpha_k)]}{F'(\alpha_k)(b - \sigma\varphi(\alpha_k))} \\ &= \left[(\sigma-1)\frac{\varphi(\alpha_k)}{F'(\alpha_k)} - \alpha_k\right] \frac{b - F(\alpha_k)}{b - \sigma\varphi(\alpha_k)} \\ &= [T(\alpha_k) - \alpha_k] \frac{b - F(\alpha_k)}{b - \sigma\varphi(\alpha_k)},\end{aligned}$$

where the operator $T(\alpha)$ is defined by

$$T(\alpha) = (\sigma - 1)\frac{\varphi(\alpha)}{F'(\alpha)}. \tag{13.4.5}$$

The auxiliary operator T can be regarded as the asymptotic of the operator $\alpha_k \frac{N_k}{D_k}$ in (13.4.4) as the scalar b tends to $+\infty$. For $b > \sigma\|y^\delta\|_{L^1}$, the inequality

$$\omega_k := \frac{b - F(\alpha_k)}{b - \sigma\varphi(\alpha_k)} > 0$$

holds true, and the fixed point iteration (13.4.4) can be rewritten as

$$\alpha_{k+1} = \omega_k T(\alpha_k) + (1 - \omega_k)\alpha_k.$$

Therefore, the fixed point iteration (13.4.4) can be regarded as a relaxation of the iteration $\alpha_{k+1} = T(\alpha_k)$ with a dynamically updated relaxation parameter ω_k. Note that both iterations have the same fixed point. Moreover we have

$$\omega_k < 1 \quad \text{if and only if} \quad \alpha_k F'(\alpha_k) > (\sigma - 1)\varphi(\alpha_k).$$

The next result shows the monotonicity of the operator T.

Lemma 13.4.4. *The operator T is monotone in the sense that if $0 < \alpha_0 < \alpha_1$, then*

$$T(\alpha_0) \leqslant T(\alpha_1).$$

Proof. By Theorem 13.2.2, we have

$$\varphi(\alpha_0) \leqslant \varphi(\alpha_1), \quad F'(\alpha_0) \geqslant F'(\alpha_1).$$

The result now follows directly from the definition of the operator T, see (13.4.5). □

The next lemma shows the monotonic convergence of the sequence $\{T^k(\alpha_0)\}$. This iteration itself is of independent interest because of its simplicity and practically desirable monotonic convergence.

Lemma 13.4.5. *For any initial guess α_0, the sequence $\{T^k(\alpha_0)\}$ is monotonic. Furthermore, it is monotonically decreasing (respectively increasing) if $r(\alpha_0) > 0$ (respectively $r(\alpha_0) < 0$).*

Proof. Let $\alpha_k = T^k(\alpha_0)$. Then we have

$$\begin{aligned}
\alpha_{k+1} - \alpha_k &= T^{k+1}(\alpha_0) - T^k(\alpha_0) \\
&= (\sigma - 1)\frac{\varphi(\alpha_k)}{F'(\alpha_k)} - (\sigma - 1)\frac{\varphi(\alpha_{k-1})}{F'(\alpha_{k-1})} \\
&= (\sigma - 1)\frac{\varphi(\alpha_k)F'(\alpha_{k-1}) - \varphi(\alpha_{k-1})F'(\alpha_k)}{F'(\alpha_{k-1})F'(\alpha_k)} \\
&= (\sigma - 1)\frac{\varphi(\alpha_k)[F'(\alpha_{k-1}) - F'(\alpha_k)] + F'(\alpha_k)[\varphi(\alpha_k) - \varphi(\alpha_{k-1})]}{F'(\alpha_{k-1})F'(\alpha_k)}.
\end{aligned}$$

By Theorem 13.2.2 both terms in square bracket have the same sign as $\alpha_k - \alpha_{k-1}$, which shows the desired monotonicity.

Let $r(\alpha)$ again denote the residual in the balancing equation as defined by (13.4.2). Now if $r(\alpha_0) > 0$ holds, by the definition of the function r, we have

$$\alpha_0 F'(\alpha_0) - (\sigma - 1)\varphi(\alpha_0) > 0,$$

which after rearranging the terms gives

$$\alpha_0 > (\sigma - 1)\frac{\varphi(\alpha_0)}{F'(\alpha_0)} = T(\alpha_0).$$

The second assertion follows directly from this inequality and the first statement. □

Remark 13.4.6. The iteration produces a strictly monotonic sequence before reaching a solution to (13.4.1). If two consecutive steps coincide, then a solution has been found and we can stop the iteration. Upon reaching a solution α^*, there holds $r(\alpha^*) = 0$. In our subsequent analysis, this trivial case will be excluded.

Remark 13.4.7. The sequence $\{T^k(\alpha_0)\}$ can diverge to $+\infty$. This can be remedied by further regularizing the operator T by setting

$$T_\gamma(\alpha) = (\sigma - 1)\frac{\varphi(\alpha)}{F'(\alpha) + \gamma},$$

for some small number $\gamma > 0$. This preserves the monotonicity of the iterates, and ensures the upper bound $(\sigma - 1)\|y^\delta\|_{L^1}/\gamma$, which together with the trivial lower bound 0 and the monotonicity guarantees convergence of $T^k(\alpha_0)$. Moreover, the second part of Lemma 13.4.5

classifies the positive solutions of equation (13.4.1), and the sign of the function $r(\alpha)$ provides an explicit characterization for that classification. To illustrate this point, let α^* be a solution to equation (13.4.1). The iterate $T^k(\alpha_0)$ converges to α^* for α_0 in the neighborhood of α^* if and only if

$$r(\alpha) \begin{cases} < 0, & \alpha \in (\alpha^* - \varepsilon, \alpha^*), \\ > 0, & \alpha \in (\alpha^*, \alpha^* + \varepsilon), \end{cases}$$

for sufficiently small ε, and it diverges from α^* for α_0 in the neighborhood of α^* if and only if

$$r(\alpha) \begin{cases} > 0, & \alpha \in (\alpha^* - \varepsilon, \alpha^*), \\ < 0, & \alpha \in (\alpha^*, \alpha^* + \varepsilon). \end{cases}$$

In the case that $r(\alpha)$ has the same sign on $(\alpha^*-\varepsilon, \alpha^*)$ and $(\alpha^*, \alpha^*+\varepsilon)$, the iterate can converge to α^* only for α_0 in its one-sided neighborhood. If $r(\alpha) > 0$ on $(\alpha^* - \varepsilon, \alpha^*) \cup (\alpha^*, \alpha^* + \varepsilon)$, then the iterates $T^k(\alpha_0)$ converge if $r(\alpha_0) > 0$, and vice versa for $r < 0$.

We shall also need a "sign-preserving" property of the operator T: the function $r(\alpha)$ cannot vanish on the open interval between α_0 and the limit α^* of the sequence $\{T^k(\alpha_0)\}$.

Lemma 13.4.8. *For any α_0 such that $\{T^k(\alpha_0)\}$ converges to α^*, the function $r(\alpha)$ does not vanish on the open interval $(\min(\alpha_0, \alpha^*), \max(\alpha_0, \alpha^*))$.*

Proof. Without loss of generality we assume that $\alpha_0 < T(\alpha_0)$ as the other case can be treated similarly. Assume that the assertion is false. Then there exists an $\alpha \in (\alpha_0, \alpha^*)$ such that $r(\alpha) = 0$, i.e. $T(\alpha) = \alpha$. By Lemma 13.4.5, there exists some $k \in \mathbb{N}$ such that

$$T^k(\alpha_0) \leqslant \alpha < T^{k+1}(\alpha_0).$$

However, by Lemma 13.4.4, we have

$$\alpha < T^{k+1}(\alpha_0) \leqslant T(\alpha) \leqslant T^{k+2}(\alpha_0),$$

which is a contradiction to $T(\alpha) = \alpha$. □

We note that in Lemma 13.4.8, α^* can take the value $+\infty$, i.e. the convergence can be understood in a generalized sense. Using Lemma 13.4.5 and Lemma 13.4.8, we can now state a monotone convergence result for the fixed point algorithm.

Theorem 13.4.9. *Assume that α_0 satisfies that $r(\alpha_0) > 0$. Then the sequence $\{\alpha_k\}$ generated by the fixed point iteration (13.4.4) is monotonically decreasing and converges to a solution of equation (13.4.1).*

Proof. Since $r(\alpha_0) > 0$, we have

$$0 < \omega_0 = \frac{b - F(\alpha_0)}{b - F(\alpha_0) + r(\alpha_0)} < \frac{b - F(\alpha_0)}{b - F(\alpha_0)} = 1. \tag{13.4.6}$$

Due to Lemma 13.4.5, the auxiliary sequence $\{T^k(\alpha_0)\}$ is monotonically decreasing and bounded below by zero and thus converges to some α^*. In particular $T(\alpha_0) < \alpha_0$, which together with (13.4.6) implies that

$$\alpha_1 = \omega_0 T(\alpha_0) + (1 - \omega_0)\alpha_0 \in (T(\alpha_0), \alpha_0). \tag{13.4.7}$$

Consequently we have $\alpha^* \leqslant T(\alpha_0) < \alpha_1 < \alpha_0$. Lemma 13.4.8 and (13.4.7) imply that $r(\alpha_1) > 0$. Now assume that α_k generated by the algorithm satisfies $r(\alpha_k) > 0$. Then by the definition of the operator T, we have $T(\alpha_k) < \alpha_k$. Appealing to the preceding arguments we have $\omega_k \in (0, 1)$ and

$$\alpha_{k+1} = \omega_k T(\alpha_k) + (1 - \omega_k)\alpha_k \in (T(\alpha_k), \alpha_k).$$

and thus $\alpha^* \leqslant T(\alpha_k) < \alpha_{k+1} < \alpha_k$. This shows that the the sequence $\{\alpha_k\}_{k=0}^{\infty}$ is monotonically decreasing and bounded from below by α^*, and thus converges to some $\alpha^\dagger$. Upon convergence, the limit $\alpha^\dagger$ satisfies

$$\alpha^\dagger = \alpha^\dagger \frac{(\alpha^\dagger F'(\alpha^\dagger))^2 + (\sigma - 1)\varphi(\alpha^\dagger)[b - F(\alpha^\dagger) - \alpha^\dagger F'(\alpha^\dagger)]}{\alpha^\dagger F'(\alpha^\dagger)(b - \sigma\varphi(\alpha^\dagger))}$$

due to the continuous dependence of $F(\alpha)$, $F'(\alpha)$ and $\varphi(\alpha)$ on α (Theorem 13.2.2). Simplifying the equation shows that $\alpha^\dagger$ is a solution to equation (13.4.1). Moreover, from Lemma 13.4.8 we deduce that there is no other solution to equation (13.4.1) in the open interval (α^*, α_0), and thus that $\alpha^\dagger = \alpha^*$. □

Next we address the convergence behavior of the algorithm for the case $r(\alpha_0) < 0$.

Theorem 13.4.10. *Assume that the initial guess α_0 satisfies $r(\alpha_0) < 0$. Then the sequence $\{\alpha_k\}$ generated by the fixed point iteration (13.4.4) is either monotonically increasing or there exists some $k_0 \in \mathbb{N}$ such that $r(\alpha_k) \geqslant 0$ for all $k \geqslant k_0$.*

Proof. Since $r(\alpha_0) < 0$, we have

$$\omega_0 = \frac{b - F(\alpha_0)}{b - F(\alpha_0) + r(\alpha_0)} > \frac{b - F(\alpha_0)}{b - F(\alpha_0)} = 1.$$

From Lemma 13.4.5, we deduce that $\alpha_0 < T(\alpha_0)$ and moreover the auxiliary sequence $\{T^k(\alpha_0)\}$ is monotonically increasing. Consequently, we have

$$\alpha_1 = \omega_0 T(\alpha_0) + (1 - \omega_0)\alpha_0 = T(\alpha_0) + (\omega_0 - 1)(T(\alpha_0) - \alpha_0) > T(\alpha_0).$$

In particular, $\alpha_0 < \alpha_1$. Now α_1 can either satisfy $r(\alpha_1) < 0$ or $r(\alpha_1) > 0$. For the latter case, we appeal to Theorem 13.4.9, and we have $r(\alpha_k) \geqslant 0$ for $k \geqslant 1$. Otherwise $r(\alpha_1) < 0$ and hence as above $\alpha_1 < T(\alpha_1) < \alpha_2$. The claim now follows by induction. □

We can now show the convergence of the fixed point iteration.

Proof of Theorem 13.4.3. By Theorem 13.4.10, we have that the sequence $\{\alpha_k\}$ is monotonically increasing or that there exists some $k_0 \in \mathbb{N}$ such that $r(\alpha_k) \geqslant 0$ for all $k \geqslant k_0$. Moreover, by the definition of a regular attractor, $r(\alpha) > 0$ holds for all $\alpha \in (\alpha^*, \alpha^* + \varepsilon)$ for some $\varepsilon > 0$, and by Lemma 13.4.2, we have

$$\lim_{\alpha \to +\infty} r(\alpha) = -(\sigma - 1) \lim_{\alpha \to +\infty} F(\alpha) = -(\sigma - 1)\|y^\delta\|_{L^1} < 0.$$

Therefore, by the continuity of the function $r(\alpha)$ (cf. Theorem 13.2.2), there exists at least one solution to equation (13.4.1) on the interval $(\alpha^*, +\infty)$. Denote the smallest solution of equation (13.4.1) larger than α^* by α^{**}, and set $c = \frac{2}{b - \|y^\delta\|_{L^1}}$. Since the function $r(\alpha)$ is continuous and $r(\alpha^*) = 0$, for any $\delta > 0$ we can choose ε such that

$$|r(\alpha)| < \delta, \text{ for all } \alpha \in (\alpha^* - \varepsilon, \alpha^* + \varepsilon).$$

We now choose δ such that $\delta < \min\{\frac{\alpha^{**} - \alpha^*}{c\alpha^*}, \frac{b - \|y^\delta\|_{L^1}}{2}\}$, and pick ε accordingly. Consequently, we have for $\alpha \in (\alpha^* - \varepsilon, \alpha^* + \varepsilon)$

$$\begin{aligned} \omega(\alpha) - 1 = \frac{b - F(\alpha)}{b - F(\alpha) + r(\alpha)} - 1 &= \frac{-r(\alpha)}{b - F(\alpha) + r(\alpha)} \\ &\leqslant \frac{-r(\alpha)}{b - \|y^\delta\|_{L^1} - \delta} \leqslant \frac{\delta}{b - \|y^\delta\|_{L^1} - \delta} < c\delta, \end{aligned}$$

and in particular

$$\alpha_0 < \alpha_1 = T(\alpha_0) + (\omega_0 - 1)(T(\alpha_0) - \alpha_0) < T(\alpha_0) + c\delta T(\alpha_0) < T(\alpha_0) + c\delta\alpha^*,$$

where we have used that $\omega_0 > 1$ and $\alpha^* > T(\alpha_0) > \alpha_0$.

This implies $\alpha_1 < T(\alpha_0) + c\delta\alpha^* < \alpha^{**}$. Therefore, we have either $r(\alpha_1) < 0$ with $\alpha_0 < \alpha_1 < \alpha^*$ or $r(\alpha_1) \geqslant 0$ with $\alpha^* \leqslant \alpha_1 < \alpha^{**}$. In the latter case, the convergence of α_k to α^* follows directly from Theorem 13.4.9, and thus it suffices to consider the former case. We proceed by induction and assume that α_k satisfies $r(\alpha_k) < 0$. By repeating the preceding argument, we deduce that $\alpha_k < \alpha_{k+1}$. Again either $r(\alpha_{k+1}) < 0$ and convergence to α^* follows, or $r(\alpha_{k+1}) \geqslant 0$ and we can proceed as before. If $r(\alpha_k) < 0$ for all k, then the sequence $\{\alpha_k\}$ is monotonically increasing and bounded from above by α^*. It thus converges to some $\alpha^\dagger$. Analogous to Theorem 13.4.9, we can show that $\alpha^\dagger$ is a solution to equation (13.4.1). The conclusion $\alpha^\dagger = \alpha^*$ now follows from Lemma 13.4.8.

Remark 13.4.11. Note that our derivations are valid for other Tikhonov functionals, e.g. $L^2(\Omega)$ data fitting with total variation regularization, as well. A differentiability result of the cost functional with respect to the regularization parameter as in Theorem 13.2.2 is an essential ingredient of this approach.

Algorithm 13.3 Path-following method to solve L^1-data fitting problem for fixed α

1: Set $k = 0$, choose $\beta_0 > 0$, $q < 1$, $\beta_{min} > 0$, $\tau \gg 1$
2: **repeat**
3: Compute $p_{\beta_{k+1}}$ using semismooth Newton method with $p^0 = p_{\beta_k}$ ▷ Alg. 13.1
4: Set $\beta_{k+1} = q \cdot \beta_k$
5: Set $k = k + 1$
6: **until** $\|p_{\beta_k}\|_{L^\infty} > \tau$ or $\beta_k < \beta_{min}$
7: Set $x = \frac{1}{\alpha} K^* p_{\beta_{k-1}}$

13.4.2 CHOICE OF β BY A PATH-FOLLOWING METHOD

Since the introduction of the $H^1(\Omega)$ smoothing alters the structure of the primal problem (cf. Remark 13.3.4), the value of β should be as small as possible. However, the regularized dual problem ($\mathcal{P}^*_{\beta,c}$) becomes increasingly ill-conditioned as $\beta \to 0$ due to the ill-conditioning of the discretized operator KK^* and the rank-deficiency of the diagonal matrix corresponding to the (discrete) active set. Therefore, the respective system matrix will eventually become numerically singular for some small $\beta > 0$.

One possible remedy is a path-following strategy: Starting with a large β_0, we reduce its value as long as the system is still solvable, and take the solution corresponding to the smallest such value. The question remains how to automatically select the stopping index without *a priori* knowledge or expensive computations for estimating the condition number by, e.g., singular value decomposition. To select an appropriate stopping index, we exploit the structure of the (infinite-dimensional) box constraint problem: the correct solution should be nearly feasible for c sufficiently large, and therefore the discretized solution should satisfy $\|p_\beta\|_\infty \leqslant \tau$ for some $\tau \approx 1$. Recall that for the linear system corresponding to (13.3.8), the right hand side f satisfies $\|f\|_\infty \approx c \gg 1$, which should be balanced by the diagonal matrix $c\chi_{\mathcal{A}}$ in order for the solution to be feasible. If the matrix is nearly singular, this will no longer be the case, and the solution p blows up and consequently violates the feasibility condition, i.e., $\|p_\beta\|_\infty \gg 1$. Once this happens, we take the last iterate which is still (close to) feasible and return it as the solution.

This whole procedure is summarized in Algorithm 13.3. Here, β_{min} can be set to machine precision, and β_0 may be initialized with 1.

13.5 NUMERICAL EXAMPLES

We now present some numerical results to illustrate salient features of the semismooth Newton method as well as the adaptive regularization parameter choice rules. The first two benchmark examples, `deriv2` and `heat`, are taken from [Hansen 1998], and are available in the companion

Matlab package Regularization Tools (http://www2.imm.dtu.dk/~pch/Regutools/). The third example is an inverse source problem for the two-dimensional Laplace operator.

Unless otherwise stated, the first two examples are discretized into linear systems of size $n = 100$, and the parameters are set as follows: in the fixed point Algorithm 13.2, $\sigma = 1.05$, $\alpha = 0.01$ and $b = \|y^\delta\|_{L^1}$; in the path-following Algorithm 13.3, $\beta_0 = 1$, $q = \frac{1}{5}$, $\beta_{min} = 10^{-16}$ (floating point machine precision), and $\tau = 10$; in the semismooth Newton Algorithm 13.1, $k_{max} = 10$. The penalty parameter was chosen as $c = 10^9$.

We compare the performance of the proposed method with two other algorithms: the iteratively regularized least squares method (IRLS) [Wolke and Schwetlick 1988; Rodríguez and Wohlberg 2009] and a splitting approach using an alternating direction method (ADM) [Yang, Zhang, and Yin 2009]. Since these algorithms were not originally designed for the $L^1(\Omega)$ model under consideration and numerical implementations are not freely available, we have adapted the algorithms. The details and their respective parameters are described in Appendix 13.C. All parameters in the benchmark algorithms were chosen for optimal performance with the reconstruction error being the same as that from the path-following semismooth Newton algorithm.

The noisy data y^δ is generated pointwise by setting

$$y^\delta = \begin{cases} y^\dagger + \varepsilon\xi, & \text{with probability } r, \\ y^\dagger, & \text{otherwise,} \end{cases}$$

where ξ follows a normal distribution with mean 0 and standard deviation 1, and $\varepsilon = \epsilon \max|y^\delta|$ with ϵ being the relative noise level. Unless otherwise stated we set $r = 0.3$ and $\epsilon = 1$. All computations were performed with Matlab version 2009b on a single core of a 2.4 GHz workstation with 4 GByte RAM. Matlab codes implementing the algorithm presented in this paper can be downloaded from http://www.uni-graz.at/~clason/codes/l1fitting.zip.

13.5.1 EXAMPLE 1: deriv2

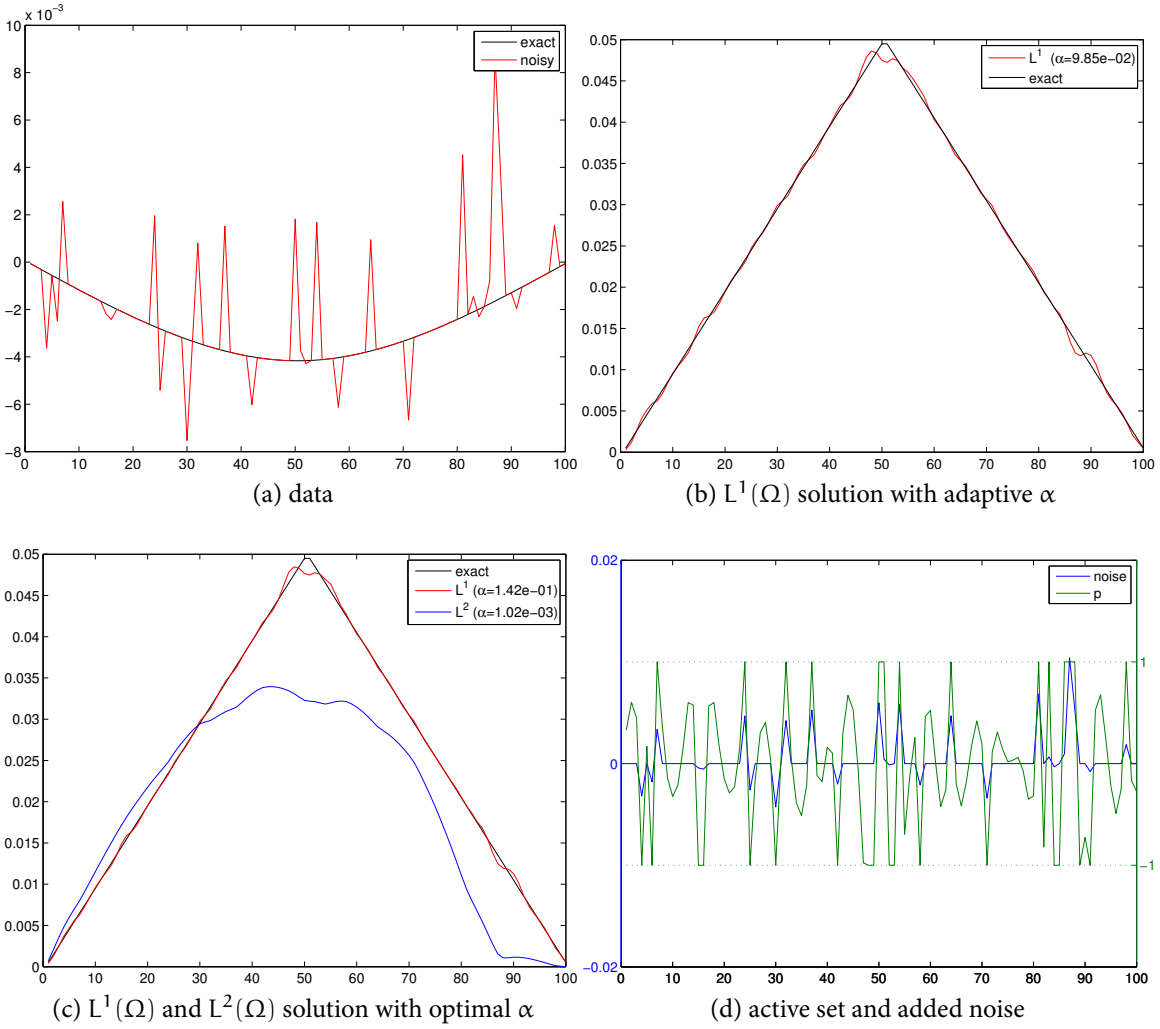
This example involves computing the second derivative of a function, i.e. the operator K is a Fredholm integral operator of the first kind:

$$(Kx)(t) = \int_0^t k(s,t)x(s)\,ds.$$

Here, the kernel $k(s,t)$ and the exact solution $x(t)$ are given by

$$k(s,t) = \begin{cases} s(t-1), & s < t, \\ t(s-1), & s \geqslant t, \end{cases} \qquad x(t) = \begin{cases} t, & t < \frac{1}{2} \\ 1-t, & \text{otherwise,} \end{cases}$$

respectively. The problem is discretized using a Galerkin method. This problem is mildly ill-posed, and the condition number of the matrix is 1.216×10^4.



(a) data

(b) $L^1(\Omega)$ solution with adaptive α

(c) $L^1(\Omega)$ and $L^2(\Omega)$ solution with optimal α

(d) active set and added noise

Figure 13.1: Results for test problem `deriv2`.

A typical realization of noisy data is displayed in Figure 13.1a. The corresponding reconstruction with the adaptively chosen parameter $\alpha_b = 9.854 \times 10^{-2}$ is shown in Figure 13.1b and agrees very well with the exact solution almost everywhere. The convergence of the fixed point algorithm is fairly fast, usually within two iterations. For comparison, we also computed the optimal value α_{opt} of the regularization parameter by sampling α at 100 points uniformly distributed over the range $[10^{-5}, 1]$ in a logarithmic scale. This yields $\alpha_{opt} = 1.418 \times 10^{-1}$, which is very close to α_b. The result, shown in Figure 13.1c, is practically identical with that by the adaptive strategy. The optimal reconstruction using L^2 data fitting, also shown in Figure 13.1c, is far inferior.

In Figure 13.1d, we show the dual solution p and the data noise. Observe that p serves as a good indicator of noise, as both location and sign of nonzero noise components are accurately detected. This numerically corroborates Corollary 13.2.8.

Table 13.1: Computing time (in seconds) and reconstruction errors for the SSN vs. IRLS and ADM methods (deriv2).

n	50	100	200	400	800	1600
t_{ssn}	0.0107	0.0223	0.0846	0.4430	3.3032	24.8826
t_{irls}	0.0131	0.0497	0.3079	2.3132	17.3116	132.2214
t_{adm}	0.0720	0.1801	0.6400	3.5907	21.5150	152.5250
e_{ssn}	1.0847e-2	3.4164e-3	7.7573e-4	5.9145e-4	2.9626e-4	3.1714e-4
e_{irls}	1.0753e-2	3.5412e-3	8.0371e-4	5.8767e-4	2.6148e-4	2.0400e-4
e_{adm}	1.0962e-2	3.8802e-3	1.2056e-3	8.3158e-4	4.7506e-4	3.9425e-4

Table 13.2: Iterates in the path-following method for β (deriv2).

β	iterations	e	$F(x)$	$\|\nabla p\|_{L^2}$
1.000e+0	2	2.860e-2	2.794e-3	2.589e-3
4.000e-2	2	2.362e-2	2.438e-3	5.155e-2
1.600e-3	2	7.729e-3	1.302e-3	2.148e-1
6.400e-5	2	7.926e-3	1.096e-3	5.760e-1
2.560e-6	7	2.096e-2	1.074e-3	4.240e+0
1.024e-7	6	9.300e-3	8.986e-4	7.423e+0
4.096e-9	4	3.681e-3	8.646e-4	6.999e+0
1.638e-10	2	1.448e-3	8.625e-4	5.994e+0
6.554e-12	3	5.334e-4	8.622e-4	6.389e+0
1.311e-12	10	3.792e-4	8.622e-4	6.884e+0

To illustrate the performance of the semismooth Newton (SSN) method, we compare the computing time and reconstruction error e, defined as $e = \|x_\alpha - x^\dagger\|_{L^2}$, for different problem sizes and $r = 0.7$ (averaged over ten runs with different noise realizations) with that of the IRLS and ADM methods in Table 13.1. For all problem sizes under consideration, the SSN method is significantly faster than the IRLS and ADM methods. The results by all three methods are very close to each other, with the ADM method showing less accuracy.

The convergence behavior of the path-following method is shown in Table 13.2. For moderate values of β, the SSN method exhibits superlinear convergence as indicated by the convergence after two iterations. This property is lost when β becomes too small, but the method still converges after very few iterations due to our path-following strategy. Interestingly, while the functional value F keeps on decreasing as β decreases, the error e experiences some transition at $\beta = 2.560 \times 10^{-6}$. This might be attributed to the change from the dominance of the H^1 term (β) to that of the $L^2(\Omega)$ term (α) in the regularized dual formulation ($\mathcal{P}^*_{\beta,c}$).

Finally, we compare the parameters chosen by the balancing principle (BP) with those obtained

Table 13.3: Comparison of balancing principle with discrepancy principle (`deriv2`).

(r, ϵ)	δ	δ_b	α_b	α_d	α_{opt}	e_b	e_d	e_{opt}
(0.3,0.1)	7.291e-5	7.284e-5	8.656e-3	1.612e-1	2.056e-1	2.849e-3	2.240e-4	1.949e-4
(0.3,0.3)	2.187e-4	2.186e-4	2.621e-2	1.125e-1	2.056e-1	9.368e-4	2.392e-4	1.949e-4
(0.3,0.5)	3.645e-4	3.644e-4	4.372e-2	1.266e-1	2.056e-1	5.971e-4	2.358e-4	1.949e-4
(0.3,0.7)	5.104e-4	5.101e-4	6.123e-2	1.458e-1	2.056e-1	3.695e-4	2.304e-4	1.949e-4
(0.3,0.9)	6.562e-4	6.557e-4	7.874e-2	1.620e-1	2.056e-1	3.851e-4	2.035e-4	1.949e-4
(0.1,0.3)	3.901e-5	3.901e-5	4.677e-3	9.543e-2	1.707e-1	7.963e-4	4.274e-5	3.904e-5
(0.3,0.3)	2.187e-4	2.186e-4	2.621e-2	1.125e-1	2.056e-1	9.368e-4	2.392e-4	1.949e-4
(0.5,0.3)	4.128e-4	4.125e-4	4.930e-2	3.048e-1	3.593e-1	1.909e-3	3.811e-4	3.711e-4
(0.7,0.3)	5.822e-4	5.798e-4	6.730e-2	4.065e-1	3.944e-1	5.438e-3	1.791e-3	1.794e-3
(0.9,0.3)	8.238e-4	7.991e-4	6.602e-2	5.147e-1	4.750e-1	1.797e-2	2.927e-3	2.859e-3

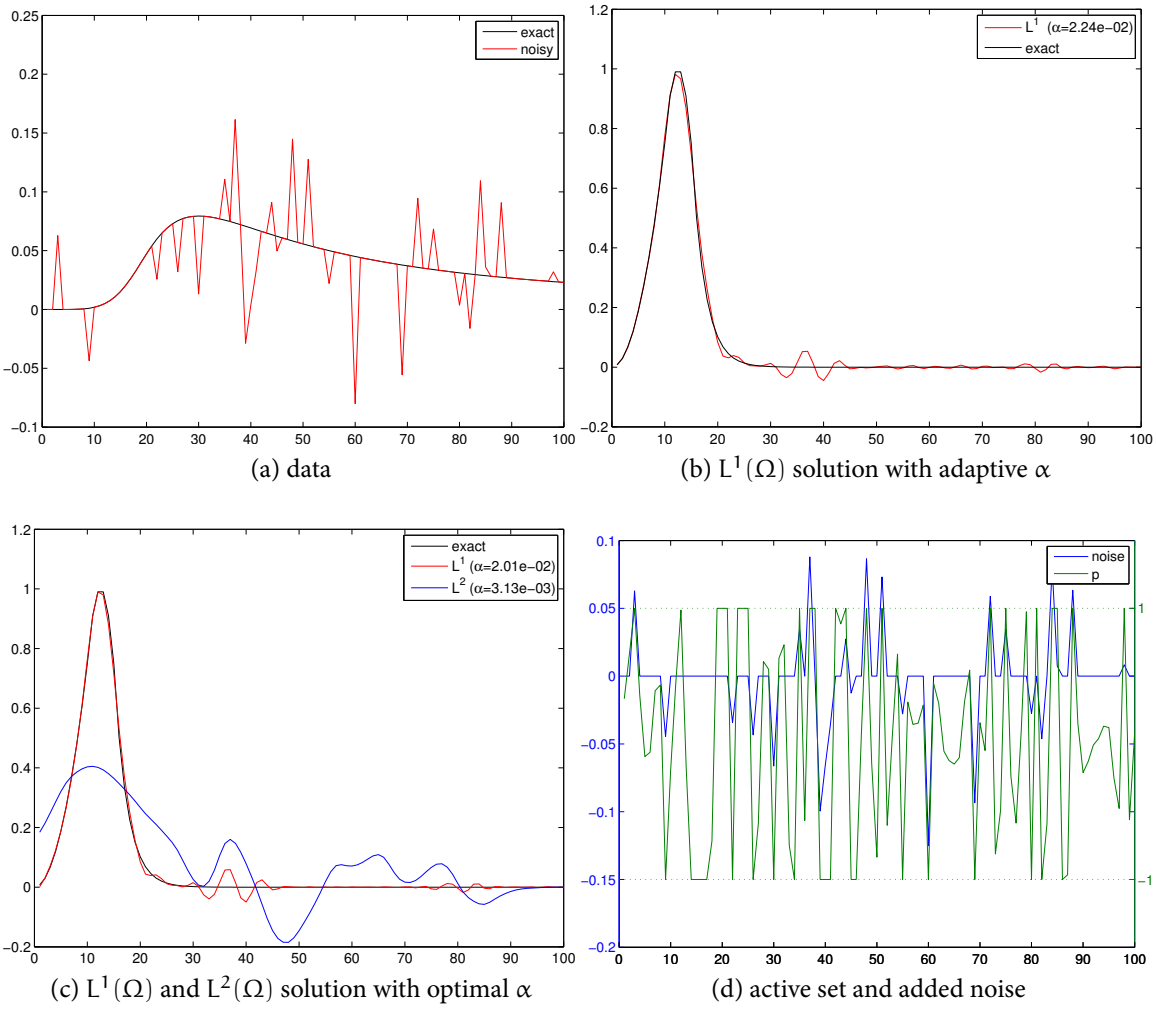
from the discrepancy principle (DP) and the optimal choice. The chosen parameters α and corresponding errors e for different noise parameters (r, ϵ) are shown in Table 13.3, where the subscript b, d and opt refer to the BP, DP and the optimal choice, respectively. For the DP, we utilize the exact noise level δ. We observe that the results by the BP and DP are largely comparable in terms of the error e despite the discrepancies in the regularization parameter. Also, the regularization parameter determined by the BP increases at the same rate of the noise level δ, whereas the one determined by the DP seems largely independent of δ, especially for fixed r. This causes slight under-regularization in the BP for low noise levels. Nonetheless, the noise level δ is estimated very accurately by δ_b. Interestingly, we observe that the two factors of the noise, i.e. r and ϵ, have drastically different effects on the inverse solution: the results seem relatively independent of the ϵ for fixed r, whereas for fixed ϵ, the error e deteriorates rapidly as noise percentage r increases. In particular, α_{opt} seems relatively independent of ϵ for fixed r, and increases at the rate of r for fixed ϵ. Finally, with the knowledge of the exact noise level δ, the DP achieves optimal convergence rate in that its error is roughly the same as that with the optimal parameter.

13.5.2 EXAMPLE 2: `heat`

This example is an inverse heat conduction problem, posed as a Volterra integral equation of the first kind. The kernel $k(s,t)$ and the exact solution $x(t)$ are given by

$$k(s,t) = \frac{(s-t)^{-\frac{3}{2}}}{2\sqrt{\pi}} e^{-\frac{1}{4(s-t)}}, \quad x(t) = \begin{cases} 75t^2, & u \leqslant 2, \\ \frac{3}{4} + (u-2)(3-u), & 2 < u \leqslant 3, \\ \frac{3}{4} e^{-2(u-3)}, & 3 < u \leqslant 10, \\ 0, & \text{otherwise}, \end{cases}$$

with $u = 20t$ and the integration interval $[0, 1]$. The integral equation is discretized using collocation and the mid-point rule. This problem is exponentially ill-posed, and the condition number is 8.217×10^{36}.



(a) data

(b) $L^1(\Omega)$ solution with adaptive α

(c) $L^1(\Omega)$ and $L^2(\Omega)$ solution with optimal α

(d) active set and added noise

Figure 13.2: Results for test problem `heat`.

The results are given in Figure 13.2. Again, the reconstruction with automatically chosen parameter $\alpha_b = 2.239 \times 10^{-2}$ is very close to the exact solution and to the optimal reconstruction with $\alpha_{opt} = 2.009 \times 10^{-2}$, while the $L^2(\Omega)$-reconstruction is vastly inferior. The performance and convergence of the path-following SSN method is similar to Example 1 (cf. Tables 13.4, 13.5). Also, the adaptive strategy yields comparable results with that for the discrepancy principle and optimal choice, see Table 13.6.

The convergence of the fixed point algorithm is now even faster: The convergence is achieved in one iteration. This may be attributed to the fact that the spectrum of the matrix spans a much broader range because of its exponential ill-posedness, and thus the residual is less sensitive to the variation of the regularization parameter. This consequently accelerates the convergence of the fixed point algorithm. Again the noise level is estimated very accurately, while the chosen regularization parameter is now closer to the optimal one compared to

Table 13.4: Computing time (in seconds) and reconstruction errors for the SSN vs. IRLS and ADM methods (heat).

n	50	100	200	400	800	1600
t_{ssn}	0.0089	0.0197	0.0808	0.4365	4.1769	27.5007
t_{irls}	0.0131	0.0515	0.3254	2.3720	18.6916	132.5247
t_{adm}	0.0729	0.1826	0.7952	4.3687	25.0411	154.9379
e_{ssn}	1.4743	2.0263e-1	1.3621e-1	1.4336e-1	1.3319e-1	1.3615e-1
e_{irls}	1.4610	1.7857e-1	1.3587e-1	1.4334e-1	1.3322e-1	1.3618e-1
e_{adm}	1.4657	1.7913e-1	1.3595e-1	1.4333e-1	1.3335e-1	1.3638e-1

Table 13.5: Iterates in the path-following method for β (heat).

β	iterations	e	$F(x)$	$\|\nabla p\|_{L^2}$
1.000e+0	2	2.248e-1	3.274e-2	2.951e-2
4.000e-2	2	1.855e-1	1.963e-2	1.090e-1
1.600e-3	2	1.610e-1	1.713e-2	2.191e-1
6.400e-5	2	1.556e-1	1.737e-2	1.408e+0
2.560e-6	6	2.250e-1	1.644e-2	5.852e+0
1.024e-7	4	7.042e-2	1.435e-2	6.436e+0
4.096e-9	3	2.043e-2	1.415e-2	6.902e+0
1.638e-10	10	1.563e-2	1.414e-2	7.361e+0
3.277e-11	10	1.546e-2	1.414e-2	8.960e+0

Example 1 and sometimes even outperforms the discrepancy principle with exact noise level (cf. Table 13.6).

13.5.3 EXAMPLE 3: INVERSE SOURCE PROBLEM IN 2D

As a two-dimensional test problem, we consider the inverse source problem for the Laplacian on the unit square $[0,1]^2$ with a homogeneous Dirichlet boundary condition, i.e. $K = (-\Delta)^{-1}$. The exact solution $x(s,t)$ is given by (cf. Figure 13.3a)

$$x(s,t) = \begin{cases} \sin 2\pi(s-\frac{1}{4}) \sin 2\pi(t-\frac{1}{4}), & |s-\frac{1}{2}| \leqslant \frac{1}{4}, |t-\frac{1}{2}| \leqslant \frac{1}{4}, \\ 0 & \text{otherwise.} \end{cases}$$

The problem is discretized on a 64×64 mesh using the standard five-point stencil, resulting in a linear system of size $n = 4096$. The problem is mildly ill-posed, and the estimated condition number is 2.689×10^3. The CPU time for one reconstruction using Algorithm 13.3 was 6.5 seconds.

Table 13.6: Comparison of balancing principle with discrepancy principle (heat).

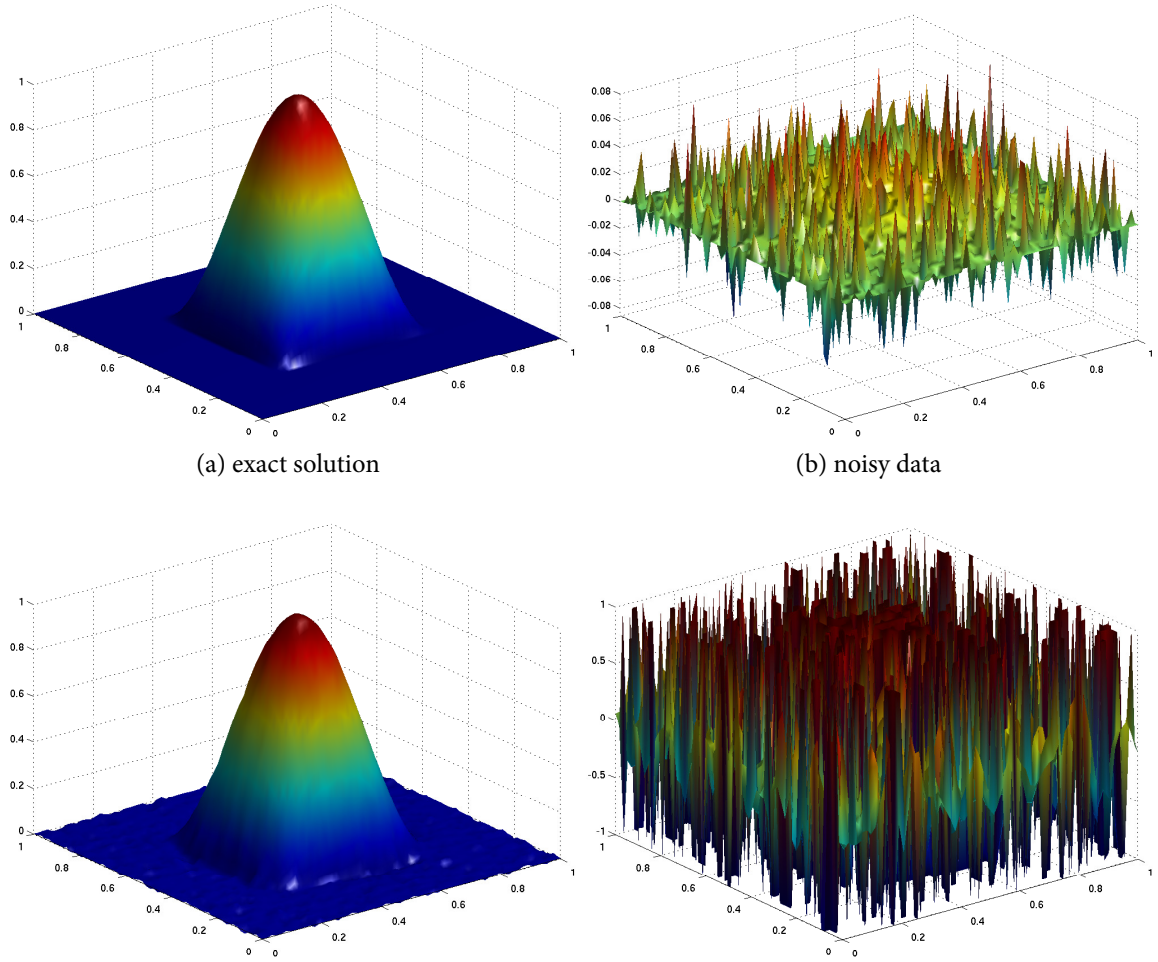
(r, ϵ)	δ	δ_b	α_b	α_d	α_{opt}	e_b	e_d	e_{opt}
(0.3,0.1)	1.390e-3	1.335e-3	1.402e-3	1.910e-2	1.830e-2	1.860e-1	2.021e-2	2.026e-2
(0.3,0.3)	4.170e-3	4.155e-3	6.638e-3	1.906e-2	1.830e-2	4.515e-2	2.021e-2	2.026e-2
(0.3,0.5)	6.950e-3	6.939e-3	1.135e-2	1.908e-2	1.830e-2	2.706e-2	2.022e-2	2.026e-2
(0.3,0.7)	9.731e-3	9.719e-3	1.604e-2	2.045e-2	1.830e-2	2.103e-2	2.032e-2	2.026e-2
(0.3,0.9)	1.251e-2	1.249e-2	2.083e-2	1.909e-2	1.830e-2	2.037e-2	2.022e-2	2.026e-2
(0.1,0.3)	7.438e-4	7.439e-4	1.227e-3	1.374e-2	7.742e-4	1.727e-3	2.736e-3	5.980e-4
(0.3,0.3)	4.170e-3	4.155e-3	6.638e-3	1.906e-2	1.830e-2	4.515e-2	2.021e-2	2.026e-2
(0.5,0.3)	7.871e-3	7.799e-3	1.225e-2	4.718e-2	3.199e-2	5.635e-2	4.026e-2	3.772e-2
(0.7,0.3)	1.110e-2	1.074e-2	2.254e-2	3.995e-2	7.390e-2	1.118e-1	1.130e-1	1.034e-1
(0.9,0.3)	1.570e-2	1.470e-2	2.247e-2	1.553e-1	5.094e-2	1.662e-1	1.487e-1	1.388e-1

The noisy data for this problem is given in Figure 13.3b. The corresponding numerical solution of the inverse source problem, shown in Figure 13.3c, is a good approximation of the exact one. Note in particular that the magnitude of the peak is correctly recovered. The L^2-norm of the reconstruction error is $e = 7.526 \times 10^{-3}$. The fixed point algorithm converged in three iterations to the value $\alpha_b = 8.797 \times 10^{-3}$. The estimated noise level was $\delta_b = 5.475 \times 10^{-3}$, which is very close to the exact one $\delta = 5.490 \times 10^{-3}$. For completeness, we show also the dual solution in Figure 13.3d.

13.6 CONCLUSION

We have presented a semismooth Newton method for the numerical solution of inverse problems with $L^1(\Omega)$ data fitting together with an adaptive method for the choice of regularization parameters. The main advantage of the adaptive strategy is that no knowledge of the noise level is necessary, and it can, in fact, provide an excellent estimate of the noise level. This is important for some practical applications. The convergence of the fixed point iteration was analyzed. In practice it is usually achieved within two or three iterations. The value for the regularization parameter obtained by the proposed technique based on the balancing principle derived from the model function approach was always fairly close to the optimal one.

Similarly, the semismooth Newton method allows an efficient numerical solution of $L^1(\Omega)$ data fitting problems. In our examples, the proposed method was significantly faster than the iteratively reweighted least squares method and the alternating direction method. In practice, implementations of the latter two methods would include early termination criteria based, e.g., on the norm of the difference of consecutive iterates, which would accelerate the methods, although at the risk of loss of accuracy. Similar strategies could of course also be applied in our method. On the other hand, we consider the fact that the semismooth Newton method



(a) exact solution

(b) noisy data

(c) $L^1(\Omega)$ reconstruction

(d) dual solution p

Figure 13.3: Results for two-dimensional inverse source problem

performs very well without the introduction of heuristic tolerance-based termination criteria to be one of its main advantages.

The path-following strategy proved to be an efficient and simple strategy to achieve the conflicting goals of minimizing the effect of the additional smoothing term on the primal problem and maintaining the numerical stability of the dual problem.

The proposed approach can also be extended to more general functionals (e.g., involving total variation terms), which will be the focus of a subsequent work.

ACKNOWLEDGMENTS

This work was carried out during the visit of the second named author at the Institute of Mathematics and Scientific Computing, Karl-Franzens-Universität Graz. He would like to thank Professor Karl Kunisch and the institute for the hospitality.

13.A CONVERGENCE OF SMOOTHING FOR PENALIZED BOX CONSTRAINTS

Here we show the convergence of the solutions of $(\mathcal{P}^*_{\beta,c})$ as β tends to zero to a solution of $(\mathcal{P}^*_c)$

$$\min_{p\in L^2(\Omega)} \frac{1}{2\alpha}\|K^*p\|^2_{L^2} - \langle p, y^\delta\rangle_{L^2} + \frac{1}{2c}\left(\|\max(0, c(p-1))\|^2_{L^2} + \|\min(0, c(p+1))\|^2_{L^2}\right).$$

For this problem, the solution might be nonunique if the operator K is not injective. Again, the functional in $(\mathcal{P}^*_c)$ is convex, as is the set of all its minimizers, and thus if problem $(\mathcal{P}^*_c)$ admits a solution in $H^1(\Omega)$, this set has an element with minimal $H^1(\Omega)$-semi-norm, denoted by $p^\dagger$.

