A PROJECTION AND A VARIATIONAL REGULARIZATION METHOD FOR SPARSE INVERSE PROBLEMS

Von der Fakultät Mathematik und Physik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

JONAS OFFTERMATT

aus Stuttgart

Hauptberichter:	Prof. Dr. Barbara Kaltenbacher
1. Mitberichter:	Prof. Dr. Bernd Hofmann
2. Mitberichter:	Prof. Dr. Kunibert G. Siebert

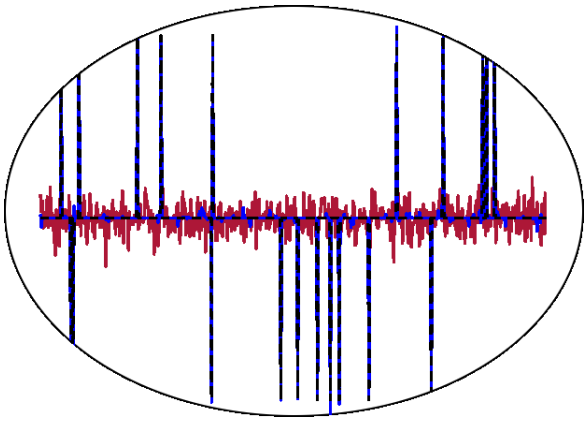
Tag der mündlichen Prüfung: 4. Mai 2012

Institut für Stochastik und Anwendungen
der Universität Stuttgart

2012

A PROJECTION AND A VARIATIONAL REGULARIZATION METHOD FOR SPARSE INVERSE PROBLEMS

JONAS OFFTERMATT



Dissertation zur Erlangung des Doktorgrades

Institut für Stochastik und Anwendungen
Fachbereich Mathematik
Universität Stuttgart

May 2012

SimTech Cluster of Excellence

University of Stuttgart
Germany

Jonas Offtermatt: *A Projection and a Variational Regularization Method for Sparse Inverse Problems,* Dissertation zur Erlangung des Doktorgrades, © May 2012

SUPERVISORS:
Prof. Dr. Barbara Kaltenbacher
Prof. Dr. Bernd Hofmann
Prof. Dr. Kunibert G. Siebert

LOCATION:
Stuttgart

TIME FRAME:
May 2012

Dedicated to my enlarging family.

ABSTRACT

The solution of sparse inverse problems has become a highly active topic over the past decade. This thesis aims at providing new methods for the regularization of sparse and possibly ill-posed inverse problems. In this work a projection and a variational regularization method for the solution of sparse inverse problems are presented. The description and analysis of each of these two methods is complemented by an additional related topic.

The projection method, developed in Chapter 4, is based on an adaptive regularization method for a distributed parameter in a parabolic Partial Differential Equation (PDE), originally introduced by Chavent and coauthors [10, 17]. Here we adapt this approach for general sparse inverse problems. Furthermore a well-definedness result is presented and it is proven that the minimizer achieved by the algorithm solves the original problem in a least squares sense. Additionally, we illustrated the efficiency of the algorithm by two numerical examples from applications in systems biology and data analysis.

The sequence of subspaces adaptively chosen by the introduced algorithm leads us to the analysis of regularization by discretization in preimage space. This regularization method is known to convergence only under additional assumptions on the solution. In Chapter 5 regularization by discretization in case of noisy data under a suitable source condition is considered. We present some results of well-definedness, stability and convergence for linear and nonlinear inverse problems in case the regularization subspace is chosen by the discrepancy principle.

In Chapter 6 the second main part of this thesis starts. There we present a variational method for sparse inverse problems. Before introducing a new regularization functional, we take a closer look at Bayesian regularization theory. We give a brief introduction and present the connection between deterministic Tikhonov regularization and stochastic Bayesian inversion in case of Gaussian densities, developed in [51]. Then we discuss the convergence results from [44] for the stochastic theory, which are based on this close connection. Also we outline a concept for a general convergence result and prove a generalization result for the existence of a $\mathcal{R}$-minimizing solution. Again we illustrate the gained results with some numerical examples.

We use the close connection between stochastic and deterministic regularization to develop a new regularization functional for sparse inverse problems in Chapter 7. There we establish well-definedness, stability and convergence proofs for this functional,

based on the results from [48]. Additionally, we prove convergence rates for the new functional. However, only in a generalized Bregman distance introduced in [31], as the generated regularization term is not convex. The proposed functional is differentiable and thus can be used in gradient based optimization methods, e. g., a Quasi Newton method. We illustrate the efficiency and accuracy of this approach again with some numerical examples.

The thesis starts with a general and detailed introduction into inverse problems. First a motivation and introduction to inverse problems is given in Chapter 1. Then a brief overview over recent results in regularization theory is presented in Chapter 2. Finally Chapter 3 closes the introductory part with a motivation and some first notations on sparsity in inverse problems.

ZUSAMMENFASSUNG

In dieser Dissertationsschrift werden neue Methoden und Aspekte der Regularisierungstheorie für dünnbesetzte, schlecht gestellte inverse Probleme diskutiert. Nachfolgend werden eine Projektions- sowie eine Variationsmethode zur Lösung von dünnbesetzten inversen Problemen vorgestellt. Ergänzend dazu wird jeweils ein verwandtes Thema besprochen.

Die Projektionsmethode wird in Kapitel 4 entwickelt und basiert auf einer adaptiven Regularisierungsmethode. Diese wurde ursprünglich von Chavent et al. [17, 10] vorgeschlagen, um einen verteilten Parameter in einer parabolischen partiellen Differentialgleichung zu identifizieren. Wir passen den Ansatz für allgemeine, dünnbesetzte, inverse Probleme an. Für den dadurch entstehenden Algorithmus zeigen wir die Wohldefiniertheit des projizierten Minimierungsproblems. Zusätzlich wird bewiesen, dass die Minimalstelle, welche der Algorithmus berechnet, das ursprüngliche Problem in Sinne des kleinsten quadratischen Fehlers löst. Um die Effizienz der Methode zu veranschaulichen, wird der Algorithmus auf ein Problem in der Systembiologie, sowie ein Problem in der Datenanalyse angewendet.

Die adaptiv erzeugte Folge von Unterräumen, welche der obige Algorithmus erzeugt, führt uns zum Thema des nächsten Kapitels, der Regularisierung durch Diskretisierung im Urbildraum. Obwohl bekannt ist, dass diese Methode nur unter Zusatzvoraussetzungen an die Lösung konvergiert, wird sie aufgrund ihrer einfachen Implementation häufig verwendet. In Kapitel 5 beweisen wir neue Ergebnisse zur Wohldefiniertheit, Stabilität und Konvergenz für Regularisierung durch Diskretisierung im Urbildraum für lineare, sowie nichtlineare inverse Probleme bei

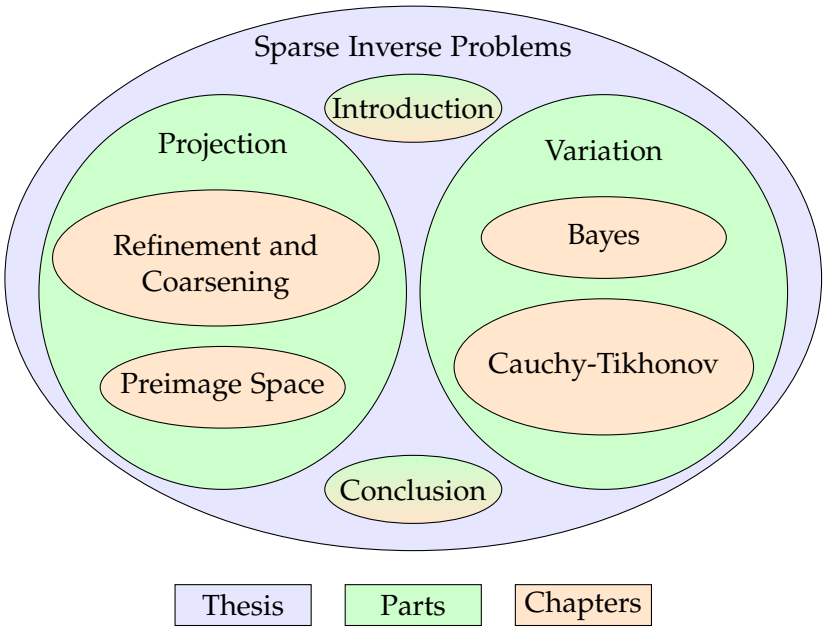
verrauschten Daten. Dabei setzen wir eine entsprechende Glattheit der Lösung voraus, sowie dass das Diskrepanzprinzip zur Wahl des regularisierenden Unterraums verwendet wird.

Mit Kapitel 6 beginnt der zweite große Teil dieser Arbeit. In diesem Teil wird ein neues Regularisierungsfunktional für dünnbesetzte inverse Probleme vorgestellt. Bevor wir allerdings dieses neue Funktional besprechen, werden wir genauer auf die stochastische Bayes Regularisierung eingehen. Zunächst gibt es eine kurze Einführung in das Thema, bevor die für das nächste Kapitel wichtige Verbindung zwischen stochastischer und deterministischer Regularisierung gezeigt wird. Diese Verbindung wurde schon in [51] von Kaipo und Somersalo ausführlich behandelt, sowie von Hofinger und Pikkarainen [44] verwendet, um Konvergenz der Posteriorverteilung gegen eine Punktverteilung zu zeigen. Wir diskutieren dieses Konvergenzkonzept und skizzieren eine mögliche Erweiterung für allgemeine Verteilungen. Zusätzlich beweisen wir eine Verallgemeinerung des Existenzresultats für $\mathcal{R}$-minimierende Lösungen (vgl. Lemma 2.4). Auch dieses Kapitel beinhaltet numerische Beispiele um die theoretischen Resultate zu veranschaulichen.

Daraufhin wird im nächsten Kapitel die Verbindung zwischen stochastischer und deterministischer Regularisierung verwendet, um ein neues Dünnbesetztheit förderndes Funktional zu erzeugen. Wir zeigen in diesem Kapitel die Wohldefiniertheit, Stabilität und Konvergenz für ein Tikhonov-Funktional mit diesem neuen Regularisierungsterm. Dazu verwenden wir die schon bekannten Ergebnisse für allgemeine Tikhonovregularisierung auf Banachräumen, vgl. [48]. Darüber hinaus beweisen wir Konvergenzraten für das so entstandene Regularisierungsfunktional. Da der erzeugte Regularisierungsterm nicht konvex ist, geschieht dies in einer verallgemeinerten Bregman-Distanz, welche von Grasmair, in [31], für nicht konvexe Funktionale eingeführt wurde. Das vorgeschlagene Funktional ist differenzierbar und kann somit in effizienten Gradienten basierten Optimierungsverfahren verwendet werden. Wir zeigen anhand einiger numerischer Beispiele, wie ein solches Verfahren funktioniert und vergleichen die Ergebnisse mit bereits bekannten Methoden für dünnbesetzte inverse Probleme.

Die vorliegende Arbeit beginnt mit einer allgemeinen Einführung in die Thematik der dünnbesetzten inversen Probleme. So wird in Kapitel 1 zuerst einmal eine Einführung, sowie Motivation für die Betrachtung von allgemeinen schlecht gestellten inversen Problemen gegeben. Daraufhin werden in Kapitel 2 erste Lösungsmethoden für inverse Probleme vorgestellt. Bevor schließlich anhand zweier Beispiele die Notationen, sowie die Relevanz von dünnbesetzten inversen Problemen veranschaulicht wird.

Contents of the Chapters



Part I *(Motivation)*

Chapter 1 General Introduction

An introduction into the theory of ill-posed inverse problem is given. The associated difficulties are demonstrated by some well-known examples.

Chapter 2 Regularization Theory

A brief introduction into regularization theory is given. The focus is on Tikhonov regularization and regularization by discretization. The previously introduced examples are solved with the proposed methods.

Chapter 3 Sparse Inverse Problems

The term sparsity is introduced and again illustrated by well-known examples. Additionally it is demonstrated, that standard techniques are not capable of solving sparse inverse problems.

Part II *(Projection)*

Chapter 4 Refinement and Coarsening Algorithm

A possible method to solve sparse inverse problems is established. The algorithm solves the original problem on adaptively chosen subspaces to generate a sparse solution.

Chapter 5 Regularization in Preimage Space

Well-posedness of regularization in preimage space is shown and illustrated with a numerical example, in case the regularization subspace is chosen by the discrepancy principle.

Part III *(Variation)*

Chapter 6 Excursus in Bayesian Inversion Theory

A brief introduction into Bayesian regularization of inverse problems is given, also a generalized convergence concept is proposed and illustrated with a numerical example.

Chapter 7 Cauchy Functional

Based on the connection between Bayesian and Tikhonov regularization, a new sparsity enforcing regularization functional is developed and well-posedness of the generated Tikhonov functional is proven.

Chapter 8 Conclusion and Outlook

A short summary over all chapters is given, as well as an outlook onto possible generalizations and enhancements.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

A Refinement and Coarsening Indicator Algorithm for Finding Sparse Solutions of Inverse Problems, Barbara Kaltenbacher and Jonas Offtermatt, *Inverse Problems and Imaging*, Volume 5, No. 2, 2011, 391-406.

A convergence analysis of regularization by discretization in preimage space, Barbara Kaltenbacher and Jonas Offtermatt, accepted by: *Mathematics of Computation*, also Preprint series of the DFG Cluster of Excellence Simulation Technology (2010 -44).

A convergence analysis of Tikhonov regularization with the Cauchy regularization term, Jonas Offtermatt and Barbara Kaltenbacher, submitted to: *Journal of Inverse and Ill-Posed Problems*, also Preprint series of the DFG Cluster of Excellence Simulation Technology.

Convergence of posteriors for structurally non-identifiable problems using results from the theory of inverse problems, Nicole Radde and Jonas Offtermatt, submitted to: *Journal of Applied Probability*, also Preprint series of the DFG Cluster of Excellence Simulation Technology.

Nur kannst du in Zeiten wie diesen,
wenn du orientierungslos durch einen dunklen Wald irrst,
der abstrakten Deduktion vertrauen.
Wenn du in die Knie gezwungen wirst,
dann knie dich hin und verehre die Doppel-S-Kurve.
Wage den Sprung des Glaubens in die Arme von Peano,
Leibniz, Hilbert und L'Hôpital. Du wirst Trost finden.
Fourier, Gauss, Laplace, Rickey. Aufgehoben.
Niemals fallen gelassen. Wiener, Riemann, Frege, Green.

— David Foster Wallace, Unendlicher Spass

ACKNOWLEDGMENTS

I want to thank my family, my colleagues and my supervisors for their constant support and their confidence in my skills.

CONTENTS

LIST OF TABLES

ACRONYMS

FEM Finite Element Method

IST Iterative Soft Thresholding

MAP Maximum A Posteriori Estimate

MLE Maximum Likelihood Estimate

MPMLE Maximum Prior Maximum Likelihood Estimate

ODE Ordinary Differential Equation

SSN Semi Smooth Newton

PDE Partial Differential Equation

Part I

INTRODUCTION TO INVERSE PROBLEMS

1 GENERAL INTRODUCTION

This work deals mainly with the solution of sparse and ill-posed inverse problems. Before going into details about the term sparse or how to solve such problems, there will be a short introduction into the topic of inverse problems. Many problems in physics and mathematics can be formulated as an operator equation

$$F(x) = y. \tag{1.1}$$

Here $F : \mathcal{X} \to \mathcal{Y}$ denotes a nonlinear[1] forward operator which maps the solution x to the data y. The operator maps from a Banach or Hilbert space $\mathcal{X}$ to a Banach or Hilbert space $\mathcal{Y}$. Usually one deals not with the exact data y, but with a noisy version y^δ, for which we assume

$$\|y - y^\delta\| \leqslant \delta, \tag{1.2}$$

with $\delta > 0$. The noise level δ indicates, how perturbed the data is. It can be interpreted as the distance between the exact and the noisy data. We will later on see, why such a bound is needed and how we can use the information given in (1.2).

Given a solution x and returning the data y, is called the forward or direct problem. Obviously we call the other way round, gathering a solution out of given data, an inverse problem. For example when dealing with a PDE containing parameters, given the parameters and calculating a solution, is the direct problem. Identifying the parameters from given measurements of the solution, is the inverse problem. In general the direct problem is a well-known and well-studied problem, whereas the inverse problem is not so well-studied and most times also not well-posed. But what does this mean "well-posed"?

A problem is called well-posed according to Hadamard [39] if:

(W1) For all admissible data, a solution exists.

(W2) For all admissible data, the solution is unique.

(W3) The solution depends continuously on the data.

A problem is called ill-posed if one of the above given items is violated.

The first item is fulfilled, if y lies in the range of F, $R(F) = \mathcal{Y}$. In other words all $y \in \mathcal{Y}$ are attainable and therefore a solution

1 Throughout this work F will generally denote a *nonlinear* operator, whereas A will denote a *linear* forward operator.

x that fulfills (1.1) always exists. The solution is unique in the linear case, if the kernel of the forward operator consists of the zero element only, i. e., $N(F) := \{x \in \mathcal{X} : F(x) = 0\} = 0$. We further note that the last condition for well-posedness, the continuous dependence of the solution on the data, is violated in case that F^{-1} is discontinuous, which can for instance occur if $\mathcal{X}$ and/or $\mathcal{Y}$ are infinite dimensional.

The condition that y is attainable, is very restrictive. We can overcome the lack of attainability by introducing a generalized notion of solution. Actually this idea was first introduced by Carl Friedrich Gauss, who used it to calculate the position of the dwarf planet Ceres. The idea is not to search for an exact solution but to search for the element of $\mathcal{X}$ which minimizes the quadratic distance to the given data. In mathematical terms:

Definition 1.1. *$\bar{x} \in \mathcal{X}$ is called a least squares solution of $F(x) = y$ if*

$$\|F(\bar{x}) - y\| = \inf\{\|F(z) - y\| \,|\, z \in \mathcal{X}\}.$$

This generalization helps if item (W1) of the above noted well-posedness definition is violated. But it will not necessarily lead to a unique solution. For a unique solution we have to further generalize the notion of a solution, which we will do in the next chapter. Before we do so we will have a look at two short examples of ill-posed operators, to illustrate what happens if the inverse of the forward operator is not continuous, i. e., (W3) from above is violated. We will come back to these examples later on. Afterwards we generalize the solution concept and introduce regularization theory, which can overcome the difficulties of ill-posed problems.

Example 1 (Numerical Differentiation). This problem is probably the best-known example of an inverse problem (see cf. [26, 47, 41]). Obviously differentiation and integration are inverse to each other. But only differentiation exhibits the typical characteristics of an ill-posed problem. Let $y(t)$ be any function in $C^1([0,1])$, $\delta \in (0,1)$, $n \in \mathbb{N}(n \geqslant 2)$ arbitrary and define the noisy data as

$$y^\delta(t) := y(t) + \delta \sin(\frac{nt}{\delta}), \quad t \in [0,1].$$

Then the derivative is given by

$$x^\delta(t) = y^{\delta\prime}(t) = y'(t) + n\cos(\frac{nx}{\delta}).$$

Whereas the exact solution is $x^\dagger = y'(t)$. Therewith the difference between noisy data and exact data in the uniform norm is just – as we assumed generally –

$$\|y(t) - y^\delta(t)\|_\infty = \delta,$$

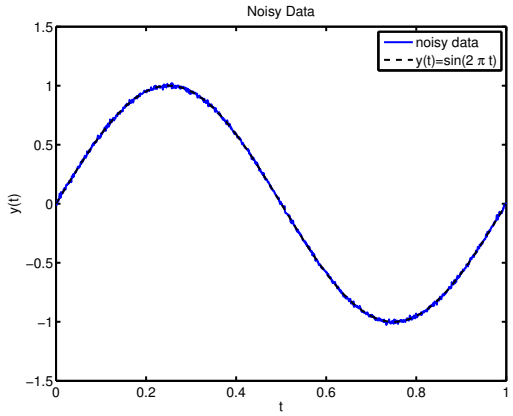


Figure 1: Plot of exact and noisy data for the numerical differentiation example. The dotted black line is the exact data $y(t) = \sin(2\pi t)$ and the solid blue line is the noisy data. A small amount of white Gaussian noise is added.

but the difference between the solutions is

$$\|x^{\delta}(t) - x^{\dagger}(t)\|_{\infty} = n.$$

And as n is chosen arbitrary, the difference can be arbitrarily large. But one has to keep in mind, that this arbitrary large difference depends heavily on the choice of the norm. Using instead of the uniform norm the C^1-norm would not lead to such a large error, see the remark on page 5 in [26].

In terms of an operator equation, differentiation can be modeled as a linear inverse problem with forward operator

$$Ax(s) = \int_{[0,s]} x(t)dt = y(t),$$

which is an integral operator of the first kind.

The phenomenen of discontinuity of the inverse of the forward operator can be pretty visualized by numerical differentiation. Lets assume we have $T = 1000$ noisy measurements of the function $y(t) := \sin(2\pi t)$ $(t \in [0, 1])$, see Figure 1, where a slight amount of Gaussian random noise is added to the exact data. However there is almost no difference visible to the naked eye between the exact data and the measurements.

Now we can apply the inverse of the forward operator A to the noisy measurements. We compare this solution to the easily calculated exact solution $x(t) = 2\pi\cos(2\pi t)$, see Figure 2. Here the difference between calculated and exact solution is massive.

So why do these small measurement errors lead to such a wrong solution? The answer is given by the condition of the inverse of the forward operator A. If we use – as we did for the plots – finite differences for the calculation of the derivative and a

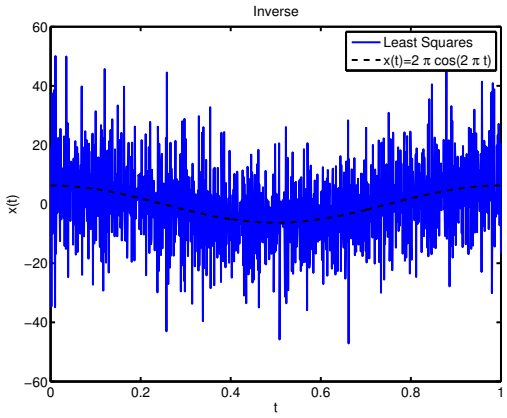


Figure 2: Solution of the inversion of A and the exact solution $x(t) = 2\pi\cos(2\pi t)$. The very small measurement errors are hugely amplified.

simple trapezoidal rule for the integration, the forward operator and its inverse are given by the matrices

$$A = h \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{pmatrix} \qquad A^{-1} = \frac{1}{h} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

both $\in \mathbb{R}^{T\times T}$ and $h = 1/T$ the distance between the measurement points. Now the only eigenvalue of A is just h and therefore the only eigenvalue of A^{-1} is $\frac{1}{h}$. So if h is very small the condition number of the matrix A is very large and small errors in the measurements are largely amplified. Such problems are called *ill-conditioned*, instead of ill-posed. As the inverse of A is not discontinuous.

Example 2 (Parameter Identification). As a second example we look at a parameter identification problem in a PDE setting. This problem is also widely considered and we will come back to it later on a few times.

We want to identify the source term q in the elliptic boundary value problem

$$\begin{aligned} \Delta u &= q && \text{in } \Omega \\ u &= 0 && \text{on } \partial\Omega \end{aligned} \tag{1.3}$$

on the unit square $\Omega = [0,1]\times[0,1] \subset \mathbb{R}^2$ from measurements of u in the whole domain Ω. This leads to a well-studied linear inverse problem

$$Aq = u\,, \tag{1.4}$$

where $q \in \mathcal{X} = L^2(\Omega)$, $u \in \mathcal{Y} = L^2(\Omega)$, and $A : \mathcal{X} \to \mathcal{Y}$.

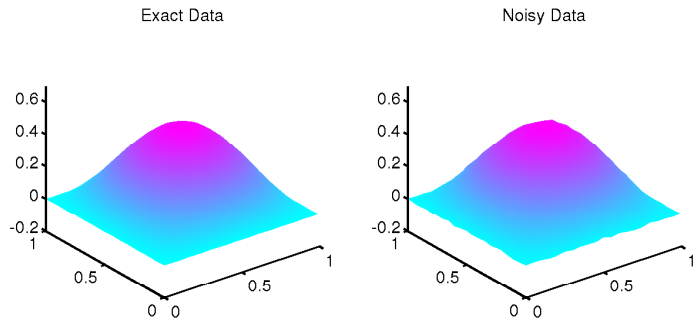


Figure 3: On the left the exact solution u of the Poisson equation (1.3) with sinusoid parameter. To the right a plot of the noisy data u^δ, with 1% random Gaussian noise.

Obviously the range of the forward operator A, mapping a parameter q to the exact solution of the Poisson problem (1.3) is actually $H^2(\Omega) \cap H_0^1(\Omega)$. On the other hand, only values but not derivatives of u can be measured, so the natural choice of Y as a Hilbert space is $L^2(\Omega)$. This discrepancy of smoothness yields an ill-posedness of degree two of the inverse problem, which can be seen directly from the fact that we consider q and u in the same space, but q is determined from u by application of the Laplace operator, i. e., twice differentation. (For a definition of the degree of ill-posedness of an inverse problem see Definition 2.42 in [46]).

Again we visualize the ill-posedness of the inverse problem with a plot of the exact parameter and the parameter obtained as a least squares solution to (1.4). We solve the forward PDE problem by using the Finite Element Method (FEM). Hence Ω is discretized with a triangle mesh and we provide an exact parameter by defining the value at the midpoint of every triangle by the function

$$q^\dagger(s,t) = 10\sin(\pi s)\sin(\pi t) \quad s \in [0,1]\,,\ t \in [0,1].$$

The corresponding exact solution of (1.3) can be seen in Figure 3. In comparison the noisy data is plotted, where we added Gaussian noise with a noise level of 1%. The least squares solution is plotted in Figure 4.

As in the above example the least squares solution and the exact parameter look completely different. Again the noise is hugely amplified, such that the result has nothing in common with the chosen exact parameter.

We have seen in these two examples that ill-posed inverse problems can not be solved by just applying the inverse of the forward operator or just solving the corresponding normal equation to get a least squares solution. So now the question arises, how to overcome the problem of discontinuity of the inverse operator. We will see in the next chapter, that regularization theory is one

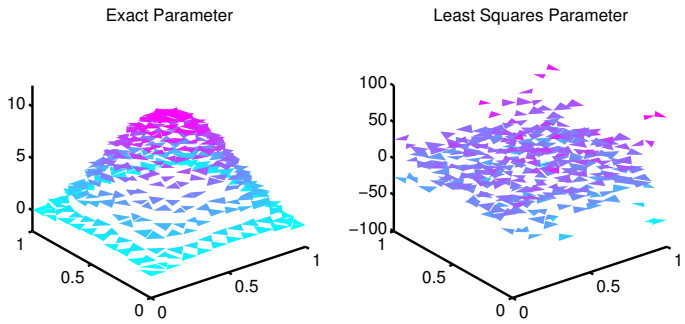


Figure 4: Left: The exact parameter $q^\dagger(s,t) = 10 \cdot \sin(\pi s)\sin(\pi t)$. Right: Least squares solution for a noise level of 1%.

way to overcome this problem. The key idea of regularization theory is to approximate the inverse of the forward operator F by a family of operators which map continuously from the data to the solution.

2 REGULARIZATION OF INVERSE PROBLEMS

In this chapter we will give a brief introduction to regularization theory of inverse problems. As here is not the place to give a comprehensive overview, we will restrict ourselves to a short overview on Tikhonov regularization and regularization by projection, two techniques we will revisit later on. The reader interested in a broader overview is referred to the monographs [26, 60, 47, 63, 4, 37]. The below given exposition of Tikhonov regularization follows very closely the analysis given in [73]. The part on regularization by projection is mainly taken from [26].

2.1 TIKHONOV REGULARIZATION

As said above the key idea of regularization theory is to approximate the (generalized) inverse of the forward operator by a family of operators, which map continuously between Y and X. In 1977 Tikhonov proposed to minimize the following functional [87, 86], instead of minimizing the least squares functional:

$$\mathcal{J}_\alpha(x) := \mathcal{S}(F(x), y^\delta) + \alpha \mathcal{R}(x) \tag{2.1}$$

Here $\mathcal{S}$ is a general data fitting term, measuring the error between $F(x)$ and the noisy data y^δ. α is the so-called regularization parameter, which so to say controls the amount of regularization. And finally $\mathcal{R}$ is a non-negative functional, called the regularization functional or regularization term, see Table 1 and Chapter 3 in [73] for examples of $\mathcal{S}$ and $\mathcal{R}$. Through adding the additional term, the approximation of the least squares solution is stabilized.

We will analyze different different types of regularization functionals $\mathcal{R}$ throughout this work, whereas $\mathcal{S}$ in the following will be the squared norm on the space Y, as proposed by Gauss, i. e., $\mathcal{S}(y, \bar{y}) = \|y - \bar{y}\|_Y^2$.

Additionally we will see, that the choice of α is a crucial task. Obviously for small α the solution will be determined by the data fitting term $\mathcal{S}$ and with growing α the influence of the additional regularization term is amplified, steering the solution towards the minimizer of $\mathcal{R}$.

With the above introduced Tikhonov functional the following questions arise:

- Is there a minimizer for every $\alpha > 0$ and every $y^\delta \in Y$?
- Does the minimizer x_δ^α depend continuously on y^δ?

Name	$\mathcal{J}_\alpha$	Description	References
Standard Tikhonov	$\mathcal{J}_\alpha := \|F(x) - y^\delta\|_Y^2 + \alpha\|x\|_X^2$	standard Tikhonov regularization with squared error norms	[26, 25]
Tikhonov with ℓ_p norm	$\mathcal{J}_\alpha := \|F(x) - y^\delta\|_{\ell_2}^2 + \alpha\|x\|_{\ell_p}^p$	the ℓ_p norm ($1 \leqslant p < 2$) enhances sparsity	[19, 34, 35]
Total Variation Regularization	$\mathcal{J}_\alpha := \frac{1}{2}\|F(x) - y^\delta\|_{L^2(\Omega)}^2 + \alpha \int_\Omega \|\nabla x\| d\mu$	introduced to remove noise from images	[78, 16]
Maximum Entropy Regularization	$\mathcal{J}_\alpha := \frac{1}{2}\|F(x) - y^\delta\|_{L^2(\Omega)}^2 + \alpha \int_\Omega x \log(\frac{x}{x_0}) d\mu$	the entropy is a measure introduced in statistics and information theory	[26, 3]
Kullback-Leibler functional	$\mathcal{J}_\alpha := KL(y^\delta, F(x)) + \alpha KL(x, x_0)$	for $\mathcal{S}$ and $\mathcal{R}$ the Kullback-Leibler divergence is used, i. e., $KL(x_1, x_2) := \int_\Omega \left(x_1 \log(\frac{x_1}{x_2}) - x_1 + x_2 \right) d\mu$	[76, 73, 50]

Table 1: Some examples of different Tikhonov type functionals

- If $\mathcal{S}(y, y^\delta) < \delta$ and $\alpha, \delta \to 0$, does the regularized solution x_δ^α converge to a solution of (1.1)?
- Can we give an estimate how fast the minimizer x_δ^α converges towards the solution?

In mathematical terms, these questions refer to the terms of *well-definedness, stability, convergence* and *convergence rates*. In [73] and [48] sufficient conditions for $\mathcal{S}, \mathcal{R}$ and F are given to answer the above questions. We recall these conditions, as we will return to them later on, in Chapter 7.

Condition 1 (Assumption 1.3 in [73])**.** *We assume*

1. *X and Y are vector spaces, with which there are associated topologies τ_X and τ_Y.*
2. *The following conditions on $\mathcal{S} : Y \times Y \to [0, \infty)$ hold:*
   a) *τ_Y is weaker than the topology induced by $\mathcal{S}$, i. e., if $\mathcal{S}(y_k, y) \to 0$ then $y_k \to_{\tau_Y} y$.*
   b) *$(y, \bar{y}) \to \mathcal{S}(y, \bar{y})$ is sequentially lower semi-continuous with respect to the τ_Y topology, i. e., for $y_k \to_{\tau_Y} y$ and $\bar{y_k} \to_{\tau_Y} \bar{y}$,*
   $$\mathcal{S}(y, \bar{y}) \leqslant \liminf_{k \to \infty} \mathcal{S}(y_k, \bar{y_k}).$$
   c) *$\lim_{k\to\infty} \mathcal{S}(y, y_k) = 0$ implies that for every $\bar{y} \in Y$ with $\mathcal{S}(\bar{y}, y) < \infty$, $\mathcal{S}(\bar{y}, y_k) \to \mathcal{S}(\bar{y}, y)$, (hence $\mathcal{S}(\bar{y}, y)$ is bounded).*
   d) *$\mathcal{S}(y, \bar{y}) = 0$ is equivalent to $y = \bar{y}$.*
3. *$F : \mathcal{D}(F) \subset X \to Y$ is continuous with respect to the topologies τ_X and τ_Y.*
4. *$\mathcal{R} : X \to [0, +\infty]$ is proper and τ_X lower semi-continuous.*
5. *$\mathcal{D}(F)$ is closed with respect to τ_X and $\mathcal{D} := \mathcal{D}(F) \cap \mathcal{D}(\mathcal{R}) \neq \emptyset$.*
6. *For every $\alpha > 0$, $y \in Y$ and $M > 0$ the level sets*
   $$\mathcal{M}_{\alpha,y} := \{x \in X : \mathcal{J}_{\alpha,y} \leqslant M\}$$
   are τ_X-sequentially compact. That is, every sequence (x_k) in $\mathcal{M}_{\alpha,y}(M)$ has a subsequence, which is convergent in U with respect to the τ_X-topology.

Before we start answering the above rose questions, we have to come back to one of the urgent difficulties with ill-posed problems. As said above one condition for well-posedness of a problem is uniqueness of the solution. In general the solution of (1.1) is not unique, and even the introduced generalized solution concept, i. e., the least squares solution, is not unique. Therefore we introduce the term of $\mathcal{R}$-*minimizing* solutions, that is:

Definition 2.1. *We call an element $x^\dagger \in \mathcal{D}$ an $\mathcal{R}$-minimizing solution if*

$$\mathcal{R}(x^\dagger) = \min\{\mathcal{R}(x) : F(x) = y\} < \infty.$$

The concept of an $\mathcal{R}$-minimizing solution generalizes the concept of a minimum norm solution, used in the general Hilbert space setting.

Now we will answer the question if Tikhonov regularization is well-defined.

2.1.1 *Well-definedness*

Lemma 2.2 (Theorem 1.6 in [73])**.** *Let Condition 1 hold. Assume that $\alpha > 0$ and $y^\delta \in \mathcal{Y}$, then there exists a minimizer of $\mathcal{J}_\alpha$.*

As this and the following lemmas are nowadays quite common results, we will only state the key ingredients of the proofs, rather than citing the full proof.

Proof. As $\mathcal{D} \neq \emptyset$ and $y^\delta \in \mathcal{Y}$ there is at least one $x \in \mathcal{X}$ such that $\mathcal{J}_\alpha < \infty$. Therefore there is a sequence $(x_k) \in \mathcal{D}$ such that

$$\lim_{k\to\infty} \mathcal{J}_\alpha(x_k) = c,$$

with $c := \inf\{\mathcal{J}_\alpha(x) : x \in \mathcal{D}\}$. This sequence is bounded and therefore has a $\tau_\mathcal{X}$ convergent subsequence, whose limit is denoted by $\hat{x}$. Now with the $\tau_\mathcal{X}$ lower semi continuity, continuity of F, and the $\tau_\mathcal{X}, \tau_\mathcal{Y}$ sequentially lower continuity of $\mathcal{R}$ and $\mathcal{S}(F(\cdot), y^\delta)$ it follows

$$\begin{aligned}\mathcal{S}(F(\hat{x}), y^\delta) + \alpha\mathcal{R}(\hat{x}) &\leqslant \liminf_{k\to\infty} \mathcal{S}(F(x_k), y^\delta) + \alpha \liminf_{k\to\infty} \mathcal{R}(x_k) \\ &\leqslant \liminf_{k\to\infty} \left(\mathcal{S}(F(x_k), y^\delta) + \alpha\mathcal{R}(x_k)\right).\end{aligned}$$

Therefore $\hat{x}$ minimizes $\mathcal{J}_\alpha$. □

2.1.2 *Stability*

Lemma 2.3 (Theorem 1.7 in [73])**.** *Let Condition 1 hold. Assume $\lim_{k\to\infty} \mathcal{S}(y_k, y^\delta) \to 0$, then every sequence (x_k) satisfying*

$$x_k \in \arg\min\{\mathcal{S}(F(x), y_k) + \alpha\mathcal{R}(\hat{x}) : x \in \mathcal{D}\} \tag{2.2}$$

has a $\tau_\mathcal{X}$ convergent subsequence and the limit of this subsequence $\hat{x}$ is a minimizer of $\mathcal{J}_\alpha$.

Additionally for every subsequence (x_l), which converges with respect to $\tau_\mathcal{X}$, $\mathcal{R}(x_l) \to \mathcal{R}(\hat{x})$.

Proof. From the definition of x_k it follows

$$\mathcal{S}(F(x), y_k) + \alpha\mathcal{R}(x_k) \leqslant \mathcal{S}(F(x^\dagger), y_k) + \alpha\mathcal{R}(x^\dagger),$$

for $x^\dagger$ an $\mathcal{R}$-minimizing solution. Now with boundedness of $\mathcal{S}$, i. e., Item 2c) in Condition 1 and Item 6 we get that (u_k) is bounded and therefore has a τ_X convergent subsequence. The limit is denoted by $\hat{x}$.

Again with continuity of F and τ_X, τ_Y lower semi continuity of $\mathcal{S}$ and $\mathcal{R}$ we get the following inequalities

$$\begin{aligned}
\mathcal{S}(F(\hat{x}), y^\delta) + \alpha\mathcal{R}(\hat{x}) &\leqslant \liminf_{k\to\infty} \mathcal{S}(F(x_k), y_k) + \alpha \liminf_{k\to\infty} \mathcal{R}(x_k) \\
&\leqslant \limsup_{k\to\infty} \left(\mathcal{S}(F(x_k), y_k) + \alpha\mathcal{R}(x_k)\right) \\
&\leqslant \lim_{k\to\infty} \left(\mathcal{S}(F(x), y_k) + \alpha\mathcal{R}(x)\right) \\
&\leqslant \mathcal{S}(F(x), y^\delta) + \alpha\mathcal{R}(x),
\end{aligned}$$

for any $x \in \mathcal{D}$.

