Alma Mater Studiorum
Università di Bologna

Dottorato di Ricerca in
MATEMATICA

Ciclo XXV

Settore Concorsuale di afferenza: 01/A5
Settore Scientifico disciplinare: MAT/08

Iterative regularization methods for ill-posed problems

Tesi di Dottorato presentata da: Ivan Tomba

Coordinatore Dottorato:
Chiar.mo Prof.
Alberto Parmeggiani

Relatore:
Chiar.ma Prof.ssa
Elena Loli Piccolomini

Esame Finale anno 2013

Contents

Introduction

Inverse and ill-posed problems are nowadays a very important field of research in applied mathematics and numerical analysis. The main reason for this large interest is the wide number of applications, ranging from medical imaging, via material testing, seismology, inverse scattering and financial mathematics, to weather forecasting, just to cite some of the most famous.
Typically, in these problems some fundamental information is not available and the solution does not depend continuously on the data. As a consequence of this lack of stability, even very small errors in the data can cause very large errors in the results. Thus, the problems have to be regularized by inserting some additional information in the data to obtain reasonable approximations of the sought for solution. On the other hand, it is important to keep the computational cost of the corresponding algorithms as low as possible, since in practical applications the total amount of data to be processed is usually very high.

The main topic of this Ph.D thesis is the regularization of ill-posed problems by means of iterative regularization techniques. The principal advantage of the iterative methods is that the regularized solutions are obtained by arresting the methods at an early stage, which often allows to spare time in the computations. On the other side, the main difficulty in their use is the choice of the stopping index of the iteration: an early stopping produces an over-regularized solution, whereas a late stopping computes a noisy solution.
In particular, we shall focus on the conjugate gradient type methods for regularizing linear ill-posed problems from the classical Hilbert space set-

ting point of view, and on a new inner-outer Newton-Iteratively Regularized Landweber method for solving nonlinear ill-posed problems in the Banach space framework.

Regarding the conjugate gradient type methods, we propose three new automatic[1] stopping rules for the Conjugate Gradient method applied to the Normal Equation in the discrete setting, based on the regularizing properties of the method in the continuous setting. These stopping rules are tested in a series of numerical simulations, including some problems of tomographic images reconstruction.

Regarding the Newton-Iteratively Regularized Landweber method, we define both the iteration and the stopping rules showing convergence and convergence rates results.

In detail, the thesis is constituted by six chapters.

- In Chapter 1 we recall the basic notions of the regularization theory in the Hilbert space framework. Revisiting the regularization theory, we mainly follow the famous book of Engl, Hanke and Neubauer [17]. Some examples are added, some others corrected and some proofs completed.
- Chapter 2 is dedicated to the definition of the conjugate gradient type methods and the analysis of their regularizing properties. A comparison of these methods is made by means of numerical simulations.
- In Chapter 3 we motivate, define and analyze the new stopping rules for the Conjugate Gradient applied to the Normal Equation. The stopping rules are tested in many different examples, including some image deblurring test problems.
- In Chapter 4 we consider some applications in tomographic problems. Some theoretical properties of the Radon Transform are studied and then used in the numerical tests to implement the stopping rules defined in Chapter 3.

[1] i.e. that can be defined precisely by a software

- Chapter 5 is a survey of the regularization theory in the Banach space framework. The main advantages of working in a Banach space setting instead of a Hilbert space setting are explained in the introduction. Then, the fundamental tools and results of this framework are summarized following [82]. The regularizing properties of some important iterative regularization methods in the Banach space framework, such as the Landweber and Iteratively Regularized Landweber methods and the Iteratively Regularized Gauss-Newton method, are described in the last section.

- The main results about the new inner-outer Newton-Iteratively Regularized Landweber iteration are presented in the conclusive part of the thesis, in Chapter 6.

Chapter 1

Regularization of ill-posed problems in Hilbert spaces

The fundamental background of this thesis is the regularization theory for (linear) ill-posed problems in the Hilbert space setting. In this introductory chapter we are going to revisit and summarize the basic concepts of this theory, which is nowadays well-established. To this end, we will mainly follow the famous book of Engl, Hanke and Neubauer [17].

Starting from some very famous examples of inverse problems (differentiation and integral equations of the first kind) we will review the notions of regularization method, stopping rule and order optimality. Then we will consider a class of finite dimensional problems arising from the discretization of ill-posed problems, the so called *discrete ill-posed problems*. Finally, in the last two sections of the chapter we will recall the basic properties of the Tikhonov and the Landweber methods.

Apart from [17], general references for this chapter are [20], [21], [22], [61], [62], [90] and [91] and, concerning the part about finite dimensional problems, [36] and the references therein.

1.1 Fundamental notations

In this section, we fix some important notations that will we used throughout the thesis.

First of all, we shall denote by $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$, the sets of the integer, real and complex numbers respectively. In $\mathbb{C}$, the imaginary unit will be denoted by the symbol $\imath$. The set of strictly positive integers will be denoted by $\mathbb{N}$ or $\mathbb{Z}^+$, the set of positive real numbers by $\mathbb{R}^+$.

If not stated explicitly, we shall denote by $\langle \cdot, \cdot \rangle$ and by $\| \cdot \|$ the standard euclidean scalar product and euclidean norm on the cartesian product $\mathbb{R}^D$, $D \in \mathbb{N}$ respectively. Moreover, $\mathbb{S}^{D-1} := \{\mathbf{x} \in \mathbb{R}^D \mid \|\mathbf{x}\| = 1\}$ will be the unit sphere in $\mathbb{R}^D$.

For i and $j \in \mathbb{N}$, we will denote by $\mathbb{M}_{i,j}(\mathbb{R})$ (respectively, $\mathbb{M}_{i,j}(\mathbb{C})$) the space of all matrices with i rows and j columns with entries in $\mathbb{R}$ (respectively, $\mathbb{C}$) and by $\mathbb{GL}_j(\mathbb{R})$ (respectively, $\mathbb{GL}_j(\mathbb{C})$) the space of all square invertible matrices on $\mathbb{R}$ ($\mathbb{C}$).

For an appropriate subset $\Omega \subseteq \mathbb{R}^D$, $D \in \mathbb{N}$, $\mathcal{C}(\Omega)$ will be the set of continuous functions on Ω. Analogously, $\mathcal{C}^k(\Omega)$ will denote the set of differentiable functions with k continuous derivatives on Ω, $k = 1, 2, ..., \infty$, and $\mathcal{C}_0^k(\Omega)$ will be the corresponding sets of functions with compact support. For $p \in [1, \infty]$, we will write $\mathcal{L}^p(\Omega)$ for the Lebesgue spaces with index p on Ω, and $\mathcal{W}^{p,k}(\Omega)$ for the corresponding Sobolev spaces, with the special cases $\mathcal{H}^k(\Omega) = \mathcal{W}^{2,k}(\Omega)$. The space of rapidly decreasing functions on $\mathbb{R}^D$ will be denoted by $\mathcal{S}(\mathbb{R}^D)$.

1.2 Differentiation as an inverse problem

In this section we present a fundamental example for the study of ill-posed problems: the computation of the derivative of a given differentiable function. Let f be any function in $\mathcal{C}^1([0,1])$. For every $\delta \in (0,1)$ and every $n \in \mathbb{N}$ define

$$f_n^\delta(t) := f(t) + \delta \sin \frac{nt}{\delta}, \ t \in [0,1]. \tag{1.1}$$

Then

$$\frac{d}{dt}f_n^\delta(t) = \frac{d}{dt}f(t) + n\cos\frac{nt}{\delta},\ t \in [0,1],$$

hence, in the uniform norm,

$$\|f - f_n^\delta\|_{\mathcal{C}([0,1])} = \delta$$

and

$$\|\frac{d}{dt}f - \frac{d}{dt}f_n^\delta\|_{\mathcal{C}([0,1])} = n.$$

Thus, if we consider f and f_n^δ the exact and perturbed data, respectively, of the problem *compute the derivative* $\frac{df}{dt}$ *of the data* f, for an arbitrary small perturbation of the data δ we can obtain an arbitrary large error n in the result. Equivalently, the operator

$$\frac{d}{dt} : (\mathcal{C}^1([0,1]), \|\ \|_{\mathcal{C}([0,1])}) \longrightarrow (\mathcal{C}([0,1]), \|\ \|_{\mathcal{C}([0,1])})$$

is not continuous. Of course, it is possible to enforce continuity by measuring the data in the $\mathcal{C}^1$-norm, but this would be like cheating, since to calculate the error in the data one should calculate the derivative, namely the result. It is important to notice that $\frac{df}{dt}$ solves the integral equation

$$K[x](s) := \int_0^s x(t)dt = f(s) - f(0), \tag{1.2}$$

i.e. the result can be obtained by inverting the operator K. More precisely, we have:

Proposition 1.2.1. *The linear operator* $K : \mathcal{C}([0,1]) \longrightarrow \mathcal{C}([0,1])$ *defined by* (1.2) *is continuous, injective and surjective onto the linear subspace of* $\mathcal{C}([0,1])$ *denoted by* $\mathcal{W} := \{x \in \mathcal{C}^1([0,1]) \mid x(0) = 0\}$. *The inverse of* K

$$\frac{d}{dt} :\ \mathcal{W} \longrightarrow \mathcal{C}([0,1])$$

is unbounded.

If K *is restricted to*

$$\mathcal{S}_\gamma := \{x \in \mathcal{C}^1([0,1]) \mid \|x\|_{\mathcal{C}([0,1])} + \|\frac{dx}{dt}\|_{\mathcal{C}([0,1])} \leq \gamma,\ \gamma > 0\},$$

then $(K|_{\mathcal{S}_\gamma})^{-1}$ *is bounded.*

Proof. The first part is obvious. For the second part, it is enough to observe that $\|x\|_{\mathcal{C}([0,1])} + \|\frac{dx}{dt}\|_{\mathcal{C}([0,1])} \leq \gamma$ ensures that $\mathcal{S}_\gamma$ is bounded and equicontinuous in $\mathcal{C}([0,1])$, thus, according to the Ascoli-Arzelà Theorem, $\mathcal{S}_\gamma$ is relatively compact. Hence $(K|_{\mathcal{S}_\gamma})^{-1}$ is bounded because it is the inverse of a bijective and continuous operator defined on a relatively compact set. □

The last statement says that we can *restore stability* by assuming a-priori bounds for f' and f''.

Suppose now we want to calculate the derivative of f via central difference quotients with step size σ and let f^δ be its noisy version with

$$\|f^\delta - f\|_{\mathcal{C}([0,1])} \leq \delta. \tag{1.3}$$

If $f \in \mathcal{C}^2[0,1]$, a Taylor expansion gives

$$\frac{f(t+\sigma) - f(t-\sigma)}{2\sigma} = f'(t) + O(\sigma),$$

but if $f \in \mathcal{C}^3[0,1]$ the second derivative can be eliminated, thus

$$\frac{f(t+\sigma) - f(t-\sigma)}{2\sigma} = f'(t) + O(\sigma^2).$$

Remembering that we are dealing with perturbed data

$$\frac{f^\delta(t+\sigma) - f^\delta(t-\sigma)}{2\sigma} \sim \frac{f(t+\sigma) - f(t-\sigma)}{2\sigma} + \frac{\delta}{\sigma},$$

the total error behaves like

$$O(\sigma^\nu) + \frac{\delta}{\sigma}, \tag{1.4}$$

where $\nu = 1, 2$ if $f \in \mathcal{C}^2[0,1]$ or $f \in \mathcal{C}^3[0,1]$ respectively.

A remarkable consequence of this is that for fixed δ, when σ is too small, the total error is large, because of the term $\frac{\delta}{\sigma}$, the propagated data error. Moreover, there exists an optimal discretization parameter $\sigma^\sharp$, which cannot be computed explicitly, since it depends on unavailable information about the exact data, i.e. the smoothness.

However, if $\sigma \sim \delta^\mu$ one can search the power μ that minimizes the total error with respect to δ, obtaining $\mu = \frac{1}{2}$ if $f \in \mathcal{C}^2[0,1]$ and $\mu = \frac{1}{3}$ if $f \in \mathcal{C}^3[0,1]$,

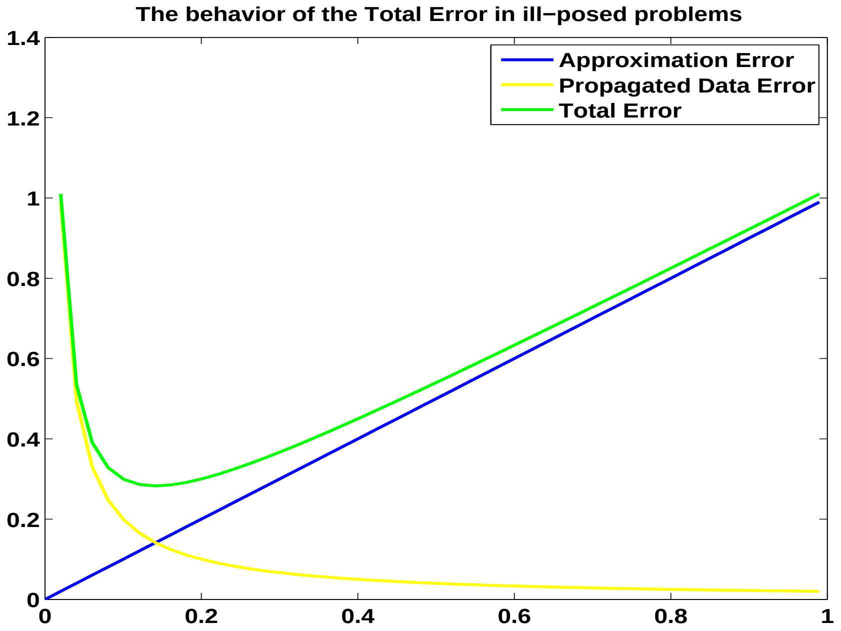


Figure 1.1: The typical behavior of the total error in ill-posed problems

with a resulting total error of the order of $O(\sqrt{\delta})$ and $O(\delta^{\frac{2}{3}})$ respectively. Thus, in the best case, the total error $O(\delta^{\frac{2}{3}})$ tends to 0 slower than the data error δ and it can be shown that this result cannot be improved unless f is a quadratic polynomial: this means that there is an intrinsic loss of information.

Summing up, in this simple example we have seen some important features concerning ill-posed problems:

- amplification of high-frequency errors;
- restoration of stability by a-priori assumptions;
- two error terms of different nature, adding up to a total error as in Figure 1.1;
- appearance of an optimal discretization parameter, depending on a-priori information;

- loss of information even under optimal circumstances.

1.3 Abel integral equations

When dealing with inverse problems, one has often to solve an integral equation. In this section we present an example which can be described mathematically by means of the *Abel Integral Equation.* The name is in honor of the famous Norwegian mathematician N. H. Abel, who studied this problem for the first time.[1]

Let a mass element move on the plane $\mathbb{R}^2_{\mathrm{x}_1,\mathrm{x}_2}$ along a curve $\mathbf{\Gamma}$ from a point $\mathbf{P}_1$ on level $\mathsf{h} > 0$ to a point $\mathbf{P}_0$ on level 0. The only force acting on the mass element is the gravitational force mg.

The direct problem is to determine the time τ in which the element moves from $\mathbf{P}_1$ to $\mathbf{P}_0$ when the curve $\mathbf{\Gamma}$ is given. In the inverse problem, one measures the time $\tau = \tau(\mathsf{h})$ for several values of h and tries to determine the curve $\mathbf{\Gamma}$. Let the curve be parametrized by $\mathrm{x}_1 = \psi(\mathrm{x}_2)$. Then $\mathbf{\Gamma} = \mathbf{\Gamma}(\mathrm{x}_2)$ and

$$\mathbf{\Gamma}(\mathrm{x}_2) = \begin{pmatrix} \psi(\mathrm{x}_2) \\ \mathrm{x}_2 \end{pmatrix}, \quad d\Gamma(\mathrm{x}_2) = \sqrt{1 + \psi'(\mathrm{x}_2)^2}.$$

According to the conservation of energy,

$$\frac{\mathsf{m}}{2} v^2 + \mathsf{mg}\mathrm{x}_2 = \mathsf{mgh},$$

thus the velocity verifies

$$v(\mathrm{x}_2) = \sqrt{2\mathsf{g}(\mathsf{h} - \mathrm{x}_2)}.$$

The total time τ from $\mathbf{P}_1$ to $\mathbf{P}_0$ is

$$\tau = \tau(\mathsf{h}) = \int_{\mathbf{P}_1}^{\mathbf{P}_0} \frac{d\Gamma(\mathrm{x}_2)}{v(\mathrm{x}_2)} = \int_0^{\mathsf{h}} \sqrt{\frac{1 + \psi'(\mathrm{x}_2)^2}{2\mathsf{g}(\mathsf{h} - \mathrm{x}_2)}} d\mathrm{x}_2, \quad \mathsf{h} > 0.$$

[1] This example is taken from [61].

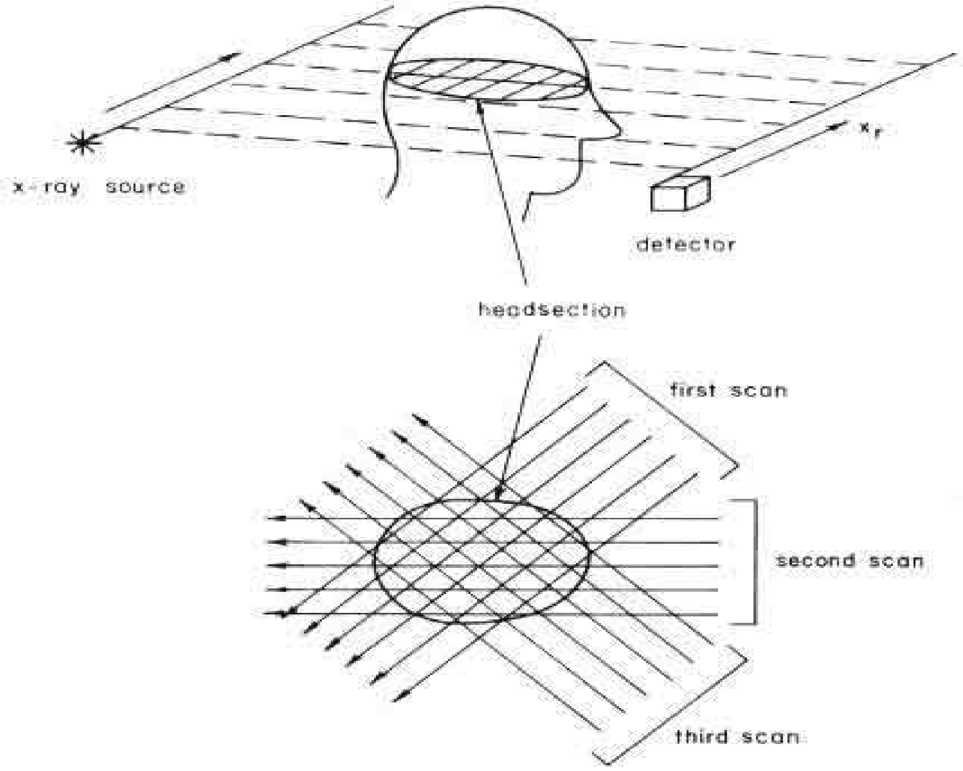


Figure 1.2: A classical example of computerized tomography.

Set $\phi(\mathrm{x}_2) := \sqrt{1+\psi'(\mathrm{x}_2)^2}$ and let $f(\mathsf{h}) := \tau(\mathsf{h})\sqrt{2\mathsf{g}}$ be known (measured). Then the problem is to determine the unknown function ϕ from Abel's integral equation

$$\int_0^{\mathsf{h}} \frac{\phi(\mathrm{x}_2)}{\sqrt{\mathsf{h}-\mathrm{x}_2}} d\mathrm{x}_2 = f(\mathsf{h}), \quad \mathsf{h} > 0. \tag{1.5}$$

1.4 Radon inversion (X-ray tomography)

We consider another very important example widely studied in medical applications, arising in Computerized Tomography, which can lead to an Abel integral equation.

Let $\Omega \subseteq \mathbb{R}^2$ be a compact domain with a spatially varying density $f : \Omega \to \mathbb{R}$ (in medical applications, Ω represents the section of a human body, see Figure 1.2). Let $\mathbb{L}$ be any line in the plane and suppose we direct a thin beam of

X-rays into the body along $\mathbb{L}$ and measure how much intensity is attenuated by going through the body. Let $\mathbb{L}$ be parametrized by its normal versor $\boldsymbol{\theta} \in \mathbb{S}^1$ and its distance $s > 0$ from the origin. If we assume that the decay $-\Delta\mathsf{I}$ of an X-ray beam along a distance Δt is proportional to the intensity I, the density f and to Δt, we obtain

$$\Delta\mathsf{I}(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp) = -\mathsf{I}(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp) f(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp)\Delta t,$$

where $\boldsymbol{\theta}^\perp$ is a unit vector orthogonal to $\boldsymbol{\theta}$. In the limit for $t \to 0$, we have

$$\frac{d}{dt}\mathsf{I}(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp) = -\mathsf{I}(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp) f(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp).$$

Thus, if $\mathsf{I}_{\mathbb{L}}(s, \boldsymbol{\theta})$ and $\mathsf{I}_0(s, \boldsymbol{\theta})$ denote the intensity of the X-ray beam measured at the detector and the emitter, respectively, where the detector and the emitter are connected by the line parametrized by s and $\boldsymbol{\theta}$ and are located outside of Ω, an integration along the line yields

$$\mathscr{R}[f](s, \boldsymbol{\theta}) := \int f(s\boldsymbol{\theta} + t\boldsymbol{\theta}^\perp)dt = -\log \frac{\mathsf{I}_{\mathbb{L}}(s, \boldsymbol{\theta})}{\mathsf{I}_0(s, \boldsymbol{\theta})}, \quad \boldsymbol{\theta} \in \mathbb{S}^1, \ s > 0, \tag{1.6}$$

where the integration can be made over $\mathbb{R}$, since obviously $f = 0$ outside of Ω.

The inverse problem of determining the density distribution f from the X-ray measurements is then equivalent to solve the integral equation of the first kind (1.6). The operator $\mathscr{R}$ is called the *Radon Transform* in honor of the Austrian mathematician J. Radon, who studied the problem of reconstructing a function of two variables from its line integrals already in 1917 (cf. [75]).

The problem simplifies in the following special case (which is of interest, e.g. in material testing), where Ω is a circle of radius R, f is radially symmetric, i.e. $f(r, \boldsymbol{\theta}) = \psi(r)$, $0 < r \leq \mathsf{R}$, $\|\boldsymbol{\theta}\| = 1$ for a suitable function ψ, and we choose only horizontal lines. If

$$g(s) := -\log \frac{\mathsf{I}_{\mathbb{L}}(s, \boldsymbol{\theta}_0)}{\mathsf{I}_0(s, \boldsymbol{\theta}_0)} \tag{1.7}$$

denotes the measurements in this situation, with $\boldsymbol{\theta}_0 = (0, \pm 1)$, for $0 < s \leq \mathsf{R}$ we have

$$\begin{aligned}\mathscr{R}[f](s, \boldsymbol{\theta}_0) &= \int_{\mathbb{R}} f(s\boldsymbol{\theta}_0 + t\boldsymbol{\theta}_0^{\perp})dt = \int_{\mathbb{R}} \psi(\sqrt{s^2+t^2})dt \\ &= \int_{-\sqrt{\mathsf{R}^2-s^2}}^{\sqrt{\mathsf{R}^2-s^2}} \psi(\sqrt{s^2+t^2})dt = 2\int_0^{\sqrt{\mathsf{R}^2-s^2}} \psi(\sqrt{s^2+t^2})dt \qquad (1.8)\\ &= 2\int_s^{\mathsf{R}} \frac{r\psi(r)}{\sqrt{r^2-s^2}}dr.\end{aligned}$$

Thus we obtain another Abel integral equation of the first kind:

$$\int_s^{\mathsf{R}} \frac{r\psi(r)}{\sqrt{r^2-s^2}}dr = \frac{g(s)}{2}, \quad 0 < s \leq \mathsf{R}. \qquad (1.9)$$

In the case $g(\mathsf{R}) = 0$, the Radon Transform can be explicitly inverted and

$$\psi(r) = -\frac{1}{\pi}\int_r^{\mathsf{R}} \frac{g'(s)}{\sqrt{s^2-r^2}}ds. \qquad (1.10)$$

We observe from the last equation that the inversion formula involves the derivative of the data g, which can be considered as an indicator of the ill-posedness of the problem. However, here the data is integrated and thus smoothed again, but the kernel of this integral operator has a singularity for $s = r$, so we expect the regularization effect of integration to be only partial. This heuristic statement can be made more precise, as we shall see later in Chapter 4.

1.5 Integral equations of the first kind

In Section 1.3, starting from a physical problem, we have constructed a very simple mathematical model based on the integral equation (1.5), where we have to recover the unknown function ϕ from the data f. Similarly, in Section 1.4 we have seen that an analogous equation is obtained to recover the function ψ from the measured data g.

As a matter of fact, very often ill-posed problems lead to integral equations. In particular, Abel integral equations such as (1.5) and (1.10) fall into the class of the Fredholm equations of the first kind whose definition is recalled below.

Definition 1.5.1. *Let $s_1 < s_2$ be real numbers and let $\varkappa$, f and ϕ be real valued functions defined respectively on $[s_1, s_2]^2$, $[s_1, s_2]$ and $[s_1, s_2]$. A Fredholm equation of the first kind is an equation of the form*

$$\int_{s_1}^{s_2} \varkappa(s,t)\phi(t)dt = f(s) \quad s \in [s_1, s_2]. \tag{1.11}$$

Fredholm integral equations of the first kind must be treated accurately (see [22] as a general reference): if $\varkappa$ is continuous and ϕ is integrable, then it is easy to see that f is also continuous, thus if the data is not continuous while the kernel is, then (1.11) has no integrable solution. This means that the question of existence is not trivial and requires investigation. Concerning the uniqueness of the solutions, take for example $\varkappa(s,t) = s \sin t$, $f(s) = s$ and $[s_1, s_2] = [0, \pi]$: then $\phi(t) = 1/2$ is a solution of (1.11), but so is each of the functions $\phi_n(t) = 1/2 + \sin nt$, $n \in \mathbb{N}$.
Moreover, we also observe that if $\varkappa$ is square integrable, as a consequence of the Riemann-Lebesgue lemma, there holds

$$\int_0^\pi \varkappa(s,t)\sin(nt)dt \to 0 \quad \text{as} \quad n \to +\infty. \tag{1.12}$$

Thus if ϕ is a solution of (1.11) and C is arbitrarily large

$$\int_0^\pi \varkappa(s,t)(\phi(t) + C\sin(nt))dt \to f(s) \quad \text{as} \quad n \to +\infty \tag{1.13}$$

and for large values of n the slightly perturbed data

$$\tilde{f}(s) := f(s) + C \int_0^\pi \varkappa(s,t)\sin(nt)dt \tag{1.14}$$

corresponds to a solution $\tilde{\phi}(t) = \phi(t) + C\sin(nt)$ which is arbitrarily distant from ϕ. In other words, as in the example considered in Section 1.2, the solution doesn't depend continuously on the data.

1.6 Hadamard's definition of ill-posed problems

Integral equations of the first kind are the most famous example of ill-posed problems. The definition of ill-posedness goes back to the beginning of the 20-th century and was stated by J. Hadamard as follows:

Definition 1.6.1. *Let F be a mapping from a topological space $\mathcal{X}$ to another topological space $\mathcal{Y}$ and consider the abstract equation*

$$F(x) = y.$$

The equation is said to be well-posed if

(i) for each $y \in \mathcal{Y}$, a solution $x \in \mathcal{X}$ of $F(x) = y$ exists;

(ii) the solution x is unique in $\mathcal{X}$;

(iii) the dependence of x upon y is continuous.

The equation is said to be ill-posed if it is not well-posed.

Of course, the definition of well-posedness above is equivalent to the request that F is surjective and injective and that the inverse mapping F^{-1} is continuous.
For example, due to the considerations made in the previous sections, integral equations of the first kind are examples ill-posed equations. If $\mathcal{X} = \mathcal{Y}$ is a Hilbert space and $F = A$ is a linear, self-adjoint operator with its spectrum contained in $[0, +\infty[$, the equation of the second kind

$$y = Ax + tx$$

is well-posed for every $t > 0$, since the operator $A + t$ is invertible and its inverse $(A + t)^{-1}$ is also continuous.

1.7 Fundamental tools in the Hilbert space setting

So far, we have seen several examples of ill-posed problems. It is obvious from Hadamard's definition of ill-posedness that an exhaustive mathematical treatment of such problems should be based on a different notion of solution of the abstract equation $F(x) = y$ to achieve existence and uniqueness. For linear problems in the Hilbert space setting, this is done with the Moore-Penrose Generalized Inverse.
At first, we fix some standard definitions and notations. General references for this section are [17] and [21].

1.7.1 Basic definitions and notations

Let A be a linear bounded (continuous) operator between two Hilbert Spaces $\mathcal{X}$ and $\mathcal{Y}$. To simplify the notations, the scalar products in $\mathcal{X}$ and $\mathcal{Y}$ and their induced norms will be denoted by the same symbols $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively. For $\bar{x} \in \mathcal{X}$ and $\delta > 0$,

$$\mathcal{B}_\delta(\bar{x}) := \{x \in \mathcal{X} \mid \|x - \bar{x}\| < \delta\} \tag{1.15}$$

is the open ball centered in $\bar{x}$ with radius δ and $\overline{\mathcal{B}_\delta(\bar{x})}$ or $\overline{\mathcal{B}}_\delta(\bar{x})$ is its closure with respect to the topology of $\mathcal{X}$.
We denote by $\mathcal{R}(A)$ the range of A:

$$\mathcal{R}(A) := \{y \in \mathcal{Y} \mid \exists\, x \in \mathcal{X} : \ y = Ax\} \tag{1.16}$$

and by $\ker(A)$ the null-space of A:

$$\ker(A) := \{x \in \mathcal{X} \mid Ax = 0\}. \tag{1.17}$$

We recall that $\mathcal{R}(A)$ and $\ker(A)$ are subspaces of $\mathcal{Y}$ and $\mathcal{X}$ respectively and that $\ker(A)$ is closed.
We also recall the following basic definitions.

Definition 1.7.1 (**Orthogonal space**). *Let $\mathcal{M} \subseteq \mathcal{X}$. The orthogonal space of $\mathcal{M}$ is the closed vector space $\mathcal{M}^\perp$ defined by:*

$$\mathcal{M}^\perp := \{x \in \mathcal{X} \mid \langle x, z \rangle = 0, \ \forall z \in \mathcal{M}\}. \tag{1.18}$$

Definition 1.7.2 (**Adjoint operator**). *The bounded operator $A^* : \mathcal{Y} \to \mathcal{X}$, defined as*

$$\langle A^* y, x \rangle = \langle y, Ax \rangle, \quad \forall x \in \mathcal{X}, \ y \in \mathcal{Y}, \tag{1.19}$$

is called the adjoint operator of A.
If $A : \mathcal{X} \to \mathcal{X}$ and $A = A^*$, then A is called self-adjoint.

Definition 1.7.3 (**Orthogonal projector**). *Let $\mathcal{W}$ be a subspace of $\mathcal{X}$.*
For every $x \in \mathcal{X}$, there exist a unique element in $\mathcal{W}$, called the projection of x onto $\mathcal{W}$, that minimizes the distance $\|x - w\|$, $w \in \mathcal{W}$.
The map $P : \mathcal{X} \to \mathcal{X}$, that associates to an element $x \in \mathcal{X}$ its projection onto $\mathcal{W}$, is called the orthogonal projector onto $\mathcal{W}$.
This is the unique linear and self-adjoint operator satisfying the relation $P = P^2$ that maps $\mathcal{X}$ onto $\mathcal{W}$.

Definition 1.7.4. *Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{X}$ and let $x \in \mathcal{X}$. The sequence x_n is said to converge weakly to x if, for every $z \in \mathcal{X}$, $\langle x_n, z \rangle$ converges to $\langle x, z \rangle$. In this case, we shall write*

$$x_n \rightharpoonup x.$$

1.7.2 The Moore-Penrose generalized inverse

We are interested in solving the equation

$$Ax = y, \tag{1.20}$$

for $x \in \mathcal{X}$, but we suppose we are only given an approximation of the exact data $y \in \mathcal{Y}$, which are assumed to exist and to be fixed, but unknown.

Definition 1.7.5. *(i) A least-squares solution of the equation $Ax = y$ is an element $x \in \mathcal{X}$ such that*

$$\|Ax - y\| = \inf_{z \in \mathcal{X}} \|Az - y\|. \tag{1.21}$$

(ii) An element $x \in \mathcal{X}$ is a best-approximate solution of $Ax = y$ if it is a least-squares solution of $Ax = y$ and

$$\|x\| = \inf\{\|z\| \mid z \textit{ is a least-squares solution of } \|Ax - y\|\} \tag{1.22}$$

holds.

(iii) The Moore-Penrose (generalized) inverse $A^\dagger$ of A is the unique linear extension of $\tilde{A}^{-1}$ to

$$\mathcal{D}(A^\dagger) := \mathcal{R}(A) + \mathcal{R}(A)^\perp \tag{1.23}$$

with

$$\ker(A^\dagger) = \mathcal{R}(A)^\perp, \tag{1.24}$$

where

$$\tilde{A} := A|_{\ker(A)^\perp} : \ \ker(A)^\perp \ \to \ \mathcal{R}(A). \tag{1.25}$$

$A^\dagger$ is well defined: in fact, it is trivial to see that $\ker(\tilde{A}) = \{0\}$ and $\mathcal{R}(\tilde{A}) = \mathcal{R}(A)$, so $\mathcal{R}(\tilde{A})^{-1}$ exists. Moreover, since $\mathcal{R}(A) \bigcap \mathcal{R}(A)^\perp = \{0\}$, every $y \in \mathcal{D}(A^\dagger)$ can be written in a unique way as $y = y_1 + y_2$, with $y_1 \in \mathcal{R}(A)$ and $y_2 \in \mathcal{R}(A)^\perp$, so using (1.24) and the requirement that $A^\dagger$ is linear, one can easily verify that $A^\dagger y = \tilde{A}^{-1} y_1$.

The Moore Penrose generalized inverse can be characterized as follows.

Proposition 1.7.1. *Let now and below P and Q be the orthogonal projectors onto* $\ker(A)$ *and* $\overline{\mathcal{R}(A)}$*, respectively. Then $\mathcal{R}(A^\dagger) = \ker(A)^\perp$ and the four Moore-Penrose equations hold:*

$$AA^\dagger A = A, \tag{1.26}$$

$$A^\dagger A A^\dagger = A^\dagger, \tag{1.27}$$

$$A^\dagger A = I - P, \tag{1.28}$$

$$AA^\dagger = Q|_{\mathcal{D}(A^\dagger)}. \tag{1.29}$$

Here and below, the symbol I denotes the identity map.
If a linear operator $\breve{A} : \mathcal{D}(A^\dagger) \to \mathcal{X}$ verifies (1.28) *and* (1.29)*, then $\breve{A} = A^\dagger$.*

Proof. For the first part, see [17]. We show the second part.
Since $\breve{A}A\breve{A} = \breve{A}Q|_{\mathcal{D}(A^\dagger)} = \breve{A}$ and $A\breve{A}A = A(I-P) = A - AP = A$, all Moore-Penrose equations hold for $\breve{A}$. Then, keeping in mind that $I - P$ is the orthogonal projector onto $\ker(A)^\perp$, for every $y \in \mathcal{D}(A^\dagger)$ we have: $\breve{A}y = \breve{A}A\breve{A}y = (I-P)\breve{A}y = \tilde{A}^{-1}A(I-P)\breve{A}y = \tilde{A}^{-1}A\breve{A}y = \tilde{A}^{-1}Qy = A^\dagger y$. □

An application of the Closed Graph Theorem leads to the following important fact.

Proposition 1.7.2. *The Moore-Penrose generalized inverse $A^\dagger$ has a closed graph $gr(A^\dagger)$. Furthermore, $A^\dagger$ is continuous if and only if $\mathcal{R}(A)$ is closed.*

We use this to give another characterization of the Moore-Penrose pseudoinverse:

Proposition 1.7.3. *There can be only one linear bounded operator*
$\breve{A} : \mathcal{Y} \to \mathcal{X}$ that verifies (1.26) *and* (1.27) *and such that $\breve{A}A$ and $A\breve{A}$ are self-adjoint. If such an $\breve{A}$ exists, then $A^\dagger$ is also bounded and $\breve{A} = A^\dagger$. Moreover, in this case, the Moore-Penrose generalized inverse of the adjoint of A, $(A^*)^\dagger$, is bounded too and*

$$(A^*)^\dagger = (A^\dagger)^*. \tag{1.30}$$

Proof. Suppose $\breve{A}, \breve{B} : \mathcal{Y} \to \mathcal{X}$ are linear bounded operators that verify (1.26) and (1.27) and such that $\breve{A}A$, $\breve{B}A$, $A\breve{A}$ and $A\breve{B}$ are self-adjoint.
Then

$$\breve{A}A = A^*\breve{A}^* = (A^*\breve{B}^*A^*)\breve{A}^* = (A^*\breve{B}^*)(A^*\breve{A}^*) = (\breve{B}A)(\breve{A}A) = \breve{B}(A\breve{A}A) = \breve{B}A$$

and in a similar way $A\check{B} = A\check{A}$. Thus we obtain

$$\check{A} = \check{A}(A\check{A}) = \check{A}(A\check{B}) = (\check{A}A)\check{B} = (\check{B}A)\check{B} = \check{B}.$$

Suppose now that such an operator $\check{A}$ exists. For every $z \in \mathcal{Y}$ and every $y \in \overline{\mathcal{R}(A)}$, $y = \lim_{n\to+\infty} Ax_n$, $x_n \in \mathcal{X}$, we have

$$\langle y, z \rangle = \lim_{n\to+\infty} \langle Ax_n, z\rangle = \lim_{n\to+\infty} \langle x_n, A^* z\rangle =$$

$$= \lim_{n\to+\infty} \langle x_n, A^* \check{A}^* A^* z\rangle = \langle A\check{A}y, z\rangle,$$

so $y = A\check{A}y$ and y lies in $\mathcal{R}(A)$. This means that $\mathcal{R}(A)$ is closed, thus according to Proposition 1.7.2 $A^\dagger$ is bounded and for the first part of the proof $A^\dagger = \check{A}$.

Finally, to prove the last statement it is enough to verify that for the linear bounded operator $(A^\dagger)^*$ conditions (1.26) and (1.27) hold with A replaced by A^*, together with the the correspondent self-adjointity conditions, which consists just of straightforward calculations. □

The definitions of least-squares solution and best-approximate solution make sense too: if $y \in \mathcal{D}(A^\dagger)$, the set $\mathcal{S}$ of the least-squares solutions of $Ax = y$ is non-empty and its best-approximate solution turns out to be unique and strictly linked to the operator $A^\dagger$. More precisely, we have:

Proposition 1.7.4. *Let $y \in \mathcal{D}(A^\dagger)$. Then:*

(i) *$Ax = y$ has a unique best-approximate solution, which is given by*

$$x^\dagger := A^\dagger y. \tag{1.31}$$

(ii) *The set $\mathcal{S}$ of the least-squares solutions of $Ax = y$ is equal to $x^\dagger +$ $\ker(A)$ and every $x \in \mathcal{X}$ lies in $\mathcal{S}$ if and only if the normal equation*

$$A^* A x = A^* y \tag{1.32}$$

holds.

(iii) $\mathcal{D}(A^\dagger)$ *is the natural domain of definition for* $A^\dagger$*, in the sense that*

$$y \notin \mathcal{D}(A^\dagger) \Rightarrow \mathcal{S} = \emptyset. \tag{1.33}$$

Proof. See [17] for (i) and (ii). Here we show (iii). Suppose $x \in \mathcal{X}$ is a least-squares solution of $Ax = y$. Then Ax is the closest element in $\mathcal{R}(A)$ to y, so $Ax - y \in \mathcal{R}(A)^\perp$. This implies that $Q(Ax - y) = 0$, but $QAx = Ax$, thus we deduce that $Qy \in \mathcal{R}(A)$ and $y = Qy + (I - Q)y \in \mathcal{D}(A^\dagger)$. □

Thus the introduction of the concept of best-approximate solution, although it enforces uniqueness, does not always lead to a solvable problem and is no remedy for the lack of continuous dependance from the data in general.

1.8 Compact operators: SVD and the Picard criterion

Among linear and bounded operators, compact operators are of special interest, since many integral operators are compact.
We recall that a compact operator is a linear operator that maps any bounded subset of $\mathcal{X}$ into a relatively compact subset of $\mathcal{Y}$.
For example, suppose that $\Omega \subseteq \mathbb{R}^D$, $D \geq 1$, is a nonempty compact and Jordan measurable set that coincides with the closure of its interior. It is well known that if the kernel $\varkappa$ is either in $\mathcal{L}^2(\Omega \times \Omega)$ or weakly singular, i.e. $\varkappa$ is continuous on $\{(s,t) \in \Omega \times \Omega \mid s \neq t\}$ and for all $s \neq t \in \Omega$

$$|\varkappa(s,t)| \leq \frac{C}{|s-t|^{D-\epsilon}} \tag{1.34}$$

with $C > 0$ and $\epsilon > 0$, then the operator $K : \mathcal{L}^2(\Omega) \to \mathcal{L}^2(\Omega)$ defined by

$$K[x](s) := \int_\Omega \varkappa(s,t)x(t)dt \tag{1.35}$$

is compact (see e.g. [20]).[2]

If a compact linear operator K is also self-adjoint, the notion of eigensystem is well-defined (a proof of the existence of an eigensystem and a more exhaustive treatment of the argument can be found e.g. in [62]): an eigensystem $\{(\lambda_j; v_j)\}_{j\in\mathbb{N}}$ of the operator K consists of all nonzero eigenvalues $\lambda_j \in \mathbb{R}$ of K and a corresponding complete orthonormal set of eigenvectors v_j. Then K can be diagonalized by means of the formula

$$Kx = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \quad \forall x \in \mathcal{X} \tag{1.36}$$

and the nonzero eigenvalues of K converge to 0.

If $K : \mathcal{X} \to \mathcal{Y}$ is not self-adjoint, then observing that the operators $K^*K : \mathcal{X} \to \mathcal{X}$ and $KK^* : \mathcal{Y} \to \mathcal{Y}$ are positive semi-definite and self-adjoint compact operators with the same set of nonzero eigenvalues written in nondecreasing order with multiplicity

$$\lambda_1^2 \geq \lambda_2^2 \geq \lambda_3^2 ..., \quad \lambda_j > 0 \;\; \forall j \in \mathbb{N},$$

we define a *singular system* $\{\lambda_j; v_j, u_j\}_{j\in\mathbb{N}}$. The vectors v_j form a complete orthonormal system of eigenvectors of K^*K and u_j, defined as

$$u_j := \frac{Kv_j}{\|Kv_j\|}, \tag{1.37}$$

form a complete orthonormal system of eigenvectors of KK^*. Thus $\{v_j\}_{j\in\mathbb{N}}$ span $\overline{\mathcal{R}(K^*)} = \overline{\mathcal{R}(K^*K)}$, $\{u_j\}_{j\in\mathbb{N}}$ span $\overline{\mathcal{R}(K)} = \overline{\mathcal{R}(KK^*)}$ and the following formulas hold:

$$Kv_j = \lambda_j u_j, \tag{1.38}$$

$$K^* u_j = \lambda_j v_j, \tag{1.39}$$

$$Kx = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle u_j, \quad \forall x \in \mathcal{X}, \tag{1.40}$$

[2] Here and below, we shall denote with the symbol K a linear and bounded operator which is also compact.

$$K^*y = \sum_{j=1}^{\infty} \lambda_j \langle y, u_j \rangle v_j, \quad \forall y \in \mathcal{Y}. \tag{1.41}$$

All the series above converge in the Hilbert space norms of $\mathcal{X}$ and $\mathcal{Y}$.
Equations (1.40) and (1.41) are the natural infinite-dimensional extension of the well known singular value decomposition (SVD) of a matrix.
For compact operators, the condition for the existence of the best-approximate solution $K^\dagger y$ of the equation $Kx = y$ can be written down in terms of the singular value expansion of K. It is called the *Picard Criterion* and can be given by means of the following theorem (see [17] for the proof).

Theorem 1.8.1. *Let $\{\lambda_j; v_j, u_j\}_{j\in\mathbb{N}}$ be a singular system for the compact linear operator K an let $y \in \mathcal{Y}$. Then*

$$y \in \mathcal{D}(K^\dagger) \Longleftrightarrow \sum_{j=1}^{\infty} \frac{|\langle y, u_j \rangle|^2}{\lambda_j^2} < +\infty \tag{1.42}$$

and whenever $y \in \mathcal{D}(K^\dagger)$,

$$K^\dagger y = \sum_{j=1}^{\infty} \frac{\langle y, u_j \rangle}{\lambda_j} v_j. \tag{1.43}$$

Thus the Picard Criterion states that the best-approximate solution $x^\dagger$ of $Kx = y$ exists only if the SVD coefficients $|\langle y, u_j \rangle|$ decay fast-enough with respect to the singular values λ_j.
In the finite dimensional case, of course the sum in (1.42) is always finite, the best-approximate solution always exists and the Picard Criterion is always satisfied.
From (1.43) we can see that in the case of perturbed data, the error components with respect to the the basis $\{u_j\}$ corresponding to the small values of λ_j are amplified by the large factors λ_j^{-1}. For example, if $\dim(\mathcal{R}(K)) = +\infty$ and the perturbed data is defined by $y_j^\delta := y + \delta u_j$, then $\|y - y_j^\delta\| = \delta$, but

$$K^\dagger y - K^\dagger y_j^\delta = \lambda_j^{-1} \langle \delta u_j, u_j \rangle v_j$$

and hence

$$\|K^\dagger y - K^\dagger y_j^\delta\| = \lambda_j^{-1} \delta \to +\infty \quad \text{as} \quad j \to +\infty.$$

In the finite dimensional case there are only finitely many eigenvalues, so these amplification factors stay bounded. However, they might be still very large: this is the case of the discrete ill-posed problems, for which also a Discrete Picard Condition can be defined, as we shall see later on.

1.9 Regularization and Bakushinskii's Theorem

In the previous sections, we started discussing the problem of solving the equation $Ax = y$. In practice, the exact data y is not known precisely, but only approximations y^δ with

$$\|y^\delta - y\| \leq \delta \tag{1.44}$$

is available. In literature, $y^\delta \in \mathcal{Y}$ is called the *noisy data* and $\delta > 0$ the *noise level.*

Our purpose is to approximate the best-approximate solution $x^\dagger = A^\dagger y$ of (1.20) from the knowledge of δ, y^δ and A.

According to Proposition 1.7.2, in general the operator $A^\dagger$ is not continuous, so in the ill-posed case $A^\dagger y^\delta$ is in general a very bad approximation of $x^\dagger$ even if it exists. Roughly speaking, regularizing $Ax = y$ means essentially to construct of a family of continuous operators $\{R_\sigma\}$, depending on a certain *regularization parameter* σ, that approximate $A^\dagger$ (in some sense) and such that $x_\sigma^\delta := R_\sigma y^\delta$ satisfies the conditions above. We state this more precisely in the following fundamental definition.

Definition 1.9.1. *Let $\sigma_0 \in (0, +\infty]$. For every $\sigma \in (0, \sigma_0)$, let $R_\sigma : \mathcal{Y} \to \mathcal{X}$ be a continuous (not necessarily linear) operator.*

The family $\{R_\sigma\}$ is called a regularization operator for $A^\dagger$ if, for every $y \in \mathcal{D}(A^\dagger)$, there exists a function

$$\alpha : \ (0, +\infty) \times \mathcal{Y} \to (0, \sigma_0),$$

called parameter choice rule for y, that allows to associate to each couple (δ, y^δ) a specific operator $R_{\alpha(\delta,y^\delta)}$ and a regularized solution $x^\delta_{\alpha(\delta,y^\delta)} := R_{\alpha(\delta,y^\delta)} y^\delta$, and such that

$$\lim_{\delta \to 0} \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \alpha(\delta, y^\delta) = 0. \tag{1.45}$$

If the parameter choice rule (below, p.c.r.) α satisfies in addition

$$\lim_{\delta \to 0} \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \| R_{\alpha(\delta,y^\delta)} y^\delta - A^\dagger y \| = 0. \tag{1.46}$$

then it is said to be convergent.
For a specific $y \in \mathcal{D}(A^\dagger)$, a pair (R_σ, α) is called a (convergent) regularization method for solving $Ax = y$ if $\{R_\sigma\}$ is a regularization for $A^\dagger$ and α is a (convergent) parameter choice rule for y.

Remark 1.9.1.

- *In the definition above all possible noisy data with noise level $\leq \delta$ are considered, so the convergence is intended in a worst-case sense.*
  However, in a specific problem, a sequence of approximate solutions $x^{\delta_n}_{\alpha(\delta_n, y^{\delta_n})}$ can converge fast to $x^\dagger$ also when (1.46) fails!

- *A p.c.r. $\alpha = \alpha(\delta, y^\delta)$ depends explicitly on the noise level and on the perturbed data y^δ.*
  According to the definition above it should also depend on the exact data y, which is unknown, so this dependance can only be on some a priori knowledge about y like smoothness properties.

We distinguish between two types of parameter choice rules:

Definition 1.9.2. *Let α be a parameter choice rule according to Definition* 1.9.1. *If α does not depend on y^δ, but only on δ, then we call α an a-priori parameter choice rule and write $\alpha = \alpha(\delta)$. Otherwise, we call α an a-posteriori parameter choice rule. If α does not depend on the noise level δ, then it is said to be an heuristic parameter choice rule.*

If α does not depend on y^δ it can be defined before the actual calculations once and for all: this justifies the terminology a-priori and a-posteriori in the definition above.
For the choice of the parameter, one can also construct a p.c.r. that depends explicitly only on the perturbed data y^δ and not on the noise level. However, a famous result due to Bakushinskii shows that such a p.c.r. cannot be convergent:

Theorem 1.9.1 (**Bakushinskii**). *Suppose $\{R_\sigma\}$ is a regularization operator for $A^\dagger$ such that for every $y \in \mathcal{D}(A^\dagger)$ there exists a convergent p.c.r. α which depends only on y^δ. Then $A^\dagger$ is bounded.*

Proof. If $\alpha = \alpha(y^\delta)$, then it follows from (1.46) that for every $y \in \mathcal{D}(A^\dagger)$ we have

$$\lim_{\delta \to 0} \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \|R_{\alpha(y^\delta)} y^\delta - A^\dagger y\| = 0, \tag{1.47}$$

so that $R_{\alpha(y)} y = A^\dagger y$. Thus, if $\{y_n\}_{n\in\mathbb{N}}$ is a sequence in $\mathcal{D}(A^\dagger)$ converging to y, then $A^\dagger y_n = R_{\alpha(y_n)} y_n$ converges to $A^\dagger y$. This means that $A^\dagger$ is sequentially continuous, hence it is bounded. □

Thus, in the ill-posed case, no heuristic parameter choice rule can yield a convergent regularization method. However, this doesn't mean that such a p.c.r. cannot give good approximations of $x^\dagger$ for a fixed positive δ!

1.10 Construction and convergence of regularization methods

In general terms, regularizing an ill-posed problem leads to three questions:

1. How to construct a regularization operator?

2. How to choose a parameter choice rule to give rise to convergent regularization methods?

3. How can these steps be performed in some optimal way?

This and the following sections will deal with the answers of these problems, at least in the linear case. The following result provides a characterization of the regularization operators. Once again, we refer to [17] for more details and for the proofs.

Proposition 1.10.1. *Let $\{R_\sigma\}_{\sigma>0}$ be a family of continuous operators converging pointwise on $\mathcal{D}(A^\dagger)$ to $A^\dagger$ as $\sigma \to 0$.*
Then $\{R_\sigma\}$ is a regularization for $A^\dagger$ and for every $y \in \mathcal{D}(A^\dagger)$ there exists an a-priori p.c.r. α such that (R_σ, α) is a convergent regularization method for solving $Ax = y$.
Conversely, if (R_σ, α) is a convergent regularization method for solving $Ax = y$ with $y \in \mathcal{D}(A^\dagger)$ and α is continuous with respect to δ, then $R_\sigma y$ converges to $A^\dagger y$ as $\sigma \to 0$.

Thus the correct approach to understand the meaning of regularization is pointwise convergence. Furthermore, if $\{R_\sigma\}$ is linear and uniformly bounded, in the ill-posed case we can't expect convergence in the operator norm, since then $A^\dagger$ would have to be bounded.
We consider an example of a regularization operator which fits the definitions above. Although very similar examples can be found in literature, cf. e.g. [61], this is slightly different.

Example 1.10.1. *Consider the operator $K : \mathcal{X} := \mathcal{L}^2[0,1] \to \mathcal{Y} := \mathcal{L}^2[0,1]$ defined by*

$$K[x](s) := \int_0^s x(t)dt.$$

Then K is linear, bounded and compact and it is easily seen that

$$\mathcal{R}(K) = \{y \in \mathcal{H}^1[0,1] \mid y \in \mathcal{C}([0,1]), y(0) = 0\} \tag{1.48}$$

and that the distributional derivative from $\mathcal{R}(K)$ to $\mathcal{X}$ is the inverse of K. Since $\mathcal{R}(K) \supseteq \mathcal{C}_0^\infty[0,1]$, $\mathcal{R}(K)$ is dense in $\mathcal{Y}$ and $\mathcal{R}(K)^\perp = \{0\}$.

For $y \in \mathcal{C}([0,1])$ and for $\sigma > 0$, define

$$(R_\sigma y)(t) := \begin{cases} \frac{1}{\sigma}(y(t+\sigma) - y(t)), & \text{if} \quad 0 \leq t \leq \frac{1}{2}, \\ \frac{1}{\sigma}(y(t) - y(t-\sigma)), & \text{if} \quad \frac{1}{2} \ < t \leq 1. \end{cases} \tag{1.49}$$

Then $\{R_\sigma\}$ is a family of linear and bounded operators with

$$\|R_\sigma y\|_{\mathcal{L}^2[0,1]} \leq \frac{\sqrt{6}}{\sigma} \|y\|_{\mathcal{L}^2[0,1]} \tag{1.50}$$

defined on a dense linear subspace of $\mathcal{L}^2[0,1]$, thus it can be extended to the whole $\mathcal{Y}$ and (1.50) *is still true.*
Since the measure of $[0,1]$ is finite, for $y \in \mathcal{R}(K)$ the distributional derivative of y lies in $\mathcal{L}^1[0,1]$, so y is a function of bounded variation. Thus, according to Lebesgue's Theorem, the ordinary derivative y' exists almost everywhere in $[0,1]$ and it is equal to the distributional derivative of y as an $\mathcal{L}^2$-function. Consequently, we can apply the Dominate Convergence Theorem to show that

$$\|R_\sigma y - K^\dagger y\|_{\mathcal{L}^2[0,1]} \to 0, \ \text{as } \sigma \to 0$$

so that, according to Proposition 1.10.1*, R_σ is a regularization for the distributional derivative $K^\dagger$.*

1.11 Order optimality

We concentrate now on how to construct *optimal* regularization methods. To this aim, we shall make use of some analytical tools from the spectral theory of linear and bounded operators. For the reader who is not accustomed with the spectral theory, we refer to Appendix *A* or to the second chapter of [17], where the basic ideas and results we will need are gathered in a few pages; for a more comprehensive treatment, classical references are e.g. [2] and [44]. In principle, a (convergent) regularization method $(\bar{R}_\sigma, \bar{\alpha})$ for solving $Ax = y$ should be optimal if the quantity

$$\varepsilon_1 = \varepsilon_1(\delta, \bar{R}_\sigma, \bar{\alpha}) := \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \|\bar{R}_{\bar{\alpha}(\delta, y^\delta)} y^\delta - A^\dagger y\| \tag{1.51}$$

converges to 0 as quickly as

$$\varepsilon_2 = \varepsilon_2(\delta) := \inf_{(R_\sigma,\alpha)} \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \|R_{\alpha(\delta,y^\delta)}y^\delta - A^\dagger y\|, \tag{1.52}$$

i.e. if there are no regularization methods for which the approximate solutions converge to 0 (in the usual worst-case sense) quicker than the approximate solutions of $(\bar{R}_\sigma, \bar{\alpha})$.
Once again, it is not advisable to look for some uniformity in y, as we can see from the following result.

Proposition 1.11.1. *Let $\{R_\sigma\}$ be a regularization for $A^\dagger$ with $R_\sigma(0) = 0$, let $\alpha = \alpha(\delta, y^\delta)$ be a p.c.r. and suppose that $\mathcal{R}(A)$ is non closed. Then there can be no function $f : (0, +\infty) \to (0, +\infty)$ with $\lim_{\delta\to 0} f(\delta) = 0$ such that*

$$\varepsilon_1(\delta, R_\sigma, \alpha) \le f(\delta) \tag{1.53}$$

holds for every $y \in \mathcal{D}(A^\dagger)$ with $\|y\| \le 1$ and all $\delta > 0$.

Thus convergence can be arbitrarily slow: in order to study convergence rates of the approximate solutions to $x^\dagger$ it is necessary to restrict on subsets of $\mathcal{D}(A^\dagger)$ (or of $\mathcal{X}$), i.e. to formulate some a-priori assumption on the exact data (or equivalently, on the exact solution). This can be done by introducing the so called *source sets*, which are defined as follows.

Definition 1.11.1. *Let μ, $\rho > 0$. An element $x \in \mathcal{X}$ is said to have a source representation if it belongs to the source set*

$$\mathcal{X}_{\mu,\rho} := \{x \in \mathcal{X} \mid x = (A^*A)^\mu w,\ \|w\| \le \rho\}. \tag{1.54}$$

The union with respect to ρ of all source sets is denoted by

$$\mathcal{X}_\mu := \bigcup_{\rho>0} \mathcal{X}_{\mu,\rho} = \{x \in \mathcal{X} \mid x = (A^*A)^\mu w\} = \mathcal{R}((A^*A)^\mu). \tag{1.55}$$

Here and below, we use spectral theory to define

$$(A^*A)^\mu := \int \lambda^\mu dE_\lambda, \tag{1.56}$$

where $\{E_\lambda\}$ *is the spectral family associated to the self-adjoint* A^*A (*cf. Appendix A*) *and since* $A^*A \geq 0$ *the integration can be restricted to the compact set* $[0, \|A^*A\|] = [0, \|A\|^2]$.

Since A is usually a smoothing operator, the requirement for an element to be in $\mathcal{X}_{\mu,\rho}$ can be considered as a smoothness condition.
The notion of optimality is based on the following result about the source sets, which is stated for compact operators, but can be extended to the non compact case (see [17] and the references therein).

Proposition 1.11.2. *Let* A *be compact with* $\mathcal{R}(A)$ *being non closed and let* $\{R_\sigma\}$ *be a regularization operator for* $A^\dagger$. *Define also*

$$\Delta(\delta, \mathcal{X}_{\mu,\rho}, R_\sigma, \alpha) := \sup\{\|R_{\alpha(\delta,y^\delta)}y^\delta - x\| \mid x \in \mathcal{X}_{\mu,\rho}, y^\delta \in \overline{\mathcal{B}_\delta(Ax)}\} \quad (1.57)$$

for any fixed μ, ρ *and* δ *in* $(0, +\infty)$ (*and* α *a p.c.r. relative to* Ax). *Then there exists a sequence* $\{\delta_k\}$ *converging to* 0 *such that*

$$\Delta(\delta_k, \mathcal{X}_{\mu,\rho}, R_\sigma, \alpha) \geq \delta_k^{\frac{2\mu}{2\mu+1}} \rho^{\frac{1}{2\mu+1}}. \quad (1.58)$$

This justifies the following definition.

Definition 1.11.2. *Let* $\mathcal{R}(A)$ *be non closed, let* $\{R_\sigma\}$ *be a regularization for* $A^\dagger$ *and let* μ, $\rho > 0$.
Let α *be a p.c.r. which is convergent for every* $y \in A\mathcal{X}_{\mu,\rho}$.
We call (R_σ, α) *optimal in* $\mathcal{X}_{\mu,\rho}$ *if*

$$\Delta(\delta, \mathcal{X}_{\mu,\rho}, R_\sigma, \alpha) = \delta^{\frac{2\mu}{2\mu+1}} \rho^{\frac{1}{2\mu+1}} \quad (1.59)$$

holds for every $\delta > 0$.
We call (R_σ, α) *of optimal order in* $\mathcal{X}_{\mu,\rho}$ *if there exists a constant* $C \geq 1$ *such that*

$$\Delta(\delta, \mathcal{X}_{\mu,\rho}, R_\sigma, \alpha) \leq C\delta^{\frac{2\mu}{2\mu+1}} \rho^{\frac{1}{2\mu+1}} \quad (1.60)$$

holds for every $\delta > 0$.

The term optimality refers to the fact that if $\mathcal{R}(A)$ is non closed, then a regularization algorithm cannot converge to 0 faster than $\delta^{\frac{2\mu}{2\mu+1}}\rho^{\frac{1}{2\mu+1}}$ as $\delta \to 0$, under the a-priori assumption $x \in \mathcal{X}_{\mu,\rho}$, or (if we are concerned with the rate only), not faster than $O(\delta^{\frac{2\mu}{2\mu+1}})$ under the a-priori assumption $x \in \mathcal{X}_\mu$. In other words, we prove the following fact:

Proposition 1.11.3. *With the assumption of Definition* 1.11.2*,* (R_σ, α) *is of optimal order in* $\mathcal{X}_{\mu,\rho}$ *if and only if there exists a constant* $C \geq 1$ *such that for every* $y \in A\mathcal{X}_{\mu,\rho}$

$$\sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \|R_{\alpha(\delta,y^\delta)}y^\delta - x^\dagger\| \leq C\delta^{\frac{2\mu}{2\mu+1}}\rho^{\frac{1}{2\mu+1}} \tag{1.61}$$

holds for every $\delta > 0$*. For the optimal case an analogous result is true.*

Proof. First, we show that $y \in A\mathcal{X}_{\mu,\rho}$ if and only if $y \in \mathcal{R}(A)$ and $x^\dagger \in \mathcal{X}_{\mu,\rho}$. The sufficiency is obvious because of (1.29). For the necessity, we observe that if $x = (A^*A)^\mu w$, with $\|w\| \leq \rho$, then x lies in $(\ker A)^\perp$, since for every $z \in \ker A$ we have:

$$\begin{aligned}\langle x, z\rangle = \langle (A^*A)^\mu w, z\rangle &= \lim_{\epsilon\to 0}\int_\epsilon^{\|A\|^2} \lambda^\mu d\langle w, E_\lambda z\rangle \\ &= \lim_{\epsilon\to 0}\int_\epsilon^{\|A\|^2} \lambda^\mu d\langle w, z\rangle = 0.\end{aligned} \tag{1.62}$$

We obtain that $x^\dagger = A^\dagger y = A^\dagger A x = (I - P)x = x$ is the only element in $\mathcal{X}_{\mu,\rho} \bigcap A^{-1}(\{y\})$, thus the entire result follows immediately and the proof is complete. □

The following result due to R. Plato assures that, under very weak assumptions, the order-optimality in a source set implies convergence in $\mathcal{R}(A)$. More precisely:

Theorem 1.11.1. *Let* $\{R_\sigma\}$ *a regularization for* $A^\dagger$*. For* $s > 0$*, let* α_s *be the family of parameter choice rules defined by*

$$\alpha_s(\delta, y^\delta) := \alpha(s\delta, y^\delta), \tag{1.63}$$

where α is a parameter choice rule for $Ax = y$, $y \in \mathcal{R}(A)$.
Suppose that for every $\tau > \tau_0$, $\tau_0 \geq 1$, (R_σ, α_τ) is a regularization method of optimal order in $\mathcal{X}_{\mu,\rho}$, for some $\mu, \rho > 0$.
Then, for every $\tau > \tau_0$, (R_σ, α_τ) is convergent for every $y \in \mathcal{R}(A)$ and it is of optimal order in every $\mathcal{X}_{\nu,\rho}$, with $0 < \nu \leq \mu$.

It is worth mentioning that the source sets $\mathcal{X}_{\mu,\rho}$ are not the only possible choice one can make. They are well-suited for operators A whose spectrum decays to 0 as a power of λ, but they don't work very well in the case of exponentially ill-posed problems in which the spectrum of A decays to 0 exponentially fast. In this case, different source conditions such as logarithmic source conditions should be used, for which analogous results and definitions can be stated. In this work logarithmic and other different source conditions shall not be considered. A deeper treatment of this argument can be found, e.g., in [47].

1.12 Regularization by projection

In practice, regularization methods must be implemented in finite-dimensional spaces, thus it is important to see what happens when the data and the solutions are approximated in finite-dimensional spaces. It turns out that this important passage from infinite to finite dimensions can be seen as a regularization method itself. One approach to deal with this problem is *regularization by projection*, where the approximation is the projection onto finite-dimensional spaces alone: many important regularization methods are included in this category, such as discretization, collocation, Galerkin or Ritz approximation.
The finite-dimensional approximations can be made both in the spaces $\mathcal{X}$ and $\mathcal{Y}$: here we consider only the first one.
We approximate $x^\dagger$ as follows: given a sequence of subspaces of $\mathcal{X}$

$$\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \mathcal{X}_3 ... \tag{1.64}$$

such that

$$\overline{\bigcup_{n\in\mathbb{N}} \mathcal{X}_n} = \mathcal{X}, \tag{1.65}$$

the n-th approximation of $x^\dagger$ is

$$x_n := A_n^\dagger y, \tag{1.66}$$

where $A_n = AP_n$ and P_n is the orthogonal projector onto $\mathcal{X}_n$. Note that since A_n has finite-dimensional range, $\mathcal{R}(A_n)$ is closed and thus $A_n^\dagger$ is bounded (i.e. x_n is a stable approximation of $x^\dagger$). Moreover, it is an easy exercise to show that x_n is the minimum norm solution of $Ax = y$ in $\mathcal{X}_n$: for this reason, this method is called *least-squares projection*.

Note that the iterates x_n may not converge to $x^\dagger$ in $\mathcal{X}$, as the following example due to Seidman shows. We reconsider it entirely in order to correct a small inaccuracy which can be found in the presentation of this example in [17].

1.12.1 The Seidman example (revisited)

Suppose $\mathcal{X}$ is infinite-dimensional, let $\{\mathsf{e}_n\}$ be an orthonormal basis for $\mathcal{X}$ and let $\mathcal{X}_n := span\{\mathsf{e}_1, ...\mathsf{e}_n\}$, for every $n \in \mathbb{N}$. Then of course $\{\mathcal{X}_n\}$ satisfies (1.64) and (1.65). Define an operator $A : \mathcal{X} \to \mathcal{Y}$ as follows:

$$A(x) = A\left(\sum_{j=1}^{\infty} \mathrm{x}_j \mathsf{e}_j\right) := \sum_{j=1}^{\infty} \left(\mathrm{x}_j \mathrm{a}_j + \mathrm{b}_j \mathrm{x}_1\right) \mathsf{e}_j, \tag{1.67}$$

with

$$|\mathrm{b}_j| := \begin{cases} 0 & \text{if } j = 1, \\ j^{-1} & \text{if } j > 1, \end{cases} \tag{1.68}$$

$$|\mathrm{a}_j| := \begin{cases} j^{-1} & \text{if } j \text{ is odd}, \\ j^{-\frac{5}{2}} & \text{if } j \text{ is even}. \end{cases} \tag{1.69}$$

Then:

- A is well defined, since $|\mathrm{a}_j \mathrm{x}_j + \mathrm{b}_j \mathrm{x}_1|^2 \leq 2\left(|\mathrm{x}_j|^2 + |\mathrm{x}_1| j^{-2}\right)$ for every j and linear.

- A is injective: $Ax = 0$ implies

  $$(\mathrm{a}_1\mathrm{x}_1, \mathrm{a}_2\mathrm{x}_2 + \mathrm{b}_2\mathrm{x}_1, \mathrm{a}_3\mathrm{x}_3 + \mathrm{b}_3\mathrm{x}_1, ...) = 0,$$

  thus $\mathrm{x}_j = 0$ for every j, i.e. $x = 0$.

- A is compact: in fact, suppose $\{x_k\}_{k\in\mathbb{N}}$ is a bounded sequence in $\mathcal{X}$. Then also the first components of x_k, denoted by $\mathrm{x}_{k,1}$, form a bounded sequence in $\mathbb{C}$, which has a convergent subsequence $\{\mathrm{x}_{k_l,1}\}$ and, in correspondence, $\sum_{j=1}^{\infty} \mathrm{b}_j\mathrm{x}_{k_l,1}\mathsf{e}_j$ is a subsequence of $\sum_{j=1}^{\infty} \mathrm{b}_j\mathrm{x}_{k,1}\mathsf{e}_j$ converging in $\mathcal{X}$ to $\sum_{j=1}^{\infty} \mathrm{b}_j \lim_{l\to\infty} \mathrm{x}_{k_l,1}\mathsf{e}_j$. Consequently, the application $x \mapsto \sum_{j=1}^{\infty} \mathrm{b}_j\mathrm{x}_1\mathsf{e}_j$ is compact. Moreover, the application $x \mapsto \sum_{j=1}^{\infty} \mathrm{a}_j\mathrm{x}_j\mathsf{e}_j$ is also compact, because it is the limit of the sequence of operators defined by $A_n x := \sum_{j=1}^{n} \mathrm{a}_j\mathrm{x}_j\mathsf{e}_j$. Thus, being the sum of two compact operators, A is compact (see, e.g., [62] for the properties of the compact operators used here).