Theorem 13.A.1. *Let $\{\beta_n\}$ be a vanishing sequence. Then the sequence of minimizers $\{p_{\beta_n,c}\}$ of $(\mathcal{P}^*_{\beta,c})$ has a subsequence converging weakly to a minimizer of problem $(\mathcal{P}^*_c)$. If the operator K is injective or there exists a unique $p^\dagger$ as defined above, then the whole sequence converges weakly to $p^\dagger$.*

Proof. Let $\mathcal{A}^+ = \{x \in \Omega : p(x) > 1\}$ and $\mathcal{A}^- = \{x \in \Omega : p(x) < 1\}$. We denote the positive and negative parts of a function p by $(p)^+$ and $(p)^-$, respectively. The functional in $(\mathcal{P}^*_{\beta,c})$ can then be written as

$$\frac{1}{2\alpha}\|K^*p\|^2_{L^2} - \langle p, y^\delta\rangle_{L^2} + \frac{\beta}{2}\|\nabla p\|^2_{L^2} + \frac{1}{2c}\left(\|c(p-1)^+\|^2_{L^2} + \|c(p+1)^-\|^2_{L^2}\right).$$

Now observe that

$$\begin{aligned}\|(p-1)^+\|^2_{L^2} &= \int_\Omega ((p-1)^+)^2\,dx = \int_{\mathcal{A}^+} p^2 - 2p + 1\,dx\\ &= \|p\|^2_{L^2(\mathcal{A}^+)} + |\mathcal{A}^+| - 2\int_{\mathcal{A}^+} p\,dx,\end{aligned}$$

Note also that

$$\int_{\mathcal{A}^+} p\,dx \leqslant \|p\|_{L^2(\mathcal{A}^+)}|\mathcal{A}^+|^{1/2} \leqslant \frac{1}{4}\|p\|^2_{L^2(\mathcal{A}^+)} + |\mathcal{A}^+|.$$

Combining these two inequalities gives

$$\|(p-1)^+\|_{L^2}^2 \geqslant \frac{1}{2}\|p\|_{L^2(\mathcal{A}^+)}^2 - |\mathcal{A}^+|.$$

Similarly, we have that

$$\|(p+1)^-\|_{L^2}^2 \geqslant \frac{1}{2}\|p\|_{L^2(\mathcal{A}^-)}^2 - |\mathcal{A}^-|.$$

Without loss of generality, we may assume that $c \geqslant 1$. Then by the minimizing property of $p_n \equiv p_{\beta_n,c}$, we have that

$$\begin{aligned}\frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2 + \frac{1}{2}(\|(p_n-1)^+\|_{L^2}^2 + \|(p_n+1)^-\|_{L^2}^2)\\ \leqslant \frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2 + P_c(p_n) \leqslant 0,\end{aligned}$$

where for the sake of brevity we have set

$$P_c(p) := \frac{1}{2c}\left(\|c(p-1)^+\|_{L^2}^2 + \|c(p+1)^-\|_{L^2}^2\right).$$

This together with the inequalities above implies that

$$\begin{aligned}\frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2 + \frac{1}{4}\|p_n\|_{L^2}^2 &\leqslant |\Omega| + \langle p_n, y^\delta\rangle_{L^2}\\ &\leqslant |\Omega| + \|p_n\|_{L^2}\|y^\delta\|_{L^2}\\ &\leqslant |\Omega| + \frac{1}{8}\|p_n\|_{L^2}^2 + 2\|y^\delta\|_{L^2}^2.\end{aligned}$$

This in particular implies that the sequence $\{p_n\}$ is uniformly bounded in $L^2(\Omega)$ independently of n. Therefore, there exists a subsequence, also denoted by $\{p_n\}$, converging weakly in $L^2(\Omega)$ to some $p^* \in L^2(\Omega)$. By the weak lower semi-continuity of norms, we have

$$\|K^*p^*\|_{L^2}^2 \leqslant \liminf_{n\to\infty}\|K^*p_n\|_{L^2}^2, \quad \langle p^*, y^\delta\rangle_{L^2} = \lim_{n\to\infty}\langle p_n, y^\delta\rangle_{L^2},$$

and moreover by the convexity of the operators max and min, we have weak lower semi-continuity of the corresponding terms

$$\begin{aligned}\|(p^*-1)^+\|_{L^2}^2 &\leqslant \liminf_{n\to\infty}\|(p_n-1)^+\|_{L^2}^2,\\ \|(p^*+1)^-\|_{L^2}^2 &\leqslant \liminf_{n\to\infty}\|(p_n+1)^-\|_{L^2}^2.\end{aligned}$$

By the minimizing property of p_n, we thus have for any fixed $p \in H^1(\Omega)$ that

$$\begin{aligned}
&\frac{1}{2\alpha}\|K^*p^*\|_{L^2}^2 - \langle p^*, y^\delta\rangle_{L^2} + P_c(p^*) \\
&\leqslant \liminf_{n\to\infty}\left(\frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2 + P_c(p_n)\right) \\
&\leqslant \liminf_{n\to\infty}\left(\frac{1}{2\alpha}\|K^*p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p\|_{L^2}^2 + P_c(p)\right) \\
&= \frac{1}{2\alpha}\|K^*p\|_{L^2}^2 - \langle p, y^\delta\rangle_{L^2} + \frac{1}{2c}\left(\|c(p-1)^+\|_{L^2}^2 + \|c(p+1)^-\|_{L^2}^2\right).
\end{aligned}$$

Therefore, p^* is a minimizer of problem ($\mathcal{P}_c^*$) over $H^1(\Omega)$. Now the density of $H^1(\Omega)$ in $L^2(\Omega)$ shows that p^* is also a minimizer of problem ($\mathcal{P}_c^*$) over $L^2(\Omega)$.

Finally, by the minimizing property of $p^\dagger$ and p_n, we have

$$\begin{aligned}
&\frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p_n\|_{L^2}^2 + P_c(p_n) \\
&\qquad\qquad \leqslant \frac{1}{2\alpha}\|K^*p^\dagger\|_{L^2}^2 - \langle p^\dagger, y^\delta\rangle_{L^2} + \frac{\beta_n}{2}\|\nabla p^\dagger\|_{L^2}^2 + P_c(p^\dagger),
\end{aligned}$$

and

$$\frac{1}{2\alpha}\|K^*p^\dagger\|_{L^2}^2 - \langle p^\dagger, y^\delta\rangle_{L^2} + P_c(p^\dagger) \leqslant \frac{1}{2\alpha}\|K^*p_n\|_{L^2}^2 - \langle p_n, y^\delta\rangle_{L^2} + P_c(p_n).$$

Adding these two inequalities together, we deduce that

$$\|\nabla p_n\|_{L^2}^2 \leqslant \|\nabla p^\dagger\|_{L^2}^2,$$

which together with the weak lower-semicontinuity of the semi-norm yields

$$\|\nabla p^*\|_{L^2}^2 \leqslant \|\nabla p^\dagger\|_{L^2}^2,$$

i.e. that p^* is a minimizer with minimal $H^1(\Omega)$-semi-norm. If K is injective or $p^\dagger$ is unique, then it follows that $p^* = p^\dagger$. Consequently, each subsequence has a subsequence converging weakly to $p^\dagger$, and the whole sequence converges weakly. □

13.B PROOF OF LEMMA 13.4.2

By Theorem 13.2.2, the function $\|Kx_\alpha - y^\delta\|_{L^1}$ is continuous and increasing as a function of α. Therefore the following limits exist

$$\lim_{\alpha\to 0^+}\|Kx_\alpha - y^\delta\|_{L^1}, \quad \lim_{\alpha\to+\infty}\|Kx_\alpha - y^\delta\|_{L^1}.$$

By the minimizing property of x_α, we have

$$\|Kx_\alpha - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 \leqslant \|Kx - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x\|_{L^2}^2, \text{ for all } x \in L^2(\Omega).$$

Taking $x = 0$, this gives

$$\|Kx_\alpha - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 \leqslant \|y^\delta\|_{L^1}.$$

Letting α tend to ∞, we deduce that

$$0 \leqslant \lim_{\alpha\to+\infty} \|x_\alpha\|_{L^2}^2 \leqslant \lim_{\alpha\to+\infty} \frac{2}{\alpha}\|y^\delta\|_{L^1} = 0,$$

i.e. $\lim_{\alpha\to+\infty} x_\alpha = 0$. From this we derive that

$$\lim_{\alpha\to+\infty} \|Kx_\alpha - y^\delta\|_{L^1} = \|y^\delta\|_{L^1}.$$

Appealing again to the minimizing property, we obtain

$$\lim_{\alpha\to+\infty} \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 = 0.$$

Let $\theta = \inf_{x\in L^2(\Omega)} \|Kx - y^\delta\|_{L^1}$. By monotonicity and continuity of $\left\|Kx_\alpha - y^\delta\right\|_{L^1}$, we have that

$$\theta = \lim_{\alpha\to 0^+} \|Kx_\alpha - y^\delta\|_{L^1}. \tag{13.B.1}$$

By the definition of the infimum, there exists an x^ε such that

$$\theta \leqslant \|Kx^\varepsilon - y^\delta\|_{L^1} \leqslant \theta + \varepsilon.$$

Now the minimizing property of x_α yields

$$\theta \leqslant \|Kx_\alpha - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 \leqslant \|Kx^\varepsilon - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x^\varepsilon\|_{L^2}^2 \leqslant \theta + \varepsilon + \frac{\alpha}{2}\|x^\varepsilon\|_{L^2}^2.$$

Letting α tend to zero, we conclude

$$\theta \leqslant \lim_{\alpha\to 0^+} \left\{\|Kx_\alpha - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2\right\} \leqslant \theta + \varepsilon$$

since ε is arbitrary, we have

$$\theta \leqslant \lim_{\alpha\to 0^+} \left\{\|Kx_\alpha - y^\delta\|_{L^1} + \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2\right\} \leqslant \theta,$$

which together with equation (13.B.1) implies that $\lim_{\alpha\to 0^+} \frac{\alpha}{2}\|x_\alpha\|_{L^2}^2 = 0$.

13.C BENCHMARK ALGORITHMS

For the reader's convenience, we briefly sketch the implemented version and the values of all occurring parameters of the benchmark methods. For a detailed description, we refer to references [Wolke and Schwetlick 1988; Rodríguez and Wohlberg 2009; Yang, Zhang, and Yin 2009]. Since the proposed SSN method solves the discrete optimality system exactly, we avoided introducing early termination criteria in the benchmark algorithms to allow for a fair comparison. Instead, we have fixed the number of iterations such that their performance was optimal while still giving the same reconstruction errors as the SSN method. In practice, one would add termination criteria based, e.g., on the norm of the difference of iterates, which would accelerate the benchmark methods as well as the SSN method.

13.C.1 ITERATIVELY REWEIGHTED LEAST SQUARES

This basic idea of the approach is to approximate the (discrete) $L^1(\Omega)$ norm from above by a quadratic function $Q(x,z)$ such that $Q(x,z) \geqslant \|z\|_{L^1}$ and $Q(x,x) = \|x\|_{L^1}$. One such choice is given by

$$Q(x,z) = \|x\|_{L^1} + \tfrac{1}{2}\left(z^T W(x) z - x^T W(x) x\right),$$

where $W(x)$ is a diagonal matrix with entries $|x_i|^{-1}$. The $L^1(\Omega)$ norm is then minimized by iteratively solving for given x^k the smooth minimization problem with $Q(x^k, x)$ in place of $\|x\|_{L^1}$. To avoid division by zero in the definition of $W(x)$, an additional regularization parameter ε is introduced, which was set to $\varepsilon = 10^{-6}$, again for best performance with the reconstruction error being the same as that from the proposed algorithm. The maximum number of iterations k_{max} was set to 40.

Algorithm 13.4 IRLS

1: Choose tol, k_{max}, set $x_0 = 0$
2: **for** $k = 0, \ldots, k_{max}$ **do**
3: Set $W = \operatorname{diag}\left(\max(|Kx_k - y^\delta|, \varepsilon)^{-1}\right)$
4: Set x_{k+1} as solution of $(K^*WK + \alpha I)x = (K^*W)y^\delta$
5: **end for**

13.C.2 ALTERNATING DIRECTION MINIMIZATION

This approach consists in introducing the splitting $z = Kx - y^\delta$ and minimizing for $\beta > 0$ the functional

$$\|z\|_{L^1} + \frac{\beta}{2}\left\|Kx - y^\delta - z\right\|_{L^2}^2 + \frac{\alpha}{2}\|x\|_{L^2}^2.$$

The minimization is carried out by alternatingly minimizing with respect to z and with respect to x, which are both explicitly solvable. If the penalty parameter β goes to infinity, the solution converges to the solution x_α of Problem ($\mathcal{P}$). We therefore employ a continuation strategy for β, as was adopted in [Yang, Zhang, and Yin 2009]. The full procedure is given in Algorithm 13.5. The parameters were chosen as $\beta_0 = 1$, $\beta_{max} = 2^{16}$, $q = 2$ and $k_{max} = 30$.

Algorithm 13.5 ADM

choose k_{max}, β_0, β_{max}, q
Set $x_0 = z_0 = 0$, $\beta = \beta_0$
repeat
  for $k = 0, \dots, k_{max}$ **do**
    Set $z_{k+1} = \operatorname{sign}(Kx_k - y^\delta) \cdot \max(|Kx_k - y^\delta| - \beta^{-1}, 0)$
    Set x_{k+1} as solution of $(K^*K + \frac{\alpha}{\beta} I)x = K^*(y^\delta + z_{k+1})$
  end for
  Set $x_0 = x_k$, $\beta = q\beta$
until $\beta > \beta_{max}$ **return** x_k

A SEMISMOOTH NEWTON METHOD FOR NONLINEAR PARAMETER IDENTIFICATION PROBLEMS WITH IMPULSIVE NOISE

14

ABSTRACT

This work is concerned with nonlinear parameter identification in partial differential equations subject to impulsive noise. To cope with the non-Gaussian nature of the noise, we consider a model with L^1 fitting. However, the nonsmoothness of the problem makes its efficient numerical solution challenging. By approximating this problem using a family of smoothed functionals, a semismooth Newton method becomes applicable. In particular, its super-linear convergence is proved under a second-order condition. The convergence of the solution to the approximating problem as the smoothing parameter goes to zero is shown. A strategy for adaptively selecting the regularization parameter based on a balancing principle is suggested. The efficiency of the method is illustrated on several benchmark inverse problems of recovering coefficients in elliptic differential equations, for which one- and two-dimensional numerical examples are presented.

14.1 INTRODUCTION

We are interested in the nonlinear inverse problem

$$S(u) = y^\delta,$$

where $S : \mathcal{X} \to \mathcal{Y}$ is the parameter-to-observation mapping and y^δ represents experimental measurements corrupted by impulsive noise. Throughout we assume that the space $\mathcal{Y}$ compactly embeds into L^q for some $q > 2$, y^δ is bounded almost everywhere, and $\mathcal{X}$ is a Hilbert space. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are defined on the bounded domains $\omega \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$, respectively. Such models arise naturally in distributed parameter identification for

differential equations, where typically $\mathcal{Y}$ is $H^1(D)$ or $H^{\frac{1}{2}}(D)$ and $\mathcal{X}$ is $L^2(\omega)$ or $H^1(\omega)$ [Banks and Kunisch 1989].

The noise model for the measured data y^δ plays a critical role in formulating and solving the problem. In practice, an additive Gaussian noise model is customarily adopted, which leads to the standard L^2 fitting. However, non-Gaussian (e.g., Laplace or Cauchy) noise – which admits the presence of significant outliers – may also occur. An extreme case is impulsive noise such as salt-and-pepper or random-valued noise, which frequently occurs in digital image acquisition and processing due to, e.g., malfunctioning pixels in camera sensors, faulty memory locations in hardware, or transmission in noisy channels [Bovik 2005]. Before giving a formal definition below (§ 14.1.2), let us briefly describe its salient feature and motivate the use of L^1 fitting. The impulsive noise models considered here are characterized by the fact that only a (possibly large) number of points are subject to large errors, while the remaining data points stay intact. (In effect, such noise is "outliers only".) Such noise thus has a *sparsity* property. Since it is well known that L^1 norms as penalties promote sparse solutions [Tibshirani 1996; Chen, Donoho, and Saunders 1998], the expectation of a sparse *residual* quite naturally leads to L^1 fitting. In contrast, L^2 fitting assumes that all points are corrupted by independent and identically distributed Gaussian noise, and one single outlier can exert substantial influence on the reconstruction [Gelman et al. 2004].

These considerations motivate adopting the model

$$\min_{u \in U} \|S(u) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u\|_{\mathcal{X}}^2,$$

where the set $U \subset \mathcal{X}$ is convex and closed, representing physical constraints on the unknown u. We are mainly interested in various structural properties of the L^1-norm fitting compared with the more conventional L^2-norm counterpart. Our main goal in this work is to resolve the computational obstacles posed by the non-differentiability of the L^1-norm and nonlinearity of the operator S, such that Newton-type methods are applicable when the operator S has the necessary differentiability properties.

Due to the practical significance of L^1 models, there has been a growing interest in analyzing their properties and in developing efficient minimization algorithms, e.g., in imaging [Kärkkäinen, Kunisch, and Majava 2005; Clason, Jin, and Kunisch 2010a] as well as parameter identification [Chaabane, Ferchichi, and Kunisch 2004]. A number of recent works have addressed the analytical properties of models with L^1 fitting, explaining their superior performance over the standard model for certain types of noise and elaborating the geometrical structure of the minimizers in the context of image denoising [Chan and Esedoğlu 2005; Allard 2007/08; Yin, Goldfarb, and Osher 2007; Duval, Aujol, and Gousseau 2009], i.e., when S is the identity operator. In addition, several efficient algorithms [Yang, Zhang, and Yin 2009; Dong, Hintermüller, and Neri 2009; Clason, Jin, and Kunisch 2010a; Clason, Jin, and Kunisch 2010b] have been developed for such problems.

However, all these works are only concerned with linear inverse problems, and their analysis and algorithms are not directly applicable to the nonlinear case of our interest. The optimality

system is not differentiable in a generalized sense, and thus can not be solved directly with a (semismooth) Newton method. We consider a smoothed variant, and prove the convergence as the smoothing parameter tends to zero. The smoothed optimality system is solved by a semismooth Newton method, and its superlinear local convergence is established under a second-order condition. To the best of our knowledge, this work represents a first investigation on L^1 fitting with general nonlinear inverse problems. The applicability of the proposed approach and its numerical performance is illustrated with several benchmark problems for distributed parameter identification for elliptic partial differential equations.

The rest of this work is organized as follows. In the remainder of this section, we introduce a selection of model problems for which our approach is applicable (§ 14.1.1) and state a precise definition of the considered noise models (§ 14.1.2). In section 14.2, we discuss well-posedness and regularization properties for nonlinear L^1 fitting (§ 14.2.1), and derive the optimality system (§ 14.2.2). The approximating problem, its convergence as the smoothing parameter tends to zero, and its numerical solution using a semismooth Newton method are studied in section 14.3. We also discuss the important issue of choosing suitable regularization and smoothing parameters. Finally, in section 14.4, we present numerical results for our model problems.

14.1.1 MODEL PROBLEMS

In this part, we describe three nonlinear model problems – an inverse potential problem, an inverse Robin coefficient problem and an inverse diffusion coefficient problem – for which our semismooth Newton method is applicable.

INVERSE POTENTIAL PROBLEM A first nonlinear model problem consists in recovering the potential term in an elliptic equation. Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain with a Lipschitz boundary Γ. We consider the equation

$$\begin{cases} -\Delta y + uy = f & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} = 0 & \text{on } \Gamma. \end{cases} \tag{14.1.1}$$

The inverse problem is to recover the potential u defined on $\omega = \Omega$ from noisy observational data y^δ in the domain $D = \Omega$, i.e., S maps $u \in \mathcal{X} = L^2(\Omega)$ to the solution $y \in \mathcal{Y} = H^1(\Omega)$ of (14.1.1). Such problems arise in heat transfer, e.g., damping design [Stojanovic 1991] and identifying heat radiative coefficients [Yamamoto and Zou 2001]. We shall seek u in the admissible set $\mathcal{U} = \{u \in L^\infty(\Omega) : u \geqslant c\} \subset \mathcal{X}$ for some fixed $c > 0$.

INVERSE ROBIN COEFFICIENT PROBLEM Our second example considers the recovery of a Robin boundary condition from boundary observation. Let $\Omega \subset \mathbb{R}^2$ be an open bounded domain with a Lipschitz boundary Γ consisting of two disjoint parts Γ_i and Γ_c. We consider the equation

$$\begin{cases} -\Delta y = 0 & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} = f & \text{on } \Gamma_c, \\ \dfrac{\partial y}{\partial n} + uy = 0 & \text{on } \Gamma_i. \end{cases} \tag{14.1.2}$$

The inverse problem consists in recovering the Robin coefficient u defined on $\omega = \Gamma_i$ from noisy observational data y^δ on the boundary $D = \Gamma_c$, i.e., S maps $u \in \mathcal{X} = L^2(\Gamma_i)$ to $y|_{\Gamma_c} \in \mathcal{Y} = H^{\frac{1}{2}}(\Gamma_c)$, where $v \mapsto v|_{\Gamma_c}$ denotes the Dirichlet trace operator and y is the solution to (14.1.2). This class of problems arises in corrosion detection and thermal analysis of quenching processes [Chaabane, Ferchichi, and Kunisch 2004; Jin and Zou 2010]. We shall seek u in the admissible set $U = \{u \in L^\infty(\Gamma_i) : u \geqslant c\} \subset \mathcal{X}$ for some fixed $c > 0$.

INVERSE DIFFUSION COEFFICIENT PROBLEM Our last example, identification of a diffusion coefficient, addresses stronger regularization for the parameter. Let $\Omega \subset \mathbb{R}^2$ be an open bounded domain with a smooth boundary Γ. We consider the equation

$$\begin{cases} -\nabla \cdot (u \nabla y) = f & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma. \end{cases} \tag{14.1.3}$$

with $f \in L^q(\Omega)$ for some $q > 2$. The inverse problem consists in recovering the diffusion coefficient u within $\omega = \Omega$ from the noisy observational data y^δ in the domain $D = \Omega$, i.e., S maps $u \in \mathcal{X} = H^1(\Omega)$ to the solution $y \in \mathcal{Y} = W_0^{1,q}(\Omega)$, $q > 2$, of (14.1.3). Such problems arise in estimating the permeability of underground flow and the conductivity of heat transfer [Yeh 1986; Banks and Kunisch 1989; Chen and Zou 1999]. We shall seek u in the admissible set $U = \{u \in L^\infty(\Omega) : \lambda \leqslant u \leqslant \lambda^{-1}\} \cap \mathcal{X}$ for some fixed $\lambda \in (0, 1)$.

These model problems share the following properties, which are verified in Appendix 14.A and are sufficient to guarantee the applicability of our approach.

(A1) The operator S is uniformly bounded in $U \subset \mathcal{X}$ and completely continuous: If for $u \in U$, the sequence $\{u_n\} \subset U$ satisfies $u_n \rightharpoonup u$ in $\mathcal{X}$, then

$$S(u_n) \to S(u) \quad \text{in } L^2(D).$$

(A2) S is twice Fréchet differentiable.

(A3) There exists a constant $C > 0$ such that for all $u \in U$ and $h \in \mathcal{X}$ there holds

$$\|S'(u)h\|_{L^2} \leqslant C\|h\|_{\mathcal{X}}.$$

(A4) There exists a constant $C > 0$ such that for all $u \in U$ and $h \in X$ there holds

$$\|S''(u)(h,h)\|_{L^2} \leqslant C\|h\|_X^2.$$

The twice differentiability of S in (A2) is required for a Newton method (cf. section 14.3.2), and ensures strict differentiability required for the chain rule; cf. the proof of Theorem 14.2.7. The a priori estimate in (A3) is employed in analyzing the convergence of the approximate solutions, while (A4) will be used to show local superlinear convergence of the semismooth Newton method.

14.1.2 NOISE MODEL

We now motivate the use of L^1 fitting for impulsive noise from a statistical viewpoint (cf. [Huber 1981; Gelman et al. 2004; Hintermüller and Rincon-Camacho 2010]). The exact data $y^\dagger = S(u^\dagger)$, where $u^\dagger$ is the true solution, defined over a domain D, is corrupted by noise. The contaminated observation y^δ is formed pointwise by

$$y^\delta(x) = f(y^\dagger, \xi_r)(x) \quad x \in D,$$

where $\xi_r(x)$ is a real-valued random variable, $r \in [0,1]$ is a noise parameter, and the function f represents the noise formation mechanism. We assume that for any two distinct points $x_1, x_2 \in D$, $\xi_r(x_1)$ and $\xi_r(x_2)$ are independent. In practice, the Gaussian noise model (and hence L^2 fitting) stands out predominantly. This is often justified by appealing to the celebrated central limit theorem: a Gaussian distribution is suitable for data that are formed as the sum of a large number of independent components [Gelman et al. 2004]. Even in the absence of such justifications, this model is still often preferred due to its computational and analytical conveniences. However, it is also clear that not all real-world data can be adequately described by the Gaussian model. Here, we consider *impulsive* noise models: There exist (many) points $x \in D$ with $f(y, \xi_r)(x) = y(x)$. The two most common types of impulsive noises, e.g., arising in digital image processing [Bovik 2005], are:

SALT-AND-PEPPER NOISE This model is especially common in image processing, and it reflects a wide variety of processes that result in the same image degradation: the corrupted data points (where $\xi_r \neq 0$) only take a fixed maximum ("salt") or minimum ("pepper") value. A simple model is as follows:

$$y^\delta(x) = \begin{cases} y^\dagger(x) & \text{with probability } 1-r, \\ y_{\max} & \text{with probability } \frac{r}{2}, \\ y_{\min} & \text{with probability } \frac{r}{2}, \end{cases}$$

where $y_{\max}$ and $y_{\min}$ are the maximum and minimum of the signal, respectively, and the parameter $r \in (0,1)$ represents the percentage of corrupted data points.

RANDOM-VALUED IMPULSE NOISE (RVIN) In the context of parameter identification problems, it is more reasonable to allow arbitrary random values at the corrupted data points, which gives rise to the following RVIN model

$$y^\delta(x) = \begin{cases} y^\dagger(x) & \text{with probability } 1-r, \\ y^\dagger(x) + \xi(x) & \text{with probability } r, \end{cases}$$

where $\xi(x)$ is a random variable, e.g., normally distributed with mean zero and typically large variance. Clearly, RVIN is generated by the random variable $\xi(x)$ and reproduces the latter if $r = 1$. However, its characteristic is fundamentally different from that of $\xi(x)$ for $r < 1$: there exist data points which are not corrupted by noise, which carry a significant amount of information in the data.

Like many non-Gaussian noise models such as Laplace and Cauchy noise, impulsive noise features significant outliers, i.e., data points that lie far away from the bulk of the data. Statistically, this calls for robust methods (robust estimation in statistics [Huber 1981]). One classical approach is to first identify the outliers with noise detectors, e.g., adaptive median filter, and then perform inversion/reconstruction on the data with outliers excluded. Its success relies crucially on accurate identification of all outliers, which remains very challenging in case of multiple outliers [She and Owen 2011], and mis-identification can significantly compromise the reconstruction. The L^1 approach provides a more systematic strategy for handling outliers due to its ability to implicitly and accurately detect outliers and to automatically prune them from the inversion procedure. The use of L^1 fitting has shown very promising results in a number of practical applications [Bernstein et al. 1974; Clason, Jin, and Kunisch 2010b; Dong, Hintermüller, and Neri 2009]. There have been some theoretical justifications of these empirical observations [Hsu, Kakade, and Zhang 2011]. They are also reflected in the optimality system, where the dual variable acts as a noise detector (cf. Corollary 14.2.8). In contrast, L^2 fitting tends to place equal weight on all data points and thus suffers from a lack of robustness: One single outlier can exert significant influences globally, and may spoil the reconstruction completely [Gelman et al. 2004, p. 443].

We observe that these statistical considerations are finite-dimensional in nature. Nonetheless, they directly motivate the use of the continuous analogue, the L^1 model, for parameter identification problems. We would like to note that the model considered here remains deterministic, despite the preceding statistical motivations. In particular, we do not regard the observational data y^δ as an "impulsive" type of stochastic process in function spaces, but consider it only as a realization of such a stochastic process as is usually the case for deterministic inverse problems [Engl, Hanke, and Neubauer 1996]. However, a stochastic analogue of the L^1 model in function spaces is also of great interest. We recall that the more conventional Gaussian model in function spaces can be modeled as a Hilbert space-valued random variable – and more generally a Hilbert space process – whose properties are characterized by its covariance structure (see the nice summary in [Bissantz et al. 2007, §2.5]). It would be desirable to have analogous characterizations for the L^1 model. Some results in

this direction can be found in [Lassas, Saksman, and Siltanen 2009], where Besov priors were (formally) studied that might also be applied to impulsive noises.

14.2 L^1 FITTING FOR NONLINEAR INVERSE PROBLEMS

The above considerations motivate considering the problem

$$(\mathcal{P}) \qquad \min_{u\in U}\left\{\mathcal{J}_\alpha(u) \equiv \|S(u)-y^\delta\|_{L^1} + \frac{\alpha}{2}\|u-u_0\|_X^2\right\}$$

for the nonlinear operator $S : U \subset X \to Y$ satisfying assumptions (A1)–(A4) (although the results of this and the next section only require (A1) and (A2)) and given $y^\delta \in L^\infty(D)$. Here, u_0 is an initial guess which also plays the role of a selection criterion.

14.2.1 EXISTENCE AND REGULARIZATION PROPERTIES

We first address the well-posedness of the problem ($\mathcal{P}$). In this section, we shall denote a minimizer of the functional $\mathcal{J}_\alpha$ by u_α^δ, while u_α will be a minimizer with y^δ replaced by the exact data $y^\dagger$. We assume that $y^\dagger$ is attainable, i.e., that there exists an element $u^\dagger \in U$ such that $y^\dagger = S(u^\dagger)$. If $u^\dagger$ is not unique, it always refers to a u_0-minimum-norm solution, i.e., an element minimizing $\|u-u_0\|_X$ over the set of solutions to $S(u) = y^\dagger$. Throughout, C denotes a generic constant, whose value may differ at different occurrences.

The proof of the next result is standard (cf., e.g., [Engl, Kunisch, and Neubauer 1989],[Engl, Hanke, and Neubauer 1996, Chap. 10]) and is thus omitted.

Theorem 14.2.1. *Under Assumption (A1), problem ($\mathcal{P}$) is well-posed and consistent, i.e.,*

(i) There exists at least one minimizer $u_\alpha^\delta \in U$ to problem ($\mathcal{P}$);

(ii) For a sequence of data $\{y_n\}$ such that $y_n \to y^\delta$ in $L^1(D)$, the sequence $\{u_\alpha^n\}$ of minimizers contains a subsequence converging to u_α^δ;

(iii) If the regularization parameter $\alpha = \alpha(\delta)$ satisfies

$$\lim_{\delta\to 0}\alpha(\delta) = \lim_{\delta\to 0}\tfrac{\delta}{\alpha(\delta)} = 0,$$

then the sequence $\{u_{\alpha(\delta)}^\delta\}$ has a subsequence converging to $u^\dagger$ as $\delta \to 0$.

If we assume Lipschitz continuity of the derivative S' and a (standard) source condition, we have the following result on the convergence rate for the a priori parameter choice rule $\alpha = \alpha(\delta) \sim \delta^\varepsilon$ for any $\varepsilon \in (0,1)$ (cf. [Hofmann et al. 2007; Pöschl 2009]).

Theorem 14.2.2. *Let $y^\delta \in Y$ satisfy $\|y^\delta - y^\dagger\|_{L^1} \leqslant \delta$ and let $u^\dagger \in U$ be a u_0-minimum norm solution of $S(u) = y^\dagger$. Moreover, let the following conditions be fulfilled:*

(i) There exists an $L > 0$ *such that* $\|S'(u^\dagger) - S'(z)\|_{L^2} \leqslant L\|u^\dagger - z\|_X$ *for all* $z \in U \subset X$.

(ii) There exists a $w \in L^\infty(D) \cap L^2(D)$ *with* $L\,\|w\|_{L^2} < 1$ *satisfying* $u^\dagger - u_0 = S'(u^\dagger)^* w$.

Then for any fixed $\varepsilon \in (0,1)$*, the choice* $\alpha \sim \delta^\varepsilon$ *and* δ *sufficiently small, we have the estimate*

$$\|u_\alpha^\delta - u^\dagger\|_X \leqslant C\delta^{\frac{1-\varepsilon}{2}}.$$

Proof. By the minimizing property of u_α^δ and $\|y^\delta - y^\dagger\|_{L^1} \leqslant \delta$, we obtain

$$\|S(u_\alpha^\delta) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u_\alpha^\delta - u_0\|_X^2 \leqslant \delta + \frac{\alpha}{2}\|u^\dagger - u_0\|_X^2,$$

and hence

$$\|S(u_\alpha^\delta) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u_\alpha^\delta - u^\dagger\|_X^2 \leqslant \delta + \alpha\langle u^\dagger - u_0, u^\dagger - u_\alpha^\delta\rangle_X.$$

Now by the source condition (ii), we obtain

$$\|S(u_\alpha^\delta) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u_\alpha^\delta - u^\dagger\|_X^2 \leqslant \delta + \alpha\langle w, S'(u^\dagger)(u^\dagger - u_\alpha^\delta)\rangle_{L^2}.$$

The Fréchet differentiability of S and condition (i) imply

$$S(u_\alpha^\delta) = S(u^\dagger) + S'(u^\dagger)(u_\alpha^\delta - u^\dagger) + r(u_\alpha^\delta, u^\dagger)$$

with $\left\|r(u_\alpha^\delta, u^\dagger)\right\|_{L^2} \leqslant \frac{L}{2}\|u_\alpha^\delta - u^\dagger\|_X^2$. Combining these estimates leads to

$$\begin{aligned}\|S(u_\alpha^\delta) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u_\alpha^\delta - u^\dagger\|_X^2 &\leqslant \delta + \alpha\langle w, (y^\dagger - y^\delta) + (y^\delta - S(u_\alpha^\delta)) + r(u_\alpha^\delta, u^\dagger)\rangle_{L^2} \\ &\leqslant \delta + \alpha\,\|w\|_{L^\infty}\,\delta + \alpha\,\|w\|_{L^\infty}\,\|S(u_\alpha^\delta) - y^\delta\|_{L^1} \\ &\quad + \frac{\alpha}{2}L\,\|w\|_{L^2}\,\|u_\alpha^\delta - u^\dagger\|_X^2,\end{aligned}$$

and hence

$$(1 - \alpha\,\|w\|_{L^\infty})\|S(u_\alpha^\delta) - y^\delta\|_{L^1} + \frac{\alpha}{2}(1 - L\,\|w\|_{L^2})\|u_\alpha^\delta - u^\dagger\|_X^2 \leqslant \delta + \alpha\,\|w\|_{L^\infty}\,\delta.$$

Now the desired result follows from the condition $L\,\|w\|_{L^2} < 1$ and the choice of α such that $\alpha\,\|w\|_{L^\infty} < 1$ for δ sufficiently small. □

Remark 14.2.3. An inspection of the proof shows that a rate of order $O(\delta^{\frac{1}{2}})$ can be achieved for a choice rule $\alpha(\delta)$ for which the limit $\alpha^* = \lim_{\delta\to 0}\alpha(\delta)$ satisfies $\alpha^* < 1/\|w\|_{L^\infty}$ and $\alpha^* > 0$. We point out that the source condition $u^\dagger - u_0 = S'(u^\dagger)^* w$ might be further relaxed by utilizing the structure of the adjoint operator; cf. [Ito and Jin 2011] for relevant discussions in the context of parameter identification.

The a priori choice gives only an order of magnitude for α and is thus practically inconvenient to use. In contrast, the discrepancy principle [Morozov 1966; Jin, Zhao, and Zou 2012] enables constructing a concrete scheme for determining the regularization parameter α. Specifically, one chooses $\alpha = \alpha(\delta)$ such that

$$\|S(u_\alpha^\delta) - y^\delta\|_{L^1} = c\delta, \tag{14.2.1}$$

where $c \geqslant 1$ is a constant. Numerically, it can be realized efficiently by either a two-parameter algorithm based on model functions or the secant method [Jin, Zhao, and Zou 2012], but it requires knowledge of the noise level δ.

The next result shows a $\mathcal{O}(\delta^{\frac{1}{2}})$ convergence rate. Its proof is almost identical to that of Theorem 14.2.2 (cf. [Jin, Zhao, and Zou 2012]) and hence is omitted.

Theorem 14.2.4 (discrepancy principle). *Let conditions (i)-(ii) in Theorem 14.2.2 be fulfilled. Then for the choice α determined by (14.2.1), there holds*

$$\|u_\alpha^\delta - u^\dagger\|_X \leqslant C\delta^{\frac{1}{2}}.$$

The next result shows an interesting property of L^1 fitting (and in general, of one-homogeneous discrepancy functionals; cf. [Burger and Osher 2004]) in the case of exact data: the regularized solution u_α coincides with the exact solution $u^\dagger$ if the regularization parameter α is sufficiently small. This is in sharp contrast to quadratic L^2 fitting, where the Tikhonov minimizer is different from the true solution for every $\alpha > 0$.

Theorem 14.2.5 (exact recovery). *Let conditions (i) and (ii) in Theorem 14.2.2 be fulfilled. Then, $u_\alpha = u^\dagger$ holds for $\alpha > 0$ sufficiently small.*

Proof. We only sketch the proof. By the minimizing properties of u_α and the source condition, we arrive at

$$\|S(u_\alpha) - y^\dagger\|_{L^1} + \frac{\alpha}{2}\|u_\alpha - u^\dagger\|_X^2 \leqslant -\alpha\langle w, S'(u^\dagger)(u_\alpha - u^\dagger)\rangle_{L^2}.$$

As before, we obtain by the Fréchet differentiability of S that

$$\begin{aligned}\|S(u_\alpha) - y^\dagger\|_{L^1} + \frac{\alpha}{2}\|u_\alpha - u^\dagger\|_X^2 &\leqslant \alpha\langle w, (y^\dagger - S(u_\alpha)) + r(u_\alpha, u^\dagger)\rangle_{L^2} \\ &\leqslant \alpha\|w\|_{L^\infty}\|S(u_\alpha) - y^\dagger\|_{L^1} + \frac{\alpha}{2}L\|w\|_{L^2}\|u_\alpha - u^\dagger\|_X^2.\end{aligned}$$

Hence, for $\alpha \leqslant 1/\|w\|_{L^\infty}$, we have $\|u_\alpha - u^\dagger\|_X = 0$, i.e. $u_\alpha = u^\dagger$. □

14.2.2 OPTIMALITY SYSTEM

We next derive the necessary first-order optimality conditions for $u_\alpha := u_\alpha^\delta$ (slightly abusing the notation).

Remark 14.2.6. In this work, we assume that the true solution $u^\dagger$ of the inverse problem (and a minimizer u_α of (P)) lies in the interior U_{int} of U and do not explicitly enforce the constraint $u \in U$, in order to focus the presentation on the treatment of the nonsmoothness inherent in the L^1-fitting problem. There is no fundamental difficulty in including this constraint in the optimization, however, in which case the first equality in the optimality conditions (OS) should be replaced by a variational inequality. When the domain of definition is given by box constraints (as in the model problems), the modified optimality system can still be solved using a semismooth Newton method after applying a Moreau–Yosida regularization; cf. [Hintermüller, Ito, and Kunisch 2002].

Theorem 14.2.7. *For any local minimizer $u_\alpha \in U_{\mathrm{int}} \subset \mathcal{X}$ of problem (P) there exists a $p_\alpha \in L^\infty(D)$ with $\|p_\alpha\|_{L^\infty} \leqslant 1$ such that the following relations hold:*

$$\text{(OS)} \qquad \begin{cases} S'(u_\alpha)^* p_\alpha + \alpha j(u_\alpha - u_0) = 0, \\ \quad \langle S(u_\alpha) - y^\delta, p - p_\alpha \rangle_{L^2} \leqslant 0 \quad \textit{for all } \|p\|_{L^\infty} \leqslant 1. \end{cases}$$

Here $S'(u)^*$ denotes the adjoint of $S'(u)$ with respect to $L^2(D)$, and $j : \mathcal{X} \to \mathcal{X}^*$ is the (linear) duality mapping, i.e., $j(u) = \partial(\frac{1}{2}\|u\|_{\mathcal{X}}^2)$. Note that both $S(u)$ and y^δ are in $L^2(D)$, and hence the duality pairing $\langle S(u) - y^\delta, p \rangle_{L^1, L^\infty}$ coincides with the standard L^2-inner product.

Proof. Setting

$$\begin{aligned} \mathcal{F} &: \mathcal{X} \to \mathbb{R}, & u &\mapsto \frac{\alpha}{2}\|u - u_0\|_{\mathcal{X}}^2, \\ \mathcal{G} &: L^1(D) \to \mathbb{R}, & v &\mapsto \|v\|_{L^1}, \end{aligned}$$

we have that

$$\mathcal{J}_\alpha(u) = \mathcal{F}(u) + \mathcal{G}(S(u) - y^\delta).$$

Since the operator S is twice Fréchet differentiable ((A2), which implies strict differentiability) and $\mathcal{G}$ is real-valued and convex, the sum and chain rules for the generalized gradient [Clarke 1990, Thms. 2.3.3, 2.3.10] yield that for all $u \in \mathcal{X}$, the functional $\mathcal{J}_\alpha$ is Lipschitz continuous near u and the relation

$$\partial \mathcal{J}_\alpha(u) = \mathcal{F}'(u) + S'(u)^* \partial \mathcal{G}(S(u) - y^\delta)$$

holds. The necessary condition $0 \in \partial \mathcal{J}_\alpha(u_\alpha)$ for every local minimizer u_α of $\mathcal{J}_\alpha$ (cf., e.g., [Clarke 1990, Prop. 2.3.2]) thus implies the existence of a subgradient $p_\alpha \in \partial \mathcal{G}(S(u_\alpha) - y^\delta) \subset L^\infty(D)$ such that

$$0 = \alpha j(u_\alpha - u_0) + S'(u_\alpha)^* p_\alpha$$

holds, which is the first relation of (OS). Since $\mathcal{G}$ is convex, the generalized gradient reduces to the convex subdifferential (cf. [Clarke 1990, Prop. 2.2.7]), and by its definition we have the equivalence

$$p_\alpha \in \partial\mathcal{G}(S(u_\alpha) - y^\delta) \;\Leftrightarrow\; S(u_\alpha) - y^\delta \in \partial\mathcal{G}^*(p_\alpha),$$

where $\mathcal{G}^*$ is the Fenchel conjugate of $\mathcal{G}$ (cf., e.g., [Ekeland and Témam 1999, Chap. I.4]), given by the indicator function of the unit ball $B \equiv \{p \in L^\infty(D) : \|p\|_{L^\infty} \leqslant 1\}$. The subdifferential of $\mathcal{G}^*$ coincides with the normal cone to B. Consequently, we deduce that $p_\alpha \in \partial\mathcal{G}(S(u_\alpha) - y^\delta)$ if and only if

$$\langle S(u_\alpha) - y^\delta, p - p_\alpha\rangle_{L^2} \leqslant 0$$

holds for all $p \in L^\infty(D)$ with $\|p\|_{L^\infty} \leqslant 1$, which is the second relation of (OS). □

The following structural information for a solution u_α of problem ($\mathcal{P}$) is a direct consequence of (OS) and is of independent interest.

Corollary 14.2.8. *Let $u_\alpha \in U_{\text{int}}$ be a minimizer of problem ($\mathcal{P}$) and $p_\alpha \in L^\infty(D)$ as given by Theorem 14.2.7. Then the following relations hold:*

$$\begin{aligned} S(u_\alpha) - y^\delta &= 0 \quad \textit{a.e. on} \quad \{x \in D : |p_\alpha(x)| < 1\},\\ S(u_\alpha) - y^\delta &\geqslant 0 \quad \textit{a.e. on} \quad \{x \in D : p_\alpha(x) = 1\},\\ S(u_\alpha) - y^\delta &\leqslant 0 \quad \textit{a.e. on} \quad \{x \in D : p_\alpha(x) = -1\}. \end{aligned}$$

This can be interpreted as follows: the box constraint on the dual solution p_α is active where the data is not attained by the primal solution u_α. In particular, the dual solution p_α acts as a noise indicator.

By using a complementarity function [Chen, Nashed, and Qi 2000; Ito and Kunisch 2008], we can rewrite the second relation of (OS) as

$$S(u_\alpha) - y^\delta = \max(0, S(u_\alpha) - y^\delta + c(p_\alpha - 1)) + \min(0, S(u_\alpha) - y^\delta + c(p_\alpha + 1))$$

for any $c > 0$. This can be further discriminated by pointwise inspection to the following three cases:

(i) $(S(u_\alpha) - y^\delta)(x) > 0$ and $p_\alpha(x) = 1$,

(ii) $(S(u_\alpha) - y^\delta)(x) < 0$ and $p_\alpha(x) = -1$,

(iii) $(S(u_\alpha) - y^\delta)(x) = 0$ and $p_\alpha(x) \in [-1, 1]$.

Consequently, we have the concise relation

$$p_\alpha = \operatorname{sign}(S(u_\alpha) - y^\delta),$$

from which we obtain a reduced optimality system

(OS′) $$0 \in \alpha j(u_\alpha - u_0) + S'(u_\alpha)^*(\operatorname{sign}(S(u_\alpha) - y^\delta)).$$

14.3 SOLUTION BY SEMISMOOTH NEWTON METHOD

In view of (OS′) and the lack of smoothness of the sign function, the optimality system (OS) is not differentiable even in a generalized sense, which precludes the application of Newton-type methods. Meanwhile, gradient descent methods are inefficient unless the step lengths are chosen appropriately, which, however, necessarily requires a detailed knowledge of Lipschitz constants. Therefore, we propose to approximate ($\mathcal{P}$) using a local smoothing of the L^1 norm. For simplicity, we will only consider $u_0 = 0$ from here on.

14.3.1 APPROXIMATION

To obtain a semismooth Newton system, we wish to replace the sign function in (OS′) by a locally linear smoothing. We therefore consider for $\beta > 0$ the smoothed problem

$$(\mathcal{P}_\beta) \qquad \min_{u \in U} \|S(u) - y^\delta\|_{L^1_\beta} + \frac{\alpha}{2}\|u\|_X^2,$$

where $\|v\|_{L^1_\beta}$ is a Huber-type smoothing of the L^1 norm:

$$\|v\|_{L^1_\beta} \equiv \int_\Omega |v(x)|_\beta \, dx, \qquad |v(x)|_\beta \equiv \begin{cases} v(x) - \frac{\beta}{2} & \text{if } v(x) > \beta, \\ -v(x) - \frac{\beta}{2} & \text{if } v(x) < -\beta, \\ \frac{1}{2\beta} v(x)^2 & \text{if } |v(x)| \leqslant \beta. \end{cases}$$

The existence of a minimizer u_β of ($\mathcal{P}_\beta$) follows as before. Since the mapping $\psi : \mathbb{R} \to \mathbb{R}$, $t \mapsto |t|_\beta$, is differentiable with a globally Lipschitz continuous derivative $t \mapsto \operatorname{sign}_\beta(t)$,

$$\operatorname{sign}_\beta(t) \equiv \begin{cases} 1 & \text{if } t > \beta, \\ -1 & \text{if } t < -\beta, \\ \frac{1}{\beta} t & \text{if } |t| \leqslant \beta, \end{cases}$$

we have that ψ defines a differentiable Nemytskii operator from $L^p(D)$ to $L^2(D)$ for every $p \geqslant 4$ (cf., e.g., [Tröltzsch 2010, Chap. 4.3] and references therein) with pointwise defined derivative $\operatorname{sign}_\beta(v)h$. We thus obtain the necessary optimality conditions for a minimizer $u_\beta \in U_{\text{int}}$:

$$(\mathrm{OS}_\beta) \qquad \alpha j(u_\beta) + S'(u_\beta)^*(\operatorname{sign}_\beta(S(u_\beta) - y^\delta)) = 0.$$

Remark 14.3.1. This Huber-type smoothing (which is also used in classical robust estimation [Huber 1981]) is equivalent to an $L^2(\Omega)$-penalization of the dual variable $p \in L^\infty(D)$ in (OS). To see this, we consider (OS) as the optimality conditions of the primal-dual saddle point problem

$$\min_{u \in U} \max_{\|p\|_{L^\infty} \leqslant 1} \langle S(u) - y^\delta, p \rangle_{L^2} + \frac{\alpha}{2} \|u\|_X^2,$$

which makes use of the dual representation of the L^1-norm. We now introduce for $\beta > 0$ the penalized saddle point problem

$$\min_{u\in U}\left(\max_{\|p\|_{L^\infty}\leqslant 1}\langle S(u)-y^\delta,p\rangle_{L^2}-\frac{\beta}{2}\|p\|_{L^2}^2\right)+\frac{\alpha}{2}\|u\|_X^2.$$

The corresponding optimality conditions for minimizers in U_{int} are given by

$$\begin{cases} S'(u_\beta)^* p_\beta+\alpha j(u_\beta)=0,\\ \langle S(u_\beta)-y^\delta-\beta p_\beta,p-p_\beta\rangle_{L^2}\leqslant 0 \end{cases} \tag{14.3.1}$$

for all $p\in L^\infty(D)$ with $\|p\|_{L^\infty}\leqslant 1$. By expressing the variational inequality again using a complementarity function with $c=\beta$, we obtain by pointwise inspection that

$$p_\beta=\operatorname{sign}_\beta(S(u_\beta)-y^\delta).$$

Inserting this expression into the first relation of (14.3.1) yields precisely (OS_β).

We next show the convergence of solutions to the approximating problems ($\mathcal{P}_\beta$) to a solution to problem ($\mathcal{P}$).