Thus $\hat{x}$ is a minimizer. Setting $x = \hat{x}$ on the right hand side, leads to

$$\mathcal{S}(F(\hat{x}), y^\delta) + \alpha\mathcal{R}(\hat{x}) = \lim_{k\to\infty} \left(\mathcal{S}(F(x_k), y_k) + \alpha\mathcal{R}(x_k)\right). \tag{2.3}$$

Convergence of $\mathcal{R}(x_l)$ to $\mathcal{R}(\hat{x})$ is shown by contradiction. Please keep in mind that, because of the continuity of F and the τ_Y lower continuity of $\mathcal{S}$, the following holds:

$$\mathcal{S}(F(\hat{x}), y^\delta) \leqslant \liminf_{k\to\infty} (\mathcal{S}(F(x_k), y_k). \tag{2.4}$$

Assume now $\mathcal{R}(x_l)$ does not converge to $\mathcal{R}(\hat{x})$. Since $\mathcal{R}$ is τ_X lower semi continuous, we can conclude that

$$c := \limsup_{l\to\infty} \mathcal{R}(x_l) > \mathcal{R}(\hat{x}).$$

Now take a subsequence, for simplicity again denoted by (x_l), such that $\mathcal{R}(x_l) \to c$. But from (2.3) we deduce

$$\lim_{l\to\infty} \left(\mathcal{S}(F(x_l), y_l\right) = \mathcal{S}(F(\hat{x}), y^\delta) + \alpha(\mathcal{R}(\hat{x}) - c) < \mathcal{S}(F(\hat{x}), y^\delta).$$

This contradicts (2.4). Thus $\mathcal{R}(x_l) \to \mathcal{R}(\hat{x})$. □

2.1.3 *Existence*

Lemma 2.4 (Theorem 1.9 in [73]). *Let Condition 1 hold. If there exists a solution of (1.1) then there exists an $\mathcal{R}$-minimizing solution.*

Remark 1. In Chapter 6 we will prove a generalization of this lemma, in such a way that only the existence of a least squares solution has to be assumed, see Proposition 6.5.

Proof. The proof goes analogously to the one of Lemma 2.2 with $\mathcal{J}_\alpha$ replaced by $\mathcal{R}$ and $\mathcal{D}$ replaced by

$$\tilde{\mathcal{D}} := \{x \in \mathcal{D} : F(x) = y\},$$

using the fact that for $x \in \tilde{\mathcal{D}}$, we have $\mathcal{J}_\alpha(x) = \mathcal{R}(x)$. □

2.1.4 *Convergence*

Lemma 2.5 (Theorem 1.10 in [73]). *Let Condition 1 hold. Assume there exists a solution of (1.1). Moreover assume that*

- *the sequence (δ_k) converges monotonically to zero,*
- *$y_k := y^{\delta_k}$ satisfies $\mathcal{S}(y, y_k) \leqslant \delta_k$,*
- *the parameter choice $\alpha(\delta)$ satisfies*

$$\alpha(\delta) \to 0 \text{ and } \frac{\delta}{\alpha(\delta)} \to 0 \text{ as } \delta \to 0, \tag{2.5}$$

- *and $\alpha(\cdot)$ is monotonically decreasing.*

Then a sequence (x_k) satisfying (2.2) exhibits a $\tau_{\mathcal{X}}$ convergent subsequence, and the limit is an $\mathcal{R}$-minimizing solution. If additionally the $\mathcal{R}$-minimizing solution $x^\dagger$ is unique, then $x_k \to x^\dagger$ with respect to $\tau_{\mathcal{X}}$.

Proof. We define $\alpha_k := \alpha(\delta_k)$ and $\alpha_{\max} = \alpha_1$.

Since x_k is defined as sequence of minimal arguments, we conclude

$$\mathcal{S}(F(x_k), y_k) + \alpha_k \mathcal{R}(x_k) \leqslant \delta_k + \alpha_k \mathcal{R}(x^\dagger). \tag{2.6}$$

From this we can already deduce $\lim_{k\to\infty} \mathcal{S}(F(x_k), y) = 0$ and $\limsup_{k\to\infty} \mathcal{R}(x_k) \leqslant \mathcal{R}(x^\dagger)$. With these inequalities it follows that $\mathcal{J}_\alpha(x_k)$ is bounded. Thus according to Item 6 in Condition 1 (x_k) has a $\tau_{\mathcal{X}}$ convergent subsequence, with limit $\hat{x}$.

Now using the $\tau_{\mathcal{X}} - \tau_{\mathcal{Y}}$ continuity of F, inequality (2.6) and item 2 a) of Condition 1, we can deduce $F(\hat{x}) = y$.

With the lower semi continuity with respect to the $\tau_{\mathcal{X}}$ topology of $\mathcal{R}$ it follows that $\hat{x}$ is a $\mathcal{R}$-minimizing solution and additionally $\mathcal{R}(x_k) \to \mathcal{R}(x^\dagger)$.

Finally the strong convergence follows, as in the Hilbert space setting, with a subsequence-subsequence argument, and taking advantage of the uniqueness of the $\mathcal{R}$-minimizing solution. □

2.1.5 *Convergence Rates*

For proving convergence rates a source condition is needed, meaning that the solution is restricted to have a sufficient degree of smoothness (cf. [26]). For nonlinear operators, the operator is

assumed to be Fréchet differentiable and additionally an assumption on the nonlinearity (a tangential cone / Scherzer condition) is needed. Very recently, cf. [48], variational inequalities where introduced to show convergence rates for nonlinear Tikhonov regularization in Banach spaces. Moreover it has been shown [73], that all of the assumptions and source conditions stated above imply a variational inequality. Hence we will restrict our analysis to the assumption of a variational inequality and will only state these results without giving any proofs.

Before we introduce variational inequalities we have to introduce the definition of Bregman distances. Bregman distances can be used to measure distances in case of regularization in Banach spaces [77].

Definition 2.6. *The Bregman distance for a convex and proper functional $\mathcal{R} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, with subgradient $\xi \in \partial\mathcal{R}(x)$, is defined at $x \in \mathcal{X}$ and $\xi \in \partial\mathcal{R}(x) \subset \mathcal{X}^*$ by*

$$D_{\mathcal{R}}(\tilde{x}; x) := \mathcal{R}(\tilde{x}) - \mathcal{R}(x) - \langle \xi, \tilde{x} - x \rangle_{\mathcal{X}^*,\mathcal{X}}, \quad \tilde{x} \in \mathcal{X}.$$

The set

$$\mathcal{D}_B(\mathcal{R}) := \{x \in \mathcal{D}(\mathcal{R}) : \partial\mathcal{R}(x) \neq \emptyset\}$$

is called the Bregman domain.

Remark 2. Keep in mind that the functional in the above definition has to be convex, to ensure that the right hand side is greater or equal to zero. We will later on see, how to generalize the definition of the Bregman distance so that also non-convex functionals can be covered. See the definition of *W*-convexity in Section 7.3 of Chapter 7.

With the concept of Bregman distances, we can introduce a variational inequality, see e. g., [77, 48].

Definition 2.7. *We say a variational inequality holds, if there exist numbers $\beta_1, \beta_2 \in [0, \infty)$ with $\beta_1 < 1$ and $\xi \in \partial\mathcal{R}(x^\dagger)$ such that*

$$\langle \xi, x^\dagger - x \rangle \leqslant \beta_1 D_{\mathcal{R}}(x; x^\dagger) + \beta_2 \mathcal{S}\left(F(x), F(x^\dagger)\right) \tag{2.7}$$

for all $x \in \mathcal{M}_{\alpha,y}$.

With this variational inequality convergence rates for nonlinear Tikhonov regularization on Banach spaces can be proved. But additional to Condition 1, we have to assume that the functional $\mathcal{R}$ is convex, the data fitting functional $\mathcal{S}$ fulfills the triangle inequality, i. e.,

$$S(y_1, y_2) \leqslant S(y_1, y_2) + S(y_2, y_3), \quad \forall y_1, y_2, y_3 \in \mathcal{Y},$$

and that the $\mathcal{R}$-minimizing solution is an element of the Bregman domain of $\mathcal{R}$, see Assumption 1.13 in [73] for details.

Lemma 2.8 (Theorem 1.14 in [73]). *Let*

$$\mathcal{J}_\alpha := \mathcal{S}(F(x), y^\delta)^r + \alpha \mathcal{R}(x),$$

and Condition 1 hold. Additionally let $\mathcal{X}, \mathcal{Y}$ be Banach spaces with duals $\mathcal{X}^*, \mathcal{Y}^*$, the functional $\mathcal{R}$ be convex, and $\mathcal{S}$ satisfy the triangle inequality. If there exists a $\mathcal{R}$-minimizing solution $x^\dagger$, which is an element of the Bregman domain of $\mathcal{R}$, $\mathcal{S}(y, y^\delta) \leqslant \delta$ and if a variational inequality (2.7) holds, then we have

- *in case $r > 1$: For $\alpha : (0, \infty) \to (0, \infty)$ satisfying $c\delta^{r-1} \leqslant \alpha(\delta) \leqslant C\delta^{r-1}$ $(0 < c \leqslant C)$*

$$D_{\mathcal{R}}(x_\delta^\alpha, x^\dagger) = \mathcal{O}(\delta) \text{ and } \mathcal{S}(F(x_\delta^\alpha), y^\delta) = \mathcal{O}(\delta).$$

- *in case $r = 1$: For $\alpha : (0, \infty) \to (0, \infty)$ satisfying $c\delta^\epsilon \leqslant \alpha(\delta) \leqslant C\delta^\epsilon$ $(0 < c \leqslant C, 0 < \epsilon < 1)$*

$$D_{\mathcal{R}}(x_\delta^\alpha, x^\dagger) = \mathcal{O}(\delta^{1-\epsilon}) \text{ and } \mathcal{S}(F(x_\delta^\alpha), y^\delta) = \mathcal{O}(\delta).$$

Proof. See proof of Theorem 1.14 in [73]. □

We will see later on how this convergence rates result can be generalized also for non-convex regularization terms, see Theorem 7.13.

As we have seen in Example 1, if we just apply the inverse of the forward operator A onto the noisy data the resulting solution does not have anything in common with the exact solution. But what happens if we use Tikhonov regularization to solve the inverse problem?

It can be shown in an easy calculation (see Theorem 5.1 in [26]) that in the linear case the minimizer of the Tikhonov functional is given by

$$x_\delta^\alpha = (A^*A + \alpha I)^{-1} A^* y^\delta. \tag{2.8}$$

Now one can see the stabilizing effect of α. The inversion of A^*A is approximated through inverting $(A^*A + \alpha I)$. Additionally through adding α, the smallest singular value is bounded from below by α. See Figure 5 for a plot of the regularized solution for the numerical differentation problem using Tikhonov regularization in comparison to the least squares solution.

For Figure 5 the regularization parameter α has been tuned by hand to get the best possible result. The choice of the regularization parameter has a crucial influence on the computed approximation. In Figure 6 there are two approximations gained by Tikhonov regularization with other parameter values for the numerical differentiation example. Once the regularization parameter α is chosen very large, then the approximation is determined by the regularization term alone and therefore tends

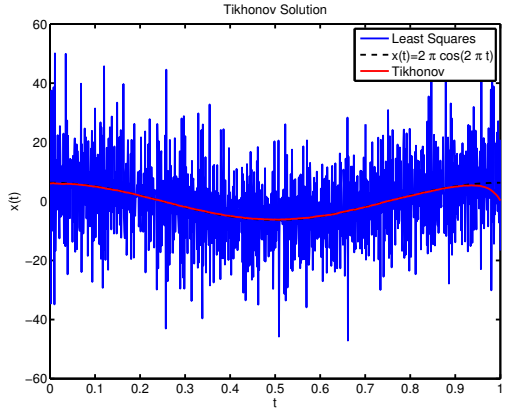


Figure 5: Tikhonov solution of the numerical differentiation example. Additionally the exact solution and the least squares solution are depicted. It can be seen clearly that the noise amplification is damped or even vanishes in the regularized solution.

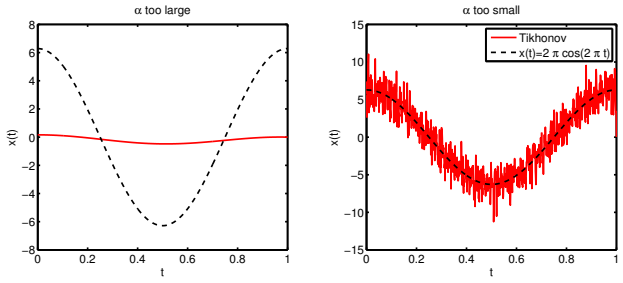


Figure 6: For the left plot the parameter α was chosen too large, such that the approximate solution x_δ^α tends towards zero. In the right picture an approximation with α chosen too small is depicted, here the solution tends towards the least squares solution.

to 0. And once the parameter is chosen very small, then the approximation tends towards the least squares solution.

So of course the question arises, how to choose the regularization parameter α? There are many possible parameter choice rules (see Chapter 4 in [26] or the review [8]). One widely used rule for theoretical results is the Discrepancy Principle [68]. As we will use it later on in Chapter 5, we will state it here. The idea is to choose α as large as possible, in such a way that the discrepancy between $F(x_\delta^\alpha)$ and y^δ is about δ. Hence α is chosen by

$$\alpha(\delta) := \sup\{\alpha > 0 \,|\, \|F(x_\delta^\alpha) - y^\delta\| \leqslant \tau\delta\}, \tag{2.9}$$

with $\tau > 1$.

In Chapter 5 we used the discrepancy principle to choose a finite subspace of $\mathcal{X}$, on which we restrict the forward operator. This technique – solving the inverse problem on different subspaces – is called regularization by projection or discretization and we will give a brief introduction into this topic in the next section.

2.2 REGULARIZATION BY PROJECTION

The key idea of regularization by projection is to project the least squares minimization on a finite-dimensional subspace and hence avoid the discontinuity of the inverse of the forward operator. We can distinguish two different cases of regularization by projection. First is regularization in preimage space and second is regularization in image space. For both we will sketch the mere idea, before in Chapter 5, we take a closer look at regularization by discretization in preimage space. There the interested reader will also find references on the detailed standard works about regularization by projection.

2.2.1 *Discretization in Preimage Space*

Let

$$\mathcal{X}_1 \subset \mathcal{X}_2 \subset \mathcal{X}_3 \subset \dots$$

be a sequence of nested finite dimensional subspaces of the preimage space, whose union is dense in $\mathcal{X}$. Now we solve the least squares problem on one of these finite dimensional subspaces:

$$x_n^\delta \in \arg\min\{\|F(z_n) - y^\delta\|^2 \,|\, z_n \in \mathcal{D} \cap \mathcal{X}_n\}.$$

In the linear case, we can define $A_n := AP_n$ with P_n the orthogonal projection on to $\mathcal{X}_n$. Now as $A_n : \mathcal{X}_n \to \mathcal{Y}$, it has a closed range, and hence the operator $A_n^\dagger$, which maps the data to the

best approximate solution of the subspace $\mathcal{X}_n$, is bounded. Therefore the operator maps continuously between $\mathcal{Y}$ and $\mathcal{X}_n$, making this approach a stable approximation of $x^\dagger$.

Also one can show, that under additional assumptions on the operator A, the sequence (x_n^δ) converges towards $x^\dagger$, see Theorem 3.20 in [26].

In Chapter 5 convergence under certain conditions is shown, if the subspace on which the inversion is carried out is chosen according to the discrepancy principle for linear as well as nonlinear operators (5 and [57]).

The concept of regularization by discretization in image space is now quite obvious.

2.2.2 *Discretization in Image Space*

Regularization by discretization in image space (also called the dual least squares method or self regularization) has been well investigated in the literature (see, e. g., [69, 72, 89] for the linear case, as well as [49, 52] for the nonlinear case). Instead of taking a sequence of subspaces of the preimage space, we take a sequence of subspaces of the image space $\mathcal{Y}$. Consider a nested sequence of finite dimensional subspaces

$$\mathcal{Y}_1 \subset \mathcal{Y}_2 \subset \mathcal{Y}_3 \subset \dots$$

of $\overline{R(A)} \subset \mathcal{Y}$. Now we define x_n^δ as the least squares solution of minimal norm of the equation

$$A_n x = y_n, \quad A_n := Q_n A,\ y_n := Q_n y,$$

with Q_n the orthogonal projection onto $\mathcal{Y}_n$. Again this is a stable approximation of $x^\dagger$.

In contrast to regularization in preimage space, regularization in image space is guaranteed to converge, without additional assumption on the forward operator, see Theorem 3.24 in [26].

We illustrate regularization by discretization in preimage space by the above introduced parameter identification problem, cf. Example 2. As we solve the underlying PDE with the FEM the different subspaces correspond to different mesh sizes. Below is a picture of the least squares solution of (1.4) on a much coarser mesh than in Example 2, see Figure 7.

The approximation on the coarser mesh is still not a very good approximation of the exact parameter, depending also on the very large mesh size. But at least the approximation on the coarse mesh resembles the exact parameter better than the approximation on the finer mesh, where strong noise propagation occurs.

In the above sections we have introduced different methods to regularize ill-posed inverse problems. In the next chapter, we

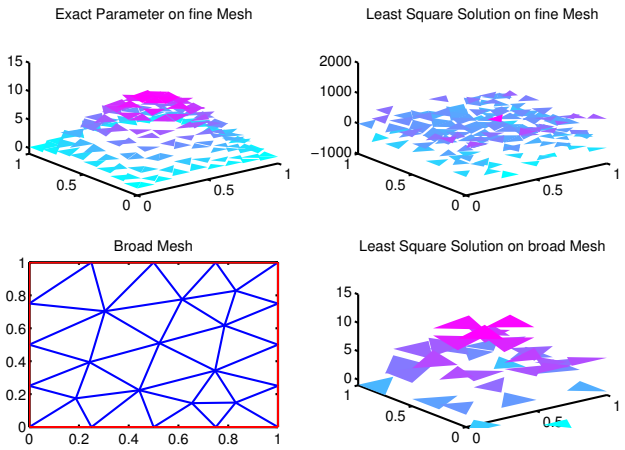


Figure 7: Comparison of least squares approximations on different meshes. Clearly the inversion on the coarser mesh is also a difficult task (lower row to the right), but the approximation is much better than the one carried out on the finer mesh (upper row right side). At least the scaling of the exact parameter (left picture upper row) is almost reached.

will consider a special class of inverse problems with special requirements on the solution, for which the above techniques fail almost in the same way as simply calculating the least squares solution to an ill-posed problem.[1]

1 In fact this statement is not exactly true. We will see later on, how Tikhonov regularization can be adapted, such that it can also handle the kind of problems we introduce in the next chapter.

3 SPARSE INVERSE PROBLEMS

In the recent years the regularization of sparse inverse problems has received much attention. Starting with carrying over the results from compressed sensing [22] onto infinite dimensional inverse problems by Daubechies et al. [19], the focus shifted towards the regularization properties of Tikhonov regularization with sparsity enforcing regularization terms, cf. [93, 33, 12, 80]. Also iterative methods [19, 11] and projection methods [58] have been analyzed.

So what makes sparse inverse problems so special? Or first of all, what is a sparse inverse problem? We consider here inverse problems as sparse, if the solution $x^\dagger$ of (1.1) has a sparse representation with respect to a given basis, that is:

$$x^\dagger = \sum_{i \in I} x_i \phi_i.$$

Here $(\phi_i)_{i \in \mathbb{N}}$ is a (orthonormal) basis of the separable Hilbert space[1] $\mathcal{X}$ and I is a finite set. This means, many of the coefficients x_i of $x^\dagger$ are zero or stated the other way round: $x^\dagger$ has only finitely many non-zero coefficients with respect to the basis $(\phi_i)_{i \in \mathbb{N}}$. The choice of the underlying basis is crucial for sparsity. A vector maybe sparse with respect to one basis, but does not have a sparse representation in a different basis.

A first example for a sparse inverse problem can be implemented with the numerical differentiation problem, introduced above in Example 1. If the exact solution $x(t)$ is, instead of the continuous sine function, a piecewise constant function (a square wave) we get a nice sparse inverse problem. See Figure 8 for a picture of the exact piecewise constant function.

Now as we know already from Example 1 above, that the inversion of the forward integral operator is an ill-conditioned problem, we use standard Tikhonov regularization with squared error norm on L^2 to stabilize the inversion. In Figure 9 you find a picture of the regularized solution, with α tuned by hand to get a best possible result.

At first sight the Tikhonov solution does not give a completely wrong approximation, but there are two main features of the exact solution the Tikhonov approximation is totally lacking. Firstly

1 In case of sparse inverse problems $\mathcal{X}$ generally denotes a separable Hilbert space, such that there are at least countable many basis functions. Please be aware that the results on regularization theory mentioned before are also valid even on non-separable Banach spaces and that some of the results stated later on, are also valid on Banach spaces.

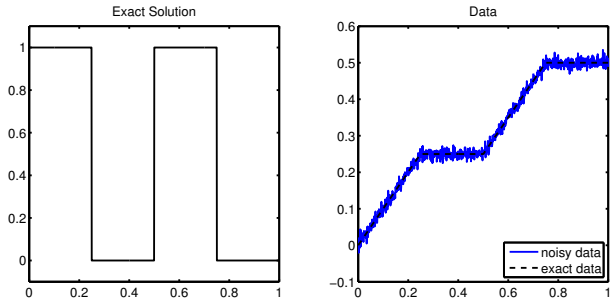


Figure 8: A sparse exact solution for the inverse problem of numerical differentiation (left) and the according exact and noisy data.

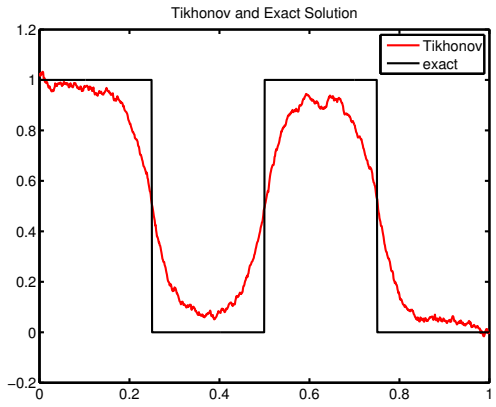


Figure 9: Comparison between Tikhonov approximation and the exact solution.

the sharp edges of the exact solution. The square wave is by no means a continuous function, whereas the approximation by Tikhonov regularization with squared error norm regularization term provides a quite smooth estimate. And secondly the exact solution is composed out of a large part, where the function is zero and a second part where it is non-zero (in this case 1). Comparing the Tikhonov approximate there are hardly any zeros at all.

This loss of zeros can be seen even better in an also very well-known and often used example of a sparse inverse problem. The problem of compressed sensing [23].

Example 3. In compressed sensing one wants to reconstruct a signal from few linear measurements. Our test case is implemented in the ℓ_1 magic packet from Candes and Romberg [15] and also used in e. g., [35, 62, 28]. We create a measurement operator $A \in \mathbb{R}^{K\times N}$, $K << N$, fill it up with standard Gaussian random numbers and orthogonalize the rows. Through this orthogonalization the matrix fulfills the restricted isometry property [7], which means that all submatrices with a small number of rows have singular values close to one.

Additionally a signal of length N, with only T randomly distributed spikes with height ± 1 is created, see Figure 10. Then the data is constructed by multiplying x by A and adding some Gaussian random noise, $y = Ax + e$.

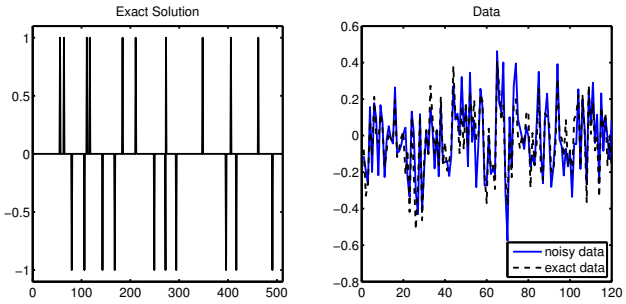


Figure 10: Plot of the exact solution (left), a signal of 20 non-zero peaks and the according exact and noisy data (right).

As before we apply standard Tikhonov regularization with squared error norm regularization functional to the data and compare the resulting approximation to the exact solution, see Figure 11.

Again the Tikhonov minimizer does not give a good approximate to the exact solution. As in the case of the numerical differentiation there are hardly any zeros and the spikes of the exact signal are not even detectable from the approximation.

So how to solve sparse inverse problems? A first hint has already been given in the introductory paragraph to this section

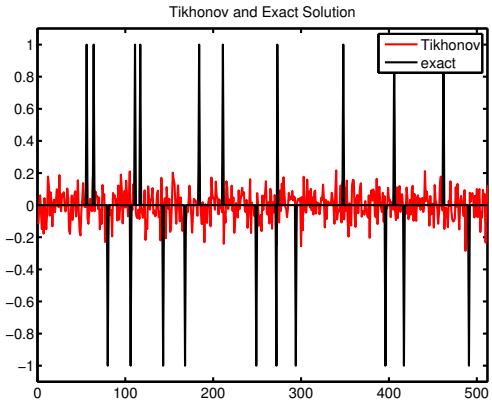


Figure 11: Comparison between Tikhonov minimizer and the exact signal.

and in Table 1: we might use Tikhonov regularization with a different regularization term $\mathcal{R}$ which enforces sparsity. Prominent examples are the ℓ_1 norm [19, 35, 85, 74] and starting from there also the ℓ_p norms with $0 \leqslant p < 2$ were examined [33, 34, 93], whereas one has to keep in mind, that for $p < 1$ the resulting regularization term is no norm and non-convex, hence the results from Chapter 2 can not be applied. In Chapter 7 of this work, we will also present a regularization term, which enforces the Tikhonov approximate to be sparse and which additionally is differentiable.

Another way to solve sparse inverse problems are greedy algorithms [65, 20, 88]. Here the idea is to search for those coefficients which will lead to the largest change of the cost functional value and only change them, leaving the others unchanged, e. g., zero – on this account they are called *greedy*. In the next chapter we will present a new algorithm for sparse inverse problems which can also be considered as a greedy type of algorithm, but in another way also as regularization by projection.

Part II

SPARSITY THROUGH PROJECTION

4

REFINEMENT AND COARSENING ALGORITHM

As we have seen in the last chapter sparse inverse problems can not be solved by standard regularization techniques. In this chapter we will introduce an algorithm for efficiently solving sparse inverse problems. This algorithm is based on a well-established adaptive method proposed by Chavent and coauthors (cf., e. g., [10], [17], see also [5]) for the identification of a distributed parameter in a parabolic PDE.

Adaptive discretization has become an important task in inverse problems for PDEs. The key idea is to use as few degrees of freedom as possible to achieve a prescribed accuracy, which obviously gives a relation to sparse inverse problems. Our aim is to both extend the ideas in [10], [17] to an adaptive method for general inverse problems and to re-interpret it as a method for finding sparse solutions.

The chapter is structured as follows: In Section 4.1 we will construct an adaptive algorithm based on refinement and coarsening indicators, analyze its well-definedness and convergence, and discuss its relation to the method from [10], [17]. Afterwards we will apply the proposed method to a sparse test problem from Systems Biology. A field of research, where many inverse problems occur, cf. [27]. Here we use the algorithm to recover the structure of gene networks, see Section 4.2. Then we will come back to the compressed sensing application, introduced in Example 3 of the previous chapter.

4.1 ADAPTIVE DISCRETIZATION

As said in the introductory part on sparse inverse problems, the forward operator F of our inverse operator equation (1.1) is now assumed to map from the separable Hilbert space $\mathcal{X}$ to the not necessarily separable Hilbert space $\mathcal{Y}$. We assume throughout this chapter that an exact solution $x^\dagger$ to (1.1) exists, but is not necessarily unique, and, as usual, that there is a bound on the noise, i. e., (1.2) holds.

As $\mathcal{X}$ is separable we can assume that there is a discretization $\mathcal{X} = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_N\}$ (possibly $N = \infty$), i. e.,

$$\forall x \in X \; \exists \mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N : \quad x = \sum_{i=1}^{N} x_i \phi_i, \tag{4.1}$$

where the sum is supposed to converge in $\mathcal{X}$ if $N = \infty$ and $\|\phi_i\| = 1$. Note that $\{\phi_1, \phi_2, \dots, \phi_N\}$ need not necessarily be linear independent and we do not impose $\mathbf{x}$ to be in ℓ_2.

We consider the least square minimization problem without regularization

$$\min_{\mathbf{x}\in\mathbb{R}^{|I|}} \underbrace{\frac{1}{2}\left\|F(\sum_{i\in I} x_i\phi_i) - y^\delta\right\|^2}_{=\mathcal{J}(\mathbf{x})} \tag{P^I}$$

or equivalently

$$\min_{x\in\operatorname{span}\{\phi_i\,:\,i\in I\}} \underbrace{\frac{1}{2}\left\|F(x) - y^\delta\right\|^2}_{=J(x)} \tag{4.2}$$

for some (finite) index set $I \subseteq \{1, 2, \dots, N\}$, where $|I| = \operatorname{card}(I)$, and denote its solution(s) by x^I, as well as

$$x^I = \sum_{i=1}^{N} x_i^I \phi_i\,.$$

For simplicity of notation we identify $\mathbf{x} \in \mathbb{R}^{|I|}$ with $\mathbf{x} \in \mathbb{R}^N$ by just filling the gaps for indices outside I with zeros.
Obviously, with

$$x^\dagger = \sum_{i=1}^{N} x_i^\dagger \phi_i\,, \tag{4.3}$$

$$I^\dagger := \{i \in \{1, \dots, N\} \mid x_i^\dagger \neq 0\}\,, \tag{4.4}$$

we have that $\mathbf{x}^\dagger$ solves $(P^{I^\dagger})$ with $\delta = 0$, i. e., $y = y^\delta$.

Starting with some small index set I^0 (corresponding to a coarse discretization), we will successively add and remove indices to obtain the correct index set $I^\dagger$ and therewith, via $(P^{I^\dagger})$, the solution $x^\dagger$.

First of all, we derive a possible generalization of the *refinement indicators* from papers by Ben Ameur, Bissell, Chavent and Jaffré.

For some current index set I^k, for which we assume that we have already obtained a solution x^{I^k} of (P^{I^k}), we want to decide, whether we should add some index i_* to decrease the data misfit

$$J(x) = \frac{1}{2}\left\|F(x) - y^\delta\right\|^2\,, \tag{4.5}$$

or in terms of the coefficients

$$\mathcal{J}(\mathbf{x}) = \frac{1}{2}\left\|F(\sum_{i=1}^{N} x_i\phi_i) - y^\delta\right\|^2\,,$$

i. e., we would like to achieve

$$\min_{x\in\mathbb{R}^{|I^k\cup\{i_*\}|}} \frac{1}{2}\left\|F(\sum_{i\in I^k\cup\{i_*\}} x_i\phi_i) - y^\delta\right\|^2 < \min_{x\in\mathbb{R}^{|I^k|}} \frac{1}{2}\left\|F(\sum_{i\in I^k} x_i\phi_i) - y^\delta\right\|^2 .$$

Note that we know the minimal value (and a minimizer) on the right hand side of this inequality, and we want to predict whether this inequality is likely to hold for i_*, without having to compute the minimizer on the left hand side. The key idea is to do some linearization.

For this purpose, we consider the constrained minimization problem

$$\min_{x\in\mathbb{R}^{|I^k\cup\{i_*\}|}} \underbrace{\frac{1}{2}\left\|F(\sum_{i\in I^k\cup\{i_*\}} x_i\phi_i) - y^\delta\right\|^2}_{=\mathcal{J}((x_i)_{i\in I^k\cup\{i_*\}})} \quad \text{s.t. } x_{i_*} = \beta \quad (P_\beta^{I^k\cup\{i_*\}})$$

for some small $\beta \in \mathbb{R}$, with the Lagrange function

$$\mathcal{L}(x,\lambda) = \frac{1}{2}\left\|F(\sum_{i\in I^k\cup\{i_*\}} x_i\phi_i) - y^\delta\right\|^2 + \lambda(\beta - x_{i_*}).$$

It is readily checked that the linear independence constraint qualification holds: there is only one constraint, whose gradient is just the i_*th unit vector. Hence, the first order necessary optimality conditions imply that for a solution x_β of $(P_\beta^{I^k\cup\{i_*\}})$, there exists a λ_β such that $\nabla_x\mathcal{L}(x_\beta,\lambda_\beta) = 0$, i. e., $(x_\beta,\lambda_\beta) = 0$, i. e.,

$$\begin{aligned} 0 =& \frac{\partial\mathcal{L}}{\partial x_{i_*}}(x_\beta,\lambda_\beta) \\ =& \langle F(\sum_{i\in I^k\cup\{i_*\}} x_{\beta,i}\phi_i) - y^\delta, F'(\sum_{i\in I^k\cup\{i_*\}} x_{\beta,i}\phi_i)\phi_{i_*}\rangle \\ & - \lambda_\beta . \end{aligned} \tag{4.6}$$

And $\forall j \in I^k$:

$$\begin{aligned} 0 =& \frac{\partial\mathcal{L}}{\partial x_j}(x_\beta,\lambda_\beta) \\ =& \langle F(\sum_{i\in I^k\cup\{i_*\}} x_{\beta,i}\phi_i) - y^\delta, F'(\sum_{i\in I^k\cup\{i_*\}} x_{\beta,i}\phi_i)\phi_j\rangle \\ 0 =& \beta - x_{i_*}. \end{aligned}$$

Now we take into account the fact that the Lagrange multiplier gives the sensitivity of the optimal value with respect to perturbations in the corresponding constraint:

$$\frac{d}{d\beta}\mathcal{J}(\mathbf{x}_\beta) = \frac{d}{d\beta}\mathcal{L}(\mathbf{x}_\beta, \lambda_\beta) = \lambda_\beta$$

where in the first equality we have used feasibility of $\mathbf{x}_\beta$ for $(P_\beta^{I^k \cup \{i_*\}})$ which implies that the term multiplied by λ_β in $\mathcal{L}(\mathbf{x}_\beta, \lambda_\beta)$ vanishes.

Thus, in the first order Taylor expansion of the optimal value we have

$$\mathcal{J}(\mathbf{x}_\beta) \approx \mathcal{J}(\mathbf{x}_0) + \frac{d}{d\beta}\mathcal{J}(\mathbf{x}_0)\,\beta = \mathcal{J}(\mathbf{x}_0) + \lambda_0\,\beta \tag{4.7}$$

The quantities $\mathcal{J}(\mathbf{x}_0)$, λ_0 are known or cheaply computable: Obviously $\mathbf{x}_0$ is just a solution of (P^{I^k}), which we have already, namely the coefficients of x^{I^k}; the Lagrange multiplier λ_0 can be computed from (4.6) with $\beta = 0$:

$$\begin{aligned}\lambda_0 &= \langle F(\sum_{i\in I^k\cup\{i_*\}} x_{0,i}\phi_i) - y^\delta, F'(\sum_{i\in I^k\cup\{i_*\}} x_{0,i}\phi_i)\phi_{i_*}\rangle \\ &= \langle F(x^{I^k}) - y^\delta, F'(x^{I^k})\phi_{i_*}\rangle\,. \end{aligned}\tag{4.8}$$

By (4.7), the magnitude of λ_0 indicates, whether there can be a potential decrease in the optimal value if we add i_* to the degrees of freedom, so we define the *refinement indicator*

$$r^{i_*} := |\lambda_0| = |\langle F(x^{I^k}) - y^\delta, F'(x^{I^k})\phi_{i_*}\rangle| = |J'(x^{I^k})\phi_{i_*}| \tag{4.9}$$

Similarly to the idea of *coarsening indicators* by Ben Ameur, Bissell, Chavent and Jaffré, one can use Lagrange multipliers and first order Taylor expansion to remove degrees of freedom if possible. For this purpose, we assume that a solution $\mathbf{x}^{\tilde{I}^k}$ of $(P^{\tilde{I}^k})$ is known and consider, for some $l_* \in \tilde{I}^k$, the constrained minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^{|\tilde{I}^k|}} \underbrace{\frac{1}{2}\left\|F(\sum_{i\in\tilde{I}^k} x_i\phi_i) - y^\delta\right\|^2}_{=\mathcal{J}((x_i)_{i\in\tilde{I}^k})} \quad \text{s.t. } x_{l_*} = \gamma \tag{$\tilde{P}_\gamma^{\tilde{I}^k}$}$$

so that $\mathbf{x}^{\tilde{I}^k}$ obviously solves this problem with

$$\gamma := \gamma_* := x_{l_*}^{\tilde{I}^k},$$

and on the other hand, a solution to $(\tilde{P}_\gamma^{\tilde{I}^k})$ with $\gamma = 0$ solves $(P^{\tilde{I}^k\setminus\{l_*\}})$.

With the Lagrange function

$$\tilde{\mathcal{L}}(\mathbf{x},\mu) = \frac{1}{2}\left\| F(\sum_{i\in\tilde{I}^k} x_i\phi_i) - y^\delta \right\|^2 + \mu(\gamma - x_{l_*})$$

we get that a solution $\mathbf{x}_\gamma$ to $(\tilde{P}_\gamma^{\tilde{I}^k})$ along with some Lagrange multiplier μ_γ satisfies the first order optimality conditions

$$\begin{aligned} 0 &= \frac{\partial\tilde{\mathcal{L}}}{\partial x_{l_*}}(\mathbf{x}_\gamma,\mu_\gamma) \\ &= \langle F(\sum_{i\in\tilde{I}^k} x_{\gamma,i}\phi_i) - y^\delta, F'(\sum_{i\in\tilde{I}^k} x_{\gamma,i}\phi_i)\phi_{i_*}\rangle - \mu_\gamma. \end{aligned} \tag{4.10}$$

Additionally we get $\forall l \in \tilde{I}^k \setminus l_*$:

$$\begin{aligned} 0 &= \frac{\partial\tilde{\mathcal{L}}}{\partial x_l}(\mathbf{x}_\gamma,\mu_\gamma) \\ &= \langle F(\sum_{i\in\tilde{I}^k} x_{\gamma,i}\phi_i) - y^\delta, F'(\sum_{i\in\tilde{I}^k} x_{\gamma,i}\phi_i)\phi_l\rangle. \end{aligned}$$

And

$$0 = \gamma - x_{i_*}.$$

Analogously to (4.7), we get

$$\mathcal{J}(\mathbf{x}_0) \approx \mathcal{J}(\mathbf{x}_\gamma) - \frac{d}{d\gamma}\mathcal{J}(\mathbf{x}_\gamma)\,\gamma$$

with

$$\frac{d}{d\gamma}\mathcal{J}(\mathbf{x}_\gamma) = \frac{d}{d\gamma}\tilde{\mathcal{L}}(\mathbf{x}_\gamma,\mu_\gamma) = \mu_\gamma\,,$$

and

$$\begin{aligned} \mu_{\gamma_*} &= \langle F(\sum_{i\in\tilde{I}^k} x_{\gamma_*,i}\phi_i) - y^\delta, F'(\sum_{i\in\tilde{I}^k} x_{\gamma_*,i}\phi_i)\phi_{i_*}\rangle \\ &= \langle F(x^{\tilde{I}^k}) - y^\delta, F'(x^{\tilde{I}^k})\phi_{l_*}\rangle \end{aligned}$$

by (4.10), so that

$$c^{l_*} := \mu_{\gamma_*}\gamma_* = \langle F(x^{\tilde{I}^k}) - y^\delta, F'(x^{\tilde{I}^k})\phi_{l_*}\rangle\gamma_* = x^{\tilde{I}^k}_{l_*} J'(x^{I^k})\phi_{l_*} \tag{4.11}$$

serves as an easily computable *coarsening indicator*.