Let $y := Ax^\dagger$ with

$$x^\dagger := \sum_{j=1}^{\infty} j^{-1}\mathsf{e}_j \tag{1.70}$$

and let $x_n := \sum_{j=1}^{\infty} \mathrm{x}_{n,j}\mathsf{e}_j$ be the best-approximate solution of $Ax = y$ in $\mathcal{X}_n$. Then it is readily seen that $\mathbf{x}_n := (\mathrm{x}_{n,1}, \mathrm{x}_{n,2}, ..., \mathrm{x}_{n,n})$ is the vector minimizing

$$\sum_{j=1}^{n} \big(\mathrm{a}_j(\mathrm{x}_j - j^{-1}) + \mathrm{b}_j(\mathrm{x}_1 - 1)\big)^2 + \sum_{j=n+1}^{\infty} j^{-2}(1 + \mathrm{a}_j - \mathrm{x}_1)^2 \tag{1.71}$$

among all $\mathbf{x} = (\mathrm{x}_1, ..., \mathrm{x}_n) \in \mathbb{C}^n$.
Imposing that the gradient of (1.71) with respect to $\mathbf{x}$ is equal to $\mathbf{0}$ for $\mathbf{x} = \mathbf{x}_n$, we obtain

$$2\mathrm{a}_1^2(\mathrm{x}_{n,1} - 1) - 2\sum_{j=n+1}^{\infty} j^{-2}(1 + \mathrm{a}_j - \mathrm{x}_{n,1}) = 0,$$

$$2\sum_{j=1}^{n} \big(\mathrm{a}_j(\mathrm{x}_{n,j} - j^{-1}) + \mathrm{b}_j(\mathrm{x}_{n,1} - 1)\big)\,\mathrm{a}_j\delta_{j,k}, \quad k = 2, ..., n.$$

Here, $\delta_{j,k}$ is the Kronecker delta, which is equal to 1 if $k = j$ and equal to 0 otherwise.

Consequently, for the first variable $\mathrm{x}_{n,1}$ we have

$$\begin{aligned} \mathrm{x}_{n,1} &= \left(1 + \sum_{j=n+1}^{\infty} (1 + \mathrm{a}_j) j^{-2}\right) \left(1 + \sum_{j=n+1}^{\infty} j^{-2}\right)^{-1} \\ &= 1 + \left(\sum_{j=n+1}^{\infty} \mathrm{a}_j j^{-2}\right) \left(1 + \sum_{j=n+1}^{\infty} j^{-2}\right)^{-1} \end{aligned} \tag{1.72}$$

and for every $k = 2, ..., n$ there holds

$$\mathrm{x}_{n,k} = (\mathrm{a}_k)^{-1} \left(k^{-1}\mathrm{a}_k - \mathrm{b}_k(\mathrm{x}_{n,1} - 1)\right) = k^{-1} - (\mathrm{a}_k k)^{-1}(\mathrm{x}_{n,1} - 1). \tag{1.73}$$

We use this to calculate

$$\begin{aligned} \|x_n - P_n x^\dagger\|^2 &= \| \sum_{j=1}^{n} (\mathrm{x}_{n,j} - j^{-1}) \mathsf{e}_j \|^2 \\ &= (\mathrm{x}_{n,1} - 1)^2 + \sum_{j=2}^{n} \left(j^{-1} - (\mathrm{a}_j j)^{-1}(\mathrm{x}_{n,1} - 1) - j^{-1}\right)^2 \\ &= (\mathrm{x}_{n,1} - 1)^2 \left(1 + \sum_{j=2}^{n} (\mathrm{a}_j j)^{-2}\right) \\ &= \left(\sum_{j=1}^{n} (\mathrm{a}_j j)^{-2}\right) \left(\sum_{j=n+1}^{\infty} \mathrm{a}_j j^{-2}\right)^2 \left(1 + \sum_{j=n+1}^{\infty} j^{-2}\right)^{-2}. \end{aligned} \tag{1.74}$$

Of these three factors, the third one is clearly convergent to 1.

The first one behaves like n^4, since, applying Cesaro's rule to

$$s_n := \frac{\sum_{j=1}^{n} j^3}{n^4},$$

we obtain

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{(n+1)^3}{(n+1)^4 - n^4} = \frac{1}{4}.$$

Similarly,

$$\sum_{j=n+1}^{\infty} \mathrm{a}_j j^{-2} \sim \sum_{j=n+1}^{\infty} j^{-3} \sim n^{-2},$$

because

$$\frac{\sum_{j=n+1}^{\infty} j^{-3}}{n^{-2}} \sim \frac{-(n+1)^{-3}}{(n+1)^{-2}-n^{-2}} = \frac{n^2(n+1)^{-1}}{(n+1)^2-n^2} \to \frac{1}{2}.$$

These calculations show that

$$||x_n - P_n x^\dagger|| \to \lambda > 0, \tag{1.75}$$

so x_n doesn't converge to $x^\dagger$, which was what we wanted to prove.

The following result gives sufficient (and necessary) conditions for convergence.

Theorem 1.12.1. *For $y \in \mathcal{D}(A^\dagger)$, let x_n be defined as above. Then*

(i) $x_n \rightharpoonup x^\dagger \iff \{\|x_n\|\}$ *is bounded;*

(ii) $x_n \to x^\dagger \iff \limsup_{n\to+\infty} \|x_n\| \leq \|x^\dagger\|$*;*

(iii) if

$$\limsup_{n\to+\infty} \|(A_n^\dagger)^* x_n\| = \limsup_{n\to+\infty} \|(A_n^*)^\dagger x_n\| < \infty, \tag{1.76}$$

then

$$x_n \to x^\dagger.$$

For the proof of this theorem and for further results about the least-squares projection method see [17].

1.13 Linear regularization: basic results

In this section we consider a class of regularization methods based on the spectral theory for linear self-adjoint operators.

The basic idea is the following one: let $\{E_\lambda\}$ be the spectral family associated to A^*A. If A^*A is continuously invertible, then $(A^*A)^{-1} = \int \frac{1}{\lambda} dE_\lambda$. Since the best-approximate solution $x^\dagger = A^\dagger y$ can be characterized by the normal equation (1.32), then

$$x^\dagger = \int \frac{1}{\lambda} dE_\lambda A^* y. \tag{1.77}$$

In the ill-posed case the integral in (1.77) does not exist, since the integrand $\frac{1}{\lambda}$ has a pole in 0. The idea is to replace $\frac{1}{\lambda}$ by a parameter-dependent family of functions $g_\sigma(\lambda)$ which are at least piecewise continuous on $[0, \|A\|^2]$ and, for convenience, continuous from the right in the points of discontinuity and to replace (1.77) by

$$x_\sigma := \int g_\sigma(\lambda) dE_\lambda A^* y. \tag{1.78}$$

By construction, the operator on the right-hand side of (1.78) acting on y is continuous, so the approximate solutions

$$x_\sigma^\delta := \int g_\sigma(\lambda) dE_\lambda A^* y^\delta, \tag{1.79}$$

can be computed in a stable way.

Of course, in order to obtain convergence as $\sigma \to 0$, it is necessary to require that $\lim_{\sigma\to 0} g_\sigma(\lambda) = \frac{1}{\lambda}$ for every $\lambda \in (0, \|A\|^2]$.

First, we study the question under which condition the family $\{R_\sigma\}$ with

$$R_\sigma := \int g_\sigma(\lambda) dE_\lambda A^* \tag{1.80}$$

is a regularization operator for $A^\dagger$.

Using the normal equation we have

$$x^\dagger - x_\sigma = x^\dagger - g_\sigma(A^*A)A^*y = (I - g_\sigma(A^*A)A^*A)x^\dagger = \int (1 - \lambda g_\sigma(\lambda)) dE_\lambda x^\dagger. \tag{1.81}$$

Hence if we set, for all (σ, λ) for which $g_\sigma(\lambda)$ is defined,

$$r_\sigma(\lambda) := 1 - \lambda g_\sigma(\lambda), \tag{1.82}$$

so that $r_\sigma(0) = 1$, then

$$x^\dagger - x_\sigma = r_\sigma(A^*A)x^\dagger. \tag{1.83}$$

In these notations, we have the following results.

Theorem 1.13.1. *Let, for all $\sigma > 0$, $g_\sigma : [0, \|A\|^2] \to \mathbb{R}$ fulfill the following assumptions:*

- *g_σ is piecewise continuous;*
- *there exists a constant $C > 0$ such that*
$$|\lambda g_\sigma(\lambda)| \leq C \tag{1.84}$$
for all $\lambda \in (0, \|A\|^2]$;
- $$\lim_{\sigma \to 0} g_\sigma(\lambda) = \frac{1}{\lambda} \tag{1.85}$$
for all $\lambda \in (0, \|A\|^2]$.

Then, for all $y \in \mathcal{D}(A^\dagger)$,

$$\lim_{\sigma \to 0} g_\sigma(A^*A)A^*y = x^\dagger \tag{1.86}$$

and if $y \notin \mathcal{D}(A^\dagger)$, then

$$\lim_{\sigma \to 0} \|g_\sigma(A^*A)A^*y\| = +\infty. \tag{1.87}$$

Remark 1.13.1. *(i) It is interesting to note that for every $y \in \mathcal{D}(A^\dagger)$ the integral $\int g_\sigma(\lambda)dE_\lambda A^*y$ is converging in $\mathcal{X}$, even if $\int \frac{1}{\lambda}dE_\lambda A^*y$ does not exist and $g_\sigma(\lambda)$ is converging pointwise to $\frac{1}{\lambda}$.*

(ii) According to Proposition 1.10.1, *in the assumptions of Theorem* 1.13.1, *$\{R_\sigma\}$ is a regularization operator for $A^\dagger$.*

Another important result concerns the so called *propagation data error* $\|x_\sigma - x_\sigma^\delta\|$:

Theorem 1.13.2. *Let g_σ and C be as in Theorem* 1.13.1, *x_σ and x_σ^δ be defined by* (1.78) *and* (1.79) *respectively. For $\sigma > 0$, let*

$$G_\sigma := \sup\{|g_\sigma(\lambda)| \mid \lambda \in [0, \|A\|^2]\}. \tag{1.88}$$

Then,

$$\|Ax_\sigma - Ax_\sigma^\delta\| \leq C\delta \tag{1.89}$$

and

$$\|x_\sigma - x_\sigma^\delta\| \leq \delta\sqrt{CG_\sigma} \tag{1.90}$$

hold.

Thus the *total error* $\|x^\dagger - x_\sigma^\delta\|$ can be estimated by

$$\|x^\dagger - x_\sigma^\delta\| \leq \|x^\dagger - x_\sigma\| + \delta\sqrt{CG_\sigma}. \tag{1.91}$$

Since $g_\sigma(\lambda) \to \frac{1}{\lambda}$ as $\sigma \to 0$, and it can be proved that the estimate (1.90) is sharp in the usual worst-case sense, the propagated data error generally explodes for fixed $\delta > 0$ as $\sigma \to 0$ (cf. [22]).
We now concentrate on the first term in (1.91), the *approximation error*. While the propagation data error can be studied by estimating $g_\sigma(\lambda)$, for the approximation error one has to look at $r_\sigma(\lambda)$:

Theorem 1.13.3. *Let g_σ fulfill the assumptions of Theorem* 1.13.1. *Let μ, ρ, σ_0 be fixed positive numbers. Suppose there exists a function $\omega_\mu : (0, \sigma_0) \to \mathbb{R}$ such that for every $\sigma \in (0, \sigma_0)$ and every $\lambda \in [0, \|A\|^2]$ the estimate*

$$\lambda^\mu |r_\sigma(\lambda)| \leq \omega_\mu(\sigma) \tag{1.92}$$

is true. Then, for $x^\dagger \in \mathcal{X}_{\mu,\rho}$,

$$\|x_\sigma - x^\dagger\| \leq \rho\, \omega_\mu(\sigma) \tag{1.93}$$

and

$$\|Ax_\sigma - Ax^\dagger\| \leq \rho\, \omega_{\mu+\frac{1}{2}}(\sigma) \tag{1.94}$$

hold.

A straightforward calculation leads immediately to an important consequence:

Corollary 1.13.1. *Let the assumptions of Theorem* 1.13.3 *hold with*

$$\omega_\mu(\sigma) = c\sigma^\mu \tag{1.95}$$

for some constant $c > 0$ and assume that

$$G_\sigma = O\left(\frac{1}{\sigma}\right), \text{ as } \sigma \to 0. \tag{1.96}$$

Then, with the parameter choice rule

$$\alpha \sim \left(\frac{\delta}{\rho}\right)^{\frac{2}{2\mu+1}}, \tag{1.97}$$

the regularization method (R_σ, α) is of optimal order in $\mathcal{X}_{\mu,\rho}$.

1.14 The Discrepancy Principle

So far, we have considered only a-priori choices for the parameter $\alpha = \alpha(\delta)$. Such a-priori choices should be based on some a-priori knowledge of the true solution, namely its smoothness, but unfortunately in practice this information is often not available. This motivates the necessity of looking for a-posteriori parameter choice rules. In this section we will discuss the most famous a-posteriori choice, the *discrepancy principle* (introduced for the first time by Morozov, cf. [67]) and some other important improved choices depending both on the noise level and on the noisy data.

Definition 1.14.1. *Let g_σ be as in Theorem 1.13.1 and such that, for every $\lambda > 0$, $\sigma \mapsto g_\sigma(\lambda)$ is continuous from the left, and let r_σ be defined by (1.82). Fix a positive number τ such that*

$$\tau > \sup\{|r_\sigma(\lambda)| \mid \sigma > 0,\ \lambda \in [0, \|A\|^2]\}. \tag{1.98}$$

For $y \in \mathcal{R}(A)$, the regularization parameter defined via the Discrepancy Principle is

$$\alpha(\delta, y^\delta) := \sup\{\sigma > 0 \mid \|Ax_\sigma^\delta - y^\delta\| \le \tau\delta\}. \tag{1.99}$$

Remark 1.14.1.

- *The idea of the Discrepancy Principle is to choose the biggest parameter for which the corresponding residual has the same order of the noise level, in order to reduce the propagated data error as much as possible.*

- *It is fundamental that $y \in \mathcal{R}(A)$. Otherwise, $\|Ax_\sigma^\delta - y^\delta\|$ can be bounded from below in the following way:*

$$\begin{aligned}\|y - Qy\| - 2\delta &\le \|y - y^\delta\| + \|Q(y^\delta - y)\| + \|y^\delta - Qy^\delta\| - 2\delta \\ &\le \delta + \delta + \|Ax_\sigma^\delta - y^\delta\| - 2\delta = \|Ax_\sigma^\delta - y^\delta\|.\end{aligned} \tag{1.100}$$

  Thus, if δ is small enough, the set

$$\{\sigma > 0 \mid \|Ax_\sigma^\delta - y^\delta\| \le \tau\delta\}$$

  is empty.

- *The assumed continuity from the left for g_σ assures that the functional $\sigma \mapsto \|Ax_\sigma^\delta - y^\delta\|$ is also continuous from the left. Therefore, if $\alpha(\delta, y^\delta)$ satisfies the Discrepancy Principle* (1.99)*, we have*
$$\|Ax^\delta_{\alpha(\delta,y^\delta)} - y^\delta\| \leq \tau\delta. \tag{1.101}$$
- *If $\|Ax_\sigma^\delta - y^\delta\| \leq \tau\delta$ holds for every $\sigma > 0$, then $\alpha(\delta, y^\delta) = +\infty$ and $x^\delta_{\alpha(\delta,y^\delta)}$ has to be understood in the sense of a limit as $\alpha \to +\infty$.*

The main convergence properties of the Discrepancy Principle are described in the following important theorem (see [17] for the long proof). A full understanding of the statement of the theorem requires the notions of saturation and qualification of a regularization method.
The term *saturation* is used to describe the behavior of some regularization operators for which
$$\|x_\sigma^\delta - x^\dagger\| = O(\delta^{\frac{2\mu}{2\mu+1}}) \tag{1.102}$$
does not hold for every μ, but only up to a finite value μ_0, called the *qualification* of the method.
More precisely, the qualification μ_0 is defined as the largest value such that
$$\lambda^\mu |r_\sigma(\lambda)| = O(\sigma^\mu) \tag{1.103}$$
holds for every $\mu \in (0, \mu_0]$.

Theorem 1.14.1. *In addition to the assumptions made for g_σ in Definition* 1.14.1*, suppose that there exists a constant $\tilde{c}$ such that G_σ satisfies*
$$G_\sigma \leq \frac{\tilde{c}}{\sigma}, \tag{1.104}$$
for every $\sigma > 0$. Assume also that the regularization method (R_σ, α) corresponding to the Discrepancy Principle has qualification $\mu_0 > \frac{1}{2}$ and that, with ω_μ defined as in Theorem 1.13.3*,*
$$\omega_\mu(\alpha) \sim \alpha^\mu, \quad \textit{for } 0 < \mu \leq \mu_0. \tag{1.105}$$
Then (R_σ, α) is convergent for every $y \in \mathcal{R}(A)$ and of optimal order in $\mathcal{X}_{\mu,\rho}$, for $\mu \in (0, \mu_0 - \frac{1}{2}]$ and for $\rho > 0$.

Thus, in general, a regularization method (R_σ, α) with α defined via the Discrepancy Principle need not be of optimal order in $\mathcal{X}_{\mu,\rho}$ for $\mu > \mu_0 - \frac{1}{2}$, as the following result for the Tikhonov method in the compact case implies:

Theorem 1.14.2 (**Groetsch**). *Let $K = A$ be compact, define $R_\sigma := (K^*K + \sigma)^{-1}K^*$ and choose the Discrepancy Principle* (1.99) *as a parameter choice rule for R_σ. If*

$$\|x^\delta_{\alpha(\delta,y^\delta)} - x^\dagger\| = o(\delta^{\frac{1}{2}}) \tag{1.106}$$

holds for every $y \in \mathcal{R}(K)$ and $y^\delta \in \mathcal{Y}$ fulfilling $\|y^\delta - y\| \leq \delta$, then $\mathcal{R}(K)$ is finite-dimensional.

Consequently, since

- $\mu_0 = 1$ for (R_σ, α) as in Theorem 1.14.2 (cf. the results in Section 1.16) and
- in the ill-posed case $\|x^\delta_{\alpha(\delta,y^\delta)} - x^\dagger\|$ does not converge faster than $O(\delta^{\frac{1}{2}})$,

(R_σ, α) cannot be of optimal order in $\mathcal{X}_{\mu,\rho}$ for $\mu > \mu_0 - \frac{1}{2}$.
This result is the motivation and the starting point for the introduction of other (improved) a-posteriori parameter choice rules defined to overcome the problem of saturation. However, we are interested mainly in iterative methods, where these rules are not needed, so we address the interested reader to [17] for more details about such rules. There, also a coverage of some of the most important heuristic parameter choices rules can be found.

1.15 The finite dimensional case: discrete ill-posed problems

In practice, ill-posed problems like integral equations of the first kind have to be approximated by a finite dimensional problem whose solution can be found by a software.
In the finite dimensional setting, the linear operator A is simply a matrix $\mathbf{A}$

$\in \mathbb{M}_{m,N}(\mathbb{R})$, the Moore Penrose Generalized Inverse $\mathbf{A}^{\dagger}$ is defined for every data $\mathbf{b} \in \mathcal{Y} = \mathbb{R}^m$ and being a linear map from $\mathbb{R}^m$ to $\ker(\mathbf{A})^{\perp} \subseteq \mathcal{X} = \mathbb{R}^N$ is continuous. Thus, according to Hadamard's definition, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ cannot be ill-posed. However, from a practical point of view, a theoretically well-posed problem can be very similar to an ill-posed one. To explain this, recall that a linear operator A is bounded if and only if there exists a constant $C > 0$ such that $\|Ax\| \leq C\|x\|$ for every $x \in \mathcal{X}$: if the constant C is too big, then this estimate is virtually useless and little perturbations in the data can generate very huge errors in the results. This concern should be even more serious if one takes into account also rounding errors due to finite arithmetics calculations. Such finite dimensional problems occur very often in practice and they are characterized by very ill-conditioned matrices.

In his book [36], P.C. Hansen distinguishes between two classes of problems where the matrix of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is highly ill-conditioned: *rank deficient* and *discrete ill-posed problems.*

In a rank deficient problem, the matrix $\mathbf{A}$ has a cluster of small eigenvalues and a well determined gap between its large and small singular values. Discrete ill-posed problems arise from the discretization of ill-posed problems such as integral equations of the first kind and their singular values typically decay gradually to zero. Although of course we shall be more interested in discrete ill-posed problems, we should keep in mind that regularization methods can also be applied with success on rank deficient problems and therefore should be considered also from this point of view.

As we have seen in Example 1.10.1, the process of discretization of an ill-posed problem is indeed a regularization itself, since it can be considered as a projection method. However, as a matter of fact, usually the regularizing process of discretization is not enough to obtain a good approximation of the exact solution and it is necessary to apply other regularization methods.

Here, we will give a very brief survey about the discretization of integral equations of the first kind. More details can be found for example in [61] (Chapter 3), in [3] and [14].

There are essentially two classes of methods for discretizing integral equations such as (1.11): quadrature methods and Galerkin methods.

In a quadrature method, one chooses m samples $f(s_i)$, $i = 1, ..., m$ of the function $f(s)$ and uses a quadrature rule with abscissas $t_1, t_2, ..., t_N$ and weights $\omega_1, \omega_2, ...\omega_N$ to calculate the integrals

$$\int_{s_1}^{s_2} \varkappa(s_i, t)\phi(t)dt \sim \sum_{j=1}^{N} \omega_j \varkappa(s_i, t_j)\phi(t_j), \quad i = 1, ..., m.$$

The result is a linear system of the type $\mathbf{Ax} = \mathbf{b}$, where the components of the vector $\mathbf{b}$ are the samples of f, the elements of the matrix $\mathbf{A} \in \mathbb{M}_{m,N}(\mathbb{R})$ are defined by $a_{i,j} = \omega_j \varkappa(s_i, t_j)$ and the unknowns x_j forming the vector $\mathbf{x}$ correspond to the values of ϕ at the abscissas t_j.

In a Galerkin method, it is necessary to fix two finite dimensional subspaces $\mathcal{X}_N \subseteq \mathcal{X}$ and $\mathcal{Y}_m \subseteq \mathcal{Y}$, $\dim(\mathcal{X}_N)= N$, $\dim(\mathcal{Y}_m)= m$ and define two corresponding sets of orthonormal basis functions ϕ_j, $j = 1, ..., N$ and ψ_i, $i = 1, ..., m$. Then the matrix and the right-hand side elements of the system $\mathbf{Ax} = \mathbf{b}$ are given by

$$a_{i,j} = \int\int_{[s_1,s_2]^2} \varkappa(s, t)\psi_i(s)\phi_j(t)dsdt \quad \text{and} \quad b_i = \int_{s_1}^{s_2} f(s)\psi_i(s)ds \quad (1.107)$$

and the unknown function ϕ depends on the solutions of the system via the formula $\phi(t) = \sum_{j=1}^{N} \mathrm{x}_j \phi_j(t)$.

If $\varkappa$ is symmetric, $\mathcal{X} = \mathcal{Y}$, $m = N$, $\mathcal{X}_N = \mathcal{Y}_N$ and $\phi_i = \psi_i$ for every i, then the matrix $\mathbf{A}$ is also symmetric and the Galerkin method is called the Rayleigh-Ritz method.

A special case of the Galerkin method is the *least-squares collocation* or *moment discretization*: it is defined for integral operators K with continuous kernel and the delta functions $\delta(s - s_i)$ as the basis functions ψ_i. In [17] it is shown that least-squares collocation is a particular projection method of the type described in Section 1.12 in which the projection is made on the space $\mathcal{Y}$ and therefore a regularization method itself.

For discrete ill-posed problems, we have already noticed that the Picard

Criterion is always satisfied. However, it is possible to state a *Discrete Picard Criterion* as follows (cf. [32] and [36]).

Definition 1.15.1 (Discrete Picard condition). *Fix a singular value decomposition of the matrix* $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$ *where* $\mathbf{U}$ *and* $\mathbf{V}$ *are constituted by the singular vectors of* $\mathbf{A}$ *thought as column vectors. The unperturbed right-hand side* $\mathbf{b}$ *in a discrete ill-posed problem satisfies the discrete Picard condition if the* SVD *coefficients* $|\langle \mathbf{u}_j, \mathbf{b}\rangle|$ *on the average decay to zero faster than the singular values* λ_j.

Unfortunately, the SVD coefficients may have a non-monotonic behavior, thus it is difficult to give a precise definition.

For many discrete ill-posed problems arising from integral equations of the first kind the Discrete Picard Criterion is satisfied for exact data. In general, it is not satisfied when the data is perturbed by the noise.

We shall return to this argument later on and we will see how the plot of the SVD coefficients may help to understand the regularizing properties of some regularization methods.

1.16 Tikhonov regularization

The most famous regularization method was introduced by A.N. Tikhonov in 1963 (cf. [90], [91]).

In the linear case, it fits the general framework of Section 1.13 and fulfills the assumptions of Theorem 1.13.1 with

$$g_\sigma(\lambda) := \frac{1}{\lambda + \sigma}. \tag{1.108}$$

The regularization parameter σ stabilizes the computation of the approximate solutions

$$x_\sigma^\delta = (A^*A + \sigma)^{-1}A^*y^\delta, \tag{1.109}$$

which can therefore be defined by the following regularized version of the normal equation:

$$A^*Ax_\sigma^\delta + \sigma x_\sigma^\delta = A^*y^\delta. \tag{1.110}$$

Tikhonov regularization can be studied from a variational point of view, which is the key to extend it to nonlinear problems:

Theorem 1.16.1. *Let x_σ be as in* (1.109)*. Then x_σ^δ is the unique minimizer of the Tikhonov functional*

$$x \mapsto \|Ax - y^\delta\|^2 + \sigma\|x\|^2. \tag{1.111}$$

As an illustrative example, we calculate the functions defined in the previous chapter in the case of Tikhonov regularization.

- Remembering that $g_\sigma(\lambda) = \dfrac{1}{\lambda + \sigma}$, we obtain immediately that

$$G_\sigma = \frac{1}{\sigma} \quad \text{and} \quad r_\sigma(\lambda) = 1 - g_\sigma(\lambda) = \frac{\sigma}{\sigma + \lambda}. \tag{1.112}$$

- The computation of $\omega_\mu(\sigma)$ requires an estimate for the function

$$h_\mu(\sigma) := \lambda^\mu \frac{\sigma}{\lambda + \sigma}. \tag{1.113}$$

Calculating the derivative of h_μ brings to

$$h_\mu'(\lambda) = r_\sigma(\lambda)\lambda^{\mu-1}(\mu - \frac{\lambda}{\lambda + \sigma}), \tag{1.114}$$

thus if $\mu < 1$ h_μ has a maximum for $\lambda = \frac{\sigma\mu}{1-\mu}$ and we obtain

$$h_\mu(\sigma) \leq \mu^\mu (1-\mu)^{1-\mu} \sigma^\mu, \tag{1.115}$$

whereas $h_\mu'(\lambda) > 0$ for $\mu \geq 1$, so h_μ assumes its maximum for $\lambda = \|A\|^2$. Putting this together, we conclude that for ω_μ we can take

$$\omega_\mu(\sigma) = \begin{cases} \sigma^\mu, & \mu \leq 1 \\ c\sigma, & \mu > 1 \end{cases} \tag{1.116}$$

with c=$\|A\|^{2\mu-2}$.

The results of Section 1.13 can thus be applied to Tikhonov regularization: in particular, according to Corollary 1.13.1, as long as $\mu \leq 1$ Tikhonov regularization with the parameter choice rule (1.97) is of optimal order in $\mathcal{X}_{\mu,\rho}$ and the best possible convergence rate, obtained when $\mu = 1$ and $\alpha \sim \left(\frac{\delta}{\rho}\right)^{\frac{2}{3}}$, is given by

$$\|x_\alpha^\delta - x^\dagger\| = O(\delta^{\frac{2}{3}}) \tag{1.117}$$

for $x^\dagger \in \mathcal{X}_{1,\rho}$.

Due to the particular form of the function ω_μ found in (1.116), the Tikhonov method saturates, since (1.103) holds only for $\mu \leq 1$. Thus Tikhonov regularization has finite qualification $\mu_0 = 1$. A result similar to Theorem 1.14.2 can be proved (see [17] or [22]).

Theorem 1.16.2. *Let K be compact with infinite-dimensional range, x_σ^δ be defined by* (1.109) *with K instead of A. Let $\alpha = \alpha(\delta, y^\delta)$ be any parameter choice rule. Then*

$$\sup\{\|x_\alpha^\delta - x^\dagger\| \ : \ \|Q(y - y^\delta)\| \leq \delta\} = o(\delta^{\frac{2}{3}}) \tag{1.118}$$

implies $x^\dagger = 0$.

The Tikhonov regularization method was also studied on convex subsets of the Hilbert space $\mathcal{X}$. This can be of particular interest in certain applications such as image deblurring where we can take $\mathcal{X} = \mathcal{L}^2([0,1]^2)$ and the solution lies in the convex set $\mathcal{C} := \{x \in \mathcal{X} \mid x \geq 0\}$. A quick treatment of the argument can be found in [17] and for details we suggest [71].

The Tikhonov method is now well understood, but has some drawbacks:

1. It is quite expensive from a computational point of view, since it requires an inversion of the operator $A^*A + \sigma$.

2. For every choice of the regularization parameter the operator to be inverted in the formula (1.109) changes, thus if α is chosen in the wrong way the computations should be restarted.

3. As a matter of fact, Tikhonov regularization calculates a smooth solution.

4. It has finite qualification.

The third drawback implies that Tikhonov regularization may not work very well if one looks for irregular solutions. For this reason, in certain problems such as image processing nowadays many researchers prefer to rely on other methods based on the minimization of a different functional. More precisely, in the objective function (1.111), $\|x\|^2$ is replaced by a term that takes into account the nature of the sought solution $x^\dagger$. For example, one can choose a version of the total variation of x, which often provides very good practical results in imaging (cf. e.g. [96]).

The fourth problem can be overcome by using a variant of the algorithm known as iterative Tikhonov regularization (cf. e.g. [17]).

The points 1 and 2 are the main reasons why we prefer iterative methods to Tikhonov regularization. Nevertheless, the Tikhonov method is still very popular. In fact, it can be combined with different methods, works well in certain applications (e.g. when the sought solution is smooth) and it remains one of the most powerful weapon against ill posed problems in the nonlinear case.

1.17 The Landweber iteration

A different way of regularizing an ill-posed problem is the approach of the iterative methods: consider the direct operator equation (1.20), i.e. the problem of calculating y from x and A. If the computation of y is easy and reasonably cheap, the iterative methods form a sequence of iterates $\{x_k\}$ based on the direct solution of (1.20) that x_k converges to $x^\dagger$.

It turns out that for many iterative methods the iteration index k plays the role of the regularization parameter σ and that these methods can be studied from the point of view of the regularization theory developed in the previous

sections.
When dealing with perturbed data y^δ, in order to use regularization techniques, one has to write the iterates x_k^δ in terms of an operator R_k (and of course of y^δ). Now, R_k may depend or not on y^δ itself. If it doesn't, then the resulting method fits completely the general theory of linear regularization: this is the case of the Landweber iteration, the subject of the present section. Otherwise, some of the basic assumptions of the regularization theory may fail to be true (e.g. the operator mapping y^δ to x_k^δ may be not continuous): conjugate gradient type methods, which will we discussed in detail in the next chapter, fit into this category.
In spite of the difficulties arising from this problem, in practice the conjugate gradient type methods are known as a very powerful weapon against ill-posed problems, whereas the Landweber iteration has the drawback of being a very slow method and is mainly used in nonlinear problems. For this reason, in this section we give only an outline of the main properties of the Landweber iteration and skip most of the proofs.
The idea of the Landweber method is to approximate $A^\dagger y$ with a sequence of iterates $\{x_k\}_{k\in\mathbb{N}}$ transforming the normal equation (1.32) into equivalent fixed point equations like

$$x = x + A^*(y - Ax) = (I - A^*A)x. \tag{1.119}$$

In practice, one is given the perturbed data y^δ instead of y and defines the Landweber iterates as follows:

Definition 1.17.1 (Landweber Iteration). *Fix an initial guess $x_0^\delta \in \mathcal{X}$ and for $k = 1, 2, 3, ...$ compute the Landweber approximations recursively via the formula*

$$x_k^\delta = x_{k-1}^\delta + A^*(y^\delta - Ax_{k-1}^\delta). \tag{1.120}$$

We observe that in the definition of the Landweber iterations we can suppose without loss of generality $\|A\| \le 1$. If this were not the case, then we would introduce a relaxation parameter ω with $0 < \omega \le \|A\|^{-2}$ in front of A^*, i.e. we would iterate

$$x_k^\delta = x_{k-1}^\delta + \omega A^*(y^\delta - Ax_{k-1}^\delta), \quad \text{k } \in \mathbb{N}. \tag{1.121}$$

In other words, we would multiply the equation $Ax = y^\delta$ by $\omega^{\frac{1}{2}}$ and iterate with (1.120).

Moreover, if $\{z_k^\delta\}$ is the sequence of the Landweber iterates with initial guess $z_0^\delta = 0$ and data $y^\delta - Ax_0^\delta$, then $x_k^\delta = x_0^\delta + z_k^\delta$, so that supposing $x_0^\delta = 0$ is no restriction too.

Since we have assumed $\|A\|^2 = 1 < 2$, then $I - A^*A$ is nonexpansive and one may apply the method of successive approximations. It is important to note that in the ill-posed case $I - A^*A$ is no contraction, because the spectrum of A^*A clusters at the origin. For example, if A is compact, then there exists a sequence $\{\lambda_n\}$ of eigenvalues of A^*A such that $|\lambda_n| \to 0$ as $n \to +\infty$ and for a corresponding sequence of eigenvectors $\{v_n\}$ one has

$$\frac{\|(I - A^*A)v_n\|}{\|v_n\|} = \frac{\|(1 - \lambda_n)v_n\|}{\|v_n\|} = |1 - \lambda_n| \longrightarrow 1 \text{ as } n \to +\infty,$$

i.e. $\|(I - A^*A)\| \geq 1$.

Despite this, already in 1956 in his work [63], Landweber was able to prove the following strong convergence result in the case of compact operators (our proof, taken from [17] makes use of the regularization theory and is valid in the general case of linear and bounded operators).

Theorem 1.17.1. *If $y \in \mathcal{D}(A^\dagger)$, then the Landweber approximations x_k corresponding to the exact data y converge to $A^\dagger y$ as $k \to +\infty$. If $y \notin \mathcal{D}(A^\dagger)$, then $\|x_k\| \to +\infty$ as $k \to +\infty$.*

Proof. By induction, the iterates x_k may be expressed in the form

$$x_k = \sum_{j=0}^{k-1} (I - A^*A)^j A^* y. \tag{1.122}$$

Suppose now $y \in \mathcal{D}(A^\dagger)$. Then since $A^*y = A^*Ax^\dagger$, we have

$$x^\dagger - x_k = x^\dagger - \sum_{j=0}^{k-1}(I - A^*A)^j A^*Ax^\dagger = x^\dagger - A^*A\sum_{j=0}^{k-1}(I - A^*A)^j x^\dagger = (I - A^*A)^k x^\dagger.$$

Thus we have found the functions g and r of Section 1.13:

$$\begin{aligned} g_k(\lambda) &= \sum_{j=0}^{k-1}(1-\lambda)^j, \\ r_k(\lambda) &= (1-\lambda)^k. \end{aligned} \tag{1.123}$$

Since $\|A\| \leq 1$ we consider $\lambda \in (0,1]$: in this interval $\lambda g_k(\lambda) = 1 - r_k(\lambda)$ is uniformly bounded and $g_k(\lambda)$ converges to $\frac{1}{\lambda}$ as $k \to +\infty$ because $r_k(\lambda)$ converges to 0. We can therefore apply Theorem 1.13.1 to prove the assertion, with k^{-1} playing the role of σ. □

The theorem above states that the approximation error of the Landweber iterates converges to zero. Next we examine the behavior of the propagated data error: on the one hand, according to the same theorem, if $y^\delta \notin \mathcal{D}(A^\dagger)$ the iterates x_k must diverge; on the other hand, the operator R_k defined by

$$R_k y^\delta := x_k^\delta \tag{1.124}$$

is equal to $\sum_{j=0}^{k-1}(I - A^*A)^j A^*$. Therefore, for fixed k, x_k^δ depends continuously on the data so that the propagation error cannot be arbitrarily large.
This leads to the following result.

Proposition 1.17.1. *Let y, y^δ be a pair of right-hand side data with $\|y - y^\delta\| \leq \delta$ and let $\{x_k\}$ and $\{x_k^\delta\}$ be the corresponding two iteration sequences. Then we have*

$$\|x_k - x_k^\delta\| \leq \sqrt{k}\delta, \quad k \geq 0. \tag{1.125}$$

Remark 1.17.1. *According to the previous results, the total error has two components, an approximation error converging (slowly) to 0 and the propagated data error of the order at most $\sqrt{k}\delta$. For small values of k the data error is negligible and the total error seems to converge to the exact solution $A^\dagger y$, but when $\sqrt{k}\delta$ reaches the order of magnitude of the approximation error, the data error is no longer hidden in x_k^δ and the total error begins to increase and eventually diverges.*

The phenomenon described in Remark 1.17.1 is called semi-convergence and is very typical of iterative methods for solving inverse problems. Thus the regularizing properties of iterative methods for ill-posed problems ultimately depend on a reliable stopping rule for detecting the transient from convergence to divergence: the iteration index k plays the part of the regularization parameter σ and the stopping rule is the counterpart of the parameter choice rule for continuous regularization methods.
As an a-posteriori stopping rule, a discrete version of the Discrepancy Principle can be considered:

Definition 1.17.2. *Fix $\tau > 1$. For $k = 0, 1, 2, ...$ let x_k^δ be the k-th iterate of an iterative method for solving $Ax = y$ with perturbed data y^δ such that $\|y - y^\delta\| \leq \delta$, $\delta > 0$. The stopping index $k_D = k_D(\delta, y^\delta)$ corresponding to the Discrepancy Principle is the biggest k such that*

$$\|y^\delta - Ax_k^\delta\| \leq \tau\delta. \tag{1.126}$$

Of course, one should prove that the stopping index k_D is well defined, i.e. there is a finite index such that the residual $\|y^\delta - Ax_k^\delta\|$ is smaller than the tolerance $\tau\delta$.
In the case of the Landweber iteration, we observe that the residual can be written in the form

$$y^\delta - Ax_k^\delta = y - Ax_{k-1}^\delta - AA^*(y^\delta - Ax_{k-1}^\delta) = (I - AA^*)(y^\delta - Ax_{k-1}^\delta),$$

hence the non-expansivity of $I - AA^*$ implies that the residual norm is monotonically decreasing. However, this is not enough to show that the discrepancy principle is well defined. For this, a more precise estimate of the residual norm is needed. This estimate can be found (cf. [17], Proposition 6.4) and leads to the following result.

Proposition 1.17.2. *The Discrepancy Principle defines a finite stopping index $k_D(\delta, y^\delta)$ with $k_D(\delta, y^\delta) = O(\delta^{-2})$.*

The regularization theory can be used to prove the order optimality of the Landweber iteration with the Discrepancy Principle as a stopping rule:

Theorem 1.17.2. *For fixed $\tau > 1$, the Landweber iteration with the discrepancy principle is convergent for every $y \in \mathcal{R}(A)$ and of optimal order in $\mathcal{X}_\mu$, for every $\mu > 0$.*
Moreover, if $A^\dagger y \in \mathcal{X}_\mu$, then $k_D(\delta, y^\delta) = O(\delta^{-\frac{2}{2\mu+1}})$ and the Landweber method has qualification $\mu_0 = +\infty$.

As mentioned earlier, the main problem of the Landweber iteration is that in practice it requires too many iterations until the stopping criterion is met. Another stopping rule has been proposed in [15], but it requires a similar number of iterations. In fact, it can be shown that the exponent $\frac{2}{2\mu+1}$ cannot be improved in general. However, it is possible to reduce it to $\frac{1}{2\mu+1}$ if more sophisticated methods are used. Such accelerated Landweber methods are the so called semi-iterative methods (with the ν-methods as significant examples): they will not be treated here, since we will focus our attention in greater detail on the conjugate gradient type methods, which are considered quicker (or at least not slower) and more flexible than the accelerated Landweber methods. For a complete coverage of the semi-iterative methods, see [17].

Chapter 2

Conjugate gradient type methods

This chapter is entirely dedicated to the conjugate gradient type methods. These methods are mostly known for being fast and robust solvers of large systems of linear equations: for example, the classical Conjugate Gradient method (CG), introduced for the first time by Hestenes and Stiefel in 1952 (see [45]), finds the exact solution of a linear system with a positive definite $N \times N$ matrix in at most N iterative steps, cf. Theorem 2.1.1 below. For this reason, the importance of these methods goes far beyond the regularization of ill-posed problems, although here they will be studied mainly from this particular point of view.

One can approach the conjugate gradient type methods in many different ways: it is possible to see them as optimization methods or as projection methods. Alternatively, one can study them from the point of view of the orthogonal polynomials. In each case, the *Krylov spaces* are fundamental in the definition of the conjugate gradient type methods, so that they are often regarded as *Krylov methods*.

Definition 2.0.3. *Let $\mathcal{V}$ be a vector space and let A be a linear map from $\mathcal{V}$ to itself. For a given vector $x_0 \in \mathcal{V}$ and for $k \in \mathbb{N}$, the k-th Krylov space* *(based on x_0) is the linear subspace of $\mathcal{V}$ defined by*

$$\mathcal{K}_{k-1}(A; x_0) := span\{x_0, Ax_0, A^2x_0, ..., A^{k-1}x_0\}. \tag{2.1}$$

A Krylov method selects the k-th iterative approximate solution x_k of $x^\dagger$ as an element of a certain Krylov space depending on A and x_0 satisfying certain conditions.

In particular, a conjugate gradient type method chooses the minimizer of a particular function in the shifted space $x_0 + \mathcal{K}_{k-1}(A; y - Ax_0)$ with respect to a particular measure.

We will introduce the subject in a finite dimensional setting with an optimization approach, but in order to understand the regularizing properties of the algorithms in the general framework of Chapter 1 the main analysis will be developed in Hilbert spaces using orthogonal polynomials. The main reference for this chapter is the book of M. Hanke [27]. For the finite dimensional introduction, we will follow [59].

2.1 Finite dimensional introduction

For $N \in \mathbb{N}$ we denote by

$$\langle \cdot, \cdot \rangle \ : \ \mathbb{R}^N \times \mathbb{R}^N \longrightarrow \mathbb{R} \tag{2.2}$$

the standard scalar product on $\mathbb{R}^N$ inducing the euclidean norm $\| \cdot \|$.

For a matrix $\mathbf{A} \in \mathbb{M}_{m,N}(\mathbb{R})$, $m \in \mathbb{N}$, $\|\mathbf{A}\|$ denotes the norm of $\mathbf{A}$ as a linear operator from $\mathbb{R}^N$ to $\mathbb{R}^m$.

For notational convenience, here and below a vector $\mathbf{x} \in \mathbb{R}^N$ will be thought as a column vector $\mathbf{x} \in \mathbb{M}_{N,1}(\mathbb{R})$, thus $\mathbf{x}^*$ will be the row vector transposed of $\mathbf{x}$.

We consider the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{2.3}$$

with $\mathbf{A} \in \mathbb{GL}_N(\mathbb{R})$ symmetric and positive definite, $\mathbf{b} \in \mathbb{R}^N$, $N >> 1$.

Definition 2.1.1. *The conjugate gradient method for solving* (2.3) *generates a sequence* $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ *in* $\mathbb{R}^N$ *such that for each* k *the* k*-th iterate* $\mathbf{x}_k$ *minimizes*

$$\phi(\mathbf{x}) := \frac{1}{2}\mathbf{x}^*\mathbf{A}\mathbf{x} - \mathbf{x}^*\mathbf{b} \tag{2.4}$$

on $\mathbf{x}_0 + \mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$*, with* $\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$.

Of course, when the minimization is made on the whole space, then the minimizer is the exact solution $\mathbf{x}^\dagger$.

Due to the assumptions made on the matrix $\mathbf{A}$, there are an orthogonal matrix $\mathbf{U} \in \mathbb{O}_N(\mathbb{R})$ and a diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, ..., \lambda_N\}$, with $\lambda_i > 0$ for every $i = 1, ..., N$, such that

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^* \tag{2.5}$$

and (2.5) can be used to define on $\mathbb{R}^N$ the so called $\mathbf{A}$-norm

$$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\mathbf{x}^*\mathbf{A}\mathbf{x}}. \tag{2.6}$$

It turns out that the minimization property of $\mathbf{x}_k$ can be read in terms of this norm:

Proposition 2.1.1. *If* $\Omega \subseteq \mathbb{R}^N$ *and* $\mathbf{x}_k$ *minimizes the function* ϕ *on* Ω*, then it minimizes also* $\|\mathbf{x}^\dagger - \mathbf{x}\|_{\mathbf{A}} = \|\mathbf{r}\|_{\mathbf{A}^{-1}}$ *on* Ω*, with* $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$.

Proof. Since $\mathbf{A}\mathbf{x}^\dagger = \mathbf{b}$ and $\mathbf{A}$ is symmetric, we have

$$\begin{aligned}\|\mathbf{x}^\dagger - \mathbf{x}\|_{\mathbf{A}}^2 &= (\mathbf{x}^\dagger - \mathbf{x})^*\mathbf{A}(\mathbf{x}^\dagger - \mathbf{x}) = \mathbf{x}^*\mathbf{A}\mathbf{x} - \mathbf{x}^*\mathbf{A}\mathbf{x}^\dagger - (\mathbf{x}^\dagger)^*\mathbf{A}\mathbf{x} + (\mathbf{x}^\dagger)^*\mathbf{A}\mathbf{x}^\dagger \\ &= \mathbf{x}^*\mathbf{A}\mathbf{x} - 2\mathbf{x}^*\mathbf{b} + (\mathbf{x}^\dagger)^*\mathbf{A}\mathbf{x}^\dagger = 2\phi(\mathbf{x}) + (\mathbf{x}^\dagger)^*\mathbf{A}\mathbf{x}^\dagger.\end{aligned} \tag{2.7}$$

Thus the minimization of ϕ is equivalent to the minimization of $\|\mathbf{x}^\dagger - \mathbf{x}\|_{\mathbf{A}}^2$ (and consequently of $\|\mathbf{x}^\dagger - \mathbf{x}\|_{\mathbf{A}}$).

Moreover, using again the symmetry of $\mathbf{A}$,

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}^\dagger\|_{\mathbf{A}}^2 &= (\mathbf{A}(\mathbf{x} - \mathbf{x}^\dagger))^*\mathbf{A}^{-1}(\mathbf{A}(\mathbf{x} - \mathbf{x}^\dagger)) = (\mathbf{A}\mathbf{x} - \mathbf{b})^*\mathbf{A}^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{\mathbf{A}^{-1}}^2\end{aligned} \tag{2.8}$$

and the proof is complete. □

Remark 2.1.1. *Proposition* 2.1.1 *has the following consequences:*

1. *The k-th iterate of* CG *minimizes the the approximation error*

$$\boldsymbol{\varepsilon}_k := \mathbf{x}_k - \mathbf{x}^\dagger$$

in the $\mathbf{A}$*-norm in the shifted Krylov space* $\mathbf{x}_0 + \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{r}_0)$.
Since a generic element $\breve{\mathbf{x}}$ *of* $\mathbf{x}_0 + \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{r}_0)$ *can be written in the form*

$$\breve{\mathbf{x}} = \mathbf{x}_0 + \sum_{j=0}^{k-1} \gamma_j \mathbf{A}^j \mathbf{r}_0 = \mathbf{x}_0 + \sum_{j=0}^{k-1} \gamma_j \mathbf{A}^{j+1}(\mathbf{x}^\dagger - \mathbf{x}_0)$$

for some coefficients $\gamma_0, ..., \gamma_{k-1}$, *if we define the polynomials*

$$\begin{aligned} \mathsf{q}_{k-1}(\lambda) &:= \sum_{j=0}^{k-1} \gamma_j \lambda^j, \\ \mathsf{p}_k(\lambda) &:= 1 - \lambda \mathsf{q}_{k-1}(\lambda), \end{aligned} \tag{2.9}$$

we obtain that

$$\begin{aligned} \mathbf{x}^\dagger - \breve{\mathbf{x}} &= \mathbf{x}^\dagger - \mathbf{x}_0 - \mathsf{q}_{k-1}(\mathbf{A})\mathbf{r}_0 = \mathbf{x}^\dagger - \mathbf{x}_0 - \mathsf{q}_{k-1}(\mathbf{A})\mathbf{A}(\mathbf{x}^\dagger - \mathbf{x}_0) \\ &= \mathsf{p}_k(\mathbf{A})(\mathbf{x}^\dagger - \mathbf{x}_0). \end{aligned} \tag{2.10}$$

Hence the minimization property of $\mathbf{x}_k$ *can also be written in the form*

$$\|\mathbf{x}^\dagger - \mathbf{x}_k\|_{\mathbf{A}} = \min_{\mathsf{p} \in \Pi_k^0} \|\mathsf{p}(\mathbf{A})(\mathbf{x}^\dagger - \mathbf{x}_0)\|_{\mathbf{A}}, \tag{2.11}$$

where Π_k^0 *is the the set of all polynomials* p *of degree equal to* k *such that* $\mathsf{p}(0) = 1$.

2. *For every* $\mathsf{p} \in \Pi_k := \{$*polynomials of degree* $k\}$ *one has*

$$\mathsf{p}(\mathbf{A}) = \mathbf{U}\mathsf{p}(\boldsymbol{\Lambda})\mathbf{U}^*.$$

Moreover, the square root of $\mathbf{A}$ *is well defined by* $\mathbf{A}^{\frac{1}{2}} := \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^*$, *with* $\boldsymbol{\Lambda}^{\frac{1}{2}} := \mathrm{diag}\{\lambda_1^{\frac{1}{2}}, ..., \lambda_N^{\frac{1}{2}}\}$ *and immediately there follows*

$$\|\mathbf{x}\|_{\mathbf{A}}^2 = \|\mathbf{A}^{\frac{1}{2}}\mathbf{x}\|^2, \;\; \mathbf{x} \in \mathbb{R}^N. \tag{2.12}$$

Consequently, since the norm of a symmetric, positive definite matrix is equal to its largest eigenvalue, we easily get:

$$\|\mathsf{p}(\mathbf{A})\mathbf{x}\|_{\mathbf{A}} = \|\mathsf{p}(\mathbf{A})\mathbf{A}^{\frac{1}{2}}\mathbf{x}\| \leq \|\mathsf{p}(\mathbf{A})\|\|\mathbf{x}\|_{\mathbf{A}}, \ \forall \mathbf{x} \in \mathbb{R}^N, \ \forall \mathsf{p} \in \Pi_k, \tag{2.13}$$

$$\|\mathbf{x}^\dagger - \mathbf{x}_k\|_{\mathbf{A}} \leq \|(\mathbf{x}^\dagger - \mathbf{x}_0)\|_{\mathbf{A}} \min_{\mathsf{p}\in\Pi_k^0} \max_{\lambda\in\mathrm{spec}(\mathbf{A})} |\mathsf{p}(\lambda)|, \tag{2.14}$$

where spec($\mathbf{A}$) *denotes the spectrum of the matrix* $\mathbf{A}$*.*

The last inequality can be reinterpreted in terms of the relative error:

Corollary 2.1.1. *Let* $\mathbf{A}$ *be symmetric and positive definite and let* $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ *be the sequence of iterates of the* CG *method. If* $k \geq 0$ *is fixed and* p *is any polynomial in* Π_k^0*, then the relative error is bounded as follows:*

$$\frac{\|\mathbf{x}^\dagger - \mathbf{x}_k\|_{\mathbf{A}}}{\|\mathbf{x}^\dagger - \mathbf{x}_0\|_{\mathbf{A}}} \leq \max_{\lambda\in\mathrm{spec}(\mathbf{A})} |\mathsf{p}(\lambda)|. \tag{2.15}$$

This leads to the most important result about the Conjugate Gradient method in $\mathbb{R}^N$.

Theorem 2.1.1. *If* $\mathbf{A} \in \mathbb{GL}_N(\mathbb{R})$ *is a symmetric and positive definite matrix and* $\mathbf{b}$ *is any vector in* $\mathbb{R}^N$*, then* CG *will find the solution* $\mathbf{x}^\dagger$ *of* (2.3) *in at most* N *iterative steps.*

Proof. It is enough to define the polynomial

$$\bar{\mathsf{p}}(\lambda) = \prod_{j=1}^{N} \frac{\lambda_j - \lambda}{\lambda_j},$$

observe that $\bar{\mathsf{p}}$ belongs to Π_N^0 and use Corollary 2.1.1: since $\bar{\mathsf{p}}$ vanishes on the spectrum of $\mathbf{A}$, $\|\mathbf{x}^\dagger - \mathbf{x}_N\|_{\mathbf{A}}$ must be equal to 0. □

This result is of course very pleasant, but not so good as it seems: first, if N is very large, N iterations can be too many. Then, we should remember that we usually have to deal with perturbed data and if $\mathbf{A}$ is ill-conditioned finding the exact solution of the perturbed system can lead to very bad results. The first problem will be considered immediately, whereas for the

second one see the next sections.
A-priori information about the data $\mathbf{b}$ and the spectrum of $\mathbf{A}$ can be very useful to improve the result stated in Theorem 2.1.1: we consider two different situations in which the same improved result can be shown.

Proposition 2.1.2. *Let $\mathbf{u}_j \in \mathbb{R}^N$, $j = 1, ..., N$, be the columns of a matrix $\mathbf{U}$ for which (2.5) holds. Suppose that $\mathbf{b}$ is a linear combination of k of these N eigenvectors of $\mathbf{A}$:*

$$\mathbf{b} = \sum_{l=1}^{k} \gamma_l \mathbf{u}_{i_l}, \quad \gamma_l \in \mathbb{R}, \quad 1 \leq i_1 < ... < i_k \leq N. \tag{2.16}$$

Then, if we set $\mathbf{x}_0 := 0$, CG *will converge in at most k iteration steps.*

Proof. For every $l = 1, ..., k$, let λ_{i_l} be the eigenvalue corresponding to the eigenvector $\mathbf{u}_{i_l}$. Then obviously

$$\mathbf{x}^\dagger = \sum_{l=1}^{k} \frac{\gamma_l}{\lambda_{i_l}} \mathbf{u}_{i_l}$$

and we proceed as in the proof of Theorem 2.1.1 defining

$$\bar{\mathsf{p}}(\lambda) = \prod_{l=1}^{k} \frac{\lambda_{i_l} - \lambda}{\lambda_{i_l}}.$$

Now $\bar{\mathsf{p}}$ belongs to Π_k^0 and vanishes on λ_{i_l} for every l, so

$$\bar{\mathsf{p}}(\mathbf{A})\mathbf{x}^\dagger = \sum_{l=1}^{k} \bar{\mathsf{p}}(\lambda_{i_l}) \frac{\gamma_l}{\lambda_{i_l}} \mathbf{u}_{i_l} = 0$$

and we use the minimization property

$$\|\mathbf{x}^\dagger - \mathbf{x}_k\|_{\mathbf{A}} \leq \|\bar{\mathsf{p}}(\mathbf{A})\mathbf{x}^\dagger\|_{\mathbf{A}} = 0$$

to conclude. □

In a similar way it is possible to prove the following statement.

Proposition 2.1.3. *Suppose that the spectrum of* $\mathbf{A}$ *consists of exactly* k *distinct eigenvalues. Then* CG *will find the solution of* (2.3) *in at most* k *iterations.*

One can also study the behavior of the relative error measured in the euclidean norm in terms of the condition number of the matrix $\mathbf{A}$:

Proposition 2.1.4. 1. *Let* $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$ *be the eigenvalues of* $\mathbf{A}$. *Then for every* $\mathbf{x} \in \mathbb{R}^N$ *we have*

$$\|\mathbf{x}\|_{\mathbf{A}} \lambda_N^{\frac{1}{2}} \leq \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{x}\|_{\mathbf{A}} \lambda_1^{\frac{1}{2}}. \tag{2.17}$$

2. *If* $\kappa_2(\mathbf{A}) := \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ *is the condition number of* $\mathbf{A}$, *then*

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|}{\|\mathbf{b}\|} \leq \sqrt{\kappa_2(\mathbf{A})} \frac{\|\mathbf{r}_0\|}{\|\mathbf{b}\|} \frac{\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathbf{A}}}{\|\mathbf{x}_0 - \mathbf{x}^\dagger\|_{\mathbf{A}}}. \tag{2.18}$$

Proof. Let $\mathbf{u}_j \in \mathbb{R}^N$ $(j = 1, ..., N)$ be the columns of a matrix $\mathbf{U}$ as in the proof of Proposition 2.1.2. Then

$$\mathbf{A}\mathbf{x} = \sum_{j=1}^{N} \lambda_j (\mathbf{u}_j^* \mathbf{x}) \mathbf{u}_j,$$

so

$$\begin{aligned} \lambda_N \|\mathbf{x}\|_{\mathbf{A}}^2 = \lambda_N \|\mathbf{A}^{\frac{1}{2}}\mathbf{x}\|_{\mathbf{A}}^2 = \lambda_N \sum_{j=1}^{N} \lambda_j (\mathbf{u}_j^* \mathbf{x})^2 \\ \leq \|\mathbf{A}\mathbf{x}\|^2 \leq \lambda_1 \sum_{j=1}^{N} \lambda_j (\mathbf{u}_j^* \mathbf{x})^2 = \lambda_1 \|\mathbf{A}^{\frac{1}{2}}\mathbf{x}\|_{\mathbf{A}}^2 = \lambda_1 \|\mathbf{x}\|_{\mathbf{A}}^2, \end{aligned} \tag{2.19}$$

which proves the first part.
For the second statement, recalling that $\|\mathbf{A}^{-1}\| = \lambda_N^{-1}$ and using the previous inequalities, we obtain

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|}{\|\mathbf{r}_0\|} = \frac{\|\mathbf{A}(\mathbf{x}^\dagger - \mathbf{x}_k)\|}{\|\mathbf{A}(\mathbf{x}^\dagger - \mathbf{x}_0)\|} \leq \sqrt{\frac{\lambda_1}{\lambda_N}} \frac{\|\mathbf{x}^\dagger - \mathbf{x}_k\|_{\mathbf{A}}}{\|\mathbf{x}^\dagger - \mathbf{x}_0\|_{\mathbf{A}}} = \sqrt{\kappa_2(\mathbf{A})} \frac{\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathbf{A}}}{\|\mathbf{x}_0 - \mathbf{x}^\dagger\|_{\mathbf{A}}}.$$

□

At last, we mention a result of J.W. Daniel (cf. [8]) that provides a bound for the relative error, which is, in some sense, as sharp as possible:

$$\frac{\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathbf{A}}}{\|\mathbf{x}_0 - \mathbf{x}^\dagger\|_{\mathbf{A}}} \leq 2\left(\frac{\sqrt{\kappa_2(\mathbf{A})}-1}{\sqrt{\kappa_2(\mathbf{A})}+1}\right)^k . \tag{2.20}$$

We conclude the section with a couple of examples that show clearly the efficiency of this method.

Example 2.1.1. *Suppose we know that the spectrum of the matrix* $\mathbf{A}$ *is contained in the interval* $\mathbb{I}_1 :=]9, 11[$. *Then, if we put* $\mathbf{x}_0 := 0$ *and*

$$\bar{\mathsf{p}}_k(\lambda) := \frac{(10-\lambda)^k}{10^k},$$

since $\bar{\mathsf{p}}_k$ *lies in* Π_k^0 *the minimization property* (2.11) *gives*

$$\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathbf{A}} \leq \|\mathbf{x}^\dagger\| \max_{9\leq\lambda\leq 11} |\bar{\mathsf{p}}_k(\lambda)| = \bar{\mathsf{p}}_k(9) = 10^{-k}. \tag{2.21}$$

Thus after k iteration steps, the relative error in the $\mathbf{A}$*-norm will be reduced of a factor* 10^{-3} *when* $10^{-k} \leq 10^{-3}$, *i.e. when* $k \geq 3$.
Observing that $\kappa_2(\mathbf{A}) \leq \frac{11}{9}$, *the estimate* (2.18) *can be used to deduce that*

$$\frac{\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\sqrt{11}}{3} 10^{-k}, \tag{2.22}$$

so the norm of the residual will be reduced of 10^{-3} *when* $10^{-k} \leq \frac{3}{\sqrt{11}} 10^{-3}$, *i.e. when* $k \geq 4$. *Moreover, since the function* $\lambda \mapsto \frac{\sqrt{\lambda}-1}{\sqrt{\lambda}+1}$ *is strictly increasing in* $]0, +\infty[$, $\frac{\sqrt{\kappa_2(\mathbf{A})}-1}{\sqrt{\kappa_2(\mathbf{A})}+1}$ *is bounded by* $\frac{\sqrt{11}-3}{\sqrt{11}+1}$ *and Daniel's inequality provides an improved version of* (2.21).

Even Daniel's estimate can be very pessimistic if we have more precise information about the spectrum of $\mathbf{A}$. For instance, if all the eigenvalues cluster in a small number of intervals, the condition number of $\mathbf{A}$ can be very huge, but CG can perform very well, as the following second example shows.

Example 2.1.2. *Suppose that the spectrum of* $\mathbf{A}$ *is contained in the intervals* $\mathbb{I}_1 := (1, 1.50)$ *and* $\mathbb{I}_2 := (399, 400)$ *and put* $\mathbf{x}_0 := 0$.
The best we can say about the condition number of $\mathbf{A}$ *is that* $\kappa_2(\mathbf{A}) \leq 400$, *which inserted in Daniel's formula gives*

$$\frac{\|\mathbf{x}_k - \mathbf{x}^\dagger\|}{\|\mathbf{x}^\dagger\|} \leq 2 \left(\frac{19}{21}\right)^k \approx 2(0.91)^k, \tag{2.23}$$

predicting a slow convergence.
However, if we take

$$\bar{\mathsf{p}}_{3k}(\lambda) := \frac{(1.25 - \lambda)^k (400 - \lambda)^{2k}}{(1.25)^k (400)^{2k}}$$

we easily see that

$$\max_{\lambda \in \mathrm{spec}(\mathbf{A})} |\bar{\mathsf{p}}_{3k}(\lambda)| \leq \left(\frac{0.25}{1.25}\right)^k = (0.2)^k, \tag{2.24}$$

providing a much sharper estimate. More precisely, in order to reduce the relative error in the $\mathbf{A}$*-norm of the factor* 10^{-3} *Daniel predicts* 83 *iteration steps, since* $2(0.91)^k < 10^{-3}$ *when* $k > -\frac{\log_{10}(2000)}{\log_{10}(0.91)} \approx 82.5$. *Instead, according to the estimate based on* $\bar{\mathsf{p}}_{3k}$ *the relative error will be reduced of the factor* 10^{-3} *after* $k = 3i$ *iterations when* $(0.2)^i < 10^{-3}$, *i.e. when* $i > -\frac{3}{\log_{10}(0.2)} \approx 4.3$, *hence it predicts only* 15 *iterations!*

In conclusion, in the finite dimensional case we have seen that the Conjugate Gradient method combines certain minimization properties in a very efficient way and that a-priori information can be used to predict the strength of its performance. Moreover, the polynomials q_k and p_k can be used to understand its behavior and can prove to be very useful in particular cases.

2.2 General definition in Hilbert spaces

In this section we define the conjugate gradient type methods in the usual Hilbert space framework. As a general reference and for the skipped proofs, we refer to [27].

If not said otherwise, here the operator A acting between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ will be self-adjoint and positive semi-definite with its spectrum contained in $[0,1]$.[1]

For $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, fix an initial guess $x_0 \in \mathcal{X}$ of the solution $A^\dagger y$ of $Ax = y$ and consider the bilinear form defined on the space of all polynomials Π_∞ by

$$\begin{aligned}[\phi, \psi]_n : &= \langle \phi(A)(y - Ax_0), A^n \psi(A)(y - Ax_0) \rangle \\ &= \int_0^\infty \phi(\lambda)\psi(\lambda)\lambda^n d\|\mathscr{E}_\lambda(y - Ax_0)\|^2,\end{aligned} \tag{2.25}$$

where $\{\mathscr{E}_\lambda\}$ denotes the spectral family associated to A.

Then from the theory of orthogonal polynomials (see, e.g., [88] Chapter II) we know that there is a well defined sequence of orthogonal polynomials $\{\mathsf{p}_k^{[n]}\}$ such that $\mathsf{p}_k^{[n]} \in \Pi_k$ and

$$[\mathsf{p}_k^{[n]}, \mathsf{p}_j^{[n]}]_n = 0, \quad k \neq j. \tag{2.26}$$

Moreover, if we force these polynomials to belong to Π_k^0, the sequence is univocally determined and satisfies a well known three-term recurrence formula, given by

$$\begin{aligned}&\mathsf{p}_0^{[n]} = 1, \ \ \mathsf{p}_1^{[n]} = 1 - \alpha_0^{[n]}\lambda, \\ &\mathsf{p}_{k+1}^{[n]} = -\alpha_k^{[n]}\lambda\mathsf{p}_k^{[n]} + \mathsf{p}_k^{[n]} - \alpha_k^{[n]}\frac{\beta_k^{[n]}}{\alpha_{k-1}^{[n]}}\left(\mathsf{p}_{k-1}^{[n]} - \mathsf{p}_k^{[n]}\right), \ k \geq 1,\end{aligned} \tag{2.27}$$

where the numbers $\alpha_k^{[n]} \neq 0$ and $\beta_k^{[n]}$, $k \geq 0$ can be computed explicitly (see below).

The k-th iterate of a conjugate gradient type method is given by

$$x_k^{[n]} := x_0 + \mathsf{q}_{k-1}^{[n]}(A)(y - Ax_0), \tag{2.28}$$

where the iteration polynomials $\{\mathsf{q}_{k-1}^{[n]}\}$ are related to the residual polynomials $\{\mathsf{p}_k^{[n]}\}$ via

$$\mathsf{q}_{k-1}^{[n]}(\lambda) = \frac{1 - \mathsf{p}_k^{[n]}}{\lambda} \in \Pi_{k-1}. \tag{2.29}$$

[1] Of course, if this is not the case, the equation $Ax = y$ can always be rescaled to guarantee $\|A\| \leq 1$.

The expression *residual polynomial* for $\mathsf{p}_k^{[n]}$ is justified by the fact that

$$\begin{aligned} y - Ax_k^{[n]} &= y - A\left(x_0 + \mathsf{q}_{k-1}^{[n]}(A)(y - Ax_0)\right) \\ &= y - Ax_0 - A\mathsf{q}_{k-1}^{[n]}(A)(y - Ax_0) \\ &= \left(I - A\mathsf{q}_{k-1}^{[n]}(A)\right)(y - Ax_0) \\ &= \mathsf{p}_k^{[n]}(A)(y - Ax_0). \end{aligned} \tag{2.30}$$

Moreover, if $y \in \mathcal{R}(A)$ and $x \in \mathcal{X}$ is such that $Ax = y$, then

$$x - x_k^{[n]} = x - x_0 - \mathsf{q}_{k-1}^{[n]}(A)A(x - x_0) = \mathsf{p}_k^{[n]}(A)(x - x_0). \tag{2.31}$$

In the following sections, in order to simplify the notations, we will omit the superscript n and the dependance of p_k and q_k from y unless strictly necessary.

2.3 The algorithms

In this section we describe how the algorithms of the conjugate gradient type methods can be derived from the general framework of the previous section. We refer basically to [27], adding a few details.
Let $n \in \mathbb{Z}$, $n \geq 0$.