Theorem 14.3.2. *As $\beta\to 0$, the family $\{u_\beta\}_{\beta>0}\subset U$ of minimizers of ($\mathcal{P}_\beta$) contains a subsequence converging in X to a minimizer of ($\mathcal{P}$).*

Proof. Note that for any $\beta>0$, there holds $|\nu(x)|_\beta\leqslant|\nu(x)|$, and consequently

$$\|S(u_\alpha)-y^\delta\|_{L^1_\beta}\leqslant\|S(u_\alpha)-y^\delta\|_{L^1}.$$

Now the minimizing property of u_β implies

$$\|S(u_\beta)-y^\delta\|_{L^1_\beta}+\frac{\alpha}{2}\|u_\beta\|_X^2\leqslant\|S(u_\alpha)-y^\delta\|_{L^1_\beta}+\frac{\alpha}{2}\|u_\alpha\|_X^2, \tag{14.3.2}$$

from which it follows that the family $\{u_\beta\}$ is uniformly bounded in U. Therefore, there exists a subsequence, also denoted by $\{u_\beta\}$, and some $u^*\in U\subset X$ such that $u_\beta\rightharpoonup u^*$ in X. By the strong continuity of S (cf. (A1)) we have $S(u_\beta)\to S(u^*)$ in L^2, and this convergence is pointwise almost everywhere after possibly passing to a further subsequence [Evans and Gariepy 1992]. In addition, since $|t|_\beta\to|t|$ as $\beta\to 0$ for every $t\in\mathbb{R}$, we have that $|S(u_\alpha)-y^\delta|_\beta$ converges pointwise to $|S(u_\alpha)-y^\delta|$. Fatou's Lemma then implies

$$\|S(u^*)-y^\delta\|_{L^1}=\int_D\lim_{\beta\to 0}|S(u_\beta)-y^\delta|_\beta\,dx\leqslant\liminf_{\beta\to 0}\|S(u_\beta)-y^\delta\|_{L^1_\beta}. \tag{14.3.3}$$

Meanwhile, by virtue of Lebesgue's dominated convergence theorem [Evans and Gariepy 1992], we deduce

$$\lim_{\beta\to 0}\|S(u_\alpha)-y^\delta\|_{L^1_\beta}=\|S(u_\alpha)-y^\delta\|_{L^1}.$$

These three relations together with the weak lower semicontinuity of norms indicate

$$\|S(u^*) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u^*\|_X^2 \leqslant \|S(u_\alpha) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u_\alpha\|_X^2.$$

This together with the minimizing property of u_α implies that u^* is a minimizer of ($\mathcal{P}$).

To conclude the proof, it suffices to show that $\limsup_{\beta\to 0}\|u_\beta\|_X \leqslant \|u^*\|_X$ holds. To this end, we assume the contrary, i.e., that there exists a subsequence of $\{u_\beta\}_{\beta>0}$, also denoted by $\{u_\beta\}$, satisfying $u_\beta \rightharpoonup u^*$ in X and $\lim_{\beta\to 0}\|u_\beta\|_X \equiv c > \|u^*\|_X$. Letting $u_\alpha = u^*$ and $\beta \to 0$ in (14.3.2), we arrive at

$$\limsup_{\beta\to 0}\|S(u_\beta) - y^\delta\|_{L^1_\beta} + \frac{\alpha}{2}c^2 \leqslant \|S(u^*) - y^\delta\|_{L^1} + \frac{\alpha}{2}\|u^*\|_X^2,$$

i.e., $\limsup_{\beta\to 0}\|S(u_\beta) - y^\delta\|_{L^1_\beta} < \|S(u^*) - y^\delta\|_{L^1}$, which is in contradiction with the weak lower semicontinuity in (14.3.3). This concludes the proof. □

14.3.2 SEMISMOOTH NEWTON METHOD

To solve the optimality system (OS_β) with a semismooth Newton method [Kummer 2000; Chen, Nashed, and Qi 2000; Ulbrich 2002; Hintermüller, Ito, and Kunisch 2002], we consider it as an operator equation $F(u) = 0$ for $F : U \subset X \to X^*$,

$$F(u) = \alpha j(u) + S'(u)^*(\operatorname{sign}_\beta(S(u) - y^\delta)).$$

We now argue the Newton differentiability of F. We recall that a mapping $F : X \to Y$ between Banach spaces X and Y is Newton differentiable at $x \in X$ if there exists a neighborhood $N(x)$ and a mapping $G : N(x) \to L(X, Y)$ with

$$\lim_{\|h\|\to 0} \frac{\|F(x+h) - F(x) - G(x+h)h\|_Y}{\|h\|_X} \to 0. \tag{14.3.4}$$

(Note that in contrast with Fréchet differentiability, the linearization is taken in a neighborhood $N(x)$ of x.) Any mapping $D_N F \in \{G(s) : s \in N(x)\}$ is then a Newton derivative of F at x.

Since $t \mapsto \operatorname{sign}_\beta(t)$ is a globally Lipschitz continuous mapping from $\mathbb{R}$ to $\mathbb{R}$, the corresponding Nemytskii operator $v \mapsto \operatorname{sign}_\beta(v)$ is Newton differentiable from L^p to L^q for any $p > q \geqslant 1$ [Ulbrich 2002; Schiela 2008], and a Newton derivative is given pointwise by

$$\big(D_N \operatorname{sign}_\beta(v)h\big)(x) = \begin{cases} 0 & \text{if } |v(x)| > \beta, \\ \frac{1}{\beta}h(x) & \text{if } |v(x)| \leqslant \beta. \end{cases}$$

This yields Newton differentiability of $\operatorname{sign}_\beta$ from $Y \hookrightarrow L^q(D)$, $q > 2$, to $L^2(D)$. By the chain rule and the Fréchet differentiability of S, it follows that $P : U \to L^2(D)$,

$$P(u) = \operatorname{sign}_\beta(S(u) - y^\delta),$$

is Newton differentiable as well, and a Newton derivative acting on a direction $v \in \mathcal{X}$ is given as

$$D_N P(u)h = \beta^{-1}\chi_{\mathcal{I}}(S'(u)h).$$

Here, $\chi_{\mathcal{I}}$ is defined pointwise for $x \in D$ by

$$\chi_{\mathcal{I}}(x) = \begin{cases} 1 & \text{if } |(S(u) - y^\delta)(x)| \leqslant \beta, \\ 0 & \text{else.} \end{cases}$$

For a given u^k, one Newton step consists in solving for the increment $\delta u \in \mathcal{X}$ in

$$\alpha j'(u^k)\delta u + (S''(u^k)\delta u)^* P(u^k) + \beta^{-1} S'(u^k)^*(\chi_{\mathcal{I}^k} S'(u^k)\delta u) = -F(u^k) \tag{14.3.5}$$

and setting $u^{k+1} = u^k + \delta u$. Given a way to compute the action of the derivatives $S'(u)h$, $S'(u)^*h$ and $[S''(u)h]^*p$ for given u, p and h (given in Appendix 14.A for the model problems), system (14.3.5) can be solved iteratively, e.g., using a Krylov method.

It remains to show the uniform well-posedness of system (14.3.5), from which superlinear convergence of the semismooth Newton method follows by standard arguments. Since the operator S is nonlinear and the functional is possibly non-convex, we assume the following condition at a minimizer u_β: There exists a constant $\gamma > 0$ such that

$$\langle S''(u_\beta)(h,h), P(u_\beta)\rangle_{L^2} + \alpha\|h\|_{\mathcal{X}}^2 \geqslant \gamma\|h\|_{\mathcal{X}}^2 \tag{14.3.6}$$

holds for all $h \in \mathcal{X}$. This is related to standard second-order sufficient optimality conditions in PDE-constrained optimization (cf., e.g., [Tröltzsch 2010, Chap. 4.10]). The condition is satisfied for either large α or sparse residual $S(u_\beta) - y^\delta$, since

$$\langle S''(u_\beta)(h,h), P(u_\beta)\rangle_{L^2} + \alpha\|h\|_{\mathcal{X}}^2 \geqslant (\alpha - C\|P(u_\beta)\|_{L^2})\|h\|_{\mathcal{X}}^2 \tag{14.3.7}$$

holds by the a priori estimate on S'' (A4). In the context of parameter identification problems, this is a reasonable assumption, since for a large noise level, α would take a large value, while a small α is chosen only for small noise levels (which, given the impulsive nature of the noise, is equivalent to strong sparsity of the residual). In the latter case, we observe that $P(u_\beta) = \operatorname{sign}_\beta(S(u_\beta) - y^\delta)$ can be expected to be small due to the L^2 smoothing of $\operatorname{sign}_\beta$ (cf. Remark 14.3.1 and note that $P(u_\beta) = p_\beta$). Condition (14.3.6) is thus satisfied if either α or β is sufficiently large. However, this property depends on β, which together with Theorem 14.3.2 motivates the use of a continuation strategy in β; cf. section 14.3.3. We remind that in general it is not possible to check such conditions a priori even for quadratic functionals.

Proposition 14.3.3. *Let $\beta > 0$ be given. If condition (14.3.6) holds, then for each $u \in U$ sufficiently close to a solution $u_\beta \in U_{\text{int}}$ of (OS_β), the mapping $D_N F : \mathcal{X} \to \mathcal{X}^*$,*

$$D_N F(u) = \alpha j'(u) + S''(u)^* P(u) + \beta^{-1} S'(u)^* \chi_I S'(u),$$

is invertible, and there exists a constant $C > 0$ independent of u such that

$$\|(D_N F)^{-1}\|_{L(\mathcal{X}^*,\mathcal{X})} \leqslant C.$$

Proof. For given $w \in \mathcal{X}^*$, we need to find $\delta u \in \mathcal{X}$ satisfying

$$\langle \alpha j'(u)\delta u + (S''(u)\delta u)^* P(u) + \beta^{-1} S'(u)^* \chi_I S'(u)\delta u, v\rangle_{\mathcal{X}^*,\mathcal{X}} = \langle w, v\rangle_{\mathcal{X}^*,\mathcal{X}}$$

for all $v \in \mathcal{X}$. Letting $v = \delta u$ and observing that $\langle j'(u)v, v\rangle_{\mathcal{X}^*,\mathcal{X}} = \|v\|_{\mathcal{X}}^2$ (since $\mathcal{X}$ is a Hilbert space), we obtain

$$\alpha \|\delta u\|_{\mathcal{X}}^2 + \langle S''(u)(\delta u, \delta u), P(u)\rangle_{L^2} + \beta^{-1} \|\chi_I S'(u)\delta u\|_{L^2}^2 = \langle w, \delta u\rangle_{\mathcal{X}^*,\mathcal{X}}.$$

Now the pointwise contraction property of the min and the max function implies

$$\begin{aligned}\|P(u_\beta) - P(u)\|_{L^2} &\leqslant \beta^{-1}\|S(u_\beta) - S(u)\|_{L^2} \\ &\quad + \beta^{-1}\|\max(0, S(u_\beta) - y^\delta - \beta) - \max(0, S(u) - y^\delta - \beta)\|_{L^2} \\ &\quad + \beta^{-1}\|\min(0, S(u_\beta) - y^\delta + \beta) - \min(0, S(u) - y^\delta + \beta)\|_{L^2} \\ &\leqslant 3\beta^{-1}\|S(u_\beta) - S(u)\|_{L^2}.\end{aligned}$$

Consequently, by the continuity of the mapping S, for sufficiently small $\|u_\beta - u\|_{\mathcal{X}}$, we have small $\|P(u_\beta) - P(u)\|_{L^2}$ as well. Thus, by condition (14.3.6) and the locally uniform boundedness of S'' (cf. (A4)), there exists an $\varepsilon > 0$ such that

$$\begin{aligned}&\alpha \|\delta u\|_{\mathcal{X}}^2 + \langle S''(u)(\delta u, \delta u), P(u)\rangle_{L^2} \\ &\qquad = \alpha \|\delta u\|_{\mathcal{X}}^2 + \langle S''(u)(\delta u, \delta u), P(u_\beta)\rangle_{L^2} + \langle S''(u)(\delta u, \delta u), P(u) - P(u_\beta)\rangle_{L^2} \\ &\qquad \geqslant \gamma \|\delta u\|_{\mathcal{X}}^2 - C\varepsilon \|\delta u\|_{\mathcal{X}}^2 \geqslant \frac{\gamma}{2} \|\delta u\|_{\mathcal{X}}^2\end{aligned}$$

holds for all u with $\|u - u_\beta\|_{\mathcal{X}} \leqslant \varepsilon$ if ε is sufficiently small.

Finally, we deduce by the Cauchy–Schwarz inequality that

$$\frac{\gamma}{4} \|\delta u\|_{\mathcal{X}}^2 \leqslant \|w\|_{\mathcal{X}^*} \|\delta u\|_{\mathcal{X}}.$$

This implies the claim. □

Newton differentiability and uniform boundedness of Newton derivatives immediately implies superlinear convergence of the semismooth Newton method (14.3.5).

Theorem 14.3.4. *Let $\beta > 0$ and condition (14.3.6) hold. Then the sequence $\{u^k\}$ of iterates in (14.3.5) converge superlinearly to a solution $u_\beta \in U_{\text{int}}$ of (OS_β), provided that u^0 is sufficiently close to u_β.*

Proof. The proof is standard [Kummer 2000; Chen, Nashed, and Qi 2000; Ulbrich 2002; Hintermüller, Ito, and Kunisch 2002] but given here for the sake of completeness. By the

definition of the Newton step $u^{k+1} = u^k - (D_N F(u^k))^{-1} F(u^k)$ and $F(u_\beta) = 0$, we obtain using Proposition 14.3.3 that

$$\begin{aligned}\|u^{k+1} - u_\beta\|_X &= \|(D_N F(u^k))^{-1}[F(u^k) - F(u_\beta) - D_N F(u^k)(u^k - u_\beta)]\|_X \\ &\leqslant C\,\|F(u^k) - F(u_\beta) - D_N F(u^k)(u^k - u_\beta)\|_{X^*} \\ &= C\|F(u_\beta + d^k) - F(u_\beta) - D_N F(u_\beta + d^k)d^k\|_{X^*}\end{aligned}$$

with $d^k := u^k - u_\beta \in \mathcal{X}$. Now the Newton differentiability of F at u_β implies that

$$\|u^{k+1} - u_\beta\|_X = o(\|u^k - u_\beta\|_X),$$

and thus there exists a neighborhood of u_β such that

$$\|u^{k+1} - u_\beta\|_X < \frac{1}{2}\|u^k - u_\beta\|_X,$$

from which convergence follows by induction. Applying (14.3.4) then yields the claimed superlinear rate. □

14.3.3 PARAMETER CHOICE

The regularized formulation ($\mathcal{P}$) of the parameter identification problem $S(u) = y^\delta$ requires specifying the regularization parameter α, whose correct choice is crucial in practice. Usually, it is determined using a knowledge of the noise level δ by, e.g., the discrepancy principle (14.2.1). However, in practice, the noise level δ may be unknown, rendering such rules inapplicable. To circumvent this issue, we propose a heuristic choice rule based on the following balancing principle [Clason, Jin, and Kunisch 2010a]: Choose α such that

$$(\sigma - 1)\|S(u_\alpha) - y^\delta\|_{L^1} - \frac{\alpha}{2}\|u_\alpha\|_X^2 = 0 \tag{14.3.8}$$

is satisfied. The underlying idea of the principle is to balance the data fitting term with the penalty term, and the weight $\sigma > 1$ controls the trade-off between them. This weight depends on the relative smoothness of residual and parameter but not on the data realization. The principle does not require knowledge of the noise level and has been successfully applied to linear inverse problems with L^1 data fitting [Clason, Jin, and Kunisch 2010a; Clason, Jin, and Kunisch 2010b]; cf. also [Ito, Jin, and Takeuchi 2011] for relevant theoretical analysis.

We compute a solution α^* to the balancing equation (14.3.8) by the following simple fixed point algorithm proposed in [Clason, Jin, and Kunisch 2010b]:

$$\alpha_{k+1} = (\sigma - 1)\frac{\|S(u_{\alpha_k}) - y^\delta\|_{L^1}}{\frac{1}{2}\|u_{\alpha_k}\|_X^2}. \tag{14.3.9}$$

This fixed point algorithm can be derived formally from the model function approach [Clason, Jin, and Kunisch 2010a]. The convergence can be proved similar to [Clason, Jin, and Kunisch 2010b], by observing that the proof given there does not depend on the linearity of the forward operator.

Theorem 14.3.5. *If the initial guess* α_0 *satisfies* $(\sigma - 1)\|S(u_{\alpha_0}) - y^\delta\|_{L^1} - \frac{\alpha_0}{2}\|u_{\alpha_0}\|_X^2 < 0$, *then the sequence* $\{\alpha_k\}$ *generated by the fixed point algorithm is monotonically decreasing and converges to a solution to* (14.3.8).

Of similar importance is the proper choice of the smoothing parameter β. If β is too large, the desirable structural property of the L^1 model will be lost. However, the second-order condition (14.3.7) depends on β and cannot be expected to hold for arbitrarily small β. In particular, the convergence radius for the semismooth Newton method is likely to shrink as β decreases to zero. These considerations motivate the following continuation strategy: Starting with a large β_0 and setting $\beta_{n+1} = q\beta_n$ for some $q \in (0, 1)$, we compute the solution u_{β_n} of (OS_β) using the previous solution u_{β_n} as an initial guess.

A crucial issue is then selecting an appropriate stopping criterion for the continuation. Since we are most interested in the L^1 structure of the problem, we base our stopping rule on the following finite termination property of the linear L^1 fitting problem [Clason, Jin, and Kunisch 2010a, Prop. 3.6]: If the active sets coincide for two consecutive iterations of the semismooth Newton method, the semismooth optimality system is solved exactly. In addition, the convergence is usually very fast due to the continuation strategy, and the required number of iterations is independent of the mesh size (this property is well-known as *mesh independence* [Hintermüller and Kunisch 2004]). Hence, if the active sets (cf. $\mathcal{A}_+^k$ and $\mathcal{A}_-^k$ in Algorithm 14.1) are still changing after a fixed number of iterations, we deduce that the semismoothness of the operator $F(u)$ might be lost and return the last feasible solution $u_{\beta_{n-1}}$ as the desired approximation. In practice, we also check for smallness of the norm of the gradient to take into account the nonlinearity of S and safeguard termination of the algorithm by stopping the continuation if a given very small value $\beta_{\min}$ is reached.

A complete description of this approach, hereafter called *path-following semismooth Newton method*, is given in Algorithm 14.1.

14.4 NUMERICAL EXAMPLES

We now present some numerical results for several benchmark parameter identification problems with one- and two-dimensional elliptic differential equations to illustrate the features of the proposed approach. In each case, the forward operator was discretized using finite elements on a uniform grid (triangular, in the case of two dimensions). We denote by P_0 the space of piecewise constant functions (on each element), while P_1 is the space of piecewise linear functions. Unless otherwise stated, the number N of grid points is 1001 in one dimension and 128×128 in two dimensions.

We implemented the semismooth Newton (SSN) method as given in Algorithm 14.1. The iteration was terminated if the active sets did not change and the norm of the gradient fell

Algorithm 14.1 Path-following semismooth Newton method.

Choose β_0, $q < 1$, $\beta_{\min} > 0$, $u_0 \in U$, $k^* > 0$, set $n = 0$
repeat
  Set $u^0 = u_n$, $k = 0$
  repeat
    Compute $y^k = S(u^k)$
    Compute active and inactive sets

$$\mathcal{A}_+^k = \{x \in D : (y^k - y^\delta)(x) > \beta\}$$
$$\mathcal{A}_-^k = \{x \in D : (y^k - y^\delta)(x) < -\beta\}$$
$$\mathcal{I}^k = \{x \in D : |(y^k - y^\delta)(x)| \leqslant \beta\}$$

    Compute $p^k = \operatorname{sign}_\beta(y^k - y^\delta)$
    Compute $F(u^k) = \alpha j(u^k) + S'(u^k)^*(p^k)$
    Compute update δu by solving

$$\alpha j'(u^k)\delta u + (S''(u^k)\delta u)^* p^k + \beta^{-1} S'(u^k)^* \chi_{\mathcal{I}^k} S'(u^k)\delta u = -F(u^k)$$

    Set $u^{k+1} = u^k + \delta u$, $k \leftarrow k + 1$.
  until ($\mathcal{A}_+^k = \mathcal{A}_+^{k-1}$ and $\mathcal{A}_-^k = \mathcal{A}_-^{k-1}$ and $\|F(u^k)\| \leqslant \text{tol}$) or $k = k^*$
  if $k < k^*$ **then**
    Set $n \leftarrow n + 1$, $u_n = u^k$, $\beta_n = q\beta_{n-1}$
  end if
until $k = k^*$ or $\beta_n < \beta_{\min}$

below 1.00×10^{-6}, or if 20 iterations were reached. In our experiments, we consider random-valued impulsive noise (cf. § 14.1.2): Given the true solution $u^\dagger$ and the corresponding exact data $y^\dagger = S(u^\dagger)$, we set

$$y^\delta = \begin{cases} y^\dagger & \text{with probability } 1 - r, \\ y^\dagger + \|y^\dagger\|_{L^\infty}\xi & \text{with probability } r, \end{cases}$$

where the random variable ξ follows the standard normal distribution and $r \in (0, 1)$ is the percentage of corrupted data points. Unless otherwise noted, we take $r = 0.3$. The exact noise level δ is defined by $\delta = \|y^\delta - y^\dagger\|_{L^1}$. The Newton system (14.3.5) is solved iteratively using BiCGstab (with tolerance 1.00×10^{-6} and maximum number of iterations 100). The reduction rate q is set to $\frac{1}{2}$.

All timing tests were performed with Matlab (R2010b) on a single core of a 2.8 GHz workstation with 24 GByte of RAM. The Matlab codes of our implementation can be downloaded from http://www.uni-graz.at/~clason/codes/l1nonlinfit.zip. To keep the presentation concise, all tables are collected in Appendix 14.B.

14.4.1 INVERSE POTENTIAL PROBLEM

This example is concerned with determining the potential $u \in L^2(\Omega)$ in (14.1.1) from noisy measurements of the state $y \in H^1(\Omega)$ in the domain Ω. The discretized operator S_h maps $u_h \in U_h = P_0$ to $y_h \in Y_h = P_1$ which satisfies

$$\langle \nabla y_h, \nabla v_h \rangle_{L^2(\Omega)} + \langle u_h y_h, v_h \rangle_{L^2(\Omega)} = \langle f, v_h \rangle_{L^2(\Omega)} \quad \text{for all } v_h \in Y_h.$$

For the automatic parameter choice using the balancing principle, we have set the weight σ to 1.03 and the initial guess α_0 to 1.

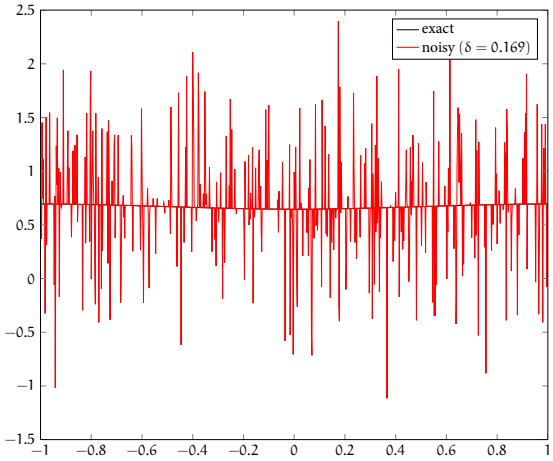
ONE-DIMENSIONAL EXAMPLE. Here, we take $\Omega = [-1, 1]$, $f(x) = 1$ and

$$u^\dagger(x) = 2 - |x| \geqslant 1.$$

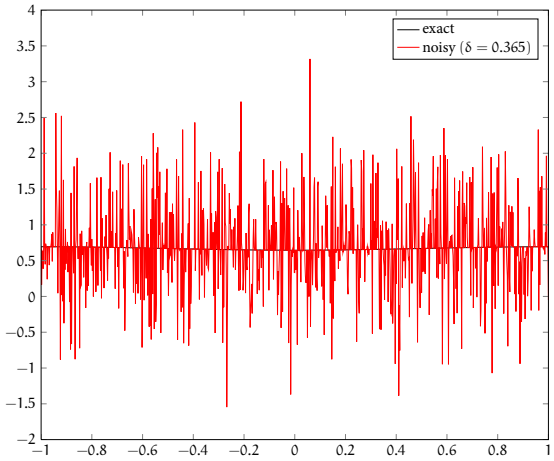
A typical realization of noisy data is displayed in Figure 14.1a for $r = 0.3$ and Figure 14.1b for $r = 0.6$. The fixed-point iteration (14.3.9) converged after 3 (4) iterations for $r = 0.3$ ($r = 0.6$), and yielded the values 4.33×10^{-3} (9.39×10^{-3}) for the regularization parameter α. The respective reconstructions u_α, shown in Figures 14.1c and 14.1d, are nearly indistinguishable from the true solution $u^\dagger$. To measure the accuracy of the solution u_α quantitatively, we compute the L^2-error $e = \|u_\alpha - u^\dagger\|_{L^2}$, which is 8.65×10^{-4} for $r = 0.3$ and 3.32×10^{-3} for $r = 0.6$. For comparison, we also show the solution by the L^2 data fitting problem (solved by a standard Newton method), where the parameter α has been chosen to give the smallest L^2 error. We observe that the L^2 reconstructions are clearly unacceptable compared to their L^1 counterparts, which illustrates the importance of a correct choice of the noise model, and especially the suitability of L^1 fitting for impulsive noise.

The performance of the balancing principle is further illustrated in Table 14.1 (cf. Appendix 14.B), where we compare the balancing parameter α_b with the "optimal", sampling-based parameter α_o for different noise levels. This parameter is obtained by sampling each interval $[0.1\alpha_b, \alpha_b]$ and $[\alpha_b, 10\alpha_b]$ uniformly with 50 parameters and taking as α_o the one with smallest L^2-error $e_o \equiv \|u_{\alpha_o} - u^\dagger\|_{L^2}$. We observe that both the regularization parameters and the reconstruction errors obtained from the two approaches are comparable. This shows the feasibility of the balancing principle for choosing an appropriate regularization parameter in nonlinear L^1 models. Table 14.1 also illustrates the fundamentally different nature of impulsive noise and L^1 fitting compared with Gaussian models, since the L^2-error does not depend linearly on the noise level or the percentage r of corrupted data. This can be attributed to the fact that the structural properties of the noise (e.g., clustering of corrupted data points, which is increasingly likely for $r \geqslant 0.5$) is more important than the noise percentage itself.

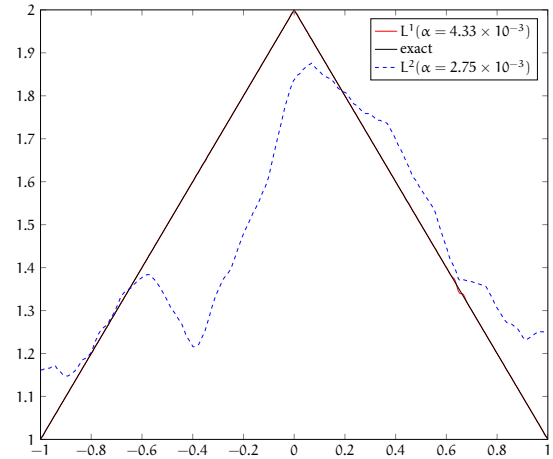
Next we study the convergence behavior of the path-following SSN method. First, the convergence behavior in the smoothing parameter β is illustrated in Table 14.2 by showing for each step in the continuation procedure the value of β, the required number of SSN iterations and



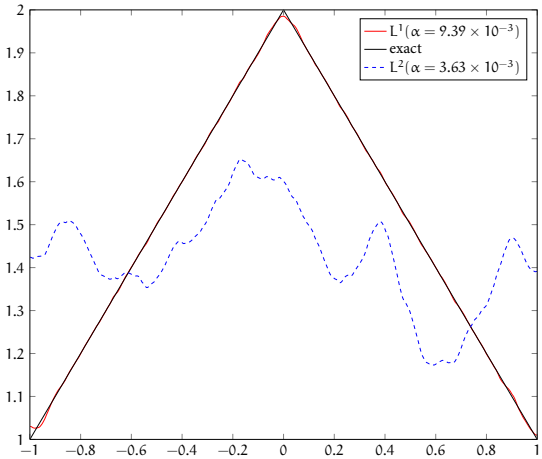
(a) exact and noisy data ($r = 0.3$)



(b) exact and noisy data ($r = 0.6$)



(c) comparison of L^1 and L^2 reconstruction

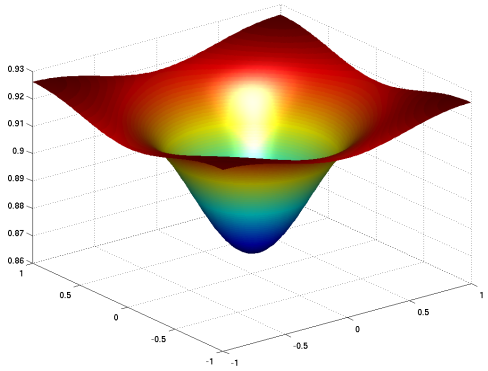


(d) comparison of L^1 and L^2 reconstruction

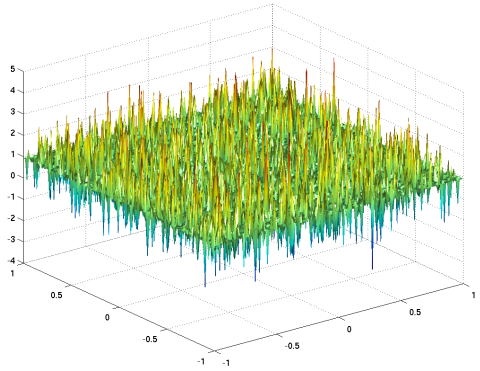
Figure 14.1: Results for 1d inverse potential problem. Left: $r = 0.3$, right: $r = 0.6$.

the L^2-error e. The required number of SSN iterations is relatively independent of the value of β provided it is sufficiently large. Then the semismoothness of the optimality system (OS_β) is gradually lost after the β value drops below 1.00×10^{-7}, and more and more iterations are required for the Krylov method to solve the Newton system (14.3.5) to the prescribed accuracy. Nonetheless, the reconstruction already represents a very reasonable approximation (in terms of the L^2-error e) at $\beta = 1.19 \times 10^{-7}$. Second, we illustrate the superlinear convergence of the SSN method by solving the optimality system (14.3.1) with fixed $r = 0.3$, $\alpha = 4.00 \times 10^{-3}$ and $\beta = 1.00 \times 10^{-1}$. Table 14.3 shows the number of elements that changed between active and inactive sets and the residual norm $\|F(u^k)\|_{L^2}$ after the kth iteration for several problem sizes N. The superlinear convergence as well as the mesh independence can be observed.

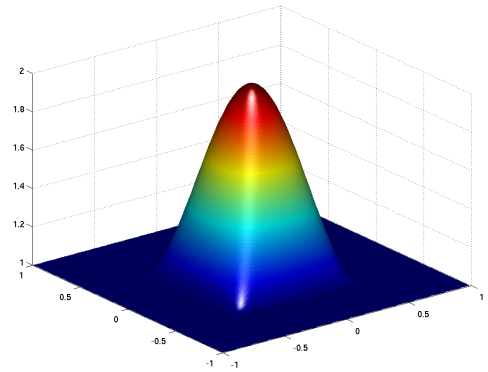
Finally, we demonstrate the scalability of the proposed approach. Table 14.4 summarizes the



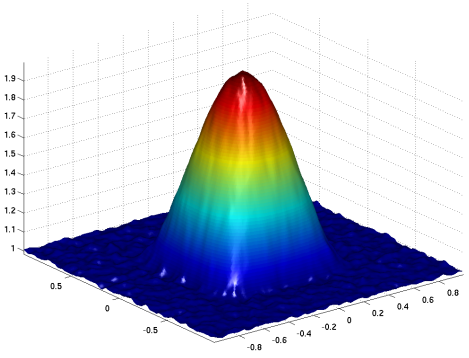
(a) exact data



(b) noisy data ($r = 0.3$)



(c) true solution



(d) L^1 reconstruction ($\alpha = 1.06 \times 10^{-2}$)

Figure 14.2: Results for the 2d inverse potential problem with $r = 0.3$ ($\delta = 2.24 \times 10^{-1}$).

computing time for one run of the path-following SSN method and for the full fixed point iteration. Since the computing time depends on the α value, we present the results with the final value of α as obtained from the fixed-point iteration (14.3.9). The presented results are the mean and standard deviation over ten noise realizations. We observe that both the fixed point iteration and the path-following SSN method scale very well with the problem size N, which corroborates the mesh independence of the SSN method [Hintermüller and Kunisch 2004]. We point out that the computational cost of calculating the balancing parameter is only two to three times that of solving the L^1 model with one fixed regularization parameter. Therefore, the balancing principle is also computationally inexpensive.

TWO-DIMENSIONAL EXAMPLE. Here, we take $\Omega = [-1, 1]^2$, $f(x_1, x_2) = 1$ and

$$u^\dagger(x_1, x_2) = 1 + \cos(\pi x_1)\cos(\pi x_2)\chi_{\{|(x_1,x_2)|_\infty < 1/2\}} \geqslant 1,$$

cf. Figure 14.2c. The exact and noisy data (with $r = 0.3$) are given in Figures 14.2a and 14.2b, respectively. The fixed point algorithm (14.3.9) converged within two iterations to the value $\alpha_b = 1.06 \times 10^{-2}$. The solution (with an L^2-error $e = 5.28 \times 10^{-3}$), shown in Figure 14.2d, accurately captures the shape as well as the magnitude of the potential $u^\dagger$ and thus represents a good approximation. The reconstruction by the L^2 model is again far from the true solution and thus is not shown here.

14.4.2 INVERSE ROBIN COEFFICIENT PROBLEM

This example, meant to illustrate coefficient recovery from boundary data, concerns reconstructing the Robin coefficient $u \in L^2(\Gamma_i)$ in (14.1.2) from noisy measurements of the Dirichlet trace of $y \in H^1(\Omega)$ on the boundary Γ_c. The discretization S_h of the forward operator S thus maps $u_h \in U_h = P_0(\Gamma_i)$ to the restriction of $y_h \in Y_h = P_1$ to the nodes on Γ_c, where y_h satisfies

$$\langle \nabla y_h, \nabla v_h \rangle_{L^2(\Omega)} + \langle u_h y_h, v_h \rangle_{L^2(\Gamma_i)} = \langle f, v_h \rangle_{L^2(\Gamma_c)} \qquad \text{for all } v_h \in Y_h.$$

Here, we take the domain $\Omega = [0,1]^2$, inaccessible boundary $\Gamma_i = \{(x_1, x_2) \in \partial\Omega : x_1 = 1\}$ and accessible (contact) boundary $\Gamma_c = \partial\Omega \setminus \Gamma_i$. Further, we set $f(x_1, x_2) = -4 + x_1$ and

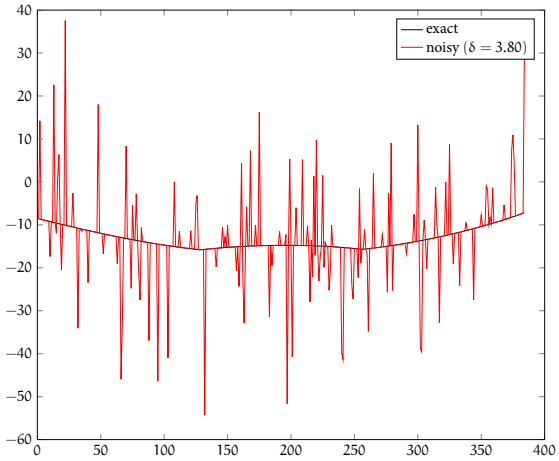
$$u^\dagger(x_2) = 1 + x_2 \geqslant 1.$$

For the automatic parameter choice using the balancing principle, we have set the weight σ to 1.03 and the initial guess α_0 to 1 as before.

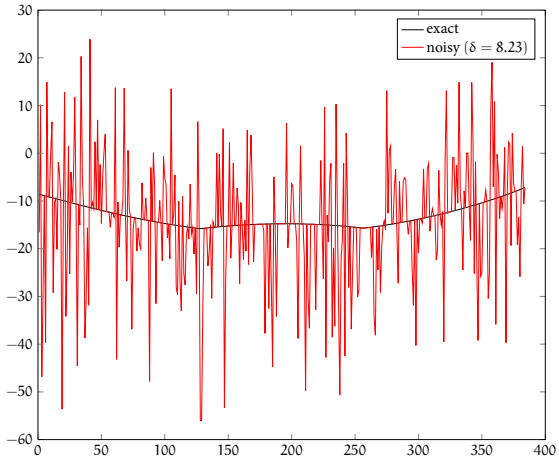
The noisy data for $r = 0.3$ and $r = 0.6$ are displayed in Figures 14.3a and 14.3b, respectively. The fixed point algorithm (14.3.9) converged after two iterations in both cases, giving a value 9.77×10^{-2} ($r = 0.3$) and 2.12×10^{-1} ($r = 0.6$) for the regularization parameter α. The corresponding reconstructions u_α, with respective L^2-error 3.13×10^{-3} and 1.05×10^{-2}, are shown in Figures 14.3c and 14.3d. Overall, the approximate solutions agree well with the true coefficient, except around the two end points, where the reconstructions suffer from pronounced boundary effect, especially in case of $r = 0.6$. Again, the reconstruction by the L^2 model (with optimal choice of α) is not acceptable and is thus not shown. A comparison of the balancing principle with the optimal choice based on sampling is given in Table 14.5. The results by these two approaches are very close to each other. From the table, we also observe the non-monotonicity of the error as a function of r, where the reconstruction error e shows a noticeable jump after $r = 0.5$.

14.4.3 INVERSE DIFFUSION COEFFICIENT PROBLEM

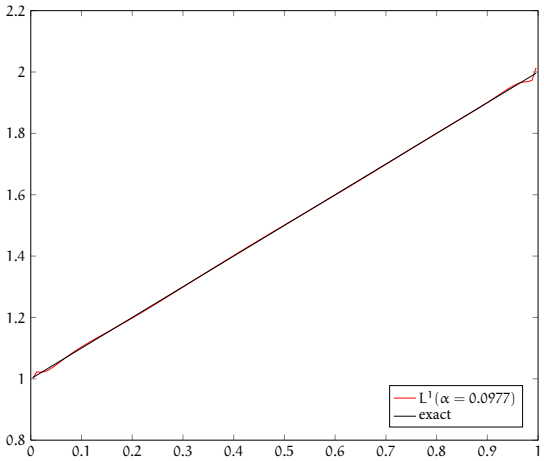
Finally, we consider the problem of determining the diffusion coefficient $u \in H^1(\Omega)$ in (14.1.3) from noisy measurements of the solution $y \in H^1_0(\Omega)$. Here we take $U_h = P_1$ and



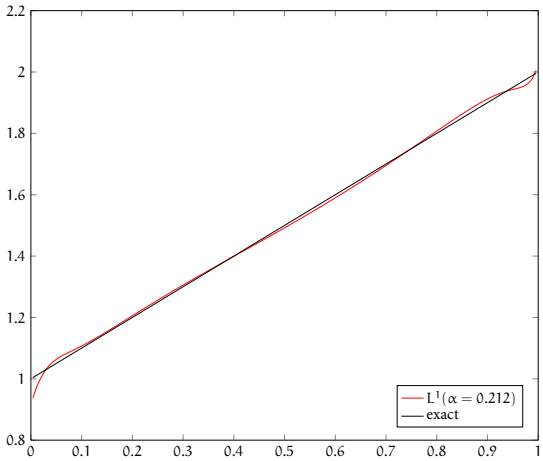
(a) exact and noisy data ($r = 0.3$)



(b) exact and noisy data ($r = 0.6$)



(c) L^1 reconstruction



(d) L^1 reconstruction

Figure 14.3: Results for the inverse Robin coefficient problem. Left: $r = 0.3$, right: $r = 0.6$.

$Y_h = P_1 \cap H_0^1(\Omega)$ and consider the discrete operator S_h as mapping $u_h \in U_h$ to $y_h \in Y_h$ satisfying

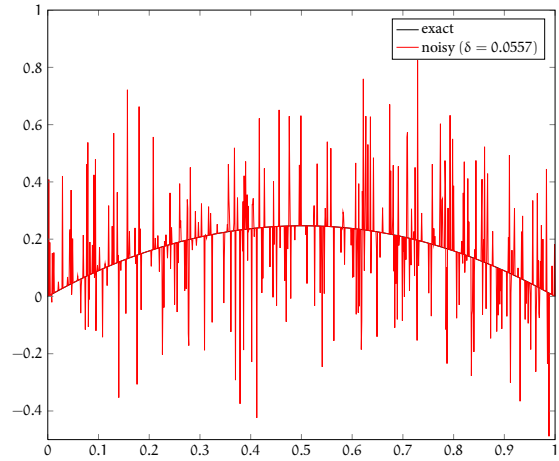
$$\langle u_h \nabla y_h, \nabla v_h \rangle_{L^2(\Omega)} = \langle f, v_h \rangle_{L^2(\Omega)} \quad \text{for all } v_h \in Y_h.$$

To accelerate the convergence of the Krylov solver, we precondition the Newton system with the inverse Helmholtz operator $(-\Delta + I)^{-1}$, i.e., the gradient $\alpha(-\Delta u + u) - \nabla y \cdot \nabla p$ is replaced by

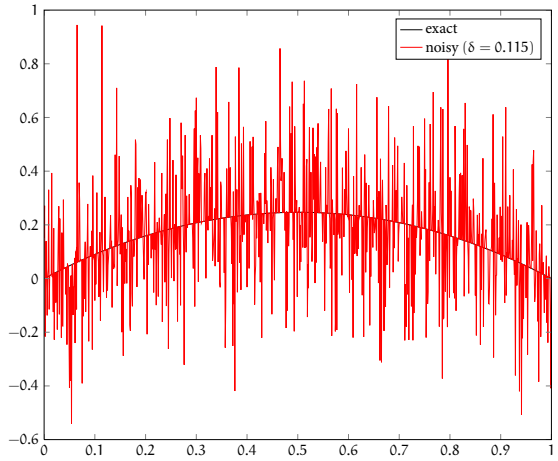
$$\alpha u - (-\Delta + I)^{-1}(\nabla y \cdot \nabla p),$$

and similarly the action of the Hessian on δu is computed as

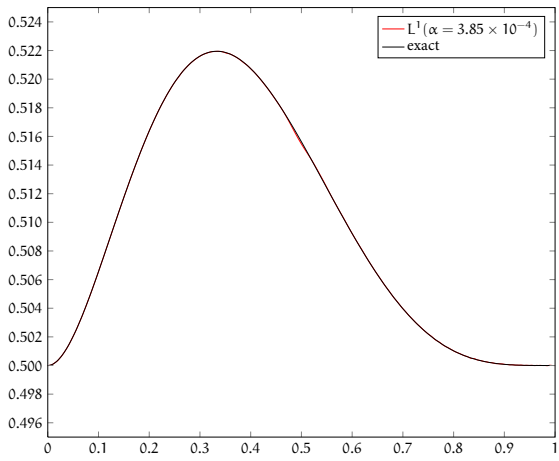
$$\alpha u - (-\Delta + I)^{-1}(\nabla \delta y \cdot \nabla p + \nabla y \cdot \nabla \delta p).$$



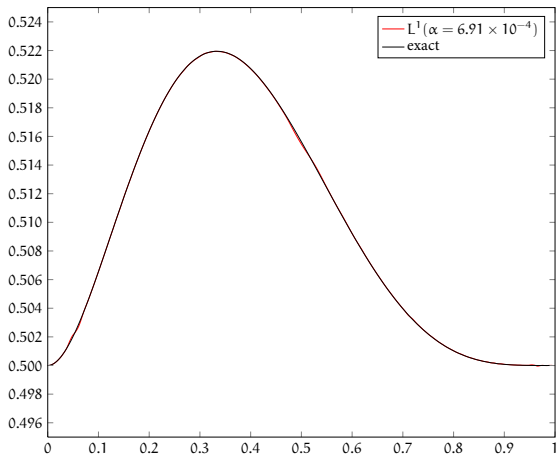
(a) exact and noisy data ($r = 0.3$)



(b) exact and noisy data ($r = 0.6$)



(c) L^1 reconstruction



(d) L^1 reconstruction

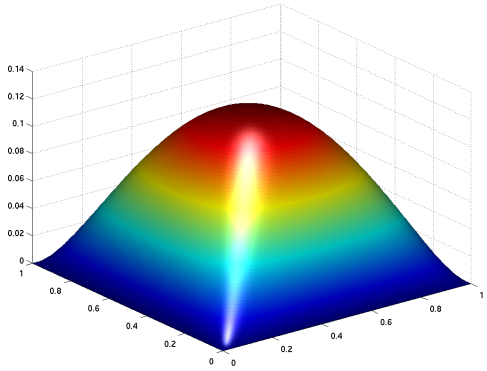
Figure 14.4: Results for the 1d inverse diffusion coefficient problem. Left: $r = 0.3$, right: $r = 0.6$.

For the automatic parameter choice using the balancing principle, we have set the weight σ to 1.001 and the initial guess α_0 to 0.1. As noted, the different weight is chosen according to the stronger smoothness assumption on u (H^1 instead of L^2 regularization).

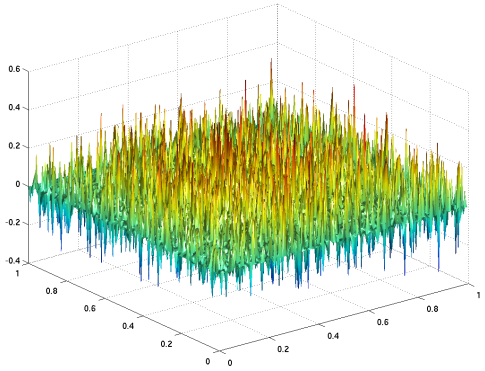
ONE-DIMENSIONAL EXAMPLE. Here, we take the domain $\Omega = [0, 1]$ and $f(x) = 1$. The exact solution $u^\dagger$ is given by

$$u^\dagger(x) = \frac{1}{2} + x^2(1-x)^4 \geqslant \frac{1}{2}.$$

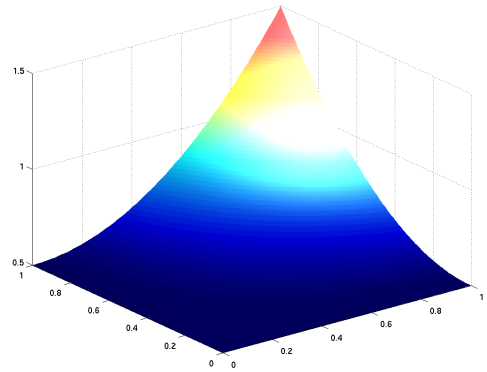
Noisy data with $r = 0.3$ and $r = 0.6$ and the reconstructions ($\alpha = 3.85 \times 10^{-4}$, L^2-error 2.77×10^{-5} and $\alpha = 6.90 \times 10^{-4}$, L^2-error 3.86×10^{-5}) are shown in Figure 14.4. In both



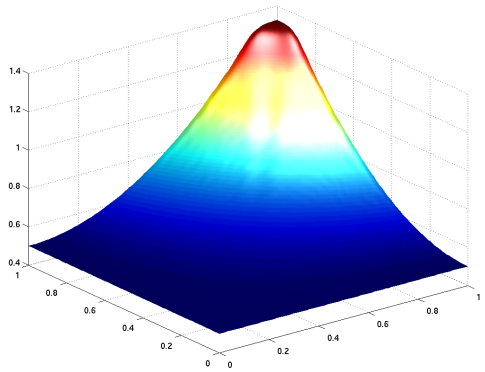
(a) exact data



(b) noisy data ($r = 0.3$)



(c) true solution



(d) L^1 reconstruction ($\alpha = 5.14 \times 10^{-5}$)

Figure 14.5: Results for the 2d inverse diffusion coefficient problem with $r = 0.3$ ($\delta = 3.06 \times 10^{-2}$).

cases, the fixed point iteration (14.3.9) converged within two iterations. The convergence of the path-following method and of the SSN method is similar to the inverse potential problem. A comparison of the balancing principle with the optimal choice based on sampling is given in Table 14.6. The results by these two approaches are very close to each other. From the table, we also observe the non-monotonicity of the error as a function of r, where the reconstruction error e remains almost constant for $r \leqslant 0.5$ and then increases quickly. Again the proposed SSN method scales very well with the problem size, as shown in Table 14.7.

TWO-DIMENSIONAL EXAMPLE. Here, we take $\Omega = [0,1]^2$, $f(x_1, x_2) = 1$ and

$$u^\dagger(x_1, x_2) = \frac{1}{2} + x_1^2 x_2^2 \geqslant \frac{1}{2};$$

cf. Figure 14.5c. The exact and noisy data ($r = 0.3$) are given in Figures 14.5a and 14.5b, respectively. The fixed point algorithm converged in seven iterations to the value $\alpha = 5.14 \times 10^{-5}$. The reconstruction, shown in Figure 14.5d, agrees well with the true solution (the L^2-error being 5.63×10^{-3}). The less accurate approximation around the corner might be attributed to the fact that the true solution does not satisfy the homogeneous Neumann conditions imposed by the Newton step. Again, we remark that the L^2 reconstruction (not presented) is far from the true solution.

14.5 CONCLUSION

In this paper we have presented a path-following semismooth Newton method for the efficient numerical solution of nonlinear parameter identification problems with impulsive noise. The method is based on a Huber-type smoothing of the L^1 fitting functional, and its superlinear convergence is proved and demonstrated numerically. Furthermore, mesh independence of the method can be observed. Several model examples for elliptic differential equations illustrate the efficiency of this approach.

The balancing principle is shown to be an effective parameter choice method, which required little a priori information such as the noise level, while adding only a small amount of computational overhead over the solution of one single minimization problem.

The presented approach can be extended in several directions. As noted in Remark 14.2.6, including constraints on the solution would be a natural progression. The extension to time-dependent problems would be straightforward but poses interesting challenges for efficient implementation. Finally, it would be worthwhile to consider mixed Gaussian and impulsive noise. While such noise is challenging for either L^1 or L^2 fitting, our robust approximation ($\mathcal{P}_\beta$) seems to be an appropriate model [Huber 1981] (cf. also Remark 14.3.1). Then the continuation $\beta \to 0$ would have to be replaced by a suitable parameter choice method for determining the optimal stopping value $\beta^* > 0$. Such an approach might also be applicable to other non-Gaussian models like Laplace and Cauchy noise.

ACKNOWLEDGMENTS

The authors are grateful to the referees for their constructive comments which have led to an improved presentation. The work of the first author was supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences"), and that of the second author was supported by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

14.A VERIFICATION OF PROPERTIES FOR MODEL PROBLEMS

For completeness, we collect in this section some results which verify the continuity and differentiability properties (A1)–(A4) for our model problems. Throughout, we shall denote by C a generic constant, which is independent of $u \in U$.