Therewith, we arrive at the following multilevel adaptive refinement and coarsening algorithm:

Algorithm 1.
1 $k := 0$.
2 Choose coarsest index set I^0.
3 Compute a solution $\mathbf{x}^0$ of P^{I^0} and set $\mathcal{J}^0 = \mathcal{J}(\mathbf{x}^0)$.
4 Compute r^{i_*}, $i_* \in \{1,2,\dots N\}\setminus I^0$ according to (4.9).

5 SET $r^0_{\max} := \max_{i_* \in \{1,2,\dots N\} \setminus I^0} r^{i_*}$.
6 WHILE $r^k_{\max} > \varepsilon$ DO
7 SET $I^* := \{i_* \in \{1,2,\dots N\} \setminus I^k \mid r^{i_*} = r^k_{\max}\}$
8 FOR ALL $i_* \in I^*$: COMPUTE A SOLUTION $\mathbf{x}^{I^k \cup \{i_*\}}$ TO $(P^{I^k \cup \{i_*\}})$
9 SELECT $i_+ \in \operatorname{argmin}_{i_* \in I^*} \mathcal{J}(\mathbf{x}^{I^k \cup \{i_*\}})$
10 SET $\tilde{I}^k = I^k \cup \{i_+\}$.
11 IF $\mathcal{J}(\mathbf{x}^{\tilde{I}^k}) < \mathcal{J}(\mathbf{x}^k)$ DO
12 COMPUTE c^{l_*}, $l_* \in I^k$ ACCORDING TO (4.11)
13 SET $c^k_{\max} := \max_{l_* \in I^k} c^{l_*}$.
14 SET $L^* := \{l_* \in \{1,2,\dots N\} \setminus I^k \mid c^{l_*} = c^k_{\max} > 0\}$
15 FOR ALL $l_* \in L^*$: COMPUTE SOLUTION $\mathbf{x}^{\tilde{I}^k \setminus \{l_*\}}$ TO $(P^{\tilde{I}^k \setminus \{l_*\}})$
16 SELECT $l_+ \in \operatorname{argmin}_{l_* \in L^*} \mathcal{J}(\mathbf{x}^{\tilde{I}^k \setminus \{l_*\}})$
17 IF $\mathcal{J}(\mathbf{x}^{\tilde{I}^k \setminus \{l_+\}}) \leqslant (1-\rho)\mathcal{J}(\mathbf{x}^k) + \rho \mathcal{J}(\mathbf{x}^{\tilde{I}^k})$ SET $I^{k+1} = \tilde{I}^k \setminus \{l_+\}$)
18 ELSE SET $I^{k+1} = \tilde{I}^k$.
19 ELSE SET $I^{k+1} = \tilde{I}^k$.
20 SET $\mathbf{x}^{k+1}$:= SOLUTION OF $(P^{I^{k+1}})$ AND $\mathcal{J}^{k+1} = \mathcal{J}(\mathbf{x}^{k+1})$.
21 COMPUTE r^{i_*}, $i_* \in \{1,2,\dots N\} \setminus I^{k+1}$ ACCORDING TO (4.9).
22 SET $r^{k+1}_{\max} := \max_{i_* \in \{1,2,\dots N\} \setminus I^{k+1}} r^{i_*}$.
23 $k = k+1$.

Remark 3. Here, line 7 might be replaced by, e. g.,
7 SELECT $I^* := \{i_* \in \{1,2,\dots N\} \setminus I^k \mid r^{i_*} \geqslant \theta r^k_{\max}\}$
or by
7 SELECT $I^* := \{i^1_*, \dots, i^s_*\}$ THE INDEX SET OF THE s LARGEST VALUES OF

$$r^{i_*},\ i_* \in \{1,2,\dots N\} \setminus I^k$$

with a typical value $\theta = 0.8$. Note that $\mathbf{x}^{k+1}$ on line 20 is one of the already computed solutions on line 8 or 15. Coarsening is only carried out if the actual misfit reduction is positive (line 11) and comparable to the reduction by refinement (line 17). If $N = \infty$ and the maximum on lines 5 and 22 is not attained, we define I^* such that $r^i \geqslant \sup_{i_* \in \mathbb{N} \setminus I^0} r^{i_*} - \varepsilon_k$ for all $i \in I^*$, where $\varepsilon_k \searrow \varepsilon$.

Note that the misfit functional J (or $\mathcal{J}$) might be replaced by any differentiable cost functional. Therewith, the proposed algorithm becomes a method for finding sparse solutions to general (unconstrained) minimization problems. Indeed, Algorithm 1 can be viewed as a special case of the following more general algorithm for minimizing a cost functional J over a separable Hilbert space X with (4.1):

Algorithm 2. .
1 CHOOSE COARSEST INDEX SET I^0.
2 COMPUTE A MINIMIZER x^0 OF J OVER $\operatorname{span}\{\phi_i : i \in I^0\}$
3 WHILE $\max\{|J'(x^k)\phi_{i_*}| : i_* \in \{1,\dots,N\} \setminus I^k\} > \varepsilon$ DO
4 COMPUTE $(\tilde{x}^k, i_+)$ AS

$$\operatorname{argmin}\{J(x) : x \in \operatorname{span}\{\phi_i : i \in I^k \cup \{i_*\}\},\ i_* \in I_*\}$$

5 WHERE $I^* = \operatorname{argmax}\{|J'(x^k)\phi_{i_*}| : i_* \in \{1,\dots,N\} \setminus I^k\}$
6 AND SET $\tilde{I}^k := I^k \cup \{i_+\}$.
7 IF $J(\tilde{x}^k) < J(x^k)$ AND $l_+ \in I^k$ CAN BE SELECTED SUCH THAT
8 $\min\{J(x) : x \in \operatorname{span}\{\phi_i : i \in \tilde{I}^k \setminus \{l_+\}\}\}$
$\leqslant (1-\rho)J(x^k) + \rho J(\tilde{x}^k)$
9 SET $I^{k+1} = \tilde{I}^k \setminus \{l_+\}$
10 AND $x^{k+1} \in \operatorname{argmin}\{J(x) : x \in \operatorname{span}\{\phi_i : i \in \tilde{I}^k \setminus \{l_+\}\}\}$
11 ELSE
12 SET $I^{k+1} = \tilde{I}^k$
13 AND $x^{k+1} = \tilde{x}^k$.

According to a standard result in optimization, the minimization problems on lines 2, 4, and 8 of Algorithm 2 (correspondingly lines 3, 8, and 15 of Algorithm 1) are solvable if J is bounded from below, coercive (i. e., from boundedness of $J(x)$ boundedness of x follows), and weakly lower semicontinuous. For the misfit functional J as in (4.5), these properties can be concluded under appropriate assumptions on F:

Lemma 4.1. *Assume that a sparse solution $x^\dagger$ to (1.1) exists, i. e., such that $I^\dagger$ in (4.3), (4.4) is finite. Let F be weakly continuous and Gateaux differentiable and let its derivative satisfy nullspace invariance*

$$\exists C > 0\, \forall x \in \mathcal{X}\, \exists R_x : \mathcal{Y} \to \mathcal{Y} : \quad F'(x) = R_x F'(x^\dagger) \text{ with } \|R_x^{-1}\| \leqslant C \tag{4.12}$$

as well as finite basis injectivity at $x^\dagger$, i. e.,

> *For any finite index set $\tilde{I} \subseteq \{1,\dots,N\}$, the restriction of $F'(x^\dagger)$ to $\operatorname{span}\{\phi_i^{(N)} : i \in \tilde{I}\}$ is injective.*

Then for any finite index set $I \subseteq \{1,\dots,N\}$ the restricted misfit functional $J|_{\operatorname{span}\{\phi_i : i \in I\}}$ with J as in (4.5) is bounded from below, coercive, and weakly lower semicontinuous.

Hence, there exists a minimizer of (4.2) and therfore, via (4.1), of (P^I).

Consequently, if I^0 is finite, then Algorithm 1 is well-defined.

Proof. J is obviously bounded from below by zero, and weak lower semicontinuity follows from weak continuity of F and weak lower semicontinuity of the norm. So it only remains to show coercivity: By the second triangle inequality and (1.2), we have

$$J(x) \geqslant \frac{1}{2}\left(\left\|F(x) - F(x^\dagger)\right\| - \delta\right)^2$$

where by Gateaux differentiability and (4.12), we can write

$$\begin{aligned}\left\|F(x) - F(x^\dagger)\right\| &= \left\|\int_0^1 R_x(x^\dagger + \theta(x - x^\dagger))\, d\theta F'(x^\dagger)(x - x^\dagger)\right\| \\ &\geqslant \frac{1}{C}\left\|F'(x^\dagger)(x - x^\dagger)\right\|\end{aligned}$$

so that finite basis injectivity applied to the index set $I \cup I^\dagger$ implies coercivity. □

Proposition 4.2. *Let the assumptions of Lemma 4.1 be satisfied and let the sequences $\mathbf{x}^k$, I^k, $\mathcal{J}^k$ be generated by Algorithm 1. Then*

(i) The sequence $\mathcal{J}^k$ of cost function values is monotonically decreasing.

(ii) Algorithm 1 stops after $K \leqslant 2^N$ (i. e., finitely many, if $\mathcal{X}$ is finite dimensional) steps.

(iii) If $\delta = 0$, $\varepsilon = 0$ then

(a) If Algorithm 1 stops after K steps (e. g., if $\mathcal{X}$ is finite dimensional, cf (ii)), then $x^K = \sum_{i \in I^K} x_i^K \phi_i$ solves (1.1) in a least squares sense, i. e., for $x^* = x^K$

$$J'(x^*) = F'(x^*)^*(F(x^*) - y) = 0 \tag{4.13}$$

(b) If Algorithm 1 does not stop after finitely many steps (hence $N = \infty$ by (ii)), then

$$\sup\{|\langle F(x^k) - y, F'(x^k)\phi_i\rangle| : i \in \mathbb{N}\} \to 0 \text{ as } k \to \infty \tag{4.14}$$

Hence, if additionally the mapping $\mathcal{X} \to l^\infty$, $x \mapsto (\langle F(x) - y, F'(x)\phi_i\rangle)_{i \in \mathbb{N}}$ is (weakly) sequentially closed, then every (weak) accumulation point x^* of $(x^k = \sum_{i \in I^k} x_i^k \phi_i)_k$ solves (4.13).

Proof. Monotonicity (i) of the cost function values follows from the fact that

- refinement increases the subspace over which the minimum is taken and
- coarsening is only done if refinement has yielded a strict decrease and coarsening does not deteriorate this decrease too much. (It is readily checked that lines 11 and 17 imply that the cost function value after coarsening is strictly lower than the one before refinement)

To see (ii) note that there are at most 2^N possiblilities for choosing I^k. Moreover, the strategy of coarsening only if the cost function is reduced, avoids cycles as follows: Assume that for some $k, m \in \mathbb{N}$, we have $I^{k+m} = I^k$, hence $\mathcal{J}^{k+m} = \mathcal{J}^k$. Due to the fact that the function values are monotonically decreasing, we have $\mathcal{J}^k = \mathcal{J}^{k+1} = \cdots = \mathcal{J}^{k+m}$, so no coarsening is done during these steps and therefore $|I^{k+m}| = |I^{k+m-1}| + 1 = \cdots |I^k| + m$, which gives a contradiction to $I^{k+m} = I^k$.

Assertion (iii)(a) follows from the fact that if the stopping criterion is reached after finitely many steps, the algorithm yields a stationary point of the misfit function: Namely, for $i \in I^K$, $\frac{\partial \mathcal{J}}{\partial x_i^K}(\mathbf{x}^K)$ vanishes since $\mathbf{x}^K$ solves (P^{I^K}), and for $i \notin I^{K_*}$, $\frac{\partial \mathcal{J}}{\partial x_i^K}(\mathbf{x}^K)$ vanishes since $0 = \lambda^i = |\frac{\partial \mathcal{J}}{\partial x_i^K}(\mathbf{x}^K)|$ by (4.6). Hence, we get, for $x^K = \sum_{i\in I^K} x_i^K \phi_i$:

$$\forall i \in \{1,\dots,N\}:\ 0 = \frac{\partial \mathcal{J}}{\partial x_i^K}(\mathbf{x}^K) = J'(x^K)\phi_i$$

hence, by taking all possible linear combinations, for all $x \in \mathcal{X}$: $J'(x^K)x = 0$, i. e., (4.13) holds.

To see assertion (iv), we use the fact that for any cost function $\hat{J}$ that is bounded from below and has Lipschitz continuous derivative on its level set $\{x \in \hat{X} :\ \hat{J}(x) \leqslant \hat{J}(x^0)\}$, to any descent direction d at x, there exists an efficient stepsize choice t_{eff} (e. g., according to the Wolfe Powell rule, see, e. g., Theorem 5.3 in [29]) such that

$$\hat{J}(\hat{x} + t_{eff} d) \leqslant \hat{J}(x) - \theta(\hat{J}'(x)d/\,\|d\|)^2$$

with $\theta \in (0,1)$ independent of the dimension of $\hat{X}$. Applying this to $\hat{X} = \operatorname{span}\{\phi_i : i \in \tilde{I}^k = I^k \cup \{i_+\}\}$ $\tilde{J} = J|_{\hat{X}}$, $x = x^k$ and $d = \phi_{i_+}$, we get that after refinement

$$\begin{aligned} J(\tilde{x}^k) &= \min\{J(x) : x \in \hat{X}\} \\ &\leqslant J(x^k + t_{eff}\phi_{i_+}) \leqslant J(x^k) - \theta|J'(x^k)\phi_{i_+}|^2 \\ &= J(x^k) - \theta(\sup\{|J'(x^k)\phi_i| : i \in \mathbb{N}\} - \varepsilon_k)^2 \end{aligned} \quad (4.15)$$

where $\tilde{x}^k = \sum_{i\in\tilde{I}^k} \tilde{x}_i^k \phi_i$, $\varepsilon_k \searrow 0$ (cf. Remark 3), and where we have used the definition of i_+ as well as the fact that $J'(x^k)\phi_i = 0$ for $i \in I^k$ in the last equality. If we do coarsening, we have

$$\begin{aligned} J(x^{k+1}) &\leqslant (1-\rho)J(x^k) + \rho J(\tilde{x}^k) \\ &\leqslant J(x^k) - \rho\theta(\sup\{|J'(x^k)\phi_i| : i \in \mathbb{N}\} - \varepsilon_k)^2 \end{aligned}$$

by (4.15), so

$$(\sup\{|J'(x^k)\phi_i| : i \in \mathbb{N}\} - \varepsilon_k)^2 \leqslant \frac{1}{\rho\theta}(J(x^k) - J(x^{k+1})\,. \quad (4.16)$$

If no coarsening is carried out, then (4.16) remains valid since $\rho \in (0,1)$. Taking the sum on both sides of (4.16) yields

$$\sum_{k=0}^{\infty} (\sup\{|J'(x^k)\phi_i| : i \in \mathbb{N}\} - \varepsilon_k)^2 \leqslant \frac{1}{\rho\theta} J(x^0)\,,$$

which implies (4.14). □

The assertions of Proposition 4.2 obviously remain valid for the more general setting of Algorithm 2 for any J that is bounded from below, coercive, and weakly lower semicontinuous.

Some further remarks are in order:

Remark 4. Obviously, the refinement indicators are just entries of the gradient vector. Nevertheless, Algorithm 1 is not just a gradient method, since no gradient steps are taken, but first order sensitivity information is only used to select degrees of freedom. The actual step is done by solving a minimization problem over the selected index set. One could possibly enhance the method by additionally using second order sensitivity information for selecting indices. On the other hand, the strategy of adding the degree of freedom that yields the largest misfit decrease reminds of a Greedy type algorithm (cf., e. g., [14], [21], [88]).

Remark 5. At this point we would like to clarify the correspondence of our setting to the concepts from [10], [9], [17]:

- The searched for quantity in [10], [9], [17] is a transmissivity function that is allowed to take different values on each *cell* of an underlying computational grid; i. e., X is the space of piecewise constant functions with possible discontinuities over all grid lines.
- In [10], [9], [17], two different ways of grouping cells are used, namely *zones* and *cuts*. There is an obviuos one-to-one correspondence between
  - specifying the values of transmissivities on each zone of a zonation and
  - specifying the jumps of these transmissivities over each interface (cut) generating this zonation.

  Here we do not distinguish between zones representation and cuts representation. Note that the ϕ_i in (4.1) need not be linear independent.
- Therewith, our degrees of freedom i correspond to cuts and the relation (4.1) corresponds to the first part of the proof of Lemma 2 in [9], which consists of showing that each cell can be represented by a linear combination of cuts.
- The natural relation between single cells (namely by spatial closeness to each other) is used to motivate coarsening in [10], [17]: There, coarsening is attempted whenever a cut divides a zone into more than two sub-zones. Since we do not distinguish between the zone representation and the cuts representation there is no analog to this condition for coarsening here. Instead we use a criterion based on cost function values (see Remark 3 above). For particular applications one might think of finding problem specific analogs to the coarsening condition from [10], [17].

Remark 6. For ill-posed operator equations

$$F(x) = y \tag{4.17}$$

on an infinite dimensional Hilbert space X_∞ with $F : X_\infty \to Y$ Algorithm 1 can be used in several ways:

- The most straightforward approach is to apply it to the Tikhonov regularized version of (4.17), i. e., to replace J by $J + \alpha\mathcal{R}$ with $\alpha > 0$ and $\mathcal{R}$ some regularization term, e. g., $\mathcal{R} = \|x - x_0\|^2$.

- As in [9], one might think of Algorithm 1 as a method for solving a finite dimensional approximation of (4.17): Considering a sequence $X_n = \operatorname{span}\{\phi_1^{(n)}, \phi_2^{(n)}, \dots, \phi_n^{(n)}\}$ of finite dimensional subspaces of X_∞, we therewith obtain a solution $x^K := x_n^*$ to the projected normal equation type problem

$$\operatorname{Proj}_{X_n} F'(x_n^*)^*(F(x_n^*) - y^\delta) = 0 \quad \text{and } x_n^* \in X_n$$

  (cf. (4.13)). Convergence to a solution of the infinite dimensional problem $F(x) = y$ in the sense of a regularization method therefore pertains to regularization by discretization in pre-image space, as introduced in Chapter 2.

- Possible regularizating properties of Algorithm 1 itself for $N = \infty$ might rely
  (a) on regularization by discretization with the special choice of the added (and removed) degrees of freedom according to the criteria on lines 7, 9, 14, 16 of Algorithm 1;
  (b) on early stopping, according to line 6 of Algorithm 1, with ε chosen in dependence of the noise level, as can be seen in the numerical experiments.

  Referring to (a) note that if coarsening is omitted, the method is similar to regularization by discretization in preimage space, but here the sequence of nested spaces

$$X^k = \operatorname{span}\{\phi_i : i \in I^k\}$$

  is not given a priori but chosen a posteriori such that $I^k = I^{k-1} + i_+$ with $|J'(x^k)\phi_{i_+}| = \max_{i\in\mathbb{N}} |J'(x^k)\phi_i|$. In this context, we mention that in the example of nonconvergence of regularization by discretization in pre-image space as given in [38], application of our refinement strategy indeed leads to convergence.

4.2 AN APPLICATION IN SYSTEMS BIOLOGY

4.2.1 *Motivation*

The reconstruction of gene networks has recently become a challenge in the area of genetics and bioinformatics. This is due

to advances in the micro-array technologies, which enable to measure gene expression levels on a genome wide scale. Using this micro-array data sets it is possible to do reverse engineering of the underlying network structures. This is in general a highly underdetermined sparse inverse problem, because only few measurements are available and not all genes interact with each other.

There have been different approaches to identify such networks, namely Bayesian methods, Boolean networks and Ordinary Differential Equation (ODE) models (see, e. g., [84] for an overview and further references)

4.2.2 *Modelling*

For modelling a gene network we used the ODE approach introduced by Yeung, Tegner and Collins [92], which was also used by [71, 84]. There, the dynamics of the gene regulation are modelled by a simple linear ODE system consisting of N equations, one for every gene. For a system of N genes, the expression level of the i-th gene is modelled by:

$$\dot{u}_i(t) = -\lambda_i u_i(t) + \sum_{j=1}^{N} w_{ij} u_j(t) + b_i(t) + \epsilon_i(t) \tag{4.18}$$

where the λ_is are the self-degradation rates of the i-th gene (cf. [2]), the b_is represent external perturbations or stimuli, and ϵ_i stands for noise. The constants w_{ij}, which denote the influence of the j-th gene on the i-th, are the most interesting ones. If $w_{ij} > 0$, the gene j activates the i-th gene, if $w_{ij} < 0$, j represses i and if $w_{ij} = 0$, there is no interaction between these two genes. As mentioned above, not all genes interact with each other, so the matrix $W = (w_{ij})_{i,j \in N}$ has only a few non-zero entries. There are some additional assumptions on gene networks. In detail, gene networks are thought as "small-world" networks, meaning they have a small characteristic path length, smaller than regular random networks, but could also have a high clustering coefficient, cf. [90, 6]. The "small-world" phenomenon was first introduced by the famous social psychologist Stanley Milgram, who examined the average path length for social networks. Another example for the "small-world" phenomena would be the Erdös Number Project, which studies research collaboration among mathematicians.

The main task in gene network reconstruction is to identify the sparse connectivity matrix $\tilde{W}$ of the N genes. With $\tilde{W} = W - \lambda I$, and $\lambda^T = (\lambda_1, .., \lambda_N)$, (4.18) can be rewritten as

$$\dot{u} = \tilde{W}u + b + \epsilon \tag{4.19}$$

where $b^T = (b_1, .., b_N)$ and $\epsilon^T = (\epsilon_1, .., \epsilon_N)$.

We test our algorithm for this model (4.19), once with zero noise and once with a small noise level. Here we set the perturbations b to zero and assume, that we have measurements for u and $\dot{u}$. Therewith, given $(u_1, \dots u_T) = (u(t_1), \dots, u(t_T))$, at T time instances $t_1 \dots, t_T$, the forward operator $A : \tilde{W} \mapsto (\dot{u}(t_1), \dots, \dot{u}(t_T))$ is linear and $\{\phi_i \ : \ i \in \{1, \dots, N^2\}\}$ is just the canonical basis of $\mathbb{R}^{N \times N}$.

4.2.3 *Network Generator*

In order to define test examples, we use the gene network generator from Steinke et al., which can be downloaded under the link given in [84]. The network generator samples small-world networks according to the description given in [1]. For the dynamics of the network see the additional file of [84]. The generator provides among others a matrix $\tilde{W}$. This matrix is sparse, incorporates "small-world" phenomena and most of its non-zero entries are in $[-1, 1]$. There are seldom entries greater than one, see the remark in the additional files to [84]. We took this matrix as the system matrix of (4.19) to get an exact solution of the inverse problem. In Figure 12 a random network with 20 genes created with the generator is visualized.

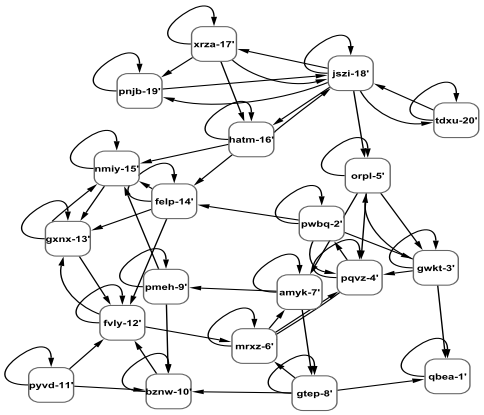


Figure 12: Picture of a random network consisting of 20 genes, created with the gene network generator in [84] and Cytoscape [82].

4.2.4 *Numerical Results*

As mentioned earlier we test Algorithm 1 for equation (4.19), with zero perturbations and first of all with exact data, assuming

that we have measurements for u and $\dot{u}$. For comparison, we also apply the wellknown Iterative Soft Thresholding (IST) algorithm from Daubechies, Defrise and DeMol, see [19], making use of the description given in [35]:

$$x_{n+1} = S_{\gamma\omega}(x_n - \gamma A^*(Ax_n - y^\delta))$$

where we choose γ such that the restriction to the forward operator $\|A^*A\| < 1$, mentioned in [19], holds. In the shrinkage operator $S_{\gamma\omega}$ defined component wise by

$$S_w(x)_i = S_w(x_i) = \max\{0, |x_i| - w\}\operatorname{sgn}(x_i)$$

we choose a small threshold $\omega = 0,001$.

As can be seen in Figure 12 every gene sampled by the network generator regulates itself. This is the case for almost every sampled network. Therefore we start Algorithm 1 not with an arbitrary index set, but with the the one corresponding to a diagonal matrix $\tilde{W}$. It is one of the advantages of the proposed method, that one can choose the coarsest index set according to some a priori knowledge about the solution. In our case, even if one of the sampled genes is not selfregulatory, the wrong edge is deleted during the iteration, because of the coarsening process.

For comparison of the two algorithms we count the number of discrepancies E between the real network $\tilde{W}_{ex}$ and the identified ones W_{RefCoa}, W_{IST}. This is done by checking the accordance of every entry of $\tilde{W}_{RefCoa}$ and $\tilde{W}_{IST}$ to the corresponding entry of $\tilde{W}_{ex}$ (the same error measure is used in [92]):

$$E = \sum_{i=1}^{N}\sum_{j=1}^{N} e_{ij}$$

with

$$e_{ij} := \begin{cases} 1 & \text{if } |\tilde{w}_{ex_{ij}} - \tilde{w}_{RefCoa_{ij}}| > \tilde{\epsilon} \\ 0 & \text{else} \end{cases}$$

and $\tilde{\epsilon}$ some prescribed value, chosen according to the noise level (for exact data we take $\tilde{\epsilon} = 10^{-3}$). In Figure 13 the number of wrong edges E versus the number of measurements T is plotted for a gene network with $N = 30$ genes. Additionally we display the required computational time for both algorithms.

In Figure 13, we can see that for a small number of measurements there are many wrong edges, whereas with increasing T the error decays until we have exact recovery of the network. The IST however is sometimes not able to detect the right solution and produces a persistently higher number of wrong edges. The computational time required with the IST grows much stronger with the number of given measurements than the time for Algorithm

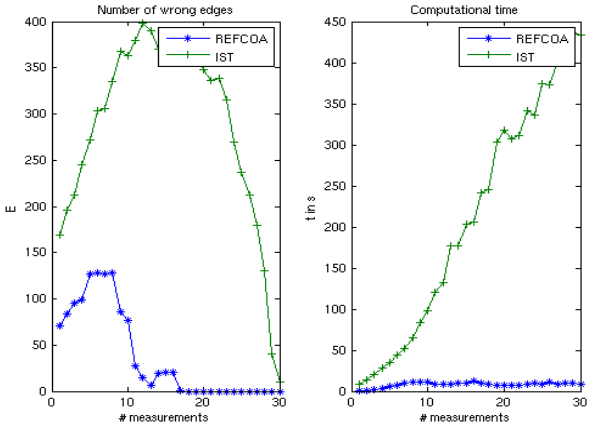


Figure 13: Number of wrong edges versus number of measurements for Algorithm 1 and for the Iterative Soft Thresholding algorithm (left); Computational time for both algorithms (right); Both for a network of 30 genes.

1. This happens because the matrix used in both algorithm gets larger with growing number of measurements. But in Algorithm 1, we only solve a smaller problem in each iteration step, depending on the used index set. With this reduced problem, the size of the used matrix in the k-th iteration step reduces from $NT \times N^2$ to $NT \times |I_k|$, with $|I_k|$ much smaller than N^2.

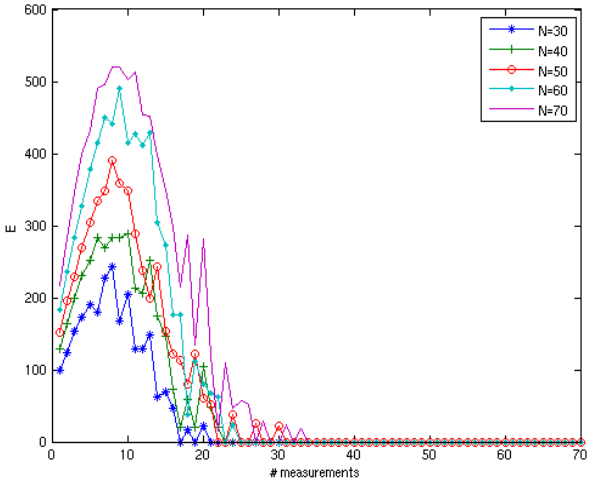


Figure 14: Number of wrong edges E for gene networks with $N = 30, 40, 50, 60, 70$ genes as a function of the number of measurements T.

In Figure 14 the numbers of wrong edges for $N = 30, 40, 50, 60$ and 70 genes are plotted for the Algorithm 1. Here, a similar behaviour can be seen as in Figure 13. For a sufficient large number of measurements the network is exactly recovered.

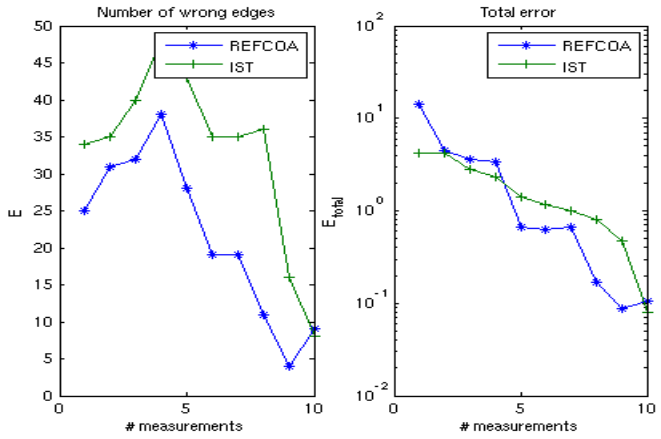


Figure 15: Comparison of Algorithm 1 with the Iterative Soft Thresholding algorithm for 1% Gaussian noise in the measurements. Number of wrong edges versus Number of measurements (left); Total error (right).

In our next test, we add a one per cent Gaussian noise to the data and compare the two algorithms for a network with 10 genes, see Figure 15. Both algorithms cannot recover the exact network. But the number of wrong edges is significantly lower for the refinement and coarsening strategy than for the iterative thresholding algorithm. The same holds true for the total error, which we measured in the Frobenius Norm

$$E_{total} = \|\tilde{W}_{ex} - \tilde{W}_{REFCOA/IST}\|_{Fro}$$

(see right side of Figure 15), so that we get the difference between the exact network entries and the computed ones. It can be seen that the error tends to almost zero for a sufficiently large number of measurements, similarly to the exact case.

For larger noise level and larger number of genes, the condition of the problem gets even worse. We applied both algorithms to a test network of $N = 40$ genes and added $\delta_{l^\infty} = 5\%$ gaussian noise to every measurement. As in the case of the smaller noise level the exact network structure is not reconstructed exactly (see Figure 16). Once again the number of wrong edges and the computational time for the Iterative Soft Thresholding is significantly higher than of Algorithm 1. IST produces more than 100 wrong edges out of 1600 possible connections, whereas the number for Algorithm 1 is less than 50, see Figure 17. The threshold $\tilde{\epsilon}$ for counting the wrong edges is set to $\tilde{\epsilon} = 0.02$ here. For both algorithms the total error, which is plotted in a logarithmic scale on right side of Figure 16, gets smaller for larger number of measurements.

To obtain these results, we have to regularize according to the noise level. We achieve the required regularization by early stopping the refinement and coarsening procedure. For this purpose,

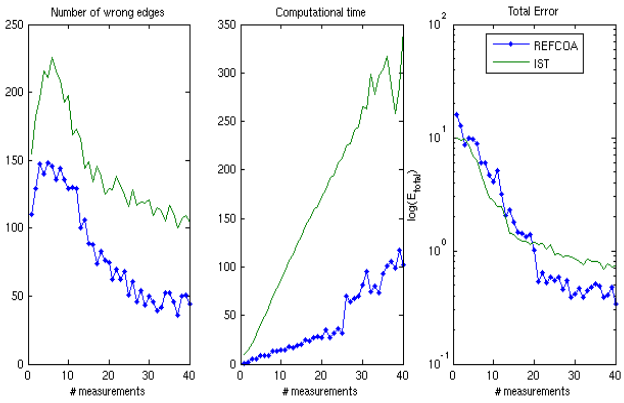


Figure 16: Comparison of Algorithm 1 to the Iterative Soft Thresholding for a network with 40 Genes and 5% gaussian noise on the data.

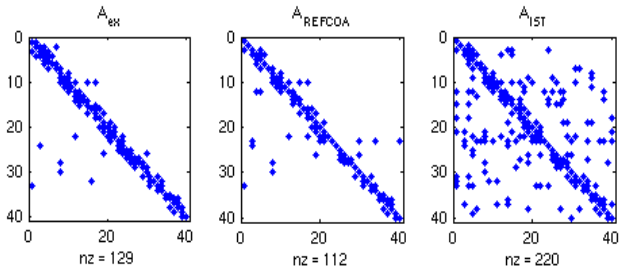


Figure 17: Nonzero entries of the connectivity matrix for the network of Figure 16 taking into account 40 measurements. (Generated with the Matlab function spy())

we choose the tolerance ε in line 6 of Algorithm 1 proportional to the noise level, realizing a generalized discrepancy principle. Note that as opposed to the classical discrepancy principle, where the stopping rule is $\left\|F(x) - y^\delta\right\|_{L^2} \leqslant \tau\delta_{L^2}$, we here have $\left\|(\langle F'(x)^* F(x) - y^\delta, \phi_i\rangle)_{i\in\{1,\dots,N\}}\right\|_{l^\infty} \leqslant \varepsilon$ with ε proportional to an appropriate power of δ_{L^2} or δ_{l^∞}, which is yet to be investigated theoretically. In our computations for Figure 16 and Figure 17, the threshold is set to $\varepsilon = 0.38 * \delta_{l^2}$. The same has to be done for the IST, where we have to choose the threshold value according to the noise level (for Figure 16 and Figure 17, the threshold is set to $\omega = 0.49 * \delta_{l^2}$).

4.3 BACK TO COMPRESSED SENSING

As we have seen the algorithm is very efficient for sparse linear inverse problems, where really few non-zero entries of $x^\dagger$ exist. So it might be a good choice for the above introduced example of compressed sensing (Example 3). In Figure 18 a comparison of the Tikhonov solution with ℓ_2 regularization term against the minimizer of the Refinement and Coarsening Algorithm in case of exact data is shown. There is almost no difference between the exact solution and the achieved minimizer of the Refinement and Coarsening Algorithm.

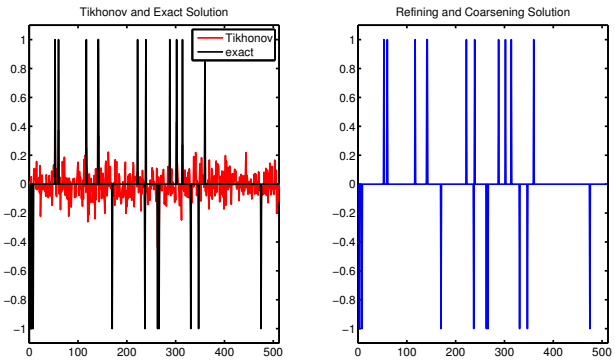


Figure 18: Comparison of Tikhonov solution with ℓ_2 squared error norm against the solution of the Refinement and Coarsening algorithm in case of noise free data. The error of the RefCoa solution is $3.3 \cdot 10^{-14}$.

In case of noisy data the Refinement and Coarsening Algorithm still provides a good approximation of the exact solution, see Figure 19. However there are some wrong spikes, with very small value outside the support of $x^\dagger$.

In Chapter 7 we will again consider the compressed sensing example. There we will reconstruct the exact signal additionally with some other wellknown algorithms for sparse inverse problems and provide an error analysis of the different minimizers.

But we want to mention already here, that with none of the other methods we used, we got as good results. Hence the Refinement and Coarsening algorithm seems to be almost perfect for the problem of compressed sensing.

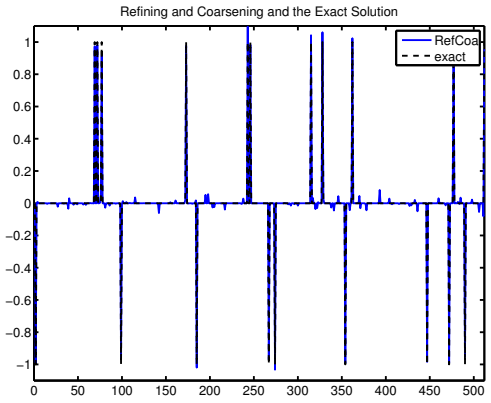


Figure 19: Solution of the Refinement and Coarsening Algorithm against the exact solution, where 5% Gaussian noise is added to the data. The exact solution (dashed line) is hardly distinguishable.

4.4 SUMMARY

In this chapter, we used ideas from [10], [9], [17] to generate an adaptive sparsity enhancing algorithm for general inverse problems. In our numerical tests we saw, firstly that the resulting method is able to determine the exact solution of a linear sparse inverse problems (or in words of the Systems Biology application, can determine the exact network structure given a sufficient large number of measurements). And secondly that the proposed method is very well suited for problems with a very small number of non-zero coefficients.

Therewith, we have introduced our first approach to solve sparse inverse problems. A deterministic projection algorithm. With this algorithm it is possible to identify gene networks, without supplementary procedures, like experimental design (cf. [84]) or robust regression (cf. [92]). Additionally we applied the algorithm to the test problem of compressed sensing, for which the algorithm yields very good reconstruction in the noise free, as well as in the noisy case.

In the next chapter we will have a closer look at regularization in preimage space, already considered in Remark 6 of this chapter as well as introduced in Chapter 1.

5

CONVERGENCE ANALYSIS OF REGULARIZATION IN PREIMAGE SPACE

In the last chapter we came across the approach of regularization by discretization in preimage space. In this chapter we will provide a convergence analysis for regularization in preimage space for the linear case, and later on for the general nonlinear situation. For the latter generalization we follow two approaches, namely a variational formulation and an operator equation formulation. This analysis is motivated by the fact that on one hand in practice very often inverse problems are just discretized and then solved, on the other hand this discretization itself can be expected to have a regularizing effect. Again we present some numerical results at the end of this chapter to illustrate the results of our convergence analysis. This time we will come back to Example 2, the parameter identification in a PDE setting.