Proposition 2.3.1. *Due to the recurrence formula* (2.27)*, the iteration polynomials satisfy*

$$\begin{aligned} &\mathsf{q}_{-1} = 0, \quad \mathsf{q}_0 = \alpha_0, \\ &\mathsf{q}_k = \mathsf{q}_{k-1} + \alpha_k\left(\mathsf{p}_k + \frac{\beta_k}{\alpha_{k-1}}(\mathsf{q}_{k-1} - \mathsf{q}_{k-2})\right), \ k \geq 1. \end{aligned} \tag{2.32}$$

Proof. By the definition of the iteration polynomials, we have

$$\lambda\mathsf{q}_{-1}(\lambda) = 1 - 1 = 0, \quad \mathsf{q}_0(\lambda) = \frac{1 - \mathsf{p}_1(\lambda)}{\lambda} = \alpha_0 \tag{2.33}$$

and for $k \geq 1$ the recurrence formula for the p_k gives

$$\begin{aligned}\mathsf{q}_k(\lambda) = \frac{1-\mathsf{p}_{k+1}(\lambda)}{\lambda} &= \frac{1+\alpha_k\lambda\mathsf{p}_k(\lambda)-\mathsf{p}_k(\lambda)+\alpha_k\alpha_{k-1}^{-1}\beta_k(\mathsf{p}_{k-1}(\lambda)-\mathsf{p}_k(\lambda))}{\lambda}\\ &= \frac{\alpha_k\lambda-\lambda^2\alpha_k\mathsf{q}_{k-1}(\lambda)+\lambda\mathsf{q}_{k-1}(\lambda)+\alpha_k\alpha_{k-1}^{-1}\beta_k(\lambda\mathsf{q}_{k-1}(\lambda)-\lambda\mathsf{q}_{k-2}(\lambda))}{\lambda}\\ &= \alpha_k\mathsf{p}_k(\lambda)+\mathsf{q}_{k-1}(\lambda)+\frac{\alpha_k}{\alpha_{k-1}}\beta_k\left(\mathsf{q}_{k-1}(\lambda)-\mathsf{q}_{k-2}(\lambda)\right).\end{aligned} \tag{2.34}$$

□

Proposition 2.3.2. *The iterates x_k of the conjugate gradient type methods can be computed with the following recursion:*

$$\begin{aligned}&\Delta x_0 = y - Ax_0, && x_1 = x_0+\alpha_0\Delta x_0,\\ &\Delta x_k = y - Ax_k + \beta_k\Delta x_{k-1}, && x_{k+1} = x_k + \alpha_k\Delta x_k, \quad k\geq 1.\end{aligned} \tag{2.35}$$

Proof. Since $\mathsf{q}_0 = \alpha_0$, the relation between x_1 and x_0 is obvious.
We proceed by induction on k. From the definitions of x_k and x_{k+1} there follows

$$x_{k+1} = x_k + (\mathsf{q}_k - \mathsf{q}_{k-1})(A)(\Delta x_0) \tag{2.36}$$

and now using Proposition 2.3.1 and the induction we have:

$$\begin{aligned}(\mathsf{q}_k-\mathsf{q}_{k-1})(A)(\Delta x_0) &= \alpha_k\mathsf{p}_k(A)(\Delta x_0)+\frac{\alpha_k}{\alpha_{k-1}}\beta_k(\mathsf{q}_{k-1}-\mathsf{q}_{k-2})(A)(\Delta x_0)\\ &= \alpha_k(y-Ax_k)+\alpha_k\beta_k\frac{(\mathsf{q}_{k-1}-\mathsf{q}_{k-2})(A)(\Delta x_0)}{\alpha_{k-1}}\\ &= \alpha_k(y-Ax_k+\beta_k\Delta x_{k-1}).\end{aligned} \tag{2.37}$$

□

Proposition 2.3.3. *Define*

$$\mathsf{s}_0 := 1, \quad \mathsf{s}_k := \mathsf{p}_k + \beta_k\mathsf{s}_{k-1}, \ k \geq 1. \tag{2.38}$$

Then for every $k \geq 0$ the following relations hold:

$$\Delta x_k = \mathsf{s}_k(A)(y-Ax_0), \tag{2.39}$$

$$\mathsf{p}_{k+1} = \mathsf{p}_k - \alpha_k\lambda\mathsf{s}_k. \tag{2.40}$$

Proof. For $k = 0$, the first relation is obviously satisfied. For $k \geq 1$, using induction again we obtain:

$$\Delta x_k = y - Ax_k + \beta_k \Delta x_{k-1} = \mathsf{p}_k(A)(\Delta x_0) + \beta_k \mathsf{s}_{k-1}(A)(\Delta x_0) = \mathsf{s}_k(A)(\Delta x_0),$$

which proves (2.39).

To see (2.40), it is enough to consider the relations

$$\frac{x_{k+1} - x_k}{\alpha_k} = \Delta x_k = \mathsf{s}_k(A)(\Delta x_0),$$
$$x_{k+1} - x_k = (\mathsf{q}_k(A) - \mathsf{q}_{k-1}(A))(\Delta x_0),$$
$$\lambda(\mathsf{q}_k(\lambda) - \mathsf{q}_{k-1}(\lambda)) = \mathsf{p}_k(\lambda) - \mathsf{p}_{k+1}(\lambda)$$

and link them together. □

Proposition 2.3.4. *The sequence* $\{\mathsf{s}_k\} = \{\mathsf{s}_k^{[n]}\}_{k\in\mathbb{N}}$ *is orthogonal with respect to the inner product* $[\cdot,\cdot]_{n+1}$. *More precisely, if* ℓ *denotes the number of the nonzero points of increase of the function* $\alpha(\lambda) = \|\mathscr{E}_\lambda(\Delta x_0)\|^2$, *then* [2]

$$\mathsf{p}_k^{[n+1]} = \frac{1}{\pi_{k,n}} \frac{\mathsf{p}_k^{[n]} - \mathsf{p}_{k+1}^{[n]}}{\lambda}, \quad \textit{with} \;\; \pi_{k,n} := (\mathsf{p}_k^{[n]})'(0) - (\mathsf{p}_{k+1}^{[n]})'(0) > 0 \tag{2.41}$$

for every $0 \leq k < \ell$.

Proof. A well known fact from the theory of orthogonal polynomials is that $\mathsf{p}_k^{[n]}$ has k simple zeros $\lambda_{j,k}^{[n]}$, $j = 1, ..., k$, with

$$0 < \lambda_{1,k}^{[n]} < \lambda_{2,k}^{[n]} < ... < \lambda_{k,k}^{[n]} \leq \|A\| \leq 1.$$

As a consequence, we obtain

$$\mathsf{p}_k^{[n]}(\lambda) = \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\lambda_{j,k}^{[n]}}\right), \quad (\mathsf{p}_k^{[n]})'(0) = -\sum_{j=1}^{k} \frac{1}{\lambda_{j,k}^{[n]}}. \tag{2.42}$$

Thus $(\mathsf{p}_k^{[n]})'(0) \leq -k$. Moreover, the zeros of two consecutive orthogonal polynomials interlace, i.e.

$$0 < \lambda_{1,k+1}^{[n]} < \lambda_{1,k}^{[n]} < \lambda_{2,k+1}^{[n]} < \lambda_{2,k}^{[n]} < ... < \lambda_{k,k}^{[n]} < \lambda_{k+1,k+1}^{[n]},$$

[2] of course, ℓ can be finite or infinity: in the ill-posed case, since the spectrum of A^*A clusters at 0, it is infinity.

so $\pi_{k,n} > 0$ holds true.
Now observe that by the definition of $\pi_{k,n}$ the right-hand side of (2.41) lies in Π_k^0. Denote this polynomial by p. For any other polynomial $\mathsf{q} \in \Pi_{k-1}$ we have

$$[\mathsf{p}, \mathsf{q}]_{n+1} = \frac{1}{\pi_{k,n}}[\mathsf{p}_k^{[n]} - \mathsf{p}_{k+1}^{[n]}, \mathsf{q}]_n = \frac{1}{\pi_{k,n}}([\mathsf{p}_k^{[n]}, \mathsf{q}]_n - [\mathsf{p}_{k+1}^{[n]}, \mathsf{q}]_n) = 0$$

and since $\mathsf{p}_k^{[n+1]}$ is the only polynomial in Π_k^0 satisfying this equation for every $\mathsf{q} \in \Pi_{k-1}^0$, $\mathsf{p} = \mathsf{p}_k^{[n+1]}$. The orthogonality of the sequence $\{\mathsf{s}_k\}$ follows immediately from Proposition 2.3.3. □

Proposition 2.3.5. *If the function $\alpha(\lambda)$ defined in Proposition* 2.3.4 *has $\ell = \infty$ points of increase, the coefficients α_k and β_k appearing in the formulas* (2.35) *of Proposition* 2.3.2 *can be computed as follows:*

$$\alpha_k = \frac{[\mathsf{p}_k, \mathsf{p}_k]_n}{[\mathsf{s}_k, \mathsf{s}_k]_{n+1}}, \quad k \geq 0, \tag{2.43}$$

$$\beta_k = \frac{1}{\alpha_{k-1}} \frac{[\mathsf{p}_k, \mathsf{p}_k]_n}{[\mathsf{s}_{k-1}, \mathsf{s}_{k-1}]_{n+1}} = \frac{[\mathsf{p}_k, \mathsf{p}_k]_n}{[\mathsf{p}_{k-1}, \mathsf{p}_{k-1}]_n}, \quad k \geq 1. \tag{2.44}$$

Otherwise, the formulas above remain valid, but the iteration must be stopped in the course of the $(\ell + 1)$-th step since $[\mathsf{s}_\ell, \mathsf{s}_\ell]_{n+1} = 0$ and α_k is undefined. In this case, we distinguish between the following possibilities:

- *if y belongs to $\mathcal{R}(A)$, for every $n \in \mathbb{N}_0$ $x_\ell^{[n]} = A^\dagger y$;*
- *if y has a non-trivial component along $\mathcal{R}(A)^\perp$ and $n \geq 1$, then $(I - \mathscr{E}_0)x_\ell = A^\dagger y$;*
- *if y has a non-trivial component along $\mathcal{R}(A)^\perp$ and $n = 0$, then the conclusion $(I - \mathscr{E}_0)x_\ell = A^\dagger y$ does not hold any more.*

Proof. Note that in the ill-posed case (2.43) and (2.44) are well defined, since all inner products are nonzero. By the orthogonality of $\{\mathsf{p}_k\}$ and Proposition 2.3.3, for every $k \geq 0$ we have

$$0 = [\mathsf{p}_{k+1}, \mathsf{s}_k]_n = [\mathsf{p}_k, \mathsf{s}_k]_n - \alpha_k[\lambda \mathsf{s}_k, \mathsf{s}_k]_n = [\mathsf{p}_k, \mathsf{p}_k] - \alpha_k[\mathsf{s}_k, \mathsf{s}_k]_{n+1},$$

which gives (2.43).

For every $k \geq 1$, the orthogonality of $\{\mathsf{s}_k\}$ with respect to $[\cdot,\cdot]_{n+1}$ yields

$$\begin{aligned} 0 = [\mathsf{s}_k, \mathsf{s}_{k-1}]_{n+1} &= [\mathsf{p}_k, \lambda \mathsf{s}_{k-1}]_n + \beta_k [\mathsf{s}_{k-1}, \mathsf{s}_{k-1}]_{n+1} \\ &= \frac{1}{\alpha_{k-1}} [\mathsf{p}_k, \mathsf{p}_{k-1} - \mathsf{p}_k]_n + \beta_k [\mathsf{s}_{k-1}, \mathsf{s}_{k-1}]_{n+1} \\ &= -\frac{1}{\alpha_{k-1}} [\mathsf{p}_k, \mathsf{p}_k]_n + \beta_k [\mathsf{s}_{k-1}, \mathsf{s}_{k-1}]_{n+1}, \end{aligned} \tag{2.45}$$

which leads to (2.44).

Now suppose that $\ell < \infty$. Then the bilinear form (2.25) turns out to be

$$[\phi, \psi]_n = \int_0^{+\infty} \lambda^n \phi(\lambda) \psi(\lambda) d\alpha(\lambda) = \sum_{j=1}^{\ell} \lambda_j^n \phi(\lambda_j) \psi(\lambda_j) \tag{2.46}$$

if $y \in \mathcal{R}(A)$ or if $n \geq 1$, whereas if neither of these two conditions is satisfied $\lambda_0 := 0$ is the $(\ell+1)$-th point of increase of $\alpha(\lambda)$ and

$$[\phi, \psi]_n = \sum_{j=0}^{\ell} \lambda_j^n \phi(\lambda_j) \psi(\lambda_j).$$

If $y \in \mathcal{R}(A)$, since there exists a unique polynomial $\mathsf{p}_k^{[n]} \in \Pi_k^0$ perpendicular to Π_{k-1} such that $\mathsf{p}_k^{[n]}(\lambda_j) = 0$ for $j = 1, ..., k$ and consequently satisfying $[\mathsf{p}_k, \mathsf{p}_k]_n = 0$, then $\|x_\ell^{[n]} - x^\dagger\|^2 = \|\mathsf{p}_\ell(A)(x_0 - x^\dagger)\|^2 = 0$. If y does not belong to $\mathcal{R}(A)$ and $n \geq 1$, since (2.46) is still valid, due to the same considerations we obtain $(I - \mathscr{E}_0) x_\ell^{[n]} = x^\dagger$. Finally, in the case $n = 0$ it is impossible to find $\mathsf{p}_k^{[n]}$ as before, thus the same conclusions cannot be deduced. □

From the orthogonal polynomial point of view, the minimization property of the conjugate gradient type methods turns out to be an easy consequence of the previous results.

Proposition 2.3.6. *Suppose $n \geq 1$, let x_k be the k-th iterate of the corresponding conjugate gradient type method and let x be any other element in the Krylov shifted subspace $x_0 + \mathcal{K}_{k-1}(A; y - Ax_0)$. Then*

$$\|A^{\frac{n-1}{2}}(y - Ax_k)\| \leq \|A^{\frac{n-1}{2}}(y - Ax)\| \tag{2.47}$$

and the equality holds if and only if $x = x_k$.
If $n = 0$ and $y \in \mathcal{R}(A^{\frac{1}{2}})$, then $\|A^{\frac{n-1}{2}}(y - Ax_k)\|$ is well defined and the same result obtained in the case $n \geq 1$ remains valid.

Proof. Consider the case $n = 1$. In terms of the residual polynomials p_k, (2.47) reads as follows:

$$[\mathsf{p}_k, \mathsf{p}_k]_{n-1} \leq [\mathsf{p}, \mathsf{p}], \quad \text{for every } \mathsf{p} \in \Pi_k^0.$$

Since for every $\mathsf{p} \in \Pi_k^0$ there exists $\mathsf{s} \in \Pi_{k-1}$ such that $\mathsf{p} - \mathsf{p}_k = \lambda\mathsf{s}$, by orthogonality we have

$$[\mathsf{p}, \mathsf{p}]_{n-1} - [\mathsf{p}_k, \mathsf{p}_k]_{n-1} = [\mathsf{p} - \mathsf{p}_k, \mathsf{p} + \mathsf{p}_k]_{n-1} = [\mathsf{s}, \lambda\mathsf{s} + 2\mathsf{p}_k]_n = [\mathsf{s}, \mathsf{s}]_{n+1} \geq 0,$$

and the equality holds if and only if $\mathsf{s} = 0$, i.e. if and only if $\mathsf{p} = \mathsf{p}_k$.
If $n = 0$ and $y \in \mathcal{R}(A^{\frac{1}{2}})$, then $[\mathsf{p}_k, \mathsf{p}_k]_{-1}$ is well defined by

$$[\mathsf{p}_k, \mathsf{p}_k]_{-1} := \int_{0+}^{\infty} \mathsf{p}_k^2(\lambda)\lambda^{-1} d\|\mathscr{E}_\lambda(y - Ax_0)\|^2$$

and the proof is the same as above. □

Note that in the case $n = 0$ this is the same result obtained in the discrete case in Proposition 2.1.1.
The computation of the coefficients α_k and β_k allows a very easy and cheap computation of the iterates of the conjugate gradient type-methods.
We focus our attention on the cases $n = 1$ and $n = 0$, corresponding respectively to the minimal residual method and the classical conjugate gradient method.

2.3.1 The minimal residual method (MR) and the conjugate gradient method (CG)

- In the case $n = 1$, from Proposition 2.3.6 we see that the corresponding method minimizes, in the shifted Krylov space $x_0 + \mathcal{K}_{k-1}(A; y - Ax_0)$, the residual norm. For this reason, this method is called *minimal residual method* (MR). Propositions 2.3.1-2.3.5 lead to Algorithm 1.

- In the case $n = 0$, using again Propositions 2.3.1-2.3.5 we find (cf. Algorithm 2) the classical Conjugate Gradient method originally proposed by Hestenes and Stiefel in [45] in 1952. If $y \in \mathcal{R}(A)$, then according to Proposition 2.3.6, the k-th iterate x_k of CG minimizes the error $x^\dagger - x_k$ in $x_0 + \mathcal{K}_{k-1}(A; y - Ax_0)$ with respect to the energy-norm $\langle x^\dagger - x_k, A(x^\dagger - x_k)\rangle$.

Looking at the algorithms, it is important to note that for every iterative step MR and CG must compute only once a product of the type Av with $v \in \mathcal{X}$.

Algorithm 1 MR

```
$r_0 = y - Ax_0;$
$d = r_0;$
$Ad = Ar_0;$
$k = 0;$
while (not stop) do
    $\alpha = \langle r_k, Ar_k\rangle / \|Ad\|^2;$
    $x_{k+1} = x_k + \alpha d;$
    $r_{k+1} = r_k - \alpha Ad;$
    $\beta = \langle r_{k+1}, Ar_{k+1}\rangle / \langle r_k, Ar_k\rangle;$
    $d = r_{k+1} + \beta d;$
    $Ad = Ar_{k+1} + \beta Ad;$
    $k = k + 1;$
end while
```

2.3.2 CGNE and CGME

Suppose that the operator A fails to be self-adjoint and semi-definite, i.e. it is of the type we discussed in Chapter 1. Then it is still possible to use the conjugate gradient type methods, seeking for the (best-approximate) solution of the equation

$$AA^*v = y \tag{2.48}$$

Algorithm 2 CG

```
$r_0 = y - Ax_0$;
$d = r_0$;
$k = 0$;
while (not stop) do
    $\alpha = \|r_k\|^2 / \langle d, Ad \rangle$;
    $x_{k+1} = x_k + \alpha d$;
    $r_{k+1} = r_k - \alpha A d$;
    $\beta = \|r_{k+1}\|^2 / \|r_k\|^2$;
    $d = r_{k+1} + \beta d$;
    $k = k + 1$;
end while
```

and putting $x = A^* \upsilon$.

In this more general case, we shall denote as usual with $\{E_\lambda\}$ the spectral family of A^*A and with $\{F_\lambda\}$ the spectral family of AA^*. All the definitions of the self-adjoint case carry over here, keeping in mind that they will always refer to AA^* instead of A and the corresponding iterates are

$$\upsilon_k = \upsilon_0 + \mathsf{q}_{k-1}(AA^*)(y - Ax_0). \tag{2.49}$$

The definition of the first iterate υ_0 is not important, since we are not interested in calculating υ_k, but we are looking for x_k. Thus we multiply both sides of the equation (2.49) by A^* and get

$$x_k = x_0 + A^* \mathsf{q}_{k-1}(AA^*)(y - Ax_0) = x_0 + \mathsf{q}_{k-1}(A^*A)A^*(y - Ax_0). \tag{2.50}$$

As in the self-adjoint case, the residual $y - Ax_k$ is expressed in terms of the residual polynomials p_k corresponding to the operator AA^* via the formula

$$y - Ax_k = \mathsf{p}_k(AA^*)(y - Ax_0) \tag{2.51}$$

and if $y = Ax$ for some $x \in \mathcal{X}$, then

$$x - x_k = \mathsf{p}_k(AA^*)(x - x_0). \tag{2.52}$$

As in the self-adjoint case, we consider the possibilities $n = 1$ and $n = 0$.

- If $n = 1$, according to Proposition 2.3.6, the iterates x_k minimize the residual norm in the Krylov shifted space $x_0 + \mathcal{K}_{k-1}(A^*A; A^*(y - Ax_0))$, cf. Algorithm 3.
  A very important fact concerning this case is that this is equal to the direct application of CG to the normal equation
  $$A^*Ax = A^*y,$$
  as one can easily verify by using Proposition 2.3.6 or by comparing the algorithms. This method is by far the most famous in literature and is usually called CGNE, i.e. CG applied to the Normal Equation.

- It is also possible to apply CG to the equation (2.48), obtaining Algorithm 4: this corresponds to the choice $n = 0$ and by Proposition 2.3.6 if $y \in \mathcal{R}(A)$ the iterates x_k minimize the error norm $\|x^\dagger - x_k\|$ in the corresponding Krylov space.[3]

We conclude this section with a remark: forming and solving the equation (2.48) can only lead to the minimal norm solution of $Ax = y$, because the iterates $x_k = A^*v_k$ lie in $\mathcal{R}(A^*) \subseteq \ker(A)^\perp$, which is closed. Thus, if one is looking for solutions different from $x^\dagger$, then should not rely on these methods.

2.3.3 Cheap Implementations

In [27] M. Hanke suggests an implementation of both gradient type methods with $n = 1$ and $n = 0$ in one scheme, which requires approximately the same computational effort of implementing only one of them. For this purpose, further results (gathered in Proposition 2.3.7 below) from the theory of orthogonal polynomials are needed.

[3] The reader should keep in mind the difference between CG and CGME: the former minimizes $x_k - x^\dagger$ in the energy norm, whereas the latter minimizes exactly the norm of the error $\|x^\dagger - x_k\|$.

Algorithm 3 CGNE

```
$r_0 = y - Ax_0$;
$d = A^* r_0$;
$k = 0$;
while (not stop) do
    $\alpha = \|A^* r_k\|^2 / \|Ad\|^2$;
    $x_{k+1} = x_k + \alpha d$;
    $r_{k+1} = r_k - \alpha Ad$;
    $\beta = \|A^* r_{k+1}\|^2 / \|A^* r_k\|^2$;
    $d = A^* r_{k+1} + \beta d$;
    $k = k + 1$;
end while
```

Algorithm 4 CGME

```
$r_0 = y - Ax_0$;
$d = A^* r_0$;
$k = 0$;
while (not stop) do
    $\alpha = \|r_k\|^2 / \|d\|^2$;
    $x_{k+1} = x_k + \alpha d$;
    $r_{k+1} = r_k - \alpha Ad$;
    $\beta = \|r_{k+1}\|^2 / \|r_k\|^2$;
    $d = A^* r_{k+1} + \beta d$;
    $k = k + 1$;
end while
```

For simplicity, in the remainder we will restrict to the case in which A is a semi-definite, self-adjoint operator and the initial guess is the origin: $x_0 = 0$. For the proof of the following facts and for a more exhaustive coverage of the argument, see [27]. The second statement has already been proved in Proposition 2.3.4.

Proposition 2.3.7. *Fix $k \in \mathbb{N}_0$, $k < \ell$. Then:*

1. *For $n \in \mathbb{N}$, the corresponding residual polynomial $\mathsf{p}_k^{[n]}$ can be written in the form*

$$\mathsf{p}_k^{[n]} = [\mathsf{p}_k^{[n]}, \mathsf{p}_k^{[n]}]_{n-1} \sum_{j=0}^{k} [\mathsf{p}_j^{[n-1]}, \mathsf{p}_j^{[n-1]}]_{n-1}^{-1} \mathsf{p}_j^{[n-1]} \tag{2.53}$$

and

$$\|A^{\frac{n-1}{2}}(y - Ax_k^{[n]})\|^2 = [\mathsf{p}_k^{[n]}, \mathsf{p}_k^{[n]}]_{n-1} = \left(\sum_{j=0}^{k} [\mathsf{p}_j^{[n-1]}, \mathsf{p}_j^{[n-1]}]_{n-1}^{-1} \right)^{-1} . \tag{2.54}$$

The same is true for $n = 0$ if and only if $\mathscr{E}_0 y = 0$, i.e. if and only if the data y has no component along $\mathcal{R}(A)^\perp$.

2. *For $n \in \mathbb{N}_0$ there holds:*

$$\mathsf{p}_k^{[n+1]} = \frac{1}{\pi_{k,n}} \frac{\mathsf{p}_k^{[n]} - \mathsf{p}_{k+1}^{[n]}}{\lambda}. \tag{2.55}$$

3. *For $n \in \mathbb{N}$, $\pi_{k,n} = (\mathsf{p}_k^{[n]})'(0) - (\mathsf{p}_{k+1}^{[n]})'(0)$ is also equal to*

$$\pi_{k,n} = \frac{[\mathsf{p}_k^{[n]}, \mathsf{p}_k^{[n]}]_{n-1} - [\mathsf{p}_{k+1}^{[n]}, \mathsf{p}_{k+1}^{[n]}]_{n-1}}{[\mathsf{p}_k^{[n+1]}, \mathsf{p}_k^{[n+1]}]_n} = \frac{[\mathsf{p}_k^{[n+1]}, \mathsf{p}_k^{[n+1]}]_n}{[\mathsf{p}_k^{[n+1]}, \mathsf{p}_k^{[n+1]}]_{n+1}}. \tag{2.56}$$

Starting from Algorithm 1 and using Proposition 2.3.7, it is not difficult to construct an algorithm which implements both MR and CG without further computational effort. The same can be done starting from CGNE. The results are summarized in Algorithm 5 (6), where x_k and z_k are the iterates corresponding respectively to CG (CGME) and MR (CGNE).
Once again, we address the reader to [27] for further details.

Algorithm 5 MR+CG

```
$x_0 = z_0$;
$r_0 = y - Az_0$;
$d = r_0$;
$p_1 = Ar_0$;
$p_2 = Ad$;
$k = 0$;
while (not stop) do
    $\alpha = \langle r_k, p_1\rangle / \|p_2\|^2$;
    $z_{k+1} = z_k + \alpha d$;
    $\pi = \|r_k\|^2 / \langle r_k, p_1\rangle$
    $x_{k+1} = x_k + \pi r_k$;
    $r_{k+1} = r_k - \alpha p_2$;
    $t = Ar_{k+1}$;
    $\beta = \langle r_{k+1}, t\rangle / \langle r_k, p_1\rangle$;
    $d = r_{k+1} + \beta d$;
    $p_1 = t$;
    $p_2 = t + \beta p_2$;
    $k = k + 1$;
end while
```

2.4 Regularization theory for the conjugate gradient type methods

This section is entirely devoted to the study of the conjugate gradient methods for ill-posed problems. Although such methods are not regularization methods in the strict sense of Definition 1.9.1, as we will see they preserve the most important regularization properties and for this reason they are usually included in the class of regularization methods. Since the results we are going to state can nowadays be considered classic and are treated in great detail both in [27] and in [17], most of the proofs will be omitted. The non-omitted

Algorithm 6 CGNE+CGME

```
$x_0 = z_0;$
$r_0 = y - Az_0;$
$d = A^* r_0;$
$p_1 = d;$
$p_2 = Ad;$
$k = 0;$
while (not stop) do
    $\alpha = \|p_1\|^2 / \|p_2\|^2;$
    $z_{k+1} = z_k + \alpha d;$
    $\pi = \|r_k\|^2 / \|p_1\|^2$
    $x_{k+1} = x_k + \pi p_1;$
    $r_{k+1} = r_k - \alpha p_2;$
    $t = A^* r_{k+1};$
    $\beta = \|t\|^2 / \|p_1\|^2;$
    $d = t + \beta d;$
    $p_1 = t;$
    $p_2 = Ad;$
    $k = k + 1;$
end while
```

proofs and calculations will serve us to define new stopping rules later on. We begin with an apparently very unpleasant result concerning the stability properties of the conjugate gradient type methods.

Theorem 2.4.1. *Let the self-adjoint semi-definite operator A be compact and non-degenerate. Then for any conjugate gradient type method with parameter $n \in \mathbb{N}_0$ and for every $k \in \mathbb{N}$, the operator $R_k = R_k^{[n]}$ that maps the data y onto the k-th iterate $x_k = x_k^{[n]}$ is discontinuous in $\mathcal{X}$.*
Moreover, even in the non compact case, R_k is discontinuous at y if and only if $\mathscr{E}_0 y$ belongs to an invariant subspace of A of dimension at most $k-1$.

Every stopping rule for a conjugate gradient type method must take into

account this phenomenon. In particular, no a-priory stopping rule $k(\delta)$ can render a conjugate gradient type method convergent (cf. [27] and [17]). At first, this seems to be discouraging, but the lack of discontinuity of R_k is not really a big problem, since it is still possible to find reliable a-posteriori stopping rules which preserve the main properties of convergence and order optimality.
Before we proceed with the analysis, we have to underline that the methods with parameter $n \geq 1$ are much easier to treat than those with $n = 0$. For this reason, we shall consider the two cases separately.

2.4.1 Regularizing properties of MR and CGNE

As usual, we begin considering the unperturbed case first.

Proposition 2.4.1. *Let $y \in \mathcal{R}(A)$ and let n_1 and n_2 be integers with $n_1 < n_2$ and $[1,1]_{n_1} < +\infty$. Then $[\mathsf{p}_k^{[n_2]}, \mathsf{p}_k^{[n_2]}]_{n_1}$ is strictly decreasing as k goes from 0 to ℓ.*

This has two important consequences:

Corollary 2.4.1. *If $y \in \mathcal{R}(A)$ and $x_k = x_k^{[n]}$ are the iterates of a conjugate gradient type method corresponding to a parameter $n \geq 1$ and right-hand side y, then*

- *The residual norm $\|y - Ax_k\|$ is strictly decreasing for $0 \leq k \leq \ell$.*
- *The iteration error $\|x^\dagger - x_k\|$ is strictly decreasing for $0 \leq k \leq \ell$.*

To obtain the most important convergence results, the following estimates play a central role. We have to distinguish between the self-adjoint case and the more general setting of Section 2.3.2. The proof of the part with the operator AA^*, which will turn out to be of great importance later, can be found entirely in [17], Theorem 7.9.

Lemma 2.4.1. *Let $\lambda_{1,k} < ... < \lambda_{k,k}$ be the the zeros of p_k. Then:*

- *In the self-adjoint case, for* $y \in \mathcal{X}$,

$$\|y - Ax_k\| \leq \|\mathscr{E}_{\lambda_{1,k}} \varphi_k(A)y\|, \tag{2.57}$$

with the function $\varphi(\lambda) := \mathsf{p}_k(\lambda)\frac{\lambda_{1,k}}{\lambda_{1,k}-\lambda}$ *satisfying*

$$0 \leq \varphi_k(\lambda) \leq 1, \quad \lambda^2 \varphi_k^2(\lambda) \leq 4|\mathsf{p}_k'(0)|^{-1}, \quad 0 \leq \lambda \leq \lambda_{1,k}. \tag{2.58}$$

- *In the general case with* AA^*, *for* $y \in \mathcal{Y}$,

$$\|y - Ax_k\| \leq \|F_{\lambda_{1,k}} \varphi_k(AA^*)y\|, \tag{2.59}$$

with the function $\varphi(\lambda) := \mathsf{p}_k(\lambda)\left(\frac{\lambda_{1,k}}{\lambda_{1,k}-\lambda}\right)^{\frac{1}{2}}$ *satisfying*

$$0 \leq \varphi_k(\lambda) \leq 1, \quad \lambda \varphi_k^2(\lambda) \leq |\mathsf{p}_k'(0)|^{-1}, \quad 0 \leq \lambda \leq \lambda_{1,k}. \tag{2.60}$$

This leads to the following convergence theorem.

Theorem 2.4.2.

- *Suppose that* A *is self-adjoint and semi-definite. If* $y \in \mathcal{R}(A)$, *then the iterates* $\{x_k\}$ *of a conjugate gradient type method with parameter* $n \geq 1$ *converge to* $A^\dagger y$ *as* $k \to +\infty$. *If* $y \notin \mathcal{R}(A)$ *and* $\ell = \infty$, *then* $\|x_k\| \to +\infty$ *as* $k \to +\infty$. *If* $y \notin \mathcal{R}(A)$ *and* $\ell < \infty$ *then the iteration terminates after* ℓ *steps,* $Ax_\ell = \mathscr{E}_0 y$ *and* $x_\ell = A^\dagger y$ *if and only if* $\ell = 0$.

- *Let* A *satisfy the assumptions of Section* 2.3.2 *and let* $\{x_k\}$ *be the iterates of a conjugate gradient type method with parameter* $n \geq 1$ *applied with* AA^*. *If* $y \in \mathcal{D}(A^\dagger)$, *then* x_k *converges to* $A^\dagger y$ *as* $k \to +\infty$, *but if* $y \notin \mathcal{D}(A^\dagger)$, *then* $\|x_k\| \to +\infty$ *as* $k \to +\infty$.

Theorem 2.4.2 implies that the iteration must be terminated appropriately when dealing with perturbed data $y^\delta \notin \mathcal{D}(A^\dagger)$, due to numerical instabilities.

Another consequence of Lemma 2.4.1 is the following one:

Lemma 2.4.2. *Let x_k^δ be the iterates of a conjugate gradient type method with parameter $n \geq 1$ corresponding to the perturbed right-hand side y^δ and the self-adjoint semi-definite operator A. If the exact right-hand side belongs to $\mathcal{R}(A)$ and $\ell = \infty$, then*

$$\limsup_{k\to+\infty} \|y^\delta - Ax_k^\delta\| \leq \|y - y^\delta\|. \tag{2.61}$$

Moreover, if the exact data satisfy the source condition

$$A^\dagger y \in \mathcal{X}_{\frac{\mu}{2},\rho}, \quad \mu > 0, \ \rho > 0, \tag{2.62}$$

then there exists a constant $C > 0$ such that

$$\|y^\delta - Ax_k^\delta\| \leq \|y - y^\delta\| + C|\mathsf{p}_k'(0)|^{-\mu-1}\rho, \quad 1 \leq k \leq \ell. \tag{2.63}$$

The same estimate is obtained for the gradient type methods working with AA^* instead of A, but the exponent $-\mu - 1$ must be replaced by $-\frac{\mu+1}{2}$.

Assuming the source condition (2.62), it is also possible to give an estimate for the error:

Lemma 2.4.3. *Let x_k^δ be the iterates of a conjugate gradient type method with parameter $n \geq 1$ corresponding to y^δ and the self-adjoint semi-definite operator A. If* (2.62) *holds, then for $0 \leq k \leq \ell$,*

$$\|A^\dagger y - x_k^\delta\| \leq C\left(\|F_{\lambda_{1,k}}(y - y^\delta)\||\mathsf{p}_k'(0)| + \rho^{\frac{1}{\mu+1}} M_k^{\frac{\mu}{\mu+1}}\right), \tag{2.64}$$

where C is a positive constant depending only on μ, and

$$M_k := \max\{\|y^\delta - Ax_k^\delta\|, \|y^\delta - y\|\}. \tag{2.65}$$

In the cases with AA^* instead of A, the same is true, but in (2.64) *$|\mathsf{p}_k'(0)|$ must be replaced by $|\mathsf{p}_k'(0)|^{\frac{1}{2}}$.*

We underline that in Hanke's statement of Lemma 2.4.3 (cf. Lemma 3.8 in [27]) the term $\|F_{\lambda_{1,k}}(y - y^\delta)\|$ in the inequality (2.64) is replaced by $\|y - y^\delta\|$. This sharper estimate follows directly from the proof of the Lemma 3.8 in [27].

Combining Lemma 2.4.2 and Lemma 2.4.3 we obtain:

Theorem 2.4.3. *If y satisfies the source condition* (2.62) *and $\|y^\delta - y\| \leq \delta$, then the iteration error of a conjugate gradient type method with parameter $n \geq 1$ associated to a self-adjoint semi-definite operator A is bounded by*

$$\|A^\dagger y - x_k^\delta\| \leq C \left(|\mathsf{p}_k'(0)|^{-\mu}\rho + |\mathsf{p}_k'(0)|\delta \right), \quad 1 \leq k \leq \ell. \tag{2.66}$$

In the cases with AA^* instead of A, the same estimate holds, but $|\mathsf{p}_k'(0)|$ must be replaced by $|\mathsf{p}_k'(0)|^{\frac{1}{2}}$.

Theorem 2.4.3 can be seen as the theoretical justification of the well known phenomenon of the semi-convergence, which is experimented in practical examples: from (2.66), we observe that for small values of k the right-hand side is dominated by $|\mathsf{p}_k'(0)|^{-\mu}\rho$, but as k increases towards $+\infty$, this term converges to 0, while $|\mathsf{p}_k'(0)|\delta$ diverges. Thus, as usual, there is a precise value of k that minimizes the error $\|A^\dagger y - x_k^\delta\|$ and it is necessary to define appropriate stopping rules to obtain satisfying results. In the case of the conjugate gradient type methods with parameter $n \geq 1$, the Discrepancy Principle proves to be an efficient one.

Definition 2.4.1 (Discrepancy Principle for MR and CGNE). *Assume $\|y^\delta - y\| \leq \delta$. Fix a number $\tau > 1$ and terminate the iteration when, for the first time, $\|y^\delta - Ax_k^\delta\| \leq \tau\delta$. Denote the corresponding stopping index with $k_D = k_D(\delta, y^\delta)$.*

A few remarks are necessary:

(i) The Discrepancy Principle is well defined. In fact, due to Lemma 2.4.2, for every δ and every y^δ such that $\|y^\delta - y\| \leq \delta$ there is always a finite stopping index such that the corresponding residual norm is smaller than $\tau\delta$.

(ii) Since the residual must be computed anyway in the course of the iteration, the Discrepancy Principle requires very little additional computational effort.

The following result is fundamental for the regularization theory of conjugate gradient type methods. For MR and CGNE it was proved for the first time by Nemirovsky in [70], our statement is taken as usual from [27], where a detailed proof using the orthogonal polynomial and spectral theory framework is also given.

Theorem 2.4.4. *Any conjugate gradient type method with parameter $n \geq 1$ with the Discrepancy Principle as a stopping rule is of optimal order, in the sense that it satisfies the conditions of Definition 1.11.2, except for the continuity of the operators R_k.*

It is not difficult to see from the proof of Plato's Theorem 1.11.1 that the discontinuity of R_k does not influence the result. Thus we obtain also a convergence result for $y \in \mathcal{R}(A)$:

Corollary 2.4.2. *Let $y \in \mathcal{R}(A)$ and $\|y^\delta - y\| \leq \delta$. If the stopping index for a conjugate gradient type method with parameter $n \geq 1$ is chosen according to the Discrepancy Principle and denoted by $k_D = k_D(\delta, y^\delta)$, then*

$$\lim_{\delta \to 0} \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \|x_{k_D}^\delta - A^\dagger y\| = 0. \tag{2.67}$$

2.4.2 Regularizing properties of CG and CGME

The case of conjugate gradient type methods with parameter $n = 0$ is much harder to study. The first difficulties arise from the fact that the residual norm is not necessarily decreasing during the iteration, as the following example shows:

Example 2.4.1. *Let $\mathbf{A} \in \mathbb{M}_2(\mathbb{R})$ be defined by*

$$\mathbf{A} = \begin{pmatrix} \tau & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.68}$$

$\tau > 0$, and let $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{y} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Then according to Algorithm 2 we have: $\mathbf{r}_0 = \mathbf{y}$, $A\mathbf{r}_0 = \begin{pmatrix} 2\tau \\ 1 \end{pmatrix}$, $\alpha = \frac{5}{4\tau+1}$ and $\mathbf{x}_1 = \frac{5}{4\tau+1}\mathbf{y}$. Therefore, if τ is

sufficiently small, we have

$$\|\mathbf{y} - A\mathbf{x}_1\| = \left\| \begin{pmatrix} 2 - \frac{10\tau}{4\tau+1} \\ 1 - \frac{5}{4\tau+1} \end{pmatrix} \right\| > \sqrt{5} = \|\mathbf{y}\| = \|\mathbf{y} - A\mathbf{x}_0\|.$$

Moreover, in the ill-posed case it is necessary to restrict to the case where the data y belongs to $\mathcal{R}(A)$ (and not to $\mathcal{D}(A^\dagger)$):

Theorem 2.4.5. *If $y \notin \mathcal{R}(A)$ and $\{x_k\}$ are the iterates of* CG (CGME) *then either the iteration breaks down in the course of the $(\ell + 1)$-th step or $\ell = +\infty$ and $\|x_k\| \to +\infty$ as $k \to +\infty$.*

However, the main problem is that there are examples showing that the Discrepancy Principle does not regularize these methods (see [27], Section 4.2). More precisely, CG and CGME with the Discrepancy Principle as a stopping rule may give rise to a sequence of iterates diverging in norm as δ goes to 0. Thus, other stopping criteria have to be formulated: one of the most important is the following.

Definition 2.4.2. *Fix $\tau > 1$ and assume $\|y - y^\delta\| \le \delta$. Terminate the* CG (CGME) *iteration as soon as $\|y^\delta - Ax_k^\delta\| = 0$, or when for the first time*

$$\sum_{j=0}^{k} \|y^\delta - Ax_j^\delta\|^{-2} \ge (\tau\delta)^{-2}. \tag{2.69}$$

According to Proposition 2.3.7, the index corresponding to this stopping rule is the smallest integer k such that $[\mathsf{p}_k^{[1]}, \mathsf{p}_k^{[1]}]_0^{\frac{1}{2}} \le \tau\delta$, i.e. it is exactly the same stopping index defined for MR (CGNE) by the Discrepancy Principle, thus we denote it again by k_D. The importance of this stopping criterion lies in the following result.

Theorem 2.4.6. *Let y satisfy* (2.62) *and let $\|y - y^\delta\| \le \delta$. If* CG *or* CGME *is applied to y^δ and terminated after k_D steps according to Definition* 2.4.2*, then there exists some uniform constant $C > 0$ such that*

$$\|A^\dagger y - x_{k_D}^\delta\| \le C\rho^{\frac{1}{\mu+1}}\delta^{\frac{\mu}{\mu+1}}. \tag{2.70}$$

Thus, due to Plato's Theorem, except for the continuity of the operator R_k, also CG and CGME are regularization methods of optimal order when they are arrested according to Definition 2.4.2.
We continue with the definition of another very important tool for regularizing ill-posed problems that will turn out to be very useful: the filter factors.

2.5 Filter factors

We have seen in Chapter 1 that the regularized solution of the equation (1.20) can be computed via a formula of the type

$$x_{reg}(\sigma) = \int g_\sigma(\lambda) dE_\lambda A^* y^\delta. \tag{2.71}$$

If the linear operator $A = K$ is compact, then using the singular value expansion of the compact operator the equation above reduces to

$$x_{reg}(\sigma) = \sum_{j=1}^{\infty} g_\sigma(\lambda_j^2)\lambda_j \langle y^\delta, u_j\rangle v_j \tag{2.72}$$

and the sum converges if g_σ satisfies the basic assumptions of Chapter 1. If we consider the operator $\mathscr{U} : \mathcal{Y} \to \mathcal{Y}$ that maps the elements e_j, $j = 1, ... + \infty$, of an orthonormal Hilbert base of $\mathcal{Y}$ into u_j, we see that for $y \in \mathcal{Y}$

$$\mathscr{U}^* y = \sum_{j=1}^{+\infty} \langle u_j, y\rangle \mathsf{e}_j.$$

Then, if $\mathscr{V} : \mathcal{X} \to \mathcal{X}$ is defined in a similar way and $\Lambda(\mathsf{e}_j) := \lambda_j \mathsf{e}_j$, (2.72) can be written in the compact form

$$x_{reg}(\sigma) = \mathscr{V} \Theta_\sigma \Lambda^\dagger \mathscr{U}^* y^\delta, \tag{2.73}$$

with $\Theta_\sigma(\mathsf{e}_j) := g_\sigma(\lambda_j^2)\lambda_j^2 \mathsf{e}_j$.
The coefficients

$$\Phi_\sigma(\lambda_j^2) := g_\sigma(\lambda_j^2)\lambda_j^2, \quad j = 1, ..., +\infty,$$

are known in literature (cf. e.g. [36]) as the *filter factors* of the regularization operator, since they attenuate the errors corresponding to the small singular values λ_j.

Filter factors are very important when dealing with ill-posed and discrete ill-posed problems, because they give an insight into the way a method regularizes the data. Moreover, they can be defined not only for linear regularization methods such as Tikhonov Regularization or Landweber type methods, but also when the solution does not depend linearly on the data, as it happens in the case of Conjugate Gradient type methods, where equation (2.71) does not hold any more. For example, from formula (2.50) we can see that for $n = 0, 1$

$$x_k^{[n]} = \mathsf{q}_{k-1}^{[n]}(A^*A)A^*y^\delta = \mathscr{V}\mathsf{q}_{k-1}^{[n]}(\Lambda^2)\mathscr{V}^*\mathscr{V}\Lambda\mathscr{U}^*y^\delta = \mathscr{V}\mathsf{q}_{k-1}^{[n]}(\Lambda^2)\Lambda^2\Lambda^\dagger\mathscr{U}^*y^\delta, \tag{2.74}$$

so the filter factors of CGME and CGNE are respectively

$$\Phi_k^{[0]}(\lambda_j^2) = \mathsf{q}_{k-1}^{[0]}(\lambda_j^2)\lambda_j^2 \tag{2.75}$$

and

$$\Phi_k^{[1]}(\lambda_j^2) = \mathsf{q}_{k-1}^{[1]}(\lambda_j^2)\lambda_j^2. \tag{2.76}$$

Later on, we shall see how this tool can be used to understand the regularizing properties of the conjugate gradient type methods.

2.6 CGNE, CGME and the Discrepancy Principle

So far, we have given a general overview of the main properties of the conjugate gradient type methods and a stopping rule for every method has been defined.

In the remainder of this chapter, we shall study the behavior of the conjugate gradient type method in discrete ill-posed problems. We will proceed as follows.

- Analyze the performances of CGNE and CGME arrested at the step $k_D = k_D(\delta, y^\delta)$, i.e. respectively with the Discrepancy Principle (cf. Definition 2.4.1) and with the a-posteriori stopping rule proposed by Hanke (cf. Definition 2.4.2). This will be the subject of the current section.

- Give an insight of the regularizing properties of CGME and CGNE by means of the filter factors (cf. Section 2.7).

- Analyze the performances obtained by the method with parameter $n = 2$ (cf. Section 2.8).

Discrete ill-posed problems are constructed very easily using P.C. Hansen's Regularization Tools [35], cf. the Appendix.
As an illustrative example, we consider the test problem $\mathsf{heat(N)}$ in our preliminary test, which will be called Test 0 below.

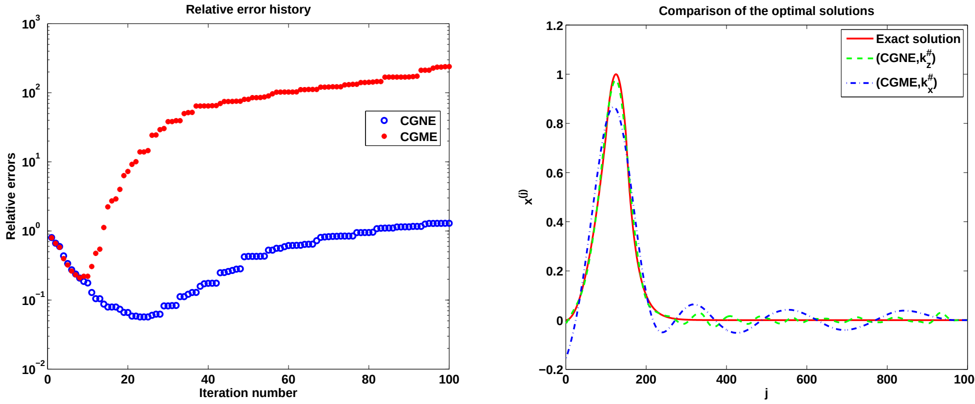
2.6.1 Test 0

The Matlab command

$$[\mathsf{A}, \mathsf{b}, \mathsf{x}] = \mathsf{heat}(\mathsf{N})$$

generates the matrix $\mathbf{A} \in \mathbb{GL}_N(\mathbb{R})$ ($\mathbf{A}$ is not symmetric in this case!), the exact solution $\mathbf{x}^\dagger$ and the right-hand side vector $\mathbf{b}$ of the artificially constructed ill-posed linear system $\mathbf{Ax} = \mathbf{b}$. More precisely, it provides the discretization of a Volterra integral equation of the first kind related to an inverse heat equation, obtained by simple collocation and midpoint rule with N points (cf. [35] and the references therein). The inverse heat equation is a well known ill-posed problem, see e.g. [17], [61] and [62].
After the construction of the exact underlying problem, we perturb the exact data with additive white noise, by generating a multivariate gaussian vector $\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, by defining a number $\varrho \in]0,1[$ representing the percentage of noise on the data and by setting

$$\mathbf{b}^\delta := \mathbf{b} + \mathbf{e}, \quad \text{with } \mathbf{e} = \frac{\varrho\|\mathbf{b}\|}{\|\mathbf{E}\|}\mathbf{E}. \tag{2.77}$$



(a) Relative errors history

(b) Optimal Solutions

Figure 2.1: Test 0: relative errors (on the left) and optimal solutions (on the right)

Here and below, $\mathbf{0}$ is the constant column vector whose components are equal to 0 and $\mathbf{I}_N$ is the identity matrix of dimension $N \times N$.

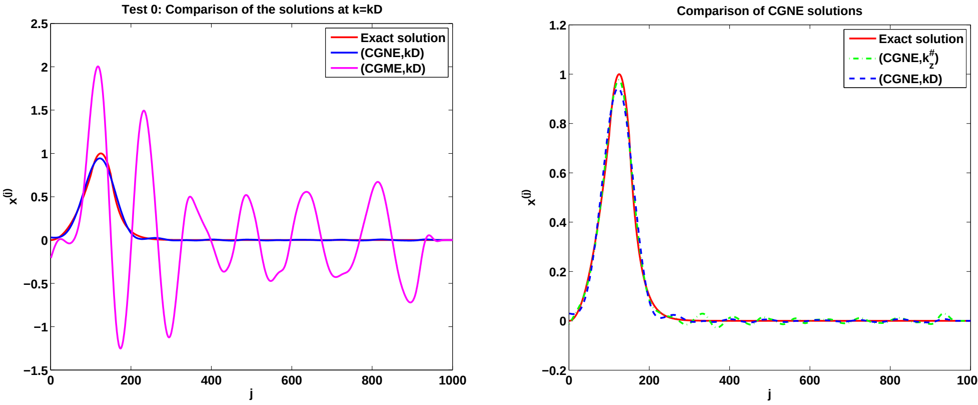
Of course, from the equation above there follows immediately that $\delta = \varrho \|\mathbf{b}\|$ and $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \delta \mathbf{I}_N)$. In this case, since $\|\mathbf{b}\| = 1.4775$ and ϱ is chosen equal to 1%, $\delta = 1.4775 \times 10^{-2}$.

Next, we solve the linear system with the noisy data $\mathbf{b}^\delta$ performing $k_{MAX} = 40$ iteration steps of algorithm 6, by means of the routine cgne_cgme defined in the Appendix. The parameter $\tau > 1$ of the Discrepancy Principle is fixed equal to 1.001. Looking at Figure 2.1 we can compare the relative errors of CGME (red stars) and CGNE (blue circles) in the first 30 iteration steps.

Denoting with $\mathbf{x}_k^\delta$ the CGME iterates and with $\mathbf{z}_k^\delta$ the CGNE iterates we observe:

1. The well known phenomenon of semi-convergence is present in both algorithms, but appears with stronger evidence in CGME than in CGNE.

2. If $k_x^\sharp(\delta)$ and $k_z^\sharp(\delta)$ are defined as the iteration indices at which, respectively, CGME and CGNE attain their best approximation of $\mathbf{x}^\dagger$, the numerical results show that $k_x^\sharp(\delta) = 8$ and $k_z^\sharp(\delta) = 24$. The correspon-



(a) Solutions at $k = k_D$

(b) Discrepancy and optimal solutions for CGNE

Figure 2.2: Test 0: comparison between the solutions of CGNE and CGME at $k = k_D$ (on the left) and between the discrepancy and the optimal solution of CGNE (on the right).

ding relative errors are approximately equal to

$$\varepsilon_x^\sharp = 0.2097, \quad \varepsilon_z^\sharp = 0.0570,$$

so CGNE achieves a better approximation, although to obtain its best result it has to perform 16 more iteration steps than CGME.

3. Calculating the iteration index defined by Morozov's Discrepancy Principle we get $k_D := k_D(\delta, \mathbf{b}^\delta) = 15$: the iterates corresponding to this index are the solutions of the regularization methods in Definition 2.4.2 and Definition 2.4.1 (respectively the a-posteriori rule proposed by Hanke and Morozov's Discrepancy Principle) and the corresponding relative errors are approximately equal to

$$\varepsilon_{xD} = 2.2347, \quad \varepsilon_{zD} = 0.0794.$$

Therefore, even if the stopping rule proposed by Hanke makes CGME a regularization method of optimal order, in this case it finds a very unsatisfying solution (cf. its oscillations in the left picture of Figure 2.2).

Moreover, from the right of Figure 2.2 we can see that CGNE arrested with the Discrepancy Principle gives a slightly oversmoothed solution compared to the optimal one, which provides a better reconstruction of the maximum and of the first components of $\mathbf{x}^\dagger$ at the price of some small oscillations in the last components.

Although we chose a very particular case, many of the facts we have described above hold in other examples as well, as we can see from the next more significant test.

2.6.2 Test 1

Test 1									
1D Test Problems									
	N	noise	$k_x^\sharp$	$k_z^\sharp$	k_D	$\varepsilon_x^\sharp$	$\varepsilon_z^\sharp$	$\varepsilon_{xD}^\sharp$	$\varepsilon_{zD}^\sharp$
Baart	1000	0.1%	3	7	4	0.1659	0.0893	1.8517	0.1148
Deriv2	1000	0.1%	9	27	21	0.2132	0.1401	1.8986	0.1460
Foxgood	1000	0.1%	2	5	3	0.0310	0.0068	0.4716	0.0070
Gravity	1000	0.1%	6	13	11	0.0324	0.0083	1.0639	0.0104
Heat	1000	0.1%	18	37	33	0.0678	0.0174	0.8004	0.0198
I-laplace	1000	0.1%	11	38	19	0.2192	0.1856	1.7867	0.1950
Phillips	1000	0.1%	4	12	9	0.0243	0.0080	0.1385	0.0089
Shaw	1000	0.1%	6	14	8	0.0853	0.0356	0.2386	0.0474
2D Test Problems									
	N	noise	$k_x^\sharp$	$k_z^\sharp$	k_D	$\varepsilon_x^\sharp$	$\varepsilon_z^\sharp$	$\varepsilon_{xD}^\sharp$	$\varepsilon_{zD}^\sharp$
Blur	2500	2.0%	9	12	7	0.1089	0.1016	0.1161	0.1180
Tomo	2500	2.0%	13	22	11	0.2399	0.2117	0.2436	0.2450

Table 2.1: Numerical results for Test 1.

We consider 10 different medium size test problems from [35]. The same algorithm of Test 0 is used apart from the choices of the test problem and of the parameters ϱ, N and $k_{MAX} = 100$. In all examples the white gaussian

noise is generated using the Matlab function rand and the seed is chosen equal to 0. The results are gathered in Table 2.1.
Looking at the data, one can easily notice that the relations

$$k_x^\sharp < k_z^\sharp, \quad \varepsilon_x^\sharp > \varepsilon_z^\sharp, \quad k_D < k_z^\sharp$$

hold true in all the examples considered. Thus it is natural to ask if they are always verified or counterexamples can be found showing opposite results. Another remark is that very often $k_D > k_x^\sharp$ and in this case the corresponding error is very huge.

2.6.3 Test 2

The following experiment allows to answer the questions asked in Test 1 and substantially confirms the general remarks we have made so far.

For each of the seven problems of Table 2.2 we choose 10 different values for each of the parameters $N \in \{100, 200, ..., 1000\}$, $\varrho \in \{0.1\%, 0.2\%, ..., 1\%\}$ and the Matlab seed $\in \{1, ..., 10\}$ for the random components of the noise on the exact data. In each case we compare the values of $k_x^\sharp$ and $k_z^\sharp$ with k_D and the values of $\varepsilon_x^\sharp$ with $\varepsilon_z^\sharp$. The left side of the table shows how many times, for each test problem, $\varepsilon_z^\sharp < \varepsilon_x^\sharp$ and vice versa. The right-hand side shows how many times, for each problem and for each method, the stopping index k_D is smaller, equal or larger than the optimal one. In this case the value of τ has been chosen equal to $1 + 10^{-15}$. From the results, summarized in Table 2.2, we deduce the following facts.

- It is possible, but very unlikely, that $\varepsilon_x^\sharp < \varepsilon_z^\sharp$ (this event has occurred only 22 times out of 7000 in Test 2 and only in the very particular test problem foxgood, which is severely ill-posed).
- The trend that emerged in Test 0 and Test 1 concerning the relation between k_D and the optimal stopping indices $k_x^\sharp$ and $k_z^\sharp$ is confirmed in Test 2. The stopping index k_D provides usually (but not always!) a

Test 2						
	Best err. perf.		Stopping			
	CGNE	CGME		$k_D < k^\sharp$	$k_D = k^\sharp$	$k_D > k^\sharp$
Baart	1000	0	CGNE	564	372	64
			CGME	0	545	455
Deriv2	1000	0	CGNE	882	112	6
			CGME	0	0	1000
Foxgood	978	22	CGNE	483	426	91
			CGME	0	532	468
Gravity	1000	0	CGNE	861	118	21
			CGME	0	1	999
Heat	1000	0	CGNE	991	9	0
			CGME	0	1	999
Phillips	1000	0	CGNE	751	207	42
			CGME	0	48	952
Shaw	1000	0	CGNE	806	185	9
			CGME	0	4	996
Total	6978	22	CGNE	5338	1429	233
			CGME	0	1031	5869

Table 2.2: Numerical results for Test 2.

slightly oversmoothed solution for CGNE and often a noise dominated solution for CGME.

In the problems with a symmetric and positive definite matrix $\mathbf{A}$, it is also possible to compare the results of CGNE and CGME with those obtained by MR and CG. This was done for phillips, shaw, deriv2 and gravity and the outcome was that CGNE attained the best performance 3939 times out of 4000, with 61 successes of MR in the remaining cases.

In conclusion, the numerical tests described above lead us to ask the following questions:

1. The relations $k_x^\sharp < k_z^\sharp$ and $\varepsilon_x^\sharp > \varepsilon_z^\sharp$ hold very often in the cases consi-

dered above. Is there a theoretical justification of this fact?

2. The conjugate gradient methods with parameter $n = 1$ seem to provide better results than those with parameter $n = 0$. What can we say about other conjugate gradient methods with parameter $n > 1$?

3. To improve the performance of CGME one can choose a larger τ. This is not in contrast with the regularization theory above. On the other hand, arresting CGNE later means stopping the iteration when the residual norm has become smaller than δ, while the Discrepancy Principle states that τ must be chosen larger than 1. How can this be justified and implemented in practice by means of a reasonable stopping rule?

We will answer the questions above in detail.

2.7 CGNE vs. CGME

In the finite dimensional setting described in Section 2.6, both iterates of CGME and CGNE will eventually converge to the vector $\tilde{\boldsymbol{x}} := \mathbf{A}^\dagger \mathbf{b}^\delta$ as described in Section 2.1, which can be very distant from the exact solution $\mathbf{x}^\dagger = \mathbf{A}^\dagger \mathbf{b}$ we are looking for, since $\mathbf{A}^\dagger$ is ill-conditioned. The problem is to understand how the iterates converge to $\tilde{\boldsymbol{x}}$ and how they reach an approximation of $\mathbf{x}^\dagger$ in their first steps.

First of all, we recall that $\mathbf{x}_k^\delta$ minimizes the norm of the error $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ in $\mathcal{K}_{k-1}(\mathbf{A}^*\mathbf{A}; \mathbf{A}^*\mathbf{b}^\delta)$, whereas $\mathbf{z}_k^\delta$ minimizes the residual norm $\|\mathbf{A}\mathbf{x} - \mathbf{b}^\delta\|$ in the same Krylov space. Thus the iterates of CGME, converging to $\tilde{\mathbf{x}}$ as fast as possible, will be the better approximations of the exact underlying solution $\mathbf{x}^\dagger$ in the very first steps, when the noise on the data still plays a secondary role. However, being the greediest approximations of the noisy solution $\tilde{\mathbf{x}}$, they will also be influenced by the noise at an earlier stage than the iterates of CGNE. This explains the relation $k_x^\sharp < k_z^\sharp$, which is often verified in the numerical experiments.

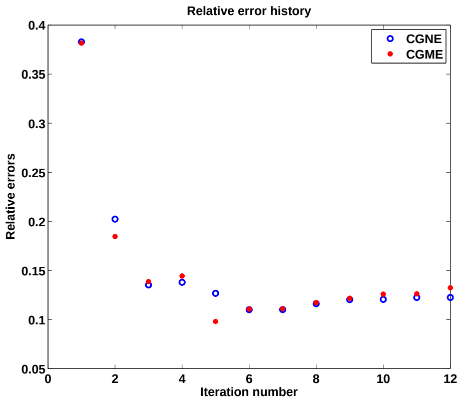


Figure 2.3: Relative error history for CGNE and CGME with a perturbation of the type $\bar{\mathbf{e}} = \mathbf{A}\bar{\mathbf{w}}$, for the problem phillips(1000). CGME achieves the better approximation ($\varepsilon_x^\sharp = 0.0980$, $\varepsilon_x^\sharp = 0.1101$).

Moreover, expanding the quantities minimized by the methods in terms of the noise $\mathbf{e}$, we get

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\mathbf{x} - \mathbf{x}^\dagger - \mathbf{A}^\dagger \mathbf{e}\|$$

for CGME and

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}^\delta\| = \|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{e}\|$$

for CGNE. In the case of CGME, the error is amplified by the multiplication with the matrix $\mathbf{A}^\dagger$. As a consequence, in general CGME will obtain a poorer reconstruction of the exact solution $\mathbf{x}^\dagger$, because its iterates will be more sensible to the amplification of the noise along the components relative to the small singular values of $\mathbf{A}$.

This justifies the relation $\varepsilon_x^\sharp > \varepsilon_z^\sharp$ verified in almost all the numerical experiments above. We observe that these considerations are based on the remark that the components of the random vector $\mathbf{e}$ are approximately of the same size. Indeed, things can change significantly if a different kind of perturbation is chosen (e.g. the SVD components of the noise $\mathbf{e}$ decay like $O(\lambda_j)$). To show this, consider the test problem phillips(1000), take $\varrho = 5\%$, define $\bar{\mathbf{e}} = \mathbf{A}\bar{\mathbf{w}}$, where $\bar{\mathbf{w}}$ is the exact solution of the problem heat(1000) and put $\mathbf{b}^\delta = \mathbf{b} + \bar{\mathbf{e}}$: from the plot of the relative errors in Figure 2.3 it is clear

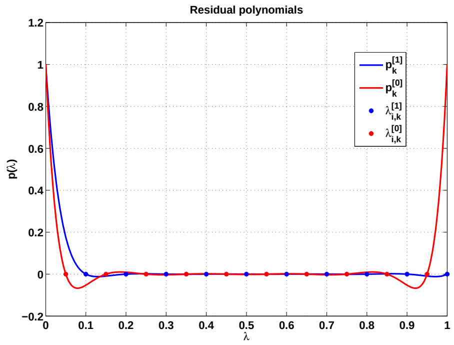


Figure 2.4: Residual polynomials for CGNE (blue line) and CGME (red line)

that in this case CGME obtains the best performance and it is not difficult to construct similar examples leading to analogous results.

This example suggests that it is almost impossible to claim that a method works better than the other one without assuming important restrictions on $\mathbf{A}$, δ and $\mathbf{b}^\delta$ and on the perturbation $\mathbf{e}$. Nevertheless, a general remark can be done from the analysis of the filter factors of the methods. In [36] P. C. Hansen describes the regularizing properties of CGNE by means of the filter factors, showing that in the first steps it tends to reconstruct the components of the solution related to the low frequency part of the spectrum. The analysis is based on the convergence of the Ritz values $\lambda_{i,k}^{[1]}$ (the zeros of the residual polynomial $\mathsf{p}_k^{[1]}$) to the singular values of the operator $\mathbf{A}$. From the plot of the residual polynomials (cf. Figure 2.4) and from the interlacing properties of their roots we can deduce that the iterates of CGME and CGNE should behave in a similar way. The main difference is the position of the roots, which allows us to compare the filter factors of the methods.

Theorem 2.7.1. *Let A be a linear compact operator between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and let y^δ be the given perturbed data of the underlying exact equation $Ax = y$. Let $\{\lambda_j; u_j, v_j\}_{j\in\mathbb{N}}$ be a singular system for A. Denote with x_k^δ and z_k^δ the iterates of* CGME *and* CGNE *corresponding to y^δ respectively, with $\mathsf{p}_k^{[0]}$ and $\mathsf{p}_k^{[1]}$ the corresponding residual polynomials and with $\Phi_k^{[0]}(\lambda_j)$ and $\Phi_k^{[1]}(\lambda_j)$ the filter factors. Let also $\lambda_{i,k}^{[0]}$, $i = 1, ..., k$ and $\lambda_{i,k}^{[1]}$, $i = 1, ..., k$ be the*

zeros of $\mathsf{p}_k^{[0]}$ *and* $\mathsf{p}_k^{[1]}$ *respectively.*
Then, for every j *such that* $\lambda_j^2 < \lambda_{1,k}^{[0]}$,

$$\Phi_k^{[0]}(\lambda_j) > \Phi_k^{[1]}(\lambda_j). \tag{2.78}$$

Proof. The filter factors of the conjugate gradient type methods are:

$$\Phi_k^{[n]}(\lambda_j) = \mathsf{q}_k^{[n]}(\lambda_j^2)\lambda_j^2, \quad n = 0, 1.$$

We recall from the theory of orthogonal polynomials that the zeros of $\mathsf{p}_k^{[0]}$ and of $\mathsf{p}_k^{[1]}$ interlace as follows:

$$\lambda_{1,k}^{[0]} < \lambda_{1,k}^{[1]} < \lambda_{2,k}^{[0]} < \lambda_{2,k}^{[1]} < ... < \lambda_{k,k}^{[0]} < \lambda_{k,k}^{[1]}. \tag{2.79}$$

Thus, writing down the residual polynomials in the form

$$\mathsf{p}_k^{[n]}(\lambda) = \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\lambda_{j,k}^{[n]}} \right), \quad n = 0, 1, \tag{2.80}$$

it is very easy to see that $\mathsf{p}_k^{[0]} < \mathsf{p}_k^{[1]}$ on $]0, \lambda_{1,k}^{[0]}]$ (cf. Figure 2.4) and consequently

$$\mathsf{q}_k^{[0]} > \mathsf{q}_k^{[1]} \quad \text{on} \quad]0, \lambda_{1,k}^{[0]}].$$

□

This result is a theoretical justification of the heuristic considerations of the beginning of this section: the iterates of CGNE filter the high frequencies of the spectrum slightly more than the iterates of CGME. Summing up:

- Thanks to its minimization properties, CGNE works better than CGME along the high frequency components, keeping the error small for a few more iteration and usually achieving the better results.

- Anyway this is not a general rule (see the counterexample of this section and the results of Test 2), because the performances of the two methods strongly depend on the matrix $\mathbf{A}$ and on the vectors $\mathbf{x}^{\dagger}$, $\mathbf{b}$, $\mathbf{e}$ and $\mathbf{x}_0$.

- Finding a particular class of problems (or data) in which CGNE always gets the better results is maybe possible, but rather difficult.

2.8 Conjugate gradient type methods with parameter n=2

We now turn to the question about the conjugate gradient type methods with parameter $n > 1$, restricting to the case $n = 2$ for AA^*.
From the implementation of the corresponding method outlined in Algorithm 7 and performed by the routine cgn2 defined in the Appendix, we can see that the computation of a new iterate requires 4 matrix-vector multiplications at each iteration step, against the only 2 needed by CGNE and CGME.

On the other hand, it is obvious that the same analysis of Section 2.7 will suggest that this method filters the high frequency components of the spectrum even better than CGNE, because of the relation $\lambda_{i,k}^{[1]} < \lambda_{i,k}^{[2]}$, valid for all $i = 1, ..., k$. As a matter of fact, the phenomenon of the semi-convergence appears more attenuate here than in the case of CGNE, as we can see e.g. from Figure 2.5, where a plot of the relative errors of both methods in the same assumptions of Test 0 of Section 2.6 has been displayed. Thus, the conjugate gradient type method with parameter $n = 2$ is more stable than

Algorithm 7 CG type method with parameter $\mathbf{n = 2}$ for $\mathbf{AA^*}$

```
$r_0 = y - Ax_0$;
$d = A^* r_0$;
$p_2 = Ad$;
$m_1 = p_2$;
$m_2 = A^* m_1$;
$k = 0$;
while (not stop) do
    $\alpha = \|p_2\|^2 / \|m_2\|^2$;
    $x_{k+1} = x_k + \alpha d$;
    $r_{k+1} = r_k - \alpha m_1$;
    $p_1 = A^* r_{k+1}$;
    $t = Ap_1$;
    $\beta = \|t\|^2 / \|p_2\|^2$;
    $d = p_1 + \beta d$;
    $m_1 = Ad$;
    $m_2 = A^* m_1$;
    $p_2 = t$;
    $k = k + 1$;
end while
```

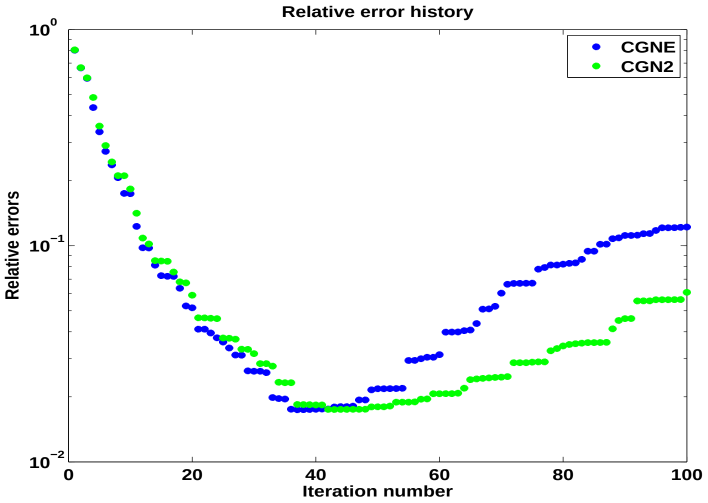


Figure 2.5: Relative error history for CGNE and the conjugate gradient type method with parameter $n = 2$ in the assumptions of Test 0 of Section 2.6.