14.A.1 ELLIPTIC POTENTIAL PROBLEM

For this model problem, S maps $u \in X = L^2(\Omega)$ to the solution $y \in Y = H^1(\Omega)$ of (14.1.1), and we take $U = \{u \in L^\infty(\Omega) : u \geqslant c\}$ for some fixed $c > 0$. The verification of properties (A1)–(A4) is analogous to [Kröner and Vexler 2009]. We therefore only give, for the sake of completeness, the explicit form of the derivatives required for the solution of the Newton system (14.3.5) using a Krylov subspace method.

For given $u \in L^2(\Omega)$, $F(u)$ is computed by the following steps:

1: Solve for $y \in H^1(\Omega)$ in

$$\langle \nabla y, \nabla v\rangle_{L^2(\Omega)} + \langle uy, v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H^1(\Omega).$$

2: Solve for $p \in H^1(\Omega)$ in

$$\langle \nabla p, \nabla v\rangle_{L^2(\Omega)} + \langle up, v\rangle_{L^2(\Omega)} = -\langle \operatorname{sign}_\beta(y - y^\delta), v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H^1(\Omega).$$

3: Set $F(u) = \alpha u + yp$.

For given $\delta u \in L^2(\Omega)$, the application of $D_N F(u)$ on δu is computed by:

1: Solve for $\delta y \in H^1(\Omega)$ in

$$\langle \nabla \delta y, \nabla v\rangle_{L^2(\Omega)} + \langle u\delta y, v\rangle_{L^2(\Omega)} = -\langle y\delta u, v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H^1(\Omega).$$

2: Solve for $\delta p \in H^1(\Omega)$ in

$$\langle \nabla \delta p, \nabla v\rangle_{L^2(\Omega)} + \langle u\delta p, v\rangle_{L^2(\Omega)} = -\langle \tfrac{1}{\beta}\chi_\mathcal{I}\delta y + p\delta u, v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H^1(\Omega).$$

3: Set $D_N F(u) = \alpha\delta u + p\delta y + y\delta p$.

14.A.2 ROBIN COEFFICIENT PROBLEM

Here, S maps $u \in X = L^2(\Gamma_i)$ to $y|_{\Gamma_c} \in Y = H^{\frac{1}{2}}(\Gamma_c)$, where y is the solution to (14.1.2). Set $U = \{u \in L^\infty(\Gamma_i) : u \geqslant c\}$ for some fixed $c > 0$. We shall denote the mapping of $u \in U$ to the solution $y \in H^1(\Omega)$ of (14.1.2) by $y(u)$. The following a priori estimate follows directly from the Lax–Milgram theorem.

Lemma 14.A.1. *For any* $u \in U$, *problem* (14.1.2) *has a unique solution* $y \in H^1(\Omega)$ *which satisfies*

$$\|y\|_{H^1(\Omega)} \leqslant C\|f\|_{H^{-\frac{1}{2}}(\Gamma_c)}.$$

Since $f \in H^{-\frac{1}{2}}(\Gamma_c)$ is fixed, the uniform boundedness of S follows from the continuity of the trace operator. We next address the complete continuity of S.

Lemma 14.A.2. *Let* $\{u_n\} \subset U$ *be a sequence converging weakly in* $L^2(\Gamma_i)$ *to* $u^* \in U$. *Then*

$$S(u_n) \to S(u^*) \quad \text{in } L^2(\Gamma_c).$$

Proof. For $u_n \in U$, set $y_n = y(u_n) \in H^1(\Omega)$. By the a priori estimate from Lemma 14.A.1, the sequence $\{y_n\}$ is uniformly bounded in $H^1(\Omega)$ and has a convergent subsequence, also denoted by $\{y_n\}$, such that there exists $y^* \in H^1(\Omega)$ with

$$y_n \rightharpoonup y^* \text{ in } H^1(\Omega).$$

The trace theorem and the Sobolev embedding theorem [Adams and Fournier 2003] imply

$$y_n \to y^* \text{ in } L^p(\Gamma_c)$$

for any $p < +\infty$. In particular, we will take $p = 4$. Then we have

$$|\langle u_n(y_n - y^*), v\rangle_{L^2(\Gamma_i)}| \leqslant \|u_n\|_{L^2(\Gamma_i)}\|y_n - y^*\|_{L^4(\Gamma_i)}\|v\|_{L^4(\Gamma_i)} \to 0$$

by the weak convergence of $\{u^n\}$ in $L^2(\Gamma_i)$ and the strong convergence of $\{y_n\}$ in $L^4(\Gamma_i)$. Therefore, we have

$$\lim_{n\to\infty} \langle u_n y_n, v\rangle_{L^2(\Gamma_i)} = \lim_{n\to\infty} \left(\langle u_n(y_n - y^*), v\rangle_{L^2(\Gamma_i)} + \langle u_n y^*, v\rangle_{L^2(\Gamma_i)} \right) = \langle u^* y^*, v\rangle_{L^2(\Gamma_i)}.$$

Now passing to the limit in the weak formulation indicates that y^* satisfies

$$\langle \nabla y^*, \nabla v\rangle_{L^2(\Omega)} + \langle u^* y^*, v\rangle_{L^2(\Gamma_i)} = \langle f, v\rangle_{L^2(\Gamma_c)} \quad \text{for all } v \in H^1(\Omega),$$

i.e., $y^* = y(u^*)$. Since every subsequence has itself a subsequence converging weakly in $H^1(\Omega)$ to $y(u^*)$, the whole sequence converges weakly. The continuity of $S : u \mapsto y(u)|_{\Gamma_c}$ then follows from the trace theorem and the Sobolev embedding theorem for $p = 2$. □

The above two statements imply that property (A1) holds. We next address the remaining properties.

Lemma 14.A.3. *The mapping* $u \mapsto y(u)$ *is twice Fréchet differentiable from* U *to* $H^1(\Omega)$, *and for every* $u \in U$ *and all directions* $h_1, h_2 \in L^2(\Gamma_i)$, *the derivatives are given by*

(i) $y'(u)h_1 \in H^1(\Omega)$ *is the solution* z *of*

$$\langle \nabla z, \nabla v\rangle_{L^2(\Omega)} + \langle uz, v\rangle_{L^2(\Gamma_i)} = -\langle h_1 y(u), v\rangle_{L^2(\Gamma_i)} \qquad \textit{for all } v \in H^1(\Omega),$$

and the following estimate holds

$$\|y'(u)h_1\|_{H^1(\Omega)} \leqslant C\|h_1\|_{L^2(\Gamma_i)}.$$

(ii) $y''(u)(h_1, h_2) \in H^1(\Omega)$ *is the solution* z *of*

$$\langle \nabla z, \nabla v\rangle_{L^2(\Omega)} + \langle uz, v\rangle_{L^2(\Gamma_i)} = -\langle h_1 y'(u)h_2, v\rangle_{L^2(\Gamma_i)} - \langle h_2 y'(u)h_1, v\rangle_{L^2(\Gamma_i)}$$

for all $v \in H^1(\Omega)$*, and the following estimate holds*

$$\|y''(u)(h_1, h_2)\|_{H^1(\Omega)} \leqslant C\|h_1\|_{L^2(\Gamma_i)}\|h_2\|_{L^2(\Gamma_i)}.$$

Proof. The characterization of the derivatives follows from direct calculation. It remains to show boundedness and continuity. By setting $v = y'(u)h_1$ in the weak formulation, Hölder's inequality, the trace theorem and the a priori estimate in Lemma 14.A.1, we have

$$\begin{aligned}\|y'(u)h_1\|^2_{H^1(\Omega)} &\leqslant C\,\|y'(u)h_1\|_{L^4(\Gamma_i)}\,\|h_1\|_{L^2(\Gamma_i)}\,\|y(u)\|_{L^4(\Gamma_i)}\\ &\leqslant C\,\|y'(u)h_1\|_{H^1(\Omega)}\,\|h_1\|_{L^2(\Gamma_i)}\,\|y(u)\|_{H^1(\Omega)}\\ &\leqslant C\,\|y'(u)h_1\|_{H^1(\Omega)}\,\|h_1\|_{L^2(\Gamma_i)}\,,\end{aligned}$$

from which the first estimate follows. Analogously we deduce that

$$\|y(u+h_1) - y(u)\|_{H^1(\Omega)} \leqslant C\,\|h_1\|_{L^2(\Gamma_i)}\,.$$

Next let $w = y(u+h_1) - y(u) - y'(u)h_1$, which satisfies

$$\langle \nabla w, \nabla v\rangle_{L^2(\Omega)} + \langle uw, v\rangle_{L^2(\Gamma_i)} = -\langle h_1(y(u+h_1) - y(u)), v\rangle_{L^2(\Gamma_i)}$$

for all $v \in H^1(\Omega)$. By repeating the proof of the preceding estimate, we deduce that

$$\|w\|_{H^1(\Omega)} \leqslant C\,\|h_1\|_{L^2(\Gamma_i)}\,\|y(u+h_1) - y(u)\|_{H^1(\Omega)}\,,$$

from which it follows directly that $y'(u)h_1$ defined above is indeed the Fréchet derivative of $y(u)$ at u. By arguing similarly and using the first assertion, the second assertion follows. □

Together with the linearity of the trace operator, we obtain $S'(u)h_1 = y'(u)h_1|_{\Gamma_c} \in H^{\frac{1}{2}}(\Gamma_c)$ and $S''(u)(h_1, h_2) = y''(u)(h_1, h_2)|_{\Gamma_c} \in H^{\frac{1}{2}}(\Gamma_c)$ and thus property (A2). Finally, properties (A3) and (A4) follow directly from the estimates in Lemma 14.A.3 and the trace theorem [Adams and Fournier 2003].

We again give the necessary steps in a Krylov subspace method for the solution to (14.3.5). For given $u \in L^2(\Gamma_i)$, $F(u)$ is computed by the following steps:

1: Solve for $y \in H^1(\Omega)$ in

$$\langle \nabla y, \nabla v \rangle_{L^2(\Omega)} + \langle uy, v \rangle_{L^2(\Gamma_i)} = \langle f, v \rangle_{L^2(\Gamma_c)} \qquad \text{for all } v \in H^1(\Omega).$$

2: Solve for $p \in H^1(\Omega)$ in

$$\langle \nabla p, \nabla v \rangle_{L^2(\Omega)} + \langle up, v \rangle_{L^2(\Gamma_i)} = -\langle \operatorname{sign}_\beta(y|_{\Gamma_c} - y^\delta), v \rangle_{L^2(\Gamma_c)} \qquad \text{for all } v \in H^1(\Omega).$$

3: Set $F(u) = \alpha u + y|_{\Gamma_i} p|_{\Gamma_i}$.

For given $\delta u \in L^2(\Gamma_i)$, the application of $D_N F(u)$ on δu is computed by:

1: Solve for $\delta y \in H^1(\Omega)$ in

$$\langle \nabla \delta y, \nabla v \rangle_{L^2(\Omega)} + \langle u \delta y, v \rangle_{L^2(\Gamma_i)} = -\langle y \delta u, v \rangle_{L^2(\Gamma_i)} \qquad \text{for all } v \in H^1(\Omega).$$

2: Solve for $\delta p \in H^1(\Omega)$ in

$$\langle \nabla \delta p, \nabla v \rangle_{L^2(\Omega)} + \langle up, v \rangle_{L^2(\Gamma_i)} = -\langle \tfrac{1}{\beta} \chi_{\mathcal{J}}(\delta y|_{\Gamma_c}), v \rangle_{L^2(\Gamma_c)} - \langle p \delta u, v \rangle_{L^2(\Gamma_i)}$$

for all $v \in H^1(\Omega)$.

3: Set $D_N F(u) = \alpha \delta u + p|_{\Gamma_i}(\delta y)|_{\Gamma_i} + y|_{\Gamma_i}(\delta p)|_{\Gamma_i}$.

14.A.3 DIFFUSION COEFFICIENT PROBLEM

In this model problem, the operator S maps $u \in \mathcal{X} = H^1(\Omega)$ to the solution $y \in \mathcal{Y} = W_0^{1,q}(\Omega)$, for some $q > 2$, of (14.1.3), and the admissible set is $U = \{u \in H^1(\Omega) : \lambda \leqslant u \leqslant \lambda^{-1}\}$ for some fixed $\lambda \in (0,1)$. The following estimate is an immediate consequence of Theorem 1 in [Meyers 1963], where $Q > 2$ is a constant depending only on λ and Ω. We shall assume $f \in L^q(\Omega)$ for some $q > Q$.

Lemma 14.A.4. *There exists a number $Q > 2$ depending only on λ and Ω such that for any $u \in U$ and $q \in (2, Q)$, problem (14.1.3) has a unique solution $y \in W_0^{1,q}(\Omega)$ which satisfies*

$$\|y\|_{W^{1,q}(\Omega)} \leqslant C \|f\|_{L^q(\Omega)}.$$

From this, the uniform boundedness of S follows since $f \in L^q(\Omega)$ is fixed. We next address the complete continuity of S.

Lemma 14.A.5. *Let $\{u_n\} \subset U$ be a sequence converging weakly in $H^1(\Omega)$ to $u^* \in U$. Then*

$$S(u_n) \to S(u^*) \quad \text{in } L^2(\Omega).$$

Proof. For $u_n \in U$, set $y_n = S(u_n) \in W_0^{1,q}(\Omega)$. By the a priori estimate from Lemma 14.A.4, the sequence $\{y_n\}$ is uniformly bounded in $W^{1,q}(\Omega)$ and has a convergent subsequence also denoted by $\{y_n\}$ such that there exists $y^* \in W_0^{1,q}(\Omega)$ with

$$y_n \rightharpoonup y^* \text{ in } W^{1,q}(\Omega).$$

The Rellich–Kondrachov embedding theorem [Adams and Fournier 2003, Thm. 6.3] implies

$$u_n \to u^* \text{ in } L^p(\Omega)$$

for any $p < +\infty$. In particular, we will take p such that $\frac{1}{2} + \frac{1}{p} + \frac{1}{q} = 1$. Then we have

$$\left|\langle (u_n - u^*)\nabla y_n, \nabla v\rangle_{L^2(\Omega)}\right| \leqslant \|u_n - u^*\|_{L^p(\Omega)} \|\nabla y_n\|_{L^q(\Omega)} \|\nabla v\|_{L^2(\Omega)} \to 0$$

by the weak convergence of $\{y_n\}$ in $W^{1,q}(\Omega)$ and the strong convergence of $\{u_n\}$ in $L^p(\Omega)$. Therefore, we have

$$\begin{aligned}\lim_{n\to\infty} \langle u_n \nabla y_n, \nabla v\rangle_{L^2(\Omega)} &= \lim_{n\to\infty} \left(\langle (u_n - u^*)\nabla y_n, \nabla v\rangle_{L^2(\Omega)} + \langle u^* \nabla y_n, \nabla v\rangle_{L^2(\Omega)} \right) \\ &= \langle u^* \nabla y^*, \nabla v\rangle_{L^2(\Omega)}.\end{aligned}$$

Now passing to the limit in the weak formulation indicates that y^* satisfies

$$\langle u^* \nabla y^*, \nabla v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega),$$

i.e., $y^* = S(u^*)$. Since every subsequence has itself a subsequence converging weakly in $W^{1,q}(\Omega)$ to $S(u^*)$, the whole sequence converges weakly. Applying again the Rellich–Kondrachov embedding theorem [Adams and Fournier 2003] for $p = 2$ completes the proof of the lemma. □

The above two statements imply that property (A1) holds. The next statement yields the remaining properties (A2), (A3) and (A4).

Lemma 14.A.6. *The operator $S : U \to W_0^{1,q}(\Omega)$ is twice Fréchet differentiable, and for every $u \in U$ and all admissible directions $h_1, h_2 \in H^1(\Omega)$, the derivatives are given by*

(i) $S'(u)h_1 \in W_0^{1,q}(\Omega)$ *is the solution z of*

$$\langle u\nabla z, \nabla v\rangle_{L^2(\Omega)} = -\langle h_1 \nabla S(u), \nabla v\rangle_{L^2(\Omega)} \qquad \textit{for all } v \in H_0^1(\Omega),$$

and the following estimate holds

$$\|S'(u)h_1\|_{W^{1,q}(\Omega)} \leqslant C \|h_1\|_{H^1(\Omega)}.$$

(ii) $S''(u)(h_1, h_2) \in W_0^{1,q}(\Omega)$ *is the solution z of*

$$\langle u\nabla z, \nabla v\rangle_{L^2(\Omega)} = -\langle h_1 \nabla S'(u)h_2 + h_2\nabla S'(u)h_1, \nabla v\rangle_{L^2(\Omega)} \qquad \textit{for all } v \in H_0^1(\Omega),$$

and the following estimate holds

$$\|S''(u)(h_1,h_2)\|_{W^{1,q}(\Omega)} \leqslant C\,\|h_1\|_{H^1(\Omega)}\,\|h_2\|_{H^1(\Omega)}.$$

Proof. Again, the characterization of the derivatives are obtained by direct calculation. Set $y = S(u) \in W_0^{1,q}(\Omega)$. By Lemma 14.A.4 and Hölder's inequality, we get

$$\begin{aligned}\|S'(u)h_1\|_{W^{1,q}(\Omega)} &\leqslant C\|h_1\nabla y\|_{L^q(\Omega)} \leqslant \|h_1\|_{L^p(\Omega)}\|\nabla y\|_{L^{q'}(\Omega)}\\ &\leqslant C\|h_1\|_{H^1(\Omega)}\|\nabla y\|_{L^{q'}(\Omega)} \leqslant C\|h_1\|_{H^1(\Omega)}\end{aligned}$$

with $q' \in (q, Q)$ and $\frac{1}{q} = \frac{1}{p} + \frac{1}{q'}$, where we have used the Sobolev embedding theorem and the estimate in Lemma 14.A.4. Analogously, we deduce that

$$\|S(u+h_1) - S(u)\|_{W^{1,\tilde q}(\Omega)} \leqslant C\|h_1\|_{H^1(\Omega)},$$

where the exponent $\tilde q$ satisfies $\tilde q \in (q, Q)$. Next let $w = S(u+h_1) - S(u) - S'(u)h_1$, which satisfies

$$\langle u\nabla w, \nabla v\rangle_{L^2(\Omega)} = -\langle h_1\nabla(S(u+h_1) - S(u)), \nabla v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega).$$

Repeating the proof of the preceding estimate, we derive

$$\|w\|_{W^{1,p}(\Omega)} \leqslant C\|h_1\|_{H^1(\Omega)}\|S(u+h_1) - S(u)\|_{W^{1,\tilde q}(\Omega)}.$$

Combining these estimates yields the first assertion, i.e., $S'(u)h_1$ defined above is indeed the Fréchet derivative of the forward operator $S(u) : H^1(\Omega) \to W_0^{1,p}(\Omega)$, and it satisfies the desired estimate. Similarly, the second assertion follows from Lemma 14.A.4 and the first assertion. □

We finally address the steps required in a Krylov subspace method for the solution to (14.3.5). For given $u \in H^1(\Omega)$, $F(u)$ is computed by the following steps:

1: Solve for $y \in H_0^1(\Omega)$ in

$$\langle u\nabla y, \nabla v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega).$$

2: Solve for $p \in H_0^1(\Omega)$ in

$$\langle u\nabla p, \nabla v\rangle_{L^2(\Omega)} = \left\langle \operatorname{sign}_\beta(y - y^\delta), v\right\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega).$$

3: Set $F(u) = \alpha(-\Delta u + u) - \nabla y \cdot \nabla p$.

For given $\delta u \in H^1(\Omega)$, the application of $D_N F(u)$ on δu is computed by:

1: Solve for $\delta y \in H_0^1(\Omega)$ in

$$\langle u\nabla\delta y, \nabla v\rangle_{L^2(\Omega)} = \langle \delta u\nabla y, \nabla v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega).$$

2: Solve for $\delta p \in H_0^1(\Omega)$ in

$$\langle u\nabla\delta p, \nabla v\rangle_{L^2(\Omega)} = -\left\langle \tfrac{1}{\beta}\chi_{\mathcal{J}}\delta y, v\right\rangle_{L^2(\Omega)} + \langle \delta u\nabla p, \nabla v\rangle_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega).$$

3: Set $D_N F(u) = \alpha(-\Delta\delta u + \delta u) - \nabla\delta y \cdot \nabla p - \nabla y \cdot \nabla\delta p$.

14.B TABLES

r	δ	α_o	α_b	e_o	e_b
0.1	6.46e-2	4.65e-3	1.66e-3	2.64e-4	5.90e-4
0.2	1.13e-1	5.53e-3	2.91e-3	3.49e-4	4.11e-4
0.3	1.68e-1	4.17e-3	4.32e-3	4.60e-4	5.06e-4
0.4	2.26e-1	3.82e-3	5.81e-3	6.26e-4	8.97e-4
0.5	2.94e-1	3.34e-3	7.56e-3	3.54e-3	5.40e-3
0.6	3.19e-1	6.72e-3	8.20e-3	1.31e-2	2.05e-2
0.7	3.86e-1	1.35e-2	9.93e-3	8.47e-3	8.95e-3
0.8	4.38e-1	6.40e-3	1.13e-2	9.27e-3	1.89e-2
0.9	4.85e-1	1.51e-2	1.28e-2	2.04e-1	2.04e-1

Table 14.1: Comparison of the balancing principle (α_b, e_b) with the sampling-based optimal choice (α_o, e_o) for the 1d inverse potential problem.

Table 14.2: Convergence of path-following method. For each step k, the parameter $\beta(k)$, number $it(k)$ of SSN iterations and L^2-error $e(k)$ are shown.

$\beta(k)$	1.00e0	5.00e-1	2.50e-1	1.25e-1	6.25e-2	3.12e-2	1.56e-2	7.81e-3	3.91e-3	1.95e-3
$it(k)$	6	4	4	3	3	3	3	3	3	3
$e(k)$	1.83e-1	1.33e-1	1.08e-1	8.77e-2	7.17e-2	5.99e-2	4.63e-2	3.43e-2	2.63e-2	2.11e-2
$\beta(k)$	9.77e-4	4.88e-4	2.44e-4	1.22e-4	6.10e-5	3.05e-5	1.53e-5	7.63e-6	3.81e-6	1.91e-6
$it(k)$	3	3	3	3	3	3	3	3	3	4
$e(k)$	1.70e-2	1.33e-2	1.03e-2	8.06e-3	6.29e-3	4.90e-3	3.80e-3	2.94e-3	2.31e-3	1.85e-3
$\beta(k)$	9.54e-7	4.77e-7	2.38e-7	1.19e-7	5.96e-8	2.98e-8	1.49e-8	7.45e-9	3.73e-9	1.86e-9
$it(k)$	4	5	5	9	14	20	20	20	20	20
$e(k)$	1.51e-3	1.28e-3	1.12e-3	1.00e-3	9.32e-4	9.08e-4	8.90e-4	8.66e-4	8.65e-4	8.66e-4

Table 14.3: Convergence behavior of the SSN method (for fixed α, β) for the 1d inverse potential problem. Shown are the problem size N, the number $n(k)$ of elements that changed between active and inactive sets and residual norm $r(k) \equiv \|F(u)\|_{L^2}$ after each iteration k.

N	k	1	2	3	4	5
101	$n(k)$	88	0	0	0	0
	$r(k)$	4.40e-2	1.51e-2	8.62e-4	6.58e-6	2.86e-10
1001	$n(k)$	791	6	5	0	0
	$r(k)$	1.23e-1	1.78e-2	2.30e-3	3.53e-5	9.99e-9
10001	$n(k)$	7803	91	16	1	0
	$r(k)$	1.21e-1	1.67e-2	1.68e-3	1.90e-5	2.47e-9

Table 14.4: Computing times (in seconds) for SSN method (t_s) and fixed-point iteration (t_b) and L^2-error e for the 1d inverse potential problem. Shown are the problem size N, the mean ($\{t_s, t_b, e\}_m$) and standard deviation ($\{t_s, t_b, e\}_s$) over ten noise realizations.

N	100	200	400	800	1600	3200	6400	12800
$t_{s,m}$	1.25	1.75	5.28	12.09	19.40	29.66	55.33	107.87
$t_{s,s}$	0.48	0.45	3.31	4.57	7.22	4.45	11.44	25.92
$t_{b,m}$	7.12	9.63	14.42	39.04	54.19	80.30	131.72	234.00
$t_{b,s}$	3.42	6.21	7.63	17.14	16.79	17.25	38.08	75.21
e_m	8.98e-1	1.51e+0	2.88e-3	9.17e-4	6.22e-4	3.52e-4	2.76e-4	2.78e-4
e_s	2.46e+0	3.16e+0	2.05e-3	5.76e-4	6.83e-4	9.36e-5	4.29e-5	6.93e-5

r	δ	α_o	α_b	e_o	e_b
0.1	1.31e+0	1.40e-1	3.38e-2	1.15e-5	4.10e-5
0.2	2.19e+0	1.07e-1	5.62e-2	4.03e-6	1.46e-5
0.3	3.41e+0	2.45e-1	8.76e-2	8.63e-4	1.22e-3
0.4	5.30e+0	6.27e-1	1.36e-1	2.64e-3	5.40e-3
0.5	6.00e+0	5.01e-1	1.54e-1	4.10e-4	1.53e-3
0.6	7.31e+0	4.41e-1	1.88e-1	3.72e-2	6.30e-2
0.7	9.13e+0	3.40e-1	2.35e-1	6.07e-3	6.36e-3
0.8	9.79e+0	2.53e-1	2.53e-1	6.59e-2	6.59e-2
0.9	1.16e+1	6.00e-1	3.16e-1	3.02e-1	3.37e-1

Table 14.5: Comparison of the balancing principle (α_b, e_b) with the sampling-based optimal choice (α_o, e_o) for the inverse Robin coefficient problem.

r	δ	α_o	α_b	e_o	e_b
0.1	2.08e-2	9.20e-5	1.52e-4	2.34e-5	2.71e-5
0.2	3.81e-2	1.43e-4	2.77e-4	2.17e-5	2.28e-5
0.3	5.82e-2	3.41e-4	3.99e-4	2.68e-5	3.98e-5
0.4	8.54e-2	2.11e-4	5.71e-4	2.59e-5	3.84e-5
0.5	9.45e-2	3.83e-4	5.98e-4	3.56e-5	4.28e-5
0.6	1.22e-1	1.28e-3	7.45e-4	2.82e-4	3.60e-4
0.7	1.42e-1	2.04e-3	8.35e-4	8.01e-4	1.31e-3
0.8	1.55e-1	1.66e-3	8.71e-4	4.82e-4	6.52e-4
0.9	1.80e-1	4.33e-3	9.42e-4	2.11e-3	6.49e-3

Table 14.6: Comparison of the balancing principle (α_b, e_b) with the sampling-based optimal choice (α_o, e_o) for the 1d inverse diffusion coefficient problem.

Table 14.7: Computing times (in seconds) for the SSN method (t_s) and fixed point iteration (t_b) and L^2-error e for the 1d inverse diffusion coefficient problem. Shown are the problem size N, the mean ($\{t_s, t_b, e\}_m$) and standard deviation ($\{t_s, t_b, e\}_s$) over ten noise realizations.

N	100	200	400	800	1600	3200	6400	12800
$t_{s,m}$	0.70	2.19	6.49	11.46	25.48	55.34	82.38	167.71
$t_{s,s}$	0.46	1.65	0.91	2.80	5.34	17.19	23.07	31.01
$t_{b,m}$	4.12	8.68	19.27	27.59	52.45	97.32	154.46	332.96
$t_{b,s}$	2.51	3.71	2.20	6.28	9.14	21.96	22.59	39.17
e_m	2.17e-1	8.03e-2	4.94e-5	3.20e-5	2.89e-5	3.33e-5	3.39e-5	3.08e-5
e_s	2.05e-1	1.53e-1	2.66e-5	5.86e-6	4.89e-6	6.74e-6	5.42e-6	3.25e-6

L^∞ FITTING FOR INVERSE PROBLEMS WITH UNIFORM NOISE

15

ABSTRACT

For inverse problems where the data are corrupted by uniform noise such as arising from quantization errors, the L^∞ norm is a more robust data fitting term than the standard L^2 norm. Well-posedness and regularization properties for linear inverse problems with L^∞ data fitting are shown, and the automatic choice of the regularization parameter is discussed. After introducing an equivalent reformulation of the problem and a Moreau–Yosida approximation, a superlinearly convergent semi-smooth Newton method becomes applicable for the numerical solution of L^∞ fitting problems. Numerical examples illustrate the performance of the proposed approach as well as the qualitative behavior of L^∞ fitting.

15.1 INTRODUCTION

This work is concerned with the inverse problem

$$Kx = y^\delta$$

for a bounded linear operator K and given data y^δ corrupted by uniformly distributed noise. Apart from being often used in numerical tests of reconstruction algorithms, such noise appears as a statistical model of quantization errors and is therefore of relevance in any inverse problem where digital acquisition and processing of measured data plays a significant role [Widrow and Kollár 2008; Shykula and Seleznjev 2006]. Although advances in the resolution of analog-to-digital converters and in the floating-point precision of microprocessors have made these concerns less important for modern measurement equipment, they have become pertinent again in the context of wireless sensor networks. These consist of a large number of small, low-cost, usually battery-powered, densely distributed sensors which transmit gathered data to a central location [Gharavi and Kumar 2003; Arampatzis, Lygeros, and Manesis 2005]. Such networks have attracted increasing attention in recent years due to their wide range of

applications, e.g., in environmental monitoring where they can be used for quickly locating sources of a contaminant from distributed measurement of its concentration [Polastre et al. 2004; Doolin and Sitar 2005]. However, their communication is limited by their low power and shared bandwidth, which requires the data to be highly compressed before transmission, leading to significant quantization errors [Niu and Varshney 2006; Schizas, Giannakis, and Luo 2007]. More robust algorithms for state estimation from quantized data would therefore allow higher compression rates and therefore extended lifetime of the sensors. In this work, we are thus especially (but not exclusively) interested in inverse problems where K is the solution operator for a (linear) partial differential equation.

Since this problem is ill-posed, regularization needs to be applied. For uniform noise, the L^∞ norm is an appropriate term for measuring the data misfit due to its connection with the maximum likelihood estimator for this noise type (see, e.g., [Boyd and Vandenberghe 2004, Chapter 7.1.1]). This leads to minimizing a Tikhonov functional of the type

$$\min_x \left\|Kx - y^\delta\right\|_{L^\infty} + \alpha \|x\|^2 \tag{15.1.1}$$

or – if the noise level δ is known – a Morozov functional of the type

$$\min_x \|x\|^2 \quad \text{subject to} \quad \left\|Kx - y^\delta\right\|_{L^\infty} \leqslant \delta.$$

(These will be made precise below, see Section 15.2.) The difficulty arises from the non-differentiability of the L^∞ norm. This may be the reason why inverse problems in L^∞ have received rather little attention in the mathematical literature, even though there has been considerable recent progress in the regularization theory in Banach spaces (see, e.g., [Burger and Osher 2004; Resmerita 2005; Resmerita and Scherzer 2006; Hofmann et al. 2007; Pöschl 2009; Scherzer et al. 2009]). Numerical methods for minimizing L^∞ functionals have been investigated in [Williams and Kalogiratou 1993a; Williams and Kalogiratou 1993b] for curve fitting and parameter estimation for ordinary differential equations and in [Grund and Rösch 2001; Prüfert and Schiela 2009; Clason, Ito, and Kunisch 2010] for optimal control of partial differential equations. There has also be some recent interest in L^∞ functionals in the context of geometric vision [Hartley and Schaffalitzky 2004; Sim and Hartley 2006; Seo and Hartley 2007].

Our main interest thus lies in deriving an efficient method for the numerical solution of inverse problems with L^∞ fitting. Following [Grund and Rösch 2001; Prüfert and Schiela 2009; Clason, Ito, and Kunisch 2010], our approach is based on an equivalent formulation of (15.1.1):

$$\min_{c,x} c + \alpha \|x\|^2 \quad \text{subject to} \quad \left\|Kx - y^\delta\right\|_{L^\infty(\Omega)} \leqslant c.$$

This can be interpreted as an "augmented Morozov regularization" for the joint estimation of the unknown parameter x and the noise level $\delta = c$. (In fact, if δ is known, the proposed approach can be used for the numerical solution of the Morozov functional by fixing $c = \delta$, see Remark 15.4.7 below.) For this reformulation, we derive optimality conditions, introduce a

Moreau–Yosida approximation and show its convergence, and prove superlinear convergence of a semi-smooth Newton method. We also address the automatic choice of the regularization parameter α using a simple fixed-point iteration.

This paper is organized as follows. In Section 15.2, we address well-posedness and convergence of a slight generalization of the Tikhonov functional (15.1.1). Section 15.3 is concerned with the fixed-point algorithm for the automatic choice of the regularization parameter. The numerical solution of the L^∞ fitting problem is discussed in Section 15.4. Finally, numerical examples for one- and two-dimensional model problems are presented in Section 15.5.

15.2 WELL-POSEDNESS AND REGULARIZATION PROPERTIES

We consider for $1 \leqslant p < \infty$ the problem

$$(\mathcal{P}) \qquad \min_{x \in \mathcal{X}} \frac{1}{p} \left\| Kx - y^\delta \right\|_{L^\infty(\Omega)}^p + \frac{\alpha}{2} \left\| x - x_0 \right\|_{\mathcal{X}}^2,$$

where $K : \mathcal{X} \to L^\infty(\Omega)$ is a bounded linear operator defined on the Hilbert space X, $\Omega \subset \mathbb{R}^n$ is a bounded domain, $x_0 \in \mathcal{X}$ is given, and $y^\delta \in L^\infty(\Omega)$ are noisy measurements with noise level $\left\| y^\dagger - y^\delta \right\|_{L^\infty(\Omega)} \leqslant \delta$ (with $y^\dagger = Kx^\dagger$ being the noise-free data). If the kernel of K is non-trivial, we denote by $x^\dagger$ the x_0-minimum norm solution, i.e., the minimizer of $\|x - x_0\|_X$ over the set $\{x \in \mathcal{X} : Kx = y^\dagger\}$. Our main assumption on K (needed for convergence of the Moreau–Yosida approximation, see Theorem 15.4.2) is that

$$(15.2.1) \qquad x_n \rightharpoonup x^\dagger \text{ in } \mathcal{X} \quad \text{implies} \quad Kx_n \to Kx^\dagger \text{ in } L^\infty(\Omega).$$

This holds if K is a compact operator or maps into a space compactly embedded into $L^\infty(\Omega)$ (as is commonly the case if K is the solution operator for a partial differential equation).

The results of this section are standard (see, e.g., [Engl, Hanke, and Neubauer 1996, Chapters 5, 10], [Scherzer et al. 2009, Chapter 3.2]), and are given here to make the presentation self-contained. The first result concerns the well-posedness of ($\mathcal{P}$).

Theorem 15.2.1. *For $\alpha > 0$ and given y^δ,*

(i) there exists a unique solution $x_\alpha^\delta \in X$ to problem ($\mathcal{P}$);

(ii) for a sequence of data $\{y_n\}_{n\in\mathbb{N}}$ such that $y_n \to y^\delta$ in $L^\infty(\Omega)$, the sequence $\{x_\alpha^n\}_{n\in\mathbb{N}}$ of minimizers contains a subsequence converging to x_α^δ;

(iii) if the regularization parameter $\alpha = \alpha(\delta)$ satisfies

$$\lim_{\delta \to 0} \alpha(\delta) = \lim_{\delta \to 0} \frac{\delta^p}{\alpha(\delta)} = 0,$$

then the family $\{x_{\alpha(\delta)}^\delta\}_{\delta>0}$ has a subsequence converging to $x^\dagger$ as $\delta \to 0$.

Rates for the convergence in (iii) can be obtained under a source condition. Here, we assume the following condition: There exists a $w \in L^\infty(\Omega)^*$, i.e., a continuous linear functional on $L^\infty(\Omega)$, such that

$$x^\dagger - x_0 = K^* w. \tag{15.2.2}$$

For a discussion of source conditions and obtainable convergence rates, see [Scherzer et al. 2009, Chapter 3.2]. More general smoothness assumptions and their relation to source conditions are discussed in [Hofmann et al. 2007; Flemming and Hofmann 2011].

Theorem 15.2.2. *If the source condition* (15.2.2) *holds, and* $\alpha = \mathcal{O}(\delta^\varepsilon)$ *with* $\varepsilon \in (0,1)$ *in case* $p = 1$ *and* $\alpha = \mathcal{O}(\delta^{p-1})$ *in case* $p > 1$*, then the minimizer* x_α^δ *of* ($\mathcal{P}$) *satisfies*

$$\|x_\alpha^\delta - x^\dagger\|_X \leqslant \begin{cases} c\delta^{\frac{1-\varepsilon}{2}} & \text{if } p = 1, \\ c\delta^{\frac{1}{2}} & \text{if } p > 1. \end{cases}$$

Proof. By the minimizing property of x_α^δ, we have

$$\begin{aligned} \frac{1}{p}\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)}^p + \frac{\alpha}{2}\|x_\alpha^\delta - x_0\|_X^2 &\leqslant \frac{1}{p}\|Kx^\dagger - y^\delta\|_{L^\infty(\Omega)}^p + \frac{\alpha}{2}\|x^\dagger - x_0\|_X^2 \\ &\leqslant \frac{\delta^p}{p} + \frac{\alpha}{2}\|x^\dagger - x_0\|_X^2 \end{aligned}$$

and hence

$$\frac{1}{p}\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)}^p + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_X^2 \leqslant \frac{\delta^p}{p} + \alpha\langle x^\dagger - x_0, x^\dagger - x_\alpha^\delta\rangle_X.$$

Now by the source condition (15.2.2), we have

$$\begin{aligned} \frac{1}{p}\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)}^p + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_X^2 &\leqslant \frac{\delta^p}{p} + \alpha\langle K^* w, x^\dagger - x_\alpha^\delta\rangle_X \\ &= \frac{\delta^p}{p} + \alpha\langle w, y^\dagger - Kx_\alpha^\delta\rangle_{L^\infty(\Omega)^*, L^\infty(\Omega)} \\ &\leqslant \frac{\delta^p}{p} + \alpha\,\|w\|_{L^\infty(\Omega)^*}\,\|Kx_\alpha^\delta - y^\dagger\|_{L^\infty(\Omega)}. \end{aligned}$$

Inserting the productive zero $0 = y^\delta - y^\delta$ on the right hand side and applying the triangle inequality yields

$$\begin{aligned} &\frac{1}{p}\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)}^p + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_X^2 \\ &\qquad\leqslant \frac{\delta^p}{p} + \alpha\,\|w\|_{L^\infty(\Omega)^*}\,(\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)} + \|y^\dagger - y^\delta\|_{L^\infty(\Omega)}). \end{aligned} \tag{15.2.3}$$

If $p = 1$, we have

$$(1 - \alpha \|w\|_{L^\infty(\Omega)^*})\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)} + \frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_X^2 \leqslant (1 + \alpha \|w\|_{L^\infty(\Omega)^*})\delta,$$

from which the desired convergence rate follows by choosing choosing $\alpha = \mathcal{O}(\delta^\varepsilon)$. For $p > 1$, we use Young's inequality $ab \leqslant \frac{1}{p}a^p + \frac{1}{p'}b^{p'}$ for $p' = \frac{p}{p-1}$, $a = \|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)}$ and $b = \alpha \|w\|_{L^\infty(\Omega)^*}$, and rearrange terms to deduce

$$-\frac{1}{p}\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)}^p \leqslant -\alpha \|w\|_{L^\infty(\Omega)^*} \|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)} + \frac{1}{p'}(\alpha \|w\|_{L^\infty(\Omega)^*})^{p'}.$$

Hence, by adding the last term on the right hand side to both sides of (15.2.3), we obtain

$$\frac{\alpha}{2}\|x_\alpha^\delta - x^\dagger\|_X^2 \leqslant \frac{\delta^p}{p} + \alpha \|w\|_{L^\infty(\Omega)^*}\delta + \frac{1}{p'}(\alpha \|w\|_{L^\infty(\Omega)^*})^{p'}.$$

Taking $\alpha = \mathcal{O}(\delta^{p-1})$ then yields the claimed estimate. □

We remark that for $p = 1$, ($\mathcal{P}$) is an exact penalization, i.e., there exists $\alpha^* > 0$ such that for all $\alpha < \alpha^*$, the minimizer x_α^0 of ($\mathcal{P}$) with exact data $y^\dagger$ satisfies $x_\alpha^0 = x^\dagger$, see [Burger and Osher 2004; Hofmann et al. 2007].

Next we consider Morozov's discrepancy principle [Morozov 1966], which consists in choosing α such that

$$\|Kx_\alpha^\delta - y^\delta\|_{L^\infty(\Omega)} = \tau\delta$$

for some $\tau > 1$.

Theorem 15.2.3. *Assume that the source condition (15.2.2) holds, and that the regularization parameter $\alpha = \alpha(\delta)$ is determined according to the discrepancy principle. Then the minimizer x_α^δ of ($\mathcal{P}$) satisfies*

$$\|x_\alpha^\delta - x^\dagger\|_X \leqslant c\delta^{\frac{1}{2}}.$$

Proof. By the minimizing property of x_α^δ and the choice of α, we have

$$\|x_\alpha^\delta - x_0\|_X^2 \leqslant \|x^\dagger - x_0\|_X^2,$$

from which it follows that

$$\begin{aligned}\|x_\alpha^\delta - x^\dagger\|_X^2 &\leqslant 2\langle x^\dagger - x_0, x^\dagger - x_\alpha^\delta\rangle_X = 2\langle K^*w, x^\dagger - x_\alpha^\delta\rangle_X \\ &\leqslant 2\|w\|_{L^\infty(\Omega)^*}\|Kx^\dagger - Kx_\alpha^\delta\|_{L^\infty(\Omega)} \\ &\leqslant 2\|w\|_{L^\infty(\Omega)^*}(\|y^\dagger - y^\delta\|_{L^\infty(\Omega)} + \|y^\delta - Kx_\alpha^\delta\|_{L^\infty(\Omega)}) \\ &\leqslant 2\|w\|_{L^\infty(\Omega)^*}(1+\tau)\delta,\end{aligned}$$

and hence we obtain the claimed estimate. □

Remark 15.2.4. If K is an adjoint operator, i.e., there exists $K_* : L^1(\Omega) \to X$ such that $(K_*)^* = K$, the source condition can be stated as: There exists $w \in L^1(\Omega)$ such that $x^\dagger - x_0 = K_*w$.

15.3 PARAMETER CHOICE

Morozov's discrepancy principle requires knowledge of the noise level, which is often not available in practice. Here, we use a heuristic choice rule derived from a balancing principle [Clason, Jin, and Kunisch 2010a; Clason, Jin, and Kunisch 2010b; Clason and Jin 2012], which involves auto-calibration of the noise level. Although there is no rigorous justification, we can give a brief motivation of this principle. Recall that the Morozov discrepancy principle chooses α such that the residual in the appropriate norm is on the order of the noise level δ. In an iterative scheme, one would start with a large parameter and reduce it until this condition is satisfied, making use of the fact that the norm of the residual is monotonically increasing as a function of α (see Lemma 15.3.1 below). On the other hand, the regularization term is monotonically decreasing; one could therefore equally choose α such that the regularization term reaches a certain value, which is proportional to the noise level δ. If δ is not known, the current residual can be used in this approach to give an estimate of the noise level. If the true data and the noise are sufficiently structurally different, it can be expected that

$$\left\|Kx_\alpha - y^\delta\right\|_{L^\infty(\Omega)} \approx \delta$$

for a reasonable range of α. (A similar assumption can be used to show convergence of minimization-based noise level-free parameter choice rules [Kindermann 2011].) This motivates considering the following heuristic principle: Choose $\alpha > 0$ such that the balancing equation

$$\frac{\alpha}{2}\|x_\alpha - x_0\|_X^2 = \sigma\|Kx_\alpha - y^\delta\|_{L^\infty(\Omega)} \tag{15.3.1}$$

is satisfied. (Note that $\|Kx^\dagger - y^\delta\|_{L^\infty(\Omega)}$ rather than $\|Kx^\dagger - y^\delta\|_{L^\infty(\Omega)}^p$ is the true noise level by definition.) Here, σ is a proportionality constant which depends on K and X, but not on δ.

We can compute a solution α^* to (15.3.1) by the following simple fixed-point algorithm proposed in [Clason, Jin, and Kunisch 2010b]:

$$\alpha_{k+1} = \sigma\frac{\|Kx_{\alpha_k} - y^\delta\|_{L^\infty(\Omega)}}{\frac{1}{2}\|x_{\alpha_k} - x_0\|_X^2}. \tag{15.3.2}$$

This fixed-point algorithm can be derived formally from the model function approach [Clason, Jin, and Kunisch 2010a]. The convergence can be proven similarly as in [Clason, Jin, and Kunisch 2010b]. We start by arguing monotonicity of the data fitting and of the regularization term.

Lemma 15.3.1. *The functions* $\|Kx_\alpha - y^\delta\|_{L^\infty(\Omega)}$ *and* $\|x_\alpha - x_0\|_X$ *are monotonic in* α*, in the sense that for* $\alpha_1, \alpha_2 > 0$*,*

$$\left(\|Kx_{\alpha_1} - y^\delta\|_{L^\infty(\Omega)} - \|Kx_{\alpha_2} - y^\delta\|_{L^\infty(\Omega)}\right)(\alpha_1 - \alpha_2) \geqslant 0$$

and

$$\left(\|x_{\alpha_1} - x_0\|_X^2 - \|x_{\alpha_2} - x_0\|_X^2\right)(\alpha_1 - \alpha_2) \leqslant 0.$$

Proof. The minimizing property of x_{α_1} and x_{α_2} yields

$$\frac{1}{p}\left\|Kx_{\alpha_1}-y^\delta\right\|^p_{L^\infty(\Omega)}+\frac{\alpha_1}{2}\left\|x_{\alpha_1}-x_0\right\|^2_X \leqslant \frac{1}{p}\left\|Kx_{\alpha_2}-y^\delta\right\|^p_{L^\infty(\Omega)}+\frac{\alpha_1}{2}\left\|x_{\alpha_2}-x_0\right\|^2_X,$$
$$\frac{1}{p}\left\|Kx_{\alpha_2}-y^\delta\right\|^p_{L^\infty(\Omega)}+\frac{\alpha_2}{2}\left\|x_{\alpha_2}-x_0\right\|^2_X \leqslant \frac{1}{p}\left\|Kx_{\alpha_1}-y^\delta\right\|^p_{L^\infty(\Omega)}+\frac{\alpha_2}{2}\left\|x_{\alpha_1}-x_0\right\|^2_X.$$

Adding these two inequalities together gives the second estimate. The first one can be obtained by dividing the two inequalities by α_1 and α_2, respectively, adding them together, and using the monotonicity of $t \mapsto t^p$ for $p \geqslant 1$ and $t \geqslant 0$. □

We shall denote by

$$r(\alpha) = \sigma\|Kx_\alpha - y^\delta\|_{L^\infty(\Omega)} - \frac{\alpha}{2}\|x_\alpha - x_0\|^2_X$$

the residual in (15.3.1). The next lemma shows the monotonicity of the iteration (15.3.2).

Lemma 15.3.2. *The sequence of regularization parameters $\{\alpha_k\}_{k\in\mathbb{N}}$ generated by the fixed-point algorithm is monotonically increasing if $r(\alpha_0) > 0$ and monotonically decreasing if $r(\alpha_0) < 0$.*

Proof. We argue by induction. If $r(\alpha_0) > 0$, then by the definition of the iteration, we have

$$\alpha_1 = \sigma\frac{\|Kx_{\alpha_0} - y^\delta\|_{L^\infty(\Omega)}}{\frac{1}{2}\|x_{\alpha_0} - x_0\|^2_X} > \alpha_0,$$

and similarly (with the opposite inequality) if $r(\alpha_0) < 0$. Now for any $k > 1$,

$$\begin{aligned}\alpha_{k+1} - \alpha_k &= \sigma\left(\frac{\|Kx_{\alpha_k} - y^\delta\|_{L^\infty(\Omega)}}{\frac{1}{2}\|x_{\alpha_k} - x_0\|^2_X} - \frac{\|Kx_{\alpha_{k-1}} - y^\delta\|_{L^\infty(\Omega)}}{\frac{1}{2}\|x_{\alpha_{k-1}} - x_0\|^2_X}\right)\\ &= \frac{2\sigma}{\|x_{\alpha_k} - x_0\|^2_X\|x_{\alpha_{k-1}} - x_0\|^2_X}\\ &\quad\Big[\|Kx_{\alpha_k} - y^\delta\|_{L^\infty(\Omega)}(\|x_{\alpha_{k-1}} - x_0\|^2_X - \|x_{\alpha_k} - x_0\|^2_X)\\ &\qquad + \|x_{\alpha_k} - x_0\|^2_X(\|Kx_{\alpha_k} - y^\delta\|_{L^\infty(\Omega)} - \|Kx_{\alpha_{k-1}} - y^\delta\|_{L^\infty(\Omega)})\Big].\end{aligned}$$

By Lemma 15.3.1, the two terms in parentheses both have the sign of $(\alpha_k - \alpha_{k-1})$, and thus the whole sequence is monotonic. □

Theorem 15.3.3. *If the initial guess α_0 satisfies $r(\alpha_0) < 0$, then the sequence $\{\alpha_k\}$ generated by the fixed-point algorithm converges to a solution to* (15.3.1).

Proof. By Lemma 15.3.2 and $r(\alpha_0) < 0$, the sequence $\{\alpha_k\}$ is monotonically decreasing. Since by definition (15.3.2) it is clearly bounded from below by zero, convergence follows. □

Note that Theorem 15.3.3 gives a constructive method of choosing a suitable parameter σ: Set α_0 sufficiently large (e.g., $\alpha_0 = 1$) and select σ small enough such that $r(\alpha_0) < 0$ is satisfied.

Remark 15.3.4. The convergence of the fixed-point iteration solely depends on the monotonicity properties of the fitting and of the regularization term. It can therefore be applied for finding solutions to the balancing equation

$$\alpha \mathcal{R}(x_\alpha) = \sigma \mathcal{F}(x_\alpha; y^\delta)$$

for minimizers x_α of the Tikhonov functional

$$\varphi\left(\mathcal{F}(x; y^\delta)\right) + \alpha \mathcal{R}(x),$$

where φ is any monotone, real-valued function and $\mathcal{R}$, $\mathcal{F}$ are arbitrary functionals for which the minimization problem is well-posed.