5.1 INTRODUCTION

Regularization by discetization in *preimage space* (called least squares method in [69]) is often used in practice due to its easy implementation. Additionally it is of particular interest due to the possibility of using problem adapted ansatz functions for the inverse problem solution. As already mentioned in the introduction, previous results in the literature on this approach for linear ill-posed problems [38, 40, 64], and especially a counterexample by Seidman [81] indicate, that convergence can not occur for general $x^\dagger \in \mathcal{X}$ but only under special assumptions on the solution.

Here we formulate a convergence condition that assumes sufficiently fast convergence of the approximation error in image space to compensate for the growth of the norm of the inverse of the projected forward operator. This sufficiently fast convergence of the approximation error in its turn will be implied by a sufficiently strong source condition and hence is satisfied for sufficiently regular solutions, along with appropriate approximation properties of the ansatz spaces $\mathcal{X}_n$.

We wish to mention that the regularizing effect of discretization has been studied in a very general setting with possible discretization of both preimage and image space, e. g., in Chapter 3 of [60] and in [66] for the linear case. The latter also provides results on convergence and convergence rates for noisy data with

an a priori and a Lepskii type a posteriori discretization level choice.

The results of this chapter especially pertain to taking into account noisy data by means of a discrepancy principle type a posteriori discretization level choice, as well as to extension to the nonlinear situation.

Before we go on, we recall some of the basics already introduced in Chapter 2. We consider a sequence $(\mathcal{X}_n)_{n\in\mathbb{N}}$ of finite dimensional subspaces of the preimage space $\mathcal{X}_n \subseteq \mathcal{X}$, along with the corresponding orthogonal projections $P_n := \mathrm{Proj}_{\mathcal{X}_n}$, and assume that $\|(I - P_n)x^\dagger\| \to 0$ as $n \to \infty$. (The latter is the case e. g., if $\mathcal{X}_{n+1} \subseteq \mathcal{X}_n$ and $\overline{\bigcup_{n\in\mathbb{N}} \mathcal{X}_n} = \mathcal{X}$.) Therewith we define a regularized approximation as a solution of the finite dimensional minimization problem

$$x_n^\delta \in \operatorname{argmin}\{\|F(z_n) - y^\delta\|^2 : z_n \in \mathcal{D} \cap \mathcal{X}_n\}, \tag{5.1}$$

an approach which is also often called least squares projection.

The discretization level n has to be chosen appropriately in order to balance between the approximation error, that decays as $n \to \infty$ and the noise propagation, that becomes stronger for larger n. We here determine n_* by the discrepancy principle

$$n_* = \min\{n \in \mathbb{N} : \|F(x_n^\delta) - y^\delta\| \leqslant \tau\delta\} \tag{5.2}$$

(with $\tau > 1$ in the linear case, and $\tau > \frac{1+\eta}{1-\eta}$, with $\eta \in (0, 1)$ as in (5.3) below in the nonlinear case). This corresponds to the discrepancy principle for choosing the regularization parameter α, as stated in the introduction see (2.9).

A very efficient approach especially in the context of nonlinear problems is to apply a multilevel strategy for the computation of the regularized solution, i. e., to start on the coarsest level $n = 1$ of projection and successively compute the solution on level n making use of information on the solution on level $n - 1$, see, e. g., [53, 59, 54, 55, 56]. When doing so, the discrepancy principle just acts as a stopping rule for this multilevel iteration.

It is well-known that a convergence analysis for nonlinear ill-posed problems requires some assumption on the structure of the nonlinearity. Here we will assume that the often used Scherzer condition (also called tangential cone condition, cf. [79]) holds:

$$\left\|F(x) - F(\bar{x}) - F'(x)(x - \bar{x})\right\| \leqslant \eta \left\|F(x) - F(\bar{x})\right\| \quad \forall x, \bar{x} \in \mathcal{D} \tag{5.3}$$

with $\eta \in (0, 1)$, which implies

$$\|F(x) - F(\bar{x})\| \leqslant \frac{1}{1-\eta} \left\|F'(x)(x - \bar{x})\right\|$$

and,

$$\left\|F'(x)(x - \bar{x})\right\| \leqslant (1 + \eta) \left\|F(x) - F(\bar{x})\right\|.$$

Note that for this purpose $F'(x)$ needs not necessarily be a Fréchet derivative. It suffices to have a bounded linear operator called $F'(x)$ mapping between $\mathcal{X}$ and $\mathcal{Y}$ and such that (5.3) holds.

In what follows we will assume that an exact solution $x^\dagger$ of (1.1) exists, i. e., $F(x^\dagger) = y$. Moreover, to simplify the exposition we assume that

$$\forall x \in \mathcal{D} : \quad \mathcal{N}(F'(x)) = \{0\}$$

which implies uniqueness of the solution to $F(x) = y$ in the linear case and by (5.3) also in the nonlinear case: From

$$0 = \|F(x) - F(\tilde{x})\| \geqslant \frac{1}{1+\eta}\|F'(x)(x-\tilde{x})\|$$

it follows that $x - \tilde{x} \in \mathcal{N}(F'(x)) = \{0\}$. Note however, that most of our results can be generalized to operators with nontrivial nullspaces.

5.2 THE LINEAR CASE

In this section we first of all consider the linear case $A \in L(\mathcal{X}, \mathcal{Y})$, so that (1.1) becomes

$$Ax = y\,. \tag{5.4}$$

The first order necessary condition for a minimizer of (5.1), i. e., of

$$\min_{z_n \in \mathcal{X}_n} \|Az_n - y^\delta\|^2 = \min_{z_n \in \mathcal{X}_n} \|AP_n z_n - y^\delta\|^2 \tag{5.5}$$

implies

$$(AP_n)^*(AP_n z_n - y^\delta) = 0 \tag{5.6}$$

hence

$$AP_n z_n - y^\delta \in \mathcal{N}((AP_n)^*) = R(AP_n)^\perp = \mathcal{Y}_n^\perp\,.$$

i. e.,

$$Q_n(AP_n x_n^\delta - y^\delta) = 0 \tag{5.7}$$

with the orthogonal projections $Q_n := \mathrm{Proj}_{\mathcal{Y}_n}$, and the subspaces $\mathcal{Y}_n := A\mathcal{X}_n \subseteq \mathcal{Y}$, see also [60, 61, 66] for a convergence analysis of x_n^δ defined by (5.7) with general $\mathcal{Y}_n$, i. e., not necessarily $\mathcal{Y}_n = A\mathcal{X}_n$.

5.2.1.1 *Well-definedness*

Lemma 5.1. *The index n_* according to the discrepancy principle (5.2) with $\tau > 1$ is well-defined.*

Proof. The residual can be represented as follows:

$$\begin{aligned} Ax_n^\delta - y^\delta =& (I - Q_n)(Ax_n^\delta - y^\delta) = -(I - Q_n)y^\delta \\ =& -(I - Q_n)Ax^\dagger + (I - Q_n)(y - y^\delta) \\ =& -(I - Q_n)A(I - P_n)x^\dagger + \\ & (I - Q_n)(y - y^\delta) \end{aligned} \tag{5.8}$$

where we have used $Ax_n^\delta \in \mathcal{Y}_n$ in the second and $AP_n x^\dagger \in \mathcal{Y}_n$ in the fourth equality. From (5.8), and $\|I - Q_n\| \leqslant 1$ we get

$$\|Ax_n^\delta - y^\delta\| \leqslant \|A\| \underbrace{\|(I - P_n)x^\dagger\|}_{\to 0 \text{ as } n\to\infty} + \delta$$

hence $\|Ax_n^\delta - y^\delta\| \leqslant \tau\delta$ for n sufficiently large. □

Denote

$$\alpha_n := \|(I - P_n)x^\dagger\|, \quad \beta_n := \|A(I - P_n)x^\dagger\| \tag{5.9}$$

$$\begin{aligned} \gamma_n :=& \inf_{\substack{z_n \in \mathcal{X}_n \\ z_n \neq 0}} \frac{\|Az_n\|}{\|z_n\|} = \inf_{\substack{z_n \in \mathcal{X}_n \\ z_n \neq 0}} \frac{\|Az_n\|}{\|(AP_n)^\dagger AP_n z_n\|} = \\ =& \inf_{\substack{w_n \in A\mathcal{X}_n \\ w_n \neq 0}} \frac{\|w_n\|}{\|(AP_n)^\dagger w_n\|} = \frac{1}{\|(AP_n)^\dagger\|}, \end{aligned} \tag{5.10}$$

where we have used the fact that by $\mathcal{N}(A) = \{0\}$ we have

$$\mathcal{N}(AP_n)^\perp = \overline{\mathcal{R}(P_n A^*)} = P_n \overline{\mathcal{R}(A^*)} = P_n \mathcal{N}(A)^\perp = P_n \mathcal{X} = \mathcal{X}_n. \tag{5.11}$$

In the following proofs we will use the error decomposition

$$\begin{aligned} x_n^\delta - x^\dagger =& x_n^\delta - P_n x^\dagger - (I - P_n)x^\dagger \\ =& (AP_n)^\dagger AP_n(x_n^\delta - P_n x^\dagger) - (I - P_n)x^\dagger \\ =& (AP_n)^\dagger(y^\delta - AP_n x^\dagger) - (I - P_n)x^\dagger \\ =& (AP_n)^\dagger(y^\delta - y) + \\ & (AP_n)^\dagger A(I - P_n)x^\dagger - (I - P_n)x^\dagger, \end{aligned} \tag{5.12}$$

where we have used (5.7) and (5.11).

5.2.1.2 *Stability for fixed discretization level*

Proposition 5.2. *Let n be fixed and let $(y^k)_{k\in\mathbb{N}}$ be a sequence converging to y^δ: $\|y^k - y^\delta\| \to 0$ as $k \to \infty$ and denote by x_n^k the corresponding regularized solutions according to (5.7) with y^δ replaced by y^k.*

Then

$$\|x_n^k - x_n^\delta\| \to 0 \ \text{ as } k \to \infty$$

Proof. From (5.12) we obtain

$$\|x_n^k - x_n^\delta\| = \|(AP_n)^\dagger(y^k - y^\delta)\| \leqslant \|(AP_n)^\dagger\|\|y^k - y^\delta\| \to 0$$

as $k \to \infty$. □

5.2.1.3 *Convergence*

Theorem 5.3. *If*

$$\frac{\beta_n}{\gamma_n} \to 0\,, \quad \frac{\beta_{n-1}}{\gamma_n} \to 0\,, \quad \alpha_n \to 0\,, \quad \text{as } n \to \infty \tag{5.13}$$

then for $n_* = n_*(\delta)$ according to (5.2)

$$\|x_{n_*(\delta)}^\delta - x^\dagger\| \to 0 \ \text{ as } \delta \to 0\,.$$

Proof. From (5.12) we get

$$\|x_n^\delta - x^\dagger\| \leqslant \frac{\delta}{\gamma_n} + \frac{\beta_n}{\gamma_n} + \alpha_n\,. \tag{5.14}$$

Let δ_m be an arbitrary sequence converging to zero and y_m a sequence of data satisfying $\|y_m - y\| \leqslant \delta_m$. We denote by n_m the corresponding index according to the discrepancy principle and by $x_{n_m}^{\delta_m}$ the approximation on level n_m according to (5.7) with y^δ replaced by y_m.

Consider first the case that $n_m \to \infty$ as $m \to \infty$. Then according to (5.8) we can estimate as follows:

$$\tau\delta_m \leqslant \|Ax_{n_m-1}^{\delta_m} - y_m\| \leqslant \|(I - Q_{n_m-1})A(I - P_{n_m-1})x^\dagger\| + \delta_m\,, \tag{5.15}$$

hence

$$\frac{\delta_m}{\gamma_{n_m}} \leqslant \frac{1}{\tau - 1}\frac{\beta_{n_m-1}}{\gamma_{n_m}}\,, \tag{5.16}$$

so by (5.14) we get

$$\|x_{n_m}^{\delta_m} - x^\dagger\| \leqslant \frac{1}{\tau - 1}\frac{\beta_{n_m-1}}{\gamma_{n_m}} + \frac{\beta_{n_m}}{\gamma_{n_m}} + \alpha_{n_m} \to 0$$

as $m \to \infty$.

In the alternative case that (n_m) has a finite accumulation point there exists an $N \in \mathbb{N}$ and a subsequence $(m_k)_{k\in\mathbb{N}}$ such that

$$\forall k \in \mathbb{N} : \quad n_{m_k} = N .$$

From (5.8) we get

$$\tau\delta_{m_k} \geqslant \|Ax^{\delta_{m_k}}_{n_{m_k}} - y_{m_k}\| \geqslant \|(I - Q_N)A(I - P_N)x^\dagger\| - \delta_{m_k}$$

hence

$$(\tau + 1)\delta_{m_k} \geqslant \|(I - Q_N)A(I - P_N)x^\dagger\| \tag{5.17}$$

which with $k \to \infty$ implies

$$(I - Q_N)Ax^\dagger = (I - Q_N)A(I - P_N)x^\dagger = 0$$

i. e., $Ax^\dagger \in A\mathcal{X}_N$, i. e., $\exists z_N \in \mathcal{X}_N : \ Ax^\dagger = Az_N$ i. e., $\exists z_N \in \mathcal{X}_N : x^\dagger - z_N \in \mathcal{N}(A) = \{0\}$, i. e., $x^\dagger \in \mathcal{X}_N$, i. e., $(I - P_N)x^\dagger = 0$, hence by (5.14)

$$\|x^{\delta_{m_k}}_{n_{m_k}} - x^\dagger\| \leqslant \frac{\delta_{m_k}}{\gamma_N} \to 0$$

as $k \to \infty$. □

Remark 7. Note that the convergence conditions of this theorem imply the convergence criterion $\limsup_{n\to\infty} \|x_n\| \leqslant \|x^\dagger\|$ for exact data from [38], since

$$\begin{aligned} \|(AP_n)^\dagger y\| &\leqslant \|(AP_n)^\dagger AP_n x^\dagger\| + \|(AP_n)^\dagger A(I - P_n)x^\dagger\| \\ &\leqslant \|Q_n x^\dagger\| + \|(AP_n)^\dagger\| \, \|A(I - P_n)x^\dagger\| . \end{aligned}$$

The difference to the convergence results from Chapter 3 in [60] is that we consider convergence not for all $x^\dagger \in \mathcal{X}$ but for a particular $x^\dagger$ that might have higher regularity than typical elements of $\mathcal{X}$ so that divergence to infinity of the factor $\frac{1}{\gamma_n}$ due to unboundedness of $A^\dagger$ might indeed be compensated by sufficiently fast convergence of the approximation error β_n.

Indeed, note that $\|A(I - P_n)x^\dagger\|$ is likely to go to zero at a faster rate than $\|(I - P_n)x^\dagger\|$, since A typically has a smoothing property. Under an additional source condition,

$$x^\dagger = (A^*A)^\nu w$$

for some index $\nu > 0$ and some $w \in \mathcal{X}$ the rate of convergence of β_n to zero will still be improved: If the smoothing property of A can be quantified via boundedness as a mapping in a scale of spaces:

$$\|(A^*A)^\nu\|_{\mathcal{X}\to\mathcal{X}^\nu} \leqslant C_\nu$$

and the approximation property of $\mathcal{X}_n$ scales according to

$$\|I - P_{n-1}\|_{\mathcal{X}^\nu\to\mathcal{X}} \leqslant f_\nu(n) , \qquad \|I - P_{n-1}\|_{\mathcal{X}^{1/2}\to\mathcal{X}} \leqslant f_{1/2}(n) ,$$

(one might think of $\mathcal{X}_n$ being finite element spaces with mesh sizes $h = \frac{1}{n}$ and of $\mathcal{X}^\nu$ being Sobolev spaces of order $s = s(\nu)$), then

$$\begin{aligned}\left\|A(I-P_{n-1})x^\dagger\right\| &= \left\|(A^*A)^{1/2}(I-P_{n-1})^2(A^*A)^\nu w\right\| \\ &\leqslant \left\|(I-P_{n-1})(A^*A)^{1/2}\right\| \cdot \\ &\qquad \|(I-P_{n-1})(A^*A)^\nu\|\,\|w\| \\ &\leqslant f_\nu(n) f_{1/2}(n)\,\|w\| \end{aligned} \tag{5.18}$$

Typically, for sufficiently good approximation spaces (in the context of finite elements these will have to be of sufficiently high order) the decay of $f_\nu(n)$ will be faster for larger ν. On the other hand, γ_n is independent of the solution, hence the product $f_\nu(n) f_{1/2}(n) \frac{1}{\gamma_n}$ will tend to zero for ν sufficiently large, i. e., for sufficiently smooth solutions.

Note the necessity of considering a particular solution instead of norm estimates in view of the fact that e. g., in case of nested spaces $\mathcal{X}_{n-1} \subseteq \mathcal{X}_n$ we have

$$\begin{aligned}\gamma_n &\leqslant \inf_{\substack{z_n \in \mathcal{X}_n \cap \mathcal{X}_{n-1}^\perp \\ z_n \neq 0}} \frac{\|Az_n\|}{\|z_n\|} = \inf_{\substack{z_n \in \mathcal{X}_n \cap \mathcal{X}_{n-1}^\perp \\ z_n \neq 0}} \frac{\|A(I-P_{n-1})z_n\|}{\|z_n\|} \\ &\leqslant \|A(I-P_{n-1})\|\,,\end{aligned}$$

such that estimating $\beta_{n-1} \leqslant \|A(I-P_{n-1})\| \left\|x^\dagger\right\|$ would not enable (5.13).

Expressing a possible rate of convergence in (5.13) in terms of $\tilde{\beta}_n := \|(I-Q_n)A(I-P_n)x^\dagger\|$, we can easily deduce a convergence rates result in terms of δ.

Corollary 5.4. *Let, with some functions $g_1, g_2, g_3 : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying*

$$g_i(0) = 0\,,\ g_i \text{ monotonically increasing} \tag{5.19}$$

and

$$\forall C > 1\ \exists \bar{C}(C) > 0\ \forall \lambda > 0 :\ g_i(C\lambda) \leqslant \bar{C}(C) g_i(\lambda) \tag{5.20}$$

for $i = 1, 2, 3$, the following rates hold in (5.13)

$$\begin{aligned}\frac{\beta_n}{\gamma_n} = O(g_1(\tilde{\beta}_n))\,, \quad \frac{\tilde{\beta}_{n-1}}{\gamma_n} = O(g_2(\tilde{\beta}_n))\,, \\ \alpha_n = O(g_3(\tilde{\beta}_n))\,,\end{aligned} \tag{5.21}$$

as $n \to \infty$. Then

$$\|x^\delta_{n_*(\delta)} - x^\dagger\| = O(\max\{g_1(\delta), g_2(\delta), g_3(\delta)\}) \ \text{as}\ \delta \to 0$$

Proof. Analogously to (5.16), (5.17) we get

$$\frac{\delta}{\gamma_{n_*}} \leqslant \frac{1}{\tau-1}\frac{\tilde\beta_{n_*-1}}{\gamma_{n_*}}, \quad (\tau+1)\delta \geqslant \tilde\beta_{n_*}.$$

Inserting this and the rates assumptions (5.21) into (5.14), we directly get the assertion. □

Remark 8. Typical functions satisfying (5.19), (5.20) are $\sigma \mapsto \sigma^\kappa$, $\sigma \mapsto |\ln\sigma|^{-q}$ with $\kappa, q > 0$.

Consider the special case of A being compact and the subspaces $\mathcal{X}_n = \operatorname{span}\{u_1,\dots,u_n\}$, where $(\sigma_j, u_j, v_j)_{j\in\mathbb{N}}$ is a singular system for A (i. e., the method reduces to truncated singular value decomposition). Then a source condition of the form $x^\dagger = f((A^*A))w$, with an index function f, (i. e., $f(0) = 0$, f monotonically increasing) implies

$$\begin{aligned}
\tilde\beta_n &= \beta_n = \Big(\sum_{j=n+1}^{\infty} \sigma_j^2 f(\sigma_j^2)^2 \langle w, u_j\rangle^2\Big)^{1/2},\\
\gamma_n &\geqslant \sigma_n,\\
\alpha_n &= \Big(\sum_{j=n+1}^{\infty} f(\sigma_j^2)^2 \langle w, u_j\rangle^2\Big)^{1/2},
\end{aligned}$$

so provided the function $\Theta(\lambda) = \sqrt{\lambda} f(\lambda)$ is strictly monotonically increasing (e. g., if $f(\lambda) > 0$ for $\lambda > 0$), we obtain that (5.21) holds with $g_1 = g_2 = g_3 = f \circ \Theta^{-1}$, which yields the usual rates under such source conditions, see, e. g., Corollary 1 in [66].

5.3 THE NONLINEAR CASE

In the linear case we have seen that the essential quantities for showing convergence are α_n, β_n, γ_n defined as in (5.9), (5.10). To deal with the nonlinear situation we define $\alpha_n, \beta_n(x), \gamma_n(x)$ as in (5.9), (5.10) with A replaced by $F'(x)$.

In view of the two formulations (5.5), (5.6) we consider two possible generalizations of the linear case, namely a variational one and one based on a projected operator equation.

5.3.1 *Nonlinear case via global minimizer*

5.3.1.1 *Well-definedness*

First of all we show well-definedness of x_n^δ according to (5.1).

Proposition 5.5. *Assume that $\mathcal{D} \cap \mathcal{X}_n \neq \emptyset$, and that F is weakly sequentially closed, i. e., for all sequences $(x_k)_{k\in\mathbb{N}} \subseteq \mathcal{D} \cap \mathcal{X}_n$*

$$\Big(x_k \rightharpoonup x \wedge F(x_k) \rightharpoonup f\Big) \quad \Rightarrow \quad \Big(x \in \mathcal{D} \wedge F(x) = f\Big).$$

Then there exists a minimizer of the cost function $J_n : \mathcal{D} \cap \mathcal{X}_n \to \mathbb{R}$, $J_n(z_n) = \|F(z_n) - y^\delta\|$.

Note that by membership of x_k in the finite dimensional space $\mathcal{X}_n$, continuity of F and closedness of $\mathcal{D} \cap \mathcal{X}_n$ (see the assumptions of Proposition 5.7 and Theorem 5.8 below) are sufficient for the assumtions of Proposition 5.5.

Proof. J_n is coercive (i. e., boundedness of cost function values implies boundedness of arguments): By (5.3) we have for all $z_n \in \mathcal{D} \cap \mathcal{X}_n$ and some fixed $x_n \in \mathcal{D} \cap \mathcal{X}_n$:

$$\begin{aligned} \|F(z_n) - y^\delta\| &= \|F(z_n) - F(x_n) + F(x_n) - y^\delta\| \\ &\geqslant \frac{1}{1+\eta}\|F'(z_n)P_n(z_n - x_n)\| - \|F(x_n) - y^\delta\| \end{aligned}$$

hence by (5.11) (with A replaced by $F'(x_n)$) we get

$$\begin{aligned} \|z_n - x_n\| &= \|(F'(x_n)P_n)^\dagger F'(x_n)P_n(z_n - x_n)\| \\ &\leqslant \frac{1}{\gamma_n(x_n)}((1+\eta)J_n(z_n) + \|F(x_n) - y^\delta\|)\,. \end{aligned}$$

Moreover, J_n is weakly lower semicontinuous due to weak sequential closedness of F and weak lower semicontinuity of the norm.

By a standard argument we therefore get existence of a minimizer. □

In the rest of this subsection we assume that $x_n^\delta \in \mathcal{X}_n$ is defined as a (not necessarily unique) global minimizer of (5.1) according to Proposition 5.5 and that

$$P_n x^\dagger \in \mathcal{D}\,, \quad n \in \mathbb{N} \tag{5.22}$$

Lemma 5.6. *Let F be continuous, $P_n x^\dagger \to x^\dagger$ as $n \to \infty$, and (5.22) hold, and let for all n x_n^δ be defined by (5.1).*

Then the index n_* according to the discrepancy principle (5.2) with $\tau > 1$ is well-defined.

Proof. By minimality of x_n^δ and (5.22) we get the following estimate of the residual

$$\begin{aligned} \|F(x_n^\delta) - y^\delta\| &\leqslant \|F(P_n x^\dagger) - y^\delta\| \\ &\leqslant \|F(P_n x^\dagger) - F(x^\dagger)\| + \delta \end{aligned} \tag{5.23}$$

Hence, by continuity of F and $P_n x^\dagger \to x^\dagger$ as $n \to \infty$ we get $\|F(x_n^\delta) - y^\delta\| \leqslant \tau\delta$ for n sufficiently large. □

The following estimate of the error will be useful in our stability and convergence proofs: By (5.3) we have

$$\begin{aligned} &\|F'(x_n^\delta)P_n(x_n^\delta - x^\dagger)\| - \|F'(x_n^\delta)(I - P_n)x^\dagger\| \\ &\leqslant \|F'(x_n^\delta)(x_n^\delta - x^\dagger)\| \leqslant (1+\eta)(\|F(x_n^\delta) - y^\delta\| + \delta) \end{aligned} \tag{5.24}$$

hence by (5.10) with A replaced by $F'(x_n^\delta)$, (5.11), and (5.23)

$$\begin{aligned}\|x_n^\delta - x^\dagger\| \leqslant& \|x_n^\delta - P_n x^\dagger\| + \|(I-P_n)x^\dagger\| \\ \leqslant& \frac{1}{\gamma_n(x_n^\delta)}\Big(\|F'(x_n^\delta)(I-P_n)x^\dagger\| + \\ & (1+\eta)(\|F(x_n^\delta) - y^\delta\| + \delta)\Big) + \|(I-P_n)x^\dagger\|\end{aligned} \tag{5.25}$$

5.3.1.2 *Stability for fixed discretization level*

Proposition 5.7. *Let F be continuous and satisfy (5.3), fix n, and let $(y^k)_{k\in\mathbb{N}}$ be a sequence converging to y^δ: $\|y^k - y^\delta\| \to 0$ as $k \to \infty$, denote by x_n^k the corresponding regularized solutions according to (5.1) with y^δ replaced by y^k and assume that $\mathcal{D} \cap \mathcal{X}_n$ is closed and nonempty.*

Then $(x_n^k)_{k\in\mathbb{N}}$ has a convergent subsequence and the limit of any convergent subsequence of $(x_n^k)_{k\in\mathbb{N}}$ is a global minimizer of (5.1). If this global minimizer is unique, then

$$\|x_n^k - x_n^\delta\| \to 0 \ \text{ as } k \to \infty\,.$$

Proof. Analogously to (5.23) we get by minimality of x_n^k and x_n^δ on one hand

$$\|F(x_n^k) - y^k\| \leqslant \|F(x_n^\delta) - y^k\| \leqslant \|F(x_n^\delta) - y^\delta\| + \|y^k - y^\delta\|$$

$$\|F(x_n^\delta) - y^\delta\| \leqslant \|F(x_n^k) - y^\delta\| \leqslant \|F(x_n^k) - y^k\| + \|y^k - y^\delta\|$$

hence

$$\big|\|F(x_n^k) - y^k\| - \|F(x_n^\delta) - y^\delta\|\big| \leqslant \|y^k - y^\delta\| \to 0 \text{ as } k \to \infty\,. \tag{5.26}$$

On the other hand, for some fixed $x_n \in \mathcal{D} \cap \mathcal{X}_n$, we get by minimality,

$$\begin{aligned}\|F'(x_n^\delta)(x_n^k - x_n^\delta)\| \leqslant& (1+\eta)\|F(x_n^k) - F(x_n^\delta)\| \\ \leqslant& (1+\eta)\Big(\|F(x_n^k) - y^k\| + \\ & \|F(x_n^\delta) - y^\delta\| + \|y^k - y^\delta\|\Big) \\ \leqslant& (1+\eta)\Big(\|F(x_n) - y^k\| \\ & + \|F(x_n) - y^\delta\| + \|y^k - y^\delta\|\Big) \\ \leqslant& 2(1+\eta)\Big(\|F(x_n) - y^\delta\| + c\Big)\end{aligned}$$

where c is such that (by convergence) $\|y^k - y^\delta\| \leqslant c$, hence by (5.10) with A replaced by $F'(x_n^\delta)$ and (5.11),

$$\|x_n^k - x_n^\delta\| \leqslant \frac{1+\eta}{\gamma_n(x_n^\delta)}(2\|F(x_n) - y^\delta\| + 2c) =: r\,,$$

so the sequence x_n^k lies in the bounded closed and finite dimensional – hence compact – set $\overline{B_r(x_n^\delta)} \cap \mathcal{X}_n$ and therefore has a convergent subsequence $(x_n^{k_m})_{m\in\mathbb{N}}$. For any convergent subsequence $(x_n^{k_m})_{m\in\mathbb{N}}$ of $(x_n^k)_{k\in\mathbb{N}}$ with limit $\tilde{x}$, by (5.26) and a standard argument (see [25]) from $\|y^k - y^\delta\| \to 0$, continuity of F and closedness of $\mathcal{D}\cap\mathcal{X}_n$ it follows that $\tilde{x}$ lies in $\mathcal{D}\cap\mathcal{X}_n$ and is a global minimizer of $\|F(\cdot) - y^\delta\|$ over $\mathcal{D}\cap\mathcal{X}_n$. Namely, we have by minimality of x_n^k for all $\forall x_n \in \mathcal{D}\cap\mathcal{X}_n$:

$$\begin{aligned}\|F(x_n) - y^\delta\| &\geqslant \limsup_{k\to\infty}\Big(\|F(x_n) - y^k\| - \|y^k - y^\delta\|\Big)\\ &\geqslant \limsup_{k\to\infty}\Big(\|F(x_n^k) - y^k\| - \|y^k - y^\delta\|\Big)\\ &\geqslant \|F(\tilde{x}) - y^\delta\|\,.\end{aligned}$$

In case of uniqueness of this global minimizer, a subsequence subsequence argument yields convergence. □

5.3.1.3 *Convergence of global minimizers*

Theorem 5.8. *Let F be continuous and satisfy (5.3), assume that for all $n \in \mathbb{N}$, $\mathcal{D}\cap\mathcal{X}_n$ is closed and nonempty, that (5.22) holds and that x_n^δ is defined as a solution to (5.1).*

If

$$\sup_{x\in\mathcal{D}\cap\mathcal{X}_n}\frac{\beta_n(x)}{\gamma_n(x)} \to 0\,,\qquad \sup_{\tilde{x},x\in\mathcal{D}\cap\mathcal{X}_n}\frac{\beta_{n-1}(\tilde{x})}{\gamma_n(x)} \to 0\,,\quad \alpha_n \to 0 \tag{5.27}$$

as $n\to\infty$, then for $n_* = n_*(\delta)$ according to (5.2)

$$\|x^\delta_{n_*(\delta)} - x^\dagger\| \to 0 \ \text{ as } \delta \to 0$$

Remark 9. Instead of (5.27) we only need

$$\frac{\beta_n(x_n^\delta)}{\gamma_n(x_n^\delta)} \to 0\,,\qquad \frac{\beta_{n-1}(P_{n-1}x^\dagger)}{\gamma_n(x_n^\delta)} \to 0\,,\quad \alpha_n \to 0\,,\quad \text{as } n\to\infty$$

Proof. Analogously to the linear case we distinguish between two cases

If $n_m \to \infty$ as $m\to\infty$, then we use (5.23) to obtain

$$\begin{aligned}\tau\delta_m &\leqslant \|F(x^{\delta_m}_{n_m-1}) - y_m\|\\ &\leqslant \|F(P_{n_m-1}x^\dagger) - F(x^\dagger)\| + \delta_m\\ &\leqslant \frac{1}{1-\eta}\|F'(P_{n_m-1}x^\dagger)(P_{n_m-1}x^\dagger - x^\dagger)\| + \delta_m\end{aligned} \tag{5.28}$$

hence

$$\frac{\delta_m}{\gamma_{n_m}(x^{\delta_m}_{n_m})} \leqslant \frac{1}{\tau-1}\frac{1}{1-\eta}\frac{\beta_{n_m-1}(P_{n_m-1}x^\dagger)}{\gamma_{n_m}(x^{\delta_m}_{n_m})}\,,$$

so by (5.25) we get

$$\|x_{n_m}^{\delta_m} - x^\dagger\| \leqslant \underbrace{\frac{\beta_{n_m}(x_{n_m}^{\delta_m}))}{\gamma_{n_m}(x_{n_m}^{\delta_m})} + \frac{1+\eta}{1-\eta}\frac{\tau+1}{\tau-1}\frac{\beta_{n_m-1}(P_{n_m-1}x^\dagger)}{\gamma_{n_m}(x_{n_m}^{\delta_m})} + \alpha_{n_m}}_{\to 0}$$

as $m \to \infty$.

If there exists an $N \in \mathbb{N}$ and a subsequence $(m_k)_{k\in\mathbb{N}}$ such that $n_{m_k} = N$ for all $k \in \mathbb{N}$, we get

$$\tau\delta_{m_k} \geqslant \|F(x_{n_{m_k}}^{\delta_{m_k}}) - y_{m_k}\| = \|F(x_N^{\delta_{m_k}}) - y_{m_k}\|$$

hence taking the limit on both sides and using stability for fixed N (see Proposition 5.7) we get that $x_N^{\delta_{m_k}}$ has a convergent subsequence and the limit $\tilde{x}$ of any convergent subsequence of $x_N^{\delta_{m_k}}$ satisfies $F(\tilde{x}) = y$, so $\tilde{x}$ is a solution.

By uniqueness of this solution, a subsequence subsequence argument yields convergence. □

5.3.2 *Nonlinear case via Euler equation*

5.3.2.1 *Well-definedness*

If the global minimizer according to Proposition 5.5 lies in the interior of $\mathcal{D}$, it will also satisfy the Euler equation

$$(F'(x_n^\delta)P_n)^*(F(x_n^\delta) - y^\delta) = 0\,, \tag{5.29}$$

which implies

$$Q_n(x_n^\delta)(F(x_n^\delta) - y^\delta) = 0 \tag{5.30}$$

with $Q_n(x) = \mathrm{Proj}_{y_n(x)}$ and $y_n(x) := F'(x)\mathcal{X}_n$. Our following investigations on the formulation (5.29) are also motivated by the fact that a numerical optimization method will usually not yield a global minimizer but a stationary point.

In this subsection we assume that x_n^δ is defined (not necessarily uniquely) by (5.30), which is possible, e. g., under the assumptions of Proposition 5.5 provided the minimizer lies in the interior of $\mathcal{D}$, or alternatively by a fixed point argument analogously to Lemma 2 in [52].

Again we start with a result on the discrepancy principle.

Lemma 5.9. *Let (5.3) be satisfied, $F'(x)$ be uniformly bounded on $\mathcal{D}$, $P_n x^\dagger \to x^\dagger$ as $n \to \infty$, and let for all n x_n^δ be defined by (5.30).*

Then the index n_* according to the discrepancy principle (5.2) with $\tau > \frac{1+\eta}{1-\eta}$ is well-defined.

Proof.

$$\begin{aligned}F(x_n^\delta)-y^\delta =&(I-Q_n(x_n^\delta))(F(x_n^\delta)-y^\delta)\\ =&(I-Q_n(x_n^\delta))\cdot\Big(F'(x_n^\delta)(x_n-x^\dagger)+\\ &F(x_n^\delta)-F(x^\dagger)-F'(x_n^\delta)(x_n-x^\dagger)+y-y^\delta\Big)\\ =&(I-Q_n(x_n^\delta))\cdot\Big(F'(x_n^\delta)(P_nx^\dagger-x^\dagger)+\\ &F(x_n^\delta)-F(x^\dagger)-F'(x_n^\delta)(x_n-x^\dagger)+y-y^\delta\Big)\end{aligned} \tag{5.31}$$

since $(I-Q_n(x_n^\delta))F'(x_n^\delta)x_n^\delta = 0 = (I-Q_n(x_n^\delta))F'(x_n^\delta)P_nx^\dagger$, and hence, due to (5.3)

$$\begin{aligned}\|F(x_n^\delta)-y^\delta\| \leqslant &\frac{1}{1-\eta}\cdot\\ &\Big(\|(I-Q_n(x_n^\delta))(F'(x_n^\delta)(I-P_n)x^\dagger\|+(1+\eta)\delta\Big)\end{aligned} \tag{5.32}$$

and therewith $\|F(x_n^\delta)-y^\delta\| \leqslant \tau\delta$ for n sufficiently large. □

The error decomposition (5.12) in the nonlinear case becomes

$$\begin{aligned}x_n^\delta-x^\dagger = &(F'(x_n^\delta)P_n)^\dagger(y^\delta-y)+\\ &(F'(x_n^\delta)P_n)^\dagger F'(x_n^\delta)(I-P_n)x^\dagger-(I-P_n)x^\dagger+\\ &(F'(x_n^\delta)P_n)^\dagger(F(x^\dagger)-F(x_n^\delta)-F'(x_n^\delta)(x^\dagger-x_n^\delta)),\end{aligned} \tag{5.33}$$

where we have used the identity

$$(F'(x_n^\delta)P_n)^\dagger(y^\delta-F(x_n^\delta)) = (F'(x_n^\delta)P_n)^\dagger Q_n(x_n^\delta)(y^\delta-F(x_n^\delta)) = 0$$

and (5.11).

5.3.2.2 *Stability for fixed discretization level*

Proposition 5.10. *Let n be fixed and let $(y^k)_{k\in\mathbb{N}}$ be a sequence converging to y^δ: $\|y^k-y^\delta\|\to 0$ as $k\to\infty$ and denote by x_n^k the corresponding regularized solutions according to (5.30) with y^δ replaced by y^k. Additionally, assume that F, F' are continuous, satisfy (5.3), that $\mathcal{D}\cap\mathcal{X}_n$ is closed and that*

$$\exists\underline{\gamma}_n>0\ \forall x\in\mathcal{D}\cap\mathcal{X}_n\ :\ \gamma_n(x)\geqslant\underline{\gamma}_n>0 \tag{5.34}$$

as well as

$$\exists\overline{\beta}_n>0\ \forall x\in\mathcal{D}\cap\mathcal{X}_n\ :\ \beta_n(x)\leqslant\overline{\beta}_n \tag{5.35}$$

holds.