CGNE with respect to the iteration index k (exactly for the same reasons why we have seen that CGNE is more stable than CGME). This could be an advantage especially when the data are largely contaminated by the noise. For example, if we consider the test problem blur(500), with $\varrho = 10\%$, we can see from Figure 2.6 that the optimal reconstructed solutions of both methods are similar, but the conjugate gradient type method with parameter $n = 2$ attenuates the oscillations caused by the noise in the background better than CGNE.

2.8.1 Numerical results

We compare the conjugate gradient type methods for the matrix $\mathbf{AA}^*$ with parameters 1 and 2 in the same examples of Test 2 of Section 2.6, by adding the test problem i_laplace. From the results of Table 2.3, we can see that the methods obtain quite similar results. The conjugate gradient type method with parameter $n = 2$ usually performs a little bit better, but this advantage is minimal: the average improvement of the results obtained by the method with $n = 2$, namely the difference of the total sums of the relative errors

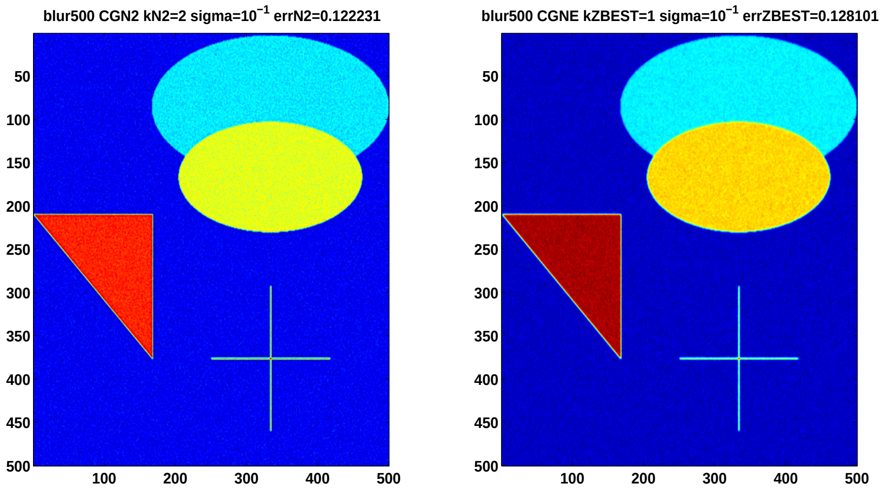


Figure 2.6: Comparison between the conjugate gradient type method with parameter $n = 2$ and CGNE for the test problem blur(500), with $\varrho = 10\%$.

divided by the total sum of the relative errors of CGNE, is equal to

$$\frac{|683.74 - 686.74|}{686.74} \cong 0.4\%.$$

Concerning the stopping index, we observe that in both cases the Discrepancy Principle stops the iteration earlier than the optimal stopping index in the large majority of the considered cases. We shall return to this important topic later.

Our numerical experiments confirm the trend also for a larger noise. Performing the same test with $\varrho \in \{10^{-2}, 2 \times 10^{-2}, ..., 10^{-1}\}$ instead of $\varrho \in \{10^{-3}, 2 \times 10^{-3}, ..., 10^{-2}\}$, we obtain that the method with parameter $n = 2$ achieves the better relative error in 4763 cases (59.5% of the times) and the overall sums of the relative errors are 1101.8 for $n = 2$ and 1115.5 for $n = 1$. Thus the average improvement obtained by the method with $n = 2$ is 1% in this case.

In conclusion, the conjugate gradient type method with parameter $n = 2$ has nice regularizing properties: in particular, it filters the high frequency components of the noise even better than CGNE. Consequently, it often achieves

Comparison of CG type methods: $\mathbf{n = 1}$ and $\mathbf{n = 2}$						
	Best err. perf.	Average rel. err.	n	Discrepancy Stopping		
				$k_D < k^\sharp$	$k_D = k^\sharp$	$k_D > k^\sharp$
Baart	598	0.13404	$n = 1$	564	372	64
	402	0.13482	$n = 2$	616	321	63
Deriv2	197	0.19916	$n = 1$	882	112	6
	803	0.19792	$n = 2$	965	33	2
Foxgood	565	0.01862	$n = 1$	483	426	91
	435	0.01876	$n = 2$	547	362	91
Gravity	386	0.02179	$n = 1$	861	118	21
	614	0.02151	$n = 2$	838	147	15
Heat	294	0.05709	$n = 1$	991	9	0
	706	0.05657	$n = 2$	996	4	0
I-laplace	447	0.18091	$n = 1$	957	30	13
	553	0.18036	$n = 2$	971	18	11
Phillips	441	0.01774	$n = 1$	751	207	42
	559	0.01731	$n = 2$	775	204	21
Shaw	548	0.05739	$n = 1$	806	185	9
	452	0.05649	$n = 2$	837	156	7
Total	3476	0.08584	$n = 1$	6285	1459	256
	4524	0.08546	$n = 2$	6545	1245	210

Table 2.3: Comparison between the CG type methods for $\mathbf{AA}^*$ with parameters 1 and 2. We emphasize that k_D is not the same stopping index here for $n = 1$ and $n = 2$.

the better results and keeps the phenomenon of the semi-convergence less pronounced, especially for large errors in the data. On the other hand, it is more expensive than CGNE from a computational point of view, because it usually performs more steps to reach the optimal solution and in each step it requires 4 matrix-vector multiplications (against the only 2 required by CGNE). Despite the possible advantages described above, in our numerical tests the improvements were minimal, even in the case of a large δ. For this reason, we believe that it should be rarely worth it, to prefer the method with parameter $n = 2$ to CGNE.

Chapter 3

New stopping rules for CGNE

In the last sections of Chapter 2 we have seen that CGNE, being both efficient and precise, is one of the most promising conjugate gradient type methods when dealing with (discrete) ill-posed problems.

The general theory suggests the Discrepancy Principle as a very reliable stopping rule, which makes CGNE a regularization method of optimal order[1]. Of course, the stopping index of the Discrepancy Principle is not necessarily the best possible for a given noise level $\delta > 0$ and a given perturbed data y^δ. Indeed, as we have seen in the numerical tests of Chapter 2, it usually provides a slightly oversmoothed solution. Moreover, in practice the noise level is often unknown: in this case it is necessary to define heuristic stopping rules. Due to Bakushinskii's Theorem a method arrested with an heuristic stopping rule cannot be convergent, but in some cases it can give more satisfactory results than other methods arrested with a sophisticated stopping rule of optimal order based on the knowledge of the noise level.

When dealing with discrete ill-posed problems (e.g. arising from the discretization of ill-posed problems defined in a Hilbert space setting), it is very important to rely on many different stopping rules, in order to choose the best one depending on the particular problem and data: among the most famous

[1] except for the continuity of the operator that maps the data into the k-th iterate of CGNE, cf. Chapter 2.

stopping rules that can be found in literature, apart from the Discrepancy Principle, we mention the Monotone Error Rule (cf. [24], [23], [25], [36]) and, as heuristic stopping rules, the L-curve ([35], [36]), the Generalized Cross Validation ([17] and the references therein) and the Hanke-Raus Criterion ([27], [24]).

In this chapter, three new stopping rules for CGNE will be proposed, analyzed and tested. All these rules rely on a general analysis of the residual norm of the CGNE iterates.

The first one, called the *Approximated Residual L-Curve Criterion*, is an heuristic stopping rule based on the global behavior of the residual norms with respect to the iteration index.

The second one, called the *Projected Data Norm Criterion*, is another heuristic stopping rule that relates the residual norms of the CGNE iterates to the residual norms of the truncated singular value decomposition.

The third one, called the *Projected Noise Norm Criterion*, is an a-posteriori stopping rule based on a statistical approach, intended to overcome the oversmoothing effect of the Discrepancy Principle and mainly bound to large scale problems.

3.1 Residual norms and regularizing properties of CGNE

This section is dedicated to a general analysis of the regularizing properties of CGNE, by linking the relative error with the residual norm in the case of perturbed data.

In their paper [60] Kilmer and Stewart related the residual norm of the minimal residual method to the norm of the relative error.

Theorem 3.1.1 (Kilmer, Stewart). *Let the following assumptions hold:*

- *The matrix $\mathbf{A} \in \mathbb{GL}_N(\mathbb{R})$ is symmetric and positive definite, and in the coordinate system of its eigenvectors the exact linear system can be* *written in the form*

$$\mathbf{\Lambda x} = \mathbf{b}, \tag{3.1}$$

where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, ..., \lambda_N\}$, $1 = \lambda_1 > \lambda_2 > ... > \lambda_N > 0$.

- *The exact data are perturbed by an additive noise* $\mathbf{e} \in \mathbb{R}^N$ *such that its components* e_i *are random variables with mean* 0 *and standard deviation* $\nu > 0$.

- *For a given* $\delta > 0$*, let* $\mathbf{y}$ *be the purported solution with residual norm* δ *minimizing the distance from the exact solution* $\mathbf{x}$*, i.e.* $\mathbf{y}$ *solves the problem*[2]

$$\begin{aligned} &\textit{minimize} \quad \|\mathbf{x} - \mathbf{y}\| \\ &\textit{subject to} \quad \|\mathbf{b}^\delta - \mathbf{\Lambda y}\|^2 = \delta^2. \end{aligned} \tag{3.2}$$

If $c > -1$ *solves the equation*

$$\sum_{i=1}^{N} \frac{\mathrm{e}_i^2}{(1 + c\lambda_i^2)^2} = \delta^2, \tag{3.3}$$

then the vector $\mathbf{y}(c)$*, with components*

$$\mathrm{y}_i(c) := \mathrm{x}_i + \frac{c\lambda_i \mathrm{e}_i}{1 + c\lambda_i^2}, \tag{3.4}$$

is a solution of (3.2) *and*

$$\|\mathbf{x} - \mathbf{y}(c)\|^2 = \sum_{i=1}^{N} \left(\frac{c\lambda_i \mathrm{e}_i}{1 + c\lambda_i^2} \right)^2. \tag{3.5}$$

Note that the solution of (3.2) is Tikhonov's regularized solution with parameter $\delta^2 > 0$, where δ satisfies (3.3).
As c varies from -1 to $+\infty$, the residual norm decreases monotonically from $+\infty$ to 0 and the error norm $\|\mathbf{x} - \mathbf{y}(c)\|$ decreases from ∞ to 0 at $c = 0$ when $\delta = \|\mathbf{e}\|$, but then increases rapidly (for further details, see the

[2] here the perturbed data $\mathbf{b}^\delta = \mathbf{b} + \mathbf{e}$ does not necessarily satisfy $\|\mathbf{b}^\delta - \mathbf{b}\| = \delta$.

considerations after Theorem 3.1 in [60]). As a consequence, choosing a solution with residual norm smaller than $\|\mathbf{e}\|$ would result in large errors, so the theorem provides a theoretical justification for the discrepancy principle in the case of the Tikhonov method.

However, the solution $\mathbf{y}(c)$ can differ significantly from the iterates of the conjugate gradient type methods.

The following simulation shows that the results of Kilmer and Stewart cannot be applied directly to the CGNE method and introduces the basic ideas behind the new stopping rules that are going to be proposed.

Fix $N = 1000$ and $p = 900$ and let

$$\mathbf{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_p, \lambda_{p+1}, ..., \lambda_N\} \tag{3.6}$$

be the diagonal matrix such that

$$\lambda_1 > ... > \lambda_p >> \lambda_{p+1} > ... > \lambda_N > 0 \tag{3.7}$$

and

$$\lambda_i \sim \begin{cases} 10^{-2} & i = 1, ..., p, \\ 10^{-8} & i = p+1, ..., N. \end{cases} \tag{3.8}$$

Let $\boldsymbol{\lambda}$ be the vector whose components are the λ_i and indicate

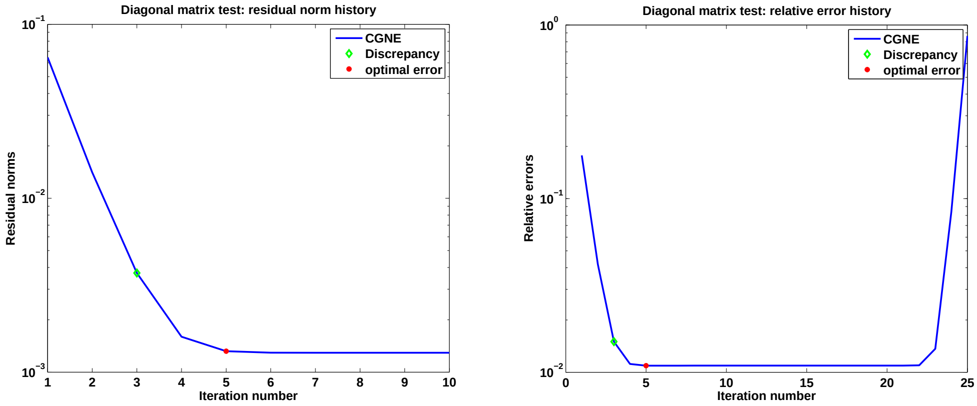
$$\boldsymbol{\lambda}_p := \begin{pmatrix} \lambda_1 \\ ... \\ \lambda_p \end{pmatrix}, \qquad \boldsymbol{\lambda}_{N-p} := \begin{pmatrix} \lambda_{p+1} \\ ... \\ \lambda_N \end{pmatrix}. \tag{3.9}$$

Accordingly, set also

$$\mathbf{e} = \begin{pmatrix} \mathbf{e}_p \\ \mathbf{e}_{N-p} \end{pmatrix}, \qquad \mathbf{b}^\delta = \mathbf{b} + \mathbf{e} = \begin{pmatrix} \mathbf{b}_p^\delta \\ \mathbf{b}_{N-p}^\delta \end{pmatrix}, \tag{3.10}$$

where $\mathbf{b} = \mathbf{\Lambda}\mathbf{x}^\dagger$ and $\mathbf{x}^\dagger$ is the exact solution of the test problem gravity from P.C. Hansen's Regularization Tools.

The left picture of Figure 3.1 shows the graphic of the residual norm of the CGNE iterates with respect to the iteration index for this test problem with noise level $\varrho = 1\%$: we note that this graphic has the shape of an L.



(a) Residual norm history

(b) Relative error norm history

Figure 3.1: Residual and relative error norm history in a diagonal matrix test problem with two cluster of singular values and $\varrho = 1\%$.

In general, a similar L-shape is observed in the discrete ill-posed problems, thanks to the rapid decay of the singular values, as described in [76] and [77]. In fact, in the general case of a non diagonal and non symmetric matrix $\mathbf{A}$, for $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ and $\|\mathbf{b} - \mathbf{b}^\delta\| \leq \delta$, $\delta > 0$, combining (2.59) and (2.60) we have

$$\|\mathbf{A}\mathbf{z}_k^\delta - \mathbf{b}^\delta\| \leq \|F_{\lambda_{1,k}}\mathbf{e}\| + |\mathsf{p}_k'(0)|^{-1/2}\|\mathbf{x}^\dagger\| \leq \delta + |\mathsf{p}_k'(0)|^{-1/2}\|\mathbf{x}^\dagger\|. \tag{3.11}$$

Since $|\mathsf{p}_k'(0)|$ is the sum of the reciprocals of the Ritz values at the k-th step and $\lambda_{1,k}$ is always smaller than λ_k, a very rough estimate of the residual norm is given by $\delta + \|\mathbf{x}^\dagger\|\lambda_k^{1/2}$, which has an L-shape if the eigenvalues of $\mathbf{A}$ decay quickly enough. Thus the residual norm curve must lie below this L-shaped curve and for this reason it is often L-shaped too.

We consider now the numerical results of the simulation. Comparing the solution obtained by the Discrepancy Principle (denoted by the subscript D) with the optimal solution (denoted with the superscript $\sharp$) we have:

- $k^\sharp = 5$ and $k_D = 3$ for the stopping indices;
- $\|\mathbf{b}^\delta - \mathbf{\Lambda}\mathbf{z}_{k^\sharp}^\delta\| \sim 1.32 \times 10^{-3}$ and $\|\mathbf{b}^\delta - \mathbf{\Lambda}\mathbf{z}_{k_D}^\delta\| \sim 3.71 \times 10^{-3}$ for the residual norms;

- $\varepsilon^\sharp \sim 1.09 \times 10^{-2}$ and $\varepsilon_D \sim 1.50 \times 10^{-2}$ for the relative error norms (a plot of the relative error norm history is shown in the right picture of Figure 3.1).

Note that $\|\mathbf{e}_{N-p}\| \sim 1.29 \times 10^{-3}$ is very close to $\|\mathbf{b}^\delta - \mathbf{\Lambda}\mathbf{z}^\delta_{k^\sharp}\|$: this suggests to stop the iteration as soon as the residual norm is lower or equal to $\tau\|\mathbf{e}_{N-p}\|$ for a suitable constant $\tau > 1$ (instead of $\tau\delta$, as in the Discrepancy Principle). This remark can be extended to the general case of a discrete ill-posed problem. In fact, the stopping index of the discrepancy principle is chosen large enough so that $\|\mathbf{x}^\dagger\||\mathsf{p}'_k(0)|^{-1/2}$ is lower than $(\tau - 1)\delta$ in the residual norm estimate (3.11) and small enough so that the term $\delta|\mathsf{p}'_k(0)|^{1/2}$ is as low as possible in the error norm estimate (2.64) in Chapter 2. However, in the sharp versions of these estimates with δ replaced by $\|F_{\lambda_{1,k}}(\mathbf{b}^\delta - \mathbf{b})\|$, when k is close to the optimal stopping index $k^\sharp$, $F_{\lambda_{1,k}}$ is the projection of the noise onto the high frequency part of the spectrum and the quantity $\|F_{\lambda_{1,k}}\mathbf{e}\|$ is a reasonable approximation of the residual norm threshold $\mathbf{e}_{N-p}$ considered in our simulation with the diagonal matrix.
Summarizing:

- the behavior of the residual norm plays an important role in the choice of the stopping index: usually its plot with respect to k has the shape of an L (the so called *Residual L-curve*);

- the norm of the projection of the noise onto the high frequency part of the spectrum may be chosen to replace the noise level δ as a residual norm threshold for stopping the iteration.

3.2 SR1: Approximated Residual L-Curve Criterion

In Section 3.1 we have seen that the residual norms of the iterates of CGNE tend to form an L-shaped curve. This curve, introduced for the first time by Reichel and Sadok in [76], differs from the famous *Standard L-Curve* considered e.g. in [33], [34] and [36], which is defined by the points

$$(\eta_k, \rho_k) := (\|\mathbf{z}_k^\delta\|, \|\mathbf{b}^\delta - \mathbf{A}\mathbf{z}_k^\delta\|), \quad k = 1, 2, 3, ... \tag{3.12}$$

Usually a log-log scale is used for the Standard L-Curve, i.e. instead of (η_k, ρ_k) the points $(\log(\eta_k), \log(\rho_k))$ are considered. In the case of the Residual L-Curve, different choices are possible, cf. [76] and [77]: we shall plot the Residual L-Curve in a semi-logarithmic scale, i.e. by connecting the points $(k, \log(\rho_k))$, $k = 1, 2, 3, ...$

In contrast to the Discrepancy Principle and the Residual L-Curve, the Standard L-Curve explicitly takes into account the growth of the norm of the computed approximate solutions (cf. e.g. [36] and [87]) as k increases. In his illuminating description of the properties of the Standard L-Curve for Tikhonov regularization and other methods in [36], P. C. Hansen suggests to define the stopping index as the integer k corresponding to the corner of the L-curve, characterized by the point of maximal curvature (the so called *L-Curve Criterion*).

Castellanos et al. [9] proposed a scheme for determining the corner of a discrete L-curve by forming a sequence of triangles with vertices at the points of the curve and then determining the desired vertex of the L from the shape of these triangles. For obvious reasons, this algorithm is known in literature as the *Triangle method*.

Hansen et al. [37] proposed an alternative approach for determining the vertex of the L: they constructed a sequence of pruned L-curves, removing an increasing number of points, and considered a list of candidate vertices produced by two different selection algorithms. The vertex of the L is selected from this list by taking the last point before reaching the part of the L-curve, where the norm of the computed approximate solution starts to increase rapidly and the norm of the associated residual vectors stagnates. This is usually called the *Pruning method* or the *L-Corner method*.

The Standard L-Curve has been applied successfully to the solution of many linear discrete ill-posed problems and is a very popular method for choos-

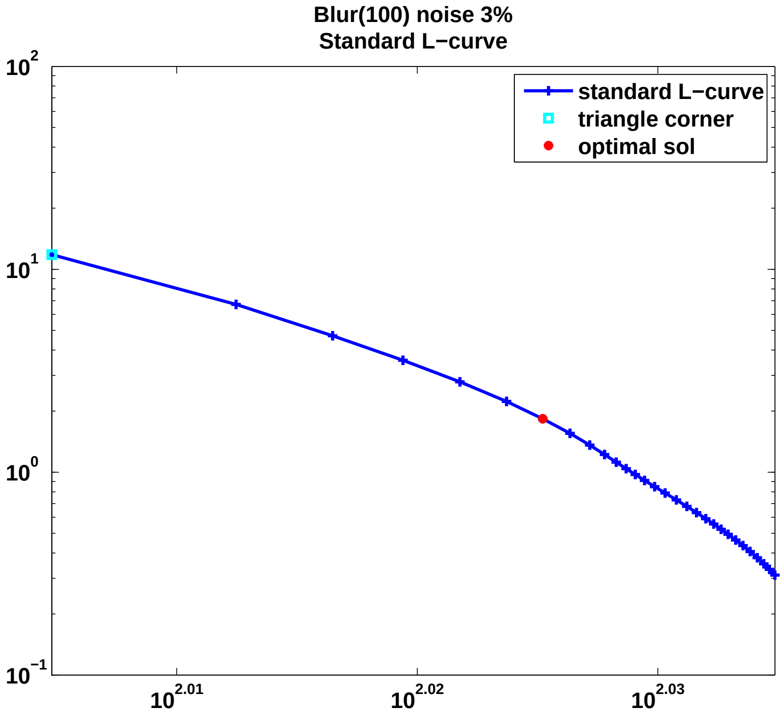


Figure 3.2: L-curves for blur(100), $\varrho = 3\%$. The L-curve is simply not L-shaped.

ing the regularization parameter, also thanks to its simplicity. However, it has some well known drawbacks, as shown by Hanke [28] and Vogel [95]. A practical difficulty, as pointed out by Reichel et al. in [76] and in the recent paper [77], is that the discrete points of the Standard L-Curve may be irregularly spaced, the distance between pairs of adjacent points may be small for some values of k and it can be difficult to define the vertex in a meaningful way. Moreover, sometimes the L-curve may not be sufficiently pronounced to define a reasonable vertex (cf., e.g., Figure 3.2).

In their paper [76], Reichel and Sadok defined the Residual L-Curve for the TSVD in a Hilbert space setting seeking to circumvent the difficulties caused by the cluster of points near the corner of the Standard L-Curve and showing that it often achieved the better numerical results. Among all heuristic methods considered in the numerical tests of [77], the Residual L-Curve proved to be one of the best heuristic stopping rules for the TSVD, but it also obtained the worst results in the case of CGNE, providing oversmoothed solutions.

Two reasons for this oversmoothing effect are the following:

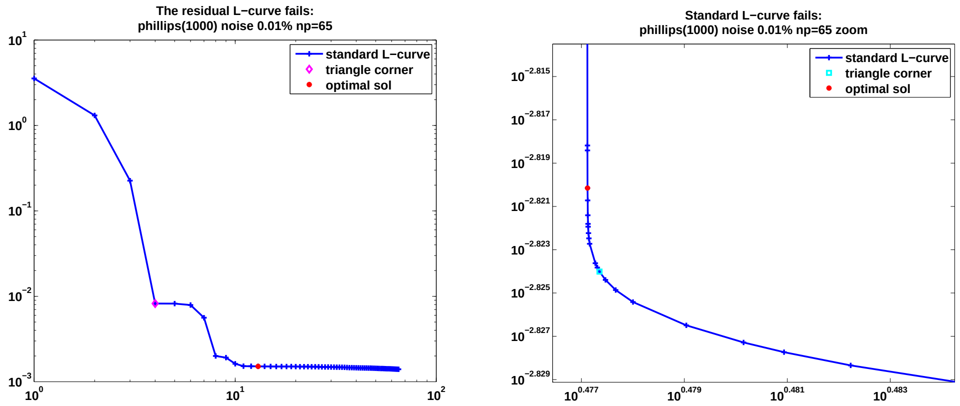
- the Residual L-Curve in the case of CGNE sometimes presents some kinks before getting flat, thus the corner may be found at an early stage;
- the residual norm of the solution is often too to large at the corner of the Residual L-Curve: it is preferable to stop the iteration as soon as the term $\|\mathbf{x}^{\dagger}\||\mathsf{p}_{\mathsf{k}}'(0)|^{-1/2}$ is neglibible in the residual norm estimate (3.11), i.e., when the curve begins to be flat.

In Figure 3.3 we show the results of the test problem phillips(1000), with noise level $\varrho = 0.01\%$. In this example, both L-curve methods fail: as expected by Hanke in [28], the Standard L-curve stops the iteration too late, giving an undersmoothed solution; on the other hand, due to a very marked step at an early stage, the Residual L-Curve provides a very oversmoothed solution.

We propose to approximate the Residual L-Curve by a smoother curve. More precisely, let n_{pt} be the total number of iterations performed by CGNE. For obvious reasons, to obtain a reasonable plot of the L-curves, we must perform enough iterations, i.e. $n_{pt} > k^{\sharp}$. We approximate the data points $\{(k, \log(\rho_k))\}_{k=1,\dots,n_{pt}}$ with cubic B-splines using the routine data_approx defined in the Appendix, obtaining a new (smoother) set of data points $\{(k, \log(\tilde{\rho}_k))\}_{k=1,\dots,n_{pt}}$.

We call the curve obtained by connecting the points $(k, \log(\tilde{\rho}_k))$ with straight lines the *Approximated Residual L-Curve* and we denote by k_L, k_{rL} and k_{arL} the indices determined by the triangle method for the Standard L-Curve, the Residual L-Curve and the Approximated Residual L-Curve respectively.

In Figure 3.4 we can see 2 approximate residual L-curves. Typically, the approximation has the following properties:



(a) Residual L-curve

(b) Standard L-curve

Figure 3.3: Residual L-Curve and Standard L-Curve for the test problem phillips(1000), $\varrho = 0.01\%$.

(i) it tends to remove or at least smooth the steps of the Residual L-Curve when they are present (cf. the picture on the left in Figure 3.4);

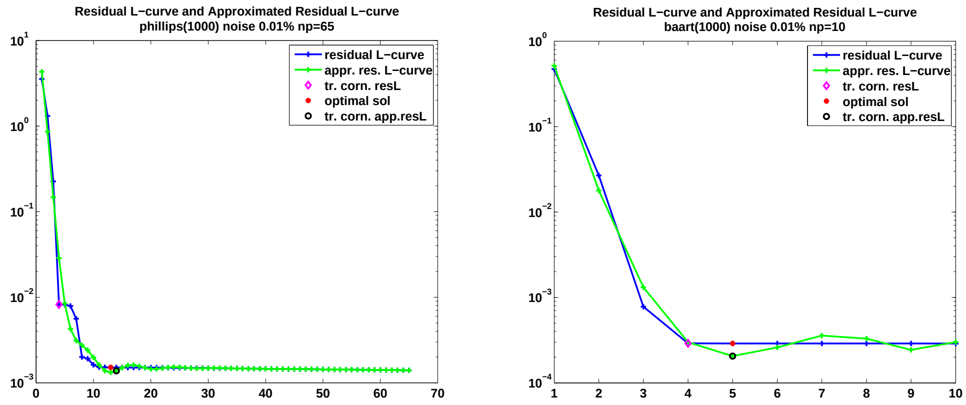
(ii) when the Residual L-Curve has a very marked L-shape it tends to have a minimum in correspondence to the plateau of the Residual L-Curve (cf. the picture on the right in Figure 3.4);

(iii) when the Residual L-Curve is smooth the shape of both curves is similar.

As a consequence, very often we have $k_{rL} < k_{arL}$ and k_{arL} corresponds to the plateau of the Residual L-Curve. This should indeed improve the performances, because it allows to push the iteration a little bit further, overcoming the oversmoothing effects described above.

We are ready to define the first of the three stopping rules (SR) for CGNE.

Definition 3.2.1 (**Approximated Residual L-Curve Criterion**). *Consider the sequence of points* $(k, \log(\rho_k))$ *obtained by performing* n_{pt} *steps of* CGNE *and let* $(k, \log(\tilde{\rho}_k))$ *be the sequence obtained by approximating* $(k, \log(\rho_k))$ *by means of the routine* data_approx. *Compute the corners* k_{rL}



(a) Phillips

(b) Baart

Figure 3.4: Residual L-Curve and Approximated Residual L-Curve for 2 different test problems. Fixed values: $N = 1000$, $\varrho = 0.01\%$, $\mathsf{seed} = 1$.

and k_{arL} *using the triangle method and let* k_0 *be the first index such that* $\tilde{\rho}_{k_0} = \min\{\tilde{\rho}_k \mid k = 1, ..., n_{pt}\}$*. Then, as a stopping index for the iterations of* CGNE*, choose*

$$k_{SR1} := \max\{k_{rL}, \check{k}\}, \qquad \check{k} = \min\{k_{arL}, k_0\}. \tag{3.13}$$

This somewhat articulated definition of the stopping index avoids possible errors caused by an undesired approximation of the Residual L-Curve or by an undesired result in the computation of the corner of the Approximated Residual L-Curve. Below, we will analyze and compare the stopping rules defined by k_L, k_{rL} and k_{SR1}.

3.3 SR1: numerical experiments

This section is dedicated to show the performances of the stopping rule SR1. In all examples below, in order to avoid some problems caused by rounding errors, the function lsqr_b from [35] has been used with parameter reorth $= 1$. For more details on this function and rounding errors in the CGNE algorithm, see Appendix D.

3.3.1 Test 1

In order to test the stopping rule of Definition 3.2.1, we consider 10 different test problems from P.C. Hansen's Regularization Tools [35]. For each test problem, we fix the number n_{pt} in such a way that both the standard and the residual L-curves can be visualized appropriately and take 2 different values for the dimension and the noise level. In each of the possible cases, we run the algorithm with 25 different Matlab seeds, for a total of 1000 different examples.

In Table 3.1, for each test problem and for each couple (N_i, ϱ_j), i,$j \in \{1, 2\}$, we show the average relative errors obtained respectively by the stopping indices k_L (Standard L-Curve), k_{rL} (Residual L-Curve) and k_{SR1} for all possible seeds. In round brackets we collect the number of failures, i.e. how many times the relative error obtained by the stopping rule is at least 5 times larger than the relative error obtained by the optimal stopping index $k^{\sharp}$. We can see that stopping rule associated to the index k_{SR1} improves the results of the Residual L-Curve in almost all the cases.

This stopping rule proves to be reliable also when the noise level is smaller and the Standard L-Curve fails, as we will see below.

3.3.2 Test 2

In this example we test the robustness of the method when the noise level is small and with respect to the number of points n_{pt}. As we have seen, this is a typical case in which the Standard L-Curve method may fail to obtain acceptable results.

We consider the test problems gravity, heat and phillips, with $N = 1000$, $\varrho \in \{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$, seed $\in \{1, 2, ..., 25\}$. For each case, we also take 3 different values of n_{pt} (the smallest one is only a little bit larger than the optimal index $k^{\sharp}$), in order to analyze the dependence of the methods on this particular parameter.

The results of Table 3.2 clearly show that the Approximated Residual L-

Test 1: results			
		Average Rel. Err. (no. of failures) for k_L, k_{rL}, k_{SR1}	
Problem[n_{pt}]	N	ϱ_1	ϱ_2
Baart[8]	N_1	0.1216(0),0.1660 (0),0.1216(0)	0.1688(0),0.2024(0),0.1676(0)
	N_2	0.1156(0),0.1656(0),0.1156(0)	0.1680(0),0.1872(0),0.1660(0)
Deriv2[30]	N_1	0.2024(0),0.1920(0),0.1872(0)	0.3024(0),0.2732(0),0.2632(0)
	N_2	0.1488(0),0.1924(0),0.1924(0)	0.2184(0),0.2772(0),0.2268(0)
Foxgood[6]	N_1	0.0104 (0),0.0308(2),0.0104 (0)	0.0704(0),0.0324(0),0.1192(2)
	N_2	0.0076 (0),0.0308 (3),0.0076(0)	0.0300(0),0.0308(0),0.0400(0)
Gravity[20]	N_1	0.0732(11),0.0232(0),0.0104(0)	0.1080(4),0.0596(0),0.0388(0)
	N_2	0.0176 (0),0.0220(0),0.0156(0)	0.0368(0),0.0580(0),0.0320(0)
Heat[50]	N_1	0.2160 (0),0.0440(0),0.0436(0)	0.3296(0),0.1232(0),0.1300(0)
	N_2	0.0680 (0),0.0404(0),0.0404(0)	0.0748(0),0.1124(0),0.1040(0)
I-Laplace[20]	N_1	0.1156 (0),0.1224(0),0.1028(0)	0.1964(0),0.1600(0),0.1584(0)
	N_2	0.1904 (0),0.2164(0),0.2020(0)	0.2128(0),0.2488(0),0.2488(0)
Phillips[30]	N_1	0.1036 (23),0.0240(0),0.0236(0)	0.0908(2),0.0276(0),0.0272(0)
	N_2	0.0204 (1),0.0240(0),0.0240(0)	0.0328(0),0.0248(0),0.0244(0)
Shaw[15]	N_1	0.1008(1),0.0596(0),0.0492(0)	0.1400(0),0.1680(0),0.1284(0)
	N_2	0.0536(0),0.0592(0),0.0476(0)	0.0636(0),0.1676(0),0.0672(0)
Blur(50,3,1)[200]	N_1	0.3280(0),0.2324(0),0.2308(0)	0.3540(0),0.3536(0),0.3536(0)
	N_2	0.2556 (0),0.1980(0), 0.1976(0)	0.3040(0),0.1776(0),0.1768(0)
Tomo[200]	N_1	0.6292 (0),0.3732(0),0.2768(0)	0.8228(0),0.3780(0),0.3808(0)
	N_2	0.6892 (0),0.3732(0),0.3748(0)	0.6424(0),0.1776(0),0.1768(0)

Table 3.1: General test for the L-curves: numerical results. In the 1D test problems $N_1 = 100$, $N_2 = 1000$, $\varrho_1 = 0.1\%$, $\varrho_2 = 1\%$; in the 2D test problems $N_1 = 900$, $N_2 = 2500$, $\varrho_1 = 1\%$, $\varrho_2 = 5\%$.

Curve method is by far the best in this case, not only because it gains the better results in terms of the relative error (cf. the sums of the relative errors for all possible $\mathsf{seed} = 1, ..., 25$), but also because it is more stable with respect to the parameter n_{pt}.

Concerning the number of failures of this example, the Standard L-Curve fails in the 66% of the cases, the Residual L-Curve in the 24.7% and the Approximated Residual L-Curve only in the 1% of the cases. We also note that for the Residual L-Curve and the Approximated Residual L-Curve methods the results tend to improve for large values of n_{pt}.

Test 2: results			
	n_{pt}: $\sum$ Rel. Err. (no. of failures) for k_L, k_{rL}, k_{SR1}		
	Gravity	Heat	Phillips
ϱ_1	20: 0.44(19),0.13(0),0.09(0) 30: 0.49(22),0.13(0),0.06(0) 40: 0.85(25),0.13(0),0.06(0)	80: 3.39(25),0.31(0),0.30(0) 120: 1.50(25),0.31(0),0.30(0) 160: 1.73(25),0.31(0),0.30(0)	45: 1.91(25),0.49(22),0.04(0) 60: 0.87(25),0.40(18),0.04(0) 45: 0.76(25),0.40(18),0.04(0)
ϱ_2	18: 0.36(10),0.20(0),0.13(0) 24: 0.33(7),0.20(0),0.10(0) 30: 0.54(17),0.19(0),0.13(0)	60: 3.47(25),0.39(0),0.36(0) 90: 1.47(19),0.39(0),0.36(0) 160: 1.61(22),0.36(0),0.36(0)	35: 1.25(25),0.60(25),0.05(0) 45: 0.59(25),0.60(25),0.05(0) 55: 0.63(25),0.60(25),0.05(0)
ϱ_3	15: 0.46(6),0.27(0),0.19(0) 20: 0.18(0),0.27(0),0.19(0) 25: 0.43(4),0.27(0),0.19(0)	50: 2.33(25),0.42(0),0.39(0) 70: 1.94(23),0.40(0),0.39(0) 90: 1.46(4),0.39(0),0.39(0)	25: 1.23(25),0.60(25),0.07(0) 32: 0.95(25),0.60(25),0.07(0) 55: 0.51(23),0.60(25),0.07(0)
ϱ_4	15: 0.32(0),0.55(0),0.28(0) 20: 0.35(0),0.55(0),0.28(0) 25: 0.60(4),0.55(0),0.28(0)	40: 3.72(25),0.88(0),0.75(0) 60: 1.55(0),0.88(0),0.46(0) 80: 1.61(0),0.88(0),0.45(0)	20: 1.30(25),0.61(5),0.60(5) 27: 0.49(2),0.61(5),0.53(2) 35: 0.56(8),0.61(5),0.53(2)

Table 3.2: Second test for the approximated Residual L-Curve Criterion: numerical results with small values of δ.

3.4 SR2: Projected Data Norm Criterion

The diagonal matrix example and the observation of Section 3.1 suggest to replace the classic threshold of the Discrepancy Principle $\|\mathbf{e}\|$ with the norm of the projection of the noise onto the high frequency part of the spectrum. However, in practice a direct computation of this quantity is impossible, because the noise is unknown (only information about its norm and its stochastic distribution is usually available) and because the Ritz values are too expensive to be calculated during the iteration.

To overcome these difficulties, we propose the following strategy, based on the singular value decomposition of the matrix $\mathbf{A}$.

Let $\mathbf{A} \in \mathbb{M}_{m,N}(\mathbb{R})$, $m \geq N$, $\mathrm{rank}(\mathbf{A}) = N$, let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$ be a SVD of $\mathbf{A}$ and suppose that the singular values of $\mathbf{A}$ may be divided into a set of large singular values $\lambda_p \leq ... \leq \lambda_1$ and a set of $N - p$ small singular values $\lambda_N \leq ... \leq \lambda_{p+1}$, with $\lambda_{p+1} < \lambda_p$. If the exact data $\mathbf{b}$ satisfy the Discrete Picard condition, then the SVD coefficients $|\mathbf{u}_i^*\mathbf{b}|$ are very small for $i > p$.

Therefore, if

$$\mathbf{x}_p^{\mathrm{TSVD}} := \sum_{j=1}^{p} \frac{\mathbf{u}_j^* \mathbf{b}^\delta}{\lambda_j} \mathbf{v}_j = \mathbf{V} \mathbf{\Lambda}_p^\dagger \mathbf{U}^* \mathbf{b}^\delta, \tag{3.14}$$

with $\mathbf{\Lambda}_p^\dagger$ being the pseudo inverse matrix of

$$\mathbf{\Lambda}_p = \begin{pmatrix} \lambda_1 & & & & & \\ & ... & & & & \\ & & \lambda_p & & & \\ & & & 0 & & \\ & & & & ... & \\ & & & & & 0 \\ \mathbf{0} & & & ... & & \mathbf{0} \end{pmatrix} \in \mathbb{M}_{m,N}(\mathbb{R}), \tag{3.15}$$

then

$$\begin{aligned} \|\mathbf{b}^\delta - \mathbf{A}\mathbf{x}_p^{\mathrm{TSVD}}\| &= \|\mathbf{U}^*\mathbf{b}^\delta - \mathbf{\Lambda}\mathbf{V}^*\mathbf{x}_p^{\mathrm{TSVD}}\| = \|\mathbf{U}^*\mathbf{b}^\delta - \mathbf{U}_p^*\mathbf{b}^\delta\| \\ &= \|\mathbf{U}_{m-p}^*\mathbf{b}^\delta\| \sim \|\mathbf{U}_{m-p}^*\mathbf{e}\|, \end{aligned} \tag{3.16}$$

where $\mathbf{U}_p$, $\mathbf{U}_{m-p} \in \mathbb{M}_m(\mathbb{R})$, depending on the column vectors $\mathbf{u}_i$ of the matrix $\mathbf{U}$, are defined by $(\mathbf{u}_1, .., \mathbf{u}_p, \mathbf{0}, .., \mathbf{0})$ and $(\mathbf{0}, .., \mathbf{0}, \mathbf{u}_{p+1}, .., \mathbf{u}_m)$ respectively. The right-hand side is exactly the projection of the noise onto the high frequency part of the spectrum, so we can interpret (3.16) as a relation between the residual norm of the truncated singular value decomposition and this quantity.

The equation (3.16) and the considerations of Section 3.1 suggest to calculate the regularized solution $\mathbf{x}_p^{\mathrm{TSVD}}$ of the perturbed problem $\mathbf{A}\mathbf{x} = \mathbf{b}^\delta$ using the truncated singular value decomposition, by stopping the iteration of CGNE as soon as the residual norm becomes smaller than $\|\mathbf{b}^\delta - \mathbf{A}\mathbf{x}_p^{\mathrm{TSVD}}\| = \|\mathbf{U}_{m-p}^*\mathbf{b}^\delta\|$.

The following numerical simulation on 8 problems of P.C. Hansen's Regularization Tools confirms the statement above. We fix the dimension $N = 1000$, $\varrho = 0.1\%$ and the constant of the Discrepancy Principle $\tau = 1.001$, run lsqr_b with reorthogonalization for each problem with 25 different Matlab seeds and compare the Discrepancy Principle solutions with those obtained by arresting the iteration of CGNE at the first index such that the residual norm is lower

Residual thresholds for stopping CGNE				
Problem	Avg. rel. err. CGNE			
	$\tau\|\mathbf{U}^*_{m-p_\sharp}\mathbf{b}^\delta\|$	$\tau\delta$	Opt. err.	$\tau\|\mathbf{U}^*_{m-p_\sharp}\mathbf{e}\|$
Baart	0.1158	0.1158	0.1041	0.1158
Deriv2	0.1456	0.1517	0.1442	0.1459
Foxgood	0.0079	0.0079	0.0076	0.0079
Gravity	0.0129	0.0144	0.0111	0.0124
Heat	0.0228	0.0281	0.0225	0.0228
I_Laplace	0.1916	0.1952	0.1870	0.1898
Phillips	0.0078	0.0087	0.0075	0.0086
Shaw	0.0476	0.0476	0.0440	0.0480

Table 3.3: Comparison between different residual norm thresholds for stopping CGNE, with p as in 3.17.

or equal to $\tau\|\mathbf{U}^*_{m-p_\sharp}\mathbf{b}^\delta\|$, with $p_\sharp$ minimizing the error of the truncated singular value decomposition:

$$\|\mathbf{x}^{\mathrm{TSVD}}_{p_\sharp} - \mathbf{x}^\dagger\| = \min_j \|\mathbf{x}^{\mathrm{TSVD}}_j - \mathbf{x}^\dagger\|. \tag{3.17}$$

The results, summarized in Table 3.3, show that this gives an extremely precise solution in a very large number of cases. Moreover, the corresponding stopping index is equal to $k^\sharp$ (the optimal stopping index of CGNE) in the 53% of the considered examples.

In the table, we also consider the results obtained by arresting the iteration when the residual norm is lower or equal to $\tau\|\mathbf{U}^*_{m-p_\sharp}\mathbf{e}\|$. As a matter of fact, the residual norm corresponding to the optimal stopping index is very well approximated by this quantity in the large majority of the considered examples.

Performing the same simulation with the same parameters except for $\varrho = 1\%$ leads to similar results: the method based on the optimal solution of the TSVD obtains the better performance in 86 cases on 200 and the worse performance only 24 times and in a very large number of examples (45%) its stopping index is equal to $k^\sharp$.

These considerations justify the following heuristic stopping rule for CGNE.

Definition 3.4.1 (**Projected Data Norm Criterion**). *Let $\mathbf{z}_k^\delta$ be the iterates of* CGNE *for the perturbed problem*

$$\mathbf{A}\mathbf{x} = \mathbf{b}^\delta. \tag{3.18}$$

Let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$ be a SVD *of $\mathbf{A}$. Let p be a regularization index for the* TSVD *relative to the data $\mathbf{b}^\delta$ and to the matrix $\mathbf{A}$ and fix $\tau > 1$. Then stop* CGNE *at the index*

$$k_{\mathrm{SR2}} := \min\{k \in \mathbb{N} \mid \|\mathbf{b}^\delta - \mathbf{A}\mathbf{z}_k^\delta\| \leq \tau \|\mathbf{U}_{m-p}^* \mathbf{b}^\delta\|\}. \tag{3.19}$$

3.4.1 Computation of the index p of the SR2

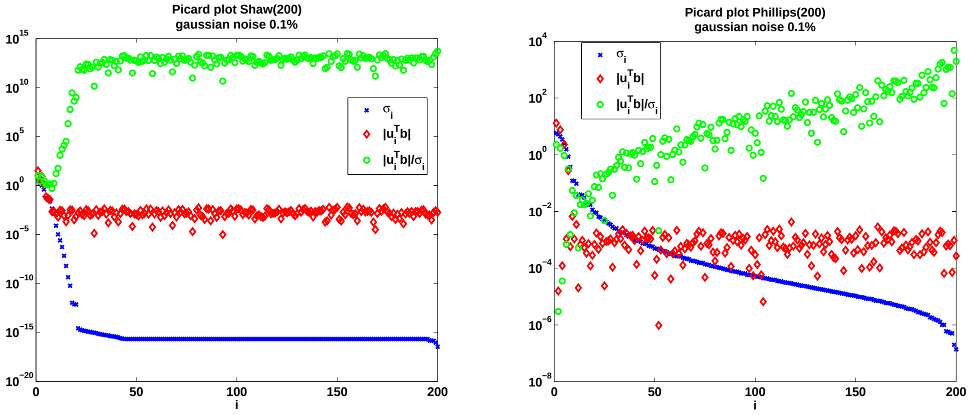
Obviously, in practice the optimal index of the truncated singular value decomposition is not available, since the exact solution is unknown. However, for discrete ill-posed problems it is well known that a good index can be chosen by analyzing the plot of the SVD coefficients $|\mathbf{u_i}^*\mathbf{b}^\delta|$ of the perturbed problem $\mathbf{A}\mathbf{x} = \mathbf{b}^\delta$. The behavior of the SVD coefficients in the case of white Gaussian noise given by $\mathbf{e} \sim N(\mathbf{0}, \nu^2\mathbf{I}_m)$, $\nu > 0$, is analyzed by Hansen in [36]: as long as the unperturbed data $\mathbf{b}$ satisfy the discrete Picard condition, the coefficients $|\mathbf{u}_i^*\mathbf{b}|$ decay on the average to 0 at least as fast as the singular values λ_i. On the other hand, the coefficients $|\mathbf{u}_i^*\mathbf{b}^\delta|$ decay, on the average, only for the first small values of i, because for large i the noisy components $\mathbf{u}_i^*\mathbf{e}$ dominate, thus after a certain critical index $i_{\mathbf{b}^\delta}$ they begin level off at the level

$$\mathbb{E}(\mathbf{u}_i^*\mathbf{e}) = \nu. \tag{3.20}$$

To compute a good index p for the truncated singular value decomposition, one must also consider machine errors if the last singular values are very small: as pointed out in [36], pg. $70 - 71$, *the number of terms that can be safely included in the solution is such that:*

$$p \leq \min\{i_{\mathbf{A}}, i_{\mathbf{b}^\delta}\}, \tag{3.21}$$

where $i_{\mathbf{A}}$ is the index at which λ_i begin to level off and $i_{\mathbf{b}^\delta}$ is the index at which $|\mathbf{u}_i^*\mathbf{b}^\delta|$ begin to level off. The value $i_{\mathbf{A}}$ is proportional to the error



(a) Shaw(200) (b) Phillips(200)

Figure 3.5: Plot of the singular values λ_i (blue cross), of the SVD coefficients $|\mathbf{u}_i^* \mathbf{b}^\delta|$ (red diamond) and of the ratios $|\mathbf{u}_i^* \mathbf{b}^\delta|/\lambda_i$ (green circle) of 2 different test problems with perturbed data: $\varrho = 0.1\%$, $\mathsf{seed} = 0$.

present in the matrix $\mathbf{A}$ (*i.e. model error*) *while the value* $i_{\mathbf{b}^\delta}$ *is proportional to the errors present in the data* $\mathbf{b}^\delta$ (*i.e. noise error*). Although very often a visual inspection of the plot of the coefficients in a semi-logarithmic scale is enough to choose the index p, defining an algorithm to compute the index p automatically is not easy, because the decay of the SVD coefficients may not be monotonic and indeed is often affected by outliers (cf. Figure 3.5). A rule based on the moving geometric mean has been proposed in [32] and implemented in the file picard.m of [35]. Here we suggest to use the *Modified Min_Max Rule*, defined in Section C.4 of the Appendix.

3.5 SR2: numerical experiments

We test the stopping rule of Definition 3.4.1 in 4000 different examples. For each of 8 test problems we choose 2 values of N, 10 values of ϱ and 25 values of seed. We compare the results obtained by the stopping index k_{SR2} with those obtained by the Discrepancy Principle. In Table 3.4 we summarize the main results of the test.

Numerical test for SR2						
Problem	Dim	Avg. rel. err. CGNE			Failures	k_{SR2} vs. k_D ($k_{\mathrm{SR2}} = k^\sharp$)
		k_{SR2}	k_D	$k^\sharp$	$\varepsilon_{\mathrm{SR2}} >5(10)\varepsilon^\sharp$	
Baart	100	0.1742	0.1750	0.1469	0(0)	20,201,29 (135)
Baart	1000	0.1612	0.1586	0.1357	0(0)	8,234,8 (89)
Deriv2	100	0.3533	0.2352	0.2242	7(1)	100,11,139 (60)
Deriv2	1000	0.2596	0.1992	0.1874	9(0)	146,32,72 (57)
Foxgood	100	0.0390	0.0312	0.0232	7(1)	41,146,63 (130)
Foxgood	1000	0.0267	0.0270	0.0166	4(0)	14,226,10 (96)
Gravity	100	0.0332	0.0349	0.0276	0(0)	146,24,80 (115)
Gravity	1000	0.0236	0.0256	0.0192	0(0)	81,163,6 (67)
Heat	100	0.1065	0.0921	0.0803	0(0)	148,4,98 (62)
Heat	1000	0.0530	0.0598	0.0476	0(0)	215,14,21 (79)
I-laplace	100	0.1360	0.1370	0.1242	0(0)	91,116,43 (114)
I-laplace	1000	0.2140	0.2134	0.2004	0(0)	84,122,44 (13)
Phillips	100	0.0311	0.0244	0.0215	0(0)	97,15,138 (74)
Phillips	1000	0.0183	0.0200	0.0156	0(0)	138,102,10 (74)
Shaw	100	0.0890	0.0973	0.0760	0(0)	96,112,42 (82)
Shaw	1000	0.0724	0.0635	0.0520	0(0)	22,168,60 (53)

Table 3.4: Comparison between the numerical results obtained by k_{SR2}, k_D and the optimal index $k^\sharp$.

The constant τ, chosen equal to 1.001 when $N = 100$ and equal to 1.005 when $N = 1000$, is always the same for k_{SR2} and k_D. The columns 3, 4 and 5 of the table collect the average relative errors for all values of $\varrho = 10^{-3}, 2 \times 10^{-3}, ..., 10^{-2}$ and $\mathtt{seed} = 1, ..., 25$ obtained by k_{SR2}, k_D and $k^\sharp$ respectively. The column 6 contains the number of times the relative error corresponding to k_{SR2}, denoted by $\varepsilon_{\mathrm{SR2}}$, is larger than 5(10) times the optimal error $\varepsilon^\sharp$. The numbers in the last column count how many times $\varepsilon_{\mathrm{SR2}} > \varepsilon_D$, how many times $\varepsilon_{\mathrm{SR2}} = \varepsilon_D$ and how many times $\varepsilon_{\mathrm{SR2}} < \varepsilon_D$ respectively. Finally, the fourth number in round brackets counts how many times $\varepsilon_{\mathrm{SR2}} = \varepsilon^\sharp$.

The results clearly show that the stopping rule is very reliable for discrete ill-posed problems of medium size. It is remarkable that in the 4000 examples considered it failed (that is, $\varepsilon_{\mathrm{SR2}} > 10\varepsilon^\sharp$) only twice (cf., e.g., the results obtained by the heuristic stopping rules in [77], Table 2). Moreover, in many cases it even improves the results of the Discrepancy Principle, which is based on the knowledge of the noise level.

3.6 Image deblurring

One of the most famous applications of the theory of ill-posed problems is to recover a sharp image from its blurry observation, i.e. *image deblurring*. It frequently arises in imaging sciences and technologies, including optical, medical, and astronomical applications and is crucial for allowing to detect important features and patterns such as those of a distant planet or some microscopic tissue.

Due to its importance, this subject has been widely studied in literature: without any claim to be exhaustive, we point out at some books [10], [38], [48], [100], or chapters of books [4], [96] dedicated to this problem.

In most applications, blurs are introduced by three different types of physical factors: optical, mechanical, or medium-induced, which could lead to familiar out-of-focus blurs, motion blurs, or atmospheric blurs respectively. We refer the reader to [10] for a more detailed account on the associated physical processes.

Mathematically, a continuous (analog) image is described by a nonnegative function $f = f(\mathbf{x})$ on $\mathbb{R}^2$ supported on a (rectangular) 2D domain Ω and the blurring process is either a linear or nonlinear operator K acting on the some functional space. Since we shall focus only on linear deblurring problems, K is assumed to be linear.

Among all linear blurs, the most frequently encountered type is the shift invariant blur, i.e. a linear blur $K = K[f]$ such that for any shift vector $\mathbf{y} \in \mathbb{R}^2$,

$$g(\mathbf{x}) = K\left[f(\mathbf{x})\right] \quad \Longrightarrow \quad g(\mathbf{x}-\mathbf{y}) = K\left[f(\mathbf{x}-\mathbf{y})\right], \mathbf{x} \in \Omega. \tag{3.22}$$

It is well known in signal processing as well as system theory [72] that a shift-invariant linear operator must be in the form of convolution:

$$g(\mathbf{x}) = K[f](\mathbf{x}) = \varkappa * f(\mathbf{x}) = \int_{\Omega} \varkappa(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y}, \tag{3.23}$$

for some suitable kernel function $\varkappa(\mathbf{x})$, or the point spread function (PSF). The function $g(\mathbf{x})$ is the blurred analog image that is converted into a digital image through a digitalization process (or sampling).

A digital image is typically recorded by means of a CCD (charge-coupled device), which is an array of tiny detectors (potential wells), arranged in a rectangular grid, able to record the amount, or intensity, of the light that hits each detector.

Thus, a digital grayscale image

$$\mathbf{G} = (g_{j,l}), \quad j = 1, ..., J, \quad l = 1, ..., L \tag{3.24}$$

is a rectangular array, whose entries represent the (nonnegative) light intensities captured by each detector.

The PSF is described by a matrix $\mathbf{H} = (h_{j,l})$ of the same size of the image, whose entries are all zero except for a very small set of pixels (j, l) distributed around a certain pixel (j_c, l_c) which is the center of the blur. Since we are assuming spatial invariant PSFs, the center of the PSF corresponds to the center of the 2D array.

In some cases the PSF can be described analytically and $\mathbf{H}$ can be constructed from a function, rather than through experimentation (e.g. the horizontal and vertical motion blurs are constructed in this way).

In other cases, the knowledge of the physical process that causes the blur provides an explicit formulation of the PSF. In this case, the elements of the PSF array are given by a precise mathematical expression: e.g. the out-of-focus blur is given by the formula

$$h_{j,l} = \begin{cases} \frac{1}{\pi r^2} & \text{if } (j - j_c)^2 + (l - l_c)^2 \leq r^2, \\ 0 & \text{otherwise,} \end{cases} \tag{3.25}$$

where $r > 0$ is the radius of the blur.

For other examples, such as the blur caused by atmospheric turbulence or the PSF associated to an astronomical telescope, we refer to [38] and the references therein.

As a consequence of the digitalization process, the continuous model described by (3.23) has to be adapted to the discrete setting as well. To do this, we consider first the 1D case

$$g(t) = \int \varkappa(t-s) f(s) ds. \tag{3.26}$$

To fix the ideas, we assume that J is even and that the function $f(s)$ is defined in the interval $[-\frac{J-1}{2}, \frac{J-1}{2}]$. Let

$$s_j = -\frac{J-1}{2} + j - 1, \quad j = 1, ..., J \tag{3.27}$$

be the J points in which the interval is subdivided and discretize $\varkappa$ and f in such a way that $\varkappa(s) = \varkappa(s_j) = h_j$ if $|s - s_j| < \frac{1}{2}$ or $s = s_j + \frac{1}{2}$ and analogously for $\varkappa$. Approximating (3.26) with the trapezoidal rule

$$g(t) \cong \sum_{j'=1}^{J} \varkappa(t - s_{j'}) f(s_{j'}) \tag{3.28}$$

and recomputing in the points s_j, we obtain the components of the discretized version $\mathbf{g}$ of the function g:

$$g_j = \sum_{j'=1}^{J} \varkappa(s_j - s_{j'}) f(s_{j'}), \quad j = 1, ...J. \tag{3.29}$$

As a consequence of the assumptions we have made, (3.29) can be rewritten into

$$g_j = \sum_{j'=1}^{J} h_{j-j'+\frac{J}{2}} f_{j'}, \quad j = 1, ...J, \tag{3.30}$$

which is the componentwise expression of the discrete convolution between the column vectors $\mathbf{h} = (h_j)$ and $\mathbf{f} = (f_j)$. We observe that some terms in the sum in the right-hand side of (3.30) may be not defined: this happens because the support of the convolution between $\varkappa$ and f is larger than the supports of $\varkappa$ and f. The problem is solved by extending the vector $\mathbf{h}$ to the larger vector

$$\tilde{\mathbf{h}} = \begin{pmatrix} h_{-\frac{J}{2}+1} \\ \dots \\ h_0 \\ \mathbf{h} \\ h_{J+1} \\ \dots \\ h_{J+\frac{J}{2}} \end{pmatrix}, \quad h_j = h_{j+J}, \quad j = -\frac{J}{2}+1, ..., \frac{J}{2} \tag{3.31}$$

and substituting $\mathbf{h}$ with $\tilde{\mathbf{h}}$ in (3.30), which is equivalent to extend $\varkappa$ periodically on the real line. The convolution (3.30) may also be expressed in the form

$$\mathbf{g} = \mathbf{A}\mathbf{f}, \quad \mathbf{A} = (a_{i,j}), \quad a_{i,j} = h_{i-j+\frac{J}{2}}, \quad i, j = 1, ..., J. \tag{3.32}$$

In the 2D case, proceeding in an analogous way, we get

$$g_{j,l} = \sum_{j'=1}^{J} \sum_{l'=1}^{L} h_{j-j'+\frac{J}{2}, l-l'+\frac{L}{2}} f_{j',l'}. \tag{3.33}$$

The equation above (3.33) is usually written in the form

$$\mathbf{G} = \mathbf{H} * \mathbf{F}, \tag{3.34}$$

where $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{F}$ are here matrices in $\mathbb{M}_{J,L}(\mathbb{R})$. If $\mathbf{g}$ and $\mathbf{f}$ are the column vectors obtained by concatenating the columns of $\mathbf{G}$ and $\mathbf{F}$ respectively, then (3.34) can be rewritten in the form

$$\mathbf{g} = \mathbf{A}\mathbf{f}. \tag{3.35}$$

In the case of an image f with 1024 × 1024 pixels, the system (3.35) has then more than one million unknowns. For generic problems of this size, the computation of the singular value decomposition is not usually possible.

However, if we set $N := JL$, then $\mathbf{A} \in \mathbb{M}_N(\mathbb{R})$ is a circulant matrix with circulant blocks (BCCB). It is well known (cf. e.g. [4] or [38]) that BCCB matrices are normal (that is, $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$) and that may be diagonalized by

$$\mathbf{A} = \mathbf{\Phi}^*\mathbf{\Lambda}\mathbf{\Phi}, \tag{3.36}$$

where $\mathbf{\Phi}$ is the two-dimensional unitary matrix unitary Discrete Fourier Transform (DFT) matrix.

We recall that the DFT of a vector $\mathbf{z} \in \mathbb{C}^N$ is the vector $\hat{\mathbf{z}}$ whose components are defined by

$$\hat{\mathrm{z}}_j = \frac{1}{\sqrt{N}} \sum_{j'=1}^{N} \mathrm{z}_{j'} e^{-2\pi\imath(j'-1)(j-1)/N} \tag{3.37}$$

and the inverse DFT of a vector $\mathbf{w} \in \mathbb{C}^N$ is the vector $\tilde{\mathbf{w}}$, whose components are calculated via the formula

$$\tilde{\mathrm{w}}_j = \frac{1}{\sqrt{N}} \sum_{j'=1}^{N} \mathrm{w}_{j'} e^{2\pi\imath(j'-1)(j-1)/N}. \tag{3.38}$$

The two-dimensional DFT of a 2D array can be obtained by computing the DFT of its columns, followed by a DFT of its rows. A similar approach is used for the inverse two-dimensional DFT. The DFT and inverse DFT computations can be written as matrix-vector multiplication operations, which may be computed by means of the FFT algorithm, see e.g. [12], [94] and Matlab's documentation on the routines fft, ifft, fft2 and ifft2. In general, the speed of the algorithms depends on the size of the vector $\mathbf{x}$: they are most efficient if the dimensions have only small prime factors. In particular, if N is a power of 2, the cost is $N\log_2(N)$.

Thus, matrix-vector multiplications with $\mathbf{\Phi}$ and $\mathbf{\Phi}^*$ may be performed quickly without constructing the matrices explicitly and since the first column of $\mathbf{\Phi}$ is the vector of all ones scaled by the square root of the dimension, denoting with $\mathbf{a}_1$ and $\boldsymbol{\phi}_1$ the first column of $\mathbf{A}$ and $\mathbf{\Phi}$ respectively, it follows that

$$\mathbf{\Phi}\mathbf{a}_1 = \mathbf{\Lambda}\boldsymbol{\phi}_1 = \frac{1}{\sqrt{N}}\boldsymbol{\lambda}, \tag{3.39}$$

where $\boldsymbol{\lambda}$ is the column vector of the eigenvalues of $\mathbf{A}$.
As a consequence, even in the 2D case, the spectral decomposition of the matrix $\mathbf{A}$ can be calculated with a reasonable computational effort, so it is possible to apply the techniques we have seen in Chapter 3 for this particular problem. In particular, we shall be able to compute the SVD coefficients (*Fourier coefficients*) $|\boldsymbol{\phi}_j^*\mathbf{g}|$ and the ratios $|\boldsymbol{\phi}_j^*\mathbf{g}|/\lambda_j$, and arrest the CGNE method according to the stopping rule of Definition 3.4.1.
Moreover, when the spectral decomposition of $\mathbf{A}$ is known, matrix-vector multiplications can be performed very efficiently in Matlab using the DFT. For example, as shown in [38], given the PSF matrix $\mathbf{H}$ and an image $\mathbf{F}$:

- to compute the eigenvalues of $\mathbf{A}$, use

$$\mathsf{S} = \mathsf{fft2(fftshift(H))}; \tag{3.40}$$

- to compute the blurred image $\mathbf{G} = \mathbf{H} * \mathbf{F}$, use[3]

$$\mathsf{G} = \mathsf{real(ifft2(H} \circledcirc \mathsf{fft2(F)))}. \tag{3.41}$$

3.7 SR3: Projected Noise Norm Criterion

In this section we define an a-posteriori stopping rule, based on a statistical approach, suited mainly for large scale problems. The aim is again to approximate the norm of the projection of the noise onto the high frequency components

$$\|\mathbf{U}_{m-p}^*\mathbf{e}\|, \tag{3.42}$$

where $p \geq 0$ is the number of the low frequency components of the problem. We assume that p can be determined by means of an algorithm: for example, in the case of image deblurring, the algorithm of Section 3.4 can be applied. We simply consider a modified version of the Discrepancy Principle with δ replaced by the expected value of $\|\mathbf{U}_{m-p}^*\mathbf{e}\|$:

[3] the operation $\circledcirc$ is the componentwise multiplication of two matrices.

Definition 3.7.1 (Projected Noise Norm Criterion). *Suppose the matrix* $\mathbf{A} \in \mathbb{M}_{m,N}(\mathbb{R})$ *has* $m-p$ *noise-dominated* SVD *coefficients. Fix* $\tau > 1$ *and stop the iteration of* CGNE *as soon as the norm of the residual is lower or equal to* $\tau\bar{\delta}$ *with* $\bar{\delta} := \delta\sqrt{(m-p)/m}$*:*

$$k_{\mathrm{SR3}} := \min\{k \in \mathbb{N} \mid \|\mathbf{A}\mathbf{z}_k^\delta - \mathbf{b}^\delta\| \leq \tau\bar{\delta}\}. \tag{3.43}$$

Note that the definition does not require a SVD of the matrix $\mathbf{A}$, but only a knowledge about p.

The following result provides a theoretical justification for the definition above: if m is big, with high probability $\bar{\delta}$ is not smaller than $\|\mathbf{U}_{m-p}^*\mathbf{e}\|$.

Theorem 3.7.1. *Let* $\epsilon_1 > 0$ *and* $\epsilon_2 > 0$ *be small positive numbers and let* $\alpha \in (0, 1/2)$*. Then there exists a positive integer* $\bar{m} = \bar{m}(\epsilon_1, \epsilon_2, \alpha)$ *such that for every* $m > \bar{m}$ *the estimate*

$$\mathbb{P}\left(\|\mathbf{U}_{m-p}^*\mathbf{e}\|^2 - \epsilon_1 > \bar{\delta}^2\right) \leq \epsilon_2 \tag{3.44}$$

holds whenever the following conditions are satisfied:

(i) $p \leq \alpha m$;

(ii) $\mathbf{e} \sim \mathcal{N}(0, \nu^2\mathbf{I}_m)$, $\nu > 0$;

(iii) $\delta^2 = \|\mathbf{e}\|^2$.

Before proving the theorem, a few remarks:

- The theorem is based on the simple idea that if $\mathbf{e} \sim \mathcal{N}(0, \nu^2\mathbf{I}_m)$, then $\mathbf{U}^*\mathbf{e}$ has the same distribution: this argument fails if the perturbation has a different distribution!

- In principle it is possible, but maybe hard, to calculate $\bar{m}$ in terms of ϵ_1, ϵ_2 and α. For this reason we do not recommend this stopping rule for the small and medium size problems of Section 3.4.

- The assumption on p restricts the validity of the statement: luckily enough, in most cases p is small with respect to m. When this does not happen ($p >> m/2$), the choice $p = m/2$ should improve the performances anyway.