15.4 NUMERICAL SOLUTION

The numerical solution of problem ($\mathcal{P}$) is based on a sequence of Moreau–Yosida approximations of an equivalent formulation of ($\mathcal{P}$), which can be solved using a superlinearly convergent semi-smooth Newton method.

15.4.1 REFORMULATION

We begin by introducing an equivalent formulation of ($\mathcal{P}$) that allows making use of techniques developed for optimization problems for partial differential equations with state constraints (see [Grund and Rösch 2001; Prüfert and Schiela 2009; Clason, Ito, and Kunisch 2010]). Since we wish to apply a Newton-type method, we fix $p = 2$ from here on (guaranteeing positive definiteness of the Hessian; see Theorem 15.4.4). Note that the value of p only influences the trade-off between minimizing the L^∞ norm of the residual and minimizing the norm of x, but not the relevant structural properties of the functional (in particular the geometry of the unit ball with respect to $\|\cdot\|_{L^\infty}^p$). Without loss of generality, we also set $x_0 = 0$ and consider

$$(\mathcal{P}_c) \qquad \min_{(x,c)\in X\times\mathbb{R}} \frac{c^2}{2} + \frac{\alpha}{2}\|x\|_X^2 \quad \text{subject to} \quad \left\|Kx - y^\delta\right\|_{L^\infty(\Omega)} \leqslant c.$$

The strict convexity of the equivalent problem ($\mathcal{P}$) directly yields the existence of a unique minimizer (x^*, c^*) with $x^* = x_\alpha^\delta$ from Theorem 15.2.1 and $c^* = \left\|Kx_\alpha^\delta - y^\delta\right\|_{L^\infty(\Omega)}$.

For a *differentiable* strictly convex functional J, the minimizer x^* can be found by computing a stationary point, which satisfies the optimality condition $J'(x^*) = 0$. For nondifferentiable problems such as ($\mathcal{P}_c$), a similar equivalence holds, although the optimality conditions become more complicated and a so-called regular point condition needs to be satisfied (see, e.g., [Maurer and Zowe 1979; Ito and Kunisch 2008]).

In the following, $j : \mathcal{X} \to \mathcal{X}^*$ denotes the (linear) duality mapping of the Hilbert space $\mathcal{X}$, i.e., $j(u) = \partial\left(\frac{1}{2}\|\cdot\|_{\mathcal{X}}^2\right)(u)$, and $\langle\cdot,\cdot\rangle_{L^\infty(\Omega)^*,L^\infty(\Omega)}$ the duality pairing between $L^\infty(\Omega)$ and its topological dual.

Theorem 15.4.1. *Let $(x^*, c^*) \in \mathcal{X} \times \mathbb{R}$ be the solution to ($\mathcal{P}_c$). Then there exist $\lambda_1, \lambda_2 \in L^\infty(\Omega)^*$ with*

$$\langle \lambda_1, \varphi\rangle_{L^\infty(\Omega)^*,L^\infty(\Omega)} \leqslant 0, \qquad \langle \lambda_2, \varphi\rangle_{L^\infty(\Omega)^*,L^\infty(\Omega)} \geqslant 0 \tag{15.4.1}$$

for all $\varphi \in L^\infty(\Omega)$ with $\varphi \geqslant 0$ such that

$$\text{(OS)} \qquad \begin{cases} \alpha j(x^*) = K^*(\lambda_1 + \lambda_2), \\ c^* = \langle \lambda_1 - \lambda_2, -1\rangle_{L^\infty(\Omega)^*,L^\infty(\Omega)}, \\ 0 = \langle \lambda_1, Kx^* - y^\delta - c^*\rangle_{L^\infty(\Omega)^*,L^\infty(\Omega)}, \\ 0 = \langle \lambda_2, Kx^* - y^\delta + c^*\rangle_{L^\infty(\Omega)^*,L^\infty(\Omega)}. \end{cases}$$

Proof. Let $G : \mathcal{X} \times \mathbb{R} \to L^\infty(\Omega) \times L^\infty(\Omega)$ be defined by

$$G(x, c) = \begin{pmatrix} Kx - y^\delta - c \\ -Kx + y^\delta - c \end{pmatrix},$$

and let K denote the non-positive cone in $L^\infty(\Omega)$, i.e.,

$$K = \{y \in L^\infty(\Omega) : y \leqslant 0\}.$$

With $J : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$,

$$J(x, c) = \frac{c^2}{2} + \frac{\alpha}{2}\|x\|_{\mathcal{X}}^2,$$

we can express ($\mathcal{P}_c$) as

$$\min_{(x,c)\in\mathcal{X}\times\mathbb{R}} J(x, c) \quad \text{subject to} \quad G(x, c) \in K \times K. \tag{15.4.2}$$

The regular point condition [Maurer and Zowe 1979; Ito and Kunisch 2008] for (15.4.2) is

$$0 \in \operatorname{int}\left(G(x^*, c^*) + G'(x^*, c^*)(\mathcal{X} \times \mathbb{R}) - K \times K\right), \tag{15.4.3}$$

where int denotes the topological interior and G' is the Fréchet derivative of G. To verify (15.4.3), we need to find $\bar{x} \in \mathcal{X}$ and $\bar{c} \in \mathbb{R}$ such that

$$\begin{aligned} (Kx^* - y^\delta - c^*) + K\bar{x} - \bar{c} < 0, \\ (-Kx^* + y^\delta - c^*) - K\bar{x} - \bar{c} < 0. \end{aligned}$$

Since the minimizer (x^*, c^*) satisfies the L^∞ bound, the terms in parentheses are non-positive almost everywhere, and thus these conditions are satisfied for $\bar{x} = 0$ and arbitrary $\bar{c} > 0$.

From [Maurer and Zowe 1979, Theorem 3.2], we then obtain the existence of $(\lambda_1, -\lambda_2)$ in the dual cone of $K \times K$ (i.e., satisfying (15.4.1)), such that with $Y := L^\infty(\Omega) \times L^\infty(\Omega)$,

$$J'(x^*, c^*) = \langle (\lambda_1, -\lambda_2), G'(x^*, c^*) \rangle_{y^*, y}$$

and

$$\langle (\lambda_1, -\lambda_2), G(x^*, c^*) \rangle_{y^*, y} = 0$$

hold. Inserting the explicit form of J', G, and G' yields (OS). □

15.4.2 MOREAU–YOSIDA APPROXIMATION

To avoid dealing with the dual space of $L^\infty(\Omega)$, we consider for $\gamma > 0$ the Moreau–Yosida approximation

$$(\mathcal{P}_\gamma) \qquad \min_{(x,c)\in X\times\mathbb{R}} \frac{c^2}{2} + \frac{\alpha}{2}\|x\|_X^2 + \frac{\gamma}{2}\left\|\max(0, Kx - y^\delta - c)\right\|_{L^2(\Omega)}^2 + \frac{\gamma}{2}\left\|\min(0, Kx - y^\delta + c)\right\|_{L^2(\Omega)}^2,$$

where the max and min are to be understood pointwise in Ω. Since this is a strictly convex and weakly lower semi-continuous problem in c and x, there exists a unique solution $(x_\gamma, c_\gamma) \in X \times \mathbb{R}$.

We next address convergence of (x_γ, c_γ) to the solution (x^*, c^*) to $(\mathcal{P}_c)$.

Theorem 15.4.2. *As $\gamma \to \infty$, (x_γ, c_γ) converges strongly to (x^*, c^*) in $X \times \mathbb{R}$.*

Proof. Let

$$\lambda_{\gamma,1} = \gamma \max(0, Kx_\gamma - y^\delta - c_\gamma), \qquad \lambda_{\gamma,2} = \gamma \min(0, Kx_\gamma - y^\delta + c_\gamma).$$

Due to the optimality of (x_γ, c_γ) and the feasibility of (x^*, c^*), we have for all $\gamma > 0$ that

$$\frac{(c_\gamma)^2}{2} + \frac{\alpha}{2}\|x_\gamma\|_X^2 + \frac{1}{2\gamma}\|\lambda_{\gamma,1}\|_{L^2(\Omega)}^2 + \frac{1}{2\gamma}\|\lambda_{\gamma,2}\|_{L^2(\Omega)}^2 \leqslant \frac{(c^*)^2}{2} + \frac{\alpha}{2}\|x^*\|_X^2 \tag{15.4.5}$$

and hence that the families

$$\{x_\gamma\}_{\gamma>0}, \quad \{c_\gamma\}_{\gamma>0}, \quad \left\{\gamma^{-1}\|\lambda_{\gamma,1}\|_{L^2(\Omega)}^2\right\}_{\gamma>0}, \quad \left\{\gamma^{-1}\|\lambda_{\gamma,2}\|_{L^2(\Omega)}^2\right\}_{\gamma>0}$$

are bounded. Consequently, there exists a sequence $\{\gamma_k\}_{k\in\mathbb{N}}$ and $(\hat{x}, \hat{c}) \in X \times \mathbb{R}$ such that $(x_{\gamma_k}, c_{\gamma_k})$ converges to $(\hat{x}, \hat{c})$ and

$$\left\|\max(0, Kx_{\gamma_k} - y^\delta - c_{\gamma_k})\right\|_{L^2(\Omega)}^2 \leqslant \frac{C}{\gamma_k} \to 0$$

for $k \to \infty$, and similarly for $\min(0, Kx_{\gamma_k} - y^\delta + c_{\gamma_k})$. Since $Kx_{\gamma_k} \to K\hat{x}$ strongly in $L^\infty(\Omega)$ by assumption (15.2.1), this implies that

$$\left\| K\hat{x} - y^\delta \right\|_{L^\infty(\Omega)} \leqslant \hat{c}.$$

Taking the limit in (15.4.5), we thus find that $(\hat{x}, \hat{c})$ coincides with the unique solution (x^*, c^*) to $(\mathcal{P}_c)$. Due to uniqueness of (x^*, c^*), the whole family $\{(x_\gamma, c_\gamma)\}_{\gamma>0}$ converges in $\mathcal{X} \times \mathbb{R}$ to (x^*, c^*). □

Since $(\mathcal{P}_\gamma)$ is differentiable and strictly convex, straightforward computation yields the (necessary and sufficient) optimality conditions

$$(OS_\gamma) \qquad \begin{cases} \alpha j(x_\gamma) + \gamma K^* \left(\max(0, Kx_\gamma - y^\delta - c_\gamma) + \min(0, Kx_\gamma - y^\delta + c_\gamma)\right) = 0, \\ c_\gamma + \gamma \left\langle -\max(0, Kx_\gamma - y^\delta - c_\gamma) + \min(0, Kx_\gamma - y^\delta + c_\gamma), 1 \right\rangle_{L^2(\Omega)} = 0. \end{cases}$$

Remark 15.4.3. Similarly to [Clason, Ito, and Kunisch 2010, Theorem 3.1], one can show that as $\gamma \to \infty$,

$$\begin{aligned} -\gamma \max(0, Kx_\gamma - y^\delta - c_\gamma) &\rightharpoonup \lambda_1, \\ \gamma \min(0, Kx_\gamma - y^\delta + c_\gamma) &\rightharpoonup \lambda_2, \end{aligned}$$

weakly-$\star$ in $L^\infty(\Omega)^*$, with λ_1 and λ_2 as given by Theorem 15.4.1.

15.4.3 SEMI-SMOOTH NEWTON METHOD

To solve the optimality system (OS_γ) with a semi-smooth Newton method [Kummer 1992; Chen, Nashed, and Qi 2000; Hintermüller, Ito, and Kunisch 2002; Ulbrich 2002], we consider it as an operator equation $F(x, c) = 0$ for $F : \mathcal{X} \times \mathbb{R} \to \mathcal{X}^* \times \mathbb{R}$,

$$F(x, c) = \begin{pmatrix} \alpha j(x) + \gamma K^* \left(\max(0, Kx - y^\delta - c) + \min(0, Kx - y^\delta + c)\right) \\ c + \gamma \left\langle -\max(0, Kx - y^\delta - c) + \min(0, Kx - y^\delta + c), 1 \right\rangle_{L^2(\Omega)} \end{pmatrix}.$$

We now argue the Newton differentiability of F. Recall that a mapping $F : X \to Y$ between Banach spaces X and Y is Newton differentiable at $x \in X$ if there exists a neighborhood $N(x)$ and a mapping $G : N(x) \to L(X, Y)$ with

$$\lim_{\|h\| \to 0} \frac{\|F(x + h) - F(x) - G(x + h)h\|_Y}{\|h\|_X} \to 0.$$

(Note that in contrast with Fréchet differentiability, the linearization is taken in a neighborhood $N(x)$ of x.) Any mapping $D_N F \in \{G(s) : s \in N(x)\}$ is then a Newton derivative of F at x.

Now we have (e.g., from [Ito and Kunisch 2008, Example 8.14]; see also [Schiela 2008]) that the function $z \mapsto \max(0, z)$ is Newton differentiable from $L^p(\Omega)$ to $L^q(\Omega)$ for any $p > q \geqslant 1$. Furthermore, the chain rule for Newton derivatives (e.g., [Ito and Kunisch 2008, Lemma 8.15]) yields that for a linear operator B with range contained in $L^p(\Omega)$, the Newton derivative of $\max(0, Bv)$ at v in direction δv is given pointwise almost everywhere by

$$\big(D_N \max(0, Bv)\delta v\big)(t) = \begin{cases} (B\delta v)(t), & \text{if } v(t) > 0, \\ 0, & \text{if } v(t) \leqslant 0. \end{cases}$$

Since for any $x \in X$ we have $Kx \in L^\infty(\Omega)$, the mapping

$$x \mapsto \max(0, Kx - y^\delta - c) \tag{15.4.6}$$

for fixed c is Newton differentiable from $\mathcal{X}$ to $L^1(\Omega) \subset L^\infty(\Omega)^*$ with Newton derivative in direction $\delta x \in \mathcal{X}$ given by $(K\delta x)\chi_1$, where

$$\chi_1(t) = \begin{cases} 1 & \text{if } (Kx - y^\delta - c)(t) > 0, \\ 0 & \text{if } (Kx - y^\delta - c)(t) \leqslant 0. \end{cases} \tag{15.4.7}$$

Similarly, the embedding that maps $c \in \mathbb{R}$ to the constant function $t \mapsto c \in L^\infty(\Omega)$ yields Newton differentiability of (15.4.6) with respect to c for fixed $x \in \mathcal{X}$ from $\mathbb{R}$ to $L^1(\Omega) \subset L^\infty(\Omega)^*$, with Newton derivative in direction $\delta c \in \mathbb{R}$ given by $(-\delta c)\chi_1$. One proceeds analogously for the min terms by defining

$$\chi_2(t) = \begin{cases} 1 & \text{if } (Kx - y^\delta + c)(t) < 0, \\ 0 & \text{if } (Kx - y^\delta + c)(t) \geqslant 0. \end{cases} \tag{15.4.8}$$

Altogether, F is Newton differentiable from $\mathcal{X} \times \mathbb{R} \to \mathcal{X}^* \times \mathbb{R}$ with Newton derivative at $(x, c) \in \mathcal{X} \times \mathbb{R}$ given by

$$D_N F(x, c)(\delta x, \delta c) = \begin{pmatrix} \alpha j'(x)\delta x + \gamma K^*((\chi_1 + \chi_2)K\delta x) + \gamma \delta c K^*(-\chi_1 + \chi_2) \\ \gamma \langle -\chi_1 + \chi_2, K\delta x \rangle_{L^2(\Omega)} + \left(1 + \gamma \langle \chi_1 + \chi_2, 1 \rangle_{L^2(\Omega)}\right) \delta c \end{pmatrix},$$

For given (x^k, c^k), a semi-smooth Newton step consists in solving for $(\delta x, \delta c) \in \mathcal{X} \times \mathbb{R}$ in

$$D_N F(x^k, c^k)(\delta x, \delta c) = -F(x^k, c^k) \tag{15.4.9}$$

and setting $x^{k+1} = x^k + \delta x$, $c^{k+1} = c^k + \delta c$. It remains to show uniform invertibility of the Newton step, which will imply local superlinear convergence of the sequence of iterates (x^k, c^k).

Theorem 15.4.4. *For every $\alpha, \gamma > 0$, the sequence (x^k, c^k) of iterates in (15.4.9) converges superlinearly to the solution (x_γ, c_γ) to (OS_γ), provided that (x^0, c^0) is sufficiently close to (x_γ, c_γ).*

Proof. For arbitrary $(x, c) \in \mathcal{X} \times \mathbb{R}$ and $(\delta x, \delta c) \in \mathcal{X} \times \mathbb{R}$, we have

$$\begin{aligned}\langle (\delta x, \delta c), D_N F(x, c)(\delta x, \delta c)\rangle_{\mathcal{X}\times\mathbb{R},(\mathcal{X}^*\times\mathbb{R}} &= \alpha \|\delta x\|_{\mathcal{X}}^2 + \delta c^2 \\ &+ \gamma \left(\|(\chi_1 + \chi_2) K\delta x\|_{L^2(\Omega)}^2 + 2\delta c \langle -\chi_1 + \chi_2, K\delta x\rangle_{L^2(\Omega)} + \delta c^2 \|\chi_1 + \chi_2\|_{L^2(\Omega)}^2 \right).\end{aligned}$$

Since χ_1 and χ_2 are characteristic functions of disjoint sets, we can estimate separately

$$\begin{aligned}\|\chi_1 K\delta x\|_{L^2(\Omega)}^2 - 2\langle \delta c\chi_1, K\delta x\rangle_{L^2(\Omega)} + \|\delta c\chi_1\|_{L^2(\Omega)}^2 &= \|\chi_1(K\delta x - \delta c)\|_{L^2(\Omega)}^2 \geqslant 0, \\ \|\chi_2 K\delta x\|_{L^2(\Omega)}^2 + 2\langle \delta c\chi_2, K\delta x\rangle_{L^2(\Omega)} + \|\delta c\chi_2\|_{L^2(\Omega)}^2 &= \|\chi_2(K\delta x + \delta c)\|_{L^2(\Omega)}^2 \geqslant 0.\end{aligned}$$

This implies

$$\langle (\delta x, \delta c), D_N F(x, c)(\delta x, \delta c)\rangle_{\mathcal{X}\times\mathbb{R},\mathcal{X}^*\times\mathbb{R}} \geqslant \alpha \|\delta x\|_{\mathcal{X}}^2 + \delta c^2$$

and thus that $D_N F(x, c)$ is an isomorphism independent of (x, c). The local superlinear convergence now follows from standard results (e.g., [Ito and Kunisch 2008, Theorem 8.16]). □

The following property (e.g., [Ito and Kunisch 2008, Remark 7.1.1]) yields an objective stopping criterion for the semismooth Newton method. Let

$$\begin{aligned}\mathcal{A}_1^k &= \left\{ t \in \Omega : (Kx^k - y^\delta - c^k)(t) > 0 \right\}, \\ \mathcal{A}_2^k &= \left\{ t \in \Omega : (Kx^k - y^\delta + c^k)(t) < 0 \right\}.\end{aligned}$$

denote the sets of points where the L^∞ norm bound is violated in iteration k.

Proposition 15.4.5. *If $\mathcal{A}_1^{k+1} = \mathcal{A}_1^k$ and $\mathcal{A}_2^{k+1} = \mathcal{A}_2^k$, then $F(x^{k+1}, c^{k+1}) = 0$.*

To deal with the local convergence of Newton's method, we make use of a continuation strategy in the numerical computation: Solve (OS_γ) for fixed $\gamma_k > 0$, choose $\gamma_{k+1} > \gamma_k$, and compute the next solution $(x_{\gamma_{k+1}}, c_{\gamma_{k+1}})$ using $(x_{\gamma_k}, c_{\gamma_k})$ as starting point. If γ_0 is sufficiently small (e.g., $\gamma_0 = 1$), one can expect convergence of the continuation scheme for any reasonable choice of (x_0, c_0) (e.g., $(x_0, c_0) = (0, 0)$ in the absence of a priori information). The full procedure for computing a numerical approximation of the solution to problem ($\mathcal{P}_c$) is given as Algorithm 15.1.

Finally, we remark on how the presented approach can be simplified in special cases.

Remark 15.4.6. In the case where K is the solution operator for a linear partial differential equation, i.e., $K = A^{-1}$ for a partial differential operator $A : Y \to Y^* \supset \mathcal{X}$ on the reflexive Banach space Y, (OS_γ) can be reformulated in a more convenient way by introducing $y = A^{-1}x$ as an independent variable and using a Lagrange multiplier approach to enforce the constraint $Ay = x$. This leads to a (semi-smooth) block optimality system, which in many cases can again be reduced to a pair of equations for (y, c) only. Take $\mathcal{X} = L^2(\Omega)$, i.e.,

Algorithm 15.1 Semi-smooth Newton method with continuation

Choose (x^0, c^0), γ^0, $\tau > 1$, $\varepsilon > 0$, k^*, γ^*; set $j = 0$
repeat
  Increment $j \leftarrow j + 1$
  Set $x_0 = x^{j-1}$, $k = 0$
  repeat
    Increment $k \leftarrow k + 1$
    Compute indicator function of active sets : χ_1^k, χ_2^k from (15.4.7) and (15.4.8)
    Solve for δx, δc in (15.4.9)
    Update $x^k = x^{k-1} + \delta x$, $c^k = c^{k-1} + \delta c$
  until $\chi_1^{k+1} = \chi_1^k$ and $\chi_2^{k+1} = \chi_2^k$, or $k = k^*$
  Set $x^j = x_k$, $c^j = c_k$
  Set $\gamma^j = \tau\gamma^{j-1}$
until $\|Kx^j - y^\delta\|_{L^\infty(\Omega)} < c + \varepsilon$ or $\gamma = \gamma^*$

$j(x) = x$, and assume that A is an isomorphism from $\mathcal{W} = H_0^1(\Omega) \cap H^2(\Omega)$ to $\mathcal{W}^*$. Due to the embedding $\mathcal{W} \hookrightarrow C_0(\Omega)$, we have that the range $A^{-1}(\mathcal{X})$ embeds compactly into $L^\infty(\Omega)$. Inserting $y_\gamma = A^{-1}x_\gamma \in \mathcal{W}$ into the first equation of (OS_γ) yields

$$\alpha A y_\gamma + \gamma A^{-*}\left(\max(0, y_\gamma - y^\delta - c_\gamma) + \min(0, y_\gamma - y^\delta + c_\gamma)\right) = 0.$$

Since the term in parentheses is in $L^\infty(\Omega)$, the mapping properties of A^{-*} yield that $Ay_\gamma \in \mathcal{W}$. We can thus apply A^* to the whole equation to obtain that (y_γ, c_γ) satisfies $F(y_\gamma, c_\gamma) = 0$ for $F : \mathcal{W} \times \mathbb{R} \to \mathcal{W}^* \times \mathbb{R}$,

$$F(y, c) = \begin{pmatrix} \alpha A^* A y + \gamma\left(\max(0, y - y^\delta - c) + \min(0, y - y^\delta + c)\right) \\ c + \gamma\left\langle -\max(0, y - y^\delta - c) + \min(0, y - y^\delta + c), 1\right\rangle_{L^2(\Omega)} \end{pmatrix}.$$

Since $y \in \mathcal{W}$, the function F is semi-smooth with Newton derivative

$$D_N F(y, c)(\delta y, \delta c) = \begin{pmatrix} \alpha A^* A\delta y + \gamma(\chi_1 + \chi_2)\delta y + \gamma\delta c(-\chi_1 + \chi_2) \\ \gamma\left\langle -\chi_1 + \chi_2, \delta y\right\rangle_{L^2(\Omega)} + \left(1 + \gamma\left\langle \chi_1 + \chi_2, 1\right\rangle_{L^2(\Omega)}\right)\delta c \end{pmatrix}.$$

Superlinear convergence of the semi-smooth Newton method can then be proven analogously to Theorem 15.4.4, using the fact that A is an isomorphism from $\mathcal{W}$ to $\mathcal{W}^*$. Given y_γ, we can then compute $x_\gamma = Ay_\gamma$. Note that due to the linearity of the operators, we can further reformulate the Newton step in terms of the new iterate (y^{k+1}, c^{k+1}) only:

$$\begin{pmatrix} \alpha A^* A + \gamma(\chi_2 + \chi_1) & \gamma(\chi_2 - \chi_1) \\ \gamma\left\langle \chi_2 - \chi_1, \cdot\right\rangle_{L^2(\Omega)} & 1 + \gamma\left\langle \chi_1 + \chi_2, 1\right\rangle_{L^2(\Omega)} \end{pmatrix}\begin{pmatrix} y^{k+1} \\ c^{k+1} \end{pmatrix} = \begin{pmatrix} \gamma(\chi_2 + \chi_1)y^\delta \\ \gamma(\chi_2 - \chi_1)y^\delta \end{pmatrix}.$$

Remark 15.4.7. The presented approach can also be applied to the Morozov regularization

$$\min_{x \in \mathcal{X}} \frac{1}{2}\|x\|_{\mathcal{X}}^2 \quad \text{subject to} \quad \|Kx - y^\delta\|_{L^\infty(\Omega)} \leqslant \delta$$

by fixing $c = \delta$ in the above derivations. Applying the same Moreau–Yosida regularization as above yields the optimality conditions $F(x_\gamma) = 0$ for $F : \mathcal{X} \to \mathcal{X}^*$,

$$F(x) = \alpha j(x) + \gamma K^* \left(\max(0, Kx - y^\delta - \delta) + \min(0, Kx - y^\delta + \delta) \right),$$

with Newton derivative

$$D_N F(x)\delta x = (\alpha j'(x) + \gamma K^*(\chi_1 + \chi_2)K)\, \delta x.$$

Well-posedness, convergence as $\gamma \to \infty$ and superlinear convergence of the semi-smooth Newton method can be shown as for the Tikhonov regularization (with obvious simplifications).

15.5 NUMERICAL EXAMPLES

In this section, we illustrate the effectiveness of the L^∞ fitting approach as well as some of its qualitative features by way of one- and two-dimensional model problems. The Matlab implementation for both examples can be downloaded as http://www.uni-graz.at/~clason/codes/linffitting.zip. All numerical tests were performed with Matlab (R2011b) on a single core of a 3.4 GHz workstation with 16 GByte of RAM.

15.5.1 INVERSE HEAT CONDUCTION PROBLEM

We first consider as a standard benchmark example an inverse heat conduction problem, posed as a Volterra integral equation of the first kind (problem `heat` in [Hansen 2007]).[1] Here, $\Omega = (0, 1)$, $\mathcal{X} = L^2(\Omega)$, and $(Kx)(t) = \int_0^t k(s, t)x(s)\, ds$. Hence, K is a compact linear operator from $L^2(\Omega)$ to $L^\infty(\Omega)$. The kernel $k(s, t)$ and the exact solution $x^\dagger(t)$ are given by

$$k(s,t) = \frac{(s-t)^{-\frac{3}{2}}}{2\sqrt{\pi}} e^{-\frac{1}{4(s-t)}}, \qquad x^\dagger(t) = \begin{cases} 75t^2 & 0 \leqslant t \leqslant \frac{1}{10}, \\ \frac{3}{4} + (20t - 2)(3 - 20t) & \frac{1}{10} < t \leqslant \frac{3}{20}, \\ \frac{3}{4} e^{-2(20t-3)} & \frac{3}{20} < t \leqslant \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

The noisy data are generated by setting

$$y^\delta(t) = Kx^\dagger(t) + \xi(t), \qquad t \in (0, 1),$$

where $\xi(t)$ is a uniformly distributed random value in the range $[-d\, y_{\max}, d\, y_{\max}]$ for a noise parameter $d > 0$ and $y_{\max} = \|Kx^\dagger\|_\infty$.

[1] MATLAB code (version 4.1) and documentation is available from http://www2.imm.dtu.dk/~pch/Regutools.

For the numerical solution of the inverse problem $Kx = y^\delta$, we apply Algorithm 15.1 (with $\tau = 10, k^* = 10, (x^0, c^0) = (0,0), \gamma^0 = 1$ and $\gamma^* = 10^{12}$) and discretize the integral equation using collocation and the mid-point rule at $n = 300$ points (unless stated otherwise). The parameters in the fixed-point iteration for the automatic parameter choice are set to $\alpha_0 = 0.1$ and $\sigma = 0.008$. The fixed-point iteration is terminated if the relative change in α is less than 10^{-3} or after 20 iterations.

A typical realization of noisy data is displayed in Figure 15.1a for $d = 0.3$ and Figure 15.1c for $d = 0.6$. The fixed-point iteration (15.3.2) converged after 6 (4) iterations for $d = 0.3$ ($d = 0.6$), and yielded the values 6.50×10^{-3} (1.65×10^{-2}) for the regularization parameter α. The respective reconstructions x_α are shown in Figures 15.1b and 15.1d. To measure the accuracy of the solution x_α quantitatively, we compute the L^2-error $e = \|x_\alpha - x^\dagger\|_{L^2}$, which is 2.48×10^{-2} for $d = 0.3$ and 7.56×10^{-2} for $d = 0.6$. For comparison, we also show the solution to the L^2 data fitting problem, where the parameter α has been chosen to give the smallest L^2 error. Clearly, the L^2 reconstructions are significantly less accurate than their L^∞ counterparts, especially at the "tail".

The performance of the automatic parameter choice is further illustrated in Table 15.1, which compares the balancing parameter α_b with the "optimal", sampling-based parameter α_o for different noise levels. This parameter is obtained by sampling each interval $[0.1\alpha_b, \alpha_b]$ and $[\alpha_b, 10\alpha_b]$ uniformly with 51 parameters and taking as α_o the one with smallest L^2-error $e_o = \|x_{\alpha_o} - x^\dagger\|_{L^2}$. Presented in each case are the mean and standard deviation over ten different noise realizations for a given noise level. Both the regularization parameters and the reconstruction errors agree closely, and the noise level is well estimated by the optimal L^∞ bound c_b. Table 15.1 also illustrates the robustness of the L^∞ data fitting, since the reconstruction error does not significantly increase with increasing noise level. This can be attributed to the fact that the structural properties of the noise (e.g., sign changes of the noise, which is neither more nor less likely for increasing d) is more important than the magnitude.

Finally, we address the performance of the semi-smooth Newton method. Table 15.2 shows the convergence history of the Newton iteration (in terms of the number of changed points in the active sets $\mathcal{A}_k^+, \mathcal{A}_k^-$ after each iteration) for $d = 0.3$, fixed α computed by the balancing principle and fixed $\gamma = 10^2$, corroborating both the local superlinear convergence (Theorem 15.4.4) and the finite termination property (Proposition 15.4.5). The behavior of the full continuation strategy is similarly illustrated in Table 15.3, demonstrating that a feasible solution (i.e., one attaining the L^∞ bound) is reached at $\gamma = 10^6$ with comparative computational effort.

15.5.2 INVERSE SOURCE PROBLEM IN 2D

Motivated by the problem of detecting the source of a contaminant using distributed and quantized measurements from a sensor network, we consider the inverse source problem for

Table 15.1: Comparison of automatic parameter choice (estimated noise level c_b, parameter α_b, reconstruction error e_b) with sampling based optimal choice (α_o, e_o) for different noise parameters d and noise levels δ (shown are the mean and standard deviation over ten different noise realizations)

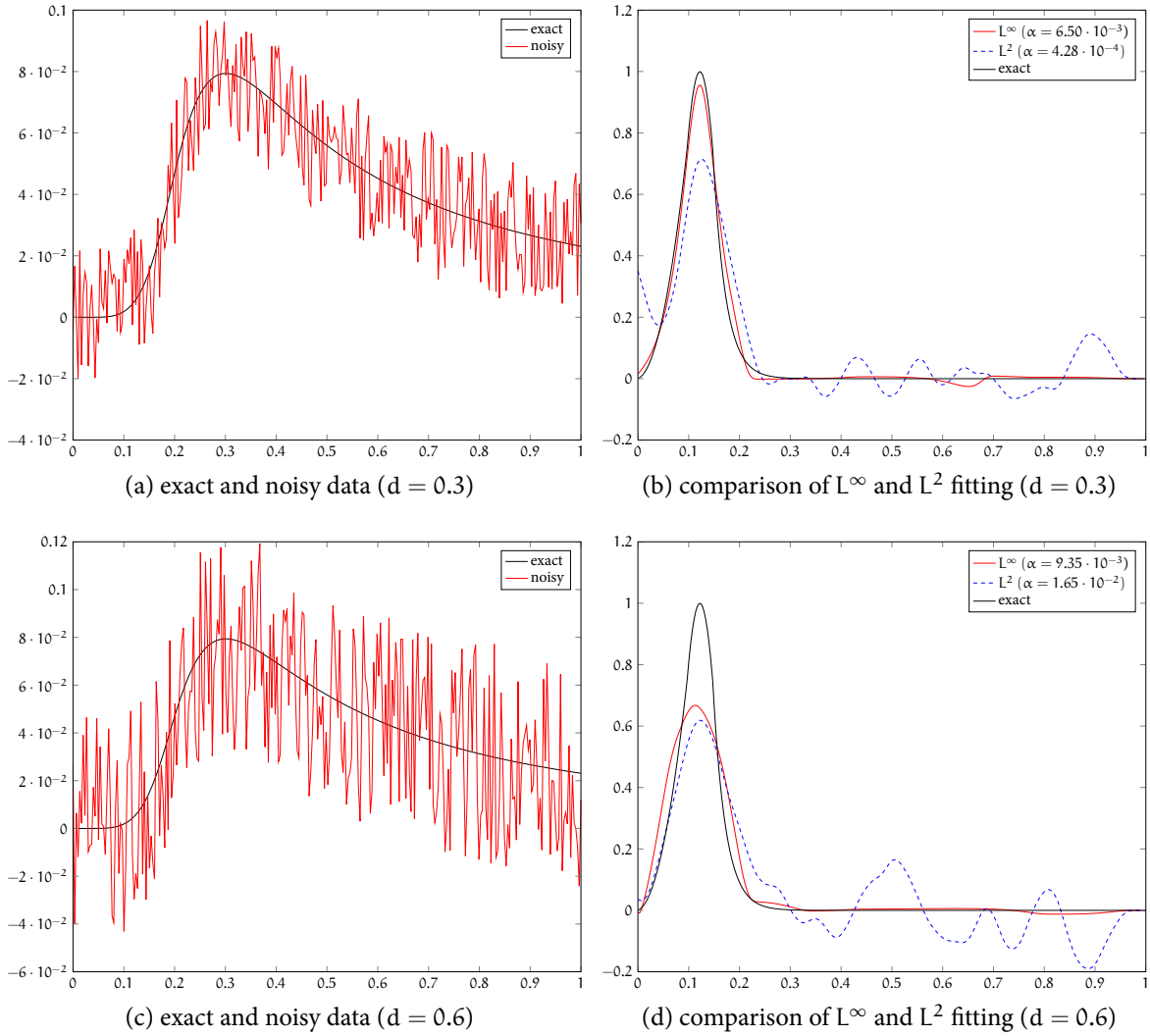
d	δ		c_b		α_b		e_b		α_o		e_o	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
0.1	7.90e-3	3.7e-5	7.79e-3	1.6e-4	2.26e-3	1.3e-4	4.06e-2	1.7e-2	2.38e-3	7.9e-4	3.76e-2	1.9e-2
0.2	1.58e-2	1.0e-4	1.57e-2	5.1e-4	4.77e-3	4.7e-4	5.06e-2	8.3e-3	4.58e-3	1.8e-3	4.47e-2	1.2e-2
0.3	2.37e-2	5.4e-5	2.35e-2	2.8e-4	7.56e-3	7.0e-4	6.81e-2	1.6e-2	5.62e-3	3.9e-3	6.48e-2	1.9e-2
0.4	3.16e-2	8.8e-5	3.10e-2	1.9e-4	9.66e-3	8.2e-4	6.06e-2	2.1e-2	8.49e-3	5.6e-3	5.58e-2	2.4e-2
0.5	3.96e-2	7.1e-5	3.90e-2	5.7e-4	1.25e-2	1.2e-3	6.29e-2	2.0e-2	1.15e-2	7.0e-3	6.21e-2	2.0e-2
0.6	4.75e-2	1.2e-4	4.67e-2	4.5e-4	1.43e-2	1.8e-3	6.43e-2	3.5e-2	1.39e-2	9.4e-3	5.80e-2	2.9e-2
0.7	5.53e-2	2.1e-4	5.47e-2	4.9e-4	1.85e-2	3.4e-3	7.77e-2	3.2e-2	2.85e-2	2.4e-2	7.27e-2	3.0e-2
0.8	6.34e-2	1.8e-4	6.27e-2	8.1e-4	2.30e-2	4.6e-3	8.81e-2	3.8e-2	2.22e-2	1.2e-2	8.72e-2	3.8e-2
0.9	7.13e-2	1.3e-4	7.03e-2	6.4e-4	2.60e-2	4.2e-3	1.15e-1	3.0e-2	3.07e-2	2.2e-2	1.13e-1	3.1e-2

Table 15.2: Convergence behavior of the semi-smooth Newton method for fixed $\gamma = 10^2$ (shown are the number of points $n(k)$ that changed in the active sets after iteration k)

k	1	2	3	4	5	6	7	8
$n(k)$	144	83	39	19	8	1	1	0

Table 15.3: Convergence behavior of the semi-smooth Newton method with continuation (shown are the number of points $n(k)$ that changed in the active sets after iteration k)

γ	1e0				1e1				1e2				1e3					1e4				1e5		1e6
k	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	1
$n(k)$	115	25	7	0	91	35	5	0	38	15	5	0	15	8	4	2	0	5	3	1	0	1	0	0



(a) exact and noisy data ($d = 0.3$)

(b) comparison of L^∞ and L^2 fitting ($d = 0.3$)

(c) exact and noisy data ($d = 0.6$)

(d) comparison of L^∞ and L^2 fitting ($d = 0.6$)

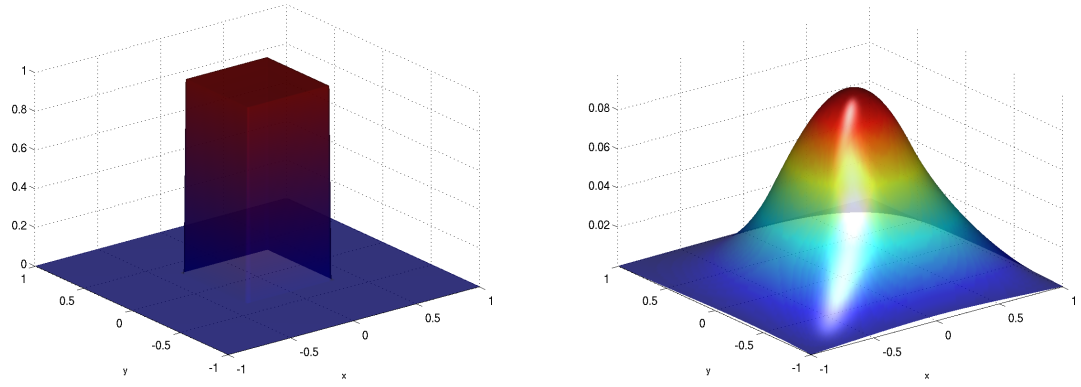
Figure 15.1: Results for inverse heat conduction problem

an elliptic partial differential operator on the domain $\Omega \subset \mathbb{R}^2$ with homogeneous Dirichlet boundary conditions, i.e., $K = A^{-1}$, $\mathcal{X} = L^2(\Omega)$, where

$$Ay = -a\Delta y + \langle b, \nabla y \rangle_{L^2(\Omega)} + fy$$

for $a \in C^{0,r}(\Omega)$ with $r > 0$ and $a \geqslant a_0 > 0$ pointwise, $b \in C^{0,r}(\Omega)^2$, $f \in L^\infty(\Omega)$ with $f - \nabla \cdot b \geqslant 0$ pointwise. This guarantees (for Ω smooth or a parallelepiped) that A is an isomorphism from $\mathcal{W} = H_0^1(\Omega) \cap H^2(\Omega)$ to $\mathcal{W}^*$ and hence that $y \in C^0(\overline{\omega})$. Here, we choose $a = 1$, $b = (-2, 0)^T$, $f = 0$, and $\Omega = [0, 1]^2$. The exact solution is given by

$$x^\dagger(t_1, t_2) = \begin{cases} 1 & \text{if } |t_1| \leqslant \frac{1}{3} \text{ and } |t_2| \leqslant \frac{1}{3}, \\ 0 & \text{otherwise,} \end{cases}$$



(a) exact solution $x^\dagger$ (b) exact data $y^\dagger$

Figure 15.2: Two-dimensional test problem: exact solution $x^\dagger$ and data $y^\dagger$

see Figure 15.2a. The exact data $y^\dagger = A^{-1}x^\dagger$ are shown in Figure 15.2b.

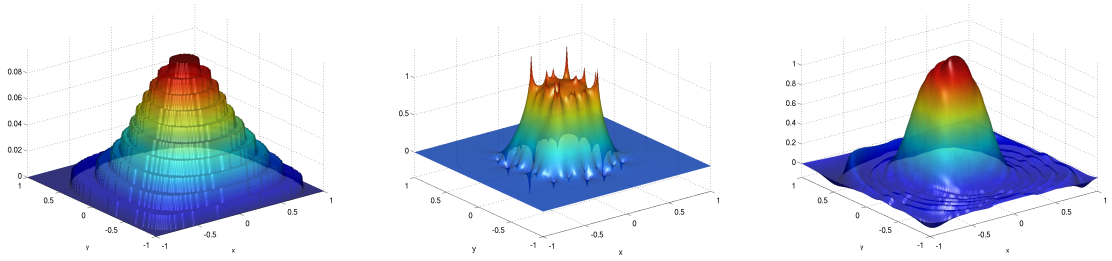
For the numerical solution of the inverse problem $u = A^{-1}y^\delta$, we apply the reformulated algorithm according to Remark 15.4.6, and discretize the differential operators using standard finite differences on a uniform mesh of size 128×128. The parameters in the semi-smooth Newton method with continuation and the fixed-point iteration are identical to the one-dimensional case.

The first example considers data subject to (deterministic) quantization errors, where we set

$$y^\delta(t) = y_s \left[\frac{y^\dagger(t)}{y_s}\right], \qquad y_s = n_b^{-1}\left(\max_{t\in\Omega}\left(y^\dagger(t)\right) - \min_{t\in\Omega}\left(y^\dagger(t)\right)\right),$$

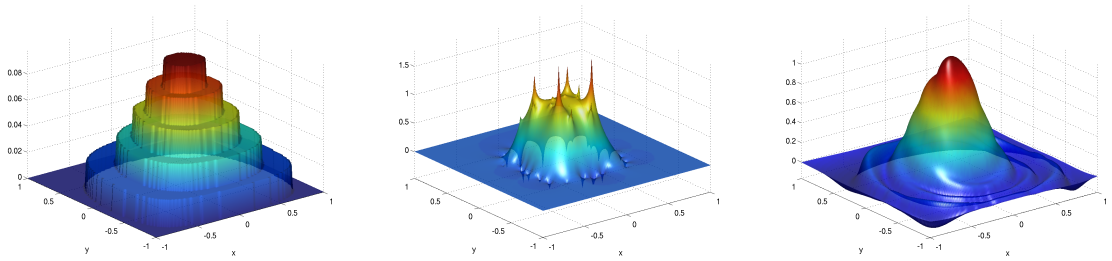
with n_b denoting the number of bins and $[s]$ denoting the nearest integer to $s \in \mathbb{R}$ (i.e., the data are rounded to n_b discrete equispaced values, see Figure 15.3a for $n_b = 10$ and Figure 15.4a for $n_b = 5$). Again we compare the solutions to the L^∞ fitting problem (where the regularization parameter is chosen using the fixed-point iteration) with reconstructions obtained from standard L^2 fitting (where the parameter is exhaustively selected to yield the lowest L^2 error) in Figures 15.3b, 15.3c and 15.4b, 15.4c. The difference in reconstruction artifacts can be observed clearly: The L^∞ artifacts are strongly localized and impulse-like, whereas the L^2 reconstruction shows typical ringing. In particular, the support of the exact solution $x^\dagger$ is accurately captured by the L^∞ reconstruction, whereas the L^2 reconstruction is non-zero everywhere.

The second example serves as a "best-case" noise for L^∞ fitting. Based on our observation in the one-dimensional case, we conjecture that the reconstruction error is largest in regions where the sign of the noise does not change. We therefore choose as additive "noise" a checkerboard



(a) data y^δ (b) L^∞ reconstruction x_α (c) L^2 reconstruction x_α

Figure 15.3: Reconstructions from quantized data ($n_b = 10$), comparing L^∞ and L^2 fitting



(a) data y^δ (b) L^∞ reconstruction x_α (c) L^2 reconstruction x_α

Figure 15.4: Reconstructions from quantized data ($n_b = 5$), comparing L^∞ and L^2 fitting

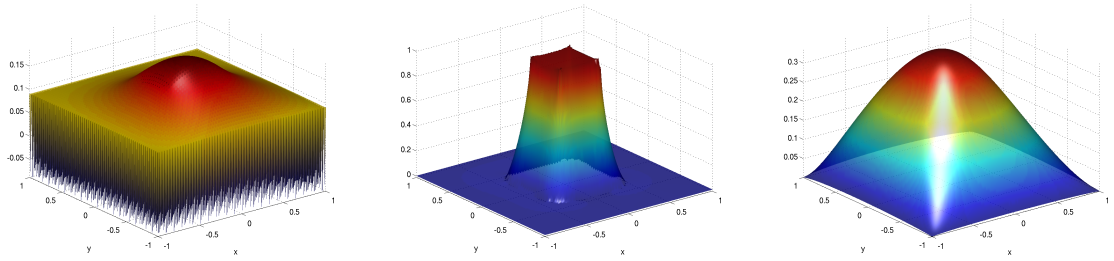
pattern on the discrete mesh of constant magnitude and alternating sign. Specifically, let $t_{ij} = (t_{1,i}, t_{2,j})$, $1 \leqslant i, j \leqslant 128$, be the grid points of the uniform mesh and set

$$y^\delta(t_{ij}) = y^\dagger(t_{ij}) + (-1)^{i+j} d \|y^\dagger\|_\infty$$

for a noise parameter $d > 0$. For data with $d = 0.9$ (Figure 15.5a), the L^∞ reconstruction is able to accurately capture support and shape of the true solution, whereas the L^2 reconstruction is far from the target. The robustness of L^∞ fitting in this case is further illustrated by Table 15.4, where it can be seen that the reconstruction error is virtually independent of the magnitude of the checkerboard noise, in contrast to L^2 fitting.

Table 15.4: Comparison of reconstruction errors for L^∞ fitting (e_∞) and L^2 fitting (e_2) for checkerboard noise of different magnitude d

d	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
e_∞	0.10961	0.1096	0.10961	0.10959	0.10963	0.10962	0.10961	0.1096	0.1096
e_2	0.15625	0.19176	0.22549	0.25568	0.26619	0.26829	0.26978	0.27086	0.27173



(a) data y^δ (b) L^∞ reconstruction x_α (c) L^2 reconstruction x_α

Figure 15.5: Reconstructions from checkerboard noise ($d = 0.9$), comparing L^∞ and L^2 fitting

15.6 CONCLUSION

For measurements subject to uniformly distributed noise, such as arising from statistical models of quantization errors, L^∞ fitting is more robust than standard L^2 fitting. The non-differentiability can be addressed by introducing a Moreau–Yosida regularization together with a continuation scheme, which allows application of a superlinearly convergent semi-smooth Newton method. The regularization parameter can be chosen automatically using a heuristic choice rule that does not require knowledge of the noise level. This approach is useful for a wide variety of linear inverse problems.

For nonlinear problems, the extension would be straightforward (subject to a usual nonlinearity and second order condition, see [Clason and Jin 2012]). By combining the methods of the current work with those of [Clason and Jin 2012], Tikhonov functionals of L^∞-L^1 type (i.e., L^∞ fitting with regularization terms of L^1 type, also known as "Dantzig selector" [Candes and Tao 2007]) could be treated. Finally, a stochastic analogue of the considered uniform noise models in function spaces would be of great interest.

ACKNOWLEDGMENTS

Part of this work was completed while visiting the Isaac Newton Institute for Mathematical Sciences, and the author thanks the institute for the hospitality. Support by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences") is gratefully acknowledged.

Part V

APPLICATIONS IN BIOMEDICAL IMAGING

A DETERMINISTIC APPROACH TO THE ADAPTED OPTODE PLACEMENT FOR ILLUMINATION OF HIGHLY SCATTERING TISSUE

16

ABSTRACT

A novel approach is presented for computing optode placements that are adapted to specific geometries and tissue characteristics, e.g., in optical tomography and photodynamic cancer therapy. The method is based on optimal control techniques together with a sparsity-promoting penalty that favors pointwise solutions, yielding both locations and magnitudes of light sources. In contrast to current discrete approaches, the need for specifying an initial set of candidate configurations as well as the exponential increase in complexity with the number of optodes are avoided. This is demonstrated with computational examples from photodynamic therapy.

16.1 INTRODUCTION

In biological and medical sciences, diagnostic and therapeutic instrumentation based on visible, near-infrared or near-ultraviolet light (referred to as optical imaging) is of special interest as it is often non-invasive, the cost for the equipment is moderate, and data acquisition is usually fast (compared to, e.g., MRI). However, the high scattering coefficient of biological tissues in the visible spectrum can be a limiting factor for this technique as it causes photons to propagate in a non-deterministic manner. This in turn makes measurements of superficial structures and the selective illumination of deeper regions significantly more complicated. Due to the stochastic nature of the photon paths, the optimal placement of the optodes is not trivial except for simple regular geometries such as cylinders or spheres.

In recent years, there has thus been increased interest in computation-driven optimization of optical hardware operating in strongly scattering tissues. For example, [Culver et al. 2001] used a singular value analysis (SVA) of the sensitivity matrix (there called "weight matrix") to compare measurement setups which differed either in the optode spacing or the measurement type (reflectance vs. transmittance setup). In [Xu, Dehghani, et al. 2003], two optode configurations for a hybrid MRI/DOT measurement device were compared, and the number of singular values above a certain threshold was used as a quality criterion. The work in [Graves et al. 2004] applied the SVA approach to assess different fluorescence tomography setups in two dimensions, while [Lasser and Ntziachristos 2007] performed comparisons of three-dimensional setups.