Then $(x_n^k)_{k\in\mathbb{N}}$ has a convergent subsequence and the limit of any convergent subsequence of $(x_n^k)_{k\in\mathbb{N}}$ solves (5.30). If this solution is unique, then

$$\|x_n^k-x_n^\delta\|\to 0 \ \text{ as } k\to\infty.$$

Proof. From (5.32), (5.33), which remain valid under the present assumptions, we obtain

$$\begin{aligned}\|x_n^k - P_n x^\dagger\| \leqslant & \frac{1}{\gamma_n(x_n^k)} \cdot \\ & \left(\|y^k - y\| + \beta_n(x_n^k) + \eta\|F(x_n^k) - F(x^\dagger)\|\right) \\ \leqslant & \frac{1}{\gamma_n(x_n^k)} \cdot \\ & \left(\delta_k + \beta_n(x_n^k) + \eta(\delta_k + \frac{1}{1-\eta}(\beta_n(x_n^k) + \right. \\ & \left. (1+\eta)\delta_k))\right)\end{aligned}$$

with $\delta_k := \|y^k - y^\delta\| + \delta$, which by (5.34), (5.35) implies boundedness of x_n^k. Hence by compactness of balls in the finite dimensional subspace $\mathcal{X}_n$ we get existence of a convergent subsequence of x_n^k. By continuity of F and F', as well as

$$(F'(x_n^k)P_n)^*(F(x_n^k) - y_n^k) = 0,$$

the limit $\tilde{x}$ of any convergent subsequence of x_n^k satisfies

$$(F'(\tilde{x})P_n)^*(F(\tilde{x}) - y^\delta) = 0.$$

□

5.3.2.3 *Convergence of solutions to Euler equations*

Theorem 5.11. *Assume that x_n^δ is defined as a solution to (5.30), that F, F' are continuous with $F'(x)$ uniformly bounded and satisfy (5.3), that $\mathcal{D} \cap \mathcal{X}_n$ is closed and that*

$$\sup_{x \in \mathcal{D} \cap \mathcal{X}_n} \frac{\beta_n(x)}{\gamma_n(x)} \to 0, \quad \sup_{\tilde{x}, x \in \mathcal{D} \cap \mathcal{X}_n} \frac{\beta_{n-1}(\tilde{x})}{\gamma_n(x)} \to 0, \quad \alpha_n \to 0 \tag{5.36}$$

as $n \to \infty$ holds.

Then for $n_* = n_*(\delta)$ according to (5.2)

$$\|x^\delta_{n_*(\delta)} - x^\dagger\| \to 0 \text{ as } \delta \to 0$$

Remark 10. More precisely we actually need

$$\frac{\beta_n(x_n^\delta)}{\gamma_n(x_n^\delta)} \to 0, \quad \frac{\beta_{n-1}(x_{n-1}^\delta)}{\gamma_n(x_n^\delta)} \to 0, \quad \alpha_n \to 0, \quad \text{as } n \to \infty$$

in place of (5.36).

Proof. With (5.33), by (5.3) we get

$$\begin{aligned}\|x_n^\delta - x^\dagger\| &\leqslant \frac{\delta}{\gamma_n(x_n^\delta)} + \frac{\beta_n(x_n^\delta)}{\gamma_n(x_n^\delta)} + \alpha_n + \frac{\eta\|F(x_n^\delta) - F(x^\dagger)\|}{\gamma_n(x_n^\delta)} \\ &\leqslant \frac{1}{1-\eta}\frac{\delta}{\gamma_n(x_n^\delta)} + \frac{1}{1-\eta}\frac{\beta_n(x_n^\delta)}{\gamma_n(x_n^\delta)} + \alpha_n \end{aligned} \tag{5.37}$$

where analogously to (5.31), (5.32) we have

$$\begin{aligned} F(x_n^\delta) - F(x^\dagger) =& (I - Q_n(x_n^\delta)) \cdot \\ & \Big(F'(x_n^\delta)(P_n x^\dagger - x^\dagger) + F(x_n^\delta) - F(x^\dagger) - \\ & F'(x_n^\delta)(x_n - x^\dagger) \Big) + Q_n(y^\delta - y) \end{aligned}$$

hence, using once more (5.3),

$$\|F(x_n^\delta) - F(x^\dagger)\| \leqslant \frac{1}{1-\eta} \Big(\|(I - Q_n(x_n^\delta))F'(x_n^\delta)(P_n x^\dagger - x^\dagger)\| + \delta \Big) .$$

The rest of the proof goes analogously to the linear case: In case of $n_m \to \infty$, in place of (5.15), we get by (5.32)

$$\begin{aligned} \tau\delta_m &\leqslant \|F(x_{n_m-1}^{\delta_m}) - y_m\| \\ &\leqslant \frac{1}{1-\eta} \Big(\|(F'(x_{n_m-1}^{\delta_m})(I - P_{n_m-1})x^\dagger\| + (1+\eta)\delta_m \Big), \end{aligned}$$

and therewith, instead of (5.16),

$$\frac{\delta_m}{\gamma_{n_m}(x_{n_m}^{\delta_m})} \leqslant \frac{1}{\tau - \frac{1+\eta}{1-\eta}} \frac{\beta_{n_m-1}(x_{n_m-1}^{\delta_m})}{\gamma_{n_m}(x_{n_m}^{\delta_m})} . \tag{5.38}$$

For the case with a finite accumulation point N of the discretization levels n_m, we estimate, using (5.31),

$$\begin{aligned} \tau\delta_{m_k} &\geqslant \|F(x_{n_{m_k}}^{\delta_{m_k}}) - y_{m_k}\| \\ &\geqslant \frac{1}{1+\eta} \|(I - Q_N(x_N^{\delta_{m_k}}))F'(x_N^{\delta_{m_k}})(I - P_N)x^\dagger\| \\ &\quad - (1+\eta)\delta_{m_k} \end{aligned}$$

$$(\tau + \eta + 1)(1+\eta)\delta_{m_k} \geqslant \|(I - Q_N(x_N^{\delta_{m_k}}))F'(x_N^{\delta_{m_k}})(I - P_N)x^\dagger\| .$$

Taking on both sides the limit along a subsequence k_l for which $x_N^{\delta_{m_{k_l}}}$ converges to some x_N^0 according to Proposition 5.10 and arguing similar to the stability proof, we get

$$0 = \|(I - Q_N(x_N^0))F'(x_N^0)(I - P_N)x^\dagger\| ,$$

which analogously to the linear case implies $(I - P_N)x^\dagger = 0$, hence convergence. □

Remark 11. By inspection of the proof we see that instead of solving (5.29) exactly, it suffices to find x_n^δ such that

$$\|(F'(x_n^\delta)P_n)^*(F(x_n^\delta) - y^\delta)\| \leqslant c\gamma_n(x_n^\delta)\delta , \tag{5.39}$$

with $0 < c < \tau(1-\eta) - (1+\eta)$. In our computations we use the heuristic tolerance

$$\|(F'(x_n^\delta)P_n)^*(F(x_n^\delta) - y^\delta)\| \leqslant c\delta^2 , \tag{5.40}$$

which can not be shown to yield convergence though, since in place of (5.38) we get an expression that is quadratic in $\frac{\delta}{\gamma_n(x_n^\delta)}$, namely

$$\frac{\delta_m}{\gamma_{n_m}(x_{n_m}^{\delta_m})} \leqslant \frac{1}{\tau - \frac{1+\eta}{1-\eta}} \frac{\beta_{n_m-1}(x_{n_m-1}^{\delta_m})}{\gamma_{n_m}(x_{n_m}^{\delta_m})} + c\left(\frac{\delta_m}{\gamma_{n_m}(x_{n_m}^{\delta_m})}\right)^2,$$

which does not exclude large values of $\frac{\delta}{\gamma_n(x_n^\delta)}$.

Condition (5.39) (or (5.40)) can be used as a stopping criterion in the iterative solution of the Euler equation.

5.4 NUMERICAL RESULTS

5.4.1 *Test Problem*

For illustrating the theoretical results from above, we consider the test problem of identifying the source term q in the PDE

$$\begin{aligned} -\Delta u + qu &= f && \text{in } \Omega \\ u &= 0 && \text{on } \partial\Omega \end{aligned} \tag{5.41}$$

on the unit square $\Omega = [0,1] \times [0,1] \subset \mathbb{R}^2$ from measurements u in the whole domain Ω. Please be aware, that in contrast to Example 2 of the introduction, defining F as the parameter-to somution map for the PDE (5.41) leads to a nonlinear inverse problem. Such that we want to solve the inverse problem,

$$F(q) = u, \tag{5.42}$$

where $q \in \mathcal{X} = L^2(\Omega)$, $u \in \mathcal{Y} = L^2(\Omega)$, and $F : \mathcal{X} \to \mathcal{Y}$, whose properties have been studied e. g., in [25], [42], where also the Scherzer condition (5.3) has been verified.

In order to achieve a nonempty interior of the domain, similarly to [42], we set

$$\begin{aligned} \mathcal{D} = \{ & q \in L^2(\Omega) \mid \|q - \hat{q}\|_{L^2(\Omega)} \leqslant \beta \\ & \text{for some } \hat{q} \in L^\infty(\Omega) \text{ with } \hat{q} \geqslant 0 \text{ a.e.}\}, \end{aligned} \tag{5.43}$$

where $\frac{1}{\beta}$ is larger than the norm of the continuous embedding $H_0^1(\Omega) \to L^4(\Omega)$.

As in Example 2 the range of the forward operator F is $H^2(\Omega) \cap H_0^1(\Omega)$ due to convexity of Ω (cf., e. g., [36]) and $\mathcal{Y}$ is the Hilbert space $L^2(\Omega)$, so this yields an ill-posedness of degree two.

We consider three test cases for the parameter q, which in the first two cases is defined as a Gaussian density function:

$$q^\dagger(x,y) = \frac{1}{2\pi\|\Sigma\|} e^{\frac{(x-\mu)^T \Sigma^{-1}(y-\mu)}{2}}$$

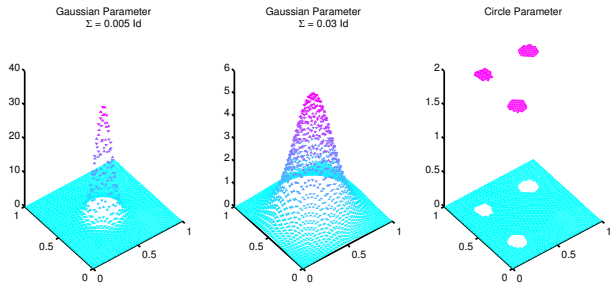


Figure 20: Exact parameters on fine grid.

with $\Sigma = 0.005\mathrm{Id}$, and $\Sigma = 0.03\mathrm{Id}$, respectively, where in both cases $\mu = (\frac{1}{2}, \frac{1}{2})$. The third test parameter is piecewise constant and supported on three non-intersecting circles

$$q^\dagger(x, y) = \begin{cases} 2 & \text{if } (x, y) \in K_1 \cup K_2 \cup K_3 \\ 0 & \text{else} \end{cases}$$

where K_1, K_2, K_3 are three circles with radius $r = 0.05$ and midpoints $M_1(\frac{1}{4}/\frac{1}{4})$, $M_2(\frac{1}{4}/\frac{3}{4})$ and $M_3(\frac{3}{4}/\frac{3}{4})$. See Figure 20 for a plot of the three test parameters on a very fine grid, with about 16.000 triangles. The corresponding solutions u of (5.41) to the different parameters are shown in fig. 21.

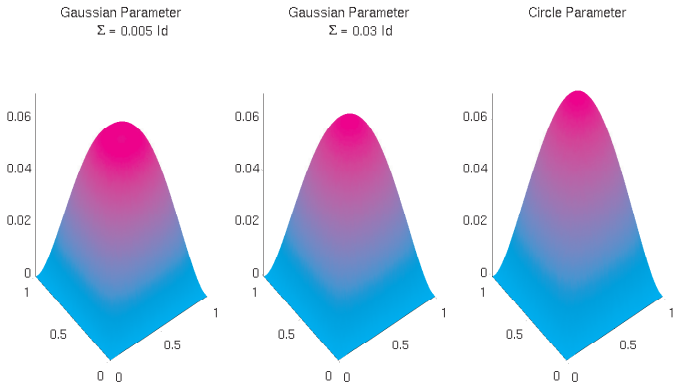


Figure 21: Exact solutions to the corresping source parameter on fine grid.

5.4.2 *Numerical Implementation*

For generating synthetic data and for solving the forward problem (5.41) we use the FEM as provided in a Matlab routine, that is capable to generate different meshes (p_n, t_n), consisting of the set of points p_n and the set of triangles t_n. Since the discretizations of the parameter q are defined on the triangles and those of the the PDE solutions on the set of points, the natural choice of $\mathcal{X}_n$ is the space of piecewise constant functions with possible jumps

over the triangle edges, whereas by choosing a very fine mesh for u we get an — up to a small approximation error — exact PDE solution. Considering L levels of discretization, labeling by 1 the coarsest and by L the finest one, we carry out the PDE solution for data generation on level L whereas all other PDE solutions are computed on level $L-1$ in order to avoid an inverse crime (see e. g., [51, 61]). For the projection spaces $\mathcal{X}_n$ the level n runs from 1 to $L-1$.

Hence the discretized forward operator consists of two components: interpolation from mesh (p_n, t_n) to mesh (p_{L-1}, t_{L-1}) and solving the underlying boundary value problem on the finest mesh.

In order to be able to transfer information on the parameter q over different grid levels in an exact manner, we generate nested grids, i. e., such that

$$p_{n-1} \subseteq p_n \text{ and } \forall \omega \in t_n \; \exists \tilde{\omega} \in t_{n-1} : \omega \subseteq \tilde{\omega}\,, \tag{5.44}$$

for $n = 2, \dots, L-1$, where $\omega, \tilde{\omega}$ denote triangles. Again, in order to avoid an inverse crime, the finest mesh for data generation is designed such that it is not part of the hierarchical mesh structure. For this purpose we initialize our computations by generating a hierarchical sequence of grids and one additional very fine grid for data generation, as well as by computing a PDE solution on the finest grid. The solution is then corrupted with random noise of a given percentage of the data norm as listed in the tables below. Then we solve (5.42) with q_n element of different subspaces $\mathcal{X}_n$ (given by respective meshes) as described in the previous section. Summarizing, we have implemented the following procedure:

1. Generate L finite element meshes $(p_1, t_1), \dots, (p_L, t_L)$ satisfying (5.44);
2. Solve (5.41) with given $q^\dagger$ on mesh L to get data u;
3. Interpolate u to the $L-1$-th mesh to avoid an inverse crime;
4. Add random noise to u according to the desired noise level δ in order to get u^δ;
5. Solve (5.1) with $\mathcal{X}_n$ defined by t_n, $n = 1, \dots, L-1$;

For validating the regularizing property of discretization in preimage space, we solve (5.42) in a least-squares sense, i. e., such that $\left\|F(q) - u^\delta\right\|^2$ is minimal, which is the method described and analyzed in the previous sections of this chapter. For the minimization of the least squares functional we used a Quasi-Newton algorithm with BFGS update, as implemented in the Matlab function *fminunc*, where we provided the gradient

$$(F'(x_n^\delta)P_n)^*(F(x_n^\delta) - y^\delta),$$

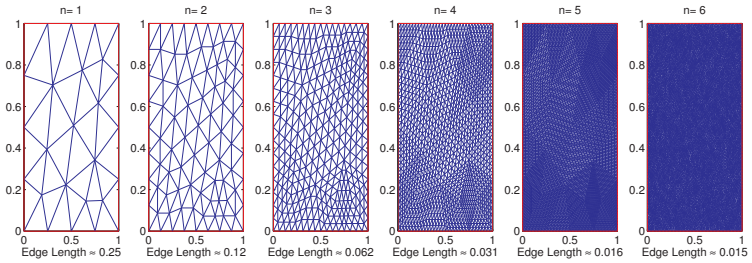


Figure 22: The $L = 6$ different meshes for the Gaussian parameter together with their corresponding maximum edge length.

which is computed using the adjoint technique. The algorithm stops if

$$\|\nabla J(x)\| = \|(F'(x_n^\delta)P_n)^*(F(x_n^\delta - y^\delta))\| \leqslant c\delta^2$$

according to equation (5.40) with $c = 8 \cdot 10^{-3}$.

5.4.3 *Results*

For each of the test examples, we generated a sequence of $L = 6$ meshes (see fig. 22) and interpolated the data to the fifth mesh, with $\approx$ 4000 triangles for the circle parameter and $\approx$ 9000 triangles in case of gaussian parameters. We corrupted the exact data with white Gaussian noise according to different noise levels $\delta = 0.5\%, 1\%, 2\%, 4\%, 8\%$. Then we solved the inverse problem on each subspace X_n, $n = 1, \ldots, L$ in the least squares sense (5.1). For comparison of the different solutions we computed the relative error:

$$e_r = \frac{\left\|q - q^\dagger\right\|}{\|q^\dagger\|}$$

with $q^\dagger$ given on (p_5, t_5). For computing this error, we interpolated q_k, $k = 1, .., 4$ to mesh 5, such that every small triangle from the fine mesh obtains the value from its surrounding triangle of the coarser mesh. This is possible because of the hierarchical structure of the mesh sequence.

In addition, for every solution we computed the corresponding ratio between discrepancy and noise level, to see if we can numerically validate the theoretically investigated discrepancy principle:

$$\tilde{\tau} = \frac{\left\|F(q) - u^\delta\right\|}{\delta}.$$

In Table 2 we list the relative error e_r and the associated ratio $\tilde{\tau}$ for the five different meshes and the different noise levels for the Gaussian parameter with the least squares solution. Additionally we mention the computation time t for the minimization process

Table 2: Gaussian Parameter with small Σ, i. e., $\Sigma = 5 \cdot 10^{-3}\mathrm{Id}$

Mesh		1	2	3	4	5
Edge Length		0.25	0.125	0.0625	0.03125	0.016
Noise Level						
8 %	e_r	**0.8570**	0.7935	0.9413	0.9999	0.9999
	$\tilde{\tau}$	1.0721	1.0742	1.8609	2.1985	2.1985
	t	1.3220	1.3203	1.2083	0.1461	0.1659
	#It	3	3	1	0	0
4 %	e_r	0.9533	**0.7809**	0.8143	0.8778	0.9999
	$\tilde{\tau}$	1.0783	1.1862	1.5464	2.4434	4.0388
	t	2.8897	1.4680	1.3593	1.4594	0.1679
	#It	15	4	2	1	0
2 %	e_r	0.9568	**0.5297**	0.7710	0.8014	0.9659
	$\tilde{\tau}$	1.2571	1.0665	1.5834	2.1592	6.9880
	t	3.0220	2.7704	1.6703	1.9957	2.5002
	#It	16	14	4	2	1
1 %	e_r	1.1083	0.5224	**0.4721**	0.7759	0.8174
	$\tilde{\tau}$	1.7815	1.2184	1.2824	2.8203	5.1237
	t	6.7834	3.0331	3.0807	2.4399	8.3380
	#It	45	16	13	3	2
0.5 %	e_r	1.2973	0.4612	0.4492	**0.4475**	0.7684
	$\tilde{\tau}$	3.0110	1.1056	1.6886	1.8145	5.2024
	t	11.4500	6.6803	3.4147	7.1335	18.6847
	#It	80	43	15	13	4

on the domain and the number of iterations the Quasi Newton solver carried out. There are three different features to mention.

Firstly, if we look at a fixed noise level, e. g., $\delta = 1\%$, we see the well-known error behavior of regularization methods for ill-posed problems (see [26]): For finer grids the error gets smaller but when the mesh size gets too small the error grows again. Thus there is an optimal discretization level, for which the error is minimal.

Secondly it can be seen, that we are able to validate the discrepancy principle also numerically. If we look at the smallest error (the bold numbers) we see that the corresponding ratio $\tilde{\tau} \approx 1$, which confirms our convergence analysis. Also for smaller noise level the best n determining the projection space $\mathcal{X}_n$ gets larger, e. g., for the gaussian parameter with $\Sigma = 0.03\mathrm{Id}$ and 0.5% noise, we achieved the smallest error on the finest mesh.

Additionally, we find that for $\delta \to 0$ the estimated parameter q tends towards the exact parameter $q^\dagger$, see the bold error numbers e. g., in Table 3

In case of the circle parameter, see Table 4, errors are higher than in the gaussian case, even in case of $\delta = 0.25\%$ noise, which is probably due to the non smoothness of the parameter.

5.5 SUMMARY

In this chapter we have seen some new results on conditional convergence and convergence rates for regularization by discretization in preimage space, including noisy data and nonlinear problems. We were motivated by the fact that in practice many ill-posed problems are regularized by discretization in preimage space and aware of the fact that this does not yield convergence in general.

In the next chapter we will have a closer look at a stochastic method to solve inverse problems, namely the Bayesian approach. We will see, how in this approach data and solution are modelled as random variables and how a point with highest probability is used as a solution. Additionally we will point out the close relation between stochastic modelling of inverse problems and deterministic Tikhonov regularization. This connection will lead us back to our main topic in Chapter 7, the solution of sparse inverse problems.

Table 3: Gaussian Parameter with larger Σ, i. e., $\Sigma = 3 \cdot 10^{-2}\mathrm{Id}$

Mesh		1	2	3	4	5
Edge Length		0.25	0.125	0.0625	0.03125	0.016
Noise Level						
8.000 %	e_r	0.3072	**0.2823**	0.9994	0.9994	0.9994
	$\tilde{\tau}$	1.0006	1.0241	1.5076	1.5076	1.5076
	t	1.3217	1.1925	0.1383	0.1454	0.1669
	#It	3	2	0	0	0
4.000 %	e_r	0.3012	**0.2119**	0.2497	0.9994	0.9994
	$\tilde{\tau}$	1.0050	1.0083	1.0797	2.4754	2.4754
	t	1.3407	1.3273	1.3917	0.1467	0.1655
	#It	3	3	2	0	0
2.000 %	e_r	0.3088	0.2159	**0.1896**	0.2176	0.9994
	$\tilde{\tau}$	1.0053	1.0073	1.0211	1.1550	4.6261
	t	1.4493	1.4552	1.5153	1.9793	0.1683
	#It	4	4	3	2	0
1.000 %	e_r	0.3017	0.2181	0.1900	**0.1848**	0.2607
	$\tilde{\tau}$	1.0022	1.0170	1.0195	1.0358	2.0177
	t	1.9879	1.4568	1.6819	2.4417	8.3386
	#It	8	4	4	3	2
0.500 %	e_r	0.2862	0.2034	0.1887	0.1838	**0.1816**
	$\tilde{\tau}$	1.0142	1.0259	1.1019	1.1643	1.3474
	t	4.6533	2.1385	1.6900	2.4584	13.4880
	#It	28	9	4	3	3

Table 4: Circle Parameter

Mesh		1	2	3	4	5
Edge Length		0.500	0.250	0.125	0.062	0.031
Noise Level						
8.000 %	e_r	1.0165	**0.7526**	0.9180	0.8370	0.9519
	$\tilde{\tau}$	1.0123	1.0081	1.0745	1.0292	1.0953
	t	0.4930	0.4989	0.5019	0.6032	0.8584
	#It	2	2	1	1	1
4.000 %	e_r	1.0182	**0.7734**	0.9251	0.8501	0.9561
	$\tilde{\tau}$	1.0287	1.0190	1.2275	1.0863	1.2891
	t	0.5498	0.4992	0.5027	0.6034	0.8595
	#It	3	2	1	1	1
2.000 %	e_r	1.3231	**0.6318**	0.7663	0.8462	0.9553
	$\tilde{\tau}$	1.0854	1.0312	1.1211	1.3619	1.9426
	t	1.0880	0.7182	0.5602	0.6040	0.8599
	#It	13	6	2	1	1
1.000 %	e_r	1.2857	**0.5813**	0.6211	0.7685	0.7739
	$\tilde{\tau}$	1.2855	1.0110	1.1025	1.3473	1.3788
	t	1.1296	1.0419	0.8014	0.7344	2.0487
	#It	14	12	6	2	2
0.500 %	e_r	1.2617	0.5591	**0.5157**	0.6126	0.7751
	$\tilde{\tau}$	1.9447	1.0184	1.0481	1.3632	2.1241
	t	1.2188	1.4464	1.1634	1.2503	2.0698
	#It	15	19	12	6	2
0.250 %	e_r	1.2706	0.5562	**0.4845**	0.5097	0.6230
	$\tilde{\tau}$	3.3979	1.0105	1.0822	1.1901	2.1400
	t	1.9215	2.9264	1.5109	1.9494	6.3133
	#It	28	46	18	12	6

Part III

SPARSITY THROUGH VARIATIONAL REGULARIZATION

6 EXCURSUS IN BAYESIAN INVERSION THEORY

In this chapter we build a bridge between convergence concepts used in statistical Bayesian learning approaches and the regularization theory of ill-posed inverse problems. For this purpose, we first review the relevant results about convergence in the field of statistical learning theory in Section 6.2. Relations between statistical and deterministic inversion theory in the special case of linear inverse problems are explained in Section 6.3. We then extract ideas from both theories that can be used towards a more general theory about convergence for structurally non-identifiable problems and postulate a general convergence conjecture for the posterior distribution in this setting. This is presented in Section 6.4. There we will also give a slight generalization of the above introduced existence proof of an $\mathcal{R}$-minimizing solution, see Lemma 2.4. A conclusion and an outlook on open issues is given in Section 6.5. But first it is demonstrated, why it is of general interest to carry over convergence results from deterministic regularization theory to the stochastic inversion approach for structurally non-identifiable problems.

6.1 STRUCTURALLY AND PRACTICALLY NON-IDENTIFIABLE PROBLEMS

When considering the inverse problem of fitting a model that is parametrized by a vector[1] x to experimental data y, this is usually formulated as an optimization problem. In this formulation an objective function $J(x, y)$, for example the sum of squared errors or the likelihood function, is optimized with respect to x. These standard approaches generally give satisfactory results if all parameters are identifiable, i. e., if the objective function has a unique global optimum. In case the data do not contain enough information to identify all parameters, optimization of those standard objective functions leads to poor results. This is the case if either only few data points are available compared to the number of parameters to be estimated, or if the kind of observations generally do not allow for a unique identification (see for example [75] and references therein). An example for the former is the identification of parameters for a dynamic model, e. g.,

1 In the Bayesian context the searched for quantity is mostly denoted by θ. To keep the here given exposition consistent with the previous chapters, we denote it with x instead. Readers more familiar with the Bayesian synopsis may exchange x with θ.

parameter estimation for an ODE from longitudinal data, where the state of the system is measured at discrete time points for different initial conditions. If the number of time points is small, this can lead to *practically non-identifiable* optimization problems, where the objective function usually has a unique optimum, but with infinite confidence intervals. Practical non-identifiability can be overcome by increasing the sample size.

On the contrary, if the kind of data generally does not allow for a unique identification of parameter values independently of the sample size, this results in *structural non-identifiability*. Structurally non-identifiable problems are strongly related to ill-posed inverse problems: The objective function does not have a unique optimum, but several parameter combinations have the same objective function value. Structural non-identifiability for parameter estimation of ODEs can be caused by non-observable (hidden) variables, if only steady state information is available. Or in case, only states can be measured up to a normalization constant, as it is often the case for biological systems ([91]). In many cases the set of optima defines a manifold in the parameter space. A simple example for a structurally non-identifiable problem is the following: We use mass action kinetics to construct an ODE model for a reversible chemical reaction $A \rightleftharpoons B$, where A and B are different molecular species.

$$\begin{aligned} \dot{a} &= -k_+ a + k_- b, \\ \dot{b} &= k_+ a - k_- b. \end{aligned} \tag{6.1}$$

Here, a and b denote concentrations of A and B, and k_+ and k_- are reaction rate constants. System (6.1) describes a closed system in which the total number of molecules is conserved, $a + b = N$. The system approaches its equilibrium state from all initial conditions. We assume that we want to estimate k_+ and k_-, i. e., $x = (k_+, k_-)$, and can observe the equilibrium concentrations $\bar{a}_i$ and $\bar{b}_i$ in $i = 1, \dots T$ different experimental settings, e. g., for different initial conditions and different numbers of molecules, N. Thus y is given by $y = \{(\bar{a}_i, \bar{b}_i)\}_{i=1,\dots,T}$. Setting the left hand sides in system (6.1) to zero, the ratio between $\bar{a}$ and $\bar{b}$ is given by the equilibrium constant K_{eq}:

$$K_{eq} = \frac{k_+}{k_-} = \frac{\bar{b}}{\bar{a}} \tag{6.2}$$

Taking the sum of squared errors as objective function, this leads to

$$J(x, y) = \sum_{i=1}^{T} \left(\frac{\bar{b}_i}{\bar{a}_i} - \frac{k_+}{k_-} \right)^2, \tag{6.3}$$

whose level sets are described by $k_+ = ck_-$. The individual values k_+ and k_- are structurally not identifiable. Hence the

optimization problem does not have a unique solution, but solutions lie on a manifold described by $k_+ = c^{opt} k_-$, where $c^{opt} = T^{-1} \sum_{i=1}^{T} \frac{\bar{b}_i}{\bar{a}_i}$ is the mean of all experimentally observed ratios, which minimizes $J(x, y)$.

If we now think back to the term of *well-posedness* introduced in Chapter 1, than we can conclude that structurally non-identifiable problems belong to the class of *ill-posed* inverse problems, as the solution might not be unique. We have seen in the introduction how regularization theory and the introduction of a generalized solution concept can overcome the problems of ill-posedness and how convergence results for the regularized solution can be established.

Inverse problems can also be considered in a statistical learning framework, where the negative log-likelihood function of a stochastic model is often taken as the standard objective function. Regularization of these problems can for example be achieved in a Bayesian context, where both the data y and the parameters x are interpreted as random variables, and the objective function is the posterior distribution $p(x|y)$, see section 6.2 below.

For structurally identifiable problems, in the limit of infinite sample sizes, the posterior distribution converges to a Gaussian distribution with covariance given by the inverse of the Fisher information times the sample size, which assures that the maximum a posteriori point estimate $\hat{x}^{MAP}$ converges in probability. Since the influence of the regularization term vanishes with increasing sample size, under some regularity conditions these results are independent of the choice of this term, and derived from frequency properties of stochastic processes (see for example Chapter 4 in [30] and references therein). This concept of convergence does not work for structurally non-identifiable likelihoods, i. e., ill-posed inverse problems. Here, the influence of the regularization term does not vanish even in the limit of infinitely large sample sizes, and convergence of the posterior requires also convergence of the regularization term additional to infinite sample sizes. Such a convergence concept was applied to show convergence of the point estimator of Tikhonov regularized linear and nonlinear inverse problems to the minimum least squares solution ([26, 67]). Furthermore, this result has been used by [44] to show convergence of the posterior distribution in a respective Bayesian setting with conjugate Gaussian distributions.

We will now have a closer look at the statistical Bayesian inversion theory, before we outline the convergence results for the posterior distribution and later on point out how to generalize this convergence concept for general distributions.

6.2 STOCHASTIC BACKGROUND

6.2.1 *Statistical Bayesian approaches and asymptotic theory*

Statistical approaches assume a stochastic modeling framework, in which the data $y = \{y_i\}$ is interpreted as set of random variates drawn from an underlying distribution. Given y, the probability $p_L(y|x)$ states how likely it is to see y under model parameters x. This probability is called *likelihood function*. A standard statistical inference approach is to maximize this function or, equivalently, minimize the negative logarithm, $l_y(x) = -\log p_L(y|x)$, with respect to the parameter x. In case of sufficiently large sample sizes this *maximum likelihood estimator* (Maximum Likelihood Estimate (MLE)) $\hat{x}^{ML}$ gives often good results. It is a consistent and unbiased estimator with nice convergence properties. The MLE describes, however, a pure data fit, and hence in case of sparse data suffers from the same problems as other approaches along this line: The optimization problem may become ill-posed, i. e., for example has no unique solution, or ill-conditioned, i. e., although a unique global optimum might exist, parameters are practically non-identifiable because the confidence intervals of this optimum are infinite. In these cases, solutions have to be stabilized by an appropriate restriction of the solution space. This can be achieved by using a statistical Bayesian approach, in which the data y and the parameters x are both treated as random variables with joint distribution

$$p(y, x) = p(y)p(x|y) = p(x)p(y|x). \tag{6.4}$$

Objective function is the *posterior distribution* $p(x|y)$, the distribution over x after having seen the data y,

$$p(x|y) = \frac{p_P(x)p_L(y|x)}{p(y)}, \tag{6.5}$$

which is proportional to the product of the likelihood function $p_L(y|x)$ and the *prior distribution* $p_P(x)$ that encodes our prior belief about the true parameter values. In a Bayesian learning framework $p_P(x)$ and $p_L(y|x)$ are usually given, and the *evidence* $p(y)$ is obtained by marginalizing over x, i. e.,

$$p(y) = \int_X p_P(x)p_L(y|x)dx. \tag{6.6}$$

Parameter estimation in the Bayesian framework translates to an investigation of the posterior distribution, also called *posterior inference*. For example, point estimates are the posterior mode, denoted Maximum A Posteriori Estimate (MAP)

$$\hat{x}^{MAP} = \arg\max_x p(x|y),$$

or the mean $E_{p(x|y)}(x)$. Furthermore, the posterior also contains information about the confidence of these point estimates. The posterior's variance or entropy can be used as appropriate summaries for this purpose. Local approximations also work with the Fisher information matrix.

The asymptotic theory makes statements about convergence of the posterior distribution in the limit of infinitely large sample sizes. Results hold in probability, meaning that we assume the data $y = (y^1, \ldots, y^n)$ to be generated by a stochastic process, i. e., the y^i are independently drawn from a true underlying sampling distribution $f(y^i)$, and convergence results hold for repeated sampling from this distribution and increasing the sample size n. A main result of this theory is that the posterior distribution approaches normality,

$$p(x|y) \xrightarrow{n\to\infty} N(\bar{x}, \Gamma_{\text{post}})$$

with mean $\bar{x}$ which minimizes the Kullback-Leibler distance between the true distribution $f(y) = \prod_{i=1}^{n} f(y^i)$ and the family $\mathcal{F}$ of model distributions $p_L(y|x)$,

$$\bar{x} = \operatorname{arg} \min_{x \text{ with } p_L(y|x) \in \mathcal{F}} \mathrm{KLD}(f(y) \parallel p_L(y|x)). \tag{6.7}$$

The posterior mode is a consistent estimator, meaning that its distribution converges to a point mass about $\bar{x}$ as $n \to \infty$. The covariance Γ_{post} equals the inverse of the product of sample size n and the Fisher information matrix, $\Gamma_{\text{post}} = (nI(x^{\text{post}}))^{-1}$. This can be seen by a Taylor series expansion of $\log p(x|y)$ up to second order about its mode.

The asymptotic theory holds if the prior influence vanishes in the limit of large sample sizes. This is only the case for structurally identifiable problems, where the kind of data generally allows to identify x uniquely, and if $\bar{x}$ is not excluded by the prior or is at the boundary of its support.

The latter problem can in practice easily be overcome by choosing prior distributions that assign positive probabilities even to x values that are a priori assumed to be not very plausible, and we do not further consider this problem here.

For identifiable problems the likelihood dominates the prior for large n, and convergence results are thus independent of the prior. This simplifies the whole analysis, since properties of the posterior can be inferred from those of the likelihood function regardless of the prior.

For non-identifiable problems the posterior mode is not always a consistent estimator (for practical examples see [75, 91]), and the normal approximation of the posterior can fail. Furthermore, concepts from the asymptotic theory cannot be directly applied, and a new convergence concept is needed that includes the effect

of the prior in the limit of large sample sizes. For this convergence concept we have to consider Tikhonov regularization in a Bayesian way.

6.3 BAYESIAN TIKHONOV REGULARIZATION

Tikhonov regularized linear inverse problems have been considered in a Bayesian framework, and convergence of the posterior distribution has been shown in [44]. In this Bayesian framework, the random variables X and Y are connected via the linear model

$$Y = AX + E \tag{6.8}$$

with E denoting a noise term. If we use a Gaussian distribution for the prior and assume E to be Gaussian distributed, the posterior distribution

$$p(x|y) \propto p_P(x)p_L(y|x)$$

can explicitly be computed, since Gaussian priors are the conjugate priors for Gaussian likelihoods (see e. g., the section about exponential families in Chapter 2 of [30]). With the underlying linear model we have $Y|X \sim N(y - Ax, \sigma^2 I)$, as for X fixed, Y is distributed as $E = AX - Y$.

Theorem 6.1 (Theorem 3.7 in [51]). *If $X : \Omega \to \mathbb{R}^n$ and $E : \Omega \to \mathbb{R}^m$ are mutually independent Gaussian random variables,*

$$X \sim N(x_0, \gamma^2 I) \quad E \sim N(0, \sigma^2 I).$$

And X and E are connected corresponding to (6.8), with $A \in \mathbb{R}^{m\times n}$ a known matrix. Then the posterior is Gaussian distributed,

$$p(x|y) = N(\bar{x}, \Gamma_{post})$$

with

$$\bar{x} = \left(A^T A + \frac{\sigma^2}{\gamma^2} I\right)^{-1} \cdot \left(A^T y + \frac{\sigma^2}{\gamma^2} x_0\right) \tag{6.9}$$

and

$$\Gamma_{post} = \sigma^2 \left(A^T A + \frac{\sigma^2}{\gamma^2}\right)^{-1}. \tag{6.10}$$

Proof. The proof can be carried out through tedious matrix multiplications or by the Schur identity, see Chapter 3 of [51] for details. □

Remark 12. If we now compare the deterministic Tikhonov regularized solution of (1.1), given in (2.8), with $\bar{x}$ that is given in (6.9), we see that they are equivalent. Except, the regularization parameter α now equals the quotient between noise and prior variance,$\alpha = \frac{\sigma^2}{\gamma^2}$.

Another equivalence occurs when comparing the maximization of the posterior to the minimization of the negative logarithm of the posterior. This can be seen most easily in the one dimensional case, then the posterior is given by

$$p(x|y) = \frac{1}{2\pi\sigma\gamma} e^{-\frac{(y-Ax)^2}{2\sigma^2} - \frac{x^2}{2\gamma^2}},$$

and hence taking the negative logarithm, ignoring the constant terms, leads to the minimization problem

$$\min J(x) = \min \left\{ (y - Ax)^2 + \frac{\sigma^2}{\gamma^2} x^2 \right\}.$$

Which is obviously just standard Tikhonov regularization with regularization parameter $\frac{\sigma^2}{\gamma^2}$. So taking the negative logarithm is a quite interesting feature, to link stochastic and deterministic terms. A feature which we will use in the next section, to show convergence of the posterior distribution to a point measure. Additionally we will use it in the next chapter to generate a new sparsity enforcing regularization term.

6.3.1 *Convergence results*

In this section, we give a brief summary of the obtained results in a series of works on the close connection between Tikhonov regularization and Bayesian learning. The main results were published in the article "Convergence rate for the Bayesian approach to linear inverse problems" by Hofinger and Pikkarainen [44], who consider convergence and convergence rates for Bayesian inference of linear inverse problems with independently and Gaussian distributed additive measurement noise. An extension to the multivariate case with arbitrary covariance matrix can be found in their follow-up paper ([45]) and additionally an extension to the infinite dimensional case is developed in [70]. Basic concept of these papers is to show convergence of the posterior distribution $p(x|y)$ to a point measure $\delta_{x^\dagger}$ in an appropriate metric and investigate convergence rates if the noise variance tends to zero. We will start by presenting the main convergence result and continue with the derivation of this concept.