Proof. Fix $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $\alpha \in (0, 1/2)$. Let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$ be an SVD of $\mathbf{A}$ and let $p \in \{1, ..., m-1\}$. We observe that if $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I}_m)$, $\nu > 0$, then $\mathbf{U}^*\mathbf{e} \sim \mathcal{N}(0, \nu^2\mathbf{I}_m)$, so $\mathbf{U}^*_{m-p}\mathbf{e} \sim (0, ..., 0, \mathrm{e}_{p+1}, ..., \mathrm{e}_m)$. Thus, if p, $\mathbf{e}$ and δ satisfy (i), (ii) and (iii) respectively, there holds:

$$\mathbb{P}\left(\|\mathbf{U}^*_{m-p}\mathbf{e}\|^2 - \frac{\delta^2(m-p)}{m} > \epsilon_1\right) = \mathbb{P}\left(\sum_{i=p+1}^{m} \mathrm{e}_i^2 - \frac{\delta^2(m-p)}{m} > \epsilon_1\right). \tag{3.45}$$

For $0 < t < \min\{1/(2p\nu^2)\}$, Markov's inequality and our assumptions (ii) and (iii) yield

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=p+1}^{m} \mathrm{e}_i^2 - \frac{\delta^2(m-p)}{m} > \epsilon_1\right) \\
&= \mathbb{P}\left(p\sum_{i=p+1}^{m} \mathrm{e}_i^2 - (m-p)\sum_{i=1}^{p} \mathrm{e}_i^2 > m\epsilon_1\right) \\
&= \mathbb{P}\left(\exp\left[t\left(p\sum_{i=p+1}^{m} \mathrm{e}_i^2 - (m-p)\sum_{i=1}^{p} \mathrm{e}_i^2\right)\right] > \exp[tm\epsilon_1]\right) \\
&\leq \exp[-tm\epsilon_1]\mathbb{E}\left(\exp\left[t\left(p\sum_{i=p+1}^{m} \mathrm{e}_i^2 - \sum_{i=1}^{p} \mathrm{e}_i^2\right)\right]\right) \\
&= \exp[-tm\epsilon_1](1-2ts\nu^2)^{-\frac{m-p}{2}}(1+2t(m-p)\nu^2)^{-\frac{p}{2}},
\end{aligned} \tag{3.46}$$

where in the last equality we have used the assumption (ii) and the fact that if X is a random variable with gaussian distribution $X \sim \mathcal{N}(0, \nu^2)$, then, for every $a < 1/(2\nu^2)$,

$$\mathbb{E}(\exp[aX^2]) = (1-2a\nu^2)^{-\frac{1}{2}}. \tag{3.47}$$

Putting $w := 2t\nu^2$ the right-hand side of (3.46) can be rewritten as

$$\exp\left[\frac{-m\epsilon_1 w}{2\nu^2}\right](1-sw)^{-\frac{m-p}{2}}(1+(m-p)w)^{-\frac{p}{2}}. \tag{3.48}$$

If $p \leq \sqrt{m}$, the expression above can be made $< \epsilon_2$ choosing w close enough to $1/p$ for all m sufficiently large. On the other hand, if $\sqrt{m} < p \leq \alpha m$ and $w = o(\frac{1}{m})$, a Taylor expansion of the second and the third factor gives

$$\begin{aligned}
&\exp\left[-\frac{m\epsilon_1 w}{2\nu^2} + \frac{m-p}{2}(sw + \frac{p^2w^2}{2} + O(p^3w^3))\right.\\
&\left.-\frac{p}{2}\left((m-p)w + \frac{(m-p)^2w^2}{2} + O((m-p)^3w^3)\right)\right] \\
&= \exp\left[-\frac{m\epsilon_1 w}{2\nu^2} - \frac{p(m-p)(m-2p)w^2}{4}(1+o(1))\right],
\end{aligned} \tag{3.49}$$

thus it is possible to choose w (e.g., $w \approx 1/(mp^{1/4})$) in such a way that the last expression can be made arbitrarily small for all m sufficiently large. Summing up, we have proved that there exists $\bar{m} \in \mathbb{N}$ such that for all $m > \bar{m}$ and for all $p < \alpha m$ there holds

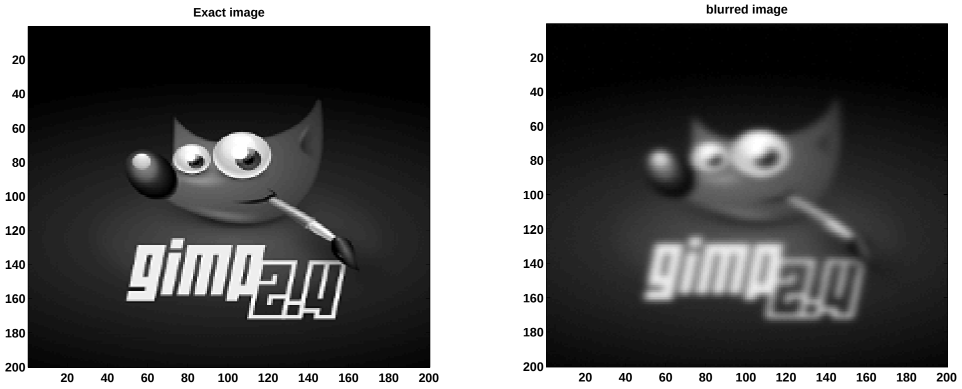
$$\inf\{t > 0 \mid \exp[-tm\epsilon_1](1 - 2ts\nu^2)^{-\frac{m-p}{2}}(1 + 2t(m-p)\nu^2)^{-\frac{p}{2}}\} \leq \epsilon_2 \tag{3.50}$$

and according to (3.45) and (3.46) this completes the proof. □

3.8 Image deblurring: numerical experiments

This section is dedicated to show the performances of the stopping rules SR2 and SR3 (with p as in the SR2) for the case of image deblurring.

In all examples below, we are given an exact image $\mathbf{F}$ and a PSF $\mathbf{H}$. Assuming periodic boundary conditions on the matrix $\mathbf{A}$ corresponding to $\mathbf{H}$ and perturbing $\mathbf{G} := \mathbf{H} * \mathbf{F}$ with white gaussian noise, we get a blurred and noisy image $\mathbf{G}^\delta = \mathbf{G} + \mathbf{E}$, where $\mathbf{E} \in \mathbb{M}_J(\mathbb{R})$ is the noise matrix. The problem fits the framework of Section 3.6: as a consequence, the singular values and the Fourier coefficients are computed by means of the 2D Discrete Fourier Transform fft2 using the routine fou_coeff from the Appendix. The stopping rules SR2 and SR3 can be applied with $\delta = \|\mathbf{e}\| = \|\mathbf{g}^\delta - \mathbf{A}\mathbf{f}\|$, where the vectors $\mathbf{e}$, $\mathbf{g}^\delta$ and $\mathbf{f}$ are the columnwise stacked versions of the matrices $\mathbf{E}$, $\mathbf{G}^\delta$ and $\mathbf{F}$ respectively.



(a) Exact image.

(b) Perturbed image: stdev $= 2.0$, $\varrho = 1.0\%$.

Figure 3.6: The exact data $\mathbf{F}$ and a perturbed data $\mathbf{G}^{\delta}$ for the gimp test problem.

We fix $\tau = 1.001$, run cgls_deb, a modified version of the function cgls from the Regularization Tools and compare the results obtained by the Discrepancy Principle, SR2 and SR3 with the optimal solution.

3.8.1 Test 1 (gimp test problem)

The dimension of the square matrix $\mathbf{F} \in \mathbb{M}_J(\mathbb{R})$ corresponding to the image gimp.png, shown on the left in Figure 3.6, is given by $J = 200$. The algorithm blurring, described in the Appendix, generates a Gaussian PSF $\mathbf{H}$ and a blurred and noisy image.

We consider different values for the standard deviation of the Gaussian PSF and for the noise level, and compare the results obtained by stopping CGNE with the discrepancy principle, SR2 and SR3.

We underline that in almost all the considered problems the computation of the index p of the stopping rule SR2 is made using only the first $N/2 = J^2/2$ Fourier coefficients, to spare time in the minimization process of the function mod_min_max. When stdv $= 1.0$ and $\varrho = 0.1, 0.5$, all the Fourier coefficients are used, since in these cases they are necessary for calculating a good value

CGNE results for the gimp test problem					
stdev	ϱ	Relative error (stopping index)			
		Discr. Princ.	SR2	SR3	Opt. Sol.
3.0	0.1%	0.2104(154)	0.2082(196)	0.2053(313)	0.2048(348)
3.0	0.5%	0.2257(49)	0.2259(48)	0.2194(86)	0.2182(112)
3.0	1.0%	0.2331(34)	0.2286(45)	0.2264(61)	0.2261(68)
3.0	5.0%	0.2733(13)	0.3079(7)	0.2613(17)	0.2583(22)
2.0	0.1%	0.1265(144)	0.1703(61)	0.1196(213)	0.1167(310)
2.0	0.5%	0.1846(36)	0.1762(54)	0.1612(96)	0.1599(109)
2.0	1.0%	0.1962(20)	0.1962(20)	0.1878(36)	0.1849(53)
2.0	5.0%	0.2199(7)	0.2228(6)	0.2160(9)	0.2149(11)
1.0	0.1%	0.0442(45)	0.0419(50)	0.0350(68)	0.0344(92)
1.0	0.5%	0.0725(15)	0.0652(22)	0.0636(28)	0.0636(28)
1.0	1.0%	0.0865(9)	0.0797(13)	0.0781(15)	0.0780(16)
1.0	5.0%	0.1244(4)	0.1435(3)	0.1194(5)	0.1194(5)

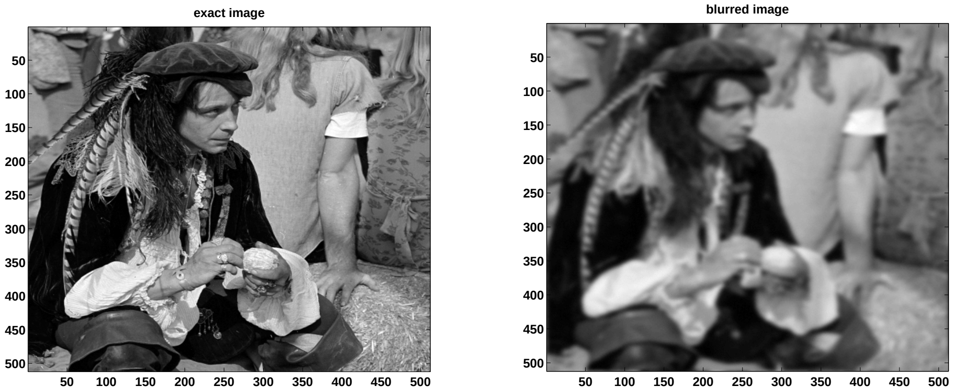
Table 3.5: Comparison between different stopping rules of CGNE for the gimp problem.

of p.

The numerical results of Table 3.5 show that the a-posteriori stopping rule SR3 always finds a relative error lower than that obtained by the Discrepancy Principle. The heuristic stopping rule SR2 gives very good results apart from some cases where it provides an over-regularized solution. Since the performance of SR3 is excellent here and changing the Matlab seed does not make things significantly different (i.e. the statistical approximation of $\|\mathbf{U}_{m-p}^{*}\mathbf{e}\|$ seems to be very solid in such a large-size problem), we deduce that the approximation of the residual norm of the TSVD solution with the norm of the projection of the noise onto the high frequency components

$$\|\mathbf{g}^{\delta} - \mathbf{A}\mathbf{f}_{p}^{\mathrm{TSVD}}\| \sim \|\mathbf{U}_{m-p}^{*}\mathbf{e}\|$$

is not very appropriate in these cases.



(a) Exact image. (b) Perturbed image: stdev $= 2.0$, $\varrho = 1.0\%$.

Figure 3.7: The exact data $\mathbf{F}$ and a perturbed data $\mathbf{G}^{\delta}$ for the pirate test problem.

3.8.2 Test 2 (pirate test problem)

The image pirate.tif, shown on the left in Figure 3.7, has a higher resolution than gimp.png: $J = 512$.

We proceed as in the gimp test problem, but with a few variations:

- instead of the values 1.0, 2.0, 3.0 for the stdev we consider the values 3.0, 4.0, 5.0;

- To compute the index p of the stopping rules SR2 and SR3, instead of considering the first $N/2$ ratios $\varphi_i = |\boldsymbol{\phi}_i^* \mathbf{g}^{\delta}|/\lambda_i$, we take only the first $N/4$. Moreover, in the computation of the curve that approximates the φ_i, we use the function data_approx with only $\lfloor N/200 \rfloor$ inner knots (instead of $\lfloor N/50 \rfloor$).

The results, summarized in Table 3.6, show that both SR2 and SR3 give excellent results, finding a relative error lower than the Discrepancy Principle in almost all cases. The phenomenon observed in the gimp test problem concerning SR2 appears to be much more attenuate here.

CGNE results for the pirate test problem					
stdev	ϱ	Relative error (stopping index)			
		Discr. Princ.	SR2	SR3	Opt. Sol.
5.0	0.1%	0.1420(171)	0.1420(170)	0.1384(330)	0.1375(484)
5.0	0.5%	0.1524(49)	0.1519(52)	0.1481(90)	0.1472(122)
5.0	1.0%	0.1579(29)	0.1551(40)	0.1529(57)	0.1524(69)
5.0	5.0%	0.1746(9)	0.1746(9)	0.1697(14)	0.1686(18)
3.0	0.1%	0.1076(104)	0.1051(153)	0.1032(233)	0.1029(280)
3.0	0.5%	0.1187(31)	0.1161(42)	0.1141(61)	0.1140(69)
3.0	1.0%	0.1251(18)	0.1251(18)	0.1204(31)	0.1198(39)
3.0	5.0%	0.1425(6)	0.1405(7)	0.1385(9)	0.1382(10)
4.0	0.1%	0.1268(142)	0.1260(158)	0.1226(290)	0.1219(397)
4.0	0.5%	0.1379(40)	0.1408(29)	0.1343(65)	0.1329(98)
4.0	1.0%	0.1432(24)	0.1418(28)	0.1389(43)	0.1385(53)
4.0	5.0%	0.1591(8)	0.1566(10)	0.1549(13)	0.1547(14)

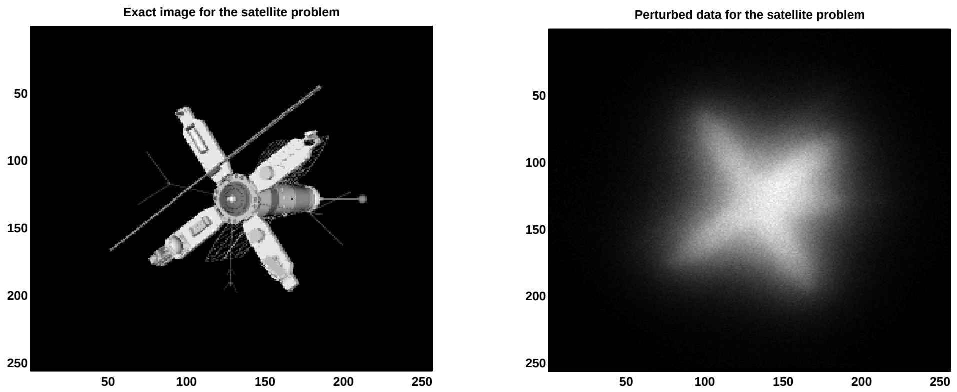
Table 3.6: Comparison between different stopping rules of CGNE for the pirate test problem.

3.8.3 Test 3 (satellite test problem)

CGNE results for the satellite problem				
	Discr. Princ.	SR2	SR3	Opt. Sol.
Stopping index	23	34	28	34
Relative Error	0.3723	0.3545	0.3592	0.3545

Table 3.7: Comparison between different stopping rules of CGNE for the satellite problem.

The data for this example were developed at the US Air Force Phillips Laboratory and have been used to test the performances of several available algorithms for computing regularized nonnegative solutions, cf. e.g. [29] and [5]. The data consist of an (unknown) exact gray-scale image $\mathbf{F}$, a space invariant point spread function $\mathbf{H}$ and a perturbed version $\mathbf{G}^\delta$ of the blurred image $\mathbf{G} = \mathbf{H} * \mathbf{F}$, see Figure 3.8. All images $\mathbf{F}$, $\mathbf{H}$ and $\mathbf{G}^\delta$ are 256×256


Exact image for the satellite problem

Perturbed data for the satellite problem

(a) Exact image

(b) Perturbed image

Figure 3.8: The exact data $\mathbf{F}$ and the perturbed data $\mathbf{G}^{\delta}$ for the satellite problem.

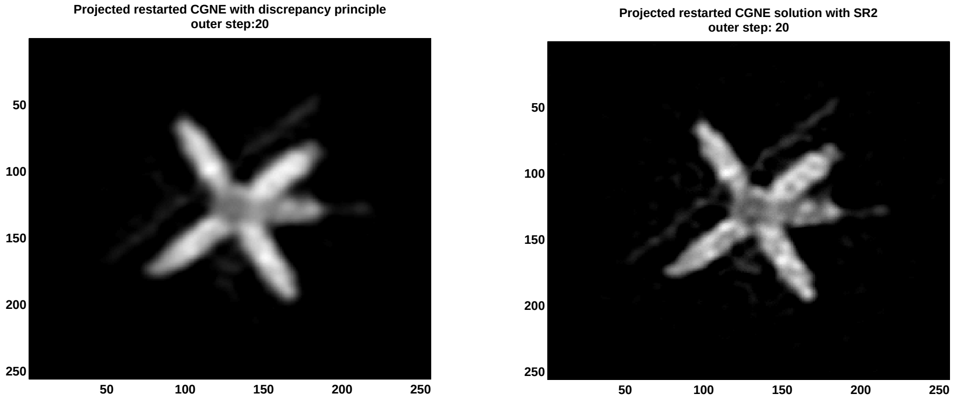
matrices with nonnegative entries.
The results, gathered in Table 3.7 show that both SR2 and SR3 improve the results of the Discrepancy Principle in this example and the stopping index of SR2 coincides with the optimal stopping index $k^{\sharp}$.

3.8.4 The new stopping rules in the Projected Restarted CGNE

These stopping rules may work very well also when CGNE is combined with other regularization methods. To show this, we consider CGNE as the inner iteration of the *Projected Restarted Algorithm* from [5]. Using the notations of Chapter 2, it is a straightforward exercise to prove that Algorithm 1 of [5] is equivalent to the following scheme:

- Fix $\tilde{\mathbf{f}}^{(0)} = \mathbf{f}^{(0)} = \mathbf{0}$ and $i = 0$.

- For every $i = 0, 1, 2, ...$, if $\mathbf{f}^{(i)}$ does not satisfy the Discrepancy Principle (respectively, SR2, SR3), compute $\tilde{\mathbf{f}}^{(i+1)}$ as the regularized solution of CGNE applied to the system $\mathbf{A}\mathbf{f} = \mathbf{g}^{\delta}$ with initial guess $\tilde{\mathbf{f}}^{(i)}$ arrested



(a) Discrepancy principle: rel. err. 0.3592. (b) SR2: rel. err. 0.3302.

Figure 3.9: The solutions of the satellite problem reconstructed by the Projected Restarted CGNE algorithm at the 20-th outer step.

with the Discrepancy Principle (respectively, SR2, SR3) and define $\mathbf{f}^{(i+1)}$ as the projection of $\tilde{\mathbf{f}}^{(i+1)}$ onto the set $\mathbb{W}$ of nonnegative vectors

$$\mathbb{W} := \left\{ \mathbf{f} \in \mathbb{R}^N \mid \mathbf{f} \geq \mathbf{0} \right\}. \tag{3.51}$$

- Stop the iteration as soon as $\mathbf{f}^{(i)}$ satisfies the Discrepancy Principle (respectively, SR2, SR3) or a prefixed number of iteration has been carried out.

An implementation of this scheme for the satellite test problem with $\tau = 1.001$ leads to the results of Table 3.8. The relative errors obtained by arresting CGNE according to SR2 and SR3 are smaller than those obtained by means of the Discrepancy Principle. We underline that the relative errors of the stopping rule SR2 are even lower than those obtained in [5] with RRGMRES instead of CGNE. The regularized solutions obtained by this projected restarted CGNE with the Discrepancy Principle and SR2 as stopping rules at the 20-th outer iteration step are shown in Figure 3.9.

Projected Restarted CGNE results for the satellite problem				
		Discr. Princ.	SR2	SR3
2 outer steps	total CGNE steps	27	45	33
	Relative Error	0.3644	0.3400	0.3499
5 outer steps	total CGNE steps	37	74	47
	Relative Error	0.3614	0.3347	0.3465
10 outer steps	total CGNE steps	46	108	64
	Relative Error	0.3601	0.3321	0.3446
20 outer steps	total CGNE steps	57	162	86
	Relative Error	0.3592	0.3302	0.3432
50 outer steps	total CGNE steps	60	274	120
	Relative Error	0.3590	0.3290	0.3422
200 outer steps	total CGNE steps	60	532	129
	Relative Error	0.3590	0.3286	0.3420

Table 3.8: Comparison between different stopping rules of CGNE as inner iteration of the Projected restarted algorithm for the satellite problem.

Chapter 4

Tomography

The term tomography is derived from the Greek word $\tau o\mu o s$, slice. It stands for a variety of different techniques for imaging two-dimensional cross sections of three-dimensional objects. The impact of these techniques in diagnostic medicine has been revolutionary, since it has enabled doctors to view internal organs with unprecedented precision and safety to the patient. We have already seen in the first Chapter that these problems can be mathematically described by means of the Radon Transform. The aim of this chapter is to analyze the main properties of the Radon Transform and to give an overview of the most important algorithms.
General references for this chapter are [19], [68], [69] and [43].

4.1 The classical Radon Transform

In this section we provide an outline of the main properties of the Radon Transform over hyperplanes of $\mathbb{R}^D$. An hyperplane $\mathbb{H}$ of $\mathbb{R}^D$ can be represented by an element of the unit cylinder

$$\mathfrak{C}^D := \{(\boldsymbol{\theta}, s) \mid \boldsymbol{\theta} \in \mathbb{S}^{D-1}, s \in \mathbb{R}\}, \tag{4.1}$$

via the formula

$$\mathbb{H}(\boldsymbol{\theta}, s) = \{\mathbf{x} \in \mathbb{R}^D \mid \langle \mathbf{x}, \boldsymbol{\theta} \rangle = s\}, \tag{4.2}$$

where $\langle\cdot,\cdot\rangle$ denotes the usual euclidean inner product of $\mathbb{R}^D$ and where we identify $\mathbb{H}(-\boldsymbol{\theta},-s)$ with $\mathbb{H}(\boldsymbol{\theta},s)$. We denote the set of all hyperplanes of $\mathbb{R}^D$ with

$$\Xi_D := \mathfrak{C}^D/\mathbb{Z}_2, \tag{4.3}$$

and define the Radon Transform of a rapidly decreasing function $f \in \mathcal{S}(\mathbb{R}^D)$ as its integral on $\mathbb{H}(\boldsymbol{\theta},s)$:

$$\mathscr{R}[f](\boldsymbol{\theta},s) := \int_{\mathbb{H}(\boldsymbol{\theta},s)} f(\mathbf{x})d\mu_{\mathbb{H}}(\mathbf{x}), \tag{4.4}$$

where $\mu_{\mathbb{H}}(\mathbf{x})$ is the Lebesgue measure on $\mathbb{H}(\boldsymbol{\theta},s)$.
Much of the theory of the Radon Transform is based on its behavior under the Fourier Transform and convolution. We recall that the Fourier Transform of a function $f \in \mathcal{L}^1(\mathbb{R}^D)$ is given by

$$\hat{f}(\mathbf{y}) = \mathscr{F}(f)(\mathbf{y}) = (2\pi)^{-D/2}\int_{\mathbb{R}^D} f(\mathbf{x})e^{-\imath\langle\mathbf{x},\mathbf{y}\rangle}d\mathbf{x}, \quad \mathbf{y}\in\mathbb{R}^D. \tag{4.5}$$

Observing that the exponential $e^{-\imath\langle\mathbf{x},\mathbf{y}\rangle}$ is constant on hyperplanes orthogonal to $\mathbf{y}$, an important relation between the Fourier and the Radon Transform is obtained integrating (4.5) along such hyperplanes. Explicitly, we write $\mathbf{y}=\xi\boldsymbol{\theta}$ for $\xi\in\mathbb{R}$ and $\boldsymbol{\theta}\in\mathbb{S}^{D-1}$ to get the famous *Projection-Slice Theorem*:

$$\begin{aligned}\hat{f}(\xi\boldsymbol{\theta}) &= (2\pi)^{-D/2}\int_{\mathbb{R}}\int_{\mathbb{H}(\boldsymbol{\theta},s)} f(\mathbf{x})e^{-\imath\xi\langle\mathbf{x},\boldsymbol{\theta}\rangle}d\mu_{\mathbb{H}}(\mathbf{x})ds\\ &= (2\pi)^{-D/2}\int_{\mathbb{R}}\left(\int_{\mathbb{H}(\boldsymbol{\theta},s)} f(\mathbf{x})d\mu_{\mathbb{H}}(\mathbf{x})\right)e^{-\imath s\xi}ds\\ &= (2\pi)^{-D/2}\int_{\mathbb{R}}\mathscr{R}[f](\boldsymbol{\theta},s)e^{-\imath s\xi}ds.\end{aligned} \tag{4.6}$$

This immediately implies that the operator $\mathscr{R}$ is injective on $\mathcal{S}(\mathbb{R}^D)$ (but also on larger spaces, e.g. on $\mathcal{L}^1(\mathbb{R}^D)$).
Moreover, let us introduce the space of Schwartz-class functions on Ξ_D. We say that a function $f\in\mathcal{C}^\infty(\mathfrak{C}^D)$ belongs to $\mathcal{S}(\mathfrak{C}^D)$ if for every k_1,$k_2\in\mathbb{N}_0$ we have

$$\sup_{(\boldsymbol{\theta},s)}(1+|s|)^{k_1}\left|\frac{\partial^{k_2}}{\partial s^{k_2}}f(\boldsymbol{\theta},s)\right| < +\infty.$$

The space $\mathcal{S}(\Xi_D)$ is the space of the even functions in $\mathcal{S}(\mathfrak{C}^D)$, i.e. $f(\boldsymbol{\theta}, s) = f(-\boldsymbol{\theta}, -s)$. The *Partial Fourier Transform* of a function $f \in \mathcal{S}(\Xi_D)$ is its Fourier Transform in the s variable:

$$f(\boldsymbol{\theta}, s) \mapsto \hat{f}(\boldsymbol{\theta}, \xi) = \mathscr{F}\left(s \mapsto f(\boldsymbol{\theta}, s)\right)(\xi) = \int_{\mathbb{R}} f(\boldsymbol{\theta}, s) e^{-\imath \xi s} ds. \tag{4.7}$$

On Ξ_D, we understand the convolution as acting on the second variable as well:

$$(g * h)(\boldsymbol{\theta}, s) = \int_{\mathbb{R}} g(\boldsymbol{\theta}, s - t) h(t) dt. \tag{4.8}$$

The Partial Fourier Transform maps $\mathcal{S}(\mathfrak{C}^D)$ into itself and $\mathcal{S}(\Xi_D)$ into itself and the Projection-Slice Theorem states that the Fourier Transform of $f \in \mathcal{S}(\mathbb{R}^D)$ is the Partial Fourier Transform of its Radon Transform. For $f \in \mathcal{S}(\mathbb{R}^D)$, if we change to polar coordinates $(\boldsymbol{\theta}, s) \mapsto s\boldsymbol{\theta}$ in $\mathbb{R}^D$, a straightforward application of the chain rule shows that its Fourier Transform lies in $\mathcal{S}(\mathfrak{C}^D)$. By the Projection-Slice Theorem then

$$\mathscr{R} : \mathcal{S}(\mathbb{R}^D) \longrightarrow \mathcal{S}(\Xi_D). \tag{4.9}$$

Another consequence of the Projection-Slice Theorem is that the Radon Transform preserves convolutions, in the sense that for every f and $g \in \mathcal{S}(\mathbb{R}^D)$ the following formula holds:

$$\mathscr{R}(f * g)(\boldsymbol{\theta}, s) = \int_{\mathbb{R}} \mathscr{R}[f](\boldsymbol{\theta}, s - t) \mathscr{R}[g](\boldsymbol{\theta}, t) dt. \tag{4.10}$$

We now introduce the *backprojection operator* $\mathscr{R}^*$ by

$$\mathscr{R}^*[g](\mathbf{x}) = \int_{\mathbb{S}^{D-1}} g(\boldsymbol{\theta}, \langle \mathbf{x}, \boldsymbol{\theta} \rangle) d\boldsymbol{\theta}, \quad g \in \mathcal{S}(\Xi_D). \tag{4.11}$$

For $g = \mathscr{R} f$, $\mathscr{R}^*[g](\mathbf{x})$ is the average of all hyperplane integrals of f through $\mathbf{x}$. Mathematically speaking, $\mathscr{R}^*$ is simply the adjoint of $\mathscr{R}$: for $\phi \in \mathcal{S}(\mathbb{R})$ and $f \in \mathcal{S}(\mathbb{R}^D)$, there holds

$$\int_{\mathbb{R}} \phi(s) \mathscr{R}[f](\boldsymbol{\theta}, s) ds = \int_{\mathbb{R}^D} \phi(\langle \mathbf{x}, \boldsymbol{\theta} \rangle) f(\mathbf{x}) d\mathbf{x} \tag{4.12}$$

and consequently for $g \in \mathcal{S}(\Xi_D)$ and $f \in \mathcal{S}(\mathbb{R}^D)$

$$\int_{\mathbb{S}^{D-1}} \int_{\mathbb{R}} g(\mathscr{R}f) d\boldsymbol{\theta} ds = \int_{\mathbb{R}^D} (\mathscr{R}^* g) f d\mathbf{x}. \tag{4.13}$$

A more general approach for studying the Radon Transform leading to the same result can be found in the first chapters of [19].

The following result is the starting point for the Filtered Backprojection algorithm, which will be discussed later.

Theorem 4.1.1. *Let $f \in \mathcal{S}(\mathbb{R}^D)$ and $\upsilon \in \mathcal{S}(\Xi_D)$. Then*[1]

$$\mathscr{R}^*(\upsilon * \mathscr{R}f) = (\mathscr{R}^*\upsilon) * f. \tag{4.14}$$

Proof. For any $\mathbf{x} \in \mathbb{R}^D$, we have

$$\begin{aligned}
\mathscr{R}^*(\upsilon * \mathscr{R}f)(\mathbf{x}) &= \int_{\mathbb{S}^{D-1}} \left(\int_{\mathbb{R}} \upsilon(\boldsymbol{\theta}, \langle \boldsymbol{\theta}, \mathbf{x} \rangle - s) \mathscr{R}[f](\boldsymbol{\theta}, s) ds \right) d\boldsymbol{\theta} \\
&= \int_{\mathbb{S}^{D-1}} \left(\int_{\mathbb{R}} \upsilon(\boldsymbol{\theta}, \langle \boldsymbol{\theta}, \mathbf{x} \rangle - s) \left(\int_{\mathbb{H}(\boldsymbol{\theta},s)} f(\mathbf{y}) d\mu_{\mathbb{H}}(\mathbf{y}) \right) ds \right) d\boldsymbol{\theta} \\
&= \int_{\mathbb{S}^{D-1}} \left(\int_{\mathbb{R}^D} \upsilon(\boldsymbol{\theta}, \langle \boldsymbol{\theta}, \mathbf{x} - \mathbf{y} \rangle) f(\mathbf{y}) d\mathbf{y} \right) d\boldsymbol{\theta} \\
&= \int_{\mathbb{R}^D} \left(\int_{\mathbb{S}^{D-1}} \upsilon(\boldsymbol{\theta}, \langle \boldsymbol{\theta}, \mathbf{x} - \mathbf{y} \rangle) d\boldsymbol{\theta} \right) f(\mathbf{y}) d\mathbf{y} \\
&= \int_{\mathbb{R}^D} \mathscr{R}^* \upsilon(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} \\
&= ((\mathscr{R}^*\upsilon) * f)(\mathbf{x}).
\end{aligned} \tag{4.15}$$

□

4.1.1 The inversion formula

We are now ready to derive the inversion formula for the Radon Transform. The proof is basically taken from [19], but here, apart from the different notations and definitions, some small errors in the resulting constants have

[1] Note the different meaning of the symbol $*$ in the formula!

been corrected. We state the general theorem exactly as in [69].
We start from the inversion formula of the classical Fourier Transform in $\mathbb{R}^D$

$$(2\pi)^{D/2} f(\mathbf{x}) = \int_{\mathbb{R}^D} \hat{f}(\mathbf{y}) e^{\imath\langle \mathbf{x},\mathbf{y}\rangle} d\mathbf{y} \tag{4.16}$$

and switch to polar coordinates $\mathbf{y} = \xi\boldsymbol{\theta}$, obtaining

$$\begin{aligned}(2\pi)^{D/2} f(\mathbf{x}) &= \int_{\mathbb{S}^{D-1}} \int_0^{+\infty} \hat{f}(\xi\boldsymbol{\theta}) e^{\imath\xi\langle \mathbf{x},\boldsymbol{\theta}\rangle} \xi^{D-1} d\xi d\boldsymbol{\theta} \\ &= \frac{1}{2}\int_{\mathbb{S}^{D-1}} \int_0^{+\infty} \hat{f}(\xi\boldsymbol{\theta}) e^{\imath\xi\langle \mathbf{x},\boldsymbol{\theta}\rangle} \xi^{D-1} d\xi d\boldsymbol{\theta} \\ &\qquad + \frac{1}{2}\int_{\mathbb{S}^{D-1}} \int_0^{+\infty} \hat{f}(\xi(-\boldsymbol{\theta})) e^{\imath\xi\langle \mathbf{x},-\boldsymbol{\theta}\rangle} \xi^{D-1} d\xi d\boldsymbol{\theta} \\ &= \frac{1}{2}\int_{\mathbb{S}^{D-1}} \int_0^{+\infty} \hat{f}(\xi\boldsymbol{\theta}) e^{\imath\xi\langle \mathbf{x},\boldsymbol{\theta}\rangle} \xi^{D-1} d\xi d\boldsymbol{\theta} \\ &\qquad + \frac{1}{2}\int_{\mathbb{S}^{D-1}} \int_{-\infty}^{0} \hat{f}(\xi\boldsymbol{\theta}) e^{\imath\xi\langle \mathbf{x},\boldsymbol{\theta}\rangle} (-\xi)^{D-1} d\xi d\boldsymbol{\theta} \\ &= \frac{1}{2}\int_{\mathbb{S}^{D-1}} \int_{-\infty}^{+\infty} \hat{f}(\xi\boldsymbol{\theta}) e^{\imath\xi\langle \mathbf{x},\boldsymbol{\theta}\rangle} |\xi|^{D-1} d\xi d\boldsymbol{\theta}.\end{aligned} \tag{4.17}$$

If D is odd, $|\xi|^{D-1} = \xi^{D-1}$, thus the Projection-Slice Theorem and the properties of the Fourier Transform in one variable imply

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2(2\pi)^D} \int_{\mathbb{S}^{D-1}} \int_{-\infty}^{+\infty} e^{\imath\xi\langle \mathbf{x},\boldsymbol{\theta}\rangle} \xi^{D-1} \left(\int_{\mathbb{R}} \mathscr{R}[f](\boldsymbol{\theta}, s) e^{-\imath s\xi} ds \right) d\xi d\boldsymbol{\theta} \\ &= c_D \int_{\mathbb{S}^{D-1}} \mathscr{F}^{-1} \left(\xi \mapsto \mathscr{F} \left(s \mapsto \frac{\partial^{D-1}}{\partial s^{D-1}} \mathscr{R}[f](\boldsymbol{\theta}, s) \right) (\xi) \right) (\langle \mathbf{x}, \boldsymbol{\theta}\rangle) d\boldsymbol{\theta} \\ &= c_D \mathscr{R}^* \left(\frac{\partial^{D-1}}{\partial s^{D-1}} \mathscr{R} f \right) (\mathbf{x}), \end{aligned} \tag{4.18}$$

where

$$c_D := \frac{2\pi}{2(\imath)^{D-1}(2\pi)^D} = \frac{1}{2}(2\pi)^{1-D}(-1)^{\frac{D-1}{2}}. \tag{4.19}$$

Suppose now D is even. To obtain a complete inversion formula, we recall a few facts from the theory of distributions (cf. [19] and, as a general reference for the theory of distributions, [80]).

1. The linear mapping $p.v.[1/t]$ from $\mathcal{S}(\mathbb{R})$ to $\mathbb{C}$ defined by
$$p.v.[1/t]h = \lim_{\epsilon\to 0^+}\int_{|t|>\epsilon}\frac{h(t)}{t}dt \tag{4.20}$$
is well defined (although $1/t$ is not locally integrable) and belongs to the dual space of $\mathcal{S}(\mathbb{R})$; that is to say, it is a tempered distribution on $\mathbb{R}$.

2. The signum function on $\mathbb{R}$ defined by
$$\operatorname{sgn}(\xi) = \begin{cases} 1 & \text{if} \quad \xi \geq 0, \\ -1 & \text{if} \quad \xi < 0 \end{cases} \tag{4.21}$$
is a tempered distribution on $\mathbb{R}$ as well, whose (distributional) Fourier Transform is related to $p.v.[1/t]$ via the formula
$$\mathscr{F}(p.v.[1/t])(\xi) = -\sqrt{\frac{\pi}{2}}\imath\operatorname{sgn}(\xi). \tag{4.22}$$

The Hilbert Transform of a function $\phi \in \mathcal{S}(\mathbb{R})$ is the convolution $\mathscr{H}\phi := \phi * \frac{1}{\pi}p.v.[1/t]$, i.e.
$$\mathscr{H}[\phi](p) = \lim_{\epsilon\to 0^+}\int_{|t|>\epsilon}\frac{\phi(p-t)}{\pi t}dt, \quad p \in \mathbb{R}. \tag{4.23}$$
As a consequence, the Fourier Transform of $\mathscr{H}\phi$ is the function
$$\mathscr{F}[\mathscr{H}\phi](\xi) = -\imath\hat{\phi}(\xi)\operatorname{sgn}(\xi). \tag{4.24}$$
Now we return to the right-hand side of (4.17), which, for D even, is equal to
$$\frac{1}{2}\int_{\mathbb{S}^{D-1}}\int_{-\infty}^{+\infty}\hat{f}(\xi\boldsymbol{\theta})e^{\imath\xi\langle\mathbf{x},\boldsymbol{\theta}\rangle}\xi^{D-1}\operatorname{sgn}(\xi)d\xi d\boldsymbol{\theta}.$$
We proceed similarly to the odd case and using (4.24) we have
$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2(2\pi)^D}\int_{\mathbb{S}^{D-1}}\int_{-\infty}^{+\infty}e^{\imath\xi\langle\mathbf{x},\boldsymbol{\theta}\rangle}\xi^{D-1}\operatorname{sgn}(\xi)\left(\int_{\mathbb{R}}\mathscr{R}[f](\boldsymbol{\theta},s)e^{-\imath s\xi}ds\right)d\xi d\boldsymbol{\theta} \\ &= c_D\int_{\mathbb{S}^{D-1}}\mathscr{F}^{-1}\left(\xi\mapsto\mathscr{F}\left(\mathscr{H}(s\mapsto\frac{\partial^{D-1}}{\partial s^{D-1}}\mathscr{R}[f])(\boldsymbol{\theta},s)\right)(\xi)\right)(\langle\mathbf{x},\boldsymbol{\theta}\rangle)d\boldsymbol{\theta} \\ &= c_D\mathscr{R}^*\left(\mathscr{H}\left(\frac{\partial^{D-1}}{\partial s^{D-1}}\mathscr{R}[f]\right)\right)(\mathbf{x}), \end{aligned} \tag{4.25}$$

where

$$c_D := \frac{2\pi}{2(\imath)^{D-2}(2\pi)^D} = \frac{1}{2}(2\pi)^{1-D}(-1)^{\frac{D-2}{2}}. \tag{4.26}$$

Altogether, we have proved the following theorem.

Theorem 4.1.2. *Let $f \in \mathcal{S}(\mathbb{R}^D)$ and let $g = \mathscr{R}f$. Then*

$$f = c_D \begin{cases} \mathscr{R}^* \mathscr{H} g^{(D-1)} & n \text{ even}, \\ \mathscr{R}^* g^{(D-1)} & n \text{ odd}, \end{cases} \tag{4.27}$$

with

$$c_D = \frac{1}{2}(2\pi)^{1-D} \begin{cases} (-1)^{\frac{D-2}{2}} & n \text{ even}, \\ (-1)^{\frac{D-1}{2}} & n \text{ odd}, \end{cases} \tag{4.28}$$

where the derivatives in Ξ_D are intended in the second variable.

We conclude the section with a remark on the inversion formula from [69]. For D odd, the equation (4.18) says that $f(\mathbf{x})$ is simply an average of $g^{(D-1)}$ over all hyperplanes through $\mathbf{x}$. Thus, in order to reconstruct f at some point $\mathbf{x}$, one needs only the integrals of f through $\mathbf{x}$. This is not true for D even. In fact, inserting the definition of the Hilbert Transform into (4.25) and changing the order of integration, we obtain[2]

$$f(\mathbf{x}) = \lim_{\epsilon \to 0^+} c_D \int_{|t|>\epsilon} \frac{1}{t} \int_{\mathbb{S}^{D-1}} g^{(D-1)}(\boldsymbol{\theta}, \langle \mathbf{x}, \boldsymbol{\theta} \rangle - t) d\boldsymbol{\theta} dt, \tag{4.29}$$

from which we can see that the computation of f at some point $\mathbf{x}$ requires integrals of f also over hyperplanes far away from $\mathbf{x}$.

4.1.2 Filtered backprojection

Rather than using the inversion formula described above, the most common method for reconstructing X-ray images is the method of the *Filtered Backprojection.* Its main advantage is the ability to cancel out high frequency noise.

[2] the equality with the right-hand side of formula (4.25) is guaranteed because g is a rapidly decreasing function.

Let $\upsilon \in \mathcal{S}(\Xi_D)$. There is a constant C such that $|\upsilon(\xi)| \leq C$ for all $\xi \in \Xi_D$ and so the definition of $\mathscr{R}^*$ implies that also $|\mathscr{R}^*\upsilon(\mathbf{x})| \leq C$ for all $\mathbf{x}$. Thus $\mathscr{R}^*\upsilon$ is a tempered distribution. Moreover, in [19] it is shown that the following relation between $\mathscr{R}^*\upsilon$ and $\hat{\upsilon}$ holds for all $\mathbf{y} \neq \mathbf{0}$ in $\mathbb{R}^D$:

$$\mathscr{F}(\mathscr{R}^*\upsilon)(\mathbf{y}) = 2(2\pi)^{(D-1)/2}\|\mathbf{y}\|^{1-D}\hat{\upsilon}\left(\frac{\mathbf{y}}{\|\mathbf{y}\|}, \|\mathbf{y}\|\right). \tag{4.30}$$

The starting point of the Filtered Backprojection algorithm is Theorem 4.1.1: we put in formula (4.14) $g = \mathscr{R}f$ and $V = \mathscr{R}^*\upsilon$ obtaining

$$V * f = \mathscr{R}^*(\upsilon * g). \tag{4.31}$$

The idea is to choose V as an approximation of the Dirac δ-function and to determine υ from $V = \mathscr{R}^*\upsilon$. Then $V * f$ is an approximation of the sought function f that is calculated in the right-hand side by backprojecting the convolution of υ with the data g.

Usually only radially symmetric functions V are chosen, i.e. $V(\mathbf{y}) = V(\|\mathbf{y}\|)$. Then $\upsilon = \upsilon(\boldsymbol{\theta}, s)$ can be assumed to be only an even function of s and formula (4.30) reads now

$$\mathscr{F}(\mathscr{R}^*\upsilon)(\mathbf{y}) = 2(2\pi)^{(D-1)/2}\|\mathbf{y}\|^{1-D}\hat{\upsilon}(\|\mathbf{y}\|). \tag{4.32}$$

Now we choose V as a band limited function by allowing a filter factor $\hat{\phi}(\mathbf{y})$ which is close to 1 for $\|\mathbf{y}\| \leq 1$ and which vanishes for $\|\mathbf{y}\| > 1$ putting

$$\hat{V}_\Upsilon(\mathbf{y}) := (2\pi)^{-D/2}\hat{\phi}(\|\mathbf{y}\|/\Upsilon), \quad \Upsilon > 0. \tag{4.33}$$

Then the corresponding function υ_Υ such that $\mathscr{R}^*\upsilon_\Upsilon = V_\Upsilon$ satisfies

$$\hat{\upsilon}_\Upsilon(\xi) = \frac{1}{2}(2\pi)^{1/2-D}\xi^{D-1}\hat{\phi}(\xi/\Upsilon), \quad \xi > 0 \tag{4.34}$$

(note that $\hat{\upsilon}_\Upsilon$ is an even function being the Fourier Transform of an even function).

Many choices of $\hat{\phi}$ can be found in literature. We mention the choice proposed by Shepp and Logan in [84], where

$$\hat{\phi}(\xi) := \begin{cases} \operatorname{sinc}(\xi\pi/2), & 0 \leq \xi \leq 1, \\ 0, & \xi > 1, \end{cases} \tag{4.35}$$

where sinc(t) is equal to $\sin(t)/t$ if $t \neq 0$, and 1 otherwise.
Once υ has been chosen, the right-hand side of equation (4.14) has to be calculated to obtain the approximation $V * f$ of f. This has to be done in a discrete setting, and the discretization of (4.14) depends on the way the function g is sampled: different samplings lead to different algorithms. Since we are going to concentrate on iterative methods, we will not proceed in the description of these algorithms and for a detailed treatment of this argument we refer to [69].

4.2 The Radon Transform over straight lines

In the previous section we studied the Radon Transform, which integrates functions on $\mathbb{R}^D$ over hyperplanes. One can also consider an integration over d-planes, with $d = 1, ..., D-1$. In this case, Ξ_D is replaced by the set of unoriented affine d-planes in $\mathbb{R}^D$, the *affine Grassmannian* $\mathbb{G}(d, D)$. For simplicity, here we will consider only the case $d = 1$: the corresponding transform is the so called *X-Ray Transform* or just the *Ray Transform.*
We identify a straight line $\mathbb{L}$ of $\mathbb{G}(1, D)$ with a direction $\boldsymbol{\theta} \in \mathbb{S}^{D-1}$ and a point $\mathbf{s} \in \boldsymbol{\theta}^{\perp}$ as $\{\mathbf{s} + t\boldsymbol{\theta},\ t \in \mathbb{R}\}$ and define the X-Ray Transform $\mathscr{P}$ by

$$\mathscr{P}[f](\boldsymbol{\theta}, \mathbf{s}) = \int_{\mathbb{R}} f(\mathbf{s} + t\boldsymbol{\theta}) dt. \tag{4.36}$$

Similarly to the case $d = D-1$ we have a Projection Slice Theorem as follows. For $f \in \mathcal{L}^1(\mathbb{R}^D)$ and $\mathbf{y} \in \mathbb{R}^D$ let $\mathbb{L} = \mathbb{L}(\boldsymbol{\theta}, \mathbf{s}) \in \mathbb{G}(1, D)$ be a straight line such that $\mathbf{y}$ lies in $\boldsymbol{\theta}^{\perp}$. Then

$$\begin{aligned} \hat{f}(\mathbf{y}) &= (2\pi)^{-D/2} \int_{\mathbb{R}^D} f(\mathbf{x}) e^{-\imath\langle \mathbf{x}, \mathbf{y}\rangle} d\mathbf{x} \\ &= (2\pi)^{-D/2} \int_{\boldsymbol{\theta}^{\perp}} \left(\int_{\mathbb{R}} f(\mathbf{s} + t\boldsymbol{\theta}) e^{-\imath\langle \mathbf{s}+t\boldsymbol{\theta}, \mathbf{y}\rangle} dt \right) d\mu_{\boldsymbol{\theta}^{\perp}}(\mathbf{s}) \\ &= (2\pi)^{-D/2} \int_{\boldsymbol{\theta}^{\perp}} \mathscr{P} f(\boldsymbol{\theta}, \mathbf{s}) e^{-\imath\langle \mathbf{s}, \mathbf{y}\rangle} d\mu_{\boldsymbol{\theta}^{\perp}}(\mathbf{s}). \end{aligned} \tag{4.37}$$

Thus $\mathscr{P}f$ is a function on $T^D := \{(\boldsymbol{\theta}, \mathbf{s})\ \boldsymbol{\theta} \in \mathbb{S}^{D-1},\ \mathbf{s} \in \boldsymbol{\theta}^\perp\}$ and $f \in \mathcal{S}(\mathbb{R}^D)$ implies $\mathscr{P}f \in \mathcal{S}(T^D)$, where

$$\mathcal{S}(T^D) = \left\{ g \in \mathcal{C}^\infty :\ \sup_{(\boldsymbol{\theta},\mathbf{s})} (1 + |\mathbf{s}|)^{k_1} \left| \frac{\partial^{k_2}}{\partial s^{k_2}} g(\boldsymbol{\theta}, \mathbf{s}) \right| < +\infty \right\}. \tag{4.38}$$

On T^D, the convolution and the Partial Fourier Transform are defined by

$$(g * h)(\boldsymbol{\theta}, \mathbf{s}) = \int_{\boldsymbol{\theta}^\perp} g(\boldsymbol{\theta}, \mathbf{s} - \mathbf{t}) h(\boldsymbol{\theta}, \mathbf{t}) d\mu_{\boldsymbol{\theta}^\perp}(\mathbf{t}), \quad \mathbf{s} \in \boldsymbol{\theta}^\perp, \tag{4.39}$$

$$\hat{g}(\boldsymbol{\theta}, \boldsymbol{\xi}) = (2\pi)^{(1-D)/2} \int_{\boldsymbol{\theta}^\perp} e^{-\imath\langle \mathbf{s}, \boldsymbol{\xi}\rangle} g(\boldsymbol{\theta}, \mathbf{s}) d\mu_{\boldsymbol{\theta}^\perp}(\mathbf{s}), \quad \boldsymbol{\xi} \in \boldsymbol{\theta}^\perp. \tag{4.40}$$

Theorem 4.2.1. *For $f, g \in \mathcal{S}(\mathbb{R}^D)$, we have*

$$\mathscr{P}f * \mathscr{P}g = \mathscr{P}(f * g). \tag{4.41}$$

As in the case $d = D - 1$, the convolution on $\mathbb{R}^D$ and on T^D are denoted by the same symbol in the theorem.
The backprojection operator $\mathscr{P}^*$ is now

$$\mathscr{P}^*[g](\mathbf{x}) = \int_{\mathbb{S}^{D-1}} g(\boldsymbol{\theta}, E_{\boldsymbol{\theta}}\mathbf{x}) d\boldsymbol{\theta}, \tag{4.42}$$

where $E_{\boldsymbol{\theta}}$ is the orthogonal projector onto $\boldsymbol{\theta}^\perp$, i.e. $E_{\boldsymbol{\theta}}\mathbf{x} = \mathbf{x} - \langle \mathbf{x}, \boldsymbol{\theta}\rangle\boldsymbol{\theta}$. Again it is the adjoint of $\mathscr{P}$:

$$\int_{\mathbb{S}^{D-1}} \int_{\boldsymbol{\theta}^\perp} g \mathscr{P} f d\boldsymbol{\theta} d\mu_{\boldsymbol{\theta}^\perp}(\mathbf{s}) = \int_{\mathbb{R}^D} f \mathscr{P}^* g d\mathbf{x}. \tag{4.43}$$

There is also an analogous version of Theorem 4.1.1:

Theorem 4.2.2. *Let $f \in \mathcal{S}(\mathbb{R}^D)$ and $g \in \mathcal{S}(T^D)$. Then*

$$\mathscr{P}^*(g * \mathscr{P}f) = (\mathscr{P}^* g) * f. \tag{4.44}$$

From (4.37) and (4.40) follows immediately that for $f \in \mathcal{S}(\mathbb{R}^D)$, $\boldsymbol{\theta} \in \mathbb{S}^{D-1}$ and $\boldsymbol{\xi} \perp \boldsymbol{\theta}$ there holds

$$\widehat{(\mathscr{P}f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) = (2\pi)^{1/2} \hat{f}(\boldsymbol{\xi}). \tag{4.45}$$

This is already enough to state the following uniqueness result in $\mathbb{R}^3$.

Theorem 4.2.3. *Let $\mathbb{S}_0^2$ be a subset of $\mathbb{S}^2$ that meets every equatorial circle of $\mathbb{S}^2$ (Orlov's condition,* 1976*). Then the knowledge of $\mathscr{P}f$ for $\boldsymbol{\theta} \in \mathbb{S}_0^2$ determines f uniquely.*

Proof. For every $\boldsymbol{\xi} \in \mathbb{R}^3$, since $\mathbb{S}_0^2$ satisfies Orlov's condition, there exists an element $\boldsymbol{\theta} \in \mathbb{S}_0^2$ such that $\boldsymbol{\theta} \perp \boldsymbol{\xi}$. Hence $\hat{f}(\boldsymbol{\xi})$ is determined by (4.45). □

An explicit inversion formula on $\mathbb{S}_0^2$ was found by the same Orlov in 1976 (see Natterer [69] for a proof).
In spherical coordinates,

$$\mathbf{x} = \begin{pmatrix} \cos\chi\cos\vartheta \\ \sin\chi\cos\vartheta \\ \sin\vartheta \end{pmatrix}, \quad 0 \le \chi < 2\pi, \quad |\vartheta| \le \frac{\pi}{2}, \tag{4.46}$$

$\mathbb{S}_0^2$ is given by $\vartheta_-(\chi) \le \vartheta \le \vartheta_+(\chi)$, $0 \le \chi < 2\pi$, where $\vartheta_\pm$ are functions such that $-\frac{\pi}{2} < \vartheta_-(\chi) < 0 < \vartheta_+(\chi) < \frac{\pi}{2}$, $0 \le \chi < 2\pi$.
Now, let $l(\mathbf{x}, \mathbf{y})$ be the length of the intersection of $\mathbb{S}_0^2$ with the plane spanned by $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^3$. According to the assumption made on $\vartheta_\pm$, $l(\mathbf{x}, \mathbf{y}) > 0$ if $\mathbf{x}$ and $\mathbf{y}$ are linearly independent.

Theorem 4.2.4 (Orlov's inversion formula). *Let $f \in \mathcal{S}(\mathbb{R}^3)$ and $g(\boldsymbol{\theta}, \mathbf{s}) = \mathscr{P}[f](\boldsymbol{\theta}, \mathbf{s})$ for $\boldsymbol{\theta} \in \mathbb{S}_0^2$ and $\mathbf{s} \perp \boldsymbol{\theta}$. Then*

$$f(\mathbf{x}) = \triangle \int_{\mathbb{S}_0^2} h(\boldsymbol{\theta}, E_{\boldsymbol{\theta}}\mathbf{x}) d\boldsymbol{\theta}, \tag{4.47}$$

where

$$h(\boldsymbol{\theta}, \mathbf{s}) = -\frac{1}{4\pi^2} \int_{\boldsymbol{\theta}^\perp} \frac{g(\boldsymbol{\theta}, \mathbf{s} - \mathbf{t})}{\|\mathbf{t}\| l(\boldsymbol{\theta}, \mathbf{t})} d\mu_{\boldsymbol{\theta}^\perp}(\mathbf{t}) \tag{4.48}$$

and $\triangle$ is the Laplace operator on $\mathbb{R}^3$.

4.2.1 The Cone Beam Transform

We define the *Cone Beam Transform* of a density function $f \in \mathcal{S}(\mathbb{R}^3)$ by

$$\mathscr{D}[f](\mathbf{x}, \boldsymbol{\theta}) = \int_0^{+\infty} f(\mathbf{x} + t\boldsymbol{\theta}) dt, \quad \mathbf{x} \in \mathbb{R}^3, \quad \boldsymbol{\theta} \in \mathbb{S}^2. \tag{4.49}$$

When $\mathbf{x}$ is the location of an X-ray source traveling along a curve $\mathbf{\Gamma}$, the operator $\mathscr{D}$ is usually called the Cone Beam Transform of f along the source curve $\mathbf{\Gamma}$. We want to invert $\mathscr{D}$ in this particular case, which is of interest in the applications.

We start by considering the case where $\mathbf{\Gamma}$ is the unit circle in the x_1-x_2 plane. A point $\mathbf{x}$ on $\mathbf{\Gamma}$ is expressed as $\mathbf{x} = (\cos\phi, \sin\phi, 0)^*$, with $\mathbf{x}^\perp = (-\sin\phi, \cos\phi, 0)^*$. An element $\boldsymbol{\theta} \in \mathbb{S}^2 \setminus \{\mathbf{x}\}$ corresponds to a couple $(\mathrm{y}_2,\mathrm{y}_3)^* \in \mathbb{R}^2$ by taking the intersection of the beam through $\boldsymbol{\theta}$ and $\mathbf{x}$ with the plane spanned by $\mathbf{x}^\perp$ and $\mathsf{e}_3 := (0, 0, 1)^*$ passing through $-\mathbf{x}$ [3]. Thus, if f vanishes outside the unit ball of $\mathbb{R}^3$, we have

$$\mathscr{D}[f](\phi, \mathrm{y}_2, \mathrm{y}_3) = \int_0^{+\infty} f((1-t)\mathbf{x} + t(\mathrm{y}_2\mathbf{x}^\perp + \mathrm{y}_3\mathsf{e}_3))dt. \tag{4.50}$$

We also define the Mellin Transform of a function $h \in \mathbb{R}$ by

$$\mathscr{M}[h](s) = \int_0^{+\infty} t^{s-1}h(t)dt. \tag{4.51}$$

Then in [69] it is shown that performing a Mellin Transform of g and f with respect to y_3 and x_3 respectively and then expanding the results in Fourier series with respect to ϕ one obtains the equations

$$g_l(\mathrm{y}_2, s) = \int_0^{+\infty} f_l\left(\sqrt{(1-t)^2 + t^2\mathrm{y}_2^2}, s\right) e^{-\imath l\alpha(t,\mathrm{y}_2)}dt, \quad l \in \mathbb{Z}, \tag{4.52}$$

where $\alpha(t, \mathrm{y}_2)$ is the argument of the point $(1-t, t\mathrm{y}_2)$ in the x_1-x_2 plane. Unfortunately an explicit solution to this equation does not exist and the entire procedure seems rather expensive from a computational point of view. This is a reason why usually nowadays different paths are used.

An explicit inversion formula was found by Tuy in 1983 (cf. [93]). It applies to paths satisfying the following condition:

Definition 4.2.1 (Tuy's condition). *Let $\mathbf{\Gamma} = \mathbf{\Gamma}(t)$, $t \in [0, 1]$ be a parametric curve on $\mathbb{R}^3$. $\mathbf{\Gamma}$ is said to satisfy Tuy's condition if it intersects each*

[3] in other words, y_2 and y_3 are just the coordinates of the stereographic projection of $\boldsymbol{\theta}$ from the projection point $\mathbf{x}$.

plane hitting $\mathrm{supp}(f)$ *transversally, i.e. if for each* $\mathbf{x} \in \mathrm{supp}(f)$ *and each* $\boldsymbol{\theta} \in \mathbb{S}^2$ *there exists* $t = t(\mathbf{x}, \boldsymbol{\theta}) \in [0, 1]$ *such that*

$$\langle \boldsymbol{\Gamma}(t), \boldsymbol{\theta} \rangle = \langle \mathbf{x}, \boldsymbol{\theta} \rangle, \quad \langle \boldsymbol{\Gamma}'(t), \boldsymbol{\theta} \rangle \neq 0. \tag{4.53}$$

Theorem 4.2.5. *Suppose that the source curve satisfies Tuy's condition. Then*

$$f(\mathbf{x}) = (2\pi)^{-3/2} \imath^{-1} \int_{\mathbb{S}^2} (\langle \boldsymbol{\Gamma}'(t), \boldsymbol{\theta} \rangle)^{-1} \frac{d}{dt} \widehat{(\mathscr{D} f)}(\boldsymbol{\Gamma}(t), \boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{4.54}$$

where $t = t(\mathbf{x}, \boldsymbol{\theta})$ *and where the Fourier Transform is performed only with respect to the first variable.*

Of course, Tuy's condition doesn't hold for the circular path described above since it lies entirely on a plane. In the pursuit of the data sufficiency condition, various scanning trajectories have been proposed, such as circle and line, circle plus arc, double orthogonal circles, dual ellipses and helix (see [50] and the references therein). Of particular interest is the helical trajectory, for which in a series of papers from 2002 to 2004 ([56], [57] and [58]) the Russian mathematician A. Katsevich found an inversion formula strictly related to the Filtered Backprojection algorithm. Although we won't investigate the implementation of these algorithms, we dedicate the following section to the basic concepts developed in those papers since they are considered a breakthrough by many experts in the field.

4.2.2 Katsevich's inversion formula

In the description of Katsevich's inversion formula we follow [97] and [98]. First of all, the source lies on an helical trajectory defined by

$$\boldsymbol{\Gamma}(t) = \left(\mathsf{R}\cos(t), \mathsf{R}\sin(t), \mathsf{P}\frac{t}{2\pi} \right)^*, \quad t \in \mathbb{I}, \tag{4.55}$$

where $\mathsf{R} > 0$ is the radius of the helix, $\mathsf{P} > 0$ is the helical pitch and $\mathbb{I} := [a, b]$, $b > a$. In medical applications, the helical path is obtained by translating the platform where the patient lies through the rotating source-detector gantry.

Thus the pitch of the helix is the displacement of the patient table per source turn. We assume that the support Ω of the function $f \in \mathcal{C}^\infty(\mathbb{R}^3)$ to be reconstructed lies strictly inside the helix, i.e. there exists a cylinder

$$\mathbb{U} := \{\mathbf{x} \in \mathbb{R}^3 \mid \mathrm{x}_1^2 + \mathrm{x}_2^2 < r\}, \quad 0 < r < \mathsf{R}, \tag{4.56}$$

such that $\Omega \subseteq \mathbb{U}$.

To understand the statement of Katsevich's formula we introduce the notions of *π-line* and *Tam-Danielsson window*.

A π-line is any line segment that connects two points on the helix which are separated by less than one helical turn (see Figure 4.1). It can be shown

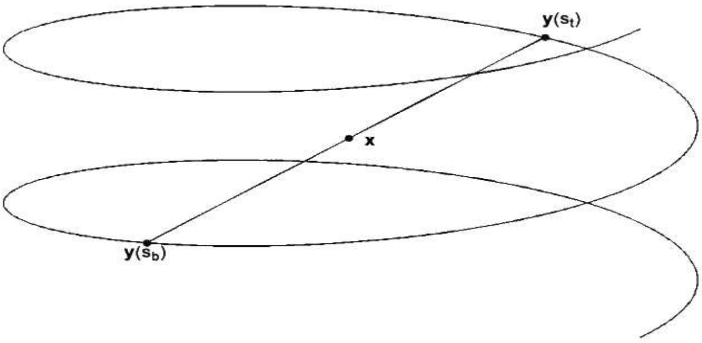


Figure 4.1: The π-line of an helix: in the figure, $\mathbf{y}(s_b)$ and $\mathbf{y}(s_t)$ correspond to $\mathbf{\Gamma}(t_0)$ and $\mathbf{\Gamma}(t_1)$ respectively.

(cf. [11]) that for every point $\mathbf{x}$ inside the helix, there is a unique π-line through $\mathbf{x}$. Let $\mathbb{I}_\pi(\mathbf{x}) = [t_0(\mathbf{x}), t_1(\mathbf{x})]$ be the parametric interval corresponding to the unique π-line passing through $\mathbf{x}$. In particular, $\mathbf{\Gamma}(t_0)$ and $\mathbf{\Gamma}(t_1)$ are the endpoints of the π-line which lie on the helix. By definition, we have $t_1 - t_0 < 2\pi$.

The region on the detector plane bounded above and below by the projections of an helix segment onto the detector plane when viewed from $\mathbf{\Gamma}(t)$ is called the Tam-Danielsson window in the literature (cf. Figure 4.2). Now, consider the ray passing through $\mathbf{\Gamma}(t)$ and $\mathbf{x}$. Let the intersection of this ray with the detector plane be denoted by $\bar{\mathbf{x}}$. Tam et al. in [89] and Danielsson et al. in

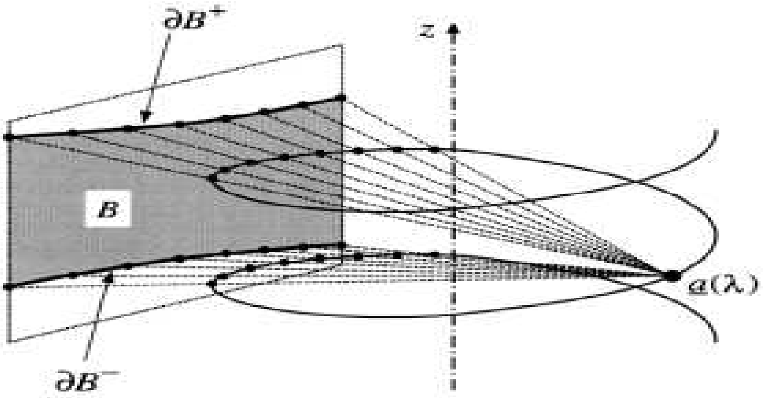


Figure 4.2: The Tam-Danielsson window: $a(\lambda)$ in the figure corresponds to $\mathbf{\Gamma}(t)$ in our notation.

[11] showed that if $\bar{\mathbf{x}}$ lies inside the Tam-Danielsson window for every $t \in \mathbb{I}_\pi$, then $f(\mathbf{x})$ may be reconstructed exactly. We define a κ-plane to be any plane that has three intersections with the helix such that one intersection is halfway between the two others. Denote the κ-plane which intersects the helix at the three points $\mathbf{\Gamma}(t)$, $\mathbf{\Gamma}(t+\psi)$ and $\mathbf{\Gamma}(t+2\psi)$ by $\kappa(t,\psi)$, $\psi \in (-\pi/2, \pi/2)$. The $\kappa(t,\psi)$-plane is spanned by the vectors $\boldsymbol{\nu}_1(t,\psi) = \mathbf{\Gamma}(t+\psi) - \mathbf{\Gamma}(t)$ and $\boldsymbol{\nu}_2(t,\psi) = \mathbf{\Gamma}(t+2\psi) - \mathbf{\Gamma}(t)$ and the unit normal vector to the $\kappa(t,\psi)$-plane is

$$\mathbf{n}(t,\psi) := \operatorname{sgn}(\psi)\frac{\boldsymbol{\nu}_1(t,\psi) \times \boldsymbol{\nu}_2(t,\psi)}{\|\boldsymbol{\nu}_1(t,\psi) \times \boldsymbol{\nu}_2(t,\psi)\|}, \quad \psi \in (-\pi/2, \pi/2), \tag{4.57}$$

where the symbol × stands for the external product in $\mathbb{R}^3$. Katsevich [58] proved that for a given $\mathbf{x}$, the κ-plane through $\mathbf{x}$ with $\psi \in (-\pi/2, \pi/2)$ is uniquely determined if the projection $\bar{\mathbf{x}}$ onto the detector plane lies in the Tam-Danielson window. A κ-line is the line of intersection of the detector plane and a κ-plane, so if $\bar{\mathbf{x}}$ lies in the Tam-Danielson window, there is a unique κ-line. We denote the unit vector from $\mathbf{\Gamma}(t)$ toward $\mathbf{x}$ by

$$\boldsymbol{\beta}(t,\mathbf{x}) = \frac{\mathbf{x} - \mathbf{\Gamma}(t)}{\|\mathbf{x} - \mathbf{\Gamma}(t)\|}. \tag{4.58}$$

For a generic $\boldsymbol{\alpha} \in \mathbb{S}^2$, let $\mathbf{m}(t,\boldsymbol{\alpha})$ be the unit normal vector for the plane $\kappa(t,\psi)$ with the smallest $|\psi|$ value that contains the line of direction $\boldsymbol{\alpha}$ which passes through $\mathbf{\Gamma}(t)$, and put $\mathbf{e}(t,\mathbf{x}) := \boldsymbol{\beta}(t,\mathbf{x}) \times \mathbf{m}(t,\boldsymbol{\beta})$. Then $\boldsymbol{\beta}(t,\mathbf{x})$ and $\mathbf{e}(t,\mathbf{x})$ span the κ-plane that we will want to use for the reconstruction of f in $\mathbf{x}$. Any direction in the plane may be expressed as

$$\boldsymbol{\theta}(t,\mathbf{x},\gamma) = \cos(\gamma)\boldsymbol{\beta}(t,\mathbf{x}) + \sin(\gamma)\mathbf{e}(t,\mathbf{x}), \quad \gamma \in [0,2\pi). \tag{4.59}$$

We can now state Katsevich's result as follows.

Theorem 4.2.6 (Katsevich). *For $f \in \mathcal{C}_0^\infty(\mathbb{U})$,*

$$f(\mathbf{x}) = -\frac{1}{2\pi^2}\int_{\mathbb{I}_\pi(\mathbf{x})} \frac{1}{\|\mathbf{x}-\mathbf{\Gamma}(t)\|} p.v. \int_0^{2\pi} \frac{\partial}{\partial q}\mathscr{D}f(\mathbf{\Gamma}(q),\boldsymbol{\theta}(t,\mathbf{x},\gamma))|_{q=t}\frac{d\gamma dt}{\sin\gamma}, \tag{4.60}$$

where p.v. stands for the principal value integral and where all the objects appearing in the formula are defined as above.

Proof. See [56], [57], [58]. □

For a fixed $\mathbf{x}$, consider the κ-plane with unit normal $\mathbf{m}(t,\boldsymbol{\beta}(t,\mathbf{x}))$. We consider a generic line in this plane with direction $\boldsymbol{\theta}_0(t,\mathbf{x}) = \cos(\gamma_0)\boldsymbol{\beta}(t,\mathbf{x}) + \sin(\gamma_0)\mathbf{e}(t,\mathbf{x}), \gamma_0 \in [0,2\pi)$ and define

$$g'(t,\boldsymbol{\theta}_0(t,\mathbf{x})) := \frac{\partial}{\partial q}\mathscr{D}f(\mathbf{\Gamma}(q),\boldsymbol{\theta}(t,\mathbf{x},\gamma)|_{q=t} \tag{4.61}$$

and

$$g^F(t,\boldsymbol{\theta}_0(t,\mathbf{x})) := p.v. \int_0^{2\pi} \frac{1}{\pi\sin\gamma} g'(t,\cos(\gamma_0-\gamma)\boldsymbol{\beta}(t,\mathbf{x}) - \sin(\gamma_0-\gamma)\mathbf{e}(s,\mathbf{x}))d\gamma. \tag{4.62}$$

Thus Katsevich's formula can be rewritten as

$$f(\mathbf{x}) = -\frac{1}{2\pi}\int_{\mathbb{I}_\pi(\mathbf{x})} \frac{1}{\|\mathbf{x}-\mathbf{\Gamma}(t)\|} g^F(t,\boldsymbol{\beta}(t,\mathbf{x}))dt. \tag{4.63}$$

Therefore, we see that Katsevich's formula may be implemented as a derivative, followed by a 1D convolution, and then a back-projection: for this reason it is usually described as a Filtered Backprojection-type formula. Further details for the implementation of Katsevich's formula can be found, e.g., in [97] and [98].