A similar problem is faced in photodynamic therapy (PDT) [Dolmans, Fukumura, and Jain 2003], which is used for dermatological and oncological treatments (e.g., esophageal cancers, especially at a stage where surgical intervention is not indicated). It appears attractive to extend PDT to other carcinomas on epi- or endothelial surfaces, and clinical trials are carried out for, e.g., cervix carcinomas. Some other potential candidates, like mesotheliomas of the thoracic cavity (which are typically difficult to treat), represent a special challenge. A major problem is the design of an appropriate light applicator. In the esophageal cavity, which has a simple geometry, the laser light can be readily applied using a cylindrical scattering device. In contrast, the geometry of the intrathoracic cavity is very complex. The application of a standard esophageal applicator is not recommendable because its area of treatment is small and curved and therefore the illumination would be inhomogeneous. Homogeneity of the irradiation is, however, a crucial design criterion, as inhomogeneities can lead to locally ineffective treatment on the one hand and local overdoses on the other hand. This can have serious consequences, including lethal overdoses [Schouwink and Baas 2004].

In the past, some designs for flexible light diffusers have been proposed, where a typical solution is to use cylindrical or spherical diffusers in a bag filled with a scattering medium. These are applicable after pneumonectomia, if blood accumulations at the surface of the bags are prevented by continuous rinsing [Dwyer et al. 2000; Friedberg et al. 2003; Krueger et al. 2003; Baas et al. 1997]. Some regions like the sinus diaphragmaticus are difficult to access and have to be illuminated separately. This can be achieved with wedge-shaped illuminators [van Veen et al. 2001]; however, positioning of these illuminators and homogenization of the illumination is difficult. Another approach is to fill the thorax with a biologically non-hazardous scattering medium (e.g., with intralipid), which can also be used for rinsing to avoid blood accumulation. Typically, a spherical diffuser is used for illumination. However, it is difficult to control the dose rate with this approach. This method has also been applied to cases where lung tissue was not resected [Friedberg et al. 2003]. In general, both methods try to achieve homogeneous fluence using real-time dosimetry and manual repositioning of the light diffuser. An interesting alternative consists in textile-based diffusers, where special optical fibers are integrated in a textile. These diffusers are very flexible but suffer from inhomogeneous illumination and a low transmission rate [Selm et al. 2007; Rothmaier et al. 2008]. Recently so called "light blankets" with arrays of cylindrical diffusers [Hu, Wang, and Zhu 2009] or a spirally-wound side-glowing fiber [Hu, Wang, and Zhu 2010] embedded in a bag

filled with intralipid were presented. They are easy to fabricate but still show inhomogeneities, especially at the corners. Due to the need for a homogeneous fluence rate, it is of great interest to optimize the placement of these fibers to obtain improved illumination using minimal energy.

The purpose of this work is thus to present a general approach to compute adapted optode locations for different geometries, tissue types, and applications. The method is based on considering this task as an optimal control problem for a partial differential equation describing the diffusion of photons in a strongly scattering medium, where the location of optodes are modeled as a continuous "source field". The crucial step – first proposed in [Stadler 2009] – is to include a penalty term that favors pointwise solutions. In this way, both locations and magnitudes of the light sources to be placed are obtained in a single step. The main advantage of this approach over previously published – discrete – methods is that no initial maximal or minimal configuration needs to be specified (although an allowable region can be enforced), and that a combinatorial problem with exponential complexity is avoided. In addition, the algorithm is not based on stochastic (e.g., Monte Carlo) methods but is fully deterministic, which eases the verification of the outcome significantly. Finally, the approach is flexible and can incorporate a wide variety of objective criteria (e.g., photon flux over a given boundary section) by changing the target functional. The proposed approach is demonstrated in the context of optimizing the illumination pattern in the photodynamic treatment of intrathoracic cancer.

The rest of this work is organized as follows. In section 16.2.1, the specific mathematical model for photon diffusion is given. Section 16.2.2 presents our optimal control framework for optode placement, whose numerical solution is described in section 16.2.3. The setup of the numerical experiments used for demonstrating our approach can be found in section 16.3, and the results are presented in section 16.4. A discussion of the proposed method in section 16.5 concludes the work.

16.2 THEORY

During photodynamic treatment of cancer, a photosensitizer such as Photofrin is injected intravenously. Afterwards, the cancerogeneous site is illuminated with red to near-infrared light from sources in a diffuser which is applied directly on the region of interest, i.e., in the intrathoracic cavity. The absorption of energy by the photo-activable drug leads to the formation of cytotoxic singlet oxygen, which destroys cancer cells selectively. The challenge is to homogenize the light intensity as both under- and overexposure can lead to ineffective treatment [Henderson et al. 2000].

16.2.1 MATHEMATICAL MODEL

We use the diffusion approximation of the radiative transfer equation to model the steady state of light propagation in a scattering medium [Arridge 1999]. This leads to a stationary elliptic partial differential equation for the photon distribution $\varphi \in H^1(\Omega)$,

$$\begin{cases} -\nabla \cdot (\kappa(x)\nabla\varphi(x)) + \mu_a(x)\varphi(x) = q(x) & \text{in } \Omega, \\ \kappa(x)\,\vec{n}(x) \cdot \nabla\varphi(x) + \rho\varphi(x) = 0 & \text{on } \Gamma. \end{cases} \tag{16.2.1}$$

The geometry of the object is given by the domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ being the number of spatial dimensions, with boundary Γ whose outward normal vector is denoted by $\vec{n}$. The medium is characterized by the absorption coefficient μ_a, the reduced scattering coefficient μ_s', and the diffusion coefficient $\kappa = \left[\frac{1}{d}(\mu_a + \mu_s')\right]^{-1}$. The coefficient ρ models the reflection of a part of the photons at the boundary due to a mismatch in the index of refraction. Finally, the source term q models the light emission of the embedded optodes.

For the optimal control approach, we also require the solution $p \in H^1(\Omega)$ of the adjoint equation

$$\begin{cases} -\nabla \cdot (\kappa(x)\nabla p(x)) + \mu_a(x)p(x) = f(x) & \text{in } \Omega, \\ \kappa(x)\,\vec{n}(x) \cdot \nabla p(x) + \rho p(x) = 0 & \text{on } \Gamma \end{cases} \tag{16.2.2}$$

for given $f \in L^2(\Omega)$. Both equations should be understood in the weak sense.

16.2.2 OPTODE PLACEMENT OPTIMIZATION

Since optodes act as discrete light sources, the source term can be modeled as $q(x) = \sum_{j=1}^{N} q_j \delta(x - x_j)$ for $q_j \in \mathbb{R}_+$ and $x_j \in \Omega$, $1 \leqslant j \leqslant N$, where δ denotes the Dirac distribution (i.e., $\int f \, d\delta(x) = f(0)$ for all continuous functions f). A straightforward approach for optimizing the placement of the optodes (as was done, e.g., in [Freiberger, Clason, and Scharfetter 2010]) would identify a set of $M \gg N$ possible optode locations $x_1, \dots, x_M$ and chose the best N locations such that a certain performance criterion $J(q)$ is minimized. The corresponding optimal source magnitudes q_j would then be computed in a second step.

To avoid the combinatorial complexity of this discrete approach, instead of specifying the optode locations beforehand, we optimize the (distributed) source term q directly while adding a penalty term that promotes *sparsity* of q, i.e., smallness of its support $\{x \in \Omega : q(x) \neq 0\}$. This has the added advantage that the number N of optodes need not be specified in advance. Since we are looking for point sources, this requires searching for q in the space of regular Borel measures (which includes the Dirac distribution). Following [Clason and Kunisch 2011], we are thus led to the optimization problem

$$\min_{q \in \mathcal{M}(\Omega)} J(q) + \alpha \|q\|_{\mathcal{M}},$$

where $\mathcal{M}(\Omega)$ is the space of regular Borel measures, i.e., the dual of the space $C_0(\Omega)$ of continuous functions with compact support on Ω, with norm

$$\|q\|_{\mathcal{M}} = \sup_{\substack{f \in C_0(\Omega) \\ \|f\|_C \leqslant 1}} \int_\Omega f \, dq,$$

which reduces to

$$\|q\|_{\mathcal{M}} = \int_\Omega |q(x)| \, dx = \|q\|_{L^1}$$

for $q \in L^1(\Omega)$. This is related to the well-known fact that L^1 norms promote sparsity in optimization. The penalty parameter α controls the sparsity of the solution: The larger α, the smaller the support of q.

Motivated by the application in PDT, we chose as performance criterion the deviation from a constant illumination z in an observation region $\omega_o \subset \Omega$ such that $J(q) := \frac{1}{2}\|\varphi|_{\omega_o} - z\|^2_{L^2(\omega_o)}$, where $\varphi|_{\omega_o}$ denotes the restriction of φ to ω_o. Due to the linearity of the forward problem, we can take $z = 1\,\mathrm{W\,m^{-2}}$ without loss of generality. After optimization, the magnitude of the resultant sources can be linearly scaled to achieve the required illumination z. In addition, we restrict the possible light source locations to a control region $\omega_q \subset \Omega$, which does not overlap with the observation region ω_o (i.e., $\overline{\omega}_q \cap \overline{\omega}_o = \emptyset$), and enforce non-negativity of the source term q (which represents the optodes). This leads to the following optimization problem:

$$\min_{\varphi \in H^1(\Omega), q \in \mathcal{M}(\omega_q)} \frac{1}{2}\|\varphi|_{\omega_o} - z\|^2_{L^2(\omega_o)} + \alpha\|q\|_{\mathcal{M}(\omega_q)} \quad \text{subject to (16.2.1) and } q \geqslant 0. \tag{16.2.3}$$

It was shown in [Clason and Kunisch 2012] that this problem has a solution $q^* \in \mathcal{M}(\omega_q)$, which can be approximated by a sequence of functions $q_\gamma \in L^2(\omega_q)$ for $\gamma \to \infty$ satisfying

$$q_\gamma + \gamma \min(0, p_\gamma + \alpha) = 0, \tag{16.2.4}$$

where p_γ is the solution of (16.2.2) with right hand side $f := \varphi_\gamma - z$ and φ_γ is the solution of (16.2.1) with right hand side q_γ. Equation (16.2.4) can be solved using a semismooth Newton method which is superlinearly convergent; see [Clason and Kunisch 2012]. To globalize the Newton method and closely approximate the solution q^* of (16.2.3), we use a continuation scheme in γ where we iteratively solve the problem for an increasing sequence γ_n, using the previous solution as initial guess.

16.2.3 FINITE ELEMENT DISCRETIZATION

The discretization needs to account for the fact that the functions q_γ converge to measures as γ increases. We therefore employ the finite element discretization proposed in [Casas,

Clason, and Kunisch 2012], where the photon density φ_γ and the adjoint variable p_γ are discretized using piecewise linear elements, while the source term q_γ is discretized using linear combinations of Dirac distributions centered at the vertices x_i, $1 \leqslant i \leqslant N(T)$, of the triangulation T:

$$q_\gamma = \sum_{i=1}^{N(T)} q_i \delta(x - x_i).$$

In practice, the number of nodes $N(T)$ will be determined by the need to resolve the geometry of the domain and the required accuracy of the solution of the forward model (16.2.1). Although further refinement of the triangulation increases the number of possible optode locations, the sparsity-promoting property of the minimized functional discourages placing additional optodes. In fact, it was shown in [Casas, Clason, and Kunisch 2012] that for a given discretization of the forward model, the computed sources (for $\gamma \to \infty$) are optimal among all (non-discretized) measures.

Since the linear finite element basis functions form a nodal basis, the right hand side in the weak formulation of (16.2.1) for a piecewise linear basis function e_j becomes

$$\langle q_\gamma, e_j \rangle = \sum_{i=1}^{N(T)} q_i \langle \delta(x - x_i), e_j \rangle = q_j,$$

i.e., the mass matrix is the identity. Introducing the stiffness matrix A corresponding to (16.2.1) and the observation mass matrix M_o with entries $M_{ij} = \int_{\omega_o} e_i e_j \, dx$, we obtain the discrete optimality system

$$\begin{cases} A\varphi_\gamma - q_\gamma = 0, \\ -M_o \varphi_\gamma + A^T p_\gamma = -M_o z, \\ q_\gamma + \gamma \min(0, p_\gamma|_{\omega_q} + \alpha) = 0, \end{cases}$$

Eliminating q_γ using the last equation and applying a semismooth Newton method, cf. [Clason and Kunisch 2012], we have to solve for (φ^{k+1}, p^{k+1}) the block system

$$\begin{pmatrix} A & D_k \\ -M_o & A \end{pmatrix} \begin{pmatrix} \varphi^{k+1} \\ p^{k+1} \end{pmatrix} = \begin{pmatrix} -\alpha d^k \\ -M_o z \end{pmatrix}, \tag{16.2.5}$$

where D_k is a diagonal matrix with the entries of the vector d^k,

$$d_j^k = \begin{cases} \gamma & \text{if } (p^k|_{\omega_q})_j < -\alpha, \\ 0 & \text{else,} \end{cases} \tag{16.2.6}$$

on the diagonal. It can be shown that the semismooth Newton method has converged once $d^{k+1} = d^k$ holds. After the final p^k has been computed, the corresponding control can be obtained from (16.2.4).

The complete procedure is given in Algorithm 16.1.

Algorithm 16.1 Semismooth Newton method with continuation

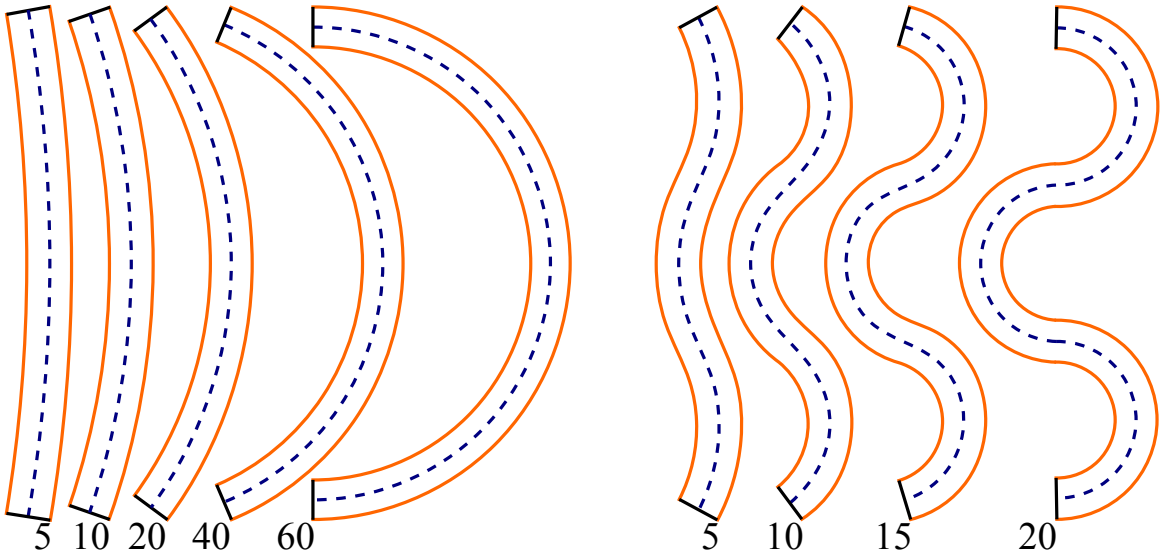
for $m = 1, \dots, m^*$ **do**
  set $\gamma = 2^{(m-1)}$, $\varphi^0 = p^0 = d^0 = 0$
  for $k = 0, \dots, k^*$ **do**
    solve (16.2.5) for φ^{k+1}, p^{k+1}
    compute d^{k+1} from (16.2.6)
    if $d^{k+1} = d^k$ **then**
      set $q^{(m)} = \gamma \min(0, p^{k+1}|_{\omega_q} + \alpha)$
      break
    end if
  end for
end for

16.3 MATERIALS AND METHODS

The optimization algorithm described in section 16.2.2 is implemented in Python using the open source finite element library FEniCS [Logg, Mardal, Wells, et al. 2012]. The parameters in Algorithm 16.1 are set to $m^* = 34$ (such that $\gamma^* \approx 10^{10}$) and $k^* = 20$. To model a textile-based diffuser, the material parameters in (16.2.1) are taken as $\mu_a = 10^{-4}\,\mathrm{mm}^{-1}$, $\mu_s' = 10^{-1}\,\mathrm{mm}^{-1}$, and $\rho = 0.1992$. The influence of the parameter α is illustrated by comparing the results for different values of α specified below.

The meshes for the light diffusers containing the optodes are created with the commercial mesh generator Hypermesh™. To demonstrate the behavior of the optimization algorithm for different geometries, we first consider simple two-dimensional spline models which represent the cross-section of an infinitely long pad. This geometry mimics that of an array of parallel cylindrical diffusers embedded in a scattering substrate. Five single-curved models and four double-curved models with increasing curvature κ were created as shown in Figure 16.1. The dimensions correspond approximately to a width of 10 mm and height of 120 mm. In all cases, the region ω_o in which the illumination should be homogenized are the left and right outer lines (indicated in orange in Figure 16.1). The region ω_q where optodes are allowed to be placed is a single line equidistant from both (indicated by a dashed line in Figure 16.1). The meshes for the single-curved models of curvature $\kappa = 5$, 10, 20, 40, and 60 consist of 61 038, 61 789, 67 160, 80 664, and 105 322 finite elements, respectively. The double-curved models of curvature $\kappa = 5$, 10, 15, and 20 are comprised of 62 349, 70 735, 82 119, and 104 220 finite elements, respectively.

The photodynamic treatment is simulated by embedding the light diffuser model in the intrapleural space of a realistic three-dimensional human thorax model that is constructed from a stack of CT images. The approximate dimensions are: height 100 mm, width 150 mm, thickness 10 mm. The observation region ω_o is defined as the outer and inner surface of the model, and ω_q is an interior manifold equidistant from both (see Figure 16.2; ω_q is indicated



(a) Single-curved models. (b) Double-curved models.

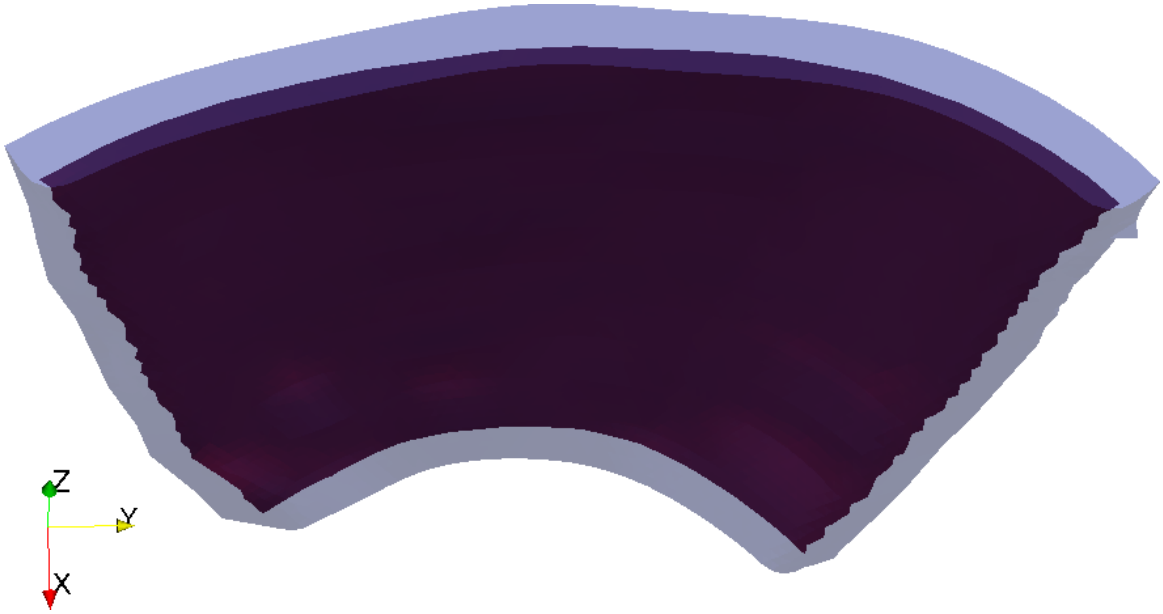
Figure 16.1: Two-dimensional model geometries (numbers denote curvature κ).

in purple). The generated mesh consists of 81 770 elements.

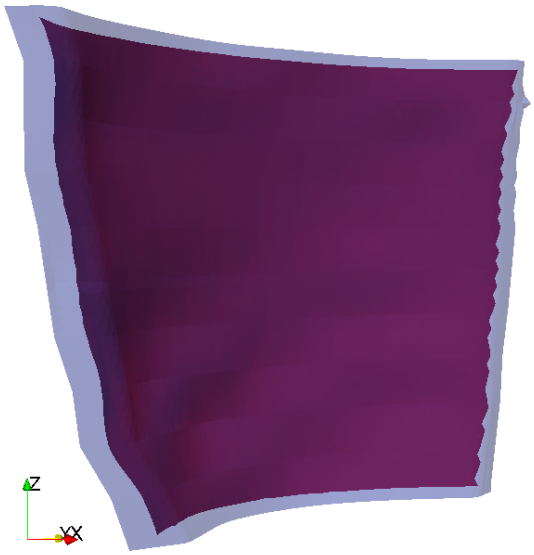
The results are evaluated quantitatively for different values of the sparsity-controlling parameter α. The coefficient of variation c_v of the resultant photon density φ_γ over the observation region ω_o and the number N of sources after the optimization procedure serve as quality measures. For the latter, the nodes in the control region ω_q satisfying $q_\gamma > 10^{-16}$ are counted. We compare the results for $\alpha \in \{0.1, 0.01, 0.001\}$ for the two-dimensional models and $\alpha \in \{0.2, 0.4, \ldots, 1.8\}$ for the three-dimensional model.

16.4 RESULTS

The quantitative results for the two-dimensional geometries are given in Table 16.1 for the single-curved models and in Table 16.2 for the double-curved models. As can be seen by comparing the number of active nodes N with the total number of nodes for each model, the algorithm indeed produces discrete sources that can be used as optode positions. The obtained coefficients of variation c_v indicate that a homogeneous illumination of the desired region is possible at least for $\alpha < 0.1$, demonstrating the feasibility of the proposed approach. The robustness of the algorithm with respect to geometry is illustrated by the fact that the achieved variations do not depend very much on the curvature. It can also be observed how the penalty parameter α determines the tradeoff between the number of active optodes and the homogeneity of the illumination in the region of interest: larger values of α yield fewer optodes but less homogeneous illumination, again independent of curvature.



(a) Top oblique view



(b) Front oblique view

(c) Side oblique view

Figure 16.2: Three-dimensional model. The admissible manifold ω_q for optodes is indicated in purple.

Table 16.1: Results for single-curved models. Shown are the number N of active nodes and the coefficient of variation c_ν of the photon density in the observation domain for different curvatures κ and values of α.

κ	5			10			20			40			60		
α	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
N	22	49	59	18	51	83	15	56	66	20	51	62	22	68	147
c_ν	2.52e-1	1.82e-2	5.40e-3	2.96e-1	2.07e-2	6.17e-3	1.78e-1	1.96e-2	7.95e-3	1.68e-1	2.53e-2	1.09e-2	1.21e-1	2.02e-2	1.57e-2

Table 16.2: Results for double-curved models. Shown are the number N of active nodes and the coefficient of variation c_ν of the photon density in the observation domain for different curvatures κ and values of α.

κ	5			10			15			20		
α	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
N	12	50	134	19	40	148	26	49	60	33	77	130
c_ν	1.65e-1	2.48e-2	1.51e-2	2.03e-1	2.88e-2	2.45e-2	2.24e-1	3.27e-2	2.94e-2	4.92e-1	3.47e-2	3.09e-2

Table 16.3: Results for three-dimensional model. Shown are the number N of active nodes and the coefficient of variation c_ν of the photon density in the observation domain for different values of α.

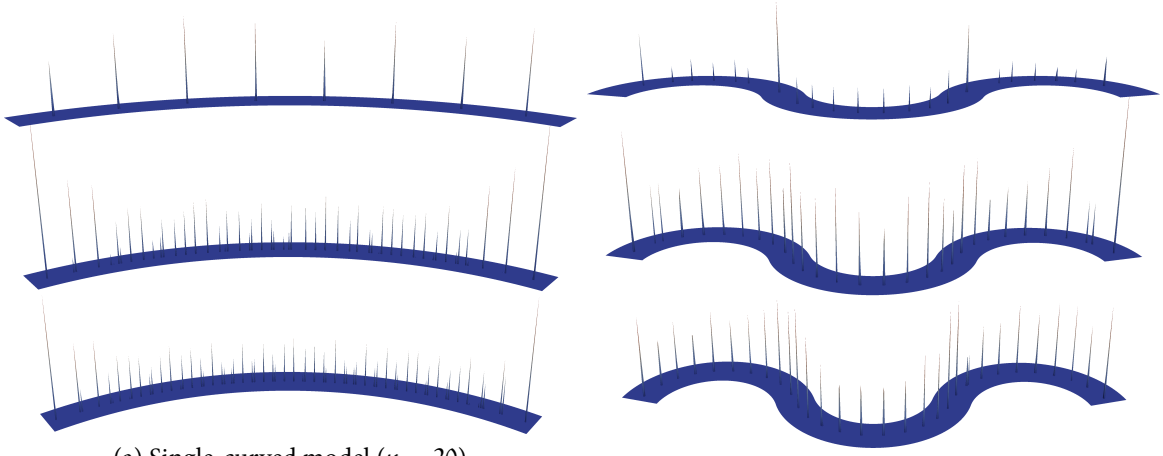
α	1.8	1.6	1.4	1.2	1.0	0.8	0.6	0.4	0.2
N	0	12	150	250	333	409	498	637	884
c_ν	—	1.85e+0	5.64e-1	3.59e-1	2.65e-1	2.04e-1	1.56e-1	1.13e-1	6.72e-2

The qualitative behavior of the computed sources for each value of α is shown in Figure 16.3a and Figure 16.3b for a representative single-curved ($\kappa = 20$) and double-curved model ($\kappa = 15$), respectively, where the relative strength of the sources is coded by height. (Note when comparing Tables 16.1 and 16.2 with Figure 16.3 that neighboring active nodes appear as a single peak and thus can be taken as a single optode.) While for the single-curved model and $\alpha = 0.1$, the distribution of optodes agrees well with the intuitive choice of equally spaced optodes of approximately equal magnitude, the other values indicate that a better illumination can be achieved with stronger sources towards the tips of the model. It should be pointed out that even in the former case, the number of optodes to be distributed is not obvious. For the double-curved models, the results indicate that optodes should be placed preferentially in regions where the curvature changes.

For reference, Figure 16.4 shows the corresponding photon densities φ_γ (in $\mathrm{W\,m^{-2}}$, normalized to unit mean) plotted along part of the observation region (left line in Figure 16.1), illustrating how the parameter α and the model geometry influence the homogeneity of the illumination in this region. As expected, photon fluence shows the most pronounced inhomogeneities close to the borders. In the case of the single curved model, a nearly sinusoidal ripple pattern arises in more than 80 % of the target region, while in the double curved model the ripple is superimposed on a step-profile with the steps located approximately at the zero-crossing points of the curvature. With $\alpha = 0.1$, the peak–peak fluctuations are still around 40 % of the mean value even far away from the borders, which may be considered as unsatisfactory. However, when decreasing alpha to 0.01 or less, the ripple remains within a few percent, which is sufficient, especially when comparing this value to other sources of fluctuations of the irradiation such as local absorption changes by tissue inhomogeneities, bleeding, or inhomogeneities of the distribution of the photosensitizer.

The quantitative results for the three-dimensional model are shown in Table 16.3. For $\alpha = 1.8$, no controls are placed and thus the photon density is zero. This is consistent with the theory, which predicts that there is a threshold value for α above which the optimal control is identically zero; cf. [Casas, Clason, and Kunisch 2012, Proposition 2.2].

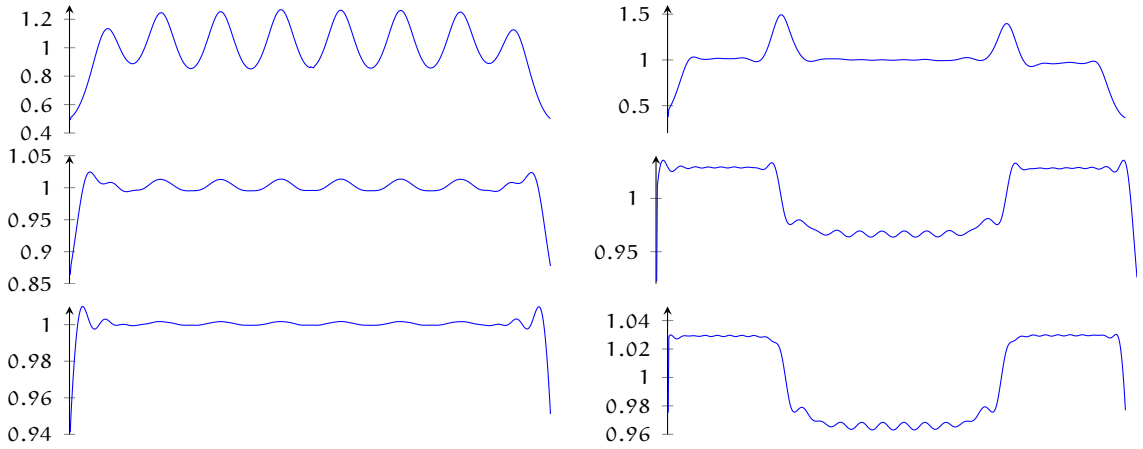
Figure 16.5 shows location and magnitude (color coded) of the computed optodes and the corresponding photon densities (in $\mathrm{W\,m^{-2}}$, normalized to unit mean) for $\alpha = 1.2$, $\alpha = 0.8$, and $\alpha = 0.4$. Due to the nonuniform curvature of the model, a homogeneous illumination is harder to achieve than in the two-dimensional case, especially at the borders of the target region. However, for $\alpha < 1.2$, the inhomogeneities in the interior are usually within 10 %, and the few hot spots of 30 % would still be acceptable. Although of course the specific placement may be difficult to realize in practice, the qualitative distribution can be useful information in the initial design process.



(a) Single-curved model ($\kappa = 20$)

(b) Double-curved model ($\kappa = 15$)

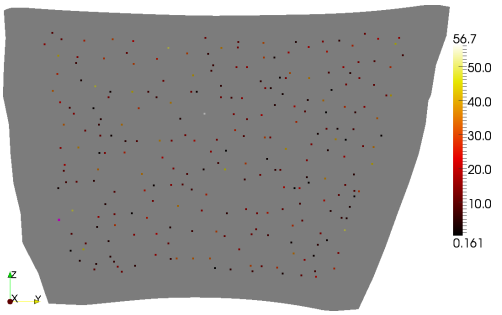
Figure 16.3: Optode positions and relative magnitudes (height-coded) for representative single-curved and double-curved models for three different values of alpha (from top to bottom: $\alpha = 0.1$, $\alpha = 0.01$, $\alpha = 0.001$).



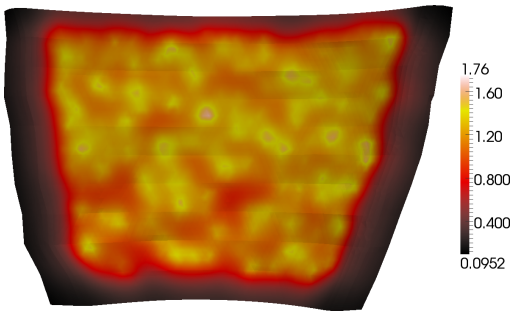
(a) Single-curved model ($\kappa = 20$)

(b) Double-curved model ($\kappa = 15$)

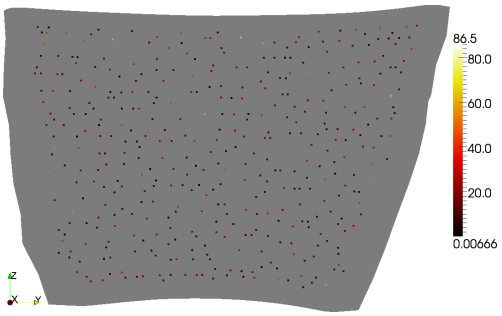
Figure 16.4: Photon densities φ_γ (in W m^{-2}, normalized to unit mean) plotted along part of the observation region (left line in Fig. 16.1) for representative single-curved and double-curved models for three different values of alpha (from top to bottom: $\alpha = 0.1$, $\alpha = 0.01$, $\alpha = 0.001$).



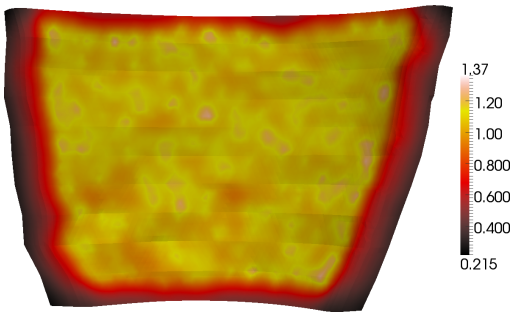
(a) optodes, $\alpha = 1.2$



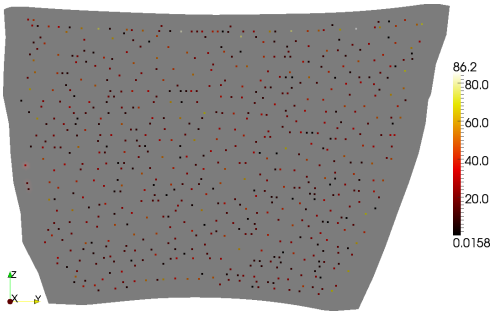
(b) photon density, $\alpha = 1.2$



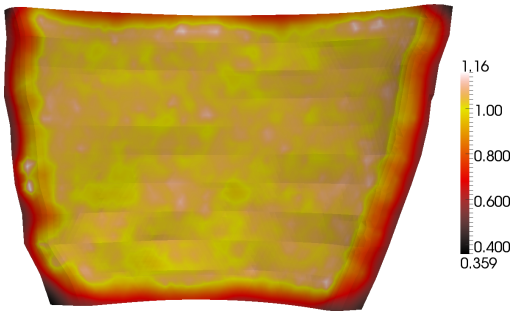
(c) optodes, $\alpha = 0.8$



(d) photon density, $\alpha = 0.8$



(e) optodes, $\alpha = 0.4$



(f) photon density, $\alpha = 0.4$

Figure 16.5: Optode positions and magnitudes (left) and photon densities (right; in W m^{-2}, normalized to unit mean) for the three-dimensional model and three different values of α.

16.5 DISCUSSION

The proposed approach is able to generate reasonable optode configurations adapted to specific geometries, even in situations (such as complex three-dimensional models) where optimal setups are not intuitively obvious. Our method also yields relative strengths of the optodes to be placed, which would otherwise have to be computed in a separate step. Furthermore, the algorithm is deterministic and does not require a-priori knowledge such as an initial set of candidate locations or the number of optodes after optimization, which on the contrary is provided by our approach. The method can be used as a tool during the initial design process to estimate the number of sources required as well as their location and relative strengths.

By formulating the optode placement problem as a continuous optimization problem, the combinatorial complexity inherent in discrete approaches is avoided. This is critical for achieving an efficient optimization technique and – to our knowledge – has not been presented before in the context of diffuse optical imaging. As an example, our Python implementation required about three minutes on a MacBook Pro (2.16 GHz Intel Core2 Duo with 2 GB RAM) for the single-curved model with $\kappa = 5$. Our approach could therefore also be used in an interactive setting, where the engineer will adapt design parameters, such as the optical coefficients of the diffuser, based on the outcome of an optimization run.

While the number of desired optodes is correlated with the penalty parameter α, it is not directly controllable. This drawback is analogous to the problem of finding the "best" regularization parameter in image reconstruction (e.g., for diffuse optical tomography), where typically the determination of the parameter is left to the user or is based on heuristics. Certainly, one could think about finding a good parameter through successive optimization runs, e.g., with decreasing values of α if the user specified an upper bound on the number of optodes.

The achieved results are satisfactory from the mathematical point of view; but of course they should also be discussed in an engineering context. In particular, it may be difficult to place many sources in a more or less irregular pattern. The number of optodes depends on the required uniformity of the surface fluence. A reasonable value in practice would be a CV of 0.05. Table 1 shows CVs (for single-curved pads) below 0.03 for around 50 optodes, but up to 0.25 for less than 25 optodes. The numbers for the double-curved pad are only slightly greater. This means that the between 40 and 50 required sources can be expected. In practice, such a design can be approximated comparatively easily with parallely arranged cylindrical polymer diffusers of sufficiently small radius, which are fed by individual optical fibers. Instead of a fixed grid of diffusers, one can imagine dense bundles of uniformly spaced diffusers where only those close to the optimal positions are connected to the laser source. This would allow a very flexible use and adaptive homogenization of the fluence dependent on the individual anatomical situation (e.g., curvature), which is certainly desirable in the context of a personalized optimization. Such a concept can be realized by using fiberoptic switches with many channels. In three dimensions, the sources may be fiber-coupled spherical diffusers or simply open-ended fibers. Due to the higher number of potential positions (637

for a CV of 0.11, see Tab. 3), the construction of a flexible structure here may be difficult, and pre-fabricated pads that are adapted to a certain anatomical target geometry appear more realistic than a truly adaptive system.

Although a rigorous sensitivity analysis has not yet been carried out, our experience indicates that the computed photon density distributions are relatively robust to small perturbations of the optode locations and magnitudes. Similarly, we did not observe significant changes in the results due to small random perturbations of the optical parameters. This can be attributed to the linearity and the strong diffusivity of the model (16.2.1). Such robustness is very important for practical implementations because it means that the result is not very sensitive to manufacturing tolerances.

One of the main advantages of the optimal control approach is its flexibility. For example, it is straightforward to extend the underlying model to include, e.g., inhomogeneous material properties or to replace the diffusion approximation by a more complicated model such as the radiative transfer equation. It is also possible to consider different objective criteria such as the photon flux through a given (boundary or internal) surface by changing the functional $J(q)$. In principle, the approach can be applied to the problem of optimal experiment design for optical tomography, if the objective $J(q)$ is based on a suitable sensitivity term. However, this extension of our method is subject to future work.

ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences").

PARALLEL IMAGING WITH NONLINEAR RECONSTRUCTION USING VARIATIONAL PENALTIES

17

ABSTRACT

A new approach based on nonlinear inversion for autocalibrated parallel imaging with arbitrary sampling patterns is presented. By extending the iteratively regularized Gauss–Newton method with variational penalties, the improved reconstruction quality obtained from joint estimation of image and coil sensitivities is combined with the superior noise suppression of total variation and total generalized variation regularization. In addition, the proposed approach can lead to enhanced removal of sampling artifacts arising from pseudorandom and radial sampling patterns. This is demonstrated for phantom and in-vivo measurements.

17.1 INTRODUCTION

It was shown recently [Bauer and Kannengiesser 2007; Ying and Sheng 2007; Uecker, Hohage, et al. 2008; Uecker, Karaus, and Frahm 2009] that nonlinear inversion can be applied successfully to image reconstruction of undersampled data from multiple coils, i.e., parallel imaging [Sodickson and Manning 1997; Pruessmann et al. 1999; Griswold et al. 2002]. The joint estimation of images and coil sensitivities, which can be achieved with an iteratively regularized Gauss–Newton (IRGN) method, leads to more accurate estimation of the coil sensitivities, and therefore yields results with improved image quality. It was also demonstrated that this method can be extended to non-Cartesian imaging [Knoll, Clason, Uecker, et al. 2009; Uecker, Zhang, and Frahm 2010]. In particular, radial sampling has the advantage that the sampling pattern automatically leads to an oversampling of the central frequencies of k-space, which eliminates the need to acquire additional reference lines when performing auto-calibrated parallel imaging. Another important characteristic of radial sampling is the fact that aliasing artifacts, which are introduced by undersampling, have a distinctively regular

appearance which is usually different from the image content. For this reason, it is possible to remove these so-called streaking artifacts during reconstruction with the integration of suitable penalties on the reconstructed image, with total variation (TV) proving particularly effective in the radial case [Chang, He, and Fang 2006; Block, Uecker, and Frahm 2007]. Successful application of TV regularization to conventional Cartesian subsampling had been reported as well [Liu et al. 2009]. This interest in total variation-type regularization has been stimulated by the success of compressed sensing [Candes, Romberg, and Tao 2006; Donoho 2006] in MRI [Lustig, Lee, et al. 2005; Lustig, Donoho, and Pauly 2007; Gamper, Boesiger, and Kozerke 2008], which is often used in combination with pseudorandom sampling (i.e., randomly selecting sampling points from a regular Cartesian grid); see also [Xiang 2005] for a related idea. Recently, compressed sensing has also been used for parallel imaging with unkown coil sensitivies by formulating the reconstruction as a low-rank matrix completion problem [Lustig and Pauly 2010a].

The purpose of this work is to demonstrate that the formulation of autocalibrated parallel imaging as a nonlinear inverse problem yields a general framework that allows the joint estimation of coil sensitivities and image content in combination with arbitrary sampling patterns and variational penalties. This is illustrated via the integration of a TV penalty in the IRGN method. The immediate benefit is that TV regularization helps to suppress the noise amplification [Rudin, Osher, and Fatemi 1992] that seriously limits parallel imaging with conventional methods (for linear reconstruction methods, the effect of noise suppression due to regularization is well known [Liang et al. 2002; Lin et al. 2004; Hoge et al. 2004; Hoge et al. 2005]). When applied to data sets obtained with radial or pseudorandom sampling patterns, TV also leads to enhanced removal of artifacts. The flexibility is further shown by replacing TV with a different convex regularization functional that is more suitable in cases when the assumption of piecewise constant images is not reasonable: total generalized variation (TGV) of second order [Bredies, Kunisch, and Pock 2010; Knoll, Bredies, et al. 2010]. A similarly general approach in terms of sampling patterns and variational penalties is followed in SPIRiT [Lustig and Pauly 2010b], which can also be formulated as a nonlinear minimization problem. In contrast to the approach considered here, calibration is carried out in a separate step, and the corresponding variant with non-quadratic regularization (L1 SPIRiT) is based on compressed sensing using a wavelet transform.

17.2 THEORY

Mathematically, parallel MR imaging can be formulated as a nonlinear inverse problem where the sampling operator $\mathcal{F}_S$ (defined by the k-space trajectory, e.g., Fourier transform followed by multiplication with a binary mask in the case of standard Cartesian subsampling) and the correspondingly acquired k-space data $g = (g_1, \ldots, g_N)^T$ from N receiver coils are given, and the spin density u and the unknown (or not perfectly known) set of coil sensitivities

$c = (c_1, \ldots, c_N)^T$ have to be found such that

$$F(u, c) := (\mathcal{F}_S(u \cdot c_1), \ldots, \mathcal{F}_S(u \cdot c_N))^T = g$$

holds. As was shown in [Uecker, Hohage, et al. 2008; Knoll, Clason, Uecker, et al. 2009], this problem can be solved using the iteratively regularized Gauss–Newton (IRGN) method [Bakushinsky and Kokurin 2004; Engl, Hanke, and Neubauer 1996; Blaschke, Neubauer, and Scherzer 1997; Hohage 1997], i.e., computing in each step k for given $x^k := (u^k, c^k)$ the solution $\delta x := (\delta u, \delta c)$ of the minimization problem

$$\min_{\delta x} \frac{1}{2}\|F'(x^k)\delta x + F(x^k) - g\|^2 + \frac{\alpha_k}{2}\mathcal{W}(c^k + \delta c) + \beta_k \mathcal{R}(u^k + \delta u) \tag{17.2.1}$$

for given $\alpha_k, \beta_k > 0$, and then setting $x^{k+1} := x^k + \delta x$, $\alpha_{k+1} := q_\alpha \alpha_k$ and $\beta_{k+1} := q_\beta \beta_k$ with $0 < q_\alpha, q_\beta < 1$. Here, $F'(x^k)$ is the Fréchet derivative of F evaluated at x^k. The term $\mathcal{W}(c) = \|Wc\|^2 = \|w \cdot \mathcal{F}c\|^2$ is a penalty on the high Fourier coefficients of the sensitivities and $\mathcal{R}$ is a regularization term for the image. So far, the application of the IRGN method to parallel imaging has been formulated with a conventional L^2 penalty [Uecker, Hohage, et al. 2008; Knoll, Clason, Uecker, et al. 2009] (i.e., $\mathcal{R}(u) = \frac{1}{2}\|u\|^2$). As demonstrated in this work, the IRGN method can also be used with other regularization terms, which can be chosen dependent on the application. For example, the stability of the method with respect to noise can be improved: Since α_k and β_k are decreasing during the iteration, the problem in (17.2.1) will become increasingly ill-conditioned. This leads to noise amplification, which can be counteracted by using a regularization term with stronger noise removal properties than the L^2 penalty. A possible choice is the total variation (TV) of the image, i.e.,

$$\mathcal{R}(u) = \int |\nabla u|_2 \, dx,$$

where $|\cdot|_2$ denotes the Euclidean norm in $\mathbb{R}^2$. To calculate the solution of (17.2.1), we make use of the dual characterization of the TV semi-norm:

$$\beta \int |\nabla u|_2 \, dx = \sup_{p \in C_\beta} \langle u, -\operatorname{div} p \rangle,$$

where $p = (p_1, p_2)^T$, $\operatorname{div} p = \partial_x p_1 + \partial_y p_2$ with appropriate boundary conditions and

$$C_\beta = \{p \in L^2(\Omega; \mathbb{C}^2) : \ \operatorname{div} p \in L^2(\Omega; \mathbb{C}), \ |p(x)|_2 \leqslant \beta \text{ for almost all } x \in \Omega\}.$$

The problem in (17.2.1) then becomes a non-smooth convex-concave saddle-point problem,

$$\min_{\delta u, \delta c} \max_{p \in C_{\beta_k}} \frac{1}{2}\|F'(x^k)\delta x + F(x^k) - g\|^2 + \frac{\alpha_k}{2}\mathcal{W}(c^k + \delta c) + \langle u^k + \delta u, -\operatorname{div} p \rangle,$$

which can be solved efficiently using a projected primal-dual extra-gradient method [Pock et al. 2009; Chambolle and Pock 2010], given as Algorithm 17.1. Since this requires only

Algorithm 17.1 Solution of TV sub-problem (17.2)

function TVSOLVE(u, c, g, α, β)
  $\delta u, \delta\bar{u}, \delta c, \delta\bar{c}, p \leftarrow 0$, choose $\sigma, \tau > 0$
  repeat
    $p \leftarrow \mathrm{proj}_\beta(p + \tau\nabla(u + \delta\bar{u}))$
    $\delta u_{old} \leftarrow \delta u,\ \delta c_{old} \leftarrow \delta c$
    $\delta u \leftarrow \delta u - \sigma(\sum_{i=1}^{N} c_i^* \cdot \mathcal{F}_s^*(\mathcal{F}_s(u \cdot \delta\bar{c}_i + c_i \cdot \delta\bar{u}) + F(u, c) - g) - \mathrm{div}\, p)$
    $\delta c \leftarrow \delta c - \sigma(u^* \cdot \mathcal{F}_s^*(\mathcal{F}_s(u \cdot \delta\bar{c}_i + c_i \cdot \delta\bar{u}) + F(u, c) - g) + \alpha W^* W(c_i + \delta\bar{c}_i))$
    $\delta\bar{u} \leftarrow 2\delta u - \delta u_{old}$
    $\delta\bar{c} \leftarrow 2\delta c - \delta c_{old}$
  until convergence
  return $\delta u, \delta c$
end function

application of $F'(x^k)$ and its adjoint $F'(x^k)^*$, the algorithm can be implemented efficiently on modern multi-core hardware such as graphics processing units (GPUs). Due to the bilinear structure of F, the action of $F'(x^k)$ and $F'(x^k)^*$ can be calculated explicitly in terms of the subsampling operator $\mathcal{F}_S$ and its adjoint $\mathcal{F}_S^*$. The projection onto the convex set C_β can be calculated pointwise by setting for all $x \in \Omega$

$$\mathrm{proj}_\beta(q)(x) = \frac{q(x)}{\max(1, \beta^{-1}(|q(x)|_2)}.$$

Since TV regularization is known to introduce staircasing artifacts if the penalty parameter is large, we also consider second order total generalized variation (TGV), which is a generalization of TV that avoids the staircasing in regions of smooth signal change [Bredies, Kunisch, and Pock 2010; Knoll, Bredies, et al. 2010]. This amounts to setting

$$\beta\mathcal{R}(u) = \inf_v \beta\|\nabla u - v\| + 2\beta\|\mathcal{E}v\|,$$

where $\mathcal{E}v = \frac{1}{2}(\nabla v + \nabla v^T)$ denotes the symmetrized gradient of the complex-valued vector field $v \in \mathcal{C}^1(\Omega; \mathbb{C}^2)$. We refer to [Bredies, Kunisch, and Pock 2010; Knoll, Bredies, et al. 2010] for a detailed description of this functional and an explanation of its properties. Using again the dual representation of the norms, the Gauss–Newton step (17.2.1) is equivalent to the saddle point problem

$$\min_{\delta u, \delta c, v} \max_{\substack{p \in C_{\beta_k} \\ q \in C^2_{\beta_k}}} \frac{1}{2}\|F'(x^k)\delta x + F(x^k) - g\|^2 + \frac{\alpha_k}{2}W(c^k + \delta c) + \langle \nabla u^k + \delta u - v, p\rangle + \langle \mathcal{E}v, q\rangle,$$

where

$$C^2_\beta = \left\{ q \in \mathcal{C}_c(\Omega, \mathcal{S}^{2\times 2}) :\ \left(|q_{11}(x)|^2 + |q_{22}(x)|^2 + 2|q_{12}(x)|^2\right)^{1/2} \leqslant 2\beta\ \forall x \in \Omega \right\}$$

Algorithm 17.2 Solution of TGV sub-problem (17.2)

function TGVSOLVE(u, c, g, α, β)
  $\delta u, \delta\bar{u}, \delta c, \delta\bar{c}, v, \bar{v}, p, q \leftarrow 0$, choose $\sigma, \tau > 0$
  repeat
    $p \leftarrow \mathrm{proj}_\beta(p + \tau(\nabla(u + \delta\bar{u}) - v))$
    $q \leftarrow \mathrm{proj}^2_\beta(q + \tau(\mathcal{E}v))$
    $\delta u_{old} \leftarrow \delta u,\ \delta c_{old} \leftarrow \delta c,\ v_{old} \leftarrow v$
    $\delta u \leftarrow \delta u - \sigma(\sum_{i=1}^{N} c_i^* \cdot \mathcal{F}_s^*(\mathcal{F}_s(u \cdot \delta\bar{c}_i + c_i \cdot \delta\bar{u}) + F(u, c) - g) - \mathrm{div}\, p)$
    $\delta c \leftarrow \delta c - \sigma(u^* \cdot \mathcal{F}_s^*(\mathcal{F}_s(u \cdot \delta\bar{c}_i + c_i \cdot \delta\bar{u}) + F(u, c) - g) + \alpha W^* W(c_i + \delta\bar{c}_i))$
    $v \leftarrow v - \sigma(-p + \mathcal{E}^* q)$
    $\delta\bar{u} \leftarrow 2\delta u - \delta u_{old}$
    $\delta\bar{c} \leftarrow 2\delta c - \delta c_{old}$
    $\bar{v} \leftarrow 2v - v_{old}$
  until convergence
  return $\delta u, \delta c$
end function

and $\mathcal{S}^{2\times2}$ denotes the set of symmetric complex matrices. The corresponding extra-gradient method is given as Algorithm 17.2, where the projection proj^2_β onto C^2_β can again be computed pointwise.