6.3.1.1 *Main result*

Theorem 6.2 ([44], Theorem 13 and 15)**.** *Let* X *and* E *be two independently Gaussian distributed random variables,*

$$p_P(x) = N(x_0, \gamma^2 I) \text{ and } p_E(e) = N(0, \sigma^2 I),$$

which are connected via (6.8), and let x_0 *be an element in the complement of* $\mathcal{N}(A)$*,* $x_0 \in \mathcal{N}(A)^\perp$*. If* $\gamma(\sigma)$ *satisfies:*

$$\frac{\sigma}{\gamma(\sigma)} \to 0 \quad \text{and} \quad \gamma(\sigma)\sqrt{-\log(C(m)\sigma^{2\kappa(m)})} \to 0 \tag{6.11}$$

as $\sigma \to 0$*, where* $C(m)$ *and* $\kappa(m)$ *are given in Proposition 6.3 (see below), then*

$$\rho_P(p(x|y), \delta_{x^\dagger}) \to 0$$

as $\sigma \to 0$*.*

Here, ρ_P denotes the *Prokhorov metric* (see [43, 24]). Comparing this theorem with Lemma 2.5, $\lim_{\sigma\to 0} \frac{\sigma}{\gamma(\sigma)} \to 0$ corresponds to $\lim_{\delta\to 0} \alpha(\delta) = 0$, since $\alpha = \frac{\sigma^2}{\delta^2}$. This means that the influence of the regularization term has to vanish in the limit of infinitely small measurement noise. Furthermore, from the condition $\lim_{\sigma\to 0} \gamma(\sigma)\sqrt{-\log(C(m)\sigma^{2\kappa(m)})} \to 0$ we can conclude the following: First,

$$\lim_{\sigma\to 0} \gamma(\sigma) \to 0,$$

since

$$\lim_{\sigma\to 0} \sqrt{-\log(C(m)\sigma^{2\kappa(m)})} \to \infty,$$

which means that the prior distribution has to become more informative in the limit. Second, this has to happen with a rate that is fast enough, such that the product still goes to zero in the limit. A possible choice in two dimensions for $\gamma(\sigma)$ is e. g., $\gamma(\sigma) = \sqrt{\sigma}$, according to Theorem 7 in [45]. The proofs of the theorem strongly exploit that the posterior is explicitly known and that it is a Gaussian distribution. Thus a direct transfer to more general settings with other distributions seems to be very difficult, which is also confirmed in the conclusions of [44].

6.3.1.2 *Derivation of Theorem 6.2*

Here we give a very coarse excerpt of the derivation and focus on the main ingredients, for more details we refer to [44]. With equation (6.11), the required ratio between noise and regularization or prior influence is already given, but as we know from Chapter 1, the deterministic theory also requires a bound on the

noise, cf. (1.2). Since in the Bayesian setting the information on the noise is only given as a random distribution an explicit bound like in Lemma 2.5 is not possible. The required bound has to be translated into a proposition on the noise distribution. This is done in the following lemma, which is crucial to prove Theorem 6.2.

Proposition 6.3 ([44], Lemma 7)**.** *Let ξ be a random variable with values in $\mathbb{R}^m$. Assume that the distribution of ξ is $N(y_0, \Sigma)$. Let us define $\kappa(m) := \max\{1, m-2\}$ and $C(m)$ to be*

$$C(m) := \begin{cases} \frac{2\pi}{(m+1)^2} & \text{if } m \text{ is odd} \\ \frac{2^m}{m^2} & \text{if } m \text{ is even} \end{cases}$$

Then there exists a positive constant $\eta(m)$ such that

$$\rho_K(\xi, y_0) \leqslant \sqrt{-\|\Sigma\| \log(C(m)\|\Sigma\|^{\kappa(m)})} \tag{6.12}$$

for all $\|\Sigma\| < \eta(m)$.

The proof of this proposition is rather technical with several different cases. It uses the probabilistic definition of the Ky Fan metric ρ_K (cf. [24]) and the equivalence of the Ky Fan and the Prokhorov metric. Convergence of a random variable in the Ky Fan metric leads to convergence in probability, and is thus equivalent to convergence of the posterior in the asymptotic theory. In the proof one needs to evaluate the probability that a random variate z of a Gaussian distribution has a distance greater than a given value c from its mean m, which is given by an integral of the Gaussian distribution over the range $\| z - m \| > c$. Most of the proof is then concerned with finding an upper bound for this integral in dependence of the covariance matrix Σ for various cases. Through the proof of the upper bound the constants $\kappa(m)$ and $C(m)$ given in proposition (6.12) and used in Theorem (6.2) are established. It is important to notice, that they depend on the dimension of the problem, which is different from deterministic inverse problems (compare Lemma 2.5).

With this upper bound and the known bounds for the Tikhonov regularization, we can obtain the convergence of the posterior distribution to a point measure. This is done by dividing their distance in the Prokhorov metric into a stochastic term and a deterministic term:

$$\begin{aligned} \rho_P(p(x|y), \delta_{x^\dagger}) &\leqslant \rho_P(p(x|y), \delta_{\bar{x}}) + \rho_P(\delta_{\bar{x}}, \delta_{x^\dagger}) \\ &= \rho_P(p(x|y), \delta_{\bar{x}}) + \rho_K(\bar{x}, x^\dagger) \\ &= \rho_P(p(x|y), \delta_{\bar{x}}) + \|\bar{x} - x^\dagger\|, \end{aligned} \tag{6.13}$$

where $\bar{x}$ is the mean of the posterior distribution, which is equal to $\hat{x}^{MAP}$ in case of a Gaussian posterior. Here we have used the fact,

that the Prokhorov metric between delta distributions equals the Ky Fan distance between the center points of those distributions. Moreover the Ky Fan distance between deterministic elements (or rather almost surely constant random variables) equals their Banach space distance, such that we can split the equation in a stochastic distance between the posterior and a delta distribution centered at the posterior mean and a deterministic part, the distance between the MAP and the minimum norm least squares solution.

Now we can estimate the first term with the bound (6.12) and the identity of $\rho_P = \rho_K$ for delta measures. The second term is a deterministic term, which we can approximate by the deterministic results, since $\bar{x}$ is the mean of the posterior distribution, which is just the solution of the Tikhonov regularization as said above.

6.3.1.3 *Connections between large sample inference and convergence in the deterministic sense*

In the following we connect the two theories further by showin the equivalence between $n \to \infty$ in the asymptotic theory of large samples and $\sigma \to 0$ in the theory of ill-posed, linear, inverse problems, in case of Gaussian distributions. Assume samples $y = (y_1, \dots, y_n)$ to be drawn independently from a Gaussian distribution,

$$Y_i \sim N(f(x^*), \sigma^2 I), \tag{6.14}$$

with $y^i \in \mathbb{R}^m$, $x \in \mathbb{R}^k$ and $f : \mathbb{R}^k \to \mathbb{R}^m$ invertible. The likelihood is given by

$$p_L(y|x) = \prod_{i=1}^{n} p(y_i|f(x)). \tag{6.15}$$

According to the central limit theorem, the sample mean $\bar{y}$ can be seen as a random variable that approaches a Gaussian distribution for infinite sample sizes with mean $f(x^*)$ and variance σ^2/nI,

$$p(\bar{y}|x^*) = N(f(x^*), \frac{\sigma^2}{n} I) \tag{6.16}$$

$$= \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left(-\frac{1}{2}\left(\frac{\bar{y} - f(x^*)}{\frac{\sigma}{\sqrt{n}}}\right)^2\right). \tag{6.17}$$

This can also be regarded as distribution over $f(x^*)$,

$$p(f(x^*)|y) = N(\bar{y}, \frac{\sigma^2}{n} I).$$

The sampling distribution of the linear inverse problem (without any regularization) is given by $p_L(y|x) = N(Ax, \sigma^2 I)$. Regarding this as a distribution over x leads to

$$p(x|y) = N(A^{-1}y, \sigma^2 I),$$

in case A would be invertible. Comparing the variances of these two expressions, it is easy to see that decreasing σ by a factor k corresponds to increasing the sample size n by a factor k, and thus the limit $n \to \infty$ in the large sample theory corresponds to the limit $\sigma^2 \to 0$ in Tikhonov regularized inverse problems, i. e., better estimates can be obtained by measuring more accurately or by increasing the sample size.

6.3.1.4 *Numerical Illustration*

We use a simple linear problem to illustrate the above explained theoretical results, where we want to identify the parameter $x \in \mathbb{R}^2$ out of noisy measurements $y^\delta \in \mathbb{R}^2$. Let $A : \mathbb{R}^2 \to \mathbb{R}^2$ be an ill-conditioned matrix

$$A := \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

For simulating the Bayesian inference approach, we sampled a random vector ϵ out of the noise distribution $p_E(e) = N(0, \sigma^2 I)$ with $\sigma = k^{-2}$ and $k = 2, \ldots, 5$ and added it to the exact data y, which was created by $y = Ax^\dagger$ with $x^\dagger = (1, -1)^T$. The posterior distribution was calculated according to equations (6.9) and (6.10), with $x_0 = (0, 0) \in \mathcal{N}(A)^\perp$. We produced a contour plot of the likelihood and the prior distribution, together with 300 samples from the posterior distribution (blue dots) and marked the position of $\hat{x}^{MAP}$ (red cross). In Figure 23 (*left*) the regularization parameter $\alpha := \frac{\sigma}{\gamma(\sigma)}$ and additionally $\gamma\sqrt{-\log(\sigma)}$ tends to zero. With this configuration the constraints on σ and γ from Theorem 6.2 are fulfilled and the posterior distribution convergences apparently towards a point mass at the true solution $x^\dagger$. Additionally we have chosen a slightly different configuration on σ and γ (Figure 23 *right*), such that the regularization parameter α tends to zero, but the second condition, that $\gamma(\sigma)\sqrt{-\log(C(m)\sigma^{2\kappa(m)})} \to 0$, is not fulfilled. In this case the posterior mass does not converge to a point, and thus the posterior does not converge in probability. The mode is defined by the prior distribution, but the mass is distributed along the minimum least squares line even for very small measurement noise.

From the above given analysis we can see, that convergence results for structurally non-identifiable problems in the Bayesian learning theory are possible, if we take in account results from the deterministic theory on ill-posed inverse problems. The addition of a prior probability distribution generally leads to well-posed inverse problems, and the connection to Tikhonov regularization leads to convergence. To use this connection we had to translate the general requirements from the deterministic convergence results and restrict ourself to Gaussian distributions for prior and likelihood. In the next section we will extend this approach.

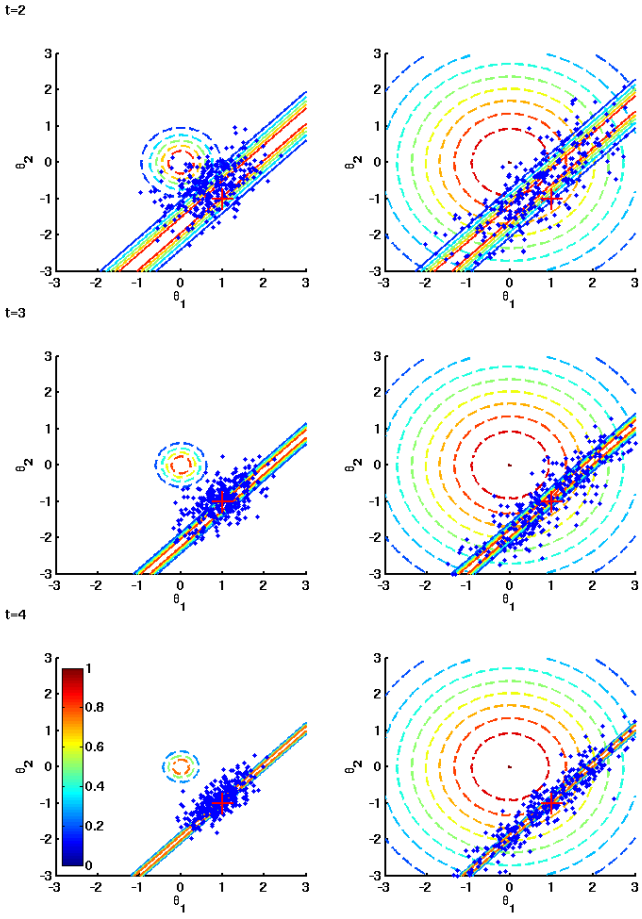


Figure 23: Contour plot of the likelihood and the prior distribution (dashed line) together with 300 samples from the posterior distribution (the blue dots) for the above explained numerical example. *Left:* $\sigma := \frac{1}{t^2}$ for $k = 2, .., 4$, $\gamma(\sigma) = \sqrt{\sigma}$ according to Theorem 7 in [45], and $x_0 := (0, 0)^T$. Clearly the posterior distribution mass gets closer and more centered at $x^\dagger = (1, -1)^T$ as $\alpha \to 0$ ($x^\dagger$ is marked with a red cross). *Right:* Same σ and x_0, but $\gamma = 2$ for all t. The posterior does not converge in probability.

6.4 TOWARDS CONVERGENCE

If we sum up the last two sections, we see that results of the asymptotic theory in the Bayesian learning framework only apply to structurally identifiable problems. In this case the prior influence vanishes in the large sample limit and we get asymptotic normality of the posterior. However, a different convergence concept is needed for structurally non-identifiable problems which includes the influence of the prior distribution in the limit $n \to \infty$. Results from the inverse problem theory provide the basis for such a new convergence concept, where mainly convergence of point estimates is considered so far. This has been extended to convergence of the whole posterior distribution in case of linear models with conjugate Gaussian distributions for prior and noise variables.

If we now want to generalize this convergence concept to other pairs of distributions, or state a generalized concept for structurally non identifiable problems, we can outline a scheme from the above given introduction. A possible result would state:

Conjecture 1. *We are given a data set $y = \{y^i\} \in Y$ that consists of $i = 1, \dots, n$ random variates drawn independently from an underlying distribution $f(y^i)$, a stochastic model class $p(y|x) = \prod_{i=1}^{n} p(y^i|x)$ with variance $V(data)$ that is parametrized by a vector x and which is assumed to describe the generation of the data set, and a prior distribution $p_P(x)$ on x whose variance $V(prior)$ can be chosen in dependence of the sample size n. Assume furthermore that*

1. *the posterior distribution $p(x|y)$ exhibits a unique maximum $\hat{x}^{MAP}$ for all possible data sets y,*
2. *the distance between a randomly drawn $Y \sim \prod_i^n p(y|x)$ and the data y^* that is most likely generated by sampling from the product $\prod_i^n p(y|x)$, i.e., $y^* = \arg\max_y \prod_{i=1}^{n} p(y|x)$, is bounded from above in the Ky Fan metric like in equation (6.12).*

We consider the large sample size limit $n \to \infty$.

If the variance of the prior distribution $V(prior)$ is appropriately chosen in dependence of the data set size such that

1. *it approaches zero for large sample sizes and*
2. *convergence to zero is in an appropriate sense slower than convergence of the variance $V(data)$ of the likelihood function,*

then

$$\rho_P(p(x|y), \delta_{\hat{x}^{MPMLE}}) \xrightarrow{n\to\infty} 0 \tag{6.18}$$

where $\delta_{\hat{x}^{MPMLE}}$ is a point measure concentrated at

$$\hat{x}^{MPMLE} \in \arg\max_x \{p(x) \mid x \in \arg\max_v p(y|v)\},$$

the Maximum Prior Maximum Likelihood Estimate (MPMLE).

The conjecture states convergence of the posterior distribution in probability or in an appropriate metric towards a point measure centered at a given point, if the noise tends to zero, or similarly, as the number of data points increase. Loosely this would mean, the mass of the posterior distribution should be concentrated at a given point, if we know the data better and better.

The parameter $\hat{x}^{\mathrm{MPMLE}}$ describes from the set of maximum likelihood estimators $\hat{x}^{\mathrm{ML}}$ the one which also maximizes the prior $p_P(x)$. The definition of $\hat{x}^{\mathrm{MPMLE}}$ is related to the definition of the $\mathcal{R}$-minimizing solution in the deterministic setting, cf. Definition (2.1), and leads to identifiability also in case of structurally non identifiable problems.

From the deterministic theory, see Chapter 2, we can explicitly state sufficient conditions for existence and uniqueness of the minimum norm least squares solution for a general regularization functional, as well as for the convergence of the point estimate to the minimum norm least squares solution, see Condition 1. This translates into conditions on the prior distribution and the posterior distribution. We will now state sufficient conditions under which an MLE and an MPMLE exist.

Proposition 6.4. *Assume the linear model (6.8) holds and* $y \in R(A)$. *If* $p_E(0) = \max\{p_E(x) | x \in \mathcal{X}\}$, *i.e., the noise distribution reaches its global maximum in zero, then there exists an MLE.*

Proof. As we assume the linear model (6.8) to hold, the likelihood $p_L(y|x)$ is is distributed like the noise $p_E(e)$ and evaluated at $e = Ax - y$.

Since $y \in R(A)$ there exists an $\tilde{x} \in \mathcal{X}$, such that $A\tilde{x} = y$ and hence

$$p_L(y|\tilde{x}) = p_E(A\tilde{x} - y) = p_E(0) \geqslant p_E(x), \quad \forall x \in \mathcal{X}.$$

Therewith $\tilde{x}$ is an MLE. □

Now we state the conditions under which an MPMLE exists.

Proposition 6.5. *Assume the linear model (6.8) holds. Let the likelihood function* $p_L(\cdot)$ *and the prior distribution* $p_P(\cdot)$ *in conjecture 1 be weakly upper semi continuous. Additionally let* $-\log(p_P(\cdot))$ *be coercive, for all* $x \in \mathcal{X}$, $\mathcal{X}$ *a reflexive Banach space. Assume there exists a MLE* $\hat{x}^{ML}$, *then there exists a MPMLE* $\hat{x}^{MPMLE}$.

Remark 13. A sufficient condition for the coercivity assumption in Proposition 6.5 is

$$p_P(x) \leqslant e^{-C\|x\|}$$

for all $x \in \mathcal{X}$ and a constant $C \geqslant 0$.

Proof. Instead of showing the existence of a Maximum Prior Maximum Likelihood Estimate, we will show the existence of a minimum element for $-\log(p_P(\cdot))$ in the set of minima of $l_y = -\log p_L(y|x)$. As this minimum element also maximizes the prior distribution and is in the set of maximum likelihood estimates.

Let $M_{MLE} := \{\tilde{x} : l_y(\tilde{x}) = \min\{l_y(x) : x \in \mathcal{X}\}\}$, the set of all minimum negative log-likelihood estimates. We first show the weak closedness of the set of MLEs.

Let $(x_k) \in M_{MLE}$, with $x_k \rightharpoonup x$. Now as $p_L(\cdot)$ is weakly upper semi continuous, $l_y(\cdot)$ is weakly lower semi continuous, because

$$\begin{aligned} p_L(x) &\geqslant \limsup p_L(x_k) \\ \log p_L(x) &\geqslant \limsup \log p_L(x_k) \\ -\log p_L(x) &\leqslant -\limsup \log p_L(x_k) \\ -\log p_L(x) &\leqslant \liminf(-\log p_L(x_k)) \end{aligned}$$

Therewith

$$l_y(x) \leqslant \liminf l_y(x_k) \leqslant l_y(\tilde{x}),$$

for all $\tilde{x} \in \mathcal{X}$. The last inequality follows, because the x_ks are minima of $l_y(\cdot)$. Hence $x \in M_{MLE}$, and the set of minima is weakly closed.

Set $c := \inf\{-\log(p_P(x)) : x \in M_{MLE}\}$. Since $M_{MLE} \neq \emptyset$, $p_P(x) \leqslant 1$ and $\sup p_P(x) > 0$ we have $c \in \mathbb{R}$. Now there is a sequence $(x_k) \in M_{MLE}$ with $-\log(p_P(x_k)) \to c$.

Hereby $-\log(p_P(x_k))$ is bounded. Hence by our coercivity assumption of $-\log(p_P(\cdot))$ also (x_k) is bounded and thus has a weak convergent subsequence $(x_{k_l}) \rightharpoonup \hat{x}$.

Again as $p_P(x)$ is weakly upper semi continuous, $-\log(p_P(\cdot))$ is weakly lower semi continuous. And hence,

$$-\log(p_P(\hat{x})) \leqslant \liminf -\log(p_P(x_{k_l})) = c.$$

Thus $-\log(p_P(\hat{x})) = c$ and as the set M_{MLE} is weakly closed, $\hat{x} \in M_{MLE}$. Now $\hat{x} = \hat{x}^{MPMLE}$ □

Remark 14. For the deterministic setting and nonlinear inverse problems Christiane Pöschl showed in Theorem 1.9 of [73] the existence of a $\mathcal{R}$-minimizing solution if a solution of the inverse problem exists, cf. Lemma 2.4 in Chapter 2. The concept of a $\mathcal{R}$-minimizing solution is equivalent to the prior maximizing solution introduced above, but for the above existence proof in the context of linear inverse problems, we only assumed that a maximum likelihood estimate exists.

In the next subsection we will use a simple numerical example to illustrate the above introduced conjecture.

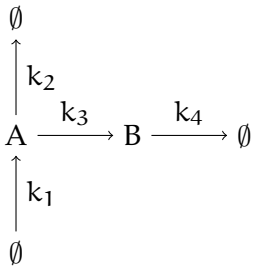


Figure 24: Chemical reaction system that is used to illustrate convergence of the posterior distribution for nonlinear problems.

6.4.1 *Numerical example*

We consider the chemical reaction system shown in Figure 24 and use again mass action kinetics to describe its dynamics with ODEs:

$$\dot{a} = k_1 - (k_2 + k_3)a \tag{6.19}$$

$$\dot{b} = k_3 a - k_4 b \tag{6.20}$$

This system approaches an equilibrium state, which is described by $\bar{a} = \frac{k_1}{k_2+k_3}$ and $\bar{b} = \frac{k_3 \bar{a}}{k_4}$. Assume that we know the value of the degradation rate constant k_4, we can measure $\bar{b}$, and we want to infer the equilibrium concentration $\bar{a}$ of species A and the conversion rate constant k_3, i. e., $x = (\bar{a}, k_3)$. The rate constants k_1 and k_2 are also unknown but not of our primary interest. The data is measured with an additive Gaussian measurement error with mean 0 and variance σ^2, and the measurement is repeated T times. The negative log likelihood function is in this case given by

$$l_y(x) \propto \sum_{i=1}^{T} \left(y_i - \frac{x_1 x_2}{k_4} \right)^2, \tag{6.21}$$

whose level sets are described by $x_1 = c/x_2$. Therefore the solution is not unique and the parameters are structurally not identifiable. Solutions lie on the manifold $x_1 = c^{\text{opt}}/x_2$ with $c^{\text{opt}} = k_4/T \sum_{i=1}^{T} y_i$.

Figure 25 shows the contour plots of a Gaussian prior $p_P(x) = N(0, \gamma^2 I)$ (*left*), the negative log likelihood $l_y(x_1, x_2)$ for $c^{\text{opt}} = 1$ (*center*) and the resulting posterior distributions (*right*) for different $\sigma = \frac{1}{n^2}$ with $n = 1, \ldots, 5$. As before, the variance γ of the prior distribution was chosen $\gamma(\sigma) = \sqrt{\sigma}$. The posterior mass cumulates at

$$\hat{x} = \arg\max_x \left(p_P(x) | x_1 x_2 = c^{\text{opt}} \right), \tag{6.22}$$

which is here given by $x = (1, 1)$ and which demonstrates convergence in probability.

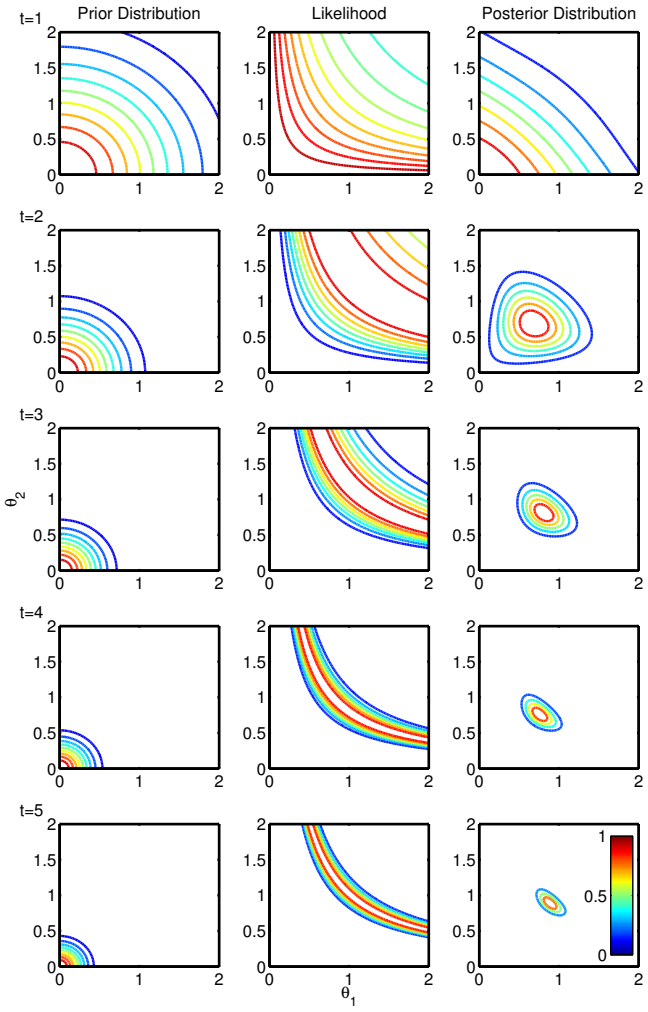


Figure 25: A non identifiable problem. Here the likelihood distribution (middle column) reaches its maximum on the manifold $x_2 = \frac{1}{x_1}$, such that a maximum likelihood estimator is not feasible. If we add a suitable prior(left column), then the posterior (right column) cumulates at $x^\dagger = (1,1)^T$. Again we set $\sigma := \frac{1}{t^2}$ for $t = 1,..,5$ and $\gamma(\sigma) = \sqrt{\sigma}$.

In comparison, Figure 26 (upper row) shows results if the variance of the prior is set to a fixed value $\gamma = 2$ and not adapted to σ. For large n the posterior is dominated by the likelihood term and the posterior mass does not cumulate about a specific value. The inverse problem is still ill-conditioned. Similarly, if γ decreases faster than σ, the posterior distribution becomes closer to the prior distribution regardless of the likelihood function, and the posterior converges to the maximum of the prior.

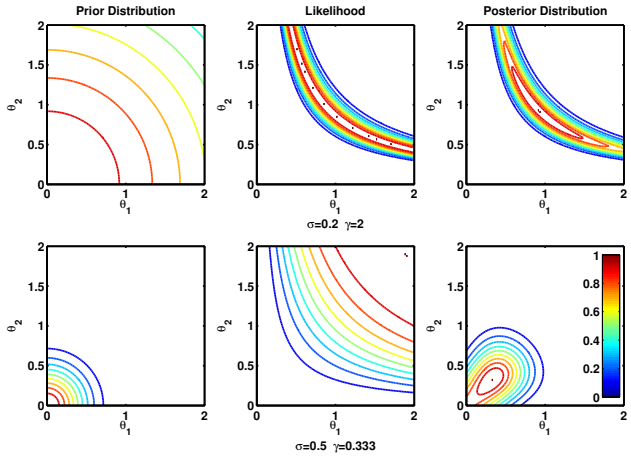


Figure 26: Here the relation between prior distribution and likelihood is not chosen correctly. In the upper row the influence of the prior is too weak, $\gamma = 2$ and $\sigma = 0.2$. Therefore the posterior resembles the likelihood. Whereas in the lower row, the influence of the prior distribution is too strong ($\gamma = \frac{1}{3}$ and $\sigma = 0.5$) and the posterior approaches the prior.

6.5 SUMMARY AND DISCUSSION

In this chapter we compared convergence concepts and results from the theory of ill-posed inverse problems and regularization theory in order to set up a similar convergence theory for the statistical inference of structurally non-identifiable problems, where the asymptotic theory fails. For these problems, the posterior distribution in a Bayesian setting depends on the prior distribution even in the large sample size limit. Moreover, for arbitrary but fixed prior distribution although the mode might converge, the posterior does usually not converge in probability, since the mass does not become more and more concentrated in a point. This has to be taken into account when setting up a framework for convergence of the posterior. Additional to the limit of large sample sizes, for convergence in probability the prior has to become more and more informative as well, which is already used in the deterministic theory of ill-posed inverse problems. While convergence of point estimates has already been studied for var-

ious objective functionals, convergence of the whole posterior distribution in a statistical Bayesian framework is only poorly investigated so far.

In the last section we presented ideas how to generalize the work of Hofinger et al. in this direction, who derived convergence results of the posterior distribution in case of linear problems and Gaussian prior and measurement noise. A generalization of the lines of proofs in their work did not seem to be possible for several reasons. Most importantly, it is exploited that the posterior is a multivariate Gaussian distribution and can explicitly be computed in this special setting. This allows to estimate an upper bound for the distance between a posterior random variable and its mean value, for which a straightforward generalization does not seem possible at the moment. Moreover, these theorems also use the fact, that only the mean of the posterior does depend on the data and is a random variable, while the covariance is deterministic. Thus, the posterior distribution itself can easily be treated as a random variable. As far as we know, this opportunity to regard the posterior as a random variable is only valid in case of Gaussian distributions and is perhaps not possible in case of other pairings of likelihood and prior.

In a non-conjugate framework, where the posterior is not given in explicit form, we need other techniques to show convergence. However, we presented ideas how the concept can be generalized, and illustrated these ideas with numerical examples. Additionally we provided a slight generalization of the existence results for maximum prior maximum likelihood estimates, or equivalently $\mathcal{R}$-minimizing solutions.

In the next chapter we will use the close connection between deterministic Tikhonov regularization and the stochastic Bayesian inversion, presented in Theorem 6.1. As we have seen, taking the negative logarithm of the posterior density function leads to an equivalent minimization problem as in the deterministic case. Such that, we could use results from deterministic theory to extend the stochastic theory. Now, we will go the other way around, using well-known sparsity enforcing prior distributions to generate a new regularization functional for deterministic Tikhonov regularization.

7 CAUCHY FUNCTIONAL

In the following we will use the close connection between stochastic and deterministic regularization theory to develop a new sparsity enforcing regularization functional, namely the Cauchy functional. It is based on the Cauchy probability density function, very slowly growing at zero and thus differentiable. We show that the generated Tikhonov functional with Cauchy regularization term is a proper regularization method in Section 7.2. Our proofs are based on the results on Tikhonov regularization in Banach spaces with general regularization term, which we introduced in Chapter 2. In Section 7.3 we show convergence rates for the new functional, but as the Cauchy term is not convex, we can not use the approaches we covered in Chapter 2. Instead we have to introduce a new notion of convexity, based on the work of Markus Grasmair [32]. Finally, we will come back to the examples of sparse inverse problems, which motivated our analysis in Chapter 3. We solve them with the proposed regularization approach and compare the numerical results to other sparsity promoting approaches.

7.1 INTRODUCTION

In the last chapter we have seen that if the likelihood and the prior distribution are Gaussian distributions, the negative logarithm of the posterior distribution directly leads to the Tikhonov functional with quadratic regularization term (Theorem 6.1). Starting from this close connection, we can search for sparsity promoting regularization functionals by considering sparsity enforcing prior distributions and we can investigate the resulting regularization functionals by taking the negative logarithm of the distribution.

One example is the betaprime distribution, given by

$$p(x) = p(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta))} x^{\alpha-1}(1+x)^{-\alpha-\beta}.$$

Setting $\alpha = 1$ and $\beta = 0$ this prior distribution results in the regularization term $\mathcal{R}(x) = \sum \log(1 + x_i)$, whose regularization properties have already been shown in Example 3.12 (A slowly growing function) of [12].

Another interesting prior distribution for sparse inverse problems is the Cauchy distribution. It is used in [51] to infer sparsity in a Bayesian setting. The Cauchy distribution is defined via:

$$p(x|\omega) = \frac{\omega}{\pi} \cdot \frac{1}{1 + \omega^2 x^2}.$$

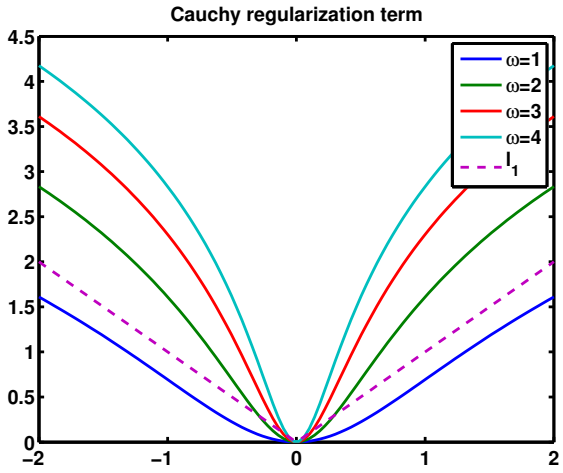


Figure 27: Plot of the one dimensional Cauchy regularization term for different ω-values. The dotted line is the absolute value for comparison.

This definition holds for one-dimensional variables and can be extended to the multivariate case by taking the product over all dimensions. Considering the negative logarithm and disregarding the constant term, then leads to:

$$\mathcal{R}_C(x) = \sum_{i=1}^{N} \log(1 + \omega^2 x_i^2).$$

This is a differentiable penalty term, where the penalization of small entries is controlled by the parameter ω. With higher values of ω, small values of x will be penalized stronger, see Figure 27.

As we want to deal with problems in possibly infinite dimensional Hilbert spaces, we further generalize the regularization term to

$$\mathcal{R}_C(x) := \sum_{i \in I} \log(1 + \omega^2 \langle x, \phi_i \rangle^2),$$

where $\phi = (\phi_i)_{i \in I}$ is a complete orthonormal basis (or more generally a frame) of a separable Hilbert space $\mathcal{X}$. We assume that I is countable.

We are now interested in the behavior of the Tikhonov functional with Cauchy regularization term, i. e.,

$$\mathcal{J}_\alpha(x) = \|F(x) - y^\delta\|_Y^2 + \mathcal{R}_C(x). \tag{7.1}$$

Together with the general assumption on the noisy data, cf. (1.2):

$$\|y - y^\delta\|_Y = \|F(x^\dagger) - y^\delta\|_Y \leqslant \delta.$$

7.2 WELLPOSEDNESS

In Condition 1 of Chapter 2 we stated sufficient assumptions for a general Tikhonov functional $\mathcal{J}_\alpha = \mathcal{S}(F(x), y^\delta) + \alpha\mathcal{R}(x)$ to be a proper regularization method. Now, we only have to verify this conditions in terms of the Cauchy-Tikhonov functional introduced in (7.1).

Assumption 1. *Throughout this chapter we assume the following:*

1. *Let τ_X and τ_Y be the weak topologies on the Hilbert spaces X and Y.*
2. *The data-fitting term $\mathcal{S}$ is the squared error norm given on the Hilbert space Y, i.e., $\mathcal{S}(y, \bar{y}) := \|y - \bar{y}\|_Y^2$.*
3. *The forward operator F is weakly continuous, $\mathcal{D}(F)$ is weakly closed and $\mathcal{D}(F) \cap \mathcal{D}(\mathcal{R}) \neq \emptyset$.*

With this choice of $\mathcal{S}$ we can state that the conditions from Condition 1 on the data term are satisfied. Obviously the conditions on F are also satisfied. Therefore we only have to verify the assumptions on $\mathcal{R}$, which are stated in item 4 and 6 of Condition 1.

Lemma 7.1. *Let Assumption 1 hold. The Cauchy regularization term $\mathcal{R}_C(x)$ is weakly lower semi continuous.*

Proof. We have to show that:

$$\liminf_{k\to\infty} \mathcal{R}_C(x_k) \geqslant \mathcal{R}_C(x).$$

Now, if $\mathcal{R}_C(x) < \infty$:

$$\begin{aligned}
\mathcal{R}_C(x) &= \sum_{i\in I} \log(1 + \omega^2 \langle x, \phi_i\rangle^2) \\
&= \sum_{i\in I} \log(1 + \omega^2 (\underbrace{\liminf_{k\to\infty} \langle x_k, \phi_i\rangle}_{=\langle x, \phi_i\rangle})^2) \\
&= \sum_{i\in I} \log\left(\liminf_{k\to\infty} \left(1 + \omega^2 \langle x_k, \phi_i\rangle^2\right)\right) \\
&= \sum_{i\in I} \liminf_{k\to\infty} \log(1 + \omega^2 \langle x_k, \phi_i\rangle^2) \\
&= \liminf_{k\to\infty} \sum_{i\in I} \log(1 + \omega^2 \langle x_k, \phi_i\rangle^2) = \liminf_{k\to\infty} \mathcal{R}_C(x_k).
\end{aligned}$$

Here the first equality holds just because of the weak convergence of $x_k \rightharpoonup x$ and the Riesz representation theorem. Then we can exchange the limit with the logarithm, because of the continuity of the quadratic and the logarithm function. Finally we can exchange summation and limit, because the sum is absolutely convergent.

Now consider the case $\mathcal{R}_C(x) = \infty$. Then for an arbitrary $M \in \mathbb{R}^+$, there exists $i_0 \in \mathbb{N}$ such that

$$\sum_{i=1}^{i_0} \log(1 + \omega^2 \langle x, \phi_i \rangle^2) > 2M.$$

Additionally, because of the weak convergence $x_k \rightharpoonup x$:

$$\forall \epsilon > 0 \; \exists k_0^i \; \forall k > k_0^i : \; |\langle x_k, \phi_i \rangle - \langle x, \phi_i \rangle| < \epsilon$$

holds. And as $f(x) := \log(1 + \omega^2 x^2)$ is continuous, there exists a $\delta > 0$ such that

$$\forall x_1, x_2 : |x_1 - x_2| < \delta \Rightarrow |f(x_1) - f(x_2)| < \frac{M}{i_0}.$$

Now let $\epsilon := \delta$, then for all $k > \max\{k_0^1, ..., k_0^{i_0}\}$:

$$\begin{aligned}
\mathcal{R}_C(x_k) &\geqslant \sum_{i=1}^{i_0} \log(1 + \omega^2 \langle x_k, \phi_i \rangle^2) \\
&= \sum_{i=1}^{i_0} \log(1 + \omega^2 \langle x, \phi_i \rangle^2) - \\
&\qquad \sum_{i=1}^{i_0} \Big(\log(1 + \omega^2 \langle x, \phi_i \rangle^2) - \\
&\qquad\qquad \log(1 + \omega^2 \langle x_k, \phi_i \rangle^2) \Big) \\
&\geqslant 2M - \\
&\qquad \sum_{i=1}^{i_0} \underbrace{\Big(\log(1 + \omega^2 \langle x, \phi_i \rangle^2) - \log(1 + \omega^2 \langle x_k, \phi_i \rangle^2) \Big)}_{\leqslant \frac{M}{i_0} \quad \forall k > \max\{k_0^1, ..., k_0^{i_0}\}} \\
&\geqslant 2M - M = M,
\end{aligned}$$

i. e., $\liminf_{k \to \infty} \mathcal{R}_C(x_k) = \infty$, since M was arbitrary. □

Lemma 7.2. *Let Assumption 1 hold. For every $\alpha > 0$, $y \in Y$ and $M > 0$ the level sets*

$$\mathcal{M}_{\alpha,y} := \{x \in X : \mathcal{J}_\alpha = \|F(x) - y^\delta\|_Y^2 + \mathcal{R}_C(x) \leqslant M\}$$

are weakly sequentially compact.