4.3 Spectral properties of the integral operator

In this section we study the spectral properties of the operator $\mathscr{R}$ defined by (4.4).

We consider the polynomials of degree k orthogonal with respect to the weight function t^l in $[0,1]$ and denote them by $\mathsf{P}_{k,l} = \mathsf{P}_{k,l}(t)$. Similarly to what we have already seen in Chapter 2, the polynomials $\mathsf{P}_{k,l}$ are well defined for every k and $l \in \mathbb{N}_0$ and the orthogonality property means that

$$\int_0^1 t^l \mathsf{P}_{k_1,l}(t)\mathsf{P}_{k_2,l}(t) = \delta_{k_1,k_2}, \tag{4.64}$$

where $\delta_{k_1,k_2} = 1$ if $k_1{=}k_2$ and 0 otherwise. In fact, up to a normalization, they are the *Jacobi polynomials* $\mathsf{G}_\mathsf{k}(l+(D-2)/2, l+(D-2)/2, t)$ (cf. Abramowitz and Stegun [1], formula 22.2.2).

We shall also need the *Gegenbauer* polynomials C^μ_m, which are defined as the orthogonal polynomials on $[-1,1]$ with weight function $(1-t^2)^{\mu-1/2}$, $\mu > -1/2$. Moreover, we recall that a *spherical harmonic* of degree l is the restriction to $\mathbb{S}^{D-1}$ of an harmonic polynomial homogeneous of degree l on $\mathbb{R}^D$ (cf. e.g. [66], [83] and [85]). There exist exactly

$$\mathsf{n}(D,l) := \frac{(2l+D-2)(D+l-3)!}{l!(D-2)!} \tag{4.65}$$

linearly independent spherical harmonics of degree l and spherical harmonics of different degree are orthogonal in $\mathcal{L}^2(\mathbb{S}^{D-1})$.

Now let $\mathsf{Y}_{l,k}$, $k = 1, ..., \mathsf{n}(D,l)$ be an orthonormal basis for the spherical harmonics of degree l. We define, for $i \geq 0$, $0 \leq l \leq i$, $1 \leq k \leq \mathsf{n}(D,l)$,

$$f_{ilk}(\mathbf{x}) = \sqrt{2}\mathsf{P}_{(i-l)/2,l+(D-2)/2}(\|\mathbf{x}\|^2)\|\mathbf{x}\|^l \mathsf{Y}_{l,k}(\mathbf{x}/\|\mathbf{x}\|) \tag{4.66}$$

and

$$g_{ilk}(\boldsymbol{\theta}, s) = c(i)w(s)^{D-1}\mathsf{C}_i^{D/2}(s)\mathsf{Y}_{l,k}(\boldsymbol{\theta}). \tag{4.67}$$

Here $w(s) := (1-s^2)^{1/2}$ and

$$c(i) = \frac{\pi 2^{1-D/2}\Gamma(i+D)}{i!(i+D/2)(\Gamma(D/2))^2}, \tag{4.68}$$

where $\mathsf{\Gamma}$ stands for Euler's Gamma function.

Theorem 4.3.1 (**Davison(1983) and Louis (1984)**). *The functions f_{ilk} and g_{ilk}, $i \geq 0$, $0 \leq l \leq i$, $1 \leq k \leq \mathsf{n}(D,l)$, are complete orthonormal families in the spaces $\mathcal{L}^2(\{\|\mathbf{x}\| < 1\})$ and $\mathcal{L}^2(\Xi^D, w^{1-D})$, respectively. The singular value decomposition of $\mathscr{R}$ as an operator between these spaces is given by*

$$\mathscr{R}f = \sum_{i=0}^{\infty} \lambda_i \sum_{\substack{0\leq l\leq i,\\ l+i \ even}} \sum_{k=1}^{\mathsf{n}(D,l)} \langle f, f_{ilk}\rangle_{\mathcal{L}^2(\{\|\mathbf{x}\|<1\})} g_{ilk}, \tag{4.69}$$

where

$$\lambda_i = \left(\frac{2^D \pi^{D-1}}{(i+1)\cdots(i+D-1)} \right)^{1/2} \tag{4.70}$$

are the singular values of $\mathscr{R}$, each being of multiplicity $\mathsf{n}(D,l)\lfloor\frac{i+2}{2}\rfloor$.

Proof. See [13] and [64]. For a proof in a simplified case with $D = 2$, cf. [4]. □

We observe that in the case $D = 2$ the singular values decay to zero rather slowly, namely $\lambda_i = O(i^{-1/2})$. This is in accordance with the remark we have made in the introductory example of Chapter 1, where we have seen that to compute an inversion of $\mathscr{R}$ the data are differentiated and smoothed again. A more precise statement to explain this can be made by means of the following theorem (see [69] for the details not specified below).

Theorem 4.3.2. *Let Ω be a bounded and sufficiently regular domain in $\mathbb{R}^D$ and let $\alpha \in \mathbb{R}$. Then there are positive constants $c(\alpha,\Omega)$ and $C(\alpha,\Omega)$ such that for all $f \in \mathcal{H}_0^\alpha(\Omega)$*

$$c(\alpha,\Omega)\|f\|_{\mathcal{H}_0^\alpha(\Omega)} \leq \|\mathscr{R}f\|_{\mathcal{H}^{\alpha+(D-1)/2}(\Xi_D)} \leq C(\alpha,\Omega)\|f\|_{\mathcal{H}_0^\alpha(\Omega)}. \tag{4.71}$$

Here, $\mathcal{H}_0^\alpha(\Omega)$ is the closure of $\mathcal{C}_0^\infty(\Omega)$ with respect to the norm of the Sobolev space $\mathcal{H}^\alpha(\mathbb{R}^D)$ and $\mathcal{H}^\beta(\Xi_D)$ is the space of even functions g on the cylinder $\mathfrak{C}^D$ such that

$$\|g\|_{\mathcal{H}^\beta(\mathfrak{C}^D)} := \int_{\mathbb{S}^{D-1}} \int_{\mathbb{R}} (1+\xi^2)^{\beta/2} \hat{g}(\boldsymbol{\theta},\xi)^2 d\xi d\boldsymbol{\theta}, \quad \beta \in \mathbb{R}. \tag{4.72}$$

Thus, roughly speaking, in the general case we can say that $\mathscr{R}f$ is smoother than f by an order of $(D-1)/2$.
Similar results can be found for the operator $\mathscr{P}$.

Theorem 4.3.3 (**Maass (1987)**). *With the functions f_{ilk} defined above and a certain complete orthonormal system g_{ilk} on $\mathcal{L}^2(T^D, w)$, where $w(\boldsymbol{\xi}) = (1-\|\boldsymbol{\xi}\|^2)^{1/2}$, there are positive numbers λ_{il} such that*

$$\mathscr{P}f(\boldsymbol{\theta}, s) = \sum_{i=0}^{\infty} \sum_{\substack{0\le l\le i,\\ l+i \text{ even}}} \lambda_{il} \sum_{k=1}^{\mathsf{n}(D,l)} \langle f, f_{ilk}\rangle g_{ilk}. \tag{4.73}$$

The singular values λ_{il}, each of multiplicity $\mathsf{n}(D,l)$, satisfy

$$\lambda_{il} = O(i^{-1/2}) \tag{4.74}$$

as $i \to +\infty$, uniformly in l.

Proof. See [65]. □

Theorem 4.3.4. *Let Ω be a bounded and sufficiently regular domain in $\mathbb{R}^D$ and let $\alpha \in \mathbb{R}$. Then there are positive constants $c(\alpha, \Omega)$ and $C(\alpha, \Omega)$ such that for all $f \in \mathcal{H}_0^\alpha(\Omega)$*

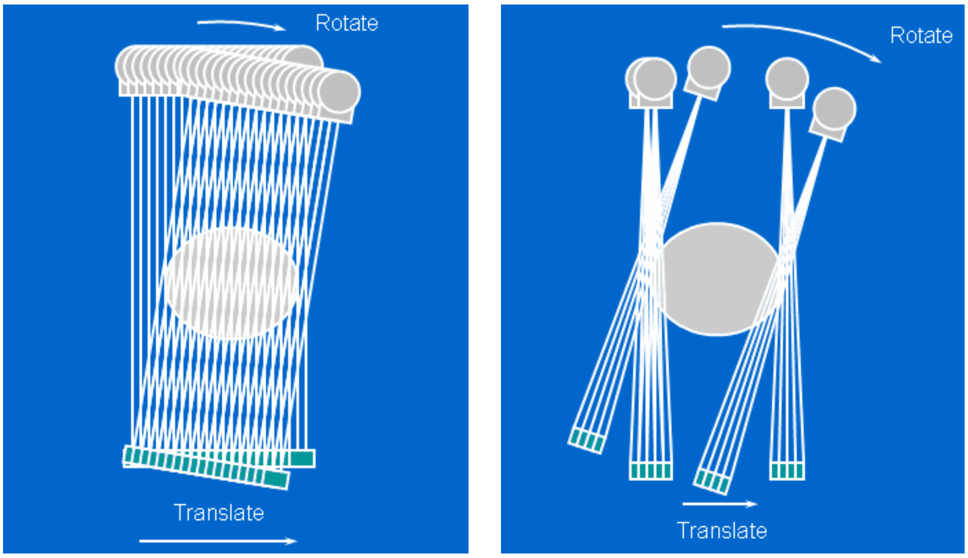
$$c(\alpha,\Omega)\|f\|_{\mathcal{H}_0^\alpha(\Omega)} \le \|\mathscr{P}f\|_{\mathcal{H}^{\alpha+1/2}(T^D)} \le C(\alpha,\Omega)\|f\|_{\mathcal{H}_0^\alpha(\Omega)}, \tag{4.75}$$

where

$$\|g\|_{\mathcal{H}^\beta(T^D)} := \int_{\mathbb{S}^{D-1}} \int_{\boldsymbol{\theta}^\perp} (1+\|\boldsymbol{\xi}\|^2)^{\beta/2} \hat{g}(\boldsymbol{\theta}, \boldsymbol{\xi})^2 d\boldsymbol{\xi} d\boldsymbol{\theta}, \quad \beta \in \mathbb{R}. \tag{4.76}$$

4.4 Parallel, fan beam and helical scanning

In this section we give a very brief survey of the different scanning geometries in computerized tomography. We distinguish between the 2D and the 3D cases.



(a) First generation: parallel, dual motion scanner.

(b) Second generation: narrow fan beam ($\sim 10°$), dual motion scanner.

Figure 4.3: First and second generation scanners.

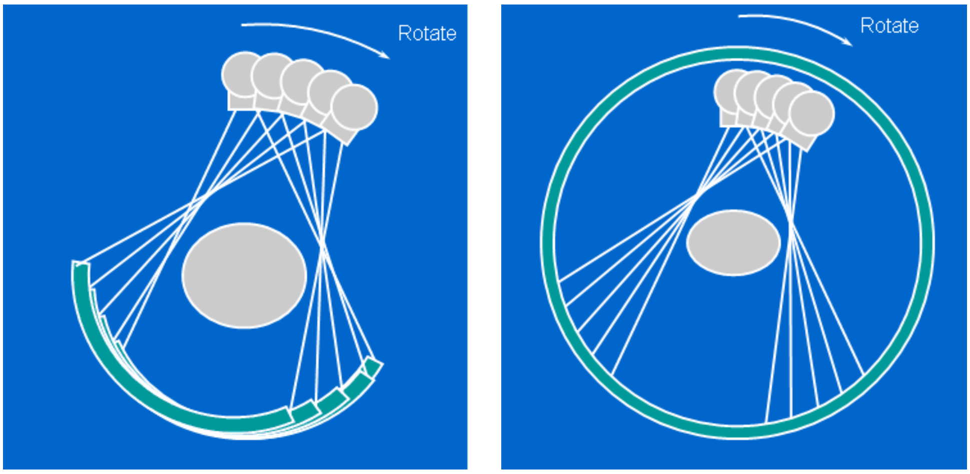
4.4.1 2D scanning geometry

In 2D geometry, only one slice of the object is scanned at a time, the reconstruction is made slice by slice by means of the classical 2D Radon Transform. In parallel scanning, the X-rays are emitted along a two-parameter family of straight lines $\mathbb{L}_{jl}$, $j = 0, ..., \bar{j} - 1$, $l = -\bar{l}, ..., \bar{l}$, where $\mathbb{L}_{jl}$ is the straight line making an angle $\phi_j = j\Delta\phi$ with the x$_2$-axis and having signed distance $s_l = l\Delta s$ from the origin, i.e., $\langle \mathbf{x}, \boldsymbol{\theta}_j \rangle = s_l$, $\boldsymbol{\theta}_j = (\cos\phi_j, \sin\phi_j)^*$. The measured values g_{jl}^{PAR} are simply

$$g_{jl}^{PAR} = \mathscr{R}[f](\boldsymbol{\theta}_j, s_l), \quad j = 0, ..., \bar{j} - 1, \quad l = -\bar{l}, ..., \bar{l}. \tag{4.77}$$

In the first CT scanners (first generation scanners) the source moves along a straight line. The X-ray is fired at each position s_l and the intensity is measured by a detector behind the object which translates simultaneously with the source. Then the same process is repeated with a new angular direction. A first improvement on this invasive and rather slow technique came with the second generation scanners, where more than one detector is used, but the number of detectors is still small (cf. Figure 4.3).

Fan beam scanning geometry is characterized by the use of many detectors. In a third generation scanner, the X-ray source and the detectors are mounted on a common rotating frame (cf. Figure 4.4). During the rotation the detectors are read out in small time intervals which is equivalent to assume that the X-rays are fired from a number of discrete source positions. Let $r\boldsymbol{\theta}(\beta_j)$, $\boldsymbol{\theta}(\beta) := (\cos\beta, \sin\beta)^*$ be the j-th source position and let α_l be the angle the l-th ray in the fan emanating from the source at $r\boldsymbol{\theta}(\beta_j)$ makes with the central ray. Then the measured valued g_{jl} correspond to the 2D Cone Beam Transform:



(a) Third generation: fan beam, rotating detectors.

(b) Fourth generation: fan beam, stationary detectors.

Figure 4.4: Fan beam geometry: third and fourth generation scanners.

$$g_{jl}^{FAN} = \mathscr{D}[f](r\boldsymbol{\theta}(\beta_j), \boldsymbol{\theta}(\alpha_l + \beta_j + \pi)), \quad j = 0, ..., \bar{j} - 1, l = -\bar{l}, ..., \bar{l}. \quad (4.78)$$

In a fourth generation scanner, the detectors are at $r\boldsymbol{\theta}(\beta_j)$, the source is rotating continuously on a circle around the origin (cf. Figure 4.4) and the detectors are read out at discrete time intervals.

4.4.2 3D scanning geometry

In 3D cone beam scanning, the source runs around the object on a curve, together with a 2D detector array. As we have already seen in the section dedicated Katsevich's formula, in the simplest case the curve is a circle and the situation can be modeled by the 3D Cone Beam Transform in the same way as for 2D third generation scanners. When the object is translated continuously in the direction of the axis of symmetry of the fan beam scanner, we obtain the case of 3D helical scanning.

The number of rays actually measured varies between 100 in industrial tomography to 10^6-10^8 in radiological applications [69]. Thus the number of data to be processed is extremely large: for this reason we shall concentrate on iterative regularization methods which are usually faster than other reconstruction techniques.

4.5 Relations between Fourier and singular functions

We have seen that in the case of image deblurring the SVD of the matrix of the underlying linear system is given by means of the DFT, according to (3.36). This allows to study the spectral properties of the problem, even when the size of the system is large.

In the tomography related problems the exact SVD of the matrix of the system is not available. Anyway, it is possible to exploit some a-priori known qualitative information to obtain good numerical results. There are two important pieces of information that are going to be used in the numerical experiments below: the first one is the decay of the singular values of the Radon operator described in Section 4.3; the second one is a general property of ill-posed problems, which is the subject of the current section.

In the paper [39] Hansen, Kilmer and Kjeldsen developed an important insight into the relationship between the SVD and discrete Fourier bases for discrete ill-posed problems arising from the discretization of Fredholm integral equations of the first kind. We reconsider their analysis and relate it to the considerations we have made in Chapter 3 and in particular in Section 3.1 and Section 3.4.

4.5.1 The case of the compact operator

Although we intend to study large scale discrete ill-posed problems, we return first to the continuous setting of Section 1.8, because the basic ideas draw upon certain properties of the underlying continuous problem involving a compact integral operator. As stated already in [39], this material is not intended as a rigorous analysis but rather as a tool to gain some insight into non-asymptotic relations between the Fourier and singular functions of the integral operator.

Consider the Hilbert space $\mathcal{X} = \mathcal{L}^2([-\pi, \pi])$ with the usual inner product $\langle \cdot, \cdot \rangle$ and denote with $\| \cdot \|$ the norm induced by the scalar product. Define

$$K[f](s) := \int_{\mathbb{I}} \varkappa(s,t) f(t) dt, \quad \mathbb{I} = [-\pi, \pi]. \tag{4.79}$$

Assume that the kernel $\varkappa$ is real, (piecewise) $\mathcal{C}^1(\mathbb{I} \times \mathbb{I})$ and non-degenerate. Moreover, assume for simplicity also that $\|\varkappa(\pi, \cdot) - \varkappa(-\pi, \cdot)\| = 0$. Then, as we have seen in Section 1.8, K is a compact operator from $\mathcal{X}$ to itself and there exist a singular system $\{\lambda_j; v_j, u_j\}$ for K such that (1.38), (1.39), (1.40) and (1.41) hold.

Define the infinite matrix B whose rows indexed by $k = -\infty, ..., +\infty$ and columns indexed by $j = 1, ..., +\infty$, with entries

$$B_{k,j} = |\langle u_j, e^{\imath ks}/\sqrt{2\pi} \rangle|.$$

Then the following phenomenon, observed in the discrete setting, can be shown:

The largest entries in the matrix B form a V-shape, with the V lying on the side and the tip of the V located at $k = 0$, $j = 1$.

This means that the function u_j is well represented only by a small number of the $e^{\imath ks}/\sqrt{2\pi}$ for some $|k|$ in a band of contiguous integers depending on j. Therefore, the singular functions are similar to the Fourier series functions in the sense that large singular values (small j) and their corresponding singular functions correspond to low frequencies, and small singular values (larger j) correspond to high frequencies. The important consequence is that it is possible to obtain at least some qualitative information about the spectral properties of an ill-posed problem without calculating the SVD of the operator but only performing a Fourier Transform.

4.5.2 Discrete ill-posed problems

As a matter of fact, the properties of the integral operator described in Section 4.5.1 are observed in practice. As already shown in the paper [39], the Fourier and the SVD coefficients of a discrete ill-posed problem $\mathbf{Ax} = \mathbf{b}$ with perturbed data $\mathbf{b}^\delta$ have a similar behavior if they are ordered in the correct way.

As a consequence of the phenomenon described in Section 4.5.1 and of the properties of the Discrete Fourier Transform, we can reorder the Fourier coefficients as follows:

$$\varphi_i := \begin{cases} (\mathbf{\Phi}^* \mathbf{b}^\delta)_{(i+1)/2} & \text{if } i \text{ is odd,} \\ (\mathbf{\Phi}^* \mathbf{b}^\delta)_{(2m-i+2)/2} & \text{if } i \text{ is even.} \end{cases} \tag{4.80}$$

In Figure 4.5 we compare the SVD and the Fourier coefficients of the test problem phillips(200) with $\varrho = 0.1\%$ and Matlab seed $= 1$. It is evident from the graphic that the Fourier coefficients, reordered according to the formula (4.80), decay very similarly to the SVD coefficients in this example.

In [39], only the case of the 1D ill-posed problems was considered in detail. Here we consider the case of a 2D tomographic test problem, where:

- the matrix $\mathbf{A}$ of the system is the discretization of a 2D integral operator acting on a function of two space variables. For example, in the case of parallel X-rays modeled by the Radon Transform, $\mathscr{R}[f](\boldsymbol{\theta}, s)$ is

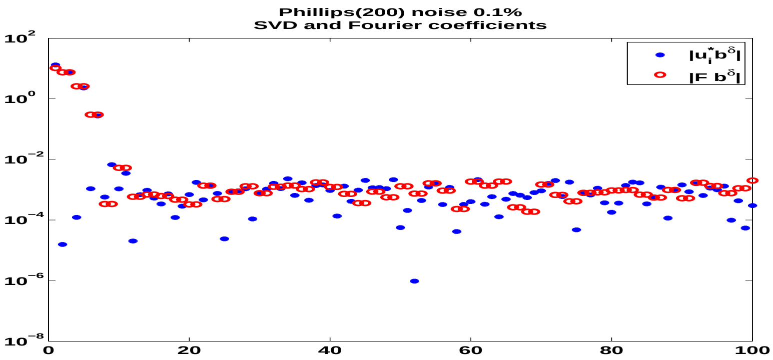


Figure 4.5: SVD(blue full circles) and Fourier (red circles) coefficients of phillips(200), noise 0.1%

simply the integral of the density function f multiplied by the Dirac δ-function supported on the straight line corresponding to $(\boldsymbol{\theta}, s)$.

- The exact data, denoted by the vector $\mathbf{g}$, is the columnwise stacked version of the matrix $\mathbf{G}$ with entries given by a formula of the type (4.77) or (4.78) (the *sinogram*).

- The exact solution $\mathbf{f}$ is the columnwise stacked version of the image $\mathbf{F}$ obtained by computing the density function f on the discretization points of the domain.

To calculate the Fourier coefficients of the perturbed problem $\mathbf{Af} = \mathbf{g}^{\delta}$, $\|\mathbf{g}^{\delta} - \mathbf{g}\| \leq \delta$, we suggest the following strategy:

(i) compute the two-dimensional Discrete Fourier Transform of the (perturbed) sinogram $\mathbf{G}^{\delta}$ corresponding to $\mathbf{g}^{\delta}$.

(ii) Consider the matrix of the Fourier coefficients obtained at the step (i) and reorder its columnwise stacked version as in the 1D case (cf. formula (4.80)).

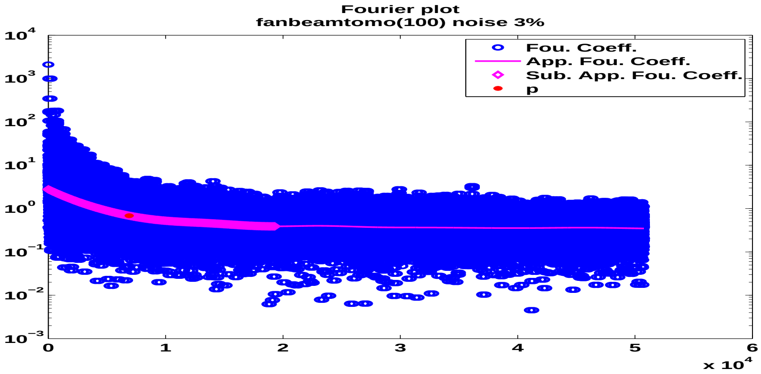


Figure 4.6: The computation of the index p for fanbeamtomo(100), with $\varrho = 3\%$ and Matlab seed $= 0$.

4.6 Numerical experiments

In this section we present some numerical experiments performed on three different test problems of P.C. Hansen's Air Tools (cf. Appendix C.6). In all the considered examples we denote with:

- $\mathbf{A}$, $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{G}^{\delta}$ the matrix of the system, the exact solution, the exact data and the perturbed data respectively;
- $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{g}^{\delta}$ the columnwise stacked versions of the matrices $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{G}^{\delta}$;
- m and N the number of rows and columns of $\mathbf{A}$ respectively (in all the test problems considered $m > N \geq 10000$);
- l_0 and j_0 the number of rows and columns of the sinogram $\mathbf{G}^{\delta}$ (see also Appendix C.6);
- J the number of rows (and columns) of the exact solution $\mathbf{F}$;
- $\boldsymbol{\varphi}$ the vector of the Fourier coefficients of the perturbed system $\mathbf{A}\mathbf{f} = \mathbf{g}^{\delta}$, reordered as described in Section 4.5.2;
- $\tilde{\boldsymbol{\varphi}}$ the approximation of the vector $\boldsymbol{\varphi}$ computed by the routine data_approx with $\lfloor m/5000 \rfloor$ inner knots.

CGNE results for the fanbeamtomo problem					
J	ϱ	Average relative error			
		Discr. Princ.	SR1(np)	SR3	Opt. Sol.
100	1%	0.0750	0.0817(100)	0.0568	0.0427
100	3%	0.1437	0.1347(75)	0.1232	0.1132
100	5%	0.1943	0.1847(50)	0.1715	0.1696
100	7%	0.2273	0.2273(25)	0.2139	0.2127
200	1%	0.1085	0.1049(80)	0.0947	0.0886
200	3%	0.1747	0.1669(60)	0.1693	0.1668
200	5%	0.2226	0.2143(40)	0.2127	0.2123
200	7%	0.2550	0.2550(20)	0.2495	0.2477

Table 4.1: Comparison between different stopping rules of CGNE for the fanbeamtomo test problem.

Using the algorithm cgls, we compute the regularized solutions of CGNE stopped according to the Discrepancy Principle and to SR1 and SR3.
The index p of SR3 is calculated as follows (see Figure 4.6):

(i) determine a subsequence of the approximated Fourier coefficients by choosing the elements $\tilde{\varphi}_1, \tilde{\varphi}_{1+J}, \tilde{\varphi}_{1+2J}, \dots$.

(ii) Discard all the elements after the first relative minimum of this subsequence.

(iii) The index p is defined by $p := 1 + (i + 1)J$ where i is the corner of the discrete curve obtained at the step (ii) determined by the algorithm triangle.

4.6.1 Fanbeamtomo

We consider the function fanbeamtomo, that generates a tomographic test problem with fan beam X-rays. We choose two different values for the dimension J, four different percentages ϱ of noise on the data, and 16 different Matlab seeds, for a total of 128 simulations.
For fixed values of J and ϱ, the average relative errors over all Matlab seeds

CGNE results for the seismictomo problem					
dim	ϱ	Average relative error			
		Discr. Princ.	SR1(np)	SR3	Opt. Sol.
100	1%	0.0985	0.1022(50)	0.0954	0.0929
100	3%	0.1344	0.1367(35)	0.1291	0.1291
100	5%	0.1544	0.1603(25)	0.1549	0.1533
100	7%	0.1713	0.1820(20)	0.1813	0.1713
200	1%	0.1222	0.1260(50)	0.1207	0.1177
200	3%	0.1503	0.1637(35)	0.1484	0.1481
200	5%	0.1699	0.1934(25)	0.1661	0.1660
200	7%	0.1830	0.1960(20)	0.1814	0.1800

Table 4.2: Comparison between different stopping rules of CGNE for the seismictomo test problem.

are summarized in Table 4.1. The results show that SR3 provides an improvement on the Discrepancy Principle, which is more significant when the noise level is small. Moreover, SR1 confirms to be a reliable heuristic stopping rule.

4.6.2 Seismictomo

The function seismictomo creates a two-dimensional seismic tomographic test problem (see Appendix C.6 and [40]).

As for the case of fanbeamtomo, we choose two different values for the dimension J, four different percentages ϱ of noise on the data, and 16 different Matlab seeds. For fixed values of J and ϱ, the average relative errors over all Matlab seeds are summarized in Table 4.2.

The numerical results show again that SR3 improves the results of the Discrepancy Principle. It is interesting to note that in this case the solutions of SR1 are slightly oversmoothed. In fact, the Approximated Residual L-Curve is very similar to the Residual L-Curve, so in this example the approximation helps very little in overcoming the oversmoothing effect of the Residual L-Curve method described in Chapter 3.

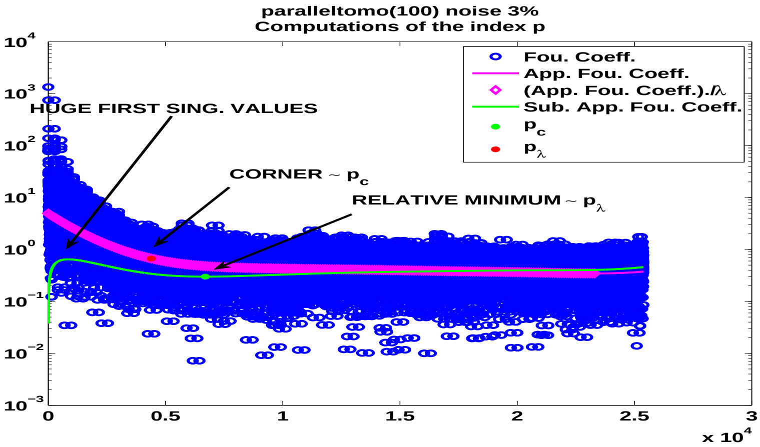


Figure 4.7: The computation of the index p_λ for paralleltomo(100), with $\varrho = 3\%$ and Matlab seed $= 0$.

4.6.3 Paralleltomo

The function paralleltomo generates a 2D tomographic test problem with parallel X-rays.

For this test problem we consider also the following variant of SR3 for computing the index p (cf. Figure 4.7). Using the Matlab function svds, we calculate the largest singular value λ_1 of the matrix $\mathbf{A}$. Assuming that the singular values of $\mathbf{A}$ decay like $O(i^{-1/2})$ we define a vector of *approximated singular values* $\tilde{\lambda}_i := \lambda_1 i^{-1/2}$ for $i \geq 1$ (the approximation is justified by Theorem 4.3.1). Typically, the graphic of the ratios $\tilde{\varphi}_i/\tilde{\lambda}_i$ is similar to that shown in Figure 4.7. Similarly to the 1D examples described in Chapter 3, this graphic has a relative minimum when the Fourier coefficients begin to level off. We denote with p_λ the index corresponding to this minimum and with p_c the index determined as in the test problems fanbeamtomo and seismictomo.

As in the previous cases, we choose two different values for the dimension J, four different percentages ϱ of noise on the data, and 16 different Matlab seeds. For fixed values of J and ϱ, the average relative errors over all Matlab seeds are summarized in Table 4.3.

CGNE results for the paralleltomo problem						
dim	ϱ	Average relative error				
		Discr. Princ.	SR1(np)	SR3$_c$	SR3$_\lambda$	Opt. Sol.
100	1%	0.1247	0.1186(100)	0.1041	0.0868	0.0759
100	3%	0.1992	0.1861(75)	0.1861	0.1759	0.1703
100	5%	0.2378	0.2378(50)	0.2299	0.2272	0.2269
100	7%	0.2724	0.2674(25)	0.2674	0.2670	0.2670
200	1%	0.1568	0.1501(80)	0.1516	0.1507	0.1465
200	3%	0.2140	0.2055(60)	0.2055	0.2051	0.2051
200	5%	0.2593	0.2517(40)	0.2523	0.2517	0.2517
200	7%	0.2962	0.2950(20)	0.2931	0.2931	0.2889

Table 4.3: Comparison between different stopping rules of CGNE for the paralleltomo test problem.

The numerical results show that the qualitative a-priori notion about the spectrum improves the results. As a matter of fact, SR3 with $p = p_\lambda$ provides excellent results, in particular when the noise level is low. The performances of the heuristic stopping rule SR1 are very good here as well.

Chapter 5

Regularization in Banach spaces

So far, we have considered only linear ill-posed problems in a Hilbert space setting. We have seen that this theory is well established since the nineties, so we focused mainly on the applications in the discrete setting.
In the past decade of research, in the area of inverse and ill-posed problems, a great deal of attention has been devoted to the regularization in the Banach space setting. The research on regularization methods in Banach spaces was driven by different mathematical viewpoints. On the one hand, there are various practical applications where models that use Hilbert spaces are not realistic or appropriate. Usually, in such applications sparse solutions[1] of linear and nonlinear operator equations are to be determined, and models working in $\mathcal{L}^p$ spaces, non-Hilbertian Sobolev spaces or continuous function spaces are preferable. On the other hand, mathematical tools and techniques typical of Banach spaces can help to overcome the limitations of Hilbert models. In the monograph [82], a series of different applications ranging from non-destructive testing, such as X-ray diffractometry, via phase retrieval, to an inverse problem in finance are presented. All these applications can be

[1] Sparsity means that the searched-for solution has only a few nonzero coefficients with respect to a specific, given, basis.

modeled by operator equations

$$F(x) = y, \tag{5.1}$$

where the so-called *forward operator* $F : \mathcal{D}(F) \subseteq \mathcal{X} \to \mathcal{Y}$ denotes a continuous linear or nonlinear mapping between Banach spaces $\mathcal{X}$ and $\mathcal{Y}$.

In the opening section of this chapter we shall describe another important example, the problem of identifying coefficients or source terms in partial differential equations (PDEs) from data obtained from the PDE solutions.

Then we will introduce the fundamental notions and tools that are peculiar of the Banach space setting and discuss the problem of regularization in this new framework.

At the end of the chapter we shall focus on the properties of some of the most important regularization methods in Banach spaces for solving nonlinear ill-posed problems, such as the Landweber-type methods and the Iteratively Regularized Gauss-Newton method.

We point out that the aim of this chapter is the introduction of the Newton-Landweber type iteration that will be discussed in the final chapter of this thesis. Thus, methods using only Tikhonov-type penalization terms shall not be considered here.

5.1 A parameter identification problem for an elliptic PDE

The problem of identifying coefficients or source terms in partial differential equations from data obtained from the PDE solution arises in a variety of applications ranging from medical imaging, via nondestructive testing, to material characterization, as well as model calibration.

The following example has been studied repeatedly in the literature (see, e.g. [7], [16], [31], [52], [78] and [82]) to illustrate theoretical conditions and numerically test the convergence of regularization methods.

Consider the identification of the space-dependent coefficient c in the elliptic boundary value problem

$$\begin{cases} -\triangle u + cu = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{5.2}$$

from measurements of u in Ω. Here, $\Omega \subseteq \mathbb{R}^D$, $D \in \mathbb{N}$ is a smooth, bounded domain and $\triangle$ is the Laplace operator on Ω.

The forward operator $F : \mathcal{D}(F) \subseteq \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are to be specified below, and its derivative can be formally written as

$$F(c) = \mathscr{A}(c)^{-1} f, \quad F'(c)h = -\mathscr{A}(c)^{-1}(hF(c)), \tag{5.3}$$

where $\mathscr{A}(c) : \mathcal{H}^2(\Omega) \cap \mathcal{H}^1_0(\Omega) \to \mathcal{L}^2(\Omega)$ is defined by $\mathscr{A}(c)u = -\triangle u + cu$.

In order to preserve ellipticity, a straightforward choice of the domain $\mathcal{D}(F)$ is

$$\mathcal{D}(F) := \{c \in \mathcal{X} \mid c \geq 0 \text{ a.e.}, \|c\|_{\mathcal{X}} \leq \gamma\}, \tag{5.4}$$

for some sufficiently small $\gamma > 0$. For the situation in which the theory requires a nonempty interior of $\mathcal{D}(F)$ in $\mathcal{X}$, the choice

$$\mathcal{D}(F) := \{c \in \mathcal{X} \mid \exists\, \hat{\vartheta} \in \mathcal{L}^\infty(\Omega),\, \hat{\vartheta} \geq 0 \text{ a.e. } : \|c - \hat{\vartheta}\|_{\mathcal{X}} \leq \gamma\}, \tag{5.5}$$

for some sufficiently small $\gamma > 0$, has been devised in [30].

The preimage and image spaces $\mathcal{X}$ and $\mathcal{Y}$ are usually both set to $\mathcal{L}^2(\Omega)$, in order to fit into the Hilbert space theory. However, as observed in [82], the choice $\mathcal{Y} = \mathcal{L}^\infty(\Omega)$ is the natural topology for the measured data and in the situation of impulsive noise the choice $\mathcal{Y} = \mathcal{L}^1(\Omega)$ provides a more robust option than the choice $\mathcal{Y} = \mathcal{L}^2(\Omega)$ (cf. also [6] and [49]). Concerning the preimage space, one often aims at actually reconstructing a uniformly bounded coefficient, or a coefficient that is sparse in some sense, suggesting the use of the $\mathcal{L}^\infty(\Omega)$ or the $\mathcal{L}^1(\Omega)$-norm.

This motivates to study the use of

$$\mathcal{X} = \mathcal{L}^p(\Omega), \quad \mathcal{Y} = \mathcal{L}^r(\Omega),$$

with general exponents $p, r \in [1, +\infty]$ within the context of this example.

Restricting to the choice (5.4) of the domain, it is possible to show the following results (cf. [82]).

Proposition 5.1.1. *Let $p, r, s \in [1, +\infty]$ and denote by $(\mathcal{W}^{2,s} \cap \mathcal{H}_0^1)(\Omega)$ the closure of the space $\mathcal{C}_0^\infty(\Omega)$ with respect to the norm*

$$\|v\|_{\mathcal{W}^{2,s}\cap\mathcal{H}_0^1} := \|\triangle v\|_{\mathcal{L}^s} + \|\nabla v\|_{\mathcal{L}^2},$$

invoking Friedrichs' inequality.
Let also $\mathcal{X} = \mathcal{L}^p(\Omega)$, $\mathcal{Y} = \mathcal{L}^r(\Omega)$.

(*i*) *If*

$$s \geq D/2 \textit{ and } \{s = 1 \textit{ or } s > \max\{1, D/2\} \textit{ or } r < +\infty\}$$

or

$$s < D/2 \textit{ and } r \leq \frac{Ds}{D - 2s},$$

then $(\mathcal{W}^{2,s} \cap \mathcal{H}_0^1)(\Omega) \subseteq \mathcal{L}^r(\Omega)$ and there exists a constant $C_{(a)}^r > 0$ such that

$$\forall v \in (\mathcal{W}^{2,s} \cap \mathcal{H}_0^1)(\Omega): \quad \|v\|_{\mathcal{L}^r} \leq C_{(a)}^r \|v\|_{\mathcal{W}^{2,s}\cap\mathcal{H}_0^1}.$$

(*ii*) *Assume $c \in \mathcal{D}(F)$ with a sufficiently small $\gamma > 0$ and let*

$$1 = s \geq \frac{D}{2} \textit{ or } s > \max\{1, \frac{D}{2}\}. \tag{5.6}$$

Then the operator $\mathscr{A}(c)^{-1} : \mathcal{L}^s(\Omega) \to (\mathcal{W}^{2,s} \cap \mathcal{H}_0^1)(\Omega)$ is well defined and bounded by some constant $C_{(d)}^s$.

(*iii*) *For any $f \in \mathcal{L}^{\max\{1,D/2\}}(\Omega)$, the operator $F : \mathcal{D}(F) \subseteq \mathcal{X} \to \mathcal{Y}$, $F(c) = \mathscr{A}(c)^{-1}f$ is well defined and bounded on $\mathcal{D}(F)$ as in (5.4) with $\gamma > 0$ sufficiently small.*

(*iv*) *For any*

$$p, r \in [1, +\infty], \; f \in \mathcal{L}^1(\Omega), \; D \in \{1, 2\}$$

or

$$p \in (D/2, \infty], \; p \geq 1, \; r \in [1, +\infty], \; f \in \mathcal{L}^{D/2+\epsilon}(\Omega), \; \epsilon > 0$$

and $c \in \mathcal{D}(F)$, the operator $F'(c) : \mathcal{X} \to \mathcal{Y}$, $F'(c) = -\mathscr{A}(c)^{-1}(hF(c))$, is well defined and bounded.

5.2 Basic tools in the Banach space setting

The aim of this section is to introduce the basic tools and fix the classical notations used in the Banach space setting for regularizing ill-posed problems. For details and proofs, cf. [82] and the references therein.

5.2.1 Basic mathematical tools

Definition 5.2.1 (Conjugate exponents, dual spaces and dual pairings). *For $p > 1$ we denote by $p^* > 1$ the conjugate exponent of p, satisfying the equation*

$$\frac{1}{p} + \frac{1}{p^*} = 1. \tag{5.7}$$

We denote by $\mathcal{X}^*$ the dual space of a Banach space $\mathcal{X}$, which is the Banach space of all bounded (continuous) linear functionals $x^* : \mathcal{X} \to \mathbb{R}$, equipped with the norm

$$\|x^*\|_{\mathcal{X}^*} := \sup_{\|x\|=1} |x^*(x)|. \tag{5.8}$$

For $x^* \in \mathcal{X}^*$ and $x \in \mathcal{X}$ we denote by $\langle x^*, x\rangle_{\mathcal{X}^* \times \mathcal{X}}$ and $\langle x, x^*\rangle_{\mathcal{X} \times \mathcal{X}^*}$ the duality pairing (duality product) defined as

$$\langle x^*, x\rangle_{\mathcal{X}^* \times \mathcal{X}} := \langle x, x^*\rangle_{\mathcal{X} \times \mathcal{X}^*} := x^*(x). \tag{5.9}$$

In norms and dual pairings, when clear from the context, we will omit the indices indicating the spaces.

Definition 5.2.2. *Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{X}$ and let $x \in \mathcal{X}$. The sequence x_n is said to converge weakly to x if, for every $x^* \in \mathcal{X}^*$, $\langle x^*, x_n\rangle$ converges to $\langle x^*, x\rangle$.*

As in the Hilbert space case, we shall denote by the symbol $\rightharpoonup$ the weak convergence in $\mathcal{X}$ and by $\to$ the strong convergence in $\mathcal{X}$.

Definition 5.2.3 (Adjoint operator). *Let A be a bounded (continuous) linear operator between two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$. Then the bounded linear operator $A^* : \mathcal{Y}^* \to \mathcal{X}^*$, defined as*

$$\langle A^*y^*, x\rangle_{\mathcal{X}^* \times \mathcal{X}} = \langle y^*, Ax\rangle_{\mathcal{Y}^* \times \mathcal{Y}}, \quad \forall x \in \mathcal{X}, y^* \in \mathcal{Y}^*,$$

is called the adjoint operator of A.

As in the Hilbert space setting, we denote by $\ker(A)$ the null-space of A and by $\mathcal{R}(A)$ the range of A.
We recall three important inequalities that are going to be used later.

Theorem 5.2.1 (Cauchy's inequality). *For $x \in \mathcal{X}$ and $x^* \in \mathcal{X}^*$ we have:*

$$|\langle x^*, x\rangle_{\mathcal{X}^* \times \mathcal{X}}| \leq \|x^*\|_{\mathcal{X}^*}\|x\|_{\mathcal{X}}.$$

Theorem 5.2.2 (Hölder's inequality). *For functions $f \in \mathcal{L}^p(\Omega)$, $g \in \mathcal{L}^{p^*}(\Omega)$, $\Omega \subseteq \mathbb{R}^D$ as in Section 5.1:*

$$\left|\int_\Omega f(\mathbf{x})g(\mathbf{x})d\mathbf{x}\right| \leq \left(\int_\Omega |f(\mathbf{x})|^p d\mathbf{x}\right)^{1/p} \left(\int_\Omega |g(\mathbf{x})|^{p^*} d\mathbf{x}\right)^{1/p^*}.$$

Theorem 5.2.3 (Young's inequality). *Let a and b denote real numbers and $p, p^* > 1$ conjugate exponents. Then*

$$ab \leq \frac{1}{p}|a|^p + \frac{1}{p^*}|b|^{p^*}.$$

5.2.2 Geometry of Banach space norms

Definition 5.2.4 (Subdifferential of a convex functional). *A functional $f : \mathcal{X} \to \mathbb{R} \cup \infty$ is called convex if*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathcal{X},\ \forall t \in [0,1].$$

In this case, an element $x^* \in \mathcal{X}^*$ is a subgradient of f in x if

$$f(y) \geq f(x) + \langle x^*, y - x\rangle, \quad \forall y \in \mathcal{X}.$$

The set $\partial f(x)$ of all subgradients of f in x is called the subdifferential of f in x.

Theorem 5.2.4 (Optimality conditions). *Let $f : \mathcal{X} \to \mathbb{R} \cup \infty$ be a convex functional and let z be such that $f(z) < \infty$. Then*

$$f(z) = \min_{x \in \mathcal{X}} f(x) \Longleftrightarrow 0 \in \partial f(z).$$

This result generalizes the classical optimality condition $f'(z) = 0$, where f' is the Fréchet derivative of f. The subdifferential also is a generalization of the differential in the sense that if f is Gateaux-differentiable, then $\partial f(x) = \{\nabla f(x)\}$.
In general the subdifferential of a function may also be the empty set. However, if the function is Lipschitz-continuous, its sudifferential is not empty.
Among the various properties of the subdifferential of a convex functional, we recall the following one.

Theorem 5.2.5 (**Monotonicity of the subgradient**). *Assume that the convex functional $f : \mathcal{X} \to \mathbb{R} \cup \infty$ is proper, i.e. the set of all $x \in \mathcal{X}$ such that $f(x) < \infty$ is non-empty. Then the following monotonicity property holds:*

$$\langle x^* - y^*, x - y\rangle \geq 0, \quad \forall x^* \in \partial f(x),\ y^* \in \partial f(y),\ x, y \in \mathcal{X}.$$

To understand the geometry of the Banach spaces, it is important to study the properties of the proper convex functional $x \mapsto \frac{1}{p}\|x\|^p$. We start introducing the so-called duality mapping $J_p^{\mathcal{X}}$.

Definition 5.2.5 (**Duality mapping**). *The set valued mapping $J_p^{\mathcal{X}} : \mathcal{X} \to 2^{X^*}$, with $p \geq 1$ defined by*

$$J_p^{\mathcal{X}}(x) := \{x^* \in \mathcal{X}^* \mid \langle x^*, x\rangle = \|x\|\|x^*\|,\ \|x^*\| = \|x\|^{p-1}\}, \tag{5.10}$$

is called the duality mapping of $\mathcal{X}$ with gauge function $t \mapsto t^{p-1}$.

By $j_p^{\mathcal{X}}$ we denote a single-valued selection of $J_p^{\mathcal{X}}$, i.e. $j_p^{\mathcal{X}} : \mathcal{X} \to \mathcal{X}^*$ is a mapping with $j_p^{\mathcal{X}} \in J_p^{\mathcal{X}}$ for all $x \in \mathcal{X}$.
The duality mapping is the subgradient of the functional above:

Theorem 5.2.6 (**Asplund**). *Let $\mathcal{X}$ be a normed space and $p \geq 1$. Then*

$$J_p^{\mathcal{X}} = \partial\left(\frac{1}{p}\|\cdot\|_{\mathcal{X}}^p\right).$$

Example 5.2.1. *For every $p \in (1, +\infty)$, $J_p : L^p(\Omega) \to L^{p^*}(\Omega)$ is given by*

$$J_p^{\mathcal{X}}(f) = |f|^{p-1}\mathrm{sgn}(f), \quad f \in L^p(\Omega). \tag{5.11}$$

From the monotonicity property of the subgradient and the Asplund Theorem, we know that for all $x, z \in \mathcal{X}$ we have

$$\frac{1}{p}\|z\|^p - \frac{1}{p}\|x\|^p - \langle J_p^{\mathcal{X}}(x), z - x\rangle \geq 0,$$

and setting $y = -(z - x)$ yields

$$\frac{1}{p}\|x - y\|^p - \frac{1}{p}\|x\|^p + \langle J_p^{\mathcal{X}}(x), y\rangle \geq 0.$$

We are interested in the upper and lower bounds of the left-hand side of the above inequality in terms of the norm of y.

Definition 5.2.6 (p-convexity and p-smoothness).

- *A Banach space $\mathcal{X}$ is said to be convex of power type p or p-convex if there exists a constant $c_p > 0$ such that*

  $$\frac{1}{p}\|x - y\|^p \geq \frac{1}{p}\|x\|^p - \langle j_p^{\mathcal{X}}(x), y\rangle + \frac{c_p}{p}\|y\|^p$$

  for all $x, y \in \mathcal{X}$ and all $j_p^{\mathcal{X}} \in J_p^{\mathcal{X}}$.

- *A Banach space $\mathcal{X}$ is said to be smooth of power type p or p-smooth if there exists a constant $G_p > 0$ such that*

  $$\frac{1}{p}\|x - y\|^p \leq \frac{1}{p}\|x\|^p - \langle j_p^{\mathcal{X}}(x), y\rangle + \frac{G_p}{p}\|y\|^p$$

  for all $x, y \in \mathcal{X}$ and all $j_p^{\mathcal{X}} \in J_p^{\mathcal{X}}$.

The p-convexity and p-smoothness properties can be regarded as an extension of the polarization identity

$$\frac{1}{2}\|x - y\|^2 = \frac{1}{2}\|x\|^2 - \langle x, y\rangle + \frac{1}{2}\|y\|^2, \tag{5.12}$$

which ensures that Hilbert spaces are 2-convex and 2-smooth.

The p-convexity and p-smoothness are related to other famous properties of convexity and smoothness.

Definition 5.2.7.

- *A Banach space $\mathcal{X}$ is said to be strictly convex if $\|\frac{1}{2}(x+y)\| < 1$ for all x, y of the unit ball of $\mathcal{X}$ satisfying the condition $x \neq y$.*

- *A Banach space $\mathcal{X}$ is said to be uniformly convex if, for the modulus of convexity $\delta_{\mathcal{X}} : [0,2] \to [0,1]$, defined by*
$$\delta_{\mathcal{X}}(\epsilon) := \inf \left\{ 1 - \|\frac{1}{2}(x+y)\| \ : \ \|x\| = \|y\| = 1, \ \ \|x-y\| \geq \epsilon \right\},$$
we have
$$\delta_{\mathcal{X}}(\epsilon) > 0, \quad 0 < \epsilon \leq 2.$$

- *A Banach space $\mathcal{X}$ is said to be smooth if, for every $x \in \mathcal{X}$ with $x \neq 0$, there is a unique $x^* \in \mathcal{X}^*$ such that $\|x^*\| = 1$ and $\langle x^*, x \rangle = \|x\|$.*

- *A Banach space $\mathcal{X}$ is said to be uniformly smooth if, for the modulus of smoothness $\rho_{\mathcal{X}} : [0, +\infty) \to [0, +\infty)$, defined by*
$$\rho_{\mathcal{X}}(\tau) := \frac{1}{2} \sup\{\|x+y\| + \|x-y\| - 2 \mid \|x\| = 1, \|y\| \leq \tau\},$$
there holds:
$$\lim_{\tau \to 0} \frac{\rho_{\mathcal{X}}(\tau)}{\tau} = 0.$$

If a Banach space is p-smooth for some $p > 1$, then $x \mapsto \|x\|^p$ is Fréchet-differentiable, hence Gateaux-differentiable and therefore $J_p^{\mathcal{X}}$ is single-valued. In the famous paper [99], Xu and Roach proved a series of important inequalities, some of which will be very useful in the proofs of the following chapter. Here we recall only the results about uniformly smooth and s-smooth Banach spaces and refer to [82] and to [99] for the analogous results about uniformly convex and s-convex Banach spaces.

Theorem 5.2.7 (**Xu-Roach inequalities I**)**.** *Let $\mathcal{X}$ be uniformly smooth, $1 < p < \infty$, and $j_p^{\mathcal{X}} \in J_p^{\mathcal{X}}$. Then, for all $x, y \in \mathcal{X}$, we have*
$$\begin{aligned} &\|x-y\|^p - \|x\|^p + p\langle j_p^{\mathcal{X}}(x), y \rangle \\ &\leq +pG_p \int_0^1 \frac{(\|x-ty\| \vee \|x\|)^p}{t} \rho_{\mathcal{X}} \left(\frac{t\|y\|}{\|x-ty\| \vee \|x\|} \right) dt, \end{aligned} \tag{5.13}$$

where $a \vee b = \max\{a, b\}$, a, $b \in \mathbb{R}$, *and where* $G_p > 0$ *is a constant that can be written explicitly* (*cf. its expression in* [*82*]).

Theorem 5.2.8 (Xu-Roach inequalities II). *The following statements are equivalent:*

(*i*) $\mathcal{X}$ *is s-smooth.*

(*ii*) *For some* $1 < p < \infty$ *the duality mapping* $J_p^{\mathcal{X}}$ *is single-valued and for all* $x, y \in \mathcal{X}$ *we have*

$$\|J_p^{\mathcal{X}}(x) - J_p^{\mathcal{X}}(y)\| \le C(\|x\| \vee \|y\|)^{p-s}\|x-y\|^{s-1}.$$

(*iii*) *The statement* (*ii*) *holds for all* $p \in (1, \infty)$.

(*iv*) *For some* $1 < p < \infty$, *some* $j_p^{\mathcal{X}} \in J_p^{\mathcal{X}}$ *and for all* x, y, *the inequality* (5.13) *holds. Moreover, the right-hand side of* (5.13) *can be estimated by*

$$C\int_0^1 t^{s-1}\left(\|x-ty\| \vee \|x\|\right)^{p-s}\|y\|^s dt.$$

(*v*) *The statement* (*iv*) *holds for all* $p \in (1, \infty)$ *and all* $j_p^{\mathcal{X}} \in J_p^{\mathcal{X}}$.

The generic constant C *can be chosen independently of* x *and* y.

An important consequence of the Xu-Roach inequalities is the following result.

Corollary 5.2.1. *If* $\mathcal{X}$ *be p-smooth, then for all* $1 < q < p$ *the space* $\mathcal{X}$ *is also q-smooth. If on the other hand* $\mathcal{X}$ *is p-convex, then for all q such that* $p < q < \infty$ *the space* $\mathcal{X}$ *is also q-convex.*
If $\mathcal{X}$ *is s-smooth and* $p > 1$, *then the duality mapping* $J_p^{\mathcal{X}}$ *is single-valued.*

The spaces that are convex or smooth of power type share many interesting properties, summarized in the following theorems.

Theorem 5.2.9. *If* $\mathcal{X}$ *is p-convex, then:*

- *$p \geq 2$;*
- *$\mathcal{X}$ is uniformly convex and the modulus of convexity satisfies $\delta_{\mathcal{X}}(\epsilon) \geq C\epsilon^p$;*
- *$\mathcal{X}$ is strictly convex;*
- *$\mathcal{X}$ is reflexive (i.e. $\mathcal{X}^{**} = \mathcal{X}$).*

If $\mathcal{X}$ is p-smooth, then:

- *$p \leq 2$;*
- *$\mathcal{X}$ is uniformly smooth and the modulus of smoothness satisfies $\rho_{\mathcal{X}}(\tau) \leq C\tau^p$;*
- *$\mathcal{X}$ is smooth;*
- *$\mathcal{X}$ is reflexive.*

Theorem 5.2.10. *There hold:*

- *$\mathcal{X}$ is p-smooth if and only if $\mathcal{X}^*$ is p^* convex.*
- *$\mathcal{X}$ is p-convex if and only if $\mathcal{X}^*$ is p^* smooth.*
- *$\mathcal{X}$ is uniformly convex (respectively uniformly smooth) if and only if $\mathcal{X}^*$ is uniformly smooth (respectively uniformly convex).*
- *$\mathcal{X}$ is uniformly convex if and only if $\mathcal{X}$ is uniformly smooth.*

Theorem 5.2.11. *The duality mappings satisfy the following assertions:*

- *For every $x \in \mathcal{X}$ the set $J_p^{\mathcal{X}}(x)$ is empty and convex.*
- *$\mathcal{X}$ is smooth if and only if the duality mapping $J_p^{\mathcal{X}}$ is single-valued.*
- *If $\mathcal{X}$ is uniformly smooth, then $J_p^{\mathcal{X}}$ is single-valued and uniformly continuous on bounded sets.*

- *If $\mathcal{X}$ is convex of power type and smooth, then $J_p^{\mathcal{X}}$ is single-valued, bijective, and the duality mapping $J_{p^*}^{\mathcal{X}^*}$ is single-valued with*

$$J_{p^*}^{\mathcal{X}^*}(J_p^{\mathcal{X}}(x)) = x.$$

An important consequence of the last statement is that for spaces being smooth of power type and convex of power type the duality mappings on the primal and dual spaces can be used to transport all elements from the primal to the dual space and vice versa. This is crucial to extend the regularization methods defined in the Hilbert space setting to the Banach space setting, as we shall see later.

The smoothness and convexity of power type properties have been studied for some important function spaces. We summarize the results in the theorem below.

Theorem 5.2.12. *Let $\Omega \subseteq \mathbb{R}^D$ be a domain. Then for $1 < r < \infty$ the spaces ℓ^r of infinite real sequences, the Lebesgue spaces $\mathcal{L}^r(\Omega)$ and the Sobolev spaces $\mathcal{W}^{m,r}(\Omega)$, equipped with the usual norms, are*

$$\max\{2, r\}\textit{-convex} \quad \textit{and} \quad \min\{2, r\}\textit{-smooth}.$$

Moreover, it is possible to show that ℓ^1 cannot be p-convex or p-smooth for any p.

5.2.3 The Bregman distance

Due to the geometrical properties of the Banach spaces, it is often more appropriate to exploit the *Bregman distance* instead of functionals like $\|x - y\|_{\mathcal{X}}^p$ or $\|j_p^{\mathcal{X}}(x) - j_p^{\mathcal{X}}(y)\|_{\mathcal{X}^*}^p$ to prove convergence of the algorithms.

Definition 5.2.8. *Let $j_p^{\mathcal{X}}$ be a single-valued selection of the duality mapping $J_p^{\mathcal{X}}$. Then, the functional*

$$D_p(x, y) := \frac{1}{p}\|x\|^p - \frac{1}{p}\|y\|^p - \langle j_p^{\mathcal{X}}(y), x - y\rangle_{\mathcal{X}^* \times \mathcal{X}}, \; x, y \in \mathcal{X} \tag{5.14}$$

is called the Bregman distance (with respect to the functional $\frac{1}{p}\|\cdot\|^p$).

The Bregman distance is not a distance in the classical sense, but has many useful properties.

Theorem 5.2.13. *Let $\mathcal{X}$ be a Banach space and $j_p^{\mathcal{X}}$ be a single-valued selection of the duality mapping $J_p^{\mathcal{X}}$. Then:*

- $D_p(x,y) \geq 0 \ \forall \ x, y \in \mathcal{X}$.
- $D_p(x,y) = 0$ *if and only if* $j_p^{\mathcal{X}}(y) \in J_p^{\mathcal{X}}(x)$.
- *If $\mathcal{X}$ is smooth and uniformly convex, then a sequence $\{x_n\} \subseteq \mathcal{X}$ remains bounded in $\mathcal{X}$ if $D_p(y, x_n)$ is bounded in $\mathbb{R}$. In particular, this is true if $\mathcal{X}$ is convex of power type.*
- *$D_p(x,y)$ is continuous in the first argument. If $\mathcal{X}$ is smooth and uniformly convex, then $J_p^{\mathcal{X}}$ is continuous on bounded subsets and $D_p(x,y)$ is continuous in its second argument. In particular, this is true if $\mathcal{X}$ is convex of power type.*
- *If $\mathcal{X}$ is smooth and uniformly convex, then the following statements are equivalent:*
  - ▶ $\lim_{n\to\infty} \|x_n - x\| = 0$;
  - ▶ $\lim_{n\to\infty} \|x_n\| = \|x\|$ *and* $\lim_{n\to\infty} \langle J_p^{\mathcal{X}}(x_n), x\rangle = \langle J_p^{\mathcal{X}}(x), x\rangle$;
  - ▶ $\lim_{n\to\infty} D_p(x, x_n) = 0$.

  In particular, this is true if $\mathcal{X}$ is convex of power type.
- *The sequence $\{x_n\}$ is a Cauchy sequence in $\mathcal{X}$ if it is bounded and for all $\epsilon > 0$ there is an $N(\epsilon) \in \mathbb{N}$ such that $D_p(x_k, x_l) < \epsilon$ for all $k, l \geq N(\epsilon)$.*
- *$\mathcal{X}$ is p-convex if and only if* $D_p(x,y) \geq c_p \|x - y\|^p$.
- *$\mathcal{X}$ is p-smooth if and only if* $D_p(x,y) \leq G_p \|x - y\|^p$.

The following property of the Bregman distance replaces the classical triangle inequality.

Proposition 5.2.1 (Three-point identity). *Let $j_p^{\mathcal{X}}$ be a single-valued selection of the duality mapping $J_p^{\mathcal{X}}$. Then:*

$$D_p(x,y) = D_p(x,z) + D_p(z,y) + \langle j_p^{\mathcal{X}}(z) - j_p^{\mathcal{X}}(y), x - z \rangle. \tag{5.15}$$

There is a close relationship between the primal Bregman distances and the related Bregman distances in the dual space.

Proposition 5.2.2. *Let $j_p^{\mathcal{X}}$ be a single-valued selection of the duality mapping $J_p^{\mathcal{X}}$. If there exists a single-valued selection $j_{p^*}^{\mathcal{X}^*}$ of $J_{p^*}^{\mathcal{X}^*}$ such that for fixed $y \in \mathcal{X}$ we have $j_{p^*}^{\mathcal{X}^*}(j_p^{\mathcal{X}}(y)) = y$, then*

$$D_p(y,x) = D_{p^*}(j_p^{\mathcal{X}}(x), j_p^{\mathcal{X}}(y)) \tag{5.16}$$

for all $x \in \mathcal{X}$.

5.3 Regularization in Banach spaces

In this section we extend the fundamental concepts of the regularization theory for linear ill-posed operator equations in Hilbert spaces discussed in Chapter 1 to the more general framework of the present chapter.

5.3.1 Minimum norm solutions

In Section 1.6 we have seen Hadamard's definition of ill-posed problems. We recall that an abstract operator equation $F(x) = y$ is well posed (in the sense of Hadamard) if for all right-hand sides y a solution of the equation exists, is unique and the solution depends continuously on the data.

For linear problems in the Hilbert space setting, we have defined the Moore-Penrose generalized inverse, that allows to overcome the problems of existence and uniqueness of the solution by defining the minimum norm (or best-approximate) solution. For general ill-posed problems in Banach spaces, it is also possible to define a minimum norm solution.

In the linear case, the definition is similar to the Hilbert space setting.

Definition 5.3.1 (Minimum norm solution). *Let A be a linear operator between Banach spaces $\mathcal{X}$ and $\mathcal{Y}$ An element $x^\dagger \in \mathcal{X}$ is called a minimum norm solution of the operator equation $Ax = y$ if*

$$Ax^\dagger = y \quad \textit{and} \quad \|x^\dagger\| = \inf\{\|\tilde{x}\| \mid \tilde{x} \in \mathcal{X},\ A\tilde{x} = y\}.$$

The following result gives a characterization of the minimum norm solution (see [82] for the proof).

Proposition 5.3.1. *Let $\mathcal{X}$ be smooth and uniformly convex and let $\mathcal{Y}$ be an arbitrary Banach space. Then, if $y \in \mathcal{R}(A)$, the minimum norm solution of $Ax = y$ is unique. Furthermore, it satisfies the condition $J_p^{\mathcal{X}}(x^\dagger) \in \overline{\mathcal{R}(A^*)}$ for $1 < p < \infty$. If additionally there is some $x \in \mathcal{X}$ such that $J_p^{\mathcal{X}}(x) \in \overline{\mathcal{R}(A^*)}$ and $x - x^\dagger \in \ker(A)$, then $x = x^\dagger$.*

In the nonlinear case, one has to face nonlinear operator equations of the type

$$F(x) = y, \quad x \in \mathcal{D}(F) \subseteq \mathcal{X}, \quad y \in F(\mathcal{D}(F)) \subseteq \mathcal{Y}, \tag{5.17}$$

where $F : \mathcal{D}(F) \subseteq \mathcal{X} \to \mathcal{Y}$ is a nonlinear mapping with domain $\mathcal{D}(F)$ and range $F(\mathcal{D}(F))$.
According to the local character of the solutions in nonlinear equations we have to focus on some neighborhood of a reference element $x_0 \in \mathcal{X}$ which can be interpreted as an initial guess for the solution to be determined. Then, one typically shifts the coordinate system from the zero to x_0 and searches for x_0-minimum norm solutions.

Definition 5.3.2 ($\mathbf{x_0}$-minimum norm solution). *An element $x^\dagger \in \mathcal{D}(F) \subseteq \mathcal{X}$ is called a x_0-minimum norm solution of the operator equation $F(x) = y$ if*

$$F(x^\dagger) = y \quad \textit{and} \quad \|x^\dagger - x_0\| = \inf\{\|\tilde{x} - x_0\| \mid \tilde{x} \in \mathcal{D}(F),\ F(\tilde{x}) = y\}.$$

To ensure that x_0-minimum norm solutions to the nonlinear operator equation (5.17) exist, some assumptions have to be made on the Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, on the domain $\mathcal{D}(F)$ and on the operator F.

Proposition 5.3.2. *Assume the following conditions hold:*

(i) *$\mathcal{X}$ and $\mathcal{Y}$ are infinite dimensional reflexive Banach spaces.*

(ii) *$\mathcal{D}(F) \subseteq \mathcal{X}$ is a convex and closed subset of $\mathcal{X}$.*

(iii) *$F : \mathcal{D}(F) \subseteq \mathcal{X} \to \mathcal{Y}$ is weak-to-weak sequentially continuous, i.e. $x_n \rightharpoonup \bar{x}$ in $\mathcal{X}$ with $x_n \in \mathcal{D}(F)$, $n \in \mathbb{N}$, and $\bar{x} \in \mathcal{D}(F)$ implies $F(x_n) \rightharpoonup F(\bar{x})$ in $\mathcal{Y}$.*

Then the nonlinear operator equation (5.17) *admits an x_0-minimum norm solution.*

Proof. See [82], Proposition 3.14. □

5.3.2 Regularization methods

As usual, we shall assume that the data y of an ill-posed (linear or nonlinear) operator equation are not given exactly, but only elements $y^\delta \in \mathcal{Y}$ satisfying the inequality $\|y - y^\delta\| \le \delta$, with noise level $\delta > 0$ are available.
Consequently, regularization approaches are required for detecting good approximate solutions. Here we give a definition of regularization methods which is analogous to that given in Chapter 1, but a little more general.

Definition 5.3.3. *Let $\sigma_0 \in (0, +\infty]$. For every $\sigma \in (0, \sigma_0)$, let $R_\sigma : \mathcal{Y} \to \mathcal{X}$ be a continuous operator.*
The family $\{R_\sigma\}$ is called a regularization operator for $A^\dagger$ if, for every $y \in \mathcal{D}(A^\dagger)$, there exists a function

$$\alpha : \ (0, +\infty) \times \mathcal{Y} \to (0, \sigma_0),$$

called parameter choice rule for y, that allows to associate to each couple (δ, y^δ) a specific operator $R_{\alpha(\delta,y^\delta)}$ and a regularized solution $x^\delta_{\alpha(\delta,y^\delta)} := R_{\alpha(\delta,y^\delta)} y^\delta$, and such that

$$\lim_{\delta \to 0} \sup_{y^\delta \in \overline{\mathcal{B}_\delta(y)}} \alpha(\delta, y^\delta) = 0. \tag{5.18}$$

If, in addition, for every sequence $\{y_n^\delta\}_{n\in\mathbb{N}} \subseteq \mathcal{Y}$ *with* $\|y_n^\delta - y\| \le \delta_n$ *and* $\delta_n \to 0$ *as* $n \to \infty$ *the regularized solutions* $x^{\delta_n}_{\alpha(y_n^\delta,\delta_n)}$ *converge in a well-defined sense to a well-defined solution* $x^\dagger$ *of* (5.17)*, then* α *is said to be convergent. Convergent regularization methods are defined accordingly, analogously to Definition* 1.9.1*. If the solution of equation* (5.17) *is not unique, convergence to solutions possessing desired properties, e.g.* x_0*-minimum norm solutions, is required.*
In the linear case, similar definitions hold.

The distinction between a-priori, a-posteriori and heuristic parameter choice rules is still valid in this context.
We have seen that in the case of linear operator equations in a Hilbert space setting the construction of regularization methods is based in general on the approximation of the Moore-Penrose approximated inverse of the linear operator A by a σ-dependent family of bounded operators with regularized solutions

$$x_\sigma^\delta = g_\sigma(A^*A)A^*y^\delta, \quad \sigma > 0.$$

However, in the Banach space setting neither $A^\dagger$ nor A^*A is available, since the adjoint operator $A^* : \mathcal{Y}^* \to \mathcal{X}^*$ maps between the dual spaces. In the case of nonlinear operator equations, a comparable phenomenon occurs, because the adjoint operator

$$F'(x^\dagger)^* : \mathcal{Y}^* \to \mathcal{X}^*$$

of a bounded linear derivative operator

$$F'(x^\dagger) : \mathcal{X} \to \mathcal{Y}$$

of F at the solution point $x^\dagger \in \mathcal{D}(F)$ also maps between the dual spaces.
Nevertheless, two large and powerful classes of regularization methods with prominent applications, for example in imaging, were recently promoted: the class of *Tikhonov-type regularization* methods in Banach spaces and the class of *iterative regularization* methods in Banach spaces.
Once again, we shall focus our attention on iterative regularization methods:

since Tikhonov-type regularization methods require the computation of a global minimizer, often the amount of work required for carrying out iterative regularization methods is much smaller than the comparable amount for a Tikhonov-type regularization.