17.3 MATERIALS AND METHODS

17.3.1 DATA ACQUISITION

Experiments were performed for 3D pseudorandom as well as 2D radial sampling patterns. All measurements were performed on a clinical 3 T system (Siemens Magnetom TIM Trio, Erlangen, Germany). Written informed consent was obtained from all volunteers prior to the examination.

Accelerated acquisition with pseudorandom sampling was tested with phantom experiments. A receive-only 12-channel head coil was used, and an SVD based coil compression [Buehrer et al. 2007] was applied to reduce the data to 8 virtual channels. Measurements were performed with a 3D FLASH sequence with the following sequence parameters: TR 20 ms, TE 5 ms, flip angle 18°, matrix size 256×256×256. The pulse sequence was modified to include a binary 2D mask defining the subsampling of both phase-encoding directions. A resolution of 1 mm× 1 mm×5 mm was used. Raw data was exported from the scanner, a 1D Fourier transform was performed along the readout direction, and partitions orthogonal to this axis were reconstructed.

Additionally, a fully sampled T_2 weighted 2D turbo spin echo data set of the brain of a healthy volunteer was acquired with a 32-channel receive coil. The data was compressed to 12 virtual channels. Sequence parameters were TR 5000 ms, TE 99 ms, turbo factor 10, matrix size 256×256, slice thickness 4 mm and an in-plane resolution of $0.86\,\text{mm} \times 0.86\,\text{mm}$. Raw data was exported from the scanner and then subsampled retrospectively with an adapted pseudorandom sampling pattern [Lustig, Donoho, and Pauly 2007], where sampling points on a regular Cartesian grid are randomly selected according to a specified probability density function which is based on the energy distribution in k-space of medical images [Knoll, Clason, Diwoky, et al. 2011].

For comparison, conventional TV filtering was applied to the magnitude images obtained from standard IRGN reconstruction of the T_2 weighted data set of the brain. This experiment was repeated for an accelerated pseudorandom in-vivo measurement of the brain of a different volunteer. The same experimental setup was used as in the accelerated phantom experiments, except that 9 SVD compressed virtual channels were used and an isotropic spatial resolution of 1 mm was achieved.

Radial sampling experiments were performed with an rf-spoiled radial FLASH sequence with sequence parameters TR=2.0 ms, TE=1.3 ms, and a flip angle of 8°. Images of a water phantom and of the heart of a healthy volunteer were made. In-vivo data was acquired with a 32-channel body array coil, and data acquisition was performed with a protocol designed for real-time imaging [Frahm, Haase, and Matthaei 1986; Riederer et al. 1988; Wright et al. 1989] without cardiac gating and during free breathing [Zhang, Block, and Frahm 2010]. After the acquisition, the data was compressed to 12 virtual channels for the in-vivo experiments and 8 virtual channels for the phantom experiments. An in-plane resolution of $2\,\text{mm} \times 2\,\text{mm}$ and a slice thickness of 8 mm was used in combination with 128×128 image matrices. Due to the two-fold oversampling, this resulted in 256 sample points for each radial spoke.

17.3.2 NONLINEAR RECONSTRUCTION

All reconstructions were performed offline using a Matlab (R2010a, The MathWorks, Natick, MA, USA) implementation of the described nonlinear inversion method. For the reconstruction of radial data sets, Fessler and Sutton's NUFFT [Fessler and Sutton 2003] code was used. To facilitate comparison, the solution of (17.2.1) with $\mathcal{R}(v) = \frac{1}{2}\|v\|^2$ was computed using the same extra-gradient scheme, which can be obtained from Algorithm 17.1 by removing step 4 and replacing the term "$-\operatorname{div} p$" with "$+\beta u + \delta\bar{u}$" in step 6. In the following, we will refer to the Gauss–Newton reconstruction using an L^2-penalty simply as IRGN, while the reconstruction using TV and TGV penalties will be denoted by IRGN-TV and IRGN-TGV, respectively.

17.3.3 PARAMETER CHOICE

The parameters in Algorithm 17.1 were chosen according to the convergence theory for the projected extra-gradient scheme. The step lengths σ and τ were selected such that $\sigma\tau L^2 < 1$ holds, where L is the Lipschitz constant of the gradient of the functional to be minimized. This constant depends on the subsampling strategy and the iterates u^k, c^k, but can be estimated using a few iterations of the power method to approximately compute the norms of the partial Fréchet derivatives of the linearized operator $F'(x)$. As the norm of the finite difference approximation of the divergence and gradient operators with mesh size 1 is $\sqrt{8}$, we set $\tau = \sigma = (8 + 2\max(|\tilde{L}_u|, |\tilde{L}_c|))^{-1/2}$ in Algorithm 17.1, where $\tilde{L}_u$, $\tilde{L}_c$ are the estimates from the power method. The step lengths in Algorithm 17.2 were set to $\tau = \sigma = (12 + 2\max(|\tilde{L}_u|, |\tilde{L}_c|))^{-1/2}$ based on the norm of the linear operator involving the symmetrized derivative. The iteration was terminated after a fixed number of iterations, since the efficiency estimate for the extra-gradient method gives an upper bound on the required number of iterations to achieve a given accuracy. Since a high accuracy is not necessary during the initial Gauss–Newton iterations with large penalties, we started with $N_0 = 20$ iterations and set $N_k = 2N_{k+1}$. These choices were stable and yielded good results for all data sets.

The parameters in the Gauss–Newton iteration were chosen according to a quasi-optimality criterion. The initial penalties α_0, β_0 were chosen such that the norm of the residual $\|F(u^1, c^1) - g\|$ after the first iteration was roughly 3/4 of the initial residual, and the reduction factors q_α, q_β were set such that each further iteration roughly reduced the residual by a factor of 1/2. The iteration was terminated once the achieved reduction factor fell below 3/4. This lead to the choice $\alpha_0 = 1$, $\beta_0 = 2$, $q_\alpha = q_\beta = 1/10$, and 5 Gauss–Newton iterations for the radial data set. For the pseudorandom data set, $\beta_0 = 1$, $q_\beta = 1/5$ and 6 Gauss–Newton iterations were used.

Since the TV regularization parameter is continually decreased during the Gauss–Newton iteration, the final reconstruction will typically not show strong signs of TV filtering such as a cartoon-like appearance. A more pronounced TV effect can be achieved if the decrease of the regularization parameter is stopped at the desired level. To illustrate this, we will also show reconstructions where we have set $\beta_{k+1} = \max(\beta_{\min}, q_\beta \beta_k)$ with $\beta_{\min} = 5 \cdot 10^{-3}$ for L^2, TV and TGV regularization (with otherwise unchanged parameters).

Due to the difference in functionals and normalization of raw data and reconstructed magnitude images, it is not sensible for comparison purposes to use the above values of $\beta_{\min}$ as regularization parameters for TV filtering. Instead, based on visual inspection, an optimal parameter was chosen for each data set. For the subsampled data with $R = 6$, this was $\beta_1^* = 5 \cdot 10^{-3}$, while the accelerated in-vivo data with $R = 4$ required $\beta_2^* = 1.5 \cdot 10^{-2}$.

17.4 RESULTS

Figure 17.1 shows a partition of a water phantom reconstructed with IRGN and IRGN-TV from pseudorandomly subsampled 3D data using acceleration factors $R = 4$ and $R = 10$. The reduced noise amplification and artifact removal characteristics of IRGN-TV are clearly visible for both acceleration factors. In the case of moderate acceleration with $R = 4$, the final TV regularization parameter $\beta_{\min}$ was set to zero. For $R = 10$, $\beta_{\min} = 5 \cdot 10^{-3}$ was used to achieve a stronger TV regularization and a better removal of artifacts. It must be noted that the phantom only consists of regions that are piecewise constant, and therefore the underlying assumption of the TV penalty is fulfilled. However, in the case of $R = 10$ with increased TV regularization, staircasing artifacts can be observed in regions where modulations caused by the inhomogeneity of the coil sensitivities affect the reconstructed image.

The above findings are confirmed for in-vivo conditions. IRGN and IRGN-TV reconstructions (with $\beta_{\min} = 0$) of the retrospectively pseudorandomly subsampled T2 weighted data of the brain are displayed in Figure 17.2, together with a sum-of-squares image obtained from the fully sampled data. Shown are results for acceleration factors of $R = 4$, $R = 6$ and $R = 8$ (defined as the ratio of total to acquired points on the underlying Cartesian sampling grid) as well as difference images to the fully sampled SOS image. With larger acceleration factors, an increasing amount of noise amplification and residual incoherent aliasing can be observed in the IRGN reconstructions, while this effect is reduced with IRGN-TV. The difference is especially noticeable in the magnified details shown in Figure 17.3.

The comparison of IRGN-TV with IRGN followed by TV filtering of the reconstructed magnitude image is shown in Figure 17.4. The conventional IRGN reconstruction of the T2 weighted brain data with $R = 6$ (see Figure 17.2) was postprocessed with a TV filter, where an optimal TV parameter was chosen based on visual inspection. To illustrate the robustness of IRGN-TV with respect to different data sets, this procedure was repeated for the accelerated brain scan obtained with the same FLASH sequence as in the phantom experiments (Figure 17.1) with $R = 4$. Although postprocessing using an optimal TV parameter yields results that are comparable to IRGN-TV, it can be seen that this parameter value is specific to each data set: The optimal value $\beta_1^* = 5 \cdot 10^{-3}$ for the subsampled data set (left column in Figure 17.4) results in insufficient filtering for the accelerated in-vivo scan, while the optimal choice $\beta_2^* = 1.5 \cdot 10^{-2}$ for the in-vivo scan (middle column in Figure 17.4) already leads to over-regularization when applied to the first data set. In contrast, the results with IRGN-TV (right column) were obtained using the same parameter set – most notably $\beta_{\min} = 0$ – in both cases.

Figure 17.5 shows two slices of a water phantom acquired with subsampled radial measurements and reconstructed with IRGN and IRGN-TV. Here, 25 spokes were acquired to reconstruct a 128×128 matrix, corresponding to an undersampling factor of approximately $R = 8$ in comparison to a fully sampled radial data set ($128 \cdot \frac{\pi}{2} \approx 201$ spokes). Because of the high undersampling, an increased value of $\beta_{\min} = 5 \cdot 10^{-3}$ was used in this example. Both slices show reduced noise and streaking artifacts when IRGN-TV is used. However, one

of the images (top row in Figure 17.5) again exhibits staircasing artifacts for the IRGN-TV solution.

Reconstructions of real-time images of the beating heart are displayed in Figure 17.6 together with plots of the signal intensities across an indicated horizontal line. Results for 25 ($R \approx 8.0$), 21 ($R \approx 9.6$), and 19 ($R \approx 10.6$) acquired spokes are shown, corresponding to image acquisition times of 50 ms, 41 ms, and 38 ms. For all reconstructions, $\beta_{\min} = 5 \cdot 10^{-3}$ was used. The image reconstructed from the 25 spokes data set does not show streaking artifacts for both IRGN and IRGN-TV. However, noise amplification is much stronger for IRGN. In contrast, reconstructions from 21 and 19 spokes show residual streaking artifacts due to increased subsampling, which are again reduced in the images reconstructed with IRGN-TV.

The effect of TGV regularization is demonstrated in Figure 17.7. It shows highlighted regions of both phantom images affected by staircasing artifacts (pseudorandom and radial sampling, see Figs. 17.1 and 17.5) using TV as well as TGV regularization, both with $\beta_{\min} = 5 \cdot 10^{-3}$. The staircasing artifact is completely removed in the reconstruction with the IRGN-TGV method.

17.5 DISCUSSION

The results from this work demonstrate that pronounced improvements in reconstruction quality of parallel imaging with nonlinear inversion can be achieved with TV and TGV based regularization instead of a simple L^2 penalty. If moderate acceleration is used (e.g. Figure 17.1, case of $R = 4$), TV serves as a stabilization term against noise amplification, which otherwise limits the practical use of parallel imaging to low acceleration factors. In cases where acceleration is pushed to its limits (Figure 17.1, case of $R = 10$; Figs. 17.5 and 17.6), TV also leads to an additional removal of undersampling artifacts when combined with trajectories that produce incoherent aliasing. However, it must be noted that in this case, small image features with low contrast may also be removed during the reconstruction. This effect can be observed for some smaller structures, which are highlighted in Figure 17.6. It can be seen in the top row that structures with low signal, which are still visible in conventional IRGN reconstructions, are removed when TV is applied. However, as indicated by the highlights in the middle row, objects with a slightly higher signal intensity, even though of same size, are preserved in both reconstructions. These effects are also represented in the cross-sectional plots. The elimination of structures like noise or residual streaking artifacts is visible as a reduced amount of high frequency oscillations in the IRGN-TV plots. Note that due to the nature of TV, blurring of sharp edges does not occur. This can be observed, e.g., at the sharp border of the ventricle, which is preserved equally well with IRGN and IRGN-TV.

As can be seen from Figure 17.4, the results of IRGN-TV are comparable to those achievable by TV filtering, since the regularization term is the same in both cases. However, this requires (manual) parameter tuning in the case of TV filtering, even though the data sets were very similar in terms of anatomy, sampling pattern and normalization. In contrast, IRGN-TV

performed similarly well on both data sets for the same choice of parameters, which can be attributed to the fact that the data fitting term in IRGN-TV considers only the actually acquired k-space coefficients. This allows more accurate discrimination between image content and reconstruction artifacts. On the other hand, the data fitting term in TV filtering is based solely on image-space contrast. Furthermore, due to the iterative nature of the IRGN method, the noise suppression properties of strong TV regularization can take effect even for a small final regularization parameter value.

Our Matlab implementations reconstructed a single slice in a few minutes, where IRGN-TV took roughly 10% more time than IRGN (and similarly, IRGN-TGV was about 10% slower than IRGN-TV). It is possible to exploit the differentiability of the data consistency term to apply more efficient minimization algorithms such as the method of conjugate gradients in the case of IRGN or order optimal convex minimization methods such as those investigated in [Chambolle and Pock 2010] for IRGN-TV and IRGN-TGV. Because this work focused on the effect of the different regularization techniques on image quality, the same primal-dual extra-gradient method was used as inner algorithm in all cases to allow a direct comparison for identical parameter choices. In this context, it should be noted that the parameters for the iteratively regularized Gauss–Newton method were independent of the chosen regularization term, and only depended on the trajectory type. Similarly, since the norm of the forward operator is estimated in the algorithm, the fixed parameters for the primal-dual extra-gradient methods were independent of the data set.

In this work, the flexibility to include different regularization terms was demonstrated on 2D examples, where each image (slice, i.e., 2D partition of a 3D data set, or frame of a time series) was reconstructed individually. While computationally more demanding, the extension of the penalties into a third space or a time dimension should further improve the image quality due to temporal redundancies in dynamic sequences – an effect that is well described in the literature [Madore, Glover, and Pelc 1999; Kellman, Epstein, and McVeigh 2001; Tsao, Boesiger, and Pruessmann 2003; Xu, King, and Liang 2007]. For example, earlier work has shown that residual streaking artifacts in radial imaging can be removed with the use of a median filter in the time dimension when using an interleaved k-space sampling scheme [Uecker, Zhang, Voit, et al. 2010]. Since the median filter can be interpreted as solving an L^1 minimization problem, it is expected that the inclusion of a corresponding penalty – either in the form of an L^1 penalty on the difference between the current and previous slice or frame, or of a higher-dimensional T(G)V penalty on the full data set – will yield even better results. The resulting convex minimization problems can be solved using the same primal-dual extra-gradient method as employed in this work.

Another possible extension of this work is the integration of additional information about the physical signal model into the functional in (17.2.1). An important application is mapping of relaxation parameters in multi echo sequences [Block, Uecker, and Frahm 2009; Doneva et al. 2010]. Here, the framework of nonlinear inverse problems makes it straightforward to perform parameter identification during the reconstruction. Since the parameters are thus estimated

directly from the raw data, instead of from reconstructed images, better quantification is possible.

17.6 CONCLUSIONS

This work describes an approach to include additional variational penalties in parallel imaging with nonlinear inversion. The presented algorithms combine the advantages of nonlinear inversion, i.e., improved image quality through a better estimation of the coil sensitivities, with the advantageous properties of TV-based regularization terms. In addition to reducing the noise, the regularization is able to remove undersampling artifacts when combined with sampling strategies that produce incoherent aliasing, such as radial and pseudorandom sampling. The approach has the additional benefit that the inclusion of physiological parameters, either as additional penalties or as unknown parameters to be reconstructed, is straightforward.

ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences").

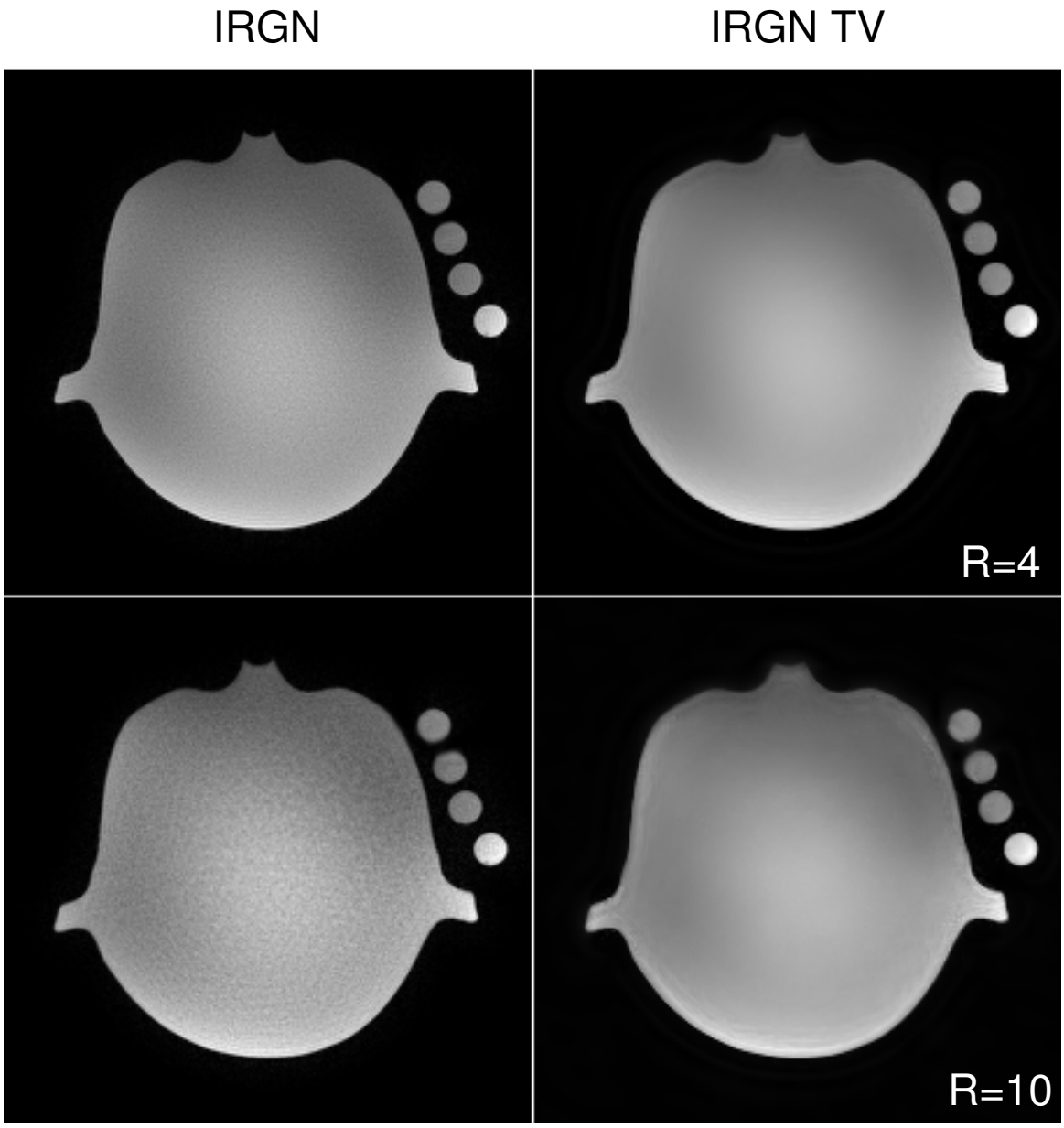


Figure 17.1: Comparison of IRGN (left) and IRGN-TV (right) for pseudorandom subsampling of a water phantom. Top: acceleration factor $R = 4$, $\beta_{min} = 0$; bottom: $R = 10$, $\beta_{min} = 5 \cdot 10^{-3}$.

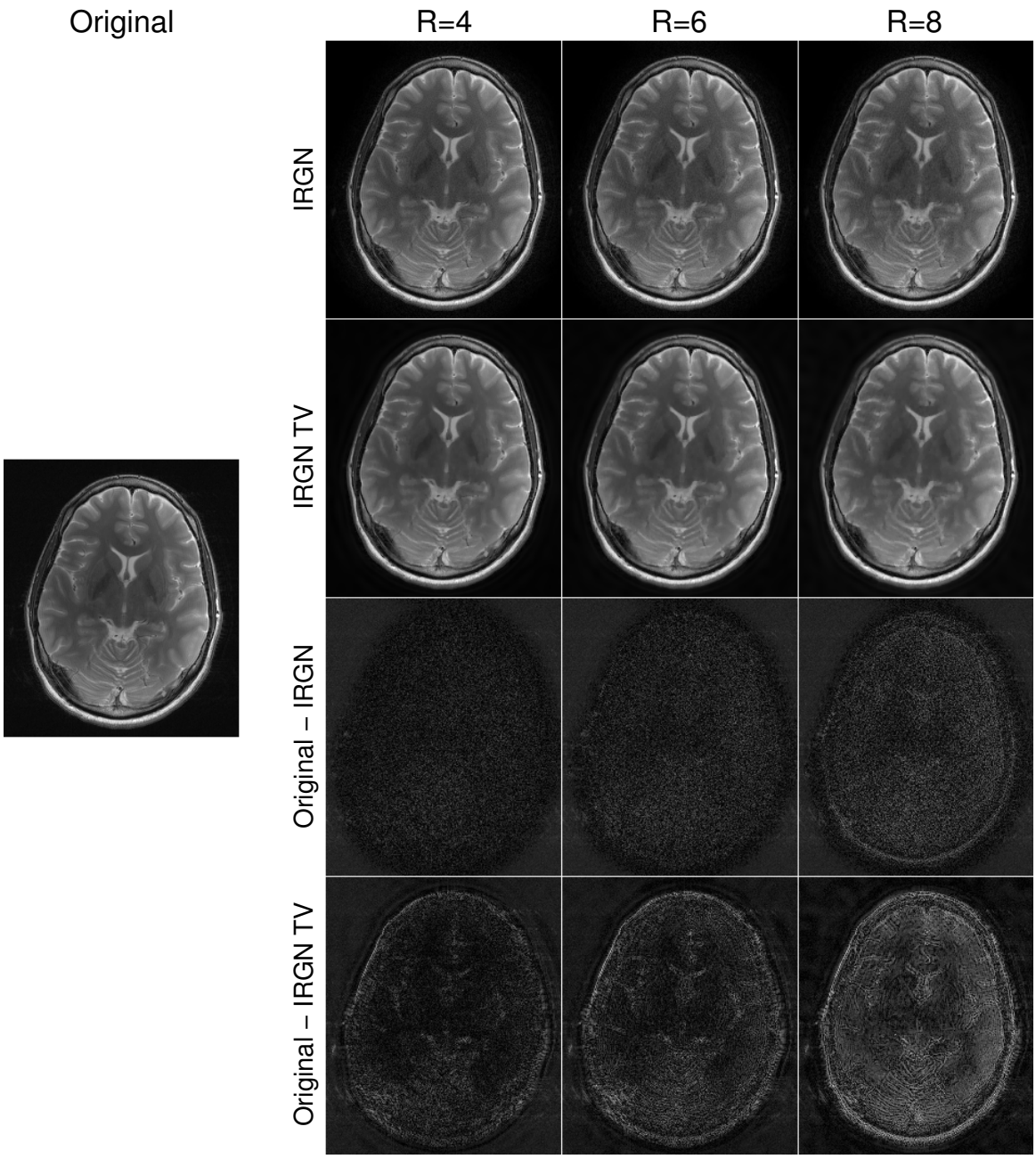


Figure 17.2: Comparison of IRGN (first row) and IRGN-TV (second row) for retrospective pseudorandom subsampling ($\beta_{\min} = 0$). From left: fully sampled acquisition, acceleration factors $R = 4$, $R = 6$, $R = 8$. Difference images to the fully sampled SOS reconstruction are shown for IRGN (third row) and IRGN-TV (fourth row) and are rescaled individually for IRGN and IRGN-TV to allow better depiction of the pixel differences.

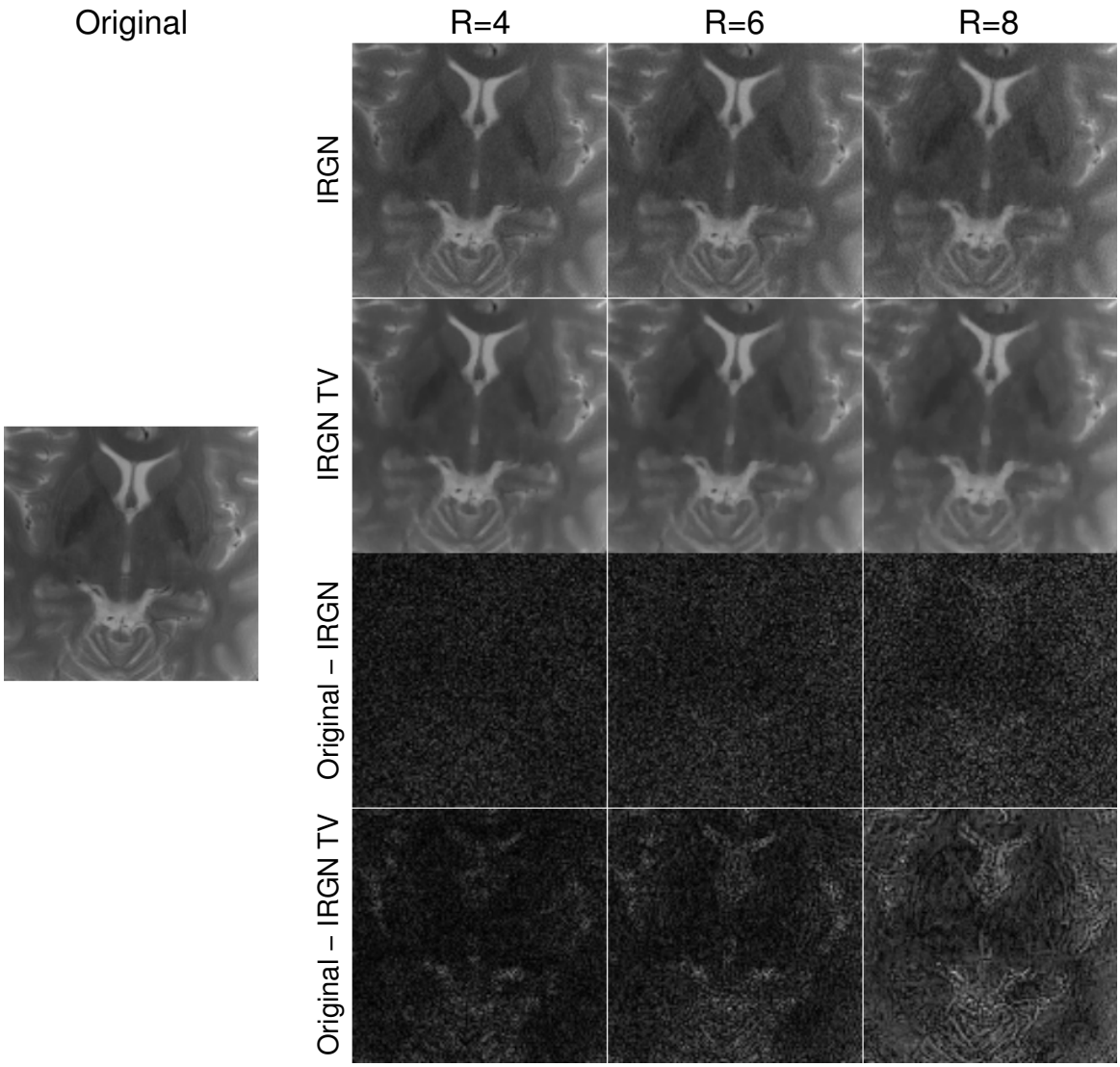


Figure 17.3: Magnified detail of Figure 17.2.

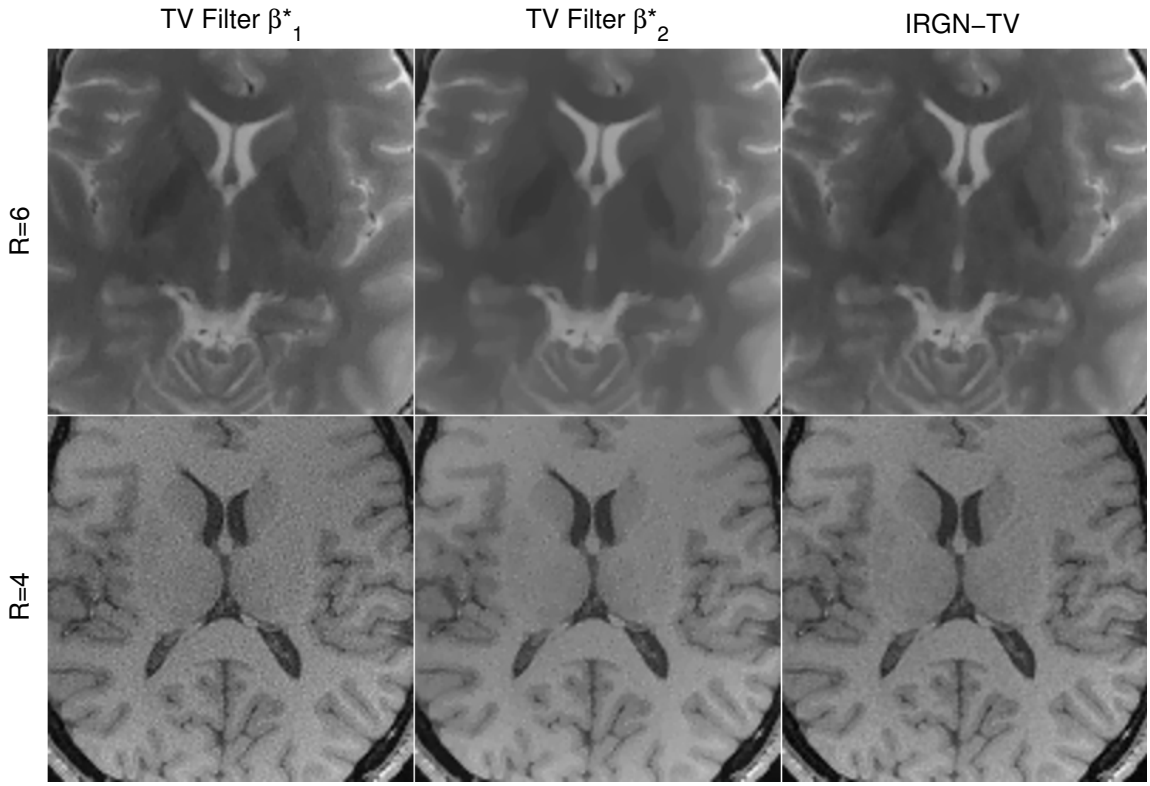


Figure 17.4: Comparison of IRGN-TV to TV filtering of conventional IRGN reconstructions for pseudorandom subsampling with $R = 6$ (top, same data as Figure 17.2) and accelerated in-vivo imaging with $R = 4$ (bottom). For TV filtering, optimal regularization parameters were identified by visual inspection: $\beta_1^* = 5 \cdot 10^{-3}$ (left, for T2 weighted TSE data) and $\beta_2^* = 1.5 \cdot 10^{-2}$ (middle, for FLASH data). Both results for IRGN-TV (right) were obtained using the same parameter set, esp. $\beta_{\min} = 0$.

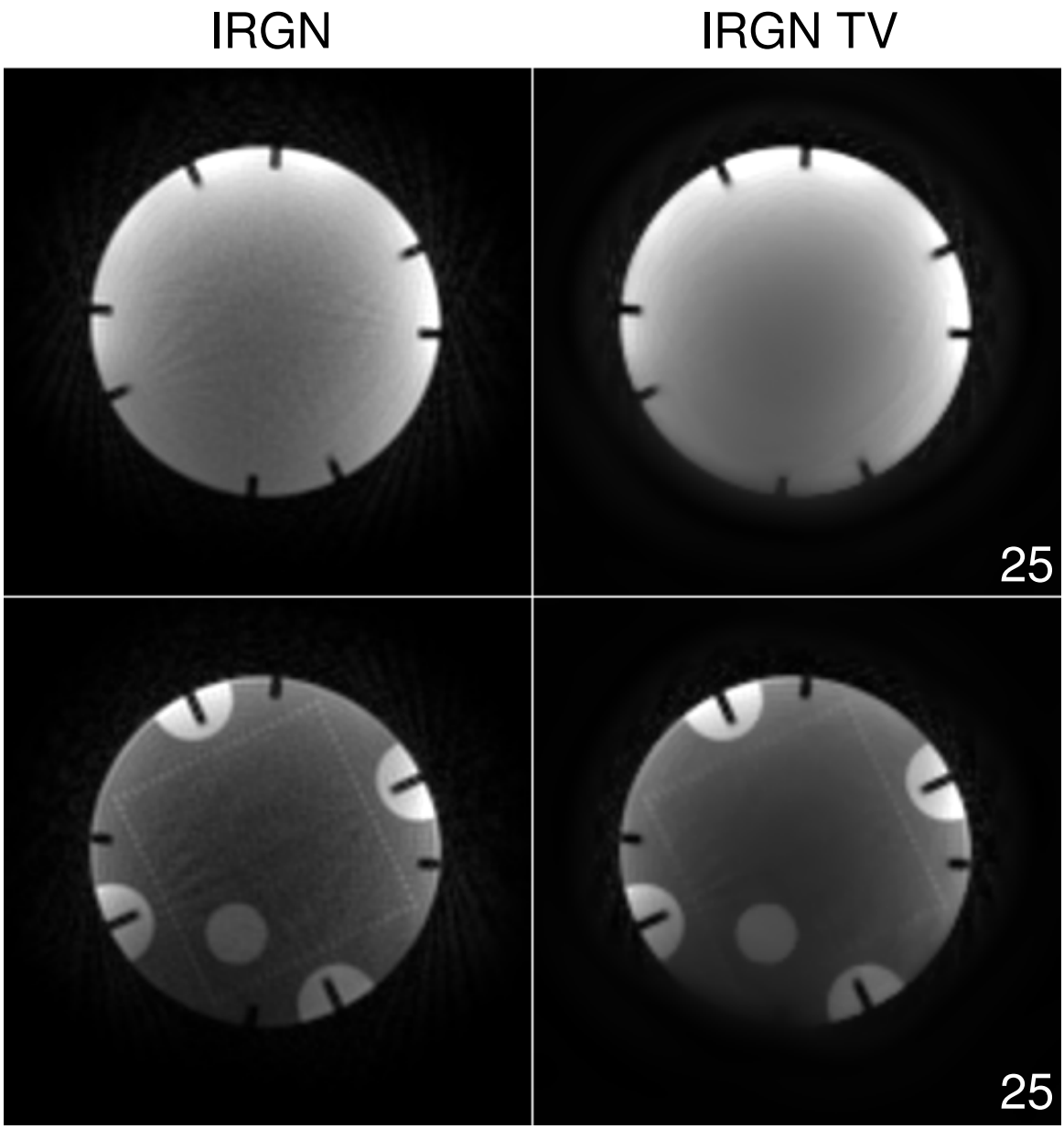


Figure 17.5: Comparison of IRGN (left) and IRGN-TV (right) for radial sampling of a phantom (25 spokes, $\beta_{\min} = 5 \cdot 10^{-3}$). Shown are two slices.

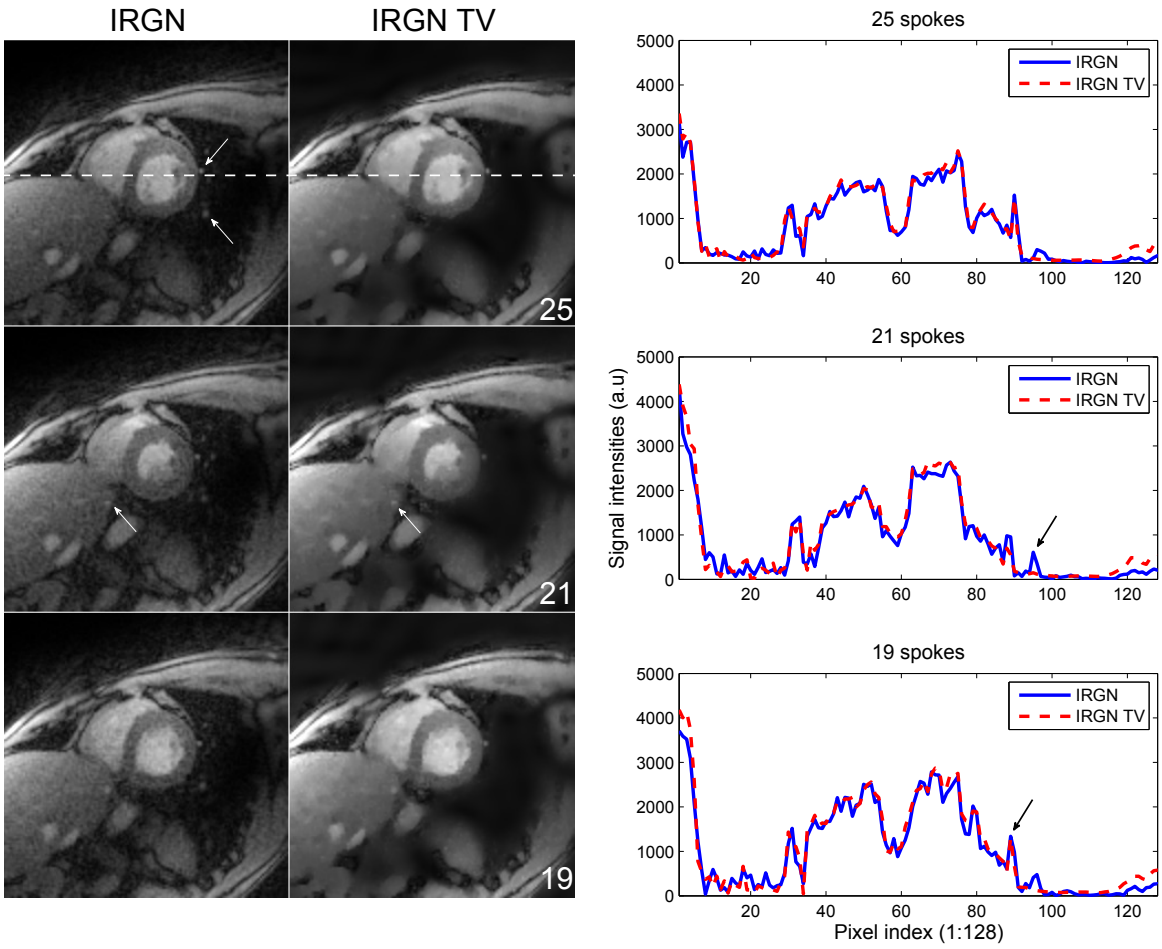


Figure 17.6: Comparison of IRGN (left) and IRGN-TV (middle) for radial sampling of a human heart ($\beta_{\min} = 5 \cdot 10^{-3}$). Top: 25 spokes. Highlighted are structures with little signal intensity that can be lost due to strong TV regularization. Middle: 21 spokes. Highlighted are structures of similar size but slightly higher signal intensity that are preserved even in case of TV regularization. Bottom: 19 spokes. The plots on the right show signal intensities across a horizontal line, indicated in the top row of the reconstruction results. The ability of IRGN-TV to preserve sharp edges is highlighted in the plot of the reconstruction from 19 spokes. The arrow marks the sharp boarder of the ventricle, which is depicted equally well with IRGN and IRGN-TV. The undesired loss of a small structure is highlighted in the plot of the reconstruction from 21 spokes. The plot crosses two adjacent vessels, which are both represented in the IRGN solution, but only the left one appears in the IRGN-TV reconstruction.

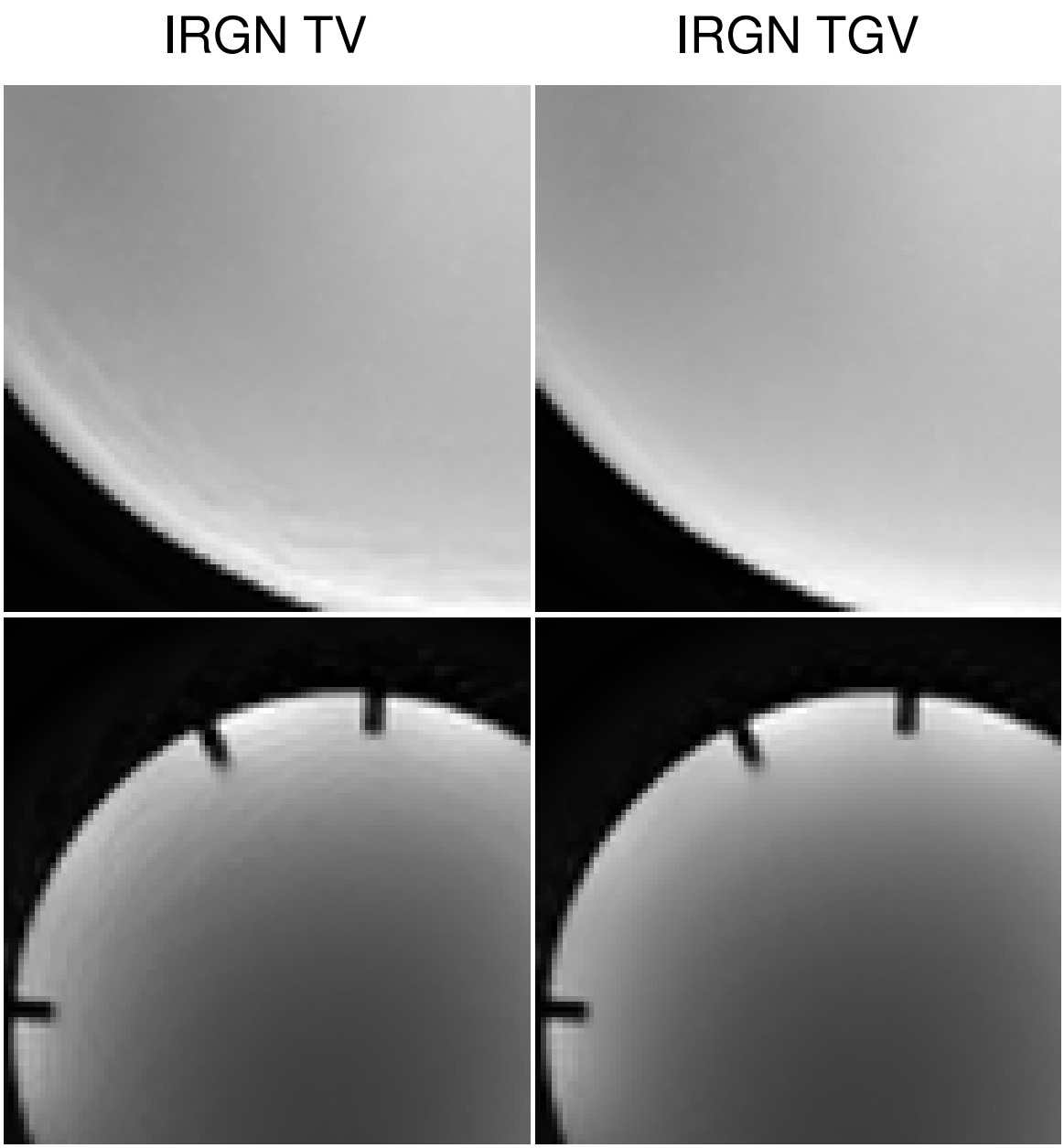


Figure 17.7: Comparison of IRGN-TV and IRGN-TGV for phantom data (top: pseudorandom sampling; bottom: radial sampling; both: $\beta_{\min} = 5 \cdot 10^{-3}$). Left: IRGN-TV (magnified details from Figures 17.1, $R = 10$ and 17.5). Modulations from the coil sensitivities lead to pronounced staircasing artifacts from TV regularization. Right: IRGN-TGV. Staircasing artifacts are completely removed for TGV regularization.

BIBLIOGRAPHY

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. 2nd ed. Vol. 140. Pure and Applied Mathematics. Elsevier/Academic Press, Amsterdam (cit. on pp. 182, 280, 281, 283).

Alibert, J.-J. and Raymond, J.-P. (1997). *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*. Numer. Funct. Anal. Optim. 18.3-4, pp. 235–250. DOI: 10.1080/01630569708816758 (cit. on pp. 5, 6).

Allard, W. K. (2007/08). *Total variation regularization for image denoising. I. Geometric theory*. SIAM J. Math. Anal. 39.4, pp. 1150–1190. DOI: 10.1137/060662617 (cit. on pp. 216, 253).

Alliney, S. (1997). *A property of the minimum vectors of a regularizing functional defined by means of the absolute norm*. IEEE Trans. Signal Process. 45, pp. 913–917. DOI: 10.1109/78.564179 (cit. on p. 216).

Alliney, S. and Ruzinsky, S. A. (1994). *An algorithm for the minimization of mixed* l_1 *and* l_2 *norms with application to Bayesian estimation*. IEEE Trans. Signal Process. 42, pp. 618–627. DOI: 10.1109/78.277854 (cit. on p. 216).

Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York (cit. on p. 61).

Amrouche, C., Ciarlet, P. G., and Ciarlet Jr., P. (2007). *Vector and scalar potentials, Poincaré's theorem and Korn's inequality*. C. R. Math. Acad. Sci. Paris 345.11, pp. 603–608. DOI: 10.1016/j.crma.2007.10.020 (cit. on p. 70).

Arampatzis, T., Lygeros, J., and Manesis, S. (2005). *A survey of applications of wireless sensors and wireless sensor networks*. In: Proceedings of the 13th IEEE International Conference on Control and Automation. IEEE, pp. 719–724. DOI: 10.1109/.2005.1467103 (cit. on p. 288).

Arridge, S. R. (1999). *Optical tomography in medical imaging*. Inverse Problems 15, R41–R93. DOI: 10.1088/0266-5611/15/2/022 (cit. on p. 313).

Attouch, H., Buttazzo, G., and Michaille, G. (2006). *Variational Analysis in Sobolev and BV Spaces*. Vol. 6. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (cit. on pp. 5, 7, 61, 96).

Baas, P., Murrer, L., Zoetmulder, F. A., Stewart, F. A., Ris, H. B., Zandwijk, N. van, Peterse, J. L., and Rutgers, E. L. (1997). *Photodynamic therapy as adjuvant therapy in surgically treated pleural malignancies*. Br. J. Cancer 76, pp. 819–826. DOI: 10.1038/bjc.1997.468 (cit. on pp. 101, 311).

Bakushinskii, A. B. (1984). *Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion*. USSR Computational Mathematics and Mathematical Physics 24.4, pp. 181–182. DOI: 10.1016/0041-5553(84)90253-2 (cit. on p. 42).

Bakushinsky, A. B. and Kokurin, M. Y. (2004). *Iterative methods for approximate solution of inverse problems*. Vol. 577. Mathematics and its Applications. Springer, Dordrecht (cit. on p. 327).

Banks, H. T. and Kunisch, K. (1989). *Estimation Techniques for Distributed Parameter Systems*. Birkhäuser, Boston (cit. on pp. 253, 255).

Bauer, F. and Kannengiesser, S. (2007). *An alternative approach to the image reconstruction for parallel data acquisition in MRI*. Math Meth Appl Sci 30, pp. 1437–1451. DOI: 10.1002/mma.848 (cit. on p. 325).

Ben-Asher, J., Cliff, E., and Burns, J. (1989). *Computational methods for the minimum effort problem with applications to spacecraft rotational maneuvers*. In: 1989 IEEE Conf. on Control and Applications, pp. 472–478. DOI: 10.1109/ICCON.1989.770562 (cit. on p. 194).