Proof. To show that the level sets are weakly sequentially compact we use the fact that every bounded sequence in a reflexive Banach space exhibits a weakly convergent subsequence. Therefore it is sufficient to show that any sequence $(x_k) \in \mathcal{M}_{\alpha,y}$ is bounded, and that the weak limit lies in $\mathcal{M}_{\alpha,y}$.

As in case of the classical Tikhonov regularization, the following inequality holds:

$$\alpha \mathcal{R}_C(x) \leqslant \mathcal{J}_{\alpha,y}(x) \leqslant M \Rightarrow \mathcal{R}_C(x) \leqslant \frac{M}{\alpha} := M \qquad \forall x \in \mathcal{M}_{\alpha,y}.$$

Therefore it is sufficient to show that from boundedness of $\mathcal{R}_C(x_k)$, boundedness of x_k follows. Which we will show by contradiction and hence assume $(x_k)_{k\in\mathbb{N}}$ to be unbounded. Thus there exists a subsequence, for simplicity denoted by $(x_k)_{k\in\mathbb{N}}$ again, such that

$$\|x_k\| \to \infty \text{ as } k \to \infty. \tag{7.2}$$

First note, that generally for every $C > 0$ there exists an $\eta > 0$ such that:

$$\log(1 + \omega^2 x^2) \geqslant \eta x^2, \tag{7.3}$$

for all $x \in [0, C]$. Which follows from the fact, that $\log(1 + \omega^2 x^2)$, as well as x^2, are monotonically increasing.

Now, we can conclude from

$$\mathcal{R}_C(x_k) := \sum_{i\in I} \log(1 + \omega^2 \langle x_k, \phi_i\rangle^2) \leqslant M, \tag{7.4}$$

that for all $k \in \mathbb{N}$ and for all $i \in I$: $\log(1 + \omega^2 \langle x_k, \phi_i\rangle^2) \leqslant M$. And therewith

$$\langle x_k, \phi_i\rangle \leqslant \sqrt{\frac{e^M - 1}{\omega^2}} =: C, \quad \forall k \in \mathbb{N}, \forall i \in I.$$

Hence, there exists an η such that:

$$\begin{aligned}\mathcal{R}_C(x_k) &= \sum_{i\in I} \log(1 + \omega^2 \langle x_k, \phi_i\rangle^2) \\ &\geqslant \sum_{i\in I} \eta \langle x_k, \phi_i\rangle^2 = \eta \|x_k\|^2 \to \infty \text{ by (7.2)},\end{aligned}$$

which contradicts (7.4). Here the first inequality follows from (7.3) and the second equality is just the Parseval equality.

Therewith $\|x_k\|$ is bounded and hence also the sequence (x_k) is bounded and exhibits a weakly convergent subsequence. By weak lower semicontinuity of $\mathcal{J}_\alpha$ (cf. point 3 in Assumption 1, as well as weak lower semicontinuity of the norm and of $\mathcal{R}_C$, according to Lemma 7.1) the weak limit lies in $\mathcal{M}_{\alpha,y}$. Thus the level sets $\mathcal{M}_{\alpha,y}$ are weakly sequentially compact. □

Remark 15. The main idea, Inequality (7.3), of the proof is taken from the proof of Lemma 3.3 in [12], where coercivity for a wide range of regularization functionals is shown. But as the Cauchy regularization term does not fit completely in the framework established in [12], see also Remark 16 below, we have given the complete proof here.

With Lemmas 7.1, 7.2, and the fact that due to $\mathcal{R}_C(0) = 0$, the functional $\mathcal{R}_C$ is proper, all points from Condition 1 are satisfied and we end up with the following theorem:

Theorem 7.3. *Let Assumption 1 hold. The Cauchy Tikhonov functional is a stable, well-posed and convergent regularization functional.*

a) (stability) *The minimizers of $\mathcal{J}_\alpha$ are stable with respect to the data y^δ. That is, for every sequence $y_k \to y$, the sequence of minimizers (x_k), $x_k := \arg\min \mathcal{J}_\alpha(x)$ has a subsequence, which converges weakly, and the limit of every subsequence is a minimizer of $\mathcal{J}_\alpha$.*

b) (well-definedness) *If there exists a solution to $F(x) = y$, then there exists a $\mathcal{R}_C$-minimizing solution, that is,*

$$x^\dagger \in \arg\min\{\mathcal{R}_C(x) : x \in \mathcal{X}, F(x) = y\}.$$

c) (weak convergence) *Assume there exists a solution to $F(x) = y$. Then, for a parameter choice with*

$$\alpha \to 0, \quad \frac{\delta^2}{\alpha} \to 0, \quad \text{for } \delta \to 0,$$

there exists a subsequence of (x_δ^α) which converges weakly to a $\mathcal{R}_C$-minimizing solution $x^\dagger$ of $F(x) = y$. Additionally

$$\mathcal{R}_C(x_\delta^\alpha) \to \mathcal{R}_C(x^\dagger).$$

Proof. For the proof of the theorem, we can refer back to the results on general Tikhonov regularization introduced in Chapter 2. The above stated results follow from Lemma 2.2, Lemma 2.3, Lemma 2.4 and Lemma 2.5, which are based on the Theorems 3.2, 3.4 and 3.5 in [48], the assertion $\mathcal{R}_C(x_\delta^\alpha) \to \mathcal{R}_C(x^\dagger)$ being contained in the proof of Theorem 3.5. □

Let from now on $\mathcal{X} = \ell_2(\mathbb{N})$, the space of square-summable sequences. If $\mathcal{X} = \ell_2(\mathbb{N})$ we can also show strong convergence of the minimizers, because the Cauchy regularization term satisfies the Kadec property.

Lemma 7.4 (Strong convergence)**.** *Let Assumption 1 hold and assume that $\mathcal{X} = \ell_2(\mathbb{N})$. Then for a parameter choice according to c) of Theorem 7.3 it follows that (x_δ^α) has a strong convergent subsequence and each limit of the subsequence is a $\mathcal{R}_C$-minimizing solution of $F(x) = y$.*

Proof. The proof follows the lines of the proof of Lemma 3.6 and Remark 3.7 in [12]. For simplicity we define $r(x) := \log(1 + \omega^2 x^2)$ and consider

$$\tilde{\mathcal{R}}_C(x) := \sum_{i \in \mathbb{N}} r(|x_i|),$$

which is equivalent to $\mathcal{R}_C(x)$. Now we can define r on $r : [0,\infty) \to [0,\infty)$, where it is strict monotonically increasing and invertible, with $r^{-1}(y)^2 = \frac{1}{\omega^2}(e^y - 1)$.

According to Theorem 7.3 c), x_δ^α has a weakly convergent subsequence, which we denote by x^n and $x^n \rightharpoonup x^\dagger$ in ℓ_2. Additionally according to Theorem 7.3 c) $\tilde{\mathcal{R}}_C(x^n) \to \tilde{\mathcal{R}}_C(x^\dagger) < \infty$.

As we are dealing with x in the Hilbert spaces ℓ_2 it is enough to show $\|x^n\| \to \|x^\dagger\|$, which implies strong convergence.

Since the x^n converge weakly, they are uniformly bounded by a constant $L > 0$, such that $|x_i^n|, |x_i^\dagger| < L$ for all i, n. Additionally $x_i^n \to x_i^\dagger$ for $n \to \infty$ holds for all i.

Now as the $|x_i^n|, |x_i^\dagger|$ are bounded and $r^{-1}(x)^2$ is locally Lipschitz continuous, there is a second constant $C(L)$ such that

$$\left|r^{-1}(|x_i^n|)^2 - r^{-1}(|x_i^\dagger|)^2\right| \leqslant C(L)\left||x_i^n| - |x_i^\dagger|\right|.$$

Set $\tilde{L} = r(L)$ and $\hat{x} := r(|x_i^n|)$, $\bar{x} := r(|x_i^\dagger|)$, then as r is strictly monotonically increasing, $|\hat{x}|, |\bar{x}| \leqslant \tilde{L}$. Hence

$$\left|r^{-1}(\hat{x})^2 - r^{-1}(\bar{x})^2\right| \leqslant C(\tilde{L})\left|\hat{x} - \bar{x}\right|,$$

which is equivalent to

$$\left||x_i^n|^2 - |x_i^\dagger|^2\right| \leqslant C(\tilde{L})\left|r(|x_i^n|) - r(|x_i^\dagger|)\right|.$$

From this we can conclude

$$\left|\|x^n\|^2 - \|x^\dagger\|^2\right| \leqslant C(\tilde{L}) \sum_{i\in\mathbb{N}} \left|r(|x_i^n|) - r(|x_i^\dagger|)\right|. \tag{7.5}$$

Define now $\tilde{x}_k^n := \min\{r(|x_k^n|), r(|x_k^\dagger|)\}$. Since r is continuous, $\lim_{n\to\infty} \tilde{x}_k^n = r(|x_k^\dagger|)$ and as $\tilde{\mathcal{R}}_C(x^\dagger) < \infty$ we get by dominated convergence

$$\lim_{n\to\infty} \sum_{i\in\mathbb{N}} \tilde{x}_k^n = \tilde{\mathcal{R}}_C(x^\dagger).$$

As $|a - b| = a + b - 2\min\{a, b\}$, we see that

$$\sum_{i\in\mathbb{N}} \left|r(|x_i^n|) - r(|x_i^\dagger|)\right| = \sum_{i\in\mathbb{N}} r(|x_i^n|) + r(|x_i^\dagger|) - 2\tilde{x}_i^n \geqslant 0.$$

Hence taking the limit $n \to \infty$ and keeping in mind that $\tilde{\mathcal{R}}_C(x^n)$ converges to $\tilde{\mathcal{R}}_C(x^\dagger)$ leads to

$$\lim_{n\to\infty} \sum_{i\in\mathbb{N}} \left|r(|x_i^n|) - r(|x_i^\dagger|)\right| = 2\tilde{\mathcal{R}}_C(x^\dagger) - 2\lim_{n\to\infty} \sum_{k\in\mathbb{N}} \tilde{x}_k^n = 0,$$

which together with (7.5) proves the assertion. □

7.3 CONVERGENCE RATE

As the Cauchy term is not convex, the standard approaches to show convergence rates (see [73]) cannot be applied, because they rely on convexity of the regularization term as a crucial assumption. Recently, Markus Grasmair has shown in [31] how it is possible to generalize the proof of convergence rates for Tikhonov functionals even in the case of a non-convex regularization term. For this purpose he had to introduce a slightly weaker form of convexity, called W-convexity. In the next few paragraphs, we will outline his approach, introduce the notion of W-convexity, and finally demonstrate how we can use this generalization to show convergence rates for the Cauchy regularization term.

7.3.1 *Generalized convexity*

The generalization of convexity mainly follows the exposition in [31], which is based on the derivation in [83]. Before giving a generalized notion of convexity, we have to introduce some notation for addition and subtraction on the extended real line $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$.

Definition 7.5 (Definition 2.1 in [31])**.** *We define the upper and lower addition and subtraction on $\mathbb{R} \cup \{\pm\infty\}$ as the extensions of the usual definitions satisfying:*

$$\begin{aligned} +\infty \dot{+} (-\infty) &= -\infty \dot{+} \infty = +\infty \\ +\infty \dot{+} (-\infty) &= -\infty \dot{+} \infty = -\infty \\ +\infty \dot{-} (+\infty) &= -\infty \dot{-} (-\infty) = +\infty \\ +\infty \dot{-} (+\infty) &= -\infty \dot{-} (-\infty) = -\infty \end{aligned}$$

Definition 7.6 (Definition 2.2 in [31])**.** *Let $\mathcal{X}$ be some set and W a family of functions $w : \mathcal{X} \to \mathbb{R}$. The (generalized) Fenchel conjugate of a function $\mathcal{R} : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ with respect to W is the function $\mathcal{R}^* : W \to \mathbb{R} \cup \{\pm\infty\}$ defined by:*

$$\mathcal{R}^*(w) := \sup_{x \in \mathcal{X}} [w(x) \dot{-} \mathcal{R}(x)].$$

The double conjugate of $\mathcal{R}$ is the function $\mathcal{R}^{**} : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$: given by

$$\mathcal{R}^{**}(x) := \sup_{w \in W} [w(x) \dot{-} \mathcal{R}^*(w)] = \sup_{w \in W} \inf_{\tilde{x} \in \mathcal{X}} [w(x) \dot{+} (\mathcal{R}(\tilde{x}) \dot{-} w(\tilde{x})].$$

Now we call a function $\mathcal{R}$ convex with respect to W, if $\mathcal{R} = \mathcal{R}^{**}$. We call it locally convex at $x \in \mathcal{X}$ with respect to W, if $\mathcal{R}(x) = \mathcal{R}^{**}(x)$.

If we look at the definition of the double conjugate in more detail, we see that we call a function locally convex at $x \in \mathcal{X}$ with respect to W, if and only if for every $\epsilon > 0$ there exists $w_\epsilon \in W$ such that

$$\mathcal{R}(\tilde{x}) \geqslant \mathcal{R}(x) \dot{+} (w_\epsilon(\tilde{x}) \dot{-} w_\epsilon(x)) - \epsilon \tag{7.6}$$

for all $\tilde{x} \in \mathcal{X}$ (cf. Remark 2.1 in [31]). With this generalized definition of the Fenchel conjugate, we can introduce a generalization of the subdifferential of a function $\mathcal{R}$.

In the following, we always assume that W is a non empty family of functions $w : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$.

Definition 7.7 (Definition 2.3 in [31]). *Let $\mathcal{R}$ be locally convex at $x \in \mathcal{X}$ with respect to W and assume $\mathcal{R}(x) \in \mathbb{R}$. The W-subdifferential of $\mathcal{R}$ at $x \in \mathcal{X}$, denoted by $\partial_W \mathcal{R}(x)$, is defined as the set of all $w \in W$ that satisfy $w(x) \in \mathbb{R}$ and*

$$\mathcal{R}(\tilde{x}) \geqslant \mathcal{R}(x) + (w(\tilde{x}) - w(x))$$

for all $\tilde{x} \in \mathcal{X}$.

Definition 7.8 (Definition 2.4 in [31]). *Let $\mathcal{R}$ be locally convex at $x \in \mathcal{X}$ with respect to W and assume $\partial_W \mathcal{R}(x) \neq \emptyset$. For $w \in \partial_W \mathcal{R}(x)$ and $\tilde{x} \in \mathcal{X}$ we define the W-Bregman distance between x and $\tilde{x}$ with respect to w as*

$$D_W^w(x; \tilde{x}) := (\mathcal{R}(\tilde{x}) - \mathcal{R}(x)) \dot{-} (w(\tilde{x}) - w(x)).$$

The W-Bregman distance is non-negative and satisfies $D_W^w(x; x) = 0$.

We will now have a closer look at a small example, once again taken from [31]. We will use the insights gained from the example later on to show that the Cauchy regularization term is locally convex with respect to W_2 introduced in Definition 7.9 below.

Example 4. Let us consider the family of all locally negative semi-definite, continuous quadratic functions on a locally convex space $\mathcal{X}$, which we denote by W. That is

$$w \in W \Leftrightarrow w(x) = c + \langle \xi, x \rangle - A(x, x) \quad \forall x \in U,$$

with $c \in \mathbb{R}$, $\xi \in \mathcal{X}^*$ and A a positive semi-definite, symmetric bounded quadratic form on $\mathcal{X}$. Here U is a neighborhood of an element $x_0 \in \mathcal{X}$.

If we now apply the definition of local convexity of (7.6), then a function $\mathcal{R} : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ is locally convex at $x \in \mathcal{X}$ with respect to W, if and only if there exists for every $\epsilon > 0$ a $\xi_\epsilon \in \mathcal{X}^*$ and a positive semi-definite, symmetric, bounded quadratic form A_ϵ on $\mathcal{X}$ such that

$$\begin{aligned} \mathcal{R}(\tilde{x}) + \epsilon &\geqslant \mathcal{R}(x) + \langle \xi_\epsilon, \tilde{x} \rangle - A_\epsilon(\tilde{x}, \tilde{x}) - \langle \xi_\epsilon, x \rangle + A_\epsilon(x, x) \\ &= \mathcal{R}(x) + \langle \xi_\epsilon, \tilde{x} - x \rangle - \\ &\quad A_\epsilon(\tilde{x} - x, \tilde{x} - x) - 2A_\epsilon(x, \tilde{x} - x) \end{aligned}$$

for all $\tilde{x} \in U$. Now we can define $\tilde{\xi_\epsilon} \in X^*$ by $\langle \tilde{\xi_\epsilon}, \hat{x} \rangle := \langle \xi_\epsilon, \hat{x} \rangle - 2A_\epsilon(x, \hat{x})$ and obtain that $\mathcal{R}$ is locally convex at $x \in X$ with respect to W if and only if for every $\epsilon > 0$ there exists a $\tilde{\xi_\epsilon} \in X^*$ and a positive semi-definite, symmetric, bounded quadratic form A_ϵ on X such that

$$\mathcal{R}(\tilde{x}) + \epsilon \geqslant \mathcal{R}(x) + \langle \tilde{\xi_\epsilon}, \tilde{x} - x \rangle - A_\epsilon(\tilde{x} - x, \tilde{x} - x) \tag{7.7}$$

for all $\tilde{x} \in U$. This directly influences the definition of the W-subdifferential. The W-subdifferential of $\mathcal{R}(x)$ at x consists of all functions $w(\tilde{x}) = a + \langle \xi, \tilde{x} - x \rangle - A(\tilde{x} - x, \tilde{x} - x)$ that satisfy

$$\mathcal{R}(\tilde{x}) \geqslant \mathcal{R}(x) + \langle \xi, \tilde{x} - x \rangle - A(\tilde{x} - x, \tilde{x} - x) \tag{7.8}$$

for all $\tilde{x} \in U$.

The last equation shows a great similarity to the notion of a proximal subdifferential of a function $\mathcal{R}(x)$ (see [18]).

We will now show the W-convexity of the Cauchy regularization term. For this purpose, differentiability of the Cauchy term will be essential.

7.3.2 *W-convexity of* $\mathcal{R}_C(x)$

First we introduce the family of functions for which the Cauchy regularization term is W-convex.

Definition 7.9 (Definition 5.1 in [31]). *We define W_2 as the set of all functions $w : \ell_2(\mathbb{N}) \to \mathbb{R}$ for which there exists $\rho > 0$, an element $x_0 \in \ell_2(\mathbb{N})$, $\xi \in \ell_2(\mathbb{N})$ and $c > 0$ such that*

$$w(\tilde{x}) = \langle \xi, \tilde{x} - x_0 \rangle - c \sum_{i \in \mathbb{N}} |\tilde{x}_i - x_{0i}|^2 \tag{7.9}$$

for all $\tilde{x} \in \ell_2(\mathbb{N})$ with $\|\tilde{x} - x_0\| < \rho$.

Using Example 4 with $a = -\langle \xi, x_0 \rangle$, $A(\tilde{x}, \bar{x}) = c \sum_{i \in \mathbb{N}} (x_{0i} - \tilde{x}_i)(x_{0i} - \bar{x}_i)$, we obtain that a functional $\mathcal{R} : X \to \mathbb{R} \cup \{\pm\infty\}$ is locally convex at $x \in X$ with respect to W_2 if and only if for every $\epsilon > 0$ there exists a $\tilde{\xi_\epsilon} \in X^*$, and $x_0 \in X$, and a $c_\epsilon > 0$ such that

$$\mathcal{R}(\tilde{x}) + \epsilon \geqslant \mathcal{R}(x) + \langle \tilde{\xi_\epsilon}, \tilde{x} - x \rangle - c_\epsilon \|x_0 - \tilde{x} + x\|^2$$

for all $\tilde{x} \in X$. Moreover, the W_2-subdifferential of such an $\mathcal{R}$ consists of all w of the form (7.9) such that for all $\tilde{x} \in \ell_2(\mathbb{N})$ with $\|\tilde{x} - x_0\| < \rho$,

$$\mathcal{R}(\tilde{x}) \geqslant \mathcal{R}(x) + \langle \xi, \tilde{x} - x \rangle - c(\|x_0 - \tilde{x}\|^2 - \|x_0 - x\|^2),$$

and therefore especially contains all w of the form (7.9) with $x_0 = x$ such that for all $\tilde{x} \in \ell_2(\mathbb{N})$ with $\|\tilde{x} - x\| < \rho$,

$$\mathcal{R}(\tilde{x}) \geqslant \mathcal{R}(x) + \langle \xi, \tilde{x} - x \rangle - c\|x - \tilde{x}\|^2.$$

Lemma 7.10. *The following holds for the Cauchy regularization term.*

i) The Cauchy regularization term

$$\mathcal{R}_C(x) = \sum_{i \in I} \log(1 + \omega^2 \langle x, \phi_i \rangle^2)$$

is locally W_2-convex.

ii) Moreover, the W_2-subdifferential of $\mathcal{R}_C$ at x contains all functions of the form $w(\tilde{x}) = \langle \xi, \tilde{x} - x \rangle - c\|\tilde{x} - x\|^2$ such that $\xi_i = \frac{2\omega^2 x_i}{1+\omega^2 x_i^2}$ and $c \geqslant \omega^2$.

The proof of Lemma 7.10 is a combination of the differentiability of $\mathcal{R}_C(x)$ and the example given above. All we have to show for i) is that there exists a ξ and a $c > 0$ such that (7.7) holds with $A(x, \tilde{x}) = c \sum_{i \in \mathbb{N}} x_i \tilde{x}_i$. It turns out that ξ is just the gradient of $\mathcal{R}_C(x)$ and c a bound for the second derivative. We will show the existence of the second derivative for more general regularization functionals of the form $\mathcal{R}(x) = \sum_{i \in I} r(x_i)$.

Lemma 7.11. *Let $\mathcal{R}(x) := \sum_{i \in \mathbb{N}} r(x_i)$, $x \in \ell_2(\mathbb{N})$, with $r(x) \geqslant 0$. If r is differentiable and*

- $\| (r'(x_i))_{i \in \mathbb{N}} \|_{\ell_2} < \infty$ *for* $x \in \ell_2(\mathbb{N})$.
- $|r'(x_1) - r'(x_2)| \leqslant L_1 |x_1 - x_2|$ *for* $x_1, x_2 \in \mathbb{R}$ *and* $L_1 > 0$

then $\mathcal{R}(x)$ is Fréchet differentiable, with

$$\mathcal{R}'(x)[h] = \sum_{i \in \mathbb{N}} r'(x_i) h_i,$$

$h = (h_i)_{i \in \mathbb{N}} \in \ell_2(\mathbb{N})$. If additionally r is twice differentiable and

- $\sum_{i \in \mathbb{N}} r''(x_i) h_{1_i} h_{2_i} < \infty$ *for* $x \in \ell_2(\mathbb{N})$ *and* $h_1, h_2 \in \ell_2(\mathbb{N})$
- $|r''(x_1) - r''(x_2)| \leqslant L_2 |x_1 - x_2|$ *for* $x_1, x_2 \in \mathbb{R}$ *and* $L_2 > 0$

then $\mathcal{R}'(x)$ is Fréchet differentiable, with

$$\mathcal{R}''(x)[h_1, h_2] = \sum_{i \in \mathbb{N}} r''(x_i) h_{1_i} h_{2_i}.$$

Proof. We will first show, that

$$|\mathcal{R}(x + h) - \mathcal{R}(x) - \mathcal{R}'(x)[h]| = o(\|h\|_{\ell_2}),$$

for $h \in \ell_2(\mathbb{N})$, with $\mathcal{R}'(x)[h] = \sum_{i \in \mathbb{N}} r'(x_i) h_i$. To do so, we can use

$$|\mathcal{R}(x+h) - \mathcal{R}(x) - \mathcal{R}'(x)[h]| = |\sum_{i \in \mathbb{N}} r(x_i + h_i) - r(x_i) - r'(x_i) h_i|,$$

as $\sum_{i\in\mathbb{N}} r'(x_i)h_i < \infty$ by the assumption $\| (r'(x_i))_{i\in\mathbb{N}} \|_{\ell_2} < \infty$ and the Cauchy Schwarz inequality:

$$\sum_{i\in\mathbb{N}} r'(x_i)h_i \leqslant \| \left(r'(x_i)\right)_{i\in\mathbb{N}} \|_{\ell_2} \|h\|_{\ell_2}.$$

Now

$$\begin{aligned}
&|\sum_{i\in\mathbb{N}} r(x_i+h_i) - r(x_i) - r'(x_i)h_i| \\
=&|\sum_{i\in\mathbb{N}} \int_0^1 r'(x_i+th_i)dth_i - r'(x_i)h_i| \\
=&|\sum_{i\in\mathbb{N}} \int_0^1 (r'(x_i+th_i) - r'(x_i))h_i dt| \\
\leqslant& \sum_{i\in\mathbb{N}} \int_0^1 (|r'(x_i+th_i) - r'(x_i)|)h_i dt \\
\leqslant& \sum_{i\in\mathbb{N}} \int_0^1 (L_1 th_i)h_i dt \\
=&\frac{L_1}{2} \sum_{i\in\mathbb{N}} h_i^2 = \frac{L_1}{2}\|h\|_{\ell_2}^2 = o(\|h\|_{\ell_2}),
\end{aligned}$$

where the second inequality hold, because of the assumed Lipschitz continuity of $r'(x)$. This shows Fréchet differentiability of $\mathcal{R}(x) = \sum_{i\in\mathbb{N}} r(x_i)$, and $\mathcal{R}'(x)[h] = \sum_{i\in\mathbb{N}} r'(x_i)h_i$. In the same way, we can show Fréchet differentiability of $\mathcal{R}'(x)$ and $\mathcal{R}''(x)[h_1,h_2] = \sum_{i\in\mathbb{N}} r''(x_i)h_{1_i}h_{2_i}$.

If $\sum_{i\in\mathbb{N}} r''(x_i)h_{1_i}h_{2_i} < \infty$

$$\begin{aligned}
&|\mathcal{R}'(x+h_2)[h_1] - \mathcal{R}'(x)[h_1] - \mathcal{R}''(x)[h_1,h_2]| \\
&= |\sum_{i\in\mathbb{N}} r'(x_i+h_{2_i})h_{1_i} - r'(x_i)h_{1_i} - r''(x_i)h_{1_i}h_{2_i}|.
\end{aligned}$$

Now as before

$$\begin{aligned}
&|\sum_{i\in\mathbb{N}} r'(x_i+h_{2_i})h_{1_i} - r'(x_i)h_{1_i} - r''(x_i)h_{1_i}h_{2_i}| \\
&= |\sum_{i\in\mathbb{N}} h_{1_i} \left(\int_0^1 r''(x_i+th_{2_i})dth_{2_i} - r''(x_i)h_{2_i} \right)| \\
&\leqslant \sum_{i\in\mathbb{N}} |h_{1_i}| \frac{L_2}{2} h_{2_i}^2 \\
&\leqslant \max_{i\in\mathbb{N}}\{|h_{1_i}|\} \frac{L_2}{2} \sum_{i\in\mathbb{N}} h_{2_i}^2 = o(\|h_2\|_{\ell_2}).
\end{aligned}$$

Once again the first inequality holds, because of the Lipschitz continuity of $r''(x)$. □

Now we can prove Lemma 7.10 by showing that the required properties from Lemma 7.11 hold for $r(x) := \log(1+\omega^2 x^2)$.

Proof of Lemma 7.10. In case of the Cauchy regularization term $r(x) = \log(1 + \omega^2 x^2)$. Therewith

$$\begin{aligned} r'(x) &= \frac{2\omega^2 x}{1 + \omega^2 x^2}, \\ r''(x) &= 2\omega^2 \frac{1 - \omega^2 x^2}{(1 + \omega^2 x^2)^2}, \\ r'''(x) &= 4\omega^4 \frac{x(\omega^2 x^2 - 3)}{(1 + \omega^2 x^2)^3}. \end{aligned}$$

Now Lipschitz continuity of $r'(x)$ follows, because $r''(x)$ is bounded from above, $|r''(x)| \leqslant 2\omega^2$. The same holds true for $r''(x)$, as $|r'''(x)| < 4\omega^3 \frac{\sqrt{3\sqrt{5}-5}}{(4-\sqrt{5})^3}$. The latter can be seen by computing the extremal values of r''' and using the fact that $r'''(x) \to 0$ for $x \to \infty$.

For $x \in \ell_2(\mathbb{N})$

$$\begin{aligned} \|(r'(x_i))_{i\in\mathbb{N}}\|_{\ell_2} &= \sum_{i\in\mathbb{N}} \left(2\omega^2 \frac{x_i}{1 + \omega^2 x_i^2}\right)^2 \leqslant 4\omega^4 \sum_{i\in\mathbb{N}} x_i^2 \\ &= 4\omega^4 \|x\|_{\ell_2}^2 < \infty. \end{aligned}$$

Additionally for $h_1, h_2 \in \ell_2$

$$\begin{aligned} \sum_{i\in\mathbb{N}} r''(x) h_{1_i} h_{2_i} &= \sum_{i\in\mathbb{N}} 2\omega^2 \underbrace{\frac{1 - \omega^2 x_i^2}{1 + \omega^2 x_i^2}}_{\leqslant 1} h_{1_i} h_{2_i} \\ &\leqslant 2\omega^2 \sum_{i\in\mathbb{N}} h_{1_i} h_{2_i} \\ &\leqslant 2\omega^2 \|h_1\|_{\ell_2} \|h_2\|_{\ell_2} < \infty. \end{aligned}$$

Hence Lemma 7.11 applies. Now $\mathcal{R}_C(\tilde{x}) - \mathcal{R}_C(x) = \sum_{i\in\mathbb{N}} r(\tilde{x}_i) - r(x_i)$ with $r(\lambda) := \log(1 + \omega^2\lambda^2)$. As $r(\lambda)$ is twice differentiable, we can consider the Taylor expansion with remainder and get:

$$\mathcal{R}_C(\tilde{x}) - \mathcal{R}_C(x) = \sum_{i\in\mathbb{N}} \left(r'(x_i)(\tilde{x}_i - x_i) + \frac{1}{2} r''(z_i)(\tilde{x}_i - x_i)^2\right),$$

with z_i an element of the line segment between $\tilde{x}_i$ and x_i. Setting $\xi = (\xi_i)_{i\in\mathbb{N}} = (r'(x_i))_{i\in\mathbb{N}}$ leads to

$$\mathcal{R}_C(\tilde{x}) - \mathcal{R}_C(x) = \langle \xi, \tilde{x} - x \rangle_{\ell_2(\mathbb{N})} + \frac{1}{2} \sum_{i\in\mathbb{N}} r''(z_i)(\tilde{x}_i - x_i)^2.$$

Additionally, $r''(\lambda)$ is bounded, namely $|r''(\lambda)| \leqslant 2\omega^2$ for all λ, and therefore

$$\begin{aligned} \mathcal{R}_C(\tilde{x}) - \mathcal{R}_C(x) &= \langle \xi, \tilde{x} - x \rangle + \frac{1}{2} \sum_{i\in\mathbb{N}} r''(z_i)(\tilde{x}_i - x_i)^2 \\ &\geqslant \langle \xi, \tilde{x} - x \rangle - \omega^2 \sum_{i\in\mathbb{N}} (\tilde{x}_i - x_i)^2, \end{aligned}$$

which proves the assertion. □

With the local W_2-convexity of $\mathcal{R}_C$ we can apply the results by Grasmair to establish a convergence rate for the Cauchy-Tikhonov regularization functional.

7.3.3 *Convergence rates under a variational inequality*

Grasmair actually shows his results for a general regularization functional, with general data-fitting term and general regularization functional:

$$\mathcal{J}_\alpha(x, y) := \mathcal{S}(F(x), y) + \alpha\mathcal{R}(x).$$

As the results stated below hold in this general setting, we will also formulate them generally here, despite the fact that our analysis is mainly concerned with the special Cauchy-Tikhonov functional introduced in (7.1).

We now introduce the generalized variational inequality necessary for the derivation of convergence rates.

Definition 7.12 (Definition 3.1 in [31]). *Let W be a family of extended real valued functions on $\mathcal{X}$ and assume that $\mathcal{R}(x)$ is W-convex at $x^\dagger$ and $\partial_W\mathcal{R}(x^\dagger) \neq \emptyset$. We say that the regularization method satisfies a variational inequality at $x^\dagger \in \mathcal{X}$ with respect to W if there exist $\beta > 0$, $\epsilon > 0$, a neighborhood U of $x^\dagger$, $w \in \partial_W\mathcal{R}(x^\dagger)$ and a concave, continuous, strictly increasing function $\Phi : [0, \infty) \to [0, \infty)$ satisfying $\Phi(0) = 0$ such that*

$$\beta D_W^w(x^\dagger; x) \leqslant (\mathcal{R}(x) - \mathcal{R}(x^\dagger)) + \Phi(\mathcal{S}(F(x), F(x^\dagger))) \tag{7.10}$$

for all $x \in \mathcal{D}(F) \cap U$ satisfying $|\mathcal{R}(x^\dagger) - \mathcal{R}(x)| < \epsilon$.

This generalized variational inequality can now be applied to prove convergence rates for Tikhonov regularization functionals, with non-convex but locally W-convex regularization terms.

Theorem 7.13 (Theorem 3.1 in [31]). *Assume that a variational inequality at $x^\dagger \in \mathcal{X}$ with respect to W is satisfied and let $\beta > 0$ and $\Phi : [0, \infty) \to [0, \infty)$ be as in definition 7.12. Let $\delta > 0$ and assume that $y^\delta \in \mathcal{Y}$ satisfies $\mathcal{S}(y, y^\delta) \leqslant \delta$. Moreover, let $x_\delta^\alpha := \arg\min_{x\in\mathcal{X}} \mathcal{J}_\alpha(x, y)$. Then for δ small enough, such that $|\mathcal{R}(x_\delta^\alpha) - \mathcal{R}(x^\dagger)| < \epsilon$ as in definition 7.12, the following hold.*

i) If $\gamma := \lim_{t\to 0^+} \frac{\Phi(t)}{t} < +\infty$ and $\alpha \leqslant \frac{1}{\gamma s}$, we have the estimate

$$\beta D_W^w(x^\dagger; x) \leqslant \frac{\delta}{\alpha} + \Phi(s\delta).$$

ii) If $\lim_{t\to 0^+} \frac{\Phi(t)}{t} = +\infty$, let $\Psi : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ be a conjugate of the convex mapping $t \mapsto \Phi^{-1}(st)$. Then we have, for a sufficiently small α, the estimate

$$\beta D_W^w(x^\dagger; x) \leqslant \frac{\delta}{\alpha} + \Phi(s\delta) + \frac{\Psi(\alpha)}{\alpha}.$$

Proof. See proof of Theorem 3.1 in [31]. Please be aware that, as the regularization method is convergent (Theorem 7.3) we assume x_δ^α to be in an ϵ-ball around $x^\dagger$ and that δ has to be small enough, such that $|\mathcal{R}(x_\delta^\alpha) - \mathcal{R}(x^\dagger)| < \epsilon$. □

Corollary 7.14 (Corollary 3.1 in [31])**.** *Let the assumptions of Theorem 7.13 be satisfied.*

(i) If $\gamma := \lim_{t\to 0^+} \frac{\Phi(t)}{t} < +\infty$ *we have for a constant parameter choice* $\alpha \leqslant \frac{1}{\gamma s}$ *the convergence rate*

$$D_W^w(x^\dagger; x_\delta^\alpha) = \mathcal{O}(\delta).$$

(ii) If $\lim_{t\to 0^+} \frac{\Phi(t)}{t} = +\infty$*, then we have for a parameter choice* $\alpha \sim \frac{\delta}{\Phi(s\delta)}$ *the convergence rate*

$$D_W^w(x^\dagger; x_\delta^\alpha) = \mathcal{O}(\Phi(s\delta)).$$

Proof. See proof of Corollary 3.1 in [31]. □

Corollary 7.15. *Let a variational inequality (7.10) with* $W := W_2$ *according to Definition 7.9 hold at* $x^\dagger$*. Then the rates given in Theorem 7.13 i), ii) and Corollary 7.14 i), ii) are also valid for the Tikhonov functional with Cauchy regularization term.*

Proof. According to Lemma 7.10, $\mathcal{R}_C$ is W_2 convex and therefore Theorem 7.13 can be applied. □

Remark 16. For results on convergence rates under a source condition we first of all refer to work by Markus Grasmair. He showed convergence rates under a source condition and the restricted injectivity condition. Additionally, he required a growth condition at zero for the regularization term (see Theorem 5.1 in [31]). Actually, this growth condition is not fulfilled by the Cauchy term.

In [32], Grasmair showed linear convergence rates for the sublinear ℓ_p regularization term with $0 < p < 1$. But again he needed a growth condition on the regularization term, which does not hold for the Cauchy functional.

Also recently, Kristian Bredies and Dirk Lorenz showed in [12] convergence rates for regularization terms with separable constraints. They considered terms of the general type $\mathcal{R}(x) = \sum_{i\in I} r(|x_i|)$, with $r(x)$ an arbitrary function. As Grasmair, they showed linear convergence rates in the l_1-norm, but demanded $r(x)$ to grow faster than $y = x$ at zero (Assumption 4.2 b) in [12]). Obviously the Cauchy term grows much slower. Another interesting approach by Bredies and Lorenz is to show convergence rates not in a given norm or other distance measure, but in the

regularization term itself, i. e., $\mathcal{R}(x_\delta^\alpha - x^\dagger) = \mathcal{O}(\delta)$. But for this approach, the regularization term has to fulfill

$$x, y \geqslant 0 \quad \Rightarrow \quad r(x) - r(y) \leqslant Cr(|x - y|),$$

for a $C \geqslant 1$ (Assumption 4.3 b) in [12]). This condition does not hold for the Cauchy term, as

$$\lim_{h \to 0} \frac{\log(1 + \omega^2 x^2) - \log(1 + \omega^2 (x + h)^2)}{\log(1 + \omega^2 h^2)} = \infty,$$

for $x \neq 0$.