For a detailed coverage of the most recent results about Tikhonov-type regularization methods, see [82].

5.3.3 Source conditions and variational inequalities

We have seen in Chapter 1 that the convergence of the regularized solutions of a regularization method to the minimum norm solution of the ill-posed operator equation $Ax = y$ can be arbitrarily slow. To obtain convergence rates

$$\varepsilon(x_\alpha^\delta, x^\dagger) = O(\varphi(\delta)) \quad \text{as } \delta \to 0 \tag{5.19}$$

for an error measure ε and an index function φ, some smoothness of the solution element $x^\dagger$ with respect to $A : \mathcal{X} \to \mathcal{Y}$ is required. In Chapter 1, we have seen a classical tool for expressing the smoothness of $x^\dagger$, the source conditions. In the Hilbert space setting, this allows to define the concept of (order) optimality of a regularization method.

In the Banach space setting, things are more complicated. The issue of the order optimality of a regularization method is, at least to the author's knowledge, still an open question in the field. The rates depend on the interplay of intrinsic smoothness of $x^\dagger$ and the smoothing properties of the operator A with non-closed range, but very often proving rates is a difficult task and it is not easy to find the correct smoothness assumptions on $x^\dagger$ to obtain convergence rates that, at least in the special case of Hilbert spaces, can be considered optimal.

In this presentation, the extension of the concept of source conditions to the Banach space setting is omitted. We only say that a wide variety of choices has been proposed and analyzed, where either $x^\dagger$ itself or an element $\xi^\dagger$ from the subdifferential of functionals of the form $\frac{1}{p}\|\cdot\|_{\mathcal{X}}^p$ in $x^\dagger$ belongs to the range of a linear operator that interacts with A in an appropriate manner. The source conditions defined in Chapter 1 are only one of these choices.

We will focus our attention on a different strategy for proving convergence rates, the use of *variational inequalities*. As many authors pointed out (cf. e.g. [46], [82] and the references therein), estimating from above a term of the form

$$|\langle J_p^{\mathcal{X}}(x^\dagger), x^\dagger - x\rangle_{\mathcal{X}^* \times \mathcal{X}}|$$

is a very powerful tool for proving convergence rates in regularization. The main advantage of this approach is that this term is contained in the Bregman distance with respect to the functional $\frac{1}{p}\|\cdot\|_{\mathcal{X}}^p$.

In the literature, many similar variational inequalities have been proposed. Essentially, the right-hand side of these inequalities contain a term with the Bregman distance between x and $x^\dagger$ and a term that depends on the operator A or, in the nonlinear case, on the forward operator F, for example:

$$|\langle J_p^{\mathcal{X}}(x^\dagger), x^\dagger - x\rangle_{\mathcal{X}^* \times \mathcal{X}}| \leq \beta_1 D_p(x, x^\dagger) + \beta_2 \|F(x) - F(x^\dagger)\|, \tag{5.20}$$

for constants $0 \leq \beta_1 < 1$ and $\beta_2 \geq 0$. The inequalities must hold for all $x \in \mathcal{M}$, with some set $\mathcal{M}$ which contains all regularized solutions of interest.

We shall not discuss these assumptions here, but we limit ourselves to state a particular variational inequality in each examined case. For a more detailed treatment of the argument, we refer to [82], where some important results that link the variational inequalities with the source conditions can also be found.

5.4 Iterative regularization methods

In this section we consider iterative regularization methods for nonlinear ill-posed operator equations (5.17). We will assume that the noise level δ is known to provide convergence and convergence rates results.

In the following, x_0 is some initial guess. We will assume that a solution to (5.17) exists: according to Proposition 5.3.2, this implies the existence of an x_0-minimum norm solution $x^\dagger$, provided that the assumptions of that proposition are satisfied.

The iterative methods discussed in this section will be either of gradient (Landweber and Iteratively Regularized Landweber) or of Newton-type (Iteratively Regularized Gauss-Newton method).

5.4.1 The Landweber iteration: linear case

Before turning to the nonlinear case, we consider the Landweber iteration for linear ill-posed problems in Banach spaces with noisy data.

In Chapter 1, we have seen that the Landweber iteration for solving linear ill-posed problems in Hilbert spaces can be expressed in the form

$$x_{k+1} = x_k - \omega A^*(Ax_k - y),$$

where $\omega > 0$ is the step size of the method. Here, we shall consider a variable step size $\omega_k > 0$ ($k \in \mathbb{N}$) in the course of the iteration, since an appropriate choice of the step size helps to prove convergence.

The generalization of the Landweber iteration to the Banach space setting requires the help of the duality mappings. As a consequence, the space $\mathcal{X}$ is assumed to be smooth and uniformly convex, whereas $\mathcal{Y}$ can be an arbitrary Banach space. Note that this implies that $j_p^{\mathcal{X}} = J_p^{\mathcal{X}}$ is single-valued, $\mathcal{X}$ is reflexive and $\mathcal{X}^*$ is strictly convex and uniformly smooth.

The Landweber algorithm

Assume that instead of the exact data $y \in \mathcal{R}(A)$ and the exact linear and bounded operator $A : \mathcal{X} \to \mathcal{Y}$, only some approximations $\{y_j\}_j$ in $\mathcal{Y}$ and $\{A_l\}_l$ in the space $\mathscr{L}(\mathcal{X}, \mathcal{Y})$ of linear and bounded operators between $\mathcal{X}$ and $\mathcal{Y}$, are available. Assume also that estimates for the deviations

$$\|y_j - y\| \le \delta_j, \quad \delta_j > \delta_{j+1} > 0, \quad \lim_{j\to\infty} \delta_j = 0, \tag{5.21}$$

and

$$\|A_l - A\| \le \eta_l, \quad \eta_l > \eta_{l+1} > 0, \quad \lim_{l\to\infty} \eta_l = 0, \tag{5.22}$$

are known. Moreover, to properly include the second case (5.22), we need an a-priori estimate for the norm of $x^\dagger$, i.e. there is a constant $\mathsf{R} > 0$ such that

$$\|x^\dagger\| \leq \mathsf{R}. \tag{5.23}$$

Further, set

$$S := \sup_{l \in \mathbb{N}} \|A_l\|. \tag{5.24}$$

(i) Fix $p, r \in (1, \infty)$. Choose constants $C, \tilde{C} \in (0,1)$ and an initial vector x_0 such that

$$j_p^{\mathcal{X}}(x_0) \in \overline{\mathcal{R}(A^*)} \quad \text{and} \quad D_p(x^\dagger, x_0) \leq \frac{1}{p}\|x^\dagger\|^p. \tag{5.25}$$

Set $j_{-1} := 0$ and $l_{-1} := 0$. For $k = 0, 1, 2, ...$ repeat

(ii) If for all $j > j_{k-1}$ and all $l > l_{k-1}$,

$$\|A_l x_k - y_j\| \leq \frac{1}{\tilde{C}}(\delta_j + \eta_l \mathsf{R}), \tag{5.26}$$

stop iterating.
Else, find $j_k > j_{k-1}$ and $l_k > l_{k-1}$ with

$$\delta_{j_k} + \eta_{l_k}\mathsf{R} \leq \tilde{C}\mathsf{R}_k$$

where

$$\mathsf{R}_k := \|A_{l_k} x_k - y_{j_k}\|.$$

Choose ω_k according to:

(a) In case $x_0 = 0$ set

$$\omega_0 := C(1-\tilde{C})^{p-1}\frac{p^{*p-1}}{S^p}\mathsf{R}_0^{p-r}.$$

(b) For all $k \geq 0$ (respectively $k \geq 1$ if $x_0 = 0$), set

$$\gamma_k := \min\left\{\rho_{\mathcal{X}^*}(1), \left(\frac{C(1-\tilde{C})\mathsf{R}_k}{2^{p^*}G_{p^*}S\|x_k\|}\right)\right\},$$

where G_{p^*} is the constant from the Xu-Roach inequalities (cf. Theorem 5.2.7), and choose $\tau_k \in (0,1]$, with

$$\frac{\rho_{\mathcal{X}^*}(\tau_k)}{\tau_k} = \gamma_k$$

and set

$$\omega_k := \frac{\tau_k}{S} \frac{\|x_k\|^{p-1}}{\mathsf{R}_k^{r-1}}. \tag{5.27}$$

Iterate

$$x_{k+1} := J_{p^*}^{\mathcal{X}^*} \left(J_p^{\mathcal{X}}(x_k) - \omega_k A_{l_k}^* j_r^{\mathcal{Y}}(A_{l_k} x_k - y_{j_k}) \right). \tag{5.28}$$

Theorem 5.4.1. *The Landweber algorithm either stops after a finite number of iterations with the minimum norm solution of $Ax = y$ or the sequence of iterates $\{x_k\}$ converges strongly to $x^\dagger$.*

Proof. See [82], Theorem 6.6. □

Let us consider now the case of noisy data y^δ and a perturbed operator A_η, with noise level

$$\|y - y^\delta\| \quad \text{and} \quad \|A - A_\eta\| \leq \eta. \tag{5.29}$$

We apply the Landweber algorithm with $\delta_j = \delta$ and $\eta_l = \eta$ for all $j, l \in \mathbb{N}$ and use the Discrepancy Principle. To that end, condition (5.26) provides us with a stopping rule: we terminate the iteration at $k_D = k_D(\delta)$, where

$$k_D(\delta) := \min\{k \in \mathbb{N} \mid \mathsf{R}_k < \frac{1}{\tilde{C}}(\delta + \eta \mathsf{R})\}. \tag{5.30}$$

The proof of Theorem 5.4.1 shows that, as long as $\mathsf{R}_k \geq \frac{1}{\tilde{C}}(\delta + \eta \mathsf{R})$, x_{k+1} is a better approximation of $x^\dagger$ than x_k. A consequence of this fact and of Theorem 5.4.1 is the stability of this method with respect to the noise.

Corollary 5.4.1. *Together with the Discrepancy Principle* (5.30) *as a stopping rule, the Landweber algorithm is a regularization method for $Ax = y$.*

We observe that since the selection $j_r^{\mathcal{Y}}$ needs not to be continuous, the method is another example of regularization with non-continuous mapping, exactly like the the conjugate gradient type methods.

5.4.2 The Landweber iteration: nonlinear case

Analogous to the Landweber method in Hilbert spaces (cf. [30]), we study a generalization of the Landweber iteration described in Section 5.4.1 to solve nonlinear problems of the form (5.17):

$$\begin{aligned} J_p^{\mathcal{X}}(x_{k+1}^\delta) &= J_p^{\mathcal{X}}(x_k^\delta) - \omega_k A_k^* j_r^{\mathcal{Y}}(F(x_k^\delta) - y^\delta), \\ x_{k+1}^\delta &= J_{p^*}^{\mathcal{X}^*}(J_p^{\mathcal{X}}(x_{k+1}^\delta)), \quad k = 0, 1, ... \end{aligned} \tag{5.31}$$

where we abbreviate $A_k = F'(x_k^\delta)$.

Of course, some assumptions are required on the spaces and on the forward operator F (see the results below). A typical assumption on the forward operator is the so-called *η-condition* (or *tangential cone condition*):

$$\|F(x) - F(\bar{x}) - F'(x)(x - \bar{x})\| \leq \|F(x) - F(\bar{x})\|, \quad \forall x, \bar{x} \in \overline{\mathcal{B}}_\rho^D(x^\dagger) \tag{5.32}$$

for some $0 < \eta < 1$, where $\overline{\mathcal{B}}_\rho^D(x^\dagger) := \{x \in \mathcal{X} \mid D_p(x^\dagger, x) \leq \rho^2, \ \rho > 0\}$.

A key point for proving convergence of the Landweber iteration is showing the monotonicity of the Bregman distances.

Proposition 5.4.1. *Assume that $\mathcal{X}$ is smooth and p-convex, that the initial guess x_0 is sufficiently close to $x^\dagger$, i.e. $x_0 \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$, that F satisfies the tangential cone condition with a sufficiently small η, that F and F' are continuous, and that*

$$\overline{\mathcal{B}}_\rho^D(x^\dagger) \subseteq \mathcal{D}(F). \tag{5.33}$$

Let τ be chosen sufficiently large, so that

$$c(\eta, \tau) := \eta + \frac{1 + \eta}{\tau} < 1. \tag{5.34}$$

Then, with the choice

$$\omega_k := \frac{p^*(1 - c(\eta, \tau))^{p-1}}{G_{p^*}^{p-1}} \frac{\|F(x_k^\delta - y^\delta\|^p - r}{\|A_k\|^p} \geq 0, \tag{5.35}$$

with G_{p^*} being the constant from the Xu-Roach inequalities (cf. Theorem 5.2.7), monotonicity of the Bregman distances

$$D_p(x^\dagger, x_{k+1}^\delta) - D_p(x^\dagger, x_k^\delta) \leq -\frac{p^*(1 - c(\eta, \tau))^p}{p(G_{p^*} p^*)^{p-1}} \frac{\|F(x_k^\delta - y^\delta\|^p}{\|A_k\|^p} \tag{5.36}$$

as well as $x_{k+1}^\delta \in \mathcal{D}(F)$ *holds for all* $k \leq k_D(\delta) - 1$*, with* $k_D(\delta)$ *satisfying the Discrepancy Principle:*

$$k_D(\delta) := \min\{k \in \mathbb{N} \mid \|F(x_k^\delta) - y^\delta\| \leq \tau\delta\}. \tag{5.37}$$

This allows to show the following convergence results for the Landweber iteration. For a proof of this theorem, as well as of Proposition 5.4.1, we refer as usual to [82].

Theorem 5.4.2. *Let the assumptions of Proposition* 5.4.1 *hold, with additionally* $\mathcal{Y}$ *being uniformly smooth and let* $k_D(\delta)$ *be chosen according to the Discrepancy Principle* (5.37)*, with* (5.34)*. Then, according to* (5.31)*, the Landweber iterates* $x_{k_D(\delta)}^\delta$ *converge to a solution of* (5.17) *as* $\delta \to 0$*. If* $\overline{\mathcal{R}(F'(x))} \subseteq \overline{\mathcal{R}(F'(x^\dagger))}$ *for all* $x \in \overline{\mathcal{B}_\rho(x_0)}$ *and* $J_p^{\mathcal{X}}(x_0) \in \overline{\mathcal{R}(F'(x^\dagger))}$*, then* $x_{k_D(\delta)}^\delta$ *converge to* $x^\dagger$ *as* $\delta \to 0$*.*

5.4.3 The Iteratively Regularized Landweber method

In the Hilbert space setting, the proof of convergence rates for the Landweber iteration under source conditions

$$x^\dagger - x_0 \ \in \ \mathcal{R}\left((F'(x^\dagger)^* F'(x^\dagger))^{\frac{\nu}{2}}\right) \tag{5.38}$$

relies on the fact that the iteration errors $x_k^\delta - x^\dagger$ remain in

$$\mathcal{R}\left((F'(x^\dagger)^* F'(x^\dagger))^{\frac{\nu}{2}}\right)$$

and their preimages under $\left((F'(x^\dagger)^* F'(x^\dagger))^{\frac{\nu}{2}}\right)$ form a bounded sequence (cf. Proposition 2.11 in [53]). In [82] is stated that this approach can hardly be carried over to the Banach space setting, unless more restrictive assumptions are made on the structure of the spaces than in the proof of convergence only, even in the case $\nu = 1$.

Therefore, an alternative version of the Landweber iteration is considered, namely the *Iteratively Regularized Landweber* method.

The iterates are now defined by

$$\begin{aligned} J_p^{\mathcal{X}}(x_{k+1}^\delta - x_0) &= (1-\alpha_k) J_p^{\mathcal{X}}(x_k^\delta - x_0) - \omega_k A_k^* j_r^{\mathcal{Y}}(F(x_k^\delta) - y^\delta), \\ x_{k+1}^\delta &= x_0 + J_{p^*}\mathcal{X}^*(J_p^{\mathcal{X}}(x_{k+1}^\delta - x_0)), \quad k = 0, 1, ... \end{aligned} \tag{5.39}$$

An appropriate choice of the sequence $\{\alpha_k\}_{k\in\mathbb{N}} \in [0,1]$, has been shown to be convergent in a Hilbert space setting (with rates under a source condition of the form $\xi^\dagger = (F'(x^\dagger))^* v$, $v \in \mathcal{Y}^*$) in [81].

In place of the Hilbert space condition (5.38) we consider the variational inequality

$$\begin{aligned} &\exists\, \beta > 0 : \forall x \in \mathcal{B}_\rho^D(x^\dagger) \\ &|\langle J_p^{\mathcal{X}}(x^\dagger - x_0), x - x^\dagger \rangle_{\mathcal{X}^*\times\mathcal{X}}| \le \beta D_p^{x_0}(x^\dagger, x)^{\frac{1-\nu}{2}} \|F'(x^\dagger)(x - x^\dagger)\|^\nu \end{aligned} \tag{5.40}$$

where

$$D_p^{x_0}(x^\dagger, x) := D_p(x^\dagger - x_0, x - x_0). \tag{5.41}$$

According to (5.40), due to the presence of additional regularity, the tangential cone condition can be relaxed to a more general condition on the degree of nonlinearity of the operator F:

$$\begin{aligned} &\left\|(F'(x^\dagger + v) - F'(x^\dagger))v\right\| \le K \left\|F'(x^\dagger)v\right\|^{c_1} D_p^{x_0}(x^\dagger, v + x^\dagger)^{c_2}, \\ &v \in X,\; x^\dagger + v \in \mathcal{B}_\rho^D(x^\dagger), \end{aligned} \tag{5.42}$$

with

$$c_1 = 1 \quad \text{or} \quad c_1 + rc_2 > 1 \quad \text{or} \quad (c_1 + rc_2 \ge 1 \;\text{ and }\; K > 0 \;\text{ sufficiently small}) \tag{5.43}$$

and

$$c_1 + c_2 \frac{2\nu}{\nu+1} \ge 1. \tag{5.44}$$

For further details on the degree of nonlinearity conditions, see [82] and the references therein.

The step size $\omega_k > 0$ is chosen such that

$$\omega_k \frac{1 - 3C(c_1)K}{3(1 - C(c_1)K)} \|F(x_k^\delta) - y^\delta\|^r - 2^{p^*+p-2} G_{p^*} \omega_k^{p^*} \|A_k^* j_r^{\mathcal{Y}}(F(x_k^\delta) - y^\delta)\|^{p^*} \ge 0, \tag{5.45}$$

where $C(c_1) = c_1^{c_1}(1-c_1)^{1-c_1}$, c_1 and K as in (5.42). This is possible, e.g. by a choice

$$0 < \omega_k \leq C_\omega \frac{\|F(x_k^\delta) - y^\delta\|^{\frac{p-r}{p-1}}}{\|A_k\|^{p^*}} =: \overline{\omega}_k,$$

with $C_\omega := \frac{2^{2-p^*-p}}{3} \frac{1-3C(c_1)K}{(1-C(c_1)K)G_{p^*}}$.

If $r \geq p$, F and F' are bounded on $\overline{\mathcal{B}}_\rho^D(x^\dagger)$, it is possible to bound ω_k from above and below, i.e. there exist $\underline{\omega}, \overline{\omega} > 0$, independent of k and δ, such that

$$0 < \underline{\omega} \leq \omega_k \leq \overline{\omega}, \tag{5.46}$$

cf. [82].

To prove convergence rates, the following a-priori stopping rule has been proposed:

$$k_*(\delta) := \min\{k \in \mathbb{N} \mid \alpha_k^{\frac{\nu+1}{r(\nu+1)-2\nu}} \leq \tau\delta\}, \tag{5.47}$$

where $\nu < 1$ is the exponent of the variational inequality (5.40).

Theorem 5.4.3. *Assume that $\mathcal{X}$ is smooth and p-convex, that $x_0 \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$, that the variational inequality* (5.40) *holds with $\nu \in (0,1]$ and β sufficiently small, that F satisfies* (5.42)*, with* (5.43) *and* (5.44)*, that F and F' are continuous and uniformly bounded in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$, that $\overline{\mathcal{B}}_\rho^D(x^\dagger) \subseteq \mathcal{D}(F)$ and that*

$$p^* \geq \frac{2\nu}{p(\nu+1) - 2\nu} + 1. \tag{5.48}$$

Let $k_*(\delta)$ be chosen according to (5.47)*, with τ sufficiently large. Moreover, assume that $r \geq p$ and that the sequence $\{\omega_k\}_{k\in\mathbb{N}}$ is chosen such that* (5.46) *holds. Finally, assume that the sequence $\{\alpha_k\}_{k\in\mathbb{N}} \subseteq [0,1]$ is chosen such that*

$$\left(\frac{\alpha_{k+1}}{\alpha_k}\right)^{\frac{2\nu}{p(\nu+1)-2\nu}} + \frac{1}{3}\alpha_k - 1 \geq c\alpha_k \tag{5.49}$$

for some $c \in (0, \frac{1}{3})$ independent of k and $\alpha_{max} = \max_{k\in\mathbb{N}} \alpha_k$ is sufficiently small.

Then, the iterates x_{k+1}^δ remain in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$ for all $k \leq k_*(\delta) - 1$, with $k_*(\delta)$ according to (5.47)*. Moreover, we obtain optimal rates*

$$D_p^{x_0}(x^\dagger, x_{k_*}) = O(\delta^{\frac{2\nu}{\nu+1}}),\ \delta \to 0 \tag{5.50}$$

as well as in the noise free case $\delta = 0$

$$D_p^{x_0}(x^\dagger, x_{k_*}) = O(\alpha_k^{\overline{\frac{2\nu}{r(\nu+1)-2\nu}}}), \ k \to \infty. \tag{5.51}$$

A possible choice of the parameters $\{\alpha_k\}_{k\in\mathbb{N}}$, satisfying (5.49), and smallness of α_{max} is given by

$$\alpha_k = \frac{\alpha_0}{(k+1)^t} \tag{5.52}$$

with $t \in (0,1]$ such that $3t\theta < \alpha_0$ sufficiently small, cf. [82].
We emphasize that in the Banach space setting an analogous of Plato's Theorem 1.11.1 is not available. Consequently, convergence rate results under source conditions or variational inequalities like (5.40) cannot be used to prove (strong) convergence results.

5.4.4 The Iteratively Regularized Gauss-Newton method

Among the iterative methods, the Iteratively Regularized Gauss-Newton (IRGN) method is one of the most important for solving nonlinear ill-posed problems.
In the Banach space setting, the $(n+1)$-th iterate of the IRGN method, denoted by $x_{n+1}^\delta = x_{n+1}^\delta(\alpha_n)$, is a minimizer $x_{n+1}^\delta(\alpha)$ of the Tikhonov functional

$$\|A_n(x-x_n^\delta)+F(x_n^\delta)-y^\delta\|^r+\alpha\|x-x_0\|^p, \quad x \in \mathcal{D}(F), \quad n = 0,1,2,...., \tag{5.53}$$

where $p, r \in (1,\infty)$, $\{\alpha_n\}$ is a sequence of regularization parameters, and $A_n = F'(x_n^\delta)$.
The regularizing properties of the IRGN method are now well understood.
If one of the assumptions

$$F'(x) : \mathcal{X} \to \mathcal{Y} \quad \text{is weakly closed} \ \ \forall x \in \mathcal{D}(F), \ \text{and } \mathcal{Y} \text{ is reflexive}, \tag{5.54}$$

$$\mathcal{D}(F) \ \ \text{is weakly closed} \tag{5.55}$$

holds, then the method is well defined (cf. Lemma 7.9 in [82]). Moreover, assuming variational inequalities similar to (5.40) and the a-priori choice (5.47) for α_n, it is possible to obtain optimal convergence rates, see [82] and the references therein.

Here we concentrate on the a-posteriori choice given by the Discrepancy Principle. More precisely, we have the following two theorems (see, as usual, [82] for the proofs).

Theorem 5.4.4. *Assume that $\mathcal{X}$ is smooth and uniformly convex and that F satisfies the tangential cone condition* (5.32) *with $\overline{\mathcal{B}}_\rho^D(x^\dagger)$ replaced by $\mathcal{D}(F) \cap \overline{\mathcal{B}}_\rho(x_0)$ and with η sufficiently small. Assume also that*

$$(x_n \rightharpoonup x \wedge F(x_n) \to f) \ \Rightarrow \ (x \in \mathcal{D}(F) \wedge F(x) = f) \tag{5.56}$$

or

$$(J_p^{\mathcal{X}}(x_n - x_0) \rightharpoonup x^* \wedge F(x_n) \to f) \ \Rightarrow \ (x := J_{p^*}^{\mathcal{X}^*}(x^*) + x_0 \in \mathcal{D}(F) \wedge F(x) = f) \tag{5.57}$$

for all $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathcal{X}$, holds, as well as (5.54) *or* (5.55)*. Let*

$$\eta < \underline{\sigma} < \overline{\sigma} < 1, \tag{5.58}$$

and let τ be chosen sufficiently large, so that

$$\eta + \frac{1+\eta}{\tau} \le \underline{\sigma} \ \text{and} \ \ \eta < \frac{1-\overline{\sigma}}{2}. \tag{5.59}$$

Choose the regularization parameters α_n such that

$$\underline{\sigma}\|F(x_n^\delta) - y^\delta\| \le \|A_n(x_{n+1}^\delta(\alpha_n) - x_n^\delta) + F(x_n^\delta) - y^\delta\| \le \overline{\sigma}\|F(x_n^\delta) - y^\delta\|, \tag{5.60}$$

if

$$\|A_n(x_0 - x_n^\delta) + F(x_n^\delta) - y^\delta\| \ge \overline{\sigma}\|F(x_n^\delta) - y^\delta\| \tag{5.61}$$

holds. Moreover, assume that

$$\delta < \frac{\|F(x_0) - y^\delta\|}{\tau} \tag{5.62}$$

and stop the iteration at the iterate $n_D = n_D(\delta)$ according to the Discrepancy Principle (5.37)*. Then, for all $n \le n_D(\delta) - 1$, the iterates*

$$x_{n+1}^\delta := \begin{cases} x_{n+1}^\delta = x_{n+1}^\delta(\alpha_n), \ \text{with } \alpha_n \text{ as in } (5.60), & \text{if } (5.61) \text{ holds} \\ x_0, & \text{else} \end{cases} \tag{5.63}$$

are well defined.
Furthermore, there exists a weakly convergent subsequence of

$$\begin{cases} x^\delta_{n_D(\delta)}, & \textit{if } (5.56) \textit{ holds} \\ J_p^{\mathcal{X}}(x^\delta_{n_D(\delta)} - x_0), & \textit{if } (5.56) \textit{ holds} \end{cases} \tag{5.64}$$

and along every weakly convergent subsequence $x_{n_D(\delta)}$ converges strongly to a solution of $F(x) = y$ as $\delta \to 0$. If the solution is unique, then $x_{n_D(\delta)}$ converges strongly to this solution as $\delta \to 0$.

The theorem above provides us with a convergence result. The following theorem gives convergence rates.

Theorem 5.4.5. *Let the assumptions of Theorem* 5.4.4 *be satisfied. Then under the variational inequality*

$$\begin{aligned} &\exists\, \beta > 0 : \forall x \in \mathcal{B}^D_\rho(x^\dagger) \\ &|\langle J_p^{\mathcal{X}}(x^\dagger - x_0), x - x^\dagger\rangle_{\mathcal{X}^* \times \mathcal{X}}| \leq \beta D_p^{x_0}(x, x^\dagger)^{\frac{1-\nu}{2}} \|F'(x^\dagger)(x - x^\dagger)\|^\nu \end{aligned} \tag{5.65}$$

with $0 < \nu < 1$, we obtain optimal convergence rates

$$D_p^{x_0}(x_{n_D}, x^\dagger) = O(\delta^{\frac{2\nu}{\nu+1}}), \ \textit{as } \delta \to 0. \tag{5.66}$$

Chapter 6

A new Iteratively Regularized Newton-Landweber iteration

The final chapter of this thesis is entirely dedicated to a new inner-outer Newton-Iteratively Regularized Landweber iteration for solving nonlinear equations of the type (5.17) in Banach spaces.

The reasons for choosing a Banach space framework have already been explained in the previous chapter. We will see the advantages of working in Banach spaces also in the numerical experiments presented later.

Concerning the method, a combination of inner and outer iterations in a Newton type framework has already been shown to be highly efficient and flexible in the Hilbert space context, see, e.g., [78] and [79].

In the recent paper [49], a Newton-Landweber iteration in Banach spaces has been considered and a weak convergence result for noisy data has been proved. However, neither convergence rates nor strong convergence results have been found. The reason for this is that the convergence rates proof in Hilbert spaces relies on the fact that the iteration errors remain in the range of the adjoint of the linearized forward operator and their preimages under this operator form a bounded sequence. Carrying over this proof to the Banach space setting would require quite restrictive assumptions on the structure of the spaces, though, which we would like to avoid, to work with as general Banach spaces as possible.

Therefore, we study here a combination of the outer Newton loop with an Iteratively Regularized Landweber iteration, which indeed allows to prove convergence rates and strong convergence.

From Section 6.1 to Section 6.5 we will study the inner-outer Newton-Iteratively Regularized Landweber method following [54]. We will see that a strategy for the stopping indices similar to that proposed in [49] leads to a weak convergence result. Moreover, always following [54], we will show a convergence rate result based on an a-priori choice of the outer stopping index.

Section 6.6 is dedicated to some numerical experiments for the elliptic PDE problem presented in Section 5.1.

In Section 6.7 we will consider a different choice of the parameters of the method that allows to show both strong convergence and convergence rates.

6.1 Introduction

In order to formulate and later on analyze the method, we recall some basic notations and concepts. For more details about the concepts appearing below, we refer to Chapter 5.

Consider, for some $p \in (1, \infty)$, the duality mapping $J_p^{\mathcal{X}}(x) := \partial \left\{ \frac{1}{p} \|x\|^p \right\}$ from $\mathcal{X}$ to its dual $\mathcal{X}^*$. To analyze convergence rates we employ the Bregman distance

$$D_p(\tilde{x}, x) = \frac{1}{p}\|\tilde{x}\|^p - \frac{1}{p}\|x\|^p - \langle j_p^{\mathcal{X}}(x), \tilde{x} - x \rangle_{\mathcal{X}^* \times \mathcal{X}}$$

(where $j_p^{\mathcal{X}}(x)$ denotes a single-valued selection of $J_p^{\mathcal{X}}(x)$) or its shifted version

$$D_p^{x_0}(\tilde{x}, x) := D_p(\tilde{x} - x_0, x - x_0)\,.$$

Throughout this paper we will assume that $\mathcal{X}$ is smooth (which implies that the duality mapping is single-valued, cf. Chapter 5) and moreover, that $\mathcal{X}$ is s-convex for some $s \in [p, \infty)$, which implies

$$D_p(x, y) \geq c_{p,s} \|x - y\|^s (\|x\| + \|y\|)^{p-s} \tag{6.1}$$

for some constant $c_{p,s} > 0$, cf. Chapter 5. As a consequence, $\mathcal{X}$ is reflexive and we also have

$$D_{p^*}(x^*, y^*) \le C_{p^*,s^*}\|x^* - y^*\|^{s^*}\left((pD_{p^*}(J_{p^*}^{\mathcal{X}^*}(x^*), 0))^{1-\frac{s^*}{p^*}} + \|x^* - y^*\|^{p^*-s^*}\right), \tag{6.2}$$

for some C_{p^*,s^*}, where s^* denotes the dual index $s^* = \frac{s}{s-1}$. The latter can be concluded from estimate (2.2) in [49], which is the first line in

$$\begin{aligned} D_{p^*}(x^*, y^*) &\le \tilde{C}_{p^*,s^*}\|x^* - y^*\|^{s^*}(\|y^*\|^{p^*-s^*} + \|x^* - y^*\|^{p^*-s^*}) \\ &\le C_{p^*,s^*}\|x^* - y^*\|^{s^*}(\|x^*\|^{p^*-s^*} + \|x^* - y^*\|^{p^*-s^*}) \\ &= C_{p^*,s^*}\|x^* - y^*\|^{s^*}(\|J_{p^*}^{\mathcal{X}^*}(x^*)\|^{(p^*-s^*)(p-1)} + \|x^* - y^*\|^{p^*-s^*}) \\ &= C_{p^*,s^*}\|x^* - y^*\|^{s^*}((pD_{p^*}(J_{p^*}^{\mathcal{X}^*}(x^*), 0))^{(p^*-s^*)\frac{p-1}{p}} + \|x^* - y^*\|^{p^*-s^*}), \end{aligned}$$

where C_{s^*,p^*} is equal to $\tilde{C}_{s^*,p^*}(1+2^{p^*-s^*-1})$ if $p^* - s^* > 1$ and is simply $2\tilde{C}_{s^*,p^*}$ otherwise.

Note that the duality mapping is bijective and $(J_p^{\mathcal{X}})^{-1} = J_{p^*}^{\mathcal{X}^*}$, the latter denoting the (by s-convexity also single-valued) duality mapping on the dual $\mathcal{X}^*$ of $\mathcal{X}$.

We will also make use of the Three-point identity (5.15) and the relation (5.16), which connects elements of the primal space with the corresponding elements of the dual space.

We here consider a combination of the Iteratively Regularized Gauss-Newton method with an Iteratively Regularized Landweber method for approximating the Newton step, using some initial guess x_0 and starting from some x_0^δ (that need not necessarily coincide with x_0)

```
For $n = 0, 1, 2 \dots$ do
    $u_{n,0} = 0$
    $z_{n,0} = x_n^\delta$
    For $k = 0, 1, 2 \dots, k_n - 1$ do
        $u_{n,k+1} = u_{n,k} - \alpha_{n,k} J_p^{\mathcal{X}}(z_{n,k} - x_0)$
                $-\omega_{n,k} A_n^* j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) - b_n)$
        $J_p^{\mathcal{X}}(z_{n,k+1} - x_0) = J_p^{\mathcal{X}}(x_n^\delta - x_0) + u_{n,k+1}$
    $x_{n+1}^\delta = z_{n,k_n}$ ,
                                                                (6.3)
```

where we abbreviate

$$A_n = F'(x_n^\delta)\,, \quad b_n = y^\delta - F(x_n^\delta)\,.$$

For obtaining convergence rates we impose a variational inequality

$$\begin{aligned} &\exists\, \beta > 0 : \ \forall x \in \overline{\mathcal{B}}_\rho^D(x^\dagger) \\ &|\langle J_p^{\mathcal{X}}(x^\dagger - x_0), x - x^\dagger \rangle_{\mathcal{X}^* \times \mathcal{X}}| \le \beta D_p^{x_0}(x^\dagger, x)^{\frac{1}{2} - \nu} \|F(x) - F(x^\dagger)\|^{2\nu}, \end{aligned} \tag{6.4}$$

with $\nu \in (0, \frac{1}{2}]$, corresponding to a source condition in the special case of Hilbert spaces, cf., e.g., [42].

Here

$$\overline{\mathcal{B}}_\rho^D(x^\dagger) = \{x \in \mathcal{X} \mid D_p^{x_0}(x^\dagger, x) \le \rho^2\}$$

with $\rho > 0$ such that $x_0 \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$.

By distinction between the cases $\|x - x_0\| < 2\|x^\dagger - x_0\|$ and $\|x - x_0\| \ge 2\|x^\dagger - x_0\|$ and the second triangle inequality we obtain from (6.1) that

$$\overline{\mathcal{B}}_\rho^D(x^\dagger) \subseteq \overline{\mathcal{B}}_{\bar\rho}(x_0) = \overline{\mathcal{B}}_{\bar\rho}^{\|\cdot\|}(x_0) = \{x \in \mathcal{X} \mid \|x - x_0\| \le \bar\rho\} \tag{6.5}$$

with $\bar\rho = \max\{2\|x^\dagger - x_0\|\,,\ \left(\frac{2^p 3^{s-p} \rho}{c_{p,s}}\right)^{1/p}\}$.

The assumptions on the forward operator besides a condition on the domain

$$\overline{\mathcal{B}}_\rho^D(x^\dagger) \subseteq \mathcal{D}(F) \tag{6.6}$$

include a structural condition on its degree of nonlinearity. For simplicity of exposition we restrict ourselves to the tangential cone condition

$$\|F(\tilde{x}) - F(x) - F'(x)(\tilde{x} - x)\| \leq \eta \, \|F(\tilde{x}) - F(x)\| \, , \; \tilde{x}, x \in \overline{\mathcal{B}}_\rho^D(x^\dagger) \, , \tag{6.7}$$

and mention in passing that most of the results shown here remain valid under a more general condition on the degree of nonlinearity already encountered in Chapter 5 (cf. [42])

$$\begin{aligned} &\left\|(F'(x^\dagger + v) - F'(x^\dagger))v\right\| \leq K \left\|F'(x^\dagger)v\right\|^{c_1} D_p^{x_0}(x^\dagger, v + x^\dagger)^{c_2} \, , \\ &v \in \mathcal{X}, \; x^\dagger + v \in \overline{\mathcal{B}}_\rho^D(x^\dagger) \, , \end{aligned} \tag{6.8}$$

with conditions on c_1, c_2 depending on the smoothness index ν in (6.4). Here F' is not necessarily the Fréchet derivative of F, but just a linearization of F satisfying the Taylor remainder estimate (6.7). Additionally, we assume that F' and F are uniformly bounded on $\overline{\mathcal{B}}_\rho^D(x^\dagger)$.

The method contains a number of parameters that have to be chosen appropriately. At this point we only state that at first the inner iteration will be stopped in the spirit of an inexact Newton method according to

$$\forall 0 \leq k \leq k_n - 1 \; : \quad \mu\|F(x_n^\delta) - y^\delta\| \leq \|A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta\| \tag{6.9}$$

for some $\mu \in (\eta, 1)$.

Since $z_{n,0} = x_n^\delta$ and $\mu < 1$, at least one Landweber step can be carried out in each Newton iteration. By doing several Landweber steps, if allowed by (6.9), we can improve the numerical performance as compared to the Iteratively Regularized Landweber iteration from [55].

Concerning the remaining parameters $\omega_{n,k}$, $\alpha_{n,k}$ and the overall stopping index n_*, we refer to the sections below for details.

Under the condition (6.9), we shall distinguish between the two cases:

(a) (6.4) holds with some $\nu > 0$;

(b) a condition like (6.4) cannot be made use of, since the exponent ν is unknown or (6.4) just fails to hold.

The results we will obtain with the choice (6.9) by distinction between a priori and a posteriori parameter choice are weaker than what one might expect at a first glance. While the Discrepancy Principle for other methods can usually be shown to yield (optimal) convergence rates if (6.4) happens to hold (even if $\nu > 0$ is not available for tuning the method but only for the convergence analysis) we here only obtain weak convergence. On the other hand, the a priori choice will only give convergence with rates if (6.4) holds with $\nu > 0$, otherwise no convergence can be shown. Still there is an improvement over, e.g, the results in [55] and [81] in the sense that there no convergence at all can be shown unless (6.4) holds with $\nu > 0$. Of course from the analysis in [49] it follows that there always exists a choice of $\alpha_{n,k}$ such that weak convergence without rates holds, namely $\alpha_{n,k} = 0$ corresponding to the Newton-Landweber iteration analyzed in [49]. What we are going to show here is that a choice of positive $\alpha_{n,k}$ is admissible, which we expect to provide improved speed of convergence for the inner iteration, as compared to pure Landweber iteration.

Later on, we will analyze a different choice of the stopping indices that leads to strong convergence.

6.2 Error estimates

For any $n \in \mathbb{N}$ we have

$$
\begin{aligned}
&D_p^{x_0}(x^\dagger, z_{n,k+1}) - D_p^{x_0}(x^\dagger, z_{n,k}) \\
&\quad = \; D_p^{x_0}(z_{n,k}, z_{n,k+1}) + \langle \underbrace{J_p^{\mathcal{X}}(z_{n,k+1} - x_0) - J_p^{\mathcal{X}}(z_{n,k} - x_0)}_{=u_{n,k+1}-u_{n,k}}, z_{n,k} - x^\dagger \rangle_{\mathcal{X}^* \times \mathcal{X}} \\
&\quad = \; \underbrace{D_p^{x_0}(z_{n,k}, z_{n,k+1})}_{(I)} \\
&\qquad - \underbrace{\omega_{n,k} \langle j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta), A_n(z_{n,k} - x^\dagger) \rangle_{\mathcal{Y}^* \times \mathcal{Y}}}_{(II)} \\
&\qquad - \underbrace{\alpha_{n,k} \langle J_p^{\mathcal{X}}(x^\dagger - x_0), z_{n,k} - x^\dagger \rangle_{\mathcal{X}^* \times \mathcal{X}}}_{(III)} \\
&\qquad - \underbrace{\alpha_{n,k} \langle J_p^{\mathcal{X}}(z_{n,k} - x_0) - J_p^{\mathcal{X}}(x^\dagger - x_0), z_{n,k} - x^\dagger \rangle_{\mathcal{X}^* \times \mathcal{X}}}_{(IV)} \, . \qquad (6.10)
\end{aligned}
$$

Assuming that $z_{n,k} \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$, we now estimate each of the terms on the right-hand side separately, depending on whether in (6.4) $\nu > 0$ is known (case a) or is not made use of (case b).

By (6.2) and (5.16) we have for the term (I)

$$
\begin{aligned}
D_p^{x_0}(z_{n,k}, z_{n,k+1}) &\le C_{p^*,s^*} \| \underbrace{J_p^{\mathcal{X}}(z_{n,k+1} - x_0) - J_p^{\mathcal{X}}(z_{n,k} - x_0)}_{=u_{n,k+1}-u_{n,k}} \|^{s^*} \\
&\quad \cdot \Big((p\rho^2)^{1-\frac{s^*}{p^*}} + \| J_p^{\mathcal{X}}(z_{n,k+1} - x_0) - J_p^{\mathcal{X}}(z_{n,k} - x_0) \|^{p^*-s^*} \Big) \\
&= C_{p^*,s^*} (p\rho^2)^{1-\frac{s^*}{p^*}} \| \alpha_{n,k} J_p^{\mathcal{X}}(z_{n,k} - x_0) \\
&\quad + \omega_{n,k} A_n^* j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta) \|^{s^*} \\
&\quad + C_{p^*,s^*} \| \alpha_{n,k} J_p^{\mathcal{X}}(z_{n,k} - x_0) + \omega_{n,k} A_n^* j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta) \|^{p^*} \\
&\le 2^{s^*-1} C_{p^*,s^*} (p\rho^2)^{1-\frac{s^*}{p^*}} \alpha_{n,k}^{s^*} \| z_{n,k} - x_0 \|^{(p-1)s^*} \\
&\quad + 2^{s^*-1} C_{p^*,s^*} (p\rho^2)^{1-\frac{s^*}{p^*}} \omega_{n,k}^{s^*} \| A_n^* j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta) \|^{s^*} \Big) \\
&\quad + 2^{p^*-1} C_{p^*,s^*} \alpha_{n,k}^{p^*} \| z_{n,k} - x_0 \|^{(p-1)p^*} \\
&\quad + 2^{p^*-1} C_{p^*,s^*} \omega_{n,k}^{p^*} \| A_n^* j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta) \|^{p^*} \Big) \\
&\le C_{p^*,s^*} \Big((p\rho^2)^{1-\frac{s^*}{p^*}} \bar{\rho}^{(p-1)s^*} 2^{s^*-1} \alpha_{n,k}^{s^*} + \bar{\rho}^p 2^{p^*-1} \alpha_{n,k}^{p^*} \Big) + \varphi(\omega_{n,k} \tilde{\mathrm{t}}_{n,k}), (6.11)
\end{aligned}
$$

where we have used the triangle inequality in $\mathcal{X}^*$ and $\mathcal{X}$, the Young's inequality

$$(a+b)^\lambda \leq 2^{\lambda-1}(a^\lambda + b^\lambda) \quad \text{for } a,b \geq 0\,,\ \lambda \geq 1\,, \tag{6.12}$$

and (6.5), as well as the abbreviations

$$\begin{aligned} \mathrm{d}_{n,k} &= D_p^{x_0}(x^\dagger, z_{n,k})^{1/2}, \\ \mathrm{t}_{n,k} &= \|A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta\|, \\ \tilde{\mathrm{t}}_{n,k} &= \|A_n^* j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta)\|, \\ &\leq \|A_n\| \mathrm{t}_{n,k}^{r-1}\,. \end{aligned} \tag{6.13}$$

Here

$$\varphi(\lambda) = 2^{s^*-1} C_{p^*,s^*} (p\rho^2)^{1-\frac{s^*}{p^*}} \lambda^{s^*} + 2^{p^*-1} C_{p^*,s^*} \lambda^{p^*}, \tag{6.14}$$

which by $p^* \geq s^* > 1$ defines a strictly monotonically increasing and convex function on $\mathbb{R}^+$.

For the term (II) in (6.10) we get, using (6.7) and (1.44),

$$\begin{aligned} &\omega_{n,k}\langle j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta), A_n(z_{n,k} - x^\dagger)\rangle_{\mathcal{Y}^*\times\mathcal{Y}} \\ &= \omega_{n,k}\mathrm{t}_{n,k}^r \\ &\quad + \omega_{n,k}\langle j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta), \\ &\qquad\qquad A_n(x_n^\delta - x^\dagger) - F(x_n^\delta) + y^\delta\rangle_{\mathcal{Y}^*\times\mathcal{Y}} \\ &\geq \omega_{n,k}\mathrm{t}_{n,k}^r - \omega_{n,k}\mathrm{t}_{n,k}^{r-1}(\eta\|F(x_n^\delta) - y^\delta\| + (1+\eta)\delta). \end{aligned} \tag{6.15}$$

Together with (6.9) this yields

$$\begin{aligned} &\omega_{n,k}\langle j_r^{\mathcal{Y}}(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta), A_n(z_{n,k} - x^\dagger)\rangle_{\mathcal{Y}^*\times\mathcal{Y}} \\ &\qquad \geq (1 - \frac{\eta}{\mu})\omega_{n,k}\mathrm{t}_{n,k}^r - (1+\eta)\omega_{n,k}\mathrm{t}_{n,k}^{r-1}\delta \qquad (6.16) \\ &\geq (1 - \frac{\eta}{\mu} - C(\tfrac{r-1}{r})\epsilon)\omega_{n,k}\mathrm{t}_{n,k}^r - C(\tfrac{r-1}{r})\frac{(1+\eta)^r}{\epsilon^{r-1}}\omega_{n,k}\delta^r\,, \qquad (6.17) \end{aligned}$$

where we have used the elementary estimate

$$a^{1-\lambda} b^\lambda \leq C(\lambda)(a+b) \quad \text{for } a,b \geq 0\,,\ \lambda \in (0,1) \tag{6.18}$$

with $C(\lambda) = \lambda^\lambda (1-\lambda)^{1-\lambda}$.

To make use of the variational inequality (6.4) for estimating (III) in case a) with $\nu > 0$, we first of all use (6.7) to conclude

$$\begin{aligned}
&\|F(z_{n,k}) - F(x^\dagger)\| \\
&\quad = \|(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta) \\
&\qquad\quad + (F(z_{n,k}) - F(x_n^\delta) - A_n((z_{n,k} - x_n^\delta) + (y^\delta - y)\| \\
&\quad \leq \mathrm{t}_{n,k} + \eta\|F(z_{n,k}) - F(x_n^\delta)\| + \delta \\
&\quad \leq \mathrm{t}_{n,k} + \eta(\|F(z_{n,k}) - F(x^\dagger)\| + \|F(x_n^\delta) - y^\delta\|) + (1+\eta)\delta\,,
\end{aligned}$$

hence by (6.9)

$$\|F(z_{n,k}) - F(x^\dagger)\| \leq \frac{1}{1-\eta}\left((1+\frac{\eta}{\mu})\mathrm{t}_{n,k} + (1+\eta)\delta\right). \tag{6.19}$$

This together with (6.4) implies

$$\begin{aligned}
&|\alpha_{n,k}\langle J_p^{\mathcal{X}}(x^\dagger - x_0), z_{n,k} - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}| \\
&\quad \leq \frac{\beta}{(1-\eta)^{2\nu}}\alpha_{n,k}\mathrm{d}_{n,k}^{1-2\nu}\left((1+\frac{\eta}{\mu})\mathrm{t}_{n,k} + (1+\eta)\delta\right)^{2\nu} \\
&\quad \leq C(\tfrac{2\nu}{r})\frac{\beta}{(1-\eta)^{2\nu}}\Big\{\omega_{n,k}\left((1+\frac{\eta}{\mu})\mathrm{t}_{n,k} + (1+\eta)\delta\right)^{r} \\
&\qquad\qquad + \Big(\omega_{n,k}^{-\frac{2\nu}{r}}\alpha_{n,k}\mathrm{d}_{n,k}^{1-2\nu}\Big)^{\frac{r}{r-2\nu}}\Big\} \\
&\quad \leq C(\tfrac{2\nu}{r})\frac{\beta}{(1-\eta)^{2\nu}}\Big\{2^{r-1}\omega_{n,k}\left((1+\frac{\eta}{\mu})^r\mathrm{t}_{n,k}^r + (1+\eta)^r\delta^r\right) \\
&\qquad\qquad + C(\tfrac{(1-2\nu)r}{2(r-2\nu)})\Big[\alpha_{n,k}\mathrm{d}_{n,k}^2 + \omega_{n,k}^{-\frac{4\nu}{r(1+2\nu)-4\nu}}\alpha_{n,k}^{\frac{r(1+2\nu)}{r(1+2\nu)-4\nu}}\Big]\Big\}\,,
\end{aligned} \tag{6.20}$$

where we have used (6.18) twice. In the case b) we simply estimate

$$|\alpha_{n,k}\langle J_p^{\mathcal{X}}(x^\dagger - x_0), z_{n,k} - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}| \leq \|x^\dagger - x_0\|^{p-1}\alpha_{n,k}(p\mathrm{d}_{n,k}^2)^{\frac{1}{p}}\,. \tag{6.21}$$

Finally, for the term (IV) we have that

$$\begin{aligned}
&\alpha_{n,k}\langle J_p^{\mathcal{X}}(x^\dagger - x_0) - J_p^{\mathcal{X}}(z_{n,k} - x_0), x^\dagger - z_{n,k}\rangle_{\mathcal{X}^*\times\mathcal{X}} \\
&\quad = \alpha_{n,k}(D_p^{x_0}(x^\dagger, z_{n,k}) + D_p^{x_0}(z_{n,k}, x^\dagger)) \;\geq\; \alpha_{n,k}\mathrm{d}_{n,k}^2\,.
\end{aligned} \tag{6.22}$$

Altogether in case a) we arrive at the estimate

$$\begin{aligned}\mathrm{d}_{n,k+1}^2 \;\leq\;& \Big(1-(1-c_0)\alpha_{n,k}\Big)\mathrm{d}_{n,k}^2 + c_1\alpha_{n,k}^{s^*} + c_2\alpha_{n,k}^{p^*} + c_3\omega_{n,k}^{-\theta}\alpha_{n,k}^{1+\theta} \\ & -(1-c_4)\omega_{n,k}\mathrm{t}_{n,k}^r + C_5\omega_{n,k}\delta^r + \varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k})\,, \qquad (6.23)\end{aligned}$$

where

$$\begin{aligned}
c_0 &= \frac{\beta}{(1-\eta)^{2\nu}}C(\tfrac{2\nu}{r})C(\tfrac{(1-2\nu)r}{2(r-2\nu)}) && (6.24)\\
c_1 &= C_{p^*,s^*}(p\rho^2)^{1-\frac{s^*}{p^*}}\bar{\rho}^{(p-1)s^*}2^{s^*-1} && (6.25)\\
c_2 &= C_{p^*,s^*}\bar{\rho}^p 2^{p^*-1} && (6.26)\\
c_3 &= c_0 && (6.27)\\
c_4 &= \frac{\eta}{\mu} + C(\tfrac{r-1}{r})\epsilon + \frac{\beta}{(1-\eta)^{2\nu}}C(\tfrac{2\nu}{r})2^{r-1}(1+\frac{\eta}{\mu})^r && (6.28)\\
C_5 &= C(\tfrac{r-1}{r})\frac{(1+\eta)^r}{\epsilon^{r-1}} + \frac{\beta}{(1-\eta)^{2\nu}}C(\tfrac{2\nu}{r})2^{r-1}(1+\eta)^r && (6.29)\\
\theta &= \frac{4\nu}{r(1+2\nu)-4\nu}\,, && (6.30)
\end{aligned}$$

(small c denoting constants that can be made small by assuming x_0 to be sufficiently close to $x^\dagger$ and therewith β, η, $\|x_0 - x^\dagger\|$ small).
In case b) we use (6.16), (6.21) instead of (6.17), (6.20), which yields

$$\begin{aligned}\mathrm{d}_{n,k+1}^2 \;\leq\;& \Big(1-\alpha_{n,k}\Big)\mathrm{d}_{n,k}^2 + \tilde{c}_0\alpha_{n,k}\mathrm{d}_{n,k}^{\frac{2}{p}} + c_1\alpha_{n,k}^{s^*} + c_2\alpha_{n,k}^{p^*} \\ & -\left(1-\frac{\eta}{\mu}\right)\omega_{n,k}\mathrm{t}_{n,k}^r + (1+\eta)\omega_{n,k}\mathrm{t}_{n,k}^{r-1}\delta + \varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k}), \quad (6.31)\end{aligned}$$

where

$$\tilde{c}_0 = \|x^\dagger - x_0\|^{p-1}p^{\frac{1}{p}}. \qquad (6.32)$$

6.3 Parameter selection for the method

To obtain convergence and convergence rates we will have to appropriately choose

- the step sizes $\omega_{n,k}$,

- the regularization parameters $\alpha_{n,k}$,
- the stopping indices k_n of the inner iteration,
- the outer stopping index n.

We will now discuss these choices in detail.

In view of estimates (6.23), (6.31) it makes sense to balance the terms $\omega_{n,k}\mathrm{t}_{n,k}^r$ and $\varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k})$. Thus for establishing convergence in case b), we will assume that in each inner Landweber iteration the step size $\omega_{n,k} > 0$ in (6.3) is chosen such that

$$\underline{c}_\omega \le \frac{\varphi\Big(\omega_{n,k}\tilde{\mathrm{t}}_{n,k}\Big)}{\omega_{n,k}\mathrm{t}_{n,k}^r} \le \overline{c}_\omega \tag{6.33}$$

with sufficiently small constants $0 < \underline{c}_\omega < \overline{c}_\omega$. In case a) of (6.4) holding true it will turn out that we do not need the lower bound in (6.33) but have to make sure that $\omega_{n,k}$ stays bounded away from zero and infinity

$$\underline{\omega} \le \omega_{n,k} \le \overline{\omega} \text{ and } \frac{\varphi\Big(\omega_{n,k}\tilde{\mathrm{t}}_{n,k}\Big)}{\omega_{n,k}\mathrm{t}_{n,k}^r} \le \overline{c}_\omega \tag{6.34}$$

for some $\overline{\omega} > \underline{\omega} > 0$.

To see that we can indeed satisfy (6.33), we rewrite it as

$$\underline{c}_\omega \le \varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k})\frac{1}{\omega_{n,k}\mathrm{t}_{n,k}^r} = \psi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k})\frac{\tilde{\mathrm{t}}_{n,k}}{\mathrm{t}_{n,k}^r} \le \overline{c}_\omega\,,$$

with sufficiently small constants $0 < \underline{c}_\omega < \overline{c}_\omega$, and $\psi(\lambda) = \frac{\varphi(\lambda)}{\lambda}$, which by (6.14) and $p^* > 1$, $s^* > 1$ defines a continuous strictly monotonically increasing function on $\mathbb{R}^+$ with $\psi(0) = 0$, $\lim_{\lambda\to+\infty}\psi(\lambda) = +\infty$, so that, after fixing $\mathrm{t}_{n,k}$ and $\tilde{\mathrm{t}}_{n,k}$, $\omega_{n,k}$ is well-defined by (6.33). An easy to implement choice of $\omega_{n,k}$ such that (6.33) holds is given by

$$\omega_{n,k} = \vartheta\min\{\mathrm{t}_{n,k}^{\frac{r}{s^*-1}}\tilde{\mathrm{t}}_{n,k}^{-s}\,,\ \mathrm{t}_{n,k}^{\frac{r}{p^*-1}}\tilde{\mathrm{t}}_{n,k}^{-p}\} \tag{6.35}$$

with ϑ sufficiently small, which is similar to the choice proposed in [49] but avoids estimating the norm of A_n. Indeed, by (6.14), with this choice, the quantity to be estimated from below and above in (6.33) becomes

$$\begin{aligned}\min\{ \quad & 2^{s^*-1}C_{p^*,s^*}(p\rho^2)^{1-\frac{s^*}{p^*}}\vartheta^{s^*-1} + 2^{p^*-1}C_{p^*,s^*}\vartheta^{p^*-1}T^{-(p^*-1)}\,,\\ & 2^{s^*-1}C_{p^*,s^*}(p\rho^2)^{1-\frac{s^*}{p^*}}\vartheta^{s^*-1}T^{(s^*-1)} + 2^{p^*-1}C_{p^*,s^*}\vartheta^{p^*-1}\}\,,\end{aligned}$$

where

$$T = \left[\mathrm{t}_{n,k}^{r(\frac{1}{p^*-1}-\frac{1}{s^*-1})}\tilde{\mathrm{t}}_{n,k}^{s-p}\right]\,.$$

This immediately implies the lower bound with

$$\underline{c}_\omega = \min\{2^{s^*-1}C_{p^*,s^*}(p\rho^2)^{1-\frac{s^*}{p^*}}\vartheta^{s^*-1}\,,\ 2^{p^*-1}C_{p^*,s^*}\vartheta^{p^*-1}\}. \tag{6.36}$$

The upper bound with

$$\overline{c}_\omega = 2^{s^*-1}C_{p^*,s^*}(p\rho^2)^{1-\frac{s^*}{p^*}}\vartheta^{s^*-1} + 2^{p^*-1}C_{p^*,s^*}\vartheta^{p^*-1} \tag{6.37}$$

follows by distinction between the cases $T \geq 1$ and $T < 1$.

For (6.34) in case a) we will need the step sizes $\omega_{n,k}$ to be bounded away from zero and infinity. For this purpose, we will assume that

$$F' \text{ and } F \text{ are uniformly bounded on } \overline{\mathcal{B}}_\rho^D(x^\dagger) \tag{6.38}$$

and that

$$r \geq s \geq p\,, \tag{6.39}$$

i.e., $r^* \leq s^* \leq p^*$. To satisfy (6.34), the choice (6.35) from case b) is modified to

$$\omega_{n,k} = \min\{\vartheta\mathrm{t}_{n,k}^{\frac{r}{s^*-1}}\tilde{\mathrm{t}}_{n,k}^{-s}\,,\ \vartheta\mathrm{t}_{n,k}^{\frac{r}{p^*-1}}\tilde{\mathrm{t}}_{n,k}^{-p}\,,\ \overline{\omega}\} \tag{6.40}$$

which, due to the fact that ψ is strictly monotonically increasing, obviously still satisfies the upper bound in (6.33) with (6.37). Using (6.13) we get

$$\begin{aligned}\mathrm{t}_{n,k}^{\frac{r}{\xi^*-1}}\tilde{\mathrm{t}}_{n,k}^{-\xi} &\geq \left(\sup\nolimits_{x\in\overline{\mathcal{B}}_\rho^D(x^\dagger)}\|F'(x)\|\right)^{-\xi}\mathrm{t}_{n,k}^{-r\xi(\frac{1}{r^*}-\frac{1}{\xi^*})}\\ &\geq \underbrace{\left(\sup_{x\in\overline{\mathcal{B}}_\rho^D(x^\dagger)}\|F'(x)\|\right)^{-\xi}\left((2+3\eta)\sup_{x\in\overline{\mathcal{B}}_\rho^D(x^\dagger)}\|F(x)-F(x^\dagger))\|+\delta\right)^{-r\xi(\frac{1}{r^*}-\frac{1}{\xi^*})}}_{=:S(\xi)}\end{aligned}$$

by (6.7), provided $z_{n,k}, x_n^\delta \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$ (a fact which we will prove by induction below). Hence, we also have that $\omega_{n,k}$ according to (6.40) satisfies $\omega_{n,k} \geq \underline{\omega}$ with $\underline{\omega} \geq \vartheta \min\{S(s)\,, S(p)\}$, thus altogether (6.34).

The regularization parameters $\{\alpha_{n,k}\}_{n\in\mathbb{N}}$ will also be chosen differently depending on whether the smoothness information ν in (6.4) is known or not. In the former case we choose $\{\alpha_{n,k}\}_{n\in\mathbb{N}}$ a priori such that

$$\frac{\mathrm{d}_{0,0}^2}{\alpha_{0,0}^\theta} \leq \bar{\gamma}\,, \tag{6.41}$$

$$\alpha_{n,k}^\theta \leq \frac{\rho^2}{\bar{\gamma}}\,, \tag{6.42}$$

$$\max_{0\leq k\leq k_n} \alpha_{n,k} \to 0 \text{ as } n \to \infty\,, \tag{6.43}$$

$$\begin{aligned}&\left\{\left(\frac{\alpha_{n,k}}{\alpha_{n,k-1}}\right)^\theta - 1 + (1-c_0)\alpha_{n,k-1}\right\}\bar{\gamma}\\ &\qquad \geq c_1\alpha_{n,k-1}^{s^*-\theta} + c_2\alpha_{n,k-1}^{p^*-\theta} + (c_3\underline{\omega}^{-\theta} + \frac{C_5\overline{\omega}}{\tau^r})\alpha_{n,k-1}\end{aligned} \tag{6.44}$$

for some $\overline{\gamma} > 0$ independent of n, k, where $c_0, C_1, c_2, c_3, C_5, \theta, \underline{\omega}, \overline{\omega}$ are as in (6.24)–(6.30), (6.34), and $\nu \in (0, \frac{1}{2}]$ is the exponent in the variational inequality (6.4). Moreover, when using (6.4) with $\nu > 0$, we are going to impose the following condition on the exponents p, r, s

$$1 + \theta = \frac{r(1+2\nu)}{r(1+2\nu) - 4\nu} \leq s^* \leq p^*\,. \tag{6.45}$$

Well definedness in case $k = 0$ is guaranteed by setting $\alpha_{n,-1} = \alpha_{n-1,k_{n-1}}$, $\omega_{n,-1} = \omega_{n-1,k_{n-1}}$, which corresponds to the last line in (6.3). To satisfy (6.41)–(6.44) for instance, we just set

$$\bar{\gamma} := \frac{\mathrm{d}_{0,0}^2}{\alpha_{0,0}^\theta}\,, \quad \alpha_{n,k} = \frac{\alpha_{0,0}}{(n+1)^\sigma} \tag{6.46}$$

with $\alpha_{0,0}^\theta \leq \frac{\rho^2}{\bar{\gamma}}$ and $\sigma \in (0, 1]$ sufficiently small. Indeed, with this choice we have

$$1 - \left(\frac{\alpha_{n,k}}{\alpha_{n,k-1}}\right)^\theta = \begin{cases} 1 - \frac{n^{\sigma\theta}}{(n+1)^{\sigma\theta}} & \text{if } k = 0 \\ 0 & \text{else.}\end{cases}$$

In the first case by the Mean Value Theorem we have for some $t \in [0,1]$

$$\begin{aligned} 1 - \left(\frac{\alpha_{n,k}}{\alpha_{n,k-1}}\right)^\theta &\leq \frac{(n+1)^{\sigma\theta} - n^{\sigma\theta}}{(n+1)^{\sigma\theta}} = \frac{\sigma\theta(n+t)^{\sigma\theta-1}}{(n+1)^{\sigma\theta}} \\ &\leq \frac{\sigma\theta}{n} = \sigma\theta\left(\frac{\alpha_{n,k-1}}{\alpha_{0,0}}\right)^{\frac{1}{\sigma}} \leq \frac{\sigma\theta}{\alpha_{0,0}}\alpha_{n,k-1}\,. \end{aligned}$$

Hence, provided (6.45) holds, a sufficient condition for (6.44) is

$$1 \geq \frac{\sigma\theta}{\alpha_{0,0}} + c_0 + \frac{1}{\bar{\gamma}}\left(c_1\alpha_{0,0}^{s^*-\theta-1} + c_2\alpha_{0,0}^{p^*-\theta-1} + c_3\underline{\omega}^{-\theta} + \frac{C_5\overline{\omega}}{\tau^r}\right),$$

which can be achieved by making $c_0, c_3, \alpha_{0,0}, \frac{\sigma\theta}{\alpha_{0,0}}$ sufficiently small and τ sufficiently large.

If ν is unknown or just zero, then in order to obtain weak convergence only, we choose $\alpha_{n,k}$ a posteriori such that

$$\alpha_{n,k} \leq \min\{1\,,\ c\omega_{n,k}\mathrm{t}_{n,k}^r\} \tag{6.47}$$

for some sufficiently small constant $c > 0$.

Also the number k_n of interior Landweber iterations in the n-th Newton step acts as a regularization parameter. We will choose it such that (6.9) holds. In case b), i.e., when we cannot make use of a $\nu > 0$ in (6.4), we also require that on the other hand

$$\mu\|F(x_n^\delta) - y^\delta\| \geq \|A_n(z_{n,k_n} - x_n^\delta) + F(x_n^\delta) - y^\delta\| = \mathrm{t}_{n,k_n}. \tag{6.48}$$

While by $z_{n,0} = x_n^\delta$ and $\mu < 1$, obviously any $k_n \geq 1$ that is sufficiently small will satisfy (6.9), existence of a finite k_n such that (6.48) holds will have to be proven below.

The stopping index n_* of the outer iteration in case a) $\nu > 0$ is known will be chosen according to

$$n_*(\delta) = \min\{n \in \mathbb{N}\ :\ \exists k \in \{0,\ldots,k_n\}\ :\ \alpha_{n,k}^{\frac{1+\theta}{r}} \leq \tau\delta\}\,. \tag{6.49}$$

with some fixed $\tau > 0$ independent of δ, and we define our regularized solution as $z_{n_*,k_{n_*}^*}$ with the index

$$k_{n_*}^* = \min\{k \in \{0,\ldots,k_{n_*}\}\ :\ \alpha_{n_*,k}^{\frac{1+\theta}{r}} \leq \tau\delta\}\,. \tag{6.50}$$

Otherwise, in case b) we use a Discrepancy Principle

$$n_*(\delta) = \min\{n \in \mathbb{N} \ : \ \|F(x_n^\delta) - y^\delta\| \leq \tau\delta\} \tag{6.51}$$

and consider $x_{n_*}^\delta = z_{n_{n^*-1},k_{n_*-1}}$ as our regularized solution.

6.4 Weak convergence

We now consider the case in which the parameter ν in (6.4) is unknown or $\nu = 0$.

Using the notations of the previous sections, we recall that $\omega_{n,k}$, $\alpha_{n,k}$, k_n and $n_*(\delta)$ are chosen as follows. For fixed Newton step n, and Landweber step k we choose the step size $\omega_{n,k} > 0$ in (6.3) is such that

$$\underline{c}_\omega \leq \frac{\varphi\Big(\omega_{n,k}\tilde{\mathrm{t}}_{n,k}\Big)}{\omega_{n,k}\mathrm{t}_{n,k}^r} \leq \overline{c}_\omega \tag{6.52}$$

i.e., (6.33) holds. We refer to Section 6.3 for well-definedness of such a step size.

Next, we select $\alpha_{n,k}$ such that

$$\alpha_{n,k} \leq \min\left\{1, \gamma_0\omega_{n,k}\mathrm{t}_{n,k}^r\right\}, \tag{6.53}$$

where $\gamma_0 > 0$ satisfies

$$\gamma_0 < \frac{1 - \frac{\eta + \frac{1+\eta}{\tau}}{\mu} - \overline{c}_\omega}{\tilde{c}_0 D_p^{x_0}(x^\dagger, x_0)^{\frac{1}{p}} + c_1 + c_2}. \tag{6.54}$$

The stopping index of the inner Landweber iteration is chosen such that

$$\forall 0 \leq k \leq k_n - 1 \ : \quad \mu\|b_n\| \leq \mathrm{t}_{n,k} \tag{6.55}$$

i.e., (6.9) holds for some $\mu \in (\eta, 1)$ and on the other hand k_n is maximal with this property, i.e.

$$\mu\|b_n\| \geq \mathrm{t}_{n,k_n}. \tag{6.56}$$

The stopping index of the Newton iteration is chosen according to the discrepancy principle (6.51)

$$n_*(\delta) = \min\{n \in \mathbb{N} \ : \ \|b_n\| \le \tau\delta\}\,. \tag{6.57}$$

In order to show weak convergence, besides the tangential cone condition (6.7) we also assume that there is a constant $\gamma_1 > 0$ such that

$$\|F'(x)\| \le \gamma_1 \tag{6.58}$$

for all $x \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$.