Bernstein, S. L., Burrows, M. L., Evans, J. E., Griffiths, A. S., Mcneill, D. A., Niessen, C. W., Richer, I., White, D. P., and Willim, D. K. (1974). *Long-range communications at extremely low frequencies*. Proc. IEEE 62.3, pp. 292–312. DOI: 10.1109/PROC.1974.9426 (cit. on p. 257).

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). *Convergence rates of general regularization methods for statistical inverse problems and applications*. SIAM J. Numer. Anal. 45.6, pp. 2610–2636. DOI: 10.1137/060651884 (cit. on p. 257).

Blaschke, B., Neubauer, A., and Scherzer, O. (1997). *On convergence rates for the iteratively regularized Gauss-Newton method*. IMA J. Numer. Anal. 17.3, pp. 421–436. URL: http://imajna.oxfordjournals.org/content/17/3/421.abstract (cit. on p. 327).

Block, K. T., Uecker, M., and Frahm, J. (2007). *Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint*. Magn Reson Med 57.6, pp. 1086–1098. DOI: 10.1002/mrm.21236 (cit. on p. 326).

Block, K. T., Uecker, M., and Frahm, J. (2009). *Model-based iterative reconstruction for radial fast spin-echo MRI*. IEEE Trans Med Imaging 28.11, pp. 1759–1769. DOI: 10.1109/TMI.2009.2023119 (cit. on p. 334).

Boccardo, L. and Gallouët, T. (1989). *Nonlinear elliptic and parabolic equations involving measure data*. J. Funct. Anal. 87.1, pp. 149–169. DOI: 10.1016/0022-1236(89)90005-0 (cit. on p. 6).

Bovik, A. C. (2005). *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*. Academic Press, Inc., Orlando (cit. on pp. 216, 253, 256).

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge (cit. on p. 289).

Bredies, K., Kunisch, K., and Pock, T. (2010). *Total Generalized Variation*. SIAM J. Imaging Sci. 3.3, pp. 492–526. DOI: 10.1137/090769521 (cit. on pp. 56, 326, 328).

Bredies, K. and Pikkarainen, H. (2012). *Inverse problems in spaces of measures*. ESAIM: Control Optim. Calc. Var. E-first. DOI: 10.1051/cocv/2011205 (cit. on pp. 92, 107).

Brezis, H. (1983). *Analyse Fonctionnelle*. Masson, Paris (cit. on pp. 67, 223).

Brezis, H. (2010). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York (cit. on pp. 3, 4, 92).

Buehrer, M., Pruessmann, K. P., Boesiger, P., and Kozerke, S. (2007). *Array compression for MRI with large coil arrays.* Magn Reson Med 57.6, pp. 1131–1139. DOI: 10.1002/mrm.21237. (Cit. on p. 329).

Burger, M. and Osher, S. (2004). *Convergence rates of convex variational regularization*. Inverse Problems 20.5, pp. 1411–1421. DOI: 10.1088/0266-5611/20/5/005 (cit. on pp. 42, 260, 289, 292).

Candes, E. J., Romberg, J., and Tao, T. (2006). *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*. IEEE Transactions on Information Theory 52.2, pp. 489–509. DOI: 10.1109/TIT.2005.862083 (cit. on p. 326).

Candes, E. and Tao, T. (2007). *The Dantzig selector: Statistical estimation when* p *is much larger than* n. Ann. Statist. 35.6, pp. 2313–2351. DOI: 10.1214/009053606000001523. (Cit. on p. 308).

Casas, E., Herzog, R., and Wachsmuth, G. (2012). *Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional*. SIAM J. Control Optim. 22.3, pp. 795–820. DOI: 10.1137/110834366 (cit. on pp. 18, 90, 107).

Casas, E. (1985). *L^2 estimates for the finite element method for the Dirichlet problem with singular data*. Numer. Math. 47, pp. 627–632. DOI: 10.1007/BF01389461 (cit. on pp. 115, 119).

Casas, E. (1986). *Control of an elliptic problem with pointwise state constraints*. SIAM J. Control Optim. 24.6, pp. 1309–1318. DOI: 10.1137/0324078 (cit. on pp. 5, 6, 108).

Casas, E. (1993). *Boundary control of semilinear elliptic equations with pointwise state constraints*. SIAM J. Control Optim. 31.4, pp. 993–1006. DOI: 10.1137/0331044 (cit. on p. 6).

Casas, E. (1997). *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*. SIAM J. Control Optim. 35.4, pp. 1297–1327. DOI: 10.1137/S0363012995283637 (cit. on pp. 6, 133).

Casas, E., Clason, C., and Kunisch, K. (2012). *Approximation of elliptic control problems in measure spaces with sparse solutions*. SIAM J. Control Optim. 50.4, pp. 1735–1752. DOI: 10.1137/110843216 (cit. on pp. 130, 139, 141, 163, 314, 315, 320).

Casas, E. and Zuazua, E. (2012). *Spike controls for elliptic and parabolic PDE*. Tech. rep. Basque Center for Applied Mathematics. URL: http://www.bcamath.org/documentos_public/archivos/publicaciones/SCL-D-11-00305R2.pdf (cit. on p. 130).

Chaabane, S., Ferchichi, J., and Kunisch, K. (2004). *Differentiability properties of the L^1-tracking functional and application to the Robin inverse problem*. Inverse Problems 20, pp. 1083–1097. DOI: 10.1088/0266-5611/20/4/006 (cit. on pp. 216, 253, 255).

Chambolle, A. and Pock, T. (2010). *A first-order primal-dual algorithm for convex problems with applications to imaging*. J Math Imaging Vis online first. DOI: 10.1007/s10851-010-0251-1 (cit. on pp. 56, 327, 334).

Chan, R. H., Dong, Y., and Hintermüller, M. (2010). *An efficient two-phase L1-TV method for restoring blurred images with impulse noise*. IEEE Transactions on Image Processing 19.4, pp. 1731–1739. DOI: 10.1109/TIP.2010.2045148 (cit. on p. 217).

Chan, T. F. and Esedoğlu, S. (2005). *Aspects of total variation regularized L^1 function approximation*. SIAM J. Appl. Math. 65.5, pp. 1817–1837. DOI: 10.1137/040604297 (cit. on pp. 216, 253).

Chang, T., He, L., and Fang, T. (2006). *MR image reconstruction from sparse radial samples using Bregman iteration*. In: Proc. Intl. Soc. Mag. Reson. Med. 14, p. 696 (cit. on p. 326).

Chavent, G. and Kunisch, K. (1997). *Regularization of linear least squares problems by total bounded variation*. ESAIM: Control Optim. Calc. Var. 2, pp. 359–376. DOI: 10.1051/cocv:1997113 (cit. on p. 68).

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). *Atomic decomposition by basis pursuit*. SIAM J. Sci. Comput. 20.1, pp. 33–61. DOI: 10.1137/S1064827596304010 (cit. on p. 253).

Chen, X., Nashed, Z., and Qi, L. (2000). *Smoothing methods and semismooth methods for nondifferentiable operator equations*. SIAM J. Numer. Anal. 38.4, pp. 1200–1216. DOI: 10.1137/S0036142999356719 (cit. on pp. 15, 262, 265, 267, 298).

Chen, Z. and Zou, J. (1999). *An augmented Lagrangian method for identifying discontinuous parameters in elliptic systems*. SIAM J. Control Optim. 37.3, pp. 892–910. DOI: 10.1137/S0363012997318602 (cit. on p. 255).

Ciarlet, P. G. (1978). *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam (cit. on pp. 113, 119, 146, 149).

Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis*. 2nd ed. Classics Appl. Math. 5. SIAM, Philadelphia (cit. on pp. 9, 261, 262).

Clason, C., Ito, K., and Kunisch, K. (2010). *Minimal invasion: An optimal L^∞ state constraint problem*. ESAIM: Math. Model. Numer. Anal. 45.3, pp. 505–522. DOI: 10.1051/m2an/2010064 (cit. on pp. 194, 289, 295, 298).

Clason, C. and Jin, B. (2012). *A semi-smooth Newton method for nonlinear parameter identification problems with impulsive noise*. SIAM J. Imaging Sci. 5, pp. 505–536. DOI: 10.1137/110826187 (cit. on pp. 293, 308).

Clason, C., Jin, B., and Kunisch, K. (2010a). *A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration*. SIAM J. Imaging Sci. 3.2, pp. 199–231. DOI: 10.1137/090758003 (cit. on pp. 253, 268, 269, 293).

Clason, C., Jin, B., and Kunisch, K. (2010b). *A duality-based splitting method for ℓ^1-TV image restoration with automatic regularization parameter choice*. SIAM J. Sci. Comput. 32.3, pp. 1484–1505. DOI: 10.1137/090768217 (cit. on pp. 43, 253, 257, 268, 293).

Clason, C. and Kunisch, K. (2011). *A duality-based approach to elliptic control problems in non-reflexive Banach spaces*. ESAIM: Control Optim. Calc. Var. 17.1, pp. 243–266. DOI: 10.1051/cocv/2010003 (cit. on pp. 90, 92, 107–109, 313).

Clason, C. and Kunisch, K. (2012). *A measure space approach to optimal source placement*. Computational Optimization and Applications 53.1, pp. 155–171. DOI: 10.1007/s10589-011-9444-9 (cit. on pp. 108, 110, 130, 314, 315).

Culver, J. P., Ntziachristos, V., Holboke, M. J., and Yodh, A. G. (2001). *Optimization of optode arrangements for diffuse optical tomography: A singular-value analysis*. Optics Letters 26, pp. 701–703. DOI: 10.1364/OL.26.000701 (cit. on p. 311).

DiBenedetto, E. (2002). *Real Analysis*. Birkhäuser, Boston, MA (cit. on p. 92).

Dolmans, D., Fukumura, D., and Jain, R. (2003). *Photodynamic therapy for cancer*. Nature Reviews Cancer 3.5, pp. 380–387. DOI: 10.1038/nrc1071 (cit. on p. 311).

Doneva, M., Boernert, P., Eggers, H., Stehning, C., Senegas, J., and Mertins, A. (2010). *Compressed sensing reconstruction for magnetic resonance parameter mapping*. Magn Reson Med 64.4, pp. 1114–1120. DOI: 10.1002/mrm.22483 (cit. on p. 334).

Dong, Y., Hintermüller, M., and Neri, M. (2009). *An efficient primal-dual method for ℓ^1 tv image restoration*. SIAM J. Imaging Sci. 2.4, pp. 1168–1189. DOI: 10.1137/090758490 (cit. on pp. 217, 253, 257).

Donoho, D. L. (2006). *Compressed sensing*. IEEE Transactions on Information Theory 52.4, pp. 1289–1306. DOI: 10.1109/TIT.2006.871582 (cit. on p. 326).

Doolin, D. and Sitar, N. (2005). *Wireless sensors for wildfire monitoring*. In: Proc. SPIE. Vol. 5765, pp. 477–484. DOI: 10.1117/12.605655 (cit. on p. 289).

Dunford, N. and Schwartz, J. T. (1988). *Linear operators. Part I*. Wiley Classics Library. John Wiley & Sons Inc., New York (cit. on p. 4).

Duval, V., Aujol, J.-F., and Gousseau, Y. (2009). *The TVL1 model: A geometric point of view*. Multiscale Model. Simul. 8.1, pp. 154–189. DOI: 10.1137/090757083 (cit. on pp. 216, 253).

Dwyer, P. J., White, W. M., Fabian, R. L., and Anderson, R. R. (2000). *Optical integrating balloon device for photodynamic therapy*. Lasers Surg Med 26.1, pp. 58–66. DOI: 10.1002/(SICI)1096-9101(2000)26:1<58::AID-LSM9>3.0.CO;2-V (cit. on p. 311).

Edwards, R. E. (1965). *Functional Analysis. Theory and Applications*. Holt, Rinehart and Winston, New York (cit. on pp. 5, 18, 132).

Ekeland, I. and Témam, R. (1999). *Convex Analysis and Variational Problems*. Classics Appl. Math. 28. SIAM, Philadelphia (cit. on pp. 7, 9, 11, 12, 62, 93, 136, 158, 196, 220, 262).

Elstrodt, J. (2005). *Maß- und Integrationstheorie*. 5th ed. Springer-Verlag, Berlin (cit. on pp. 4, 90).

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Kluwer, Dordrecht (cit. on pp. 41, 218, 219, 229, 257, 258, 290, 327).

Engl, H. W., Kunisch, K., and Neubauer, A. (1989). *Convergence rates for Tikhonov regularisation of non-linear ill-posed problems*. Inverse Problems 5.4, pp. 523–540. DOI: 10.1088/0266-5611/5/4/007 (cit. on pp. 219, 258).

Evans, L. C. and Gariepy, R. F. (1992). *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton (cit. on p. 264).

Fessler, J. A. and Sutton, B. P. (2003). *Nonuniform fast Fourier transforms using min-max interpolation*. IEEE Transactions on Signal Processing 51.2, pp. 560–574. DOI: 10.1109/TSP.2002.807005 (cit. on p. 330).

Flemming, J. and Hofmann, B. (2011). *Convergence rates in constrained Tikhonov regularization: Equivalence of projected source conditions and variational inequalities*. Inverse Problems 27.8, p. 085001. DOI: 10.1088/0266-5611/27/8/085001 (cit. on p. 291).

Frahm, J., Haase, A., and Matthaei, D. (1986). *Rapid NMR imaging of dynamic processes using the FLASH technique*. Magn Reson Med 3.2, pp. 321–327. DOI: 10.1002/mrm.1910030217 (cit. on p. 330).

Freiberger, M., Clason, C., and Scharfetter, H. (2010). *Total Variation Regularization for Nonlinear Fluorescence Tomography with an Augmented Lagrangian Splitting Approach*. Applied Optics 49.19, pp. 3741–3747. DOI: 10.1364/AO.49.003741 (cit. on p. 313).

Friedberg, J. S., Mick, R., Stevenson, J., Metz, J., Zhu, T., Buyske, J., Sterman, D. H., Pass, H. I., Glatstein, E., and Hahn, S. M. (2003). *A phase I study of Foscan-mediated photodynamic therapy and surgery in patients with mesothelioma*. Ann Thorac Surg 75.3, pp. 952–959. URL: http://www.ncbi.nlm.nih.gov/pubmed/12645723 (cit. on p. 311).

Fu, H., Ng, M. K., Nikolova, M., and Barlow, J. L. (2006). *Efficient minimization methods of mixed* l2-l1 *and* l1-l1 *norms for image restoration*. SIAM J. Sci. Comput. 27.6, pp. 1881–1902. DOI: 10.1137/040615079 (cit. on p. 217).

Gallouët, T. and Monier, A. (1999). *On the regularity of solutions to elliptic equations*. Rend. Mat. Appl. (7) 19.4, 471–488 (2000) (cit. on p. 91).

Gamper, U., Boesiger, P., and Kozerke, S. (2008). *Compressed sensing in dynamic MRI*. Magn Reson Med 59.2, pp. 365–373. DOI: 10.1002/mrm.21477 (cit. on p. 326).

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. 2nd ed. Chapman & Hall/CRC, Boca Raton (cit. on pp. 253, 256, 257).

Gharavi, H. and Kumar, S., eds. (2003). *Proceedings of the IEEE: Special issue on sensor networks and applications*. Vol. 91. 8. IEEE, pp. 1151–1163. DOI: 10.1109/JPROC.2003.814925 (cit. on p. 288).

Gilbarg, D. and Trudinger, N. S. (2001). *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Reprint of the 1998 edition. Springer-Verlag, Berlin (cit. on pp. 6, 91, 175, 195).

Graves, E. E., Culver, J. P., Ripoll, J., Weissleder, R., and Ntziachristos, V. (2004). *Singular-value analysis and optimization of experimental parameters in fluorescence molecular tomography*. J. Opt. Soc. Am. A 21, pp. 231–241. DOI: 10.1364/JOSAA.21.000231 (cit. on p. 311).

Griesse, R. and Lorenz, D. (2008). *A semismooth Newton method for Tikhonov functionals with sparsity constraints*. Inverse Problems 24.3, 035007 (19pp). DOI: 10.1088/0266-5611/24/3/035007 (cit. on p. 217).

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). *Generalized autocalibrating partially parallel acquisitions (GRAPPA)*. Magn Reson Med 47.6, pp. 1202–1210. DOI: 10.1002/mrm.10171 (cit. on p. 325).

Grund, T. and Rösch, A. (2001). *Optimal control of a linear elliptic equation with a supremum norm functional*. Optimization Methods and Software 15.3-4, pp. 299–329. DOI: 10.1080/10556780108805823 (cit. on pp. 34, 174, 194, 289, 295).

Gugat, M. and Leugering, G. (2008). *L^∞-norm minimal control of the wave equation: On the weakness of the bang-bang principle*. ESAIM Control Optim. Calc. Var. 14.2, pp. 254–283. DOI: 10.1051/cocv:2007044 (cit. on pp. 38, 194).

Haller–Dintelmann, R. and Rehberg, J. (2009). *Maximal parabolic regularity for divergence operators including mixed boundary conditions*. J. Differential Equations 247.5, pp. 1354–1396. DOI: 10.1016/j.jde.2009.06.001 (cit. on pp. 133, 134).

Hansen, P. C. (1998). *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, Philadelphia (cit. on pp. 229, 237).

Hansen, P. C. (2007). *Regularization Tools version 4.0 for Matlab 7.3*. Numer. Algorithms 46.2, pp. 189–194. DOI: 10.1007/s11075-007-9136-9. (Cit. on p. 302).

Hartley, R. I. and Schaffalitzky, F. (2004). *L-infinity Minimization in Geometric Reconstruction Problems*. In: CVPR, pp. 504–509. DOI: 10.1109/CVPR.2004.140 (cit. on p. 289).

Henderson, B. W., Busch, T. M., Vaughan, L. A., Frawley, N. P., Babich, D., Sosa, T. A., Zollo, J. D., Dee, A. S., Cooper, M. T., Bellnier, D. A., Greco, W. R., and Oseroff, A. R. (2000). *Photofrin Photodynamic Therapy Can Significantly Deplete or Preserve Oxygenation in Human Basal Cell Carcinomas during Treatment, Depending on Fluence Rate*. Cancer Treatment 60, pp. 525–529. URL: http://www.ncbi.nlm.nih.gov/pubmed/10676629 (cit. on p. 312).

Herzog, R., Stadler, G., and Wachsmuth, G. (2012). *Directional Sparsity in Optimal Control of Partial Differential Equations*. SIAM J. Control Optim. 50, pp. 943–963. DOI: 10.1137/100815037 (cit. on p. 130).

Hintermüller, M., Ito, K., and Kunisch, K. (2002). *The primal-dual active set strategy as a semismooth Newton method*. SIAM J. Optim. 13.3, 865–888 (2003). DOI: 10.1137/S1052623401383558 (cit. on pp. 15, 16, 75, 217, 227, 261, 265, 267, 298).

Hintermüller, M. and Kunisch, K. (2004). *Total bounded variation regularization as a bilaterally constrained optimization problem*. SIAM J. Appl. Math. 64.4, pp. 1311–1333. DOI: 10.1137/S0036139903422784 (cit. on pp. 92, 269, 273).

Hintermüller, M. and Kunisch, K. (2006). *Path-following methods for a class of constrained minimization problems in function space*. SIAM J. Optim. 17.1, pp. 159–187. DOI: 10.1137/040611598 (cit. on p. 187).

Hintermüller, M. and Rincon-Camacho, M. M. (2010). *Expected absolute value estimators for a spatially adapted regularization parameter choice rule in L^1-TV-based image restoration*. Inverse Problems 26.8, p. 085005. DOI: 10.1088/0266-5611/26/8/085005 (cit. on p. 256).

Hintermüller, M. and Stadler, G. (2006). *An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration*. SIAM J. Sci. Comput. 28.1, 1–23 (electronic). DOI: 10.1137/040613263 (cit. on p. 71).

Hintermüller, M. and Ulbrich, M. (2004). *A mesh-independence result for semismooth Newton methods*. Math. Program. 101.1, Ser. B, pp. 151–184. DOI: 10.1007/s10107-004-0540-9 (cit. on p. 17).

Hinze, M. (2005). *A variational discretization concept in control constrained optimization: The linear-quadratic case*. Comp. Optim. Appls. 30, pp. 45–61. DOI: 10.1007/s10589-005-4559-5 (cit. on pp. 26, 111).

Hofmann, B., Kaltenbacher, B., Pöschl, C., and Scherzer, O. (2007). *A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators*. Inverse Problems 23.3, pp. 987–1010. DOI: 10.1088/0266-5611/23/3/009 (cit. on pp. 258, 289, 291, 292).

Hoge, W. S., Brooks, D. H., Madore, B., and Kyriakos, W. (2004). *On the regularization of SENSE and Space-RIP in parallel MR imaging*. In: Proc. IEEE Int Biomedical Imaging: Nano to Macro Symp, pp. 241–244. DOI: 10.1109/ISBI.2004.1398519 (cit. on p. 326).

Hoge, W. S., Brooks, D. H., Madore, B., and Kyriakos, W. E. (2005). *A tour of accelerated parallel MR imaging from a linear systems perspective*. Concepts in Magnetic Resonance Part A 27A, pp. 17–37. DOI: 10.1002/cmr.a.20041 (cit. on p. 326).

Hohage, T. (1997). *Logarithmic convergence rates of the iteratively regularized Gauss-Newton method for an inverse potential and an inverse scattering problem*. Inverse Problems 13.5, pp. 1279–1299. DOI: 10.1088/0266-5611/13/5/012 (cit. on p. 327).

Hsu, D., Kakade, S. M., and Zhang, T. (2011). *Robust matrix decomposition with sparse corruptions*. IEEE Trans. Inf. Theory 57.11, pp. 7221–7234. DOI: 10.1109/TIT.2011.2158250 (cit. on p. 257).

Hu, Y., Wang, K., and Zhu, T. C. (2009). *A light blanket for intraoperative photodynamic therapy*. In: Photodynamic Therapy: Back to the Future. Ed. by D. H. Kessel. Vol. 7380. SPIE, 73801W. DOI: 10.1117/12.823064 (cit. on p. 311).

Hu, Y., Wang, K., and Zhu, T. C. (2010). *Pre-clinic study of uniformity of light blanket for intraoperative photodynamic therapy*. In: Optical Methods for Tumor Treatment and Detection: Mechanisms and Techniques in Photodynamic Therapy XIX. Ed. by D. H. Kessel. Vol. 7551. SPIE, p. 755112. DOI: 10.1117/12.842809 (cit. on p. 311).

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York (cit. on pp. 216, 256, 257, 263, 278).

Ito, K. and Jin, B. (2011). *A new approach to nonlinear constrained Tikhonov regularization*. Inverse Problems 27.10, p. 105005. DOI: 10.1088/0266-5611/27/10/105005 (cit. on p. 259).

Ito, K., Jin, B., and Takeuchi, T. (2011). *A regularization parameter for nonsmooth Tikhonov regularization*. SIAM J. Sci. Comput. 33.3, pp. 1415–1438. DOI: 10.1137/100790756 (cit. on p. 268).

Ito, K., Jin, B., and Zou, J. (2011). *A new choice rule for regularization parameters in Tikhonov regularization*. Applicable Analysis 90.10, pp. 1521–1544. DOI: 10.1080/00036811.2010.541450 (cit. on p. 218).

Ito, K. and Kunisch, K. (1992). *On the choice of the regularization parameter in nonlinear inverse problems*. SIAM J. Optim. 2.3, pp. 376–404. DOI: 10.1137/0802019 (cit. on pp. 217, 230).

Ito, K. and Kunisch, K. (2008). *Lagrange Multiplier Approach to Variational Problems and Applications*. Vol. 15. Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (cit. on pp. 12, 15–17, 75, 76, 98, 100, 178, 183, 187, 201, 202, 217, 221, 227, 228, 262, 295, 296, 299, 300).

Ito, K. and Kunisch, K. (2011). *Minimal effort problems and their treatment by semismooth Newton methods*. SIAM J. Control Opt. 49.5, pp. 2083–2100. DOI: 10.1137/100784667 (cit. on p. 194).

Jerison, D. and Kenig, C. E. (1995). *The inhomogeneous Dirichlet problem in Lipschitz domains*. J. Funct. Anal. 130.1, pp. 161–219. DOI: 10.1006/jfan.1995.1067 (cit. on pp. 114, 133).

Jin, B., Zhao, Y., and Zou, J. (2012). *Iterative parameter choice by discrepancy principle*. IMA Journal of Numerical Analysis. Advance Access. DOI: 10.1093/imanum/drr051 (cit. on pp. 218, 219, 260).

Jin, B. and Zou, J. (2010). *Numerical estimation of the Robin coefficient in a stationary diffusion equation*. IMA J. Numer. Anal. 30.3, pp. 677–701. DOI: 10.1093/imanum/drn066 (cit. on p. 255).

Johnston, P. R. and Gulrajani, R. M. (2002). *An analysis of the zero-crossing method for choosing regularization parameters*. SIAM J. Sci. Comput. 24.2, pp. 428–442. DOI: 10.1137/S1064827500373516 (cit. on p. 229).

Kärkkäinen, T., Kunisch, K., and Majava, K. (2005). *Denoising of smooth images using L^1-fitting*. Computing 74.4, pp. 353–376. DOI: 10.1007/s00607-004-0097-8 (cit. on pp. 216, 217, 253).

Kellman, P., Epstein, F. H., and McVeigh, E. R. (2001). *Adaptive sensitivity encoding incorporating temporal filtering (TSENSE)*. Magn Reson Med 45.5, pp. 846–852. DOI: 10.1002/mrm.1113 (cit. on p. 334).

Kindermann, S. (2011). *Convergence analysis of minimization-based noise level-free parameter choice rules for linear ill-posed problems*. Electron. Trans. Numer. Anal. 38, pp. 233–257. URL: http://etna.math.kent.edu/vol.38.2011/pp233-257.dir/ (cit. on pp. 43, 293).

Knoll, F., Bredies, K., Pock, T., and Stollberger, R. (2010). *Second Order Total Generalized Variation (TGV) for MRI*. Magn Reson Med 65.2, pp. 480–491. DOI: 10.1002/mrm.22595 (cit. on pp. 326, 328).

Knoll, F., Clason, C., Diwoky, C., and Stollberger, R. (2011). *Adapted Random Sampling Patterns for Accelerated MRI*. Magn Reson Mater Phy (MAGMA) 24.1, pp. 43–50. DOI: 10.1007/s10334-010-0234-7 (cit. on p. 330).

Knoll, F., Clason, C., Uecker, M., and Stollberger, R. (2009). *Improved Reconstruction in Non-Cartesian Parallel Imaging by Regularized Nonlinear Inversion*. In: Proc. Intl. Soc. Mag. Reson. Med. 17, p. 2721 (cit. on pp. 325, 327).

Kröner, A. and Vexler, B. (2009). *A priori error estimates for elliptic optimal control problems with a bilinear state equation*. J. Comput. Appl. Math. 230.2, pp. 781–802. DOI: 10.1016/j.cam.2009.01.023 (cit. on p. 279).

Krueger, T., Altermatt, H. J., Mettler, D., Scholl, B., Magnusson, L., and Ris, H.-B. (2003). *Experimental photodynamic therapy for malignant pleural mesothelioma with pegylated mTHPC*. Lasers Surg Med 32.1, pp. 61–68. DOI: 10.1002/lsm.10113 (cit. on p. 311).

Kummer, B. (1988). *Newton's method for non-differentiable functions*. Mathematical Research 45, pp. 114–125 (cit. on p. 14).

Kummer, B. (1992). *Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis*. In: Advances in optimization (Lambrecht, 1991). Vol. 382. Lecture Notes in Econom. and Math. Systems. Springer, Berlin, pp. 171–194 (cit. on pp. 122, 161, 298).

Kummer, B. (2000). *Generalized Newton and NCP-methods: Convergence, regularity, actions*. Discuss. Math. Differ. Incl. Control Optim. 20.2, pp. 209–244. DOI: 10.7151/dmdico.1013 (cit. on pp. 15, 265, 267).

Kunisch, K. and Zou, J. (1998). *Iterative choices of regularization parameters in linear inverse problems*. Inverse Problems 14.5, pp. 1247–1264. DOI: 10.1088/0266-5611/14/5/010 (cit. on p. 217).

Ladyzhenskaya, O. A. and Ural'tseva, N. N. (1968). *Linear and Quasilinear Elliptic Equations*. Translated from the Russian by Scripta Technica, Inc. Translation editor: Leon Ehrenpreis. Academic Press, New York (cit. on pp. 6, 195).

Lassas, M., Saksman, E., and Siltanen, S. (2009). *Discretization-invariant Bayesian inversion and Besov space priors*. Inverse Probl. Imaging 3.1, pp. 87–122. DOI: 10.3934/ipi.2009.3.87 (cit. on p. 258).

Lasser, T. and Ntziachristos, V. (2007). *Optimization of 360° projection fluorescence molecular tomography*. Medical Image Analysis 11, pp. 389–399. DOI: 10.1016/j.media.2007.04.003 (cit. on p. 311).

Liang, Z., Bammer, R., Ji, J., Pelc, N., and Glover, G. (2002). *Making better SENSE: Wavelet denoising, Tikhonov regularization, and total least squares*. In: Proc. Intl. Soc. Mag. Reson. Med. 10, p. 2388 (cit. on p. 326).

Lin, F.-H., Kwong, K. K., Belliveau, J. W., and Wald, L. L. (2004). *Parallel imaging reconstruction using automatic regularization.* Magn Reson Med 51.3, pp. 559–567. DOI: 10.1002/mrm.10718 (cit. on p. 326).

Liu, B., King, K., Steckner, M., Xie, J., Sheng, J., and Ying, L. (2009). *Regularized sensitivity encoding (SENSE) reconstruction using Bregman iterations.* Magn Reson Med 61.1, pp. 145–152. DOI: 10.1002/mrm.21799 (cit. on p. 326).

Logg, A. and Wells, G. N. (2010). *DOLFIN: Automated finite element computing.* ACM Trans. Math. Softw. 37 (2), pp. 1–28. DOI: 10.1145/1731022.1731030 (cit. on p. 103).

Logg, A., Mardal, K.-A., Wells, G. N., et al. (2012). *Automated Solution of Differential Equations by the Finite Element Method.* Software available from http://fenicsproject.org. Springer. DOI: 10.1007/978-3-642-23099-8 (cit. on p. 316).

Lustig, M., Lee, J., Donoho, D., and Pauly, J. (2005). *Faster imaging with randomly perturbed, under-sampled spirals and L1 reconstruction.* In: Proc. Intl. Soc. Mag. Reson. Med. 13, p. 685 (cit. on p. 326).

Lustig, M., Donoho, D., and Pauly, J. M. (2007). *Sparse MRI: The application of compressed sensing for rapid MR imaging.* Magn Reson Med 58.6, pp. 1182–1195. DOI: 10.1002/mrm.21391 (cit. on pp. 326, 330).

Lustig, M. and Pauly, J. M. (2010a). *Calibrationless Parallel Imaging Reconstruction by Structured Low-Rank Matrix Completion.* In: Proc. Intl. Soc. Mag. Reson. Med. 18, p. 2870 (cit. on p. 326).

Lustig, M. and Pauly, J. M. (2010b). *SPIRiT: Iterative self-consistent parallel imaging reconstruction from arbitrary k-space.* Magn Reson Med 64.2, pp. 457–471. DOI: 10.1002/mrm.22428 (cit. on p. 326).

Madore, B., Glover, G. H., and Pelc, N. J. (1999). *Unaliasing by Fourier-encoding the overlaps using the temporal dimension (UNFOLD), applied to cardiac imaging and fMRI.* Magn Reson Med 42.5, pp. 813–828. DOI: 10.1002/(SICI)1522-2594(199911)42:5<813::AID-MRM1>3.0.CO;2-S (cit. on p. 334).

Maurer, H. and Zowe, J. (1979). *First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems.* Math. Programming 16.1, pp. 98–110. DOI: 10.1007/BF01582096 (cit. on pp. 12, 160, 178, 295–297).

Meyer, C., Panizzi, L., and Schiela, A. (2011). *Uniqueness criteria for solutions of the adjoint equation in state-constrained optimal control.* Numerical Functional Analysis and Optimization 32.9, pp. 983–1007. DOI: 10.1080/01630563.2011.587074 (cit. on pp. 6, 91).

Meyers, N. G. (1963). *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations.* Ann. Scuola Norm. Sup. Pisa (3) 17, pp. 189–206. URL: http://www.numdam.org/item?id=ASNSP_1963_3_17_3_189_0 (cit. on p. 282).

Mifflin, R. (1977). *Semismooth and semiconvex functions in constrained optimization.* SIAM J. Control Optimization 15.6, pp. 959–972. DOI: 10.1137/0315061 (cit. on p. 14).

Morozov, V. A. (1966). *On the solution of functional equations by the method of regularization.* Soviet Math. Dokl. 7, pp. 414–417 (cit. on pp. 219, 260, 292).

Nečas, J. (1967). *Les Méthodes Directes en Théorie des Equations Elliptiques*. Editeurs Academia, Prague (cit. on p. 120).

Neustadt, L. W. (1962). *Minimum effort control systems*. J. SIAM Control Ser. A 1, pp. 16–31. DOI: 10.1137/0301002 (cit. on pp. 38, 194).

Nikolova, M. (2002). *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*. SIAM J. Numer. Anal. 40.3, pp. 965–994. DOI: 10.1137/S0036142901389165 (cit. on p. 216).

Nikolova, M. (2004). *A variational approach to remove outliers and impulse noise*. J. Math. Imaging Vision 20.1-2. Special issue on mathematics and image analysis, pp. 99–120. DOI: 10.1023/B:JMIV.0000011326.88682.e5 (cit. on p. 216).

Niu, R. and Varshney, P. (2006). *Target location estimation in sensor networks with quantized data*. IEEE Transactions on Signal Processing 54.12, pp. 4519–4528. DOI: 10.1109/TSP.2006.882082 (cit. on p. 289).

Pock, T., Cremers, D., Bischof, H., and Chambolle, A. (2009). *An Algorithm for Minimizing the Mumford-Shah Functional*. In: International Conference on Computer Vision (ICCV), pp. 1133–1140. DOI: 10.1109/ICCV.2009.5459348 (cit. on p. 327).

Polastre, J., Szewczyk, R., Mainwaring, A., Culler, D., and Anderson, J. (2004). *Analysis of Wireless Sensor Networks for Habitat Monitoring*. In: Wireless Sensor Networks. Ed. by C. S. Raghavendra, K. M. Sivalingam, and T. Znati. Springer US, pp. 399–423. DOI: 10.1007/978-1-4020-7884-2_18 (cit. on p. 289).

Pöschl, C. (2009). *An overview on convergence rates for Tikhonov regularization methods for non-linear operators*. J. Inverse Ill-Posed Probl. 17.1, pp. 77–83. DOI: 10.1515/JIIP.2009.009 (cit. on pp. 42, 258, 289).

Pruessmann, K. P., Weiger, M., Scheidegger, M. B., and Boesiger, P. (1999). *SENSE: Sensitivity encoding for fast MRI*. Magn Reson Med 42.5, pp. 952–962. DOI: 10.1002/(SICI)1522-2594(199911)42:5<952::AID-MRM16>3.0.CO;2-S (cit. on p. 325).

Prüfert, U. and Schiela, A. (2009). *The minimization of a maximum-norm functional subject to an elliptic PDE and state constraints*. ZAMM 89.7, pp. 536–551. DOI: 10.1002/zamm.200800097 (cit. on pp. 34, 174, 194, 289, 295).

Qi, L. Q. and Sun, J. (1993). *A nonsmooth version of Newton's method*. Math. Programming 58.3, Ser. A, pp. 353–367. DOI: 10.1007/BF01581275 (cit. on pp. 14, 122, 161).

Raviart, P.-A. and Thomas, J.-M. (1983). *Introduction à L'analyse Numérique des Equations aux Dérivées Partielles*. Masson, Paris (cit. on pp. 110, 139).

Resmerita, E. (2005). *Regularization of ill-posed problems in Banach spaces: Convergence rates*. Inverse Problems 21.4, pp. 1303–1314. DOI: 10.1088/0266-5611/21/4/007 (cit. on pp. 42, 289).

Resmerita, E. and Scherzer, O. (2006). *Error estimates for non-quadratic regularization and the relation to enhancement*. Inverse Problems 22.3, pp. 801–814. DOI: 10.1088/0266-5611/22/3/004 (cit. on p. 289).

Riederer, S. J., Tasciyan, T., Farzaneh, F., Lee, J. N., Wright, R. C., and Herfkens, R. J. (1988). *MR fluoroscopy: Technical feasibility.* Magn Reson Med 8.1, pp. 1–15. DOI: 10.1002/mrm.1910080102 (cit. on p. 330).

Ring, W. (2000). *Structural properties of solutions to total variation regularization problems.* M2AN Math. Model. Numer. Anal. 34.4, pp. 799–810. DOI: 10.1051/m2an:2000104 (cit. on pp. 61, 82).

Rodríguez, P. and Wohlberg, B. (2009). *Efficient minimization method for a generalized total variation functional.* IEEE Trans. Image Process. 18.2, pp. 322–332. DOI: 10.1109/TIP.2008.2008420 (cit. on pp. 217, 238, 250).

Rothmaier, M., Selm, B., Spichtig, S., Haensse, D., and Wolf, M. (2008). *Photonic textiles for pulse oximetry.* Opt Express 16.17, pp. 12973–12986. DOI: 10.1364/OE.16.012973 (cit. on p. 311).

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, Inc., New York, NY, USA (cit. on p. 216).

Rudin, L. I., Osher, S., and Fatemi, E. (1992). *Nonlinear total variation based noise removal algorithms.* Phys. D 60.1-4, pp. 259–268. DOI: 10.1016/0167-2789(92)90242-F (cit. on p. 326).

Rudin, W. (1970). *Real and Complex Analysis.* McGraw-Hill, London (cit. on p. 118).

Ruszczyński, A. (2006). *Nonlinear Optimization.* Princeton University Press, Princeton, NJ (cit. on p. 34).

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. (2009). *Variational Methods in Imaging.* Vol. 167. Applied Mathematical Sciences. Springer, New York (cit. on pp. 42, 289–291).

Schiela, A. (2008). *A simplified approach to semismooth Newton methods in function space.* SIAM J. on Opt. 19.3, pp. 1417–1432. DOI: 10.1137/060674375 (cit. on pp. 12, 17, 201, 265, 299).

Schirotzek, W. (2007). *Nonsmooth Analysis.* Universitext. Springer, Berlin (cit. on pp. 7, 9–11, 159).

Schizas, I., Giannakis, G., and Luo, Z. (2007). *Distributed estimation using reduced-dimensionality sensor observations.* IEEE Transactions on Signal Processing 55.8, pp. 4284–4299. DOI: 10.1109/TSP.2007.895987 (cit. on p. 289).

Schouwink, H. and Baas, P. (2004). *Foscan-mediated photodynamic therapy and operation for malignant pleural mesothelioma.* Ann Thorac Surg 78.1, 388, 388, author reply 389. DOI: 10.1016/j.athoracsur.2003.08.088 (cit. on p. 311).

Selm, B., Rothmaier, M., Camenzind, M., Khan, T., and Walt, H. (2007). *Novel flexible light diffuser and irradiation properties for photodynamic therapy.* J Biomed Opt 12.3, p. 034024. DOI: 10.1117/1.2749737 (cit. on p. 311).

Seo, Y. and Hartley, R. (2007). *A Fast Method to Minimize L^∞ Error Norm for Geometric Vision Problems.* In: ICCV, pp. 1–8. DOI: 10.1109/ICCV.2007.4408913 (cit. on p. 289).

She, S. and Owen, A. B. (2011). *Outlier detection using nonconvex penalized regression*. J. Amer. Stat. Ass. 106.494, pp. 626–639. DOI: 10.1198/jasa.2011.tm10390 (cit. on p. 257).

Showalter, R. E. (1997). *Monotone operators in Banach space and nonlinear partial differential equations*. Vol. 49. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI (cit. on p. 134).

Shykula, M. and Seleznjev, O. (2006). *Stochastic structure of asymptotic quantization errors*. Statist. Probab. Lett. 76.5, pp. 453–464. DOI: 10.1016/j.spl.2005.08.022. (Cit. on p. 288).

Sim, K. and Hartley, R. (2006). *Removing Outliers Using The L^∞ Norm*. In: CVPR. IEEE Computer Society, Washington, DC, USA, pp. 485–494. DOI: 10.1109/CVPR.2006.253 (cit. on p. 289).

Sodickson, D. K. and Manning, W. J. (1997). *Simultaneous acquisition of spatial harmonics (SMASH): Fast imaging with radiofrequency coil arrays.* Magn Reson Med 38.4, pp. 591–603. DOI: 10.1002/mrm.1910380414 (cit. on p. 325).

Stadler, G. (2009). *Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices*. Computational Optimization and Applications 44.2, pp. 159–181. DOI: 10.1007/s10589-007-9150-9 (cit. on pp. 18, 61, 68, 90, 107, 312).

Stampacchia, G. (1965). *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*. Ann. Inst. Fourier (Grenoble) 15.fasc. 1, pp. 189–258. DOI: 10.5802/aif.204 (cit. on pp. 5, 6, 63, 91).

Stojanovic, S. (1991). *Optimal damping control and nonlinear elliptic systems*. SIAM J. Control Optim. 29.3, pp. 594–608. DOI: 10.1137/0329033 (cit. on p. 254).

Sun, Z. and Zeng, J. (2010). *A damped semismooth Newton method for mixed linear complementarity problems*. Optimization Methods and Software 26.2, pp. 187–205. DOI: 10.1080/10556780903575680 (cit. on p. 204).

Témam, R. (2001). *Navier–Stokes Equations*. AMS Chelsea Publishing, Providence, RI (cit. on p. 70).

Thomée, V. (2006). *Galerkin Finite Element Methods for Parabolic Problems*. 2nd. Vol. 25. Springer Series in Computational Mathematics. Springer-Verlag, Berlin (cit. on pp. 28, 139).

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. J. Roy. Statist. Soc. Ser. B 58.1, pp. 267–288. DOI: 10.1111/j.1467-9868.2011.00771.x (cit. on p. 253).

Troianiello, G. M. (1987). *Elliptic Differential Equations and Obstacle Problems*. The University Series in Mathematics. Plenum Press, New York (cit. on pp. 6, 91, 92, 175, 195).

Tröltzsch, F. (2010). *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. Translated from the German by Jürgen Sprekels. American Mathematical Society, Providence, RI (cit. on pp. 48, 201, 263, 266).

Tsao, J., Boesiger, P., and Pruessmann, K. P. (2003). *k-t BLAST and k-t SENSE: Dynamic MRI with high frame rate exploiting spatiotemporal correlations.* Magn Reson Med 50.5, pp. 1031–1042. DOI: 10.1002/mrm.10611 (cit. on p. 334).

Uecker, M., Hohage, T., Block, K. T., and Frahm, J. (2008). *Image reconstruction by regularized nonlinear inversion – joint estimation of coil sensitivities and image content.* Magn Reson Med 60.3, pp. 674–682. DOI: 10.1002/mrm.21691 (cit. on pp. 325, 327).

Uecker, M., Karaus, A., and Frahm, J. (2009). *Inverse reconstruction method for segmented multishot diffusion-weighted MRI with multiple coils.* Magn Reson Med 62.5, pp. 1342–1348. DOI: 10.1002/mrm.22126 (cit. on p. 325).

Uecker, M., Zhang, S., and Frahm, J. (2010). *Nonlinear inverse reconstruction for real-time MRI of the human heart using undersampled radial FLASH.* Magn Reson Med 63.6, pp. 1456–1462. DOI: 10.1002/mrm.22453 (cit. on p. 325).

Uecker, M., Zhang, S., Voit, D., Karaus, A., Merboldt, K.-D., and Frahm, J. (2010). *Real-time MRI at a resolution of 20 ms.* NMR Biomed 23, pp. 986–994. DOI: 10.1002/nbm.1585 (cit. on p. 334).

Ulbrich, M. (2002). *Semismooth Newton methods for operator equations in function spaces.* SIAM J. Optim. 13.3, 805–842 (2003). DOI: 10.1137/S1052623400371569 (cit. on pp. 15, 17, 75, 217, 227, 265, 267, 298).

Ulbrich, M. (2011). *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces.* Vol. 11. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. DOI: 10.1137/1.9781611970692 (cit. on pp. 12, 14).

van Veen, P., Schouwink, J. H., Star, W. M., Sterenborg, H. J., van der Sijp, J. R., Stewart, F. A., and Baas, P. (2001). *Wedge-shaped applicator for additional light delivery and dosimetry in the diaphragmal sinus during photodynamic therapy for malignant pleural mesothelioma.* Phys Med Biol 46.7, pp. 1873–1883. DOI: 10.1088/0031-9155/46/7/310 (cit. on p. 311).

Vossen, G. and Maurer, H. (2006). *On L^1-minimization in optimal control and applications to robotics.* Optimal Control Appl. Methods 27.6, pp. 301–321. DOI: 10.1002/oca.781 (cit. on pp. 18, 61).

Wachsmuth, D. and Wachsmuth, G. (2011a). *Convergence and regularization results for optimal control problems with sparsity functional.* ESAIM: Control Optim. Calc. Var. 17, pp. 858–886. DOI: 10.1051/cocv/2010027 (cit. on pp. 18, 90, 107).

Wachsmuth, D. and Wachsmuth, G. (2011b). *Regularization error estimates and discrepancy principle for optimal control problems with inequality constraints.* Control and Cybernetics 40.4, pp. 1125–1154. URL: http://www.tu-chemnitz.de/mathematik/part_dgl/publications/Wachsmuth_Wachsmuth__On_the_regularization_of_optimization_problems_with_inequality_constraints.pdf (cit. on pp. 18, 90).

Warga, J. (1972). *Optimal Control of Differential and Functional Equations.* Academic Press, New York (cit. on pp. 5, 131).

Werner, D. (2011). *Funktionalanalysis.* 7th ed. Springer-Verlag, Berlin (cit. on p. 2).

Widrow, B. and Kollár, I. (2008). *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications.* Cambridge University Press, Cambridge, UK. URL: http://www.mit.bme.hu/books/quantization/ (cit. on p. 288).

Williams, J. and Kalogiratou, Z. (1993a). *Least squares and Chebyshev fitting for parameter estimation in ODEs*. Adv. Comput. Math. 1.3-4, pp. 357–366. DOI: 10.1007/BF02072016. (Cit. on p. 289).

Williams, J. and Kalogiratou, Z. (1993b). *Nonlinear Chebyshev fitting from the solution of ordinary differential equations*. Numer. Algorithms 5.1-4. Algorithms for approximation, III (Oxford, 1992), pp. 325–337. DOI: 10.1007/BF02108466. (Cit. on p. 289).

Wolke, R. and Schwetlick, H. (1988). *Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons*. SIAM J. Sci. Statist. Comput. 9.5, pp. 907–921. DOI: 10.1137/0909062 (cit. on pp. 238, 250).

Wright, R. C., Riederer, S. J., Farzaneh, F., Rossman, P. J., and Liu, Y. (1989). *Real-time MR fluoroscopic data acquisition and image reconstruction.* Magn Reson Med 12.3, pp. 407–415. DOI: 10.1002/mrm.1910120314 (cit. on p. 330).

Xiang, Q.-S. (2005). *Accelerating MRI by skipped phase encoding and edge deghosting (SPEED).* Magn Reson Med 54.5, pp. 1112–1117. DOI: 10.1002/mrm.20453 (cit. on p. 326).

Xie, J. and Zou, J. (2002). *An improved model function method for choosing regularization parameters in linear inverse problems*. Inverse Problems 18.3, pp. 631–643. DOI: 10.1088/0266-5611/18/3/307 (cit. on p. 217).

Xu, D., King, K. F., and Liang, Z.-P. (2007). *Improving k-t SENSE by adaptive regularization.* Magn Reson Med 57.5, pp. 918–930. DOI: 10.1002/mrm.21203 (cit. on p. 334).

Xu, H., Dehghani, H., Pogue, B. W., Springett, R., Paulsen, K. D., and Dunn, J. F. (2003). *Near-infrared imaging in the small animal brain: Optimization of fiber positions*. J. Biomed. Opt. 8.1, pp. 102–110. DOI: 10.1117/1.1528597 (cit. on p. 311).

Yamamoto, M. and Zou, J. (2001). *Simultaneous reconstruction of the initial temperature and heat radiative coefficient*. Inverse Problems 17.4, pp. 1181–1202. DOI: 10.1088/0266-5611/17/4/340 (cit. on p. 254).

Yang, J., Zhang, Y., and Yin, W. (2009). *An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise*. SIAM J. Sci. Comput. 31.4, pp. 2842–2865. DOI: 10.1137/080732894 (cit. on pp. 217, 238, 250, 251, 253).

Yeh, W. W.-G. (1986). *Review of parameter identification procedures in groundwater hydrology: The inverse problem*. Water Resource Research 22.2, pp. 95–108. DOI: 10.1029/WR022i002p00095 (cit. on p. 255).

Yin, W., Goldfarb, D., and Osher, S. (2007). *The total variation regularized L^1 model for multiscale decomposition*. Multiscale Model. Simul. 6.1, pp. 190–211. URL: http://handle.dtic.mil/100.2/ADA460529 (cit. on pp. 216, 253).

Ying, L. and Sheng, J. (2007). *Joint image reconstruction and sensitivity estimation in SENSE (JSENSE).* Magn Reson Med 57.6, pp. 1196–1202. DOI: 10.1002/mrm.21245 (cit. on p. 325).

Zhang, S., Block, K. T., and Frahm, J. (2010). *Magnetic resonance imaging in real time: Advances using radial FLASH.* J Magn Reson Imaging 31.1, pp. 101–109. DOI: 10.1002/jmri.21987 (cit. on p. 330).

Ziemer, W. P. (1989). *Weakly differentiable functions*. Vol. 120. Graduate Texts in Mathematics. Sobolev spaces and functions of bounded variation. Springer-Verlag, New York (cit. on p. 13).

Zuazua, E. (2007). *Controllability and observability of partial differential equations: Some results and open problems*. In: Handbook of differential equations: Evolutionary equations. Vol. III. Elsevier/North-Holland, Amsterdam, pp. 527–621. DOI: 10.1016/S1874-5717(07)80010-7 (cit. on pp. 38, 194).