7.4 NUMERICAL RESULTS

In this section we present numerical examples to show the performance of Tikhonov regularization with Cauchy regularization term. We compare the results to Tikhonov regularization with ℓ_1 regularization term and to the Refining and Coarsening algorithm introduced in Chapter 4. Due to differentiability of the Cauchy regularization term, it cannot be proven to yield sparse solutions. However, the results presented in this section show that it is very well able to enhance sparsity in practice, as expected from its role as a sparsity promoting prior in Bayesian inversion, cf. [51].

The MATLAB routine *fminunc* was used for minimizing the Cauchy-Tikhonov functional. As the Cauchy-Tikhonov functional is differentiable the analytical gradient of $\mathcal{J}_\alpha$ could be provided leading to a Quasi Newton method with BFGS update of the Hessian.

The ℓ_1 Tikhonov functional is minimized by using the well-known IST algorithm proposed by Daubechies, Defrise, De Mol [19]. It is known to perform slowly, but is one of the first algorithms to minimize infinite dimensional sparse inverse problems. For comparison we used the Semi Smooth Newton (SSN) method proposed by Griesse and Lorenz in [35], which applies an active set strategy and is faster than IST.

7.4.1 *Compressed Sensing*

Our first test relates back to the example of compressed sensing, cf. Example 3 of Chapter 3.

We set $N = 2048$, $K = 480$ and $T = 80$. Also we added 5% Gaussian noise to the exact data. In Figure 28, a comparison of the different minimizers is shown. For the Cauchy-Tikhonov minimization we used $\alpha = 4 \cdot 10^{-3}$ and set the sparsity promoting parameter $\omega = 35$. The shrinkage operator for the IST was set to $\bar{\gamma} = 2 \cdot 10^{-2}$. The free parameters of the Semi Smooth Newton

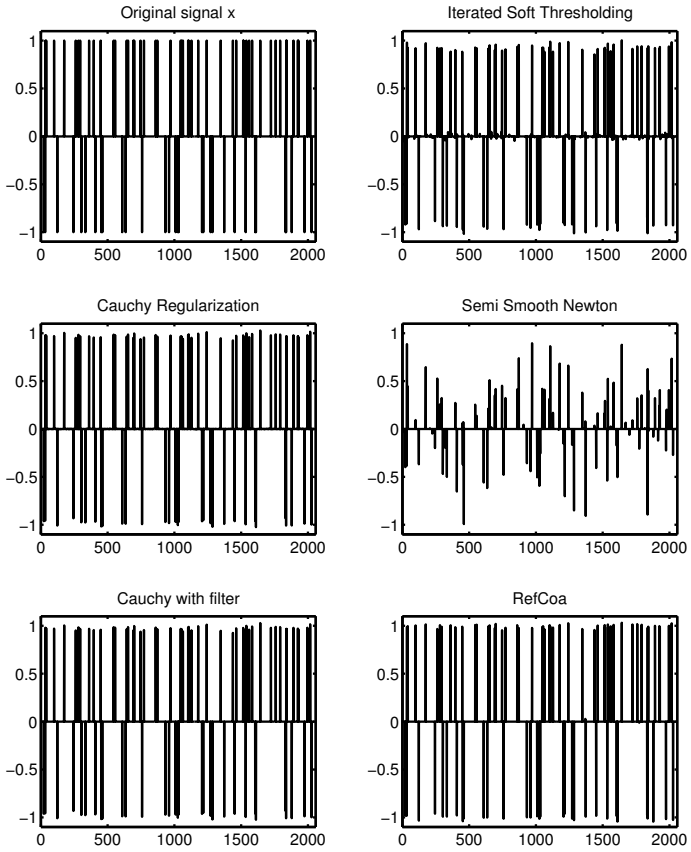


Figure 28: Comparison of the minimizers computed by the different algorithms. In the upper left, the original signal x is plotted.The upper right picture shows the minimizer obtained with Iterative Soft Thresholding. In the second row on the left is the minimizer of the Cauchy-Tikhonov functional and on the right is the minimizer of the Tikhonov functional with ℓ_1 term, obtained with the Semi Smooth Newton method. In the third row the filtered Cauchy minimizer and the solution of the Refinement and Coarsening algorithm are depicted on the left and right respectively.

method were set to $\tilde{\gamma} = 10^5$ and $\tilde{\omega} = 0.127$. We stopped the Refinement and Coarsening iteration if the maximum refining index was smaller than $\bar{r} = 0.01$. All parameters were adjusted by hand to gain the smallest possible error.

Apparently, the result of the SSN method is not as close to the exact signal as the results of the other algorithms, which can also be read off from the much higher error number listed in Table 5. However, the SSN method is much faster than the other algorithms. The minimization of the Cauchy-Tikhonov functional is faster than the IST algorithm and results in only half of the error of the IST. However as already mentioned in Chapter 4, the Refinement and Coarsening algorithm clearly provides the best result. It produces the smallest error number together with a fast computation time. None of the other strategies results in such a good approximation.

Therefore in this test case, the Cauchy-Tikhonov minimization produces an accurate solution in acceptable time. However the minimizer of the Cauchy functional is not an actual sparse solution, as can be seen by the number of non-zero entries of the minimizer, also listed in Table 5. Here, the SSN method and the Refinement and Coarsening algorithm perform best. This is because of their active set and projection strategies. Note that the coefficients of the Cauchy solution outside the support of the exact signal are very small. In this example, the maximum absolute value of a coefficient outside the support of the exact solution was $3 \cdot 10^{-3}$. Filtering the solution with a shrinkage operator as used in IST and SSN hence leads to a truly sparse solution, see also the last row of Table 5. There, the ℓ_1-error and the number of non-zero entries are listed for a filtered minimizer of the Cauchy-Tikhonov functional. We used a hard thresholding filter with the same thresholding value as for the IST, $\bar{\gamma} = 2 \cdot 10^{-2}$. The filtered Cauchy solution is then comparable to the result of the Refinement and Coarsening algorithm. It even provides a better reconstruction of the exact set of non-zero coefficients of $x^\dagger$, see the last column of table 5.

7.4.2 *Inverse Integration*

As a second example we used the Cauchy-Tikhonov regularization to solve the inverse problem of inverse integration, i. e., Example 1 introduced in Chapter 1. Again we assumed a given sparse exact solution $x^\dagger$ and sampled the noisy data by $y^\delta = Ax^\dagger + e$, where e is standard Gaussian noise. The data is given on $N = 600$ discretization points of the interval $[0, 1]$.

For the exact solution we generated a piecewise constant function, which zero is on large parts, see Figure 29. Additionally, we added 5% Gaussian noise to the exact data. The parameters in

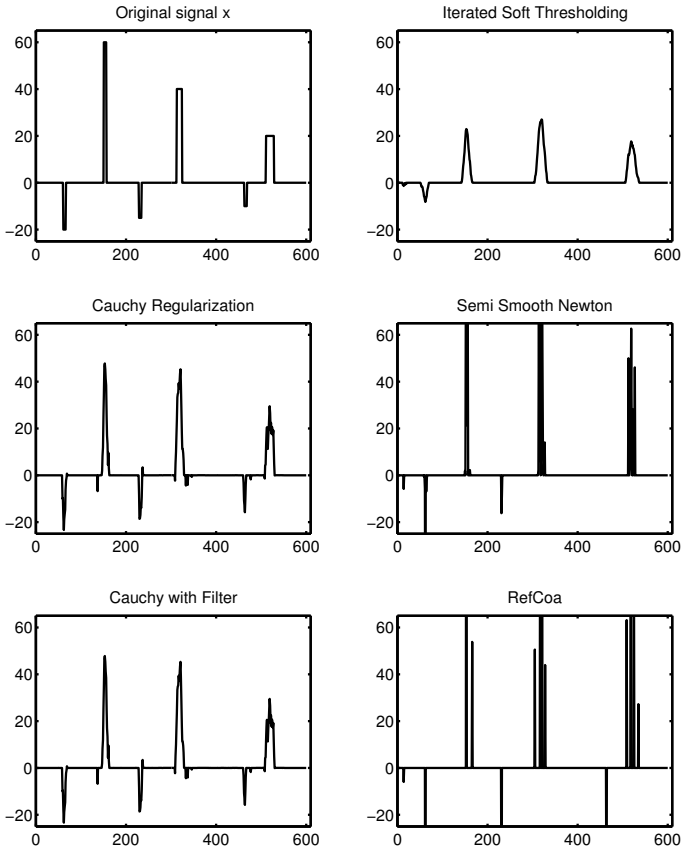


Figure 29: Different minimizers for the inverse integration example. In the first row on the left the exact signal is plotted, which consists of 6 small non-zero plateaus. The upper right plot depicts the minimizer of the iterative thresholding algorithm. In the second row again the minimizer of the Cauchy regularization and the Semi Smooth Newton algorithm are plotted. In the third row the filtered Cauchy and the Refinement and Coarsening approximates are depicted.

Table 5: The errors in ℓ_1 norm for the Compressed Sensing example. In the second column the computation time in seconds and in the third column the number of non-zero coefficients for the different strategies are given.

	$\|x^\dagger - x_\delta^\alpha\|_{\ell_1}$	t in s	nnz
Iterative Soft Thresholding	7.2670	23.0503	223
Tikhonov with Cauchy	3.4499	16.7167	2048
Semi Smooth Newton	48.0093	0.1699	96
Refinement and Coarsening	1.3347	8.3980	83
Cauchy with Filter	2.0696		80

this test case were again tuned by hand to gain the best possible results. For the IST, the thresholding parameter $\bar{\gamma} = 10^{-2}$ was used. The parameters of the Cauchy-Tikhonov functional were set to $\alpha = 10^{-3}$ and $\omega = 6$. For the SSN, we used $\tilde{\gamma} = 10^5$ and $\tilde{\omega} = 3.5 \cdot 10^{-3}$. The Refinement and Coarsening strategy stopped, if the maximum refining indicator was smaller than $\bar{r} = 0.001$.

Again we compared the four different reconstructions. This time the result of the Refinement and Coarsening algorithm poorly approximates the exact function. This can also be seen by the large relative error given in Table 6.

If one compares the plot of Cauchy minimizer with the minimizer of the SSN algorithm visually, they both give a good approximation of the exact solution. However, the minimizer achieved with the Cauchy-Tikhonov functional yields only half of the relative error of the SSN minimizer.

Comparing the time the algorithms required, the IST is, as expected, much slower than minimizing the Cauchy Tikhonov functional. Both cannot compete with the SSN or the Refinement and Coarsening approach.

This time, the Cauchy-Tikhonov functional provides the most accurate solution. But we have to admit that the SSN algorithm can be more accurate in other cases, e.g. when using the exact example function proposed by Griesse and Lorenz in [35]. There they used an exact function with only four small non-zero peaks, see Figure 1 in [35]. With this exact solution, the SSN method yields better error numbers than the Cauchy approach, see table 7.

Table 6: Illustration of the relative errors in ℓ_1 norm for the exact function of Figure 29, the computation time in seconds and the number of non-zero coefficients for the inverse integration example.

	$\frac{\|x^\dagger - x_\delta^\alpha\|_{\ell_1}}{\|x^\dagger\|_{\ell_1}}$	t in s	nnz
Iterative Soft Thresholding	0.7326	15.0984	103
Tikhonov with Cauchy	0.3457	5.2523	600
Semi Smooth Newton	0.6960	0.1487	34
Refinement and Coarsening	1.6955	0.2700	14
Cauchy with Filter	0.3446		223

Table 7: Illustration of the relative errors in ℓ_1 norm for the exact function from [35], the computation time in seconds and the number of non-zero coefficients for the inverse integration example.

	$\frac{\|x^\dagger - x_\delta^\alpha\|_{\ell_1}}{\|x^\dagger\|_{\ell_1}}$	t in s	nnz
Tikhonov with Cauchy	0.4378	5.1623	600
Semi Smooth Newton	0.2202	0.0837	27
Cauchy with Filter	0.4346		84

7.4.3 *Inverse Source Problem*

The third test case brings us back to Example 2 of Chapter 1. We use Cauchy regularization for identifying the source term q in the elliptic PDE

$$\begin{aligned} -\Delta u &= q && \text{in } \Omega \\ u &= 0 && \text{on } \partial\Omega \end{aligned} \tag{7.11}$$

on the unit square $\Omega = [0,1] \times [0,1] \subset \mathbb{R}^2$.

As before, we solved the underlying forward problem by using the MATLAB FEM routine, and minimized the Cauchy-Tikhonov functional in L^2 using the semi smooth Newton method implement in *fminunc*.

We modeled the exact parameter $q^\dagger$ by a Gaussian density function

$$q^\dagger(x,y) = \frac{1}{2\pi\|\Sigma\|} e^{\frac{(x-\mu)^T\Sigma^{-1}(y-\mu)}{2}},$$

with $\Sigma = 0.005\mathrm{Id}$, and $\mu = (\frac{1}{2}, \frac{1}{2})$. The value of each triangle of the FEM grid is assigned according to the midpoint of the triangle, thus resulting in a piecewise linear parameter. See Figure 30 for a plot of the exact parameter and the corresponding solution of (1.3), respectively the data of (1.4).

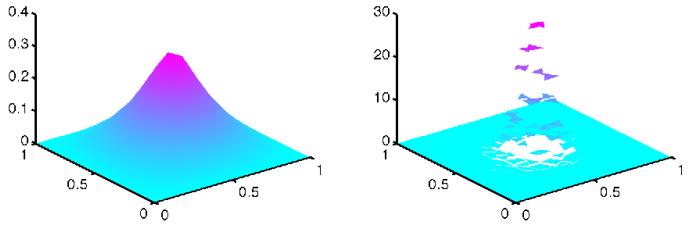


Figure 30: The exact data $u^\dagger$ (left picture) of (1.4) and the corresponding exact solution $q^\dagger$ (right picture).

To illustrate the convergence behavior of the Cauchy-Tikhonov functional, we solved the inverse problem (1.4) according to different noise levels and calculated the relative error of the produced minimizers, see Table 8. There one can see that the relative errors decreases with decreasing noise levels. For the numerical implementation we fixed $\omega = 16$ and $\alpha = 10^{-7}$. In Figure 31, we depicted the identified parameter for 1% noise added to the data.

7.5 SUMMARY

In this chapter we used the connection between Bayesian inversion and Tikhonov regularization to come back to the main

Noiselevel	$\frac{\|q^\dagger - q_\alpha^\delta\|_{L^2}}{\|q^\dagger\|_{L^2}}$
5 %	0.49535
4 %	0.41966
2 %	0.38505
1 %	0.36253
0.5 %	0.36179

Table 8: Relative errors of the calculated parameter for the inverse source problem. In the left row, the noise level δ in percent is specified.

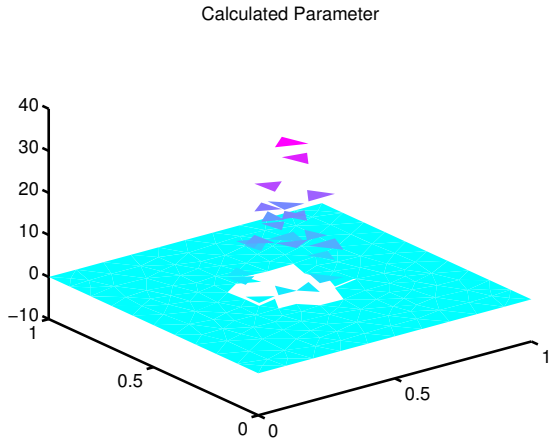


Figure 31: Plot of the identified parameter q_α^δ for 1% Gaussian noise.

topic of this work, the solution of sparse inverse problems. The proposed Cauchy functional is a promising alternative to other sparsity promoting functionals as it is differentiable and can therefore be used in gradient based minimization algorithms, e. g., Quasi-Newton methods.

We have seen that Tikhonov regularization with Cauchy penalty term is a proper regularization method and furthermore that it is possible to derive a convergence rate $\mathcal{O}(\delta)$ under a variational inequality. As the Cauchy term is non convex these rates hold only in a generalized Bregman distance and not in a given norm.

Additionally, the numerical test cases show that the Cauchy regularization term can be used to calculate accurate results in a fast way. In particular, the Cauchy-Tikhonov minimization leads to accurate results in all three test cases, whereas the other approaches produce results with varying accuracy, for the different test cases.

8 CONCLUSION AND OUTLOOK

The main topic of the thesis at hand was the solution of sparse inverse problems. Throughout the chapters of this work, we provided an overview over different methods to solve sparse inverse problems and established two new methods to identify sparse solutions of ill-posed inverse problems. Besides the projection and the regularization method discussed in Chapter 4 and Chapter 7, two other related topics were studied, namely regularization by discretization in Chapter 5 and the Bayesian inversion theory in Chapter 6.

We started our analysis motivated by the fact that standard regularization methods, as introduced in Chapter 2, are not capable of solving sparse inverse problems. As we saw, e. g., in Example 3, it is a challenging task to recover zeros or sharp edges of an exact signal. Tikhonov regularization with a standard quadratic regularization term for example is not able to recover such edges or zeros as it will lead to smooth and fully populated solutions.

Therefore, we introduced a generalization of an adaptive method proposed by Chavent et al.. This generalization leads to a projection method for sparse inverse problems which we called the Refinement and Coarsening Algorithm. The key idea of this algorithm is to solve the considered inverse problem on an adaptively chosen subspace. The algorithm does not change the parts of the solution outside this chosen subspace, but sets them to zero. We proposed two indicators for choosing the adaptive subspace. One indicating in which direction to expand the subspace and one indicating which dimensions of the subspace to abandon. Thus the gained solution is element of a subspace which is as large as needed and as small as possible. Since the underlying minimization problem is only solved on this, in general, smaller subspace, the minimization can be carried out in a very fast way. Additionally, we showed that the projected minimization problem is well-posed, i. e., there exists a minimizer, and that, if the algorithm stops after finitely many steps, the generated approximation solves the inverse problem in a least squares sense.

Starting from the idea of adaptively chosen subspaces in the Refinement and Coarsening Algorithm we took a closer look at regularization by discretization in preimage space in Chapter 5. Regularization by discretization in preimage space is known to converge only if additional assumptions on the solution or on the forward operator are valid. Here we showed convergence of the approximation sequence under a strong source condition

and under the condition that the subspace is chosen according to the discrepancy principle, cf. equation (5.2). With this assumptions we were able to show well-posedness of the regularization approach for linear as well as nonlinear forward operators. The theoretical results are illustrated by a nonlinear inverse problem similar to Example 2, which we solved numerically on different discretization levels. As postulated in the theoretical part of Chapter 5, the best possible result was achieved on the subspace, which comes closest to fulfill (5.2).

In the third part of this work we established a new regularization functional for Tikhonov regularization which promotes sparsity. This functional was generated by using the close connection between stochastic and deterministic regularization theory.

The connection between stochastic and deterministic theory was developed in Theorem 6.1 in Chapter 6. The theorem states equivalence between Bayes regularization with Gaussian distributions and Tikhonov regularization with a quadratic regularization term. Before providing this theorem, we had a closer look at the stochastic regularization approach. There, one tries to maximize the posterior distribution instead of minimizing a cost functional. The corresponding maximizer then serves as a solution, instead of the minimum least squares approximation in the deterministic case. In Chapter 6 we stated a general convergence conjecture for a posterior distribution generated by a general pair of noise and prior distribution. Additionally, we proved a slight generalization of the existence theorem for the MAP estimate, which also holds true for an $\mathcal{R}$-minimizing solution.

In the last chapter, we finally introduced the Cauchy regularization functional which occurs when considering the negative logarithm of the Cauchy distribution. The Cauchy distribution is a probability distribution used to infer sparse solutions in the Bayesian regularization theory. We examined the functional and showed well-posedness of the corresponding Tikhonov functional. For this purpose, we used the recent results for regularization in Banach spaces introduced in Chapter 2. Besides the well-posedness, we showed differentiability of the new regularization functional. One can use fast gradient based optimization methods for the minimization of the Cauchy-Tikhonov functional. However on the theoretical side, the Cauchy functional suffers one big drawback: it is not convex. Due to its non-convexity, we could not apply the results on convergence rates for general Tikhonov regularization, established in [48] and mentioned in Chapter 2. To show a convergence rate at least for a given variational inequality, we had to introduce the notion of W-convexity, a less strict version of convexity, which is fulfilled by the Cauchy term.

The Cauchy functional concluded our analysis of the solution of sparse ill-posed inverse problems. We gave two possible answers to the question raised at the beginning – how to solve sparse inverse problems – by proposing two new techniques, a projection and a variational regularization method for sparse inverse problems.

8.1 OUTLOOK

This thesis adresses a wide spectrum of topics related to sparse inverse problems. However, there are of course questions we could not answer up to now, as well as new questions that arose throughout the development of the above given results.

After showing well-posedness for the Cauchy-Tikhonov functional, the question arises, whether there are other probability density functions which can be carried over to deterministic regularization functionals (for sparsity or other purposes). Next the proof of well-posedness for the generated functionals, or better for general regularization functionals is a crucial task. There are recent works focusing on general non-convex regularization functionals, i. e., [12, 32, 31]. But as we have seen, the Cauchy regularization term does not fit completely in neither of these approaches. Thus, a theory for general non-convex regularization terms is still missing.

Another open issue is the convergence proof for arbitrary pairs of probability density functions in the Bayesian inversion theory. As we have seen the close connection can only be established in case of a pair of Gaussian distribution. The proof of convergence for a general posterior density function can be a very valuable result for both the deterministic and the stochastic approach. However, due to the above given reasons it is very challenging to prove such a general result, see Chapter 6.

Concerning the first part of this work, a theoretical investigation of the sophisticated stopping rule used in the numerical test cases in Chapter 5, cf. equation (5.40), could be of interest. As we have seen in our numerical results these inexact solution of the Euler equation leads to good results.

As indicated in Chapter 4, the Refinement and Coarsening algorithm can be used for the minimization of any differentiable misfit functional. It would hence be an interesting approach to combine a specialized Tikhonov functional, e. g., Table 1, with the refining and coarsening strategy to get a fast and efficient algorithm for solving specific inverse problems.

BIBLIOGRAPHY

[1] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[2] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits (Chapman & Hall/CRC Mathematical & Computational Biology)*. Chapman and Hall/CRC, 1 edition, 2006.

[3] U. Amato and W. Hughes. Maximum entropy regularization of fredholm integral equations of the first kind. *Inverse Problems*, 7(6):793+, 1991.

[4] A.B. Bakushinsky and M. Yu Kokurin. *Iterative Methods for Approximate Solution of Inverse Problems*. Springer, Dordrecht, 2004.

[5] W. Bangerth. *Adaptive finite element methods for the identification of distributed parameters in partial differential equations*. PhD thesis, University of Heidelberg, Germany, 2002.

[6] A. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[8] F. Bauer and M. A. Lukas. Comparing parameter choice methods for regularization of ill-posed problems. *Mathematics and Computers in Simulation*, 81(9):1795–1841, 2011.

[9] H. Ben Ameur and B. Kaltenbacher. Regularization of parameter estimation by adaptive discretization using refinement and coarsening indicators. *Journal of Inverse and Ill-Posed Problems*, 10(6), 2002.

[10] H. Ben Ameur, G. Chavent, and J. Jaffre. Refinement and coarsening indicators for adaptive parametrization: application to the estimation of hydraulic transmissivities. *Inverse Problems*, 18(3):775–794, 2002.

[11] K. Bredies and D. A. Lorenz. Iterated Hard Shrinkage for Minimization Problems with Sparsity Constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008.

[12] K. Bredies and D. A. Lorenz. Regularization with non-convex separable constraints. *Inverse Problems*, 25(8):085011+, 2009.

[13] R. Bringhurst. *The Elements of Typographic Style*. Version 2.5. Hartley & Marks, Publishers, Point Roberts, WA, USA, 2002.

[14] M. Burger and A. Hofinger. Regularized greedy algorithms for network training with data noise. *Computing*, 74:1–22, 2005.

[15] E. Candes and J. Romberg. l1-magic: Recovery of sparse signals via convex programming. Technical report, California Institute of Technology, 2005. URL `http://www-stat.stanford.edu/~candes/l1magic/`.

[16] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004.

[17] G. Chavent and R. Bissell. Indicator for the refinement of parametrization. In *Proceedings of the International Symposium in Inverse Problems in Engineering Mechanics, Nagano, Japan*, 1998.

[18] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory (Graduate Texts in Mathematics)*. Springer, 1 edition, 1997.

[19] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[20] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13(1):57–98, 1997.

[21] L. Denis, D. A. Lorenz, and D. Trede. Greedy solution of ill-posed problems: error bounds and exact inversion. *Inverse Problems*, 25(11):115017+, 2009.

[22] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 2002.

[23] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[24] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2nd edition, 2002.

[25] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems. *Inverse Problems*, 5(4):523–540, 1989.

[26] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996.

[27] H. W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, 25(12):123014+, 2009.

[28] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):586–597, 2007.

[29] C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, Berlin Heidelberg New York, 1999.

[30] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Texts in Statistical Science. Chapman & Hall, CRC, 2 edition, 2004.

[31] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Problems*, 26(11):115014+, 2010.

[32] M. Grasmair. Non-convex sparse regularisation. *Journal of Mathematical Analysis and Applications*, 365(1):19–28, 2010.

[33] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with lq penalty term. *Inverse Problems*, 24(5):055020+, 2008.

[34] M. Grasmair, O. Scherzer, and M. Haltmeier. Necessary and sufficient conditions for linear convergence of ℓ_1-regularization. *Communications on Pure and Applied Mathematics*, 64(2):161–182, 2011.

[35] R. Griesse and D. A. Lorenz. A semismooth Newton method for Tikhonov functionals with sparsity constraints. *Inverse Problems*, 24(3), 2008.

[36] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman, 1985.

[37] C.W. Groetsch. *Inverse Problems in Mathematical Sciences*. Vieweg, Braunschweig, 1993.

[38] C.W. Groetsch and A. Neubauer. Convergence of a general projection method for an operator equation of the first kind. *Houston Journal of Mathematics*, 14:201–208, 1988.

[39] J. Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13: 49–52, 1902.

[40] U. Hämarik, E. Avi, and A. Ganina. On the solution of ill-posed problems by projection methods with a posteriori choice of the discretization level. *Mathematical Modelling and Analysis*, 7:241–252, 2002.

[41] M. Hanke and O. Scherzer. Inverse Problems Light: Numerical Differentiation. *The American Mathematical Monthly*, 108 (6), 2001.

[42] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of the landweber iteration for nonlinear ill-posed problems. *Numerische Mathematik*, 72:21–37, 1995.

[43] A. Hofinger. The metrics of Prokhorov and Ky Fan for assessing uncertainty in inverse problems. Technical report, 2006.

[44] A. Hofinger and H. K. Pikkarainen. Convergence rate for the Bayesian approach to linear inverse problems. *Inverse Problems*, 23(6):2469–2484, 2007.

[45] A. Hofinger and H. K. Pikkarainen. Convergence Rates for Linear Inverse Problems in the Presence of an Additive Normal Noise. *Stochastic Analysis and Applications*, 27(2): 240–257, 2009.

[46] B. Hofmann. *Regularization for applied inverse and ill-posed problems*. Teubner Texte zur Mathematik. Teubner, Leipzig, 1986.

[47] B. Hofmann. *Mathematik inverser Probleme.* Teubner Verlag, 1999.

[48] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010, 2007.

[49] B. Hofmann, P. Mathé, and S.V. Pereverzev. Regularization by projection: Approximation theoretic aspects and distance functions. *Journal of Inverse and Ill-Posed Problems*, 15:527–545, 2007.

[50] T. Hohage and F. Werner. Iteratively regularized newton methods with general data misfit functionals and applications to poisson data. May 2011. URL http://arxiv.org/abs/1105.2690.

[51] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer, 2005.

[52] B. Kaltenbacher. Regularization by projection with a posteriori discretization level choice for linear and nonlinear ill-posed problems. *Inverse Problems*, 16(5):1523–1539, 2000.

[53] B. Kaltenbacher. On the regularizing properties of a full multigrid method for ill-posed problems. *Inverse Problems*, 17:767–788, 2001.

[54] B. Kaltenbacher. V-cycle convergence of some multigrid methods for ill-posed problems. *Mathematics of Computation*, 72:1711–1730, 2003.

[55] B. Kaltenbacher. Towards global convergence for strongly nonlinear ill-posed problems via a regularizing multilevel approach. *Numerical Functional Analysis and Optimization*, 27: 637 – 665, 2006.

[56] B. Kaltenbacher. Convergence rates of a multilevel method for the regularization of nonlinear ill-posed problems. *Journal of Integral Equations and Applications*, 20(2):201–228, 2008.

[57] B. Kaltenbacher and J. Offtermatt. A convergence analysis of regularization by discretization in preimage space. Technical report, University of Stuttgart, 2010.

[58] B. Kaltenbacher and J. Offtermatt. A Refinement and Coarsening Indicator Algorithm for Finding Sparse Solutions of Inverse Problems. *Inverse Problems and Imaging*, 5(2):391–406, 2011.

[59] B. Kaltenbacher and J. Schicho. A multi-grid method with a priori and a posteriori level choice for the regularization of nonlinear ill-posed problems. *Numerische Mathematik*, 93: 77–107, 2002.

[60] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Springer, New York, 1996.

[61] R. Kress. *Linear Integral Equations*. Springer, Heidelberg, 1989, 2nd ed. 1999.

[62] D. A. Lorenz. On the role of sparsity in inverse problems. *Journal of Inverse and Ill-posed Problems*, 17(1):61–68, 2009.

[63] A. K. Louis. *Inverse und schlecht gestellte Probleme*. Teubner, Stuttgart, 1989.

[64] G.R. Luecke and K.R. Hickey. Convergence of approximate solutions of an operator equation. *Houston Journal of Mathematics*, 11(3):345–354, 1985.

[65] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[66] P. Mathé and N. Schöne. Regularization by projection in variable Hilbert scales. *Applicable Analysis*, 87:201–219, 2008.

[67] V.A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer, 1984.

[68] V.A. Morozov. *Regularization Methods for Ill-Posed Problems*. CRC Press, Boca Raton, 1993.

[69] F. Natterer. Regularisierung schecht gestellter Probleme durch Projektionsverfahren. *Numerische Mathematik*, 28:329–341, 1977.

[70] A. Neubauer and H. K. Pikkarainen. Convergence results for the Bayesian inversion theory. *Journal of Inverse and Ill-posed Problems*, 16(6):601–613, 2008.

[71] R. Peeters and R. Westra. On the identification of sparse gene regulatory networks. In *Proc 16th Intern Symp on Mathematical Theory of Networks*, 2004.

[72] S.V. Pereverzev and S. Prössdorf. On the characterization of self-regularization properties of a fully discrete projection method for symm's integral equation. *Journal of Integral Equations and Applications*, 12(2):113–130, 2000.

[73] C. Pöschl. *Tikhonov Regularization with General Residual Term*. PhD thesis, Leopold Franzens Universität Innsbruck, 2008.

[74] R. Ramlau. Regularization Properties of Tikhonov Regularization with Sparsity Constraints. *Electronic Transactions on Numerical Analysis*, 30, 2008.

[75] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability anaylsis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15): 1923–1929, 2009.

[76] E. Resmerita and R. S. Anderssen. Joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems. *Mathematical Methods in the Applied Sciences*, 30(13):1527–1544, 2007.

[77] E. Resmerita and O. Scherzer. Error estimates for non-quadratic regularization and the relation to enhancement. *Inverse Problems*, 22(3):801+, 2006.

[78] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[79] O. Scherzer. Convergence criteria of iterative methods based on Landweber iteration for solving nonlinear problems. *Journal of Mathematical Analysis and Applications*, 194(3):911–933, 1995.

[80] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*. Springer, 2008.

[81] T.I. Seidman. Nonconvergence results for the application of least squares estimation to ill-posed problems. *Journal of Optimization Theory and Applications*, 30:535–547, 1980.

[82] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11): 2498–2504, 2003.

[83] I. Singer. *Abstract Convex Analysis*. Wiley-Interscience and Canadian Mathematics Series of Monographs and Texts, 1 edition, 1997.

[84] F. Steinke, M. Seeger, and K. Tsuda. Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. *BMC Systems Biology*, 1(51), 2007.

[85] G. Teschke and R. Ramlau. An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector-valued regimes and an application to color image inpainting. *Inverse Problems*, 23, 2007.

[86] A. N. Tikhonov and V. A. Arsenin. *Methods for Solving Ill-Posed Problems*. Nauka, Moscow, 1979.

[87] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed Problems*. John Wiley & Sons, Washington, D.C., 1977.

[88] J. A. Tropp. Greed is Good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory*, 50 (10):2231–2242, 2004.

[89] G. Vainikko and U. Hämarik. Projection methods and self-regularization in ill-posed problems. *Soviet Mathematics*, 29: 1–20, 1985. in Russian.

[90] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[91] P. Weber, J. Hasenauer, F. Allgöwer, and N. Radde. Parameter estimation and identifiability of biological networks using relative data. acccepted for Proceedings of the IFAC World Congress, 2011, 2011.

[92] M. K. Yeung, J. Tegnér, and J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6163–6168, 2002.

[93] C. A. Zarzer. On Tikhonov regularization with non-convex sparsity constraints. *Inverse Problems*, 25(2):025006+, 2009.

NOMENCLATURE

α	regularization parameter
A	linear forward operator
$\mathcal{D}$	domain of F, A or $\mathcal{D}(F) \cap \mathcal{D}(\mathcal{R})$
δ	noise bound
δ_x	point measure concentrated at x
$D_{\mathcal{R}}$	Bregman distance with respect to $\mathcal{R}$
D_W^w	W-Bregman distance with respect to w
E	random variable depicting noise
F	nonlinear forward operator
γ	prior variance
$\mathcal{J}$	general cost functional
$\mathcal{J}_\alpha$	general Tikhonov functional
KL	Kullback-Leibler functional
$\mathcal{L}$	Lagrange function
l_y	negative logarithm of the likelihood distribution
$\mathcal{M}_{\alpha,y}$	level sets of $\mathcal{J}$ with respect to α and y
$N(\mu, \sigma)$	Gaussian distribution with mean μ and variance σ
p	probability density function
p_E	noise distribution
(Φ_i)	basis of $\mathcal{X}$
p_L	likelihood distribution
P_n	projection onto the n-th subspace of $\mathcal{X}$
p_P	prior distribution
Q_n	projection onto the n-th subspace of $\mathcal{Y}$
$\mathcal{R}$	general regularization functional
$\mathcal{R}_C$	Cauchy regularization functional

ρ_K	Ky Fan metric
ρ_P	Prokhorov metric
$\mathcal{R}^*$	(generalized) Fenchel conjugate of $\mathcal{R}$
$\mathcal{R}^{**}$	double Fenchel conjugate of $\mathcal{R}$
$\mathcal{S}$	general data fitting functional
σ	noise variance
S_ω	thresholding operator
$\tau_\mathcal{X}$	topology on $\mathcal{X}$
$\tau_\mathcal{Y}$	topology on $\mathcal{Y}$
W_2	family of all negative semi-definite, continuous quadratic functions
$\mathcal{X}$	preimage space
X	random variable depicting the solution
x_i	i-th coefficient of x
$x^\dagger$	exact solution
$\hat{x}^{\text{MAP}}$	maximum a posteriori estimate (MAP)
$\hat{x}^{\text{ML}}$	maximum likelihood estimate (MLE)
$\hat{x}^{\text{MPMLE}}$	maximum prior maximum likelihood estimate (MPMLE)
$\mathcal{X}_n$	n-th subspace of $\mathcal{X}$
$\mathcal{Y}$	image space
Y	random variable depicting data
y	exact data
y^δ	noisy data
$\mathcal{Y}_n$	n-th subspace of $\mathcal{Y}$

CURRICULUM VITÆ

PERSÖNLICHE DATEN

Jonas Offtermatt

Geboren am 31.12.1982
Europäer, Deutsche Staatsbürgerschaft

verheiratet, eine Tochter

Unterer Metzgerbach 14
73728 Esslingen

jonas.offtermatt@gmx.de

AUSBILDUNG

seit Mai 2009 Promotion im Excellence Cluster Simulation Technology an der Universität Stuttgart, Projekt: *Efficient methods for Parameter Identification in Differential Equation Models*

2008 – 2009 wissenschaftliche Hilfskraft am Institut für Stochastik und Anwendungen der Universität Stuttgart

2003 – 2008 Studium der Mathematik und Philosophie/Ethik auf Lehramt an der Universität Stuttgart

2002 – 2003 Zivildienst im Kreiskrankenhaus Waiblingen

1993 – 2002 Salier Gymnasium in Waiblingen
20.06.2002 Abschluss der Schule mit dem Abitur

1989-1993 Salier Grund- und Hauptschule in Waiblingen

AUSLANDSERFAHRUNGEN

2006 Praxissemester an der deutschen Schule in Istanbul, Türkei

2011 Forschungsaufenthalt an der Universität Klagenfurt, Österreich

TEILNAHME AN WISSENSCHAFTLICHEN TAGUNGEN

1. SimTech Statusseminar, 2008, Freudenstadt, Deutschland

Summer School, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, 2009, Linz, Österreich

1. SimTech-PhD-Weekend, 2009, Hirschegg, Deutschland

2. SimTech Statusseminar, 2009, Bad Herrenalb, Deutschland

Interdisciplinary Workshop, *Sparsity and Modern Mathematical Methods for High Dimensional Data*, 2010, Brüssel, Belgien

International Conference on Inverse Problems, 2010, Wuhan, China

5th IP:M&S, *Inverse Problems: Modeling and Simulation*, 2010, Antalya, Türkei

Summer school, *Computational solution of inverse problems*, 2010, Helsinki, Finnland

2. SimTech-PhD-Weekend, 2010, Hirschegg, Deutschland

3. SimTech Statusseminar, 2010, Bad Boll, Deutschland

Workshop on Numerical Methods for Optimal Control and Inverse Problems, 2011, München, Deutschland

Introductory Workshop on Inverse Problems, 2011, Cambridge, England

4. SimTech Statusseminar, 2011, Bad Boll, Deutschland

Stuttgart, May, 2012

DECLARATION

Hiermit erkläre ich, dass ich die vorliegende Arbeit „A Projection and a Variational Regularization Method for Sparse Inverse Problems“, basierend auf den auf Seite xii angegebenen Arbeiten, selbstständig verfasst und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die Arbeit wurde an keiner anderen Universität als Dissertation eingereicht.

Stuttgart, May 2012

Jonas Offtermatt

COLOPHON

This thesis was typeset with LaTeX 2ε using Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URW Palladio L* and *FPL* were used). The listings are typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera".

(Type 1 PostScript fonts were made available by Malte Rosenau and Ulrich Dirr.)

The typographic style was inspired by Bringhurst's genius as presented in *The Elements of Typographic Style* [13]. It was made available for LaTeX via CTAN as "`classicthesis`" by Andreas Miede.

Final Version as of May 5, 2012 at 10:59.