We will now prove monotone decay of the Bergman distances between the iterates and the exact solution, cf. [30, 49]. Since n_* is chosen according to the Discrepancy Principle (6.51), and by (6.9), estimate (6.31) yields

$$\begin{aligned} \mathrm{d}_{n,k+1}^2 \ &\le \ \Big(1-\alpha_{n,k}\Big)\mathrm{d}_{n,k}^2 + \tilde{c}_0\alpha_{n,k}\mathrm{d}_{n,k}^{\frac{2}{p}} + c_1\alpha_{n,k}^{s^*} + c_2\alpha_{n,k}^{p^*} \\ &\quad - \left(1 - \frac{\eta + \frac{1+\eta}{\tau}}{\mu}\right)\omega_{n,k}\mathrm{t}_{n,k}^r + \varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k}), \end{aligned} \tag{6.59}$$

from (6.59) and the definitions of $\omega_{n,k}$ and $\alpha_{n,k}$ according to (6.52), (6.53), we infer

$$\begin{aligned} \mathrm{d}_{n,k+1}^2 - \mathrm{d}_{n,k}^2 \ &\le \ (\tilde{c}_0 D_p^{x_0}(x^\dagger, x_0)^{\frac{1}{p}} + c_1 + c_2)\alpha_{n,k} && (6.60) \\ &\quad -(1 - \frac{\eta + \frac{1+\eta}{\tau}}{\mu} - \overline{c}_\omega)\omega_{n,k}\mathrm{t}_{n,k}^r. && (6.61) \end{aligned}$$

Thus, since $\alpha_{n,k}$ is chosen smaller than $\gamma_0\omega_{n,k}\mathrm{t}_{n,k}^r$, we obtain

$$\mathrm{d}_{n,k+1}^2 - \mathrm{d}_{n,k}^2 \le -\gamma_2\omega_{n,k}\mathrm{t}_{n,k}^r, \tag{6.62}$$

with $\gamma_2 := 1 - \frac{\eta+\frac{1+\eta}{\tau}}{\mu} - \overline{c}_\omega - (\tilde{c}_0 D_p^{x_0}(x^\dagger, x_0)^{\frac{1}{p}} + c_1 + c_2)\gamma_0 > 0$ by(6.54). Summing over $k = 0, ..., k_n - 1$ we obtain

$$D_p^{x_0}(x^\dagger, x_n) - D_p^{x_0}(x^\dagger, x_{n+1}) = \sum_{k=0}^{k_n-1}(\mathrm{d}_{n,k}^2 - \mathrm{d}_{n,k+1}^2) \ge \gamma_2 \sum_{k=0}^{k_n-1}\omega_{n,k}\mathrm{t}_{n,k}^r. \tag{6.63}$$

Now, we use the definition of $\omega_{n,k}$ to observe that

$$\omega_{n,k}\mathrm{t}_{n,k}^r \geq \Phi\left(\frac{\mathrm{t}_{n,k}^r}{\tilde{\mathrm{t}}_{n,k}}\right) \geq \Phi\left(\frac{\mu\|b_n\|}{\|A_n\|}\right), \tag{6.64}$$

for $k \leq k_n - 1$, where the strictly positive and strictly monotonically increasing function $\Phi : \mathbb{R}^+ \to \mathbb{R}$ is defined by $\Phi(\lambda) = \lambda\psi^{-1}(\underline{c}_\omega\lambda)$, which yields

$$D_p^{x_0}(x^\dagger, x_n) - D_p^{x_0}(x^\dagger, x_{n+1}) \geq \gamma_2 k_n \Phi\left(\frac{\mu\|b_n\|}{\|A_n\|}\right). \tag{6.65}$$

Consequently, for every Newton step n with $b_n \neq 0$, the stopping index k_n is finite. Moreover, summing now over $n = 0, ..., n_*(\delta) - 1$ and using the assumed bound on F' (6.58) as well as (6.51) and $k_n \geq 1$, we deduce

$$D_p^{x_0}(x^\dagger, x_0) \geq D_p^{x_0}(x^\dagger, x_0) - D_p^{x_0}(x^\dagger, x_{n_*(\delta)}) \geq \gamma_2 n_*(\delta)\Phi\left(\frac{\mu\tau\delta}{\gamma_1}\right). \tag{6.66}$$

Thus, for $\delta > 0$, $n_*(\delta)$ is also finite, the method is well defined and we can directly follow the lines of the proof of Theorem 3.2 in [49] to show the weak convergence in the noisy case as stated in Theorem 6.4.1.

Besides the error estimates from Section 6.2, the key step of the proof of strong convergence as $n \to \infty$ in the noiseless case $\delta = 0$ of Theorem 6.4.1 is a Cauchy sequence argument going back to the seminal paper [30]. Since some additional terms appear in this proof due to the regularization term in the Landweber iteration, we provide this part of the proof explicitly here. By the identity

$$\begin{aligned} D_p^{x_0}(x_l, x_m) \;\; = \;\; & D_p^{x_0}(x^\dagger, x_m) - D_p^{x_0}(x^\dagger, x_l) \\ & + \langle J_p^{\mathcal{X}}(x_l - x_0) - J_p^{\mathcal{X}}(x_m - x_0), x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}} \end{aligned} \tag{6.67}$$

and the fact that the monotone decrease (6.63) and boundedness from below of the sequence $D_p^{x_0}(x^\dagger, x_m)$ implies its convergence, it suffices to prove that the last term in (6.67) tends to zero as $m < l \to \infty$. This term can be rewritten as

$$\langle J_p^{\mathcal{X}}(x_l - x_0) - J_p^{\mathcal{X}}(x_m - x_0), x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}} = \sum_{n=m}^{l-1}\sum_{k=0}^{k_n-1}\langle u_{n,k+1} - u_{n,k}, x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}$$

where

$$\begin{aligned}
&|\langle u_{n,k+1}-u_{n,k},x_l-x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}|\\
&= \quad |\alpha_{n,k}\langle J_p^{\mathcal{X}}(z_{n,k}-x_0),x_l-x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}\\
&\qquad +\omega_{n,k}\langle j_r^Y(A_n(z_{n,k}-x_n^\delta)-b_n),A_n(x_l-x^\dagger)\rangle_{\mathcal{X}^*\times\mathcal{X}}|\\
&\leq \quad \omega_{n,k}\mathrm{t}_{n,k}^{r-1}\left((2\bar\rho^p\gamma_0\mathrm{t}_{n,k}+\|A_n(x_l-x^\dagger)\|\right)
\end{aligned}$$

by our choice (6.53) of $\alpha_{n,k}$. Using (6.7), (6.48), it can be shown that

$$\|F(x_{n+1})-y\|\leq\frac{\mu+\eta}{1-\eta}\|F(x_n)-y\| \tag{6.68}$$

with a factor $\frac{\mu+\eta}{1-\eta}<1$ by our assumption $\mu<1-2\eta$, (which by continuity of F implies that a limit of x_n – if it exists – has to solve (5.17)). Hence, using again (6.7), as well as (6.68), we get

$$\|A_n(x_l-x^\dagger)\|\leq 2(1+\eta)\|F(x_n)-y\|+(1+\eta)\|F(x_l)-y\|\leq\frac{3(1+\eta)}{\mu}\mathrm{t}_{n,k}\,,$$

so that altogether we arrive at an estimate of the form

$$\begin{aligned}
&||\langle J_p^{\mathcal{X}}(x_l-x_0)-J_p^{\mathcal{X}}(x_m-x_0),x_l-x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}|\\
&\leq \quad C\sum_{n=m}^{l-1}\sum_{k=0}^{k_n-1}\omega_{n,k}\mathrm{t}_{n,k}^r\leq\frac{C}{\gamma_2}(D_p^{x_0}(x^\dagger,x_m)-D_p^{x_0}(x^\dagger,x_l))
\end{aligned}$$

by (6.63) with $l\geq n$, where the right-hand side goes to zero as $l,m\to\infty$ by the already mentioned convergence of the monotone and bounded sequence $D_p^{x_0}(x^\dagger,x_m)$.

Altogether we have proven the following result.

Theorem 6.4.1. *Assume that $\mathcal{X}$ is smooth and s-convex with $s\geq p$, that x_0 is sufficiently close to $x^\dagger$, i.e., $x_0\in\overline{\mathcal{B}}_\rho^D(x^\dagger)$, and that F satisfies* (6.7) *with* (6.6)*, that F and F' are continuous and uniformly bounded in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$. Let $\omega_{n,k},\alpha_{n,k},k_n,n_*$ be chosen according to* (6.9), (6.33), (6.47), (6.48), (6.51) *with $\eta<\frac{1}{3}$, $\eta<\mu<1-2\eta$, τ sufficiently large.*
Then, the iterates $z_{n,k}$ remain in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$ for all $n\leq n_*-1$, $k\leq k_n$, hence

any subsequence of $x^\delta_{n_*} = z_{n_*-1,k_{n_*-1}}$ *has a weakly convergent subsequence as* $\delta \to 0$. *Moreover, the weak limit of any weakly convergent subsequence solves* (5.17). *If the solution* $x^\dagger$ *to* (5.17) *is unique, then* $x^\delta_{n_*}$ *converges weakly to* $x^\dagger$ *as* $\delta \to 0$.

In the noise free case $\delta = 0$, x_n *converges strongly to a solution of* (5.17) *in* $\overline{\mathcal{B}}^D_\rho(x^\dagger)$.

6.5 Convergence rates with an a-priori stopping rule

We now consider the situation that $\nu > 0$ in (6.4) is known and recall that the parameters appearing in the methods are then chosen as follows, using the notation of Section 6.2.

First of all, for fixed Newton step n and Landweber step k we again choose the step size $\omega_{n,k} > 0$ in (6.3) such that (6.34) holds with a sufficiently small constant $\overline{c}_\omega > 0$ (see (6.69) below) which is possible as explained in Section 6.3. In order to make sure that $\omega_{n,k}$ stays bounded away from zero we assume that (6.38), (6.39) hold.

Next, we assume (6.45) and select $\alpha_{n,k}$ such that (6.41)–(6.44) holds.

Concerning the number k_n of interior Landweber iterations, we only have to make sure that (6.9) holds for some fixed $\mu \in (0,1)$ independent of n and k.

The overall stopping index n_* of the Newton iteration is chosen such that (6.49) holds.

With this n_*, our regularized solution is $z_{n_*,k^*_{n_*}}$ with the index according to (6.50).

The constants $\mu \in (0,1)$, $\tau > 0$ appearing in these parameter choices are a priori fixed, τ will have to be sufficiently large.

Moreover we assume that $\overline{c}_\omega$, β and η are small enough (the latter can be achieved by smallness of the radius ρ of the ball around $x^\dagger$ in which we will show the iterates to remain), so that we can choose $\epsilon > 0$ such that

$$c_4 + \overline{c}_\omega \leq 1 \tag{6.69}$$

with c_4 as in (6.28).

By the choice (6.34) of $\omega_{n,k}$, estimate (6.23) implies

$$\begin{aligned} \mathrm{d}^2_{n,k+1} \leq & \Big(1-(1-c_0)\alpha_{n,k}\Big)\mathrm{d}^2_{n,k} + c_1\alpha_{n,k}^{s^*} + c_2\alpha_{n,k}^{p^*} + c_3\omega_{n,k}^{-\theta}\alpha_{n,k}^{1+\theta} \\ & -(1-c_4-\overline{c}_\omega)\omega_{n,k}\mathrm{t}^r_{n,k} + C_5\omega_{n,k}\delta^r \,. \end{aligned} \tag{6.70}$$

Multiplying (6.70) with $\alpha_{n,k+1}^{-\theta}$, using (6.49), and abbreviating

$$\gamma_{n,k} := \mathrm{d}^2_{n,k}\alpha_{n,k}^{-\theta}\,,$$

we get

$$\begin{aligned} \gamma_{n,k+1} \leq & \left(\frac{\alpha_{n,k}}{\alpha_{n,k+1}}\right)^\theta \Big(\{1-(1-c_0)\alpha_{n,k}\}\,\gamma_{n,k} \\ & +(c_1\alpha_{n,k}^{s^*-\theta} + c_2\alpha_{n,k}^{p^*-\theta} + (c_3\underline{\omega}^{-\theta} + \frac{C_5\overline{\omega}}{\tau^r})\alpha_{n,k}\Big)\,. \end{aligned}$$

Using (6.44), this enables to inductively show

$$\gamma_{n,k+1} \leq \bar{\gamma}\,,$$

hence by (6.42) also

$$\mathrm{d}^2_{n,k+1} \leq \bar{\gamma}\alpha_{n,k}^\theta \leq \rho^2 \tag{6.71}$$

for all $n \leq n_* - 1$ and $k \leq k_n - 1$ as well as for $n = n_*$ and $k \leq k^*_{n_*} - 1$ according to (6.50). Inserting the upper estimate defining $k^*_{n_*}$ we therewith get

$$\mathrm{d}^2_{n_*,k^*_{n_*}} \leq \bar{\gamma}\alpha^\theta_{n_*,k^*_{n_*}} \leq \bar{\gamma}(\tau\delta)^{\frac{r\theta}{1+\theta}}\,,$$

which is the desired rate. Indeed, by (6.43), there exists a finite $k^*_{n_*} \leq k_{n_*}$ such that

$$\alpha_{n_*,k^*_{n_*}}^{\frac{1+\theta}{r}} \leq \tau\delta$$

and

$$\forall 0 \leq k \leq k^*_{n_*} - 1 \;:\quad \mu\|F(x^\delta_{n_*}) - y^\delta\| \leq \|A_n(z_{n_*,k} - x^\delta_{n_*}) + F(x^\delta_{n_*}) - y^\delta\|\,.$$

Summarizing, we arrive at

Theorem 6.5.1. *Assume that $\mathcal{X}$ is smooth and s-convex with $s \geq p$, that x_0 is sufficiently close to $x^\dagger$, i.e., $x_0 \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$, that a variational inequality* (6.4) *with $\nu \in (0,1]$ and β sufficiently small is satisfied, that F satisfies* (6.7) *with* (6.6)*, that F and F' are continuous and uniformly bounded in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$, and that* (6.45), (6.39) *hold. Let $\omega_{n,k}, \alpha_{n,k}, k_n, n_*, k^*_{n_*}$ be chosen according to* (6.9), (6.34)*,* (6.41)*–*(6.44)*,* (6.49)*,* (6.50) *with τ sufficiently large.*
Then, the iterates $z_{n,k}$ remain in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$ for all $n \leq n_* - 1$, $k \leq k_n$ and $n = n_*$, $k \leq k^*_{n_*}$. Moreover, we obtain optimal convergence rates

$$D_p^{x_0}(x^\dagger, z_{n_*,k^*_{n_*}}) = O(\delta^{\frac{4\nu}{2\nu+1}}), \quad \text{as } \delta \to 0 \tag{6.72}$$

as well as in the noise free case $\delta = 0$

$$D_p^{x_0}(x^\dagger, z_{n,k}) \quad = \quad O\left(\alpha_{n,k}^{\frac{4\nu}{r(2\nu+1)-4\nu}}\right) \tag{6.73}$$

for all $n \in \mathbb{N}$.

6.6 Numerical experiments

In this section we present some numerical experiments to test the method defined in Section 6.4. We consider the estimation of the coefficient c in the 1D boundary value problem

$$\begin{cases} -u'' + cu = f & \text{in } (0,1) \\ u(0) = g_0 & u(1) = g_1 \end{cases} \tag{6.74}$$

from the measurement of u, where g_0, g_1 and $f \in \mathcal{H}^{-1}[0,1]$ are given. Here and below, $\mathcal{H}^{-1}([0,1])$ is the dual space of the closure of $\mathcal{C}_0^\infty([0,1])$ in $\mathcal{H}^1([0,1])$, denoted by $\mathcal{H}_0^1([0,1])$, cf. e.g. [92]. We briefly recall some important facts about this problem (cf. Section 5.1):

1. For $1 \leq p \leq +\infty$ there exists a positive number γ_p such that for every c in the domain

$$\mathcal{D} := \{c \in \mathcal{L}^p[0,1] : \|c - \hat{\vartheta}\|_{\mathcal{L}^p} \leq \gamma_p, \hat{\vartheta} \geq 0 \text{ a.e.}\}$$

(6.74) has a unique solution $u = u(c) \in \mathcal{H}^1([0,1])$.

2. The nonlinear operator $F : \mathcal{D} \subseteq \mathcal{L}^p([0,1]) \to \mathcal{L}^r([0,1])$ defined as $F(c) := u(c)$ is Frechét differentiable and
$$F'(c)h = -\mathscr{A}(c)^{-1}(hu(c)), \quad F'(c)^*w = -u(c)\mathscr{A}(c)^{-1}w, \tag{6.75}$$
where $\mathscr{A}(c) : \mathcal{H}^2([0,1]) \cap \mathcal{H}^1_0([0,1]) \to \mathcal{L}^2([0,1])$ is defined by $\mathscr{A}(c)u = -u'' + cu$.

3. For every $p \in (1, +\infty)$ the duality map $J_p : \mathcal{L}^p([0,1]) \to \mathcal{L}^{p^*}([0,1])$ is given by
$$J_p(c) = |c|^{p-1}\mathrm{sgn}(c), \quad c \in \mathcal{L}^p([0,1]). \tag{6.76}$$

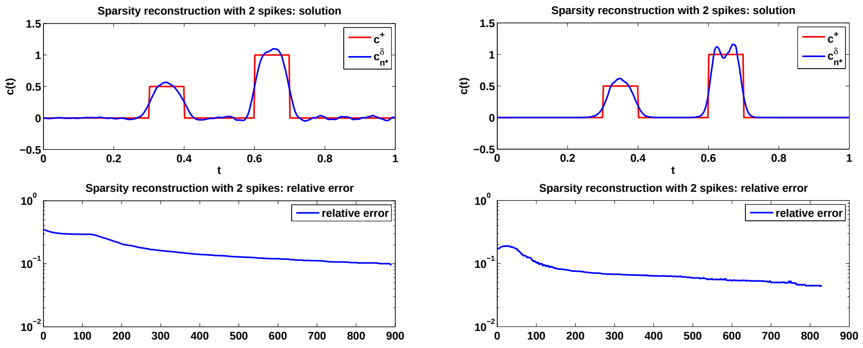
For the numerical simulations we take $\mathcal{X} = \mathcal{L}^p([0,1])$ with $1 < p < +\infty$ and $\mathcal{Y} = \mathcal{L}^r([0,1])$, with $1 \leq r \leq +\infty$ and identify c from noisy measurements u^δ of u. We solve all differential equations approximately by a finite difference method by dividing the interval $[0,1]$ into $N+1$ subintervals with equal length $1/(N+1)$; in all examples below $N = 400$. The $\mathcal{L}^p$ and $\mathcal{L}^r$ norms are calculated approximately too by means of a quadrature method.
We have chosen the same test problems as in [49]. Moreover, we added a variant of the example for sparsity reconstruction.
The parameters $\omega_{n,k}$ and $\alpha_{n,k}$ are chosen according to (6.35) and (6.53) and the outer iteration is stopped according to the Discrepancy Principle (6.51).
Concerning the stopping index of the inner iteration, in addition to the conditions (6.9) and (6.48), we require also that if $\|F(z_{n,k}) - y^\delta\| \leq \tau\delta$ then the iteration has to be stopped. More precisely,

$$k_n = \min\{k \in \mathbb{Z}, k \geq 0, \ \mid \|F(z_{n,k}) - y^\delta\| \leq \tau\delta \ \vee \ t_{n,k} \leq \mu\|b_n\|\} \tag{6.77}$$

and the regularized solution is $x^\delta_{n_*} = z_{n_{n^*-1}, k_{n_*-1}}$.

Example 6.6.1. *In the first simulation we assume that the solution is sparse:*

$$c^\dagger(t) = \begin{cases} 0.5, & 0.3 \leq t \leq 0.4, \\ 1.0, & 0.6 \leq t \leq 0.7, \\ 0.0, & \textit{elsewhere.} \end{cases} \tag{6.78}$$



(a) $p = 2$, $r = 2$.

(b) $p = 1.1$, $r = 2$.

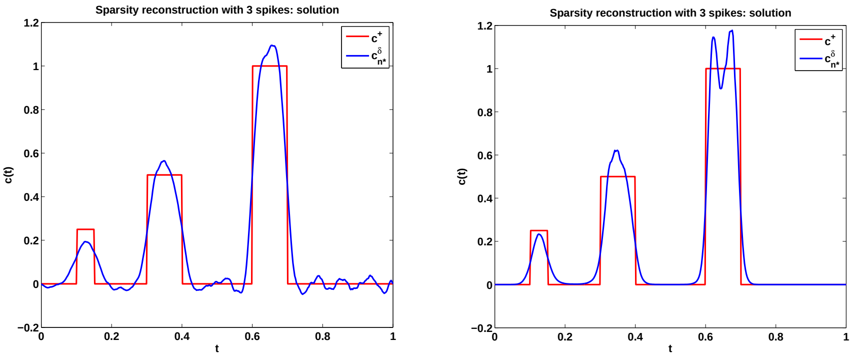
Figure 6.1: Reconstructed Solutions and relative errors for Example 6.6.1.

The test problem is constructed by taking $u(t) = u(c^{\dagger})(t) = 1 + 5t$, $f(t) = u(t)c^{\dagger}(t)$, $g_0 = 1$ *and* $g_1 = 6$. *We perturb the exact data* u *with gaussian white noise: the corresponding perturbed data* u^{δ} *satisfies* $\|u^{\delta} - u\|_{\mathcal{L}^r} = \delta$, *with* $\delta = 0.1 \times 10^{-3}$. *When applying the method of Section* 6.4, *we take* $\mu = 0.99$ *and* $\tau = 1.02$. *The upper bound* $\overline{c}_{\omega}$ *satisfies*

$$\overline{c}_{\omega} = 2^{s^*-1} C_{p^*,s^*} (p\rho^2)^{1-s^*/p^*} \hat{\vartheta}^{s^*-1} + 2^{p^*-1} C_{p^*,s^*} \hat{\vartheta}^{p^*-1}, \tag{6.79}$$

and $\hat{\vartheta}$ *is chosen as* $2^{-j^{\sharp}}$, *where* $j^{\sharp}$ *is the first index such that* $\gamma_0 := 0.99(1 - \frac{\eta}{\mu} - \frac{1+\eta}{\tau\mu} - \overline{c}_{\omega}) > 0$. *In the tests, we always choose* $\mathcal{Y} = \mathcal{L}^2([0,1])$ *and change the values of* p. *In Figure* 6.1 *we show the results obtained by our method with* $p = 2$ *and* $p = 1.1$ *respectively. From the plot of the solution we can see that the reconstruction of the sparsity in the case* $p = 1.1$ *is much better than in the case with* $p = 2$ *and the quality of the solutions is in line with what one should expect* (*cf. the solutions obtained in* [*49*]). *From the plot of the relative errors we note that a strict monotonicity of the error cannot be observed in the case with* $p = 1.1$. *The monotonicity holds instead in the case* $p = 2$. *We also underline that in this example the total number of the inner iterations*

$$N_p = \sum_{n=0}^{n^*-1} k_n$$



(a) $p = 2$, $r = 2$.

(b) $p = 1.1$, $r = 2$.

Figure 6.2: Reconstructed Solutions for Example 6.6.2, case A.

is much larger in the case $p = 2$ $(N_2 = 20141)$ *than in the case* $p = 1.1$ $(N_{1.1} = 4053)$*, thus the reconstruction with* $p = 1.1$ *is also faster.*

Example 6.6.2. *Choosing a different exact solution doesn't change the results too much. In this example we only modify* $c^{\dagger}$ *into*

$$c^{\dagger}(t) = \begin{cases} 0.25, & 0.1 \leq t \leq 0.15, \\ 0.5, & 0.3 \leq t \leq 0.4, \\ 1.0, & 0.6 \leq t \leq 0.7, \\ 0.0 & elsewhere. \end{cases} \tag{6.80}$$

and choose again $\delta = 0.1 \times 10^{-3}$.

The reconstructed solutions obtained show that choosing a p *smaller than* 2 *improves the results because the oscillations in the zero parts are damped significantly. Once again, the iteration error and the residual norms do not decrease monotonically in the case* $p = 1.1$*, but only in the average.*

We also tested the performance obtained by the method with the choice $\alpha_{n,k} = 0$ *instead of* (6.53) *and summarized the results in Table* 6.1*. Similarly to Example* 1*, a small* p *allows not only to get the better reconstruction, but also to spare time in the computation. Moreover, we notice that in this example the method with* $\alpha_{n,k} > 0$ *chosen according to* (6.53) *proved to be faster than*

Results for Example 2				
	$p = 2$, $\alpha_{n,k} > 0$	$p = 2$, $\alpha_{n,k} = 0$	$p = 1.1$, $\alpha_{n,k} > 0$	$p = 1.1$, $\alpha_{n,k} = 0$
N_p	21610	26303	4529	5701
Rel. Err.	9.8979×10^{-2}	9.8938×10^{-2}	4.9645×10^{-2}	4.9655×10^{-2}

Table 6.1: Numerical results for Example 2.

the method with $\alpha_{n,k} = 0$*, performing fewer iterations, with a gain of* 17.8% *in the case with* $p = 2$ *and* 20.5% *with* $p = 1.1$*.*

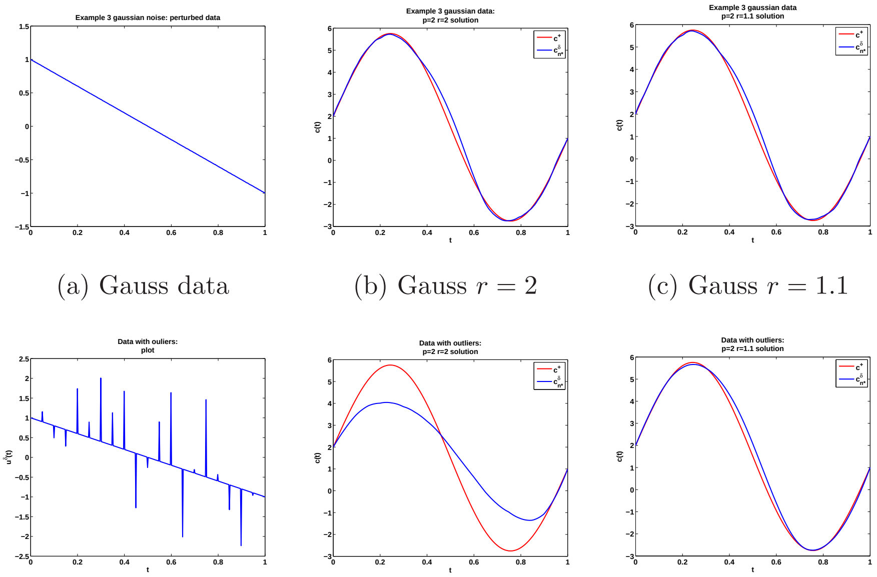
Example 6.6.3. *At last, we consider an example with noisy data where a few data points called outliers are remarkably different from other data points. This situation may arise from procedural measurement errors.*
We suppose $c^\dagger$ *to be a smooth solution*

$$c^\dagger(t) = 2 - t + 4\sin(2\pi t) \tag{6.81}$$

and take $u(c^\dagger)(t) = 1 - 2t$*,* $f(t) = (1 - 2t)(2 - t + 4\sin(2\pi t))$*,* $g_0 = 1$ *and* $g_1 = -1$ *as exact data of the problem. We start the iteration from the initial guess* $c_0(t) = 2 - t$*, fix the parameters* $\mu = 0.999$ *and* $\tau = 1.0015$ *and choose* $\overline{c}_\omega$ *and* γ_0 *as in Example* 1*.*

Case A. *At first, we assume the data are perturbed with white gaussian noise* ($\delta = 0.1 \times 10^{-2}$)*, fix* $\mathcal{X} = \mathcal{L}^2([0,1])$ *and take* $\mathcal{Y} = \mathcal{L}^r([0,1])$*, with* $r = 2$ *or* $r = 1.1$*. As we can see from Figure* 6.3*, being the data reasonably smooth, we obtain comparable reconstructions* (*in the case* $r = 2$ *the relative error is equal to* 2.1331×10^{-1}*, whereas in the case* $r = 1.1$ *we get* 2.0883×10^{-1})*.*

Case B. *The situation is much different if the perturbed data contain also a few outliers. We added* 19 *outliers to the gaussian noise perturbed data of case A obtaining the new noise level* $\delta = 0.0414$*. In this case taking* $\mathcal{Y} = \mathcal{L}^{1.1}([0,1])$ *considerably improves the results keeping the relative error reasonably small* (2.9388×10^{-1} *against* 1.1992*, cf. Figure* 6.3)*.*



(a) Gauss data

(b) Gauss $r = 2$

(c) Gauss $r = 1.1$

(d) Outliers data

(e) Outliers $r = 2$

(f) Outliers $r = 1.1$

Figure 6.3: Numerical results for Example 6.6.3: (a) and (d) are data with noise; (b) and (d) are reconstructions with $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2[0,1]$; (c) and (f) are reconstructions with $\mathcal{X} = \mathcal{L}^2[0,1]$ and $\mathcal{Y} = \mathcal{L}^{1.1}[0,1]$.

Concerning the total number of iterations N, in this example it is not subjected to remarkable variations.

To summarize, in all examples above the method obtained reasonable results, proving to be reliable, both in the sparsity reconstruction examples and when the data are affected with outliers. Concerning the total number of iterations, the introduction of the parameters $\alpha_{n,k}$ has accelerated the method in Example 2, but the issue of the speed of the algorithms requires further investigation.

6.7 A new proposal for the choice of the parameters

With the parameter choices proposed in Section 6.3, it is rather difficult to prove strong convergence and convergence rates for the Discrepancy Principle. Moreover, the numerical experiments show that the method still requires many iterations to obtain good regularized solutions.
Indeed, the choices that have been made for $\omega_{n,k}$ and $\alpha_{n,k}$ seem to make the method not flexible enough. For this reason, we propose here a different way to select these parameters. To do this, we return to the estimates of Section 6.2. Using the same notations, we estimate the term $\mathrm{d}_{n,k+1}^2$ as in (6.10) and proceed exactly as in Section 6.2 for estimating the terms (I), (II) and (IV). For the term (III), if $\nu = 0$ instead of using (6.9), we reconsider the estimate

$$\begin{aligned} &\|F(z_{n,k}) - F(x^\dagger)\| \\ &\quad = \|(A_n(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta) \\ &\qquad\quad +(F(z_{n,k}) - F(x_n^\delta) - A_n((z_{n,k} - x_n^\delta) + (y^\delta - y)\| \\ &\quad \leq \mathrm{t}_{n,k} + \eta\|F(z_{n,k}) - F(x_n^\delta)\| + \delta \\ &\quad \leq \mathrm{t}_{n,k} + \eta(\|F(z_{n,k}) - F(x^\dagger)\| + \|F(x_n^\delta) - y^\delta\|) + (1+\eta)\delta\,, \end{aligned}$$

to conclude

$$\|F(z_{n,k}) - F(x^\dagger)\| \leq \frac{1}{1-\eta}\left(\mathrm{t}_{n,k} + \eta\mathrm{r}_n + (1+\eta)\delta\right)\,. \tag{6.82}$$

This together with (6.4) implies

$$\begin{aligned}
&|\alpha_{n,k}\langle J_p^{\mathcal{X}}(x^\dagger - x_0), z_{n,k} - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}| \\
&\quad\leq \frac{\beta}{(1-\eta)^{2\nu}}\alpha_{n,k}\mathrm{d}_{n,k}^{1-2\nu}\left(\mathrm{t}_{n,k} + \eta\mathrm{r}_n + (1+\eta)\delta\right)^{2\nu} \\
&\quad\leq C(\tfrac{1}{2}+\nu)\frac{\beta}{(1-\eta)^{2\nu}}\alpha_{n,k}\left\{\mathrm{d}_{n,k}^2 + (\mathrm{t}_{n,k} + \eta\mathrm{r}_n + (1+\eta)\delta)^{\frac{4\nu}{1+2\nu}}\right\} \qquad (6.83)
\end{aligned}$$

where we have used (6.18) with $C(\lambda) = \lambda^\lambda(1-\lambda)^{1-\lambda}$.
In the case $\nu = 0$, we simply use (6.21) as in Section 6.2.
Altogether, we obtain

$$\begin{aligned}
\mathrm{d}_{n,k+1}^2 \leq\ & \Big(1 - (1-c_0)\alpha_{n,k}\Big)\mathrm{d}_{n,k}^2 + \tilde{c}_0\alpha_{n,k} + c_1\alpha_{n,k}^{s^*} + c_2\alpha_{n,k}^{p^*} \\
& + c_3\alpha_{n,k}\left(\mathrm{t}_{n,k} + \eta\mathrm{r}_n + (1+\eta)\delta\right)^{\frac{4\nu}{1+2\nu}} \\
& - \omega_{n,k}\mathrm{t}_{n,k}^r + \omega_{n,k}\mathrm{t}_{n,k}^{r-1}(\eta\mathrm{r}_n + (1+\eta)\delta) + \varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k})\,, \qquad (6.84)
\end{aligned}$$

where

$$\begin{aligned}
c_0 &= \begin{cases} 0 & \text{if } \nu = 0 \\ \frac{\beta}{(1-\eta)^{2\nu}}C(\frac{1}{2}+\nu) & \text{if } \nu > 0 \end{cases} \qquad (6.85) \\
\tilde{c}_0 &= \begin{cases} \|x^\dagger - x_0\|^{p-1}(p\rho^2)^{\frac{1}{p}} & \text{if } \nu = 0 \\ 0 & \text{if } \nu > 0 \end{cases} \\
c_1 &= C_{p^*,s^*}(p\rho^2)^{1-\frac{s^*}{p^*}}\bar{\rho}^{(p-1)s^*}2^{s^*-1} \qquad (6.86) \\
c_2 &= C_{p^*,s^*}\bar{\rho}^p 2^{p^*-1} \qquad (6.87) \\
c_3 &= \begin{cases} c_0 & \text{in case (a) } \nu,\theta > 0 \\ \|x^\dagger - x_0\|^{p-1}(p\rho^2)^{\frac{1}{p}} & \text{in case (b) } \nu,\theta = 0 \end{cases} \qquad (6.88) \\
\theta &= \frac{4\nu}{r(1+2\nu) - 4\nu}\,. \qquad (6.89)
\end{aligned}$$

Multiplying (6.84) with $\alpha_{n,k+1}^{-\theta}$ and abbreviating

$$\gamma_{n,k} := \mathrm{d}_{n,k}^2\alpha_{n,k}^{-\theta}\,,$$

we get

$$\begin{aligned}\gamma_{n,k+1}-\gamma_{n,k} \leq\ & \left(\frac{\alpha_{n,k}}{\alpha_{n,k+1}}\right)^{\theta}\Big\{\Big(1-(1-c_0)\alpha_{n,k}-\left(\frac{\alpha_{n,k+1}}{\alpha_{n,k}}\right)^{\theta}\Big)\gamma_{n,k}\\ +(\tilde{c}_0\alpha_{n,k}^{1-\theta} & + c_1\alpha_{n,k}^{s^*-\theta}+c_2\alpha_{n,k}^{p^*-\theta}+c_3\alpha_{n,k}^{1-\theta}\left(\mathrm{t}_{n,k}+\eta\mathrm{r}_n+(1+\eta)\delta\right)^{\frac{4\nu}{1+2\nu}}\\ & -\alpha_{n,k}^{-\theta}\left(\omega_{n,k}\mathrm{t}_{n,k}^{r}-\omega_{n,k}\mathrm{t}_{n,k}^{r-1}(\eta\mathrm{r}_n+(1+\eta)\delta)-\varphi(\omega_{n,k}\tilde{\mathrm{t}}_{n,k})\right)\Big\}\,.\end{aligned}$$

To obtain monotone decay of the sequence $\gamma_{n,k}$ with increasing k we choose

- $\omega_{n,k} \geq 0$ such that

$$\underline{\omega} \leq \omega_{n,k} \leq \overline{\omega} \text{ and } \frac{\varphi\Big(\omega_{n,k}\tilde{\mathrm{t}}_{n,k}\Big)}{\omega_{n,k}\mathrm{t}_{n,k}^{r}} \leq \overline{c}_{\omega} \tag{6.90}$$

for some $0 < \underline{\omega} < \overline{\omega}$, $\overline{c}_{\omega} > 0$. We will do so by setting

$$\omega_{n,k} = \vartheta \min\{\mathrm{t}_{n,k}^{\frac{r}{s^*-1}}\tilde{\mathrm{t}}_{n,k}^{-s}\,,\ \mathrm{t}_{n,k}^{\frac{r}{p^*-1}}\tilde{\mathrm{t}}_{n,k}^{-p}\,,\ \overline{\omega}\} \tag{6.91}$$

with ϑ sufficiently small, and assuming that

$$r \geq s \geq p\,, \tag{6.92}$$

- $\alpha_{n,k} \geq 0$ such that

$$\alpha_{n,k} \geq \check{\alpha}_{n,k} := \tilde{\tau}\left(\mathrm{t}_{n,k}+\eta\mathrm{r}_n+(1+\eta)\delta\right)^{\frac{r}{1+\theta}} \tag{6.93}$$

and

$$c_0+\frac{\tilde{c}_0}{\overline{\gamma}_{0,0}}+\frac{c_1}{\overline{\gamma}_{0,0}}\alpha_{n,k}^{s^*-\theta-1}+\frac{c_2}{\overline{\gamma}_{0,0}}\alpha_{n,k}^{p^*-\theta-1}+\frac{c_3}{\tilde{\tau}^{\theta}\overline{\gamma}_{0,0}}+\frac{1}{\tilde{\tau}^{1+\theta}\overline{\gamma}_{0,0}} \leq q < 1 \tag{6.94}$$

(note that in case $\theta = 0$ we have $c_0 = 0$ and vice versa, in case $\theta > 0$ we have $\tilde{c}_0 = 0$). The latter can be achieved by

$$\alpha_{n,k} \leq 1 \text{ and} \tag{6.95}$$

$$s^* \geq \theta+1\,,\quad p^* \geq \theta+1\,, \tag{6.96}$$

$$c_0, \tilde{c}_0, c_1, c_2, c_3, \tilde{\tau}^1 \text{ sufficiently small.}$$

In case (a) we additionally require

$$\alpha_{n,k+1} \geq \hat{\alpha}_{n,k+1} := \alpha_{n,k}\Big(1-(1-q)\alpha_{n,k}\Big)^{1/\theta} \tag{6.97}$$

with an upper bound $\overline{\gamma}_{0,0}$ for $\gamma_{0,0}$. Note that this just means $\alpha_{n,k+1} \geq 0$ in case (b) corresponding to $\nu = 0$, i.e., $\theta = 0$, thus an empty condition in case (b).

To meet conditions (6.93), (6.97) with a minimal $\alpha_{n,k+1}$ we set

$$\begin{aligned} \alpha_{n,k+1} &= \max\{\check{\alpha}_{n,k+1}\,,\, \hat{\alpha}_{n,k+1}\} \text{ for } k \geq 0 \\ \alpha_{n,0} &= \begin{cases} \alpha_{n-1,k^{n-1}} & \text{if } n \geq 1 \\ \alpha_{0,0} & \text{if } n = 0 \end{cases} . \end{aligned} \tag{6.98}$$

It remains to choose

- the inner stopping index k^n
- the outer stopping index n_*,

see below.

Indeed with these choices of $\omega_{n,k}$ and $\alpha_{n,k+1}$ we can inductively conclude from (6.90) that

$$\begin{aligned} \gamma_{n,k+1} - \gamma_{n,k} \;\; &\leq \;\; \left(\frac{\alpha_{n,k}}{\alpha_{n,k+1}}\right)^{\theta} \left\{\left(1-(1-q)\alpha_{n,k} - \left(\frac{\alpha_{n,k+1}}{\alpha_{n,k}}\right)^{\theta}\right)\overline{\gamma}_{0,0}\right\} \\ &\quad -\alpha_{n,k+1}^{-\theta}(1-\overline{c}_\omega)\omega_{n,k}\mathrm{t}_{n,k}^r\,, \\ &\leq \;\; -\alpha_{n,k+1}^{-\theta}(1-\overline{c}_\omega)\omega_{n,k}\mathrm{t}_{n,k}^r \leq 0\,. \end{aligned} \tag{6.99}$$

This monotonicity result holds for all $n \in \mathbb{N}$ and for all $k \in \mathbb{N}$.

By (6.99) and $\alpha_{n,k} \leq 1$ (cf. (6.95)) it can be shown inductively that all iterates remain in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$ provided

$$\overline{\gamma}_{0,0} \leq \rho^2\,. \tag{6.100}$$

Moreover, (6.99) implies that

$$\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \alpha_{n,k+1}^{-\theta}\omega_{n,k}\mathrm{t}_{n,k}^r \leq \frac{\overline{\gamma}_{0,0}}{1-\overline{c}_\omega} < \infty, \tag{6.101}$$

hence by $\alpha_{n,k+1} \leq 1$, $\omega_{n,k} \geq \underline{\omega}$

$$\mathrm{t}_{n,k} \to 0 \text{ as } k \to \infty \text{ for all } n \in \mathbb{N} \tag{6.102}$$

and

$$\sup_{k \in \mathbb{N}_0} \mathrm{t}_{n,k} \to 0 \text{ as } n \to \infty\,. \tag{6.103}$$

Especially, since $\mathrm{t}_{n,0} = \mathrm{r}_n$,

$$\mathrm{r}_n \to 0 \text{ as } n \to \infty\,. \tag{6.104}$$

To quantify the behavior of the sequence $\alpha_{n,k}$ according to (6.93), (6.97), (6.98) for fixed n we distinguish between two cases.

(i) There exists a $\underline{k}$ such that for all $k \geq \underline{k}$ we have $\alpha_{n,k} = \hat{\alpha}_{n,k}$. Considering an arbitrary accumulation point $\bar{\alpha}_n$ of $\alpha_{n,k}$ (which exists since $0 \leq \alpha_{n,k} \leq 1$) we therefore have $\bar{\alpha}_n = \bar{\alpha}_n\Big(1 - (1-q)\bar{\alpha}_n\Big)^{\frac{1}{\theta}}$, hence $\bar{\alpha}_n = 0$.

(ii) Consider the situation that (i) does not hold, i.e., there exists a subsequence k_j such that for all $j \in \mathbb{N}$ we have $\alpha_{n,k_j} = \check{\alpha}_{n,k_j}$. Then by (6.93), (6.97), and (6.102) we have $\alpha_{n,k_j} \to \tilde{\tau}\,(\eta \mathrm{r}_n + (1+\eta)\delta)^{\frac{r}{1+\theta}}$.

Altogether we have shown that

$$\limsup_{k\to\infty} \alpha_{n,k_j} \leq \tilde{\tau}\,(\eta \mathrm{r}_n + (1+\eta)\delta)^{\frac{r}{1+\theta}} \text{ for all } n \in \mathbb{N}. \tag{6.105}$$

Since η and δ can be assumed to be sufficiently small, this especially implies the bound $\alpha_{n,k} \leq 1$ in (6.95).

We consider $z_{n_*,k_*^{n_*}}$ as our regularized solution, where n_*, $k_*^{n_*}$ (and also k^n for all $n \leq n_* - 1$; note that $k_*^{n_*}$ is to be distinguished from k^{n_*} - actually the latter is not defined, since we only define k^n for $n \leq n_* - 1$!) are still to be chosen appropriately, according to the requirements from the proofs of

- convergence rates in case $\nu, \theta > 0$,
- convergence for exact data $\delta = 0$,
- convergence for noisy data as $\delta \to 0$.

6.7.1 Convergence rates in case $\nu > 0$

From (6.99) we get

$$\mathrm{d}_{n,k}^2 \leq \overline{\gamma}_{0,0}\alpha_{n,k}^\theta \text{ for all } n,k \in \mathbb{N}\,, \tag{6.106}$$

hence in order to get the desired rate

$$\mathrm{d}_{n_*,k_*^{n_*}}^2 = O(\delta^{\frac{r\theta}{1+\theta}})$$

in view of (6.105) (which is a sharp bound in case (ii) above) we need to have a bound

$$\mathrm{r}_{n*} \leq \tau\delta \tag{6.107}$$

for some constant $\tau > 0$, and we should choose $k_*^{n_*}$ large enough so that

$$\alpha_{n_*,k_*^{n_*}} \leq C_\alpha(\mathrm{r}_{n_*}+\delta)^{\frac{r}{1+\theta}} \tag{6.108}$$

which is possible with a finite $k_*^{n_*}$ by (6.105) for $C_\alpha > (\tilde{\tau}(1+\eta))^{\frac{r}{1+\theta}}$. Note that this holds without any requirements on k^n.

6.7.2 Convergence as n $\to \infty$ for exact data $\delta = 0$

To show that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence (following the seminal paper [30]), for arbitrary $m < j$, we choose the index $l \in \{m,\ldots,j\}$ such that r_l is minimal and use the identity

$$\begin{aligned} D_p^{x_0}(x_l,x_m) &= D_p^{x_0}(x^\dagger,x_m) - D_p^{x_0}(x^\dagger,x_l) \\ &\quad +\langle J_p^{\mathcal{X}}(x_l-x_0) - J_p^{\mathcal{X}}(x_m-x_0), x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}} \end{aligned} \tag{6.109}$$

and the fact that the monotone decrease and boundedness from below of the sequence $D_p^{x_0}(x^\dagger,x_m)$ implies its convergence, hence it suffices to prove that the last term in (6.109) tends to zero as $m < l \to \infty$ (analogously it can be shown that $D_p^{x_0}(x_l,x_j)$ tends to zero as $l < j \to \infty$). This term can be rewritten as

$$\begin{aligned} &\langle J_p^{\mathcal{X}}(x_l-x_0) - J_p^{\mathcal{X}}(x_m-x_0), x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}} \\ &= \sum_{n=m}^{l-1}\sum_{k=0}^{k^n-1}\langle u_{n,k+1}-u_{n,k}, x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}\,, \end{aligned}$$

where

$$
\begin{aligned}
&|\langle u_{n,k+1} - u_{n,k}, x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}}| \\
&\quad= |\alpha_{n,k}\langle J_p^{\mathcal{X}}(z_{n,k} - x_0), x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}} \\
&\qquad +\omega_{n,k}\langle j_r^Y(A_n(z_{n,k} - x_n^\delta) - b_n), A_n(x_l - x^\dagger)\rangle_{\mathcal{X}^*\times\mathcal{X}}| \\
&\quad\leq 2\bar{\rho}^p\alpha_{n,k} + \omega_{n,k}\mathrm{t}_{n,k}^{r-1}\|A_n(x_l - x^\dagger)\| \\
&\quad\leq 2\bar{\rho}^p\tilde{\tau}(\mathrm{t}_{n,k} + \eta\mathrm{r}_n)^r + \omega_{n,k}\mathrm{t}_{n,k}^{r-1}(1+\eta)(2\mathrm{r}_n + \mathrm{r}_l) \\
&\quad\leq 2\bar{\rho}^p\tilde{\tau}(\mathrm{t}_{n,k} + \eta\mathrm{r}_n)^r + 3(1+\eta)\omega_{n,k}\mathrm{t}_{n,k}^{r-1}\mathrm{r}_n
\end{aligned}
$$

by our choice of $\alpha_{n,k} = \check{\alpha}_{n,k}$ (note that $\hat{\alpha}_{n,k} = 0$ in case $\theta = 0$), condition (6.7) and the minimality of r_l.
Thus we have by $\omega_{n,k} \leq \overline{\omega}$ and Young's inequality that there exists $C > 0$ such that

$$\langle J_p^{\mathcal{X}}(x_l - x_0) - J_p^{\mathcal{X}}(x_m - x_0), x_l - x^\dagger\rangle_{\mathcal{X}^*\times\mathcal{X}} \leq C\sum_{n=m}^{l-1}\left\{\left(\sum_{k=0}^{k_n-1}\mathrm{t}_{n,k}^r\right) + k_n\mathrm{r}_n^r\right)$$

for which we can conclude convergence as $m, l \to \infty$ from (6.101) provided that

$$\sum_{n=m}^{\infty} k^n\mathrm{r}_n^r \to 0 \text{ as } m \to \infty\,,$$

which we guarantee by choosing, for an a priori fixed summable sequence $(a_n)_{n\in\mathbb{N}}$,

$$k^n := a_n\mathrm{r}_n^{-r}\,. \tag{6.110}$$

6.7.3 Convergence with noisy data as $\delta \to 0$

In case (a) $\nu, \theta > 0$ convergence follows from the convergence rates results in Subsection 6.7.1. Therefore it only remains to show convergence as $\delta \to 0$ in case $\theta = 0$.

In this section we explicitly emphasize dependence of the computed quantities on the noisy data and on the noise level by a superscript δ.

Let $\|y^{\delta_j} - y\| \leq \delta_j$ with δ_j a zero sequence and n_{*j} the corresponding stopping index. As usual [30] we distinguish between the two cases that (i) n_{*j} has a finite accumulation point and (ii) n_{*j} tends to infinity.

(i) There exists an $N \in \mathbb{N}$ and a subsequence n_{j_i} such that for all $i \in \mathbb{N}$ we have $n_{j_i} = N$. Provided

$$n_*(\delta) = N \text{ for all } \delta \;\Rightarrow\; \text{The mapping } \delta \mapsto x_N^\delta \text{ is continuous at } \delta = 0\,, \tag{6.111}$$

we can conclude that $x_N^{\delta_{j_i}} \to x_N^0$ as $i \to \infty$, and by taking the limit as $i \to \infty$ also in (6.107), x_N^0 is a solution to (5.17). Thus we may set $x^\dagger = x_N^0$ in (6.99) (with $\theta = 0$) to obtain

$$D_p^{x_0}(x_N^0, z^{\delta_{j_i}}_{n_{*j_i},k^{n_{*j_i}}_{*j_i}}) = D_p^{x_0}(x_N^0, z^{\delta_{j_i}}_{N,k^{n_{*j_i}}_{*j_i}}) \leq D_p^{x_0}(x_N^0, x_N^{\delta_{j_i}}) \to 0 \text{ as } i \to \infty$$

where we have again used the continuous dependence (6.111) in the last step.

(ii) Let $n_{*j} \to \infty$ as $j \to \infty$, and let $x^\dagger$ be a solution to (5.17). For arbitrary $\epsilon > 0$, by convergence for $\delta = 0$ (see the previous subsection) we can find n such that $D_p^{x_0}(x^\dagger, x_n^0) < \frac{\epsilon}{2}$ and, by Theorem 2.60 (d) in [82] there exists j_0 such that for all $j \geq j_0$ we have $n_{*,j} \geq n+1$ and $|D_p^{x_0}(x^\dagger, x_n^{\delta_j}) - D_p^{x_0}(x^\dagger, x_n^0)| < \frac{\epsilon}{2}$, provided

$$n \leq n_*(\delta) - 1 \text{ for all } \delta \;\Rightarrow\; \text{The mapping } \delta \mapsto x_n^\delta \text{ is continuous at } \delta = 0\,. \tag{6.112}$$

Hence, by monotonicity of the errors we have

$$\begin{aligned} & D_p^{x_0}(x^\dagger, z^{\delta_j}_{n_{*j},k^{n_{*j}}_{*j}}) \leq D_p^{x_0}(x^\dagger, x_n^{\delta_j}) \\ & \leq D_p^{x_0}(x^\dagger, x_n^0) + |D_p^{x_0}(x^\dagger, x_n^{\delta_j}) - D_p^{x_0}(x^\dagger, x_n^0)| < \epsilon\,. \end{aligned} \tag{6.113}$$

Indeed, (6.111), (6.112) can be concluded from continuity of F, F', the definition of the method (6.3), as well as stable dependence of all parameters $\omega_{n,k}$, $\alpha_{n,k}$, k^n according to (6.91), (6.93), (6.97), (6.98), (6.110) on the data y^δ.

Altogether we have derived the following algorithm.

6.7.4 Newton-Iteratively Regularized Landweber algorithm

```
Choose $\tau, \tilde{\tau}, C_\alpha$ sufficiently large, $x_0$ sufficiently close to $x^\dagger$,
$\alpha_{00} \le 1$, $\overline{\omega} > 0$, $(a_n)_{n\in\mathbb{N}_0}$ such that $\sum_{n=0}^{\infty} a_n < \infty$.
If (6.4) with $\nu \in (0,1]$ holds, set $\theta = \frac{4\nu}{r(1+2\nu)-4\nu}$, otherwise $\theta = 0$.
For $n = 0, 1, 2 \dots$ until $\mathrm{r}_n \le \tau\delta$ do
    $u_{n,0} = 0$
    $z_{n,0} = x_n^\delta$
    $\alpha_{n,0} = \alpha_{n-1,k^{n-1}}$ if $n > 0$
    For $k = 0, 1, 2 \dots$ until $\left\{ \begin{array}{ll} k = k^n - 1 = a_n \mathrm{r}_n^{-r} & \text{if } \mathrm{r}_n > \tau\delta \\ \alpha_{n_*, k_*^{n*}} \le C_\alpha(\mathrm{r}_{n_*} + \delta)^{\frac{r}{1+\theta}} & \text{if } \mathrm{r}_n \le \tau\delta \end{array} \right\}$ do
        $\omega_{n,k} = \vartheta \min\{\mathrm{t}_{n,k}^{\frac{r}{s^*-1}} \tilde{\mathrm{t}}_{n,k}^{-s}\,,\ \mathrm{t}_{n,k}^{\frac{r}{p^*-1}} \tilde{\mathrm{t}}_{n,k}^{-p}\,,\ \overline{\omega}\}$
        $u_{n,k+1} = u_{n,k} - \alpha_{n,k} J_p^{\mathcal{X}}(z_{n,k} - x_0)$
                $-\omega_{n,k} F'(x_n^\delta)^* j_r^Y (F'(x_n^\delta)(z_{n,k} - x_n^\delta) + F(x_n^\delta) - y^\delta)$
        $J_p^{\mathcal{X}}(z_{n,k+1} - x_0) = J_p^{\mathcal{X}}(x_n^\delta - x_0) + u_{n,k+1}$
        $\alpha_{n,k+1} = \max\{\check{\alpha}_{n,k+1}\,,\ \hat{\alpha}_{n,k+1}\}$ with $\check{\alpha}_{n,k+1}\,,\ \hat{\alpha}_{n,k+1}$ as in (6.93), (6.97)
    $x_{n+1}^\delta = z_{n,k^n}\,.$
```

Note that we here deal with an a priori parameter choice: θ and therefore ν has to be known, otherwise θ must be set to zero.

The analysis above yields the following convergence result.

Theorem 6.7.1. *Assume that $\mathcal{X}$ is smooth and s-convex with $s \ge p$, that x_0 is sufficiently close to $x^\dagger$, i.e., $x_0 \in \overline{\mathcal{B}}_\rho^D(x^\dagger)$, that F satisfies* (6.7) *with* (6.6)*, that F and F' are continuous and uniformly bounded in $\overline{\mathcal{B}}_\rho^D(x^\dagger)$, and that* (6.92), (6.96) *hold.*
Then, the iterates $z_{n,k}$ defined by Algorithm 6.7.4 *remain in $\mathcal{B}_\rho^D(x^\dagger)$ and converge to a solution $x^\dagger$ of* (5.17) *subsequentially as $\delta \to 0$ (i.e., there exists a convergent subsequence and the limit of every convergent subsequence is a solution).*
In case of exact data $\delta = 0$, we have subsequential convergence of x_n to a solution of (5.17) *as $n \to \infty$. If additionally a variational inequality* (6.4) *with*

$\nu \in (0,1]$ *and* β *sufficiently small is satisfied, we obtain optimal convergence rates*

$$D_p^{x_0}(x^\dagger, z_{n_*,k_*^{n_*}}) = O(\delta^{\frac{4\nu}{2\nu+1}})\,, \quad \textit{as}\;\; \delta \to 0\,. \tag{6.114}$$

Conclusions

In this short conclusive chapter, we point out the main contributions of the thesis in the area of the regularization of ill-posed problems and present some possible further developments of this work.

The three stopping rules for the Conjugate Gradient method applied to the Normal Equation presented in Chapter 3 produced very promising numerical results in the numerical experiments. In particular, SR2 provided an important insight into the regularizing properties of this method, connecting the well-known theoretical estimates of Chapter 2 with the properties of the Truncated Singular Value Decomposition method.

In the numerical experiments presented in Chapter 4, the new stopping rules defined in Chapter 3 also produced very good numerical results. Of course, these results can be considered only the starting point of a possible future work. Some further developments can be the following:

- applications of the new stopping rules in combination with more sophisticated regularization methods that make use of CGNE (e.g., the Restarted Projected CGNE described in Chapter 3);
- extension of the underlying ideas of the new stopping rules to other regularization methods (e.g., SART, Kaczmarz,...);
- analysis of the speed of the algorithms presented for computing the indices of the new stopping rules, to get improvements.

The theoretical results of Chapter 6, and in particular of Section 6.7, enhanced the regularization theory of Banach spaces. However, they have to

be tested in more serious practical examples. We believe that the new ways to arrest the iteration can indeed improve the performances significantly.
Besides the repetition of the numerical tests of Section 6.6, also two dimensional examples should be considered, as well as a comparison of the inner-outer Newton-Landweber iteration proposed here with the classical Iteratively regularized Gauss-Newton method.
Possible extensions to the case of non-reflexive Banach spaces and further simulations in different ill-posed problems should also be a subject of future research.

Appendix A

Spectral theory in Hilbert spaces

We recall briefly some fundamental results of functional calculus for self-adjoint operators in Hilbert spaces. Details and proofs can be found, e.g. in [2], [17] and [44].

Throughout this section, $\mathcal{X}$ will always denote a Hilbert space. The scalar product in $\mathcal{X}$ will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ and the norm induced by this scalar product will be denoted by $\| \cdot \|_{\mathcal{X}}$.

Definition A.0.1 (**Spectral family**). *A family $\{\mathscr{E}_\lambda\}_{\lambda \in \mathbb{R}}$ of orthogonal projectors in $\mathcal{X}$ is called a spectral family or resolution of the identity if it satisfies the following conditions:*

(i) $\mathscr{E}_\lambda \mathscr{E}_\mu = \mathscr{E}_{\min\{\lambda,\mu\}}$, λ, $\mu \in \mathbb{R}$;

(ii) $\mathscr{E}_{-\infty} = 0$, $\mathscr{E}_{+\infty} = I$, *where* $\mathscr{E}_{\pm\infty} x := \lim_{\lambda \to \pm\infty} \mathscr{E}_\lambda x$, $\forall\, x \in \mathcal{X}$, *and where* I *is the identity map on* $\mathcal{X}$.

(iii) $\mathscr{E}_{\lambda - 0} = \mathscr{E}_\lambda$, *where* $\mathscr{E}_{\lambda-0} x := \lim_{\epsilon \to 0^+} \mathscr{E}_{\lambda - \epsilon} x$, $\forall\, x \in \mathcal{X}$.

Proposition A.0.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then the limit of the Riemann sum*

$$\sum_{i=1}^{n} f(\xi_i) \left(\mathscr{E}_{\lambda_i} - \mathscr{E}_{\lambda_{i-1}} \right) x$$

exists in $\mathcal{X}$ for $|\lambda_i - \lambda_{i-1}| \to 0$, where $-\infty < a = \lambda_0 < ... < \lambda_n = b < +\infty$, $\xi_i \in (\lambda_{i-1}, \lambda_i]$, and is denoted by

$$\int_a^b f(\lambda) d\mathscr{E}_\lambda x.$$

Definition A.0.2. *For any given $x \in \mathcal{X}$ and any continuous function $f : \mathbb{R} \to \mathbb{R}$, the integral $\int_{-\infty}^{+\infty} f(\lambda) d\mathscr{E}_\lambda x$ is defined as the limit, if it exists, of $\int_a^b f(\lambda) d\mathscr{E}_\lambda x$ when $a \to -\infty$ and $b \to +\infty$.*

Since condition (i) in the definition of the spectral family is equivalent to

$$\langle \mathscr{E}_\lambda x, x \rangle \leq \langle \mathscr{E}_\mu x, x \rangle, \quad \text{for all} \;\; x \in \mathcal{X} and \lambda \leq \mu,$$

the function $\lambda \mapsto \langle \mathscr{E}_\lambda x, x \rangle = \|\mathscr{E}_\lambda x\|^2$ is monotonically increasing and due to the condition (ii) in the definition of the spectral family also continuous from the left. Hence it defines a measure on $\mathbb{R}$, denoted by $d\|\mathscr{E}_\lambda x\|^2$. Then the following connection holds:

Proposition A.0.2. *For any given $x \in \mathcal{X}$ and any continuous function $f : \mathbb{R} \to \mathbb{R}$:*

$$\int_{-\infty}^{+\infty} f(\lambda) d\mathscr{E}_\lambda x \quad \textit{exists} \quad \Longleftrightarrow \quad \int_{-\infty}^{+\infty} f^2(\lambda) d\|\mathscr{E}_\lambda x\|^2 < +\infty.$$

Proposition A.0.3. *Let A be a self-adjoint operator in $\mathcal{X}$. Then there exist a unique spectral family $\{\mathscr{E}_\lambda\}_{\lambda \in \mathbb{R}}$, called spectral decomposition of A or spectral family of A, such that*

$$\mathcal{D}(A) = \{x \in \mathcal{X} \mid \int_{-\infty}^{+\infty} \lambda^2 d\|\mathscr{E}_\lambda x\|^2 < +\infty\}$$

and

$$Ax = \int_{-\infty}^{+\infty} \lambda d\mathscr{E}_\lambda x, \quad x \in \mathcal{D}(A).$$

We write:

$$A = \int_{-\infty}^{+\infty} \lambda d\mathscr{E}_\lambda.$$

Definition A.0.3. *Let A be a self-adjoint operator in $\mathcal{X}$ with spectral family $\{\mathscr{E}_\lambda\}_{\lambda\in\mathbb{R}}$ and let f be a measurable function on $\mathbb{R}$ with respect to the measure $d\|\mathscr{E}_\lambda x\|^2$ for all $x \in \mathcal{X}$. Then $f(A)$ is the operator defined by the formula*

$$f(A)x = \int_{-\infty}^{+\infty} f(\lambda) d\mathscr{E}_\lambda x, \quad x \in \mathcal{D}(f(A)),$$

where

$$\mathcal{D}(f(A)) = \{x \in \mathcal{X} \mid \int_{-\infty}^{+\infty} f^2(\lambda) d\|\mathscr{E}_\lambda x\|^2 < +\infty\}.$$

Proposition A.0.4. *Let $\mathcal{M}_0$ be the set of all measurable functions on $\mathbb{R}$ with respect to the measure $d\|\mathscr{E}_\lambda x\|^2$ for all $x \in \mathcal{X}$ (in particular, piecewise continuous functions lie in $\mathcal{M}_0$). Let A be a self-adjoint operator in $\mathcal{X}$ with spectral family $\{\mathscr{E}_\lambda\}_{\lambda\in\mathbb{R}}$ and let f, $g \in \mathcal{M}_0$.*

(i) If $x \in \mathcal{D}(f(A))$ and $z \in \mathcal{D}(g(A))$, then

$$\langle f(A)x, g(A)z\rangle = \int_{-\infty}^{+\infty} f(\lambda)g(\lambda) d\langle \mathscr{E}_\lambda x, z\rangle.$$

(ii) If $x \in \mathcal{D}(f(A))$, then $f(A)x \in \mathcal{D}(g(A))$ if and only if $x \in \mathcal{D}((gf)(A))$. Furthermore,

$$g(A)f(A)x = (gf)(A)x.$$

(iii) If $\mathcal{D}(f(A))$ is dense in $\mathcal{X}$, then $f(A)$ is self-adjoint.

(iv) $f(A)$ commutes with $\mathscr{E}_\lambda$ for all $\lambda \in \mathbb{R}$.

Proposition A.0.5. *Let A be a self-adjoint operator in $\mathcal{X}$ with spectral family $\{\mathscr{E}_\lambda\}_{\lambda\in\mathbb{R}}$.*

(i) λ_0 lies in the spectrum of A if and only if $\mathscr{E}_{\lambda_0} \neq \mathscr{E}_{\lambda_0+\epsilon}$ for all $\epsilon > 0$.

(ii) λ_0 is an eigenvalue of A if and only if $\mathscr{E}_{\lambda_0} \neq \mathscr{E}_{\lambda_0+0} = \lim_{\epsilon\to 0} \mathscr{E}_{\lambda_0+\epsilon}$. The corresponding eigenspace is given by $(\mathscr{E}_{\lambda_0+0} - \mathscr{E}_{\lambda_0})(\mathcal{X})$.

At last, we observe that if A is a linear bounded operator, then the operator A^*A is a linear, bounded, self-adjoint and semi-positive definite operator. Let $\{E_\lambda\}$ be the spectral family of A^*A and let $\mathcal{M}$ be the set of all measurable functions on $\mathbb{R}$ with respect to the measure $d\|E_\lambda x\|^2$ for all $x \in \mathcal{X}$. Then, for all $f \in \mathcal{M}$,

$$\int_{-\infty}^{+\infty} f(\lambda) dE_\lambda x = \int_0^{\|A\|^2} f(\lambda) dE_\lambda x = \lim_{\epsilon \to 0^+} \int_0^{\|A\|^2+\epsilon} f(\lambda) dE_\lambda x.$$

Hence, the function f can be restricted to the interval $[0, \|A\|^2 + \epsilon]$ for some $\epsilon > 0$.

Appendix B

Approximation of a finite set of data with cubic B-splines

B.1 B-splines

Let $[a, b]$ be a compact interval of $\mathbb{R}$, let

$$\Delta = \{a = t_0 < t_1 < ... < t_k < t_{k+1} = b\} \tag{B.1}$$

be a partition of $[a, b]$ and let m be an integer, $m > 1$. Then the space $\mathfrak{S}_m(\Delta)$ of *polynomial splines with simple knots of order m on Δ* is the space of all function $s = s(t)$ for which there exist $k+1$ polynomials $s_0, s_1, ..., s_k$ of degree $\leq m-1$ such that

(i) $s(t) = s_j(t) \quad \text{for } t_j \leq t \leq t_{j+1}, \quad j = 0, ..., k;$

(ii) $\frac{d^i}{dt^i} s_{j-1}(t_j) = \frac{d^i}{dt^i} s_j(t_j), \quad \text{for } i = 0, ..., m-2, \quad j = 1, ..., k.$

The points t_j are called the *knots* of the spline and $t_1, ..., t_k$ are the *inner knots*.

It is well known (cf. e.g. [86]) that $\mathfrak{S}_m(\Delta)$ is a vector space of dimension $m + k$. A base of $\mathfrak{S}_m(\Delta)$ with good computational properties is given by the normalized B-splines.

An *extended partition of* $[a, b]$ *associated to* $\mathfrak{S}_m(\Delta)$ is a sequence of points

$\Delta^* = \{\tilde{t}_{-m+1} \leq ... \leq \tilde{t}_{k+m}\}$ such that $\tilde{t}_i = t_i$ for every $i = 0, ..., k+1$. There are different possible choices of the extended partition Δ^*. Here, we shall consider the choice

$$\tilde{t}_{-m+1} = ... = \tilde{t}_0 = a, \quad \tilde{t}_{k+1} = \tilde{t}_{k+2} = ... = \tilde{t}_{k+m}. \tag{B.2}$$

The *normalized B-splines on* Δ^* are the functions $\{N_{j,m}\}_{j=-m+1,...,k}$ defined recursively in the following way:

$$N_{j,1}(t) = \begin{cases} 1, & \text{for } \tilde{t}_j \leq t \leq \tilde{t}_{j+1}, \\ 0, & \text{elsewhere;} \end{cases} \tag{B.3}$$

$$N_{j,h}(t) = \begin{cases} \frac{t-\tilde{t}_j}{\tilde{t}_{j+h-1}-\tilde{t}_j} N_{j,h-1}(t) + \frac{\tilde{t}_j - t}{\tilde{t}_{j+h}-\tilde{t}_{j+1}} N_{j+1,h-1}(t), & \text{for } \tilde{t}_j \neq \tilde{t}_{j+h}, \\ 0, & \text{elsewhere} \end{cases} \tag{B.4}$$

for $h = 2, ..., m$. The cases $0/0$ must be interpreted as 0.
The functions $N_{j,h}$ have the following well known properties:

(1) Local support: $N_{j,m}(t) = 0$, $\forall\, t \notin [\tilde{t}_j, \tilde{t}_{j+m})$ if $\tilde{t}_j < \tilde{t}_{j+m}$;

(2) Non negativity: $N_{j,m}(t) > 0$, $\forall\, t \in (\tilde{t}_j, \tilde{t}_{j+m})$, $\tilde{t}_j < \tilde{t}_{j+m}$;

(3) Partition of unity: $\sum_{j=-m+1}^{k} N_{j,m}(t) = 1$, $\forall\, t \in [a, b]$.

B.2 Data approximation

Let now $(\lambda_1, \mu_1), ..., (\lambda_n, \mu_n)$, $n \in \mathbb{N}$, $n \geq m+k$, λ_j and $\mu_j \in \mathbb{R}$ such that

$$a = \lambda_1 < ... < \lambda_n = b \tag{B.5}$$

be a given set of data. We want to find a spline $s(t) \in \mathfrak{S}_m(\Delta)$,

$$s(t) = \sum_{j=-m+1}^{k} c_j N_{j,m}(t),$$

that minimizes the least-squares functional

$$\sum_{l=1}^{n} |s(\lambda_l) - \mu_l|^2 \tag{B.6}$$

on $\mathfrak{S}_m(\Delta)$. Simple calculations show that the solutions of this minimization problem are the solutions of the overdetermined linear system

$$\sum_{j=-m+1}^{k} c_j \sum_{l=1}^{n} N_{i,m}(\lambda_l) N_{j,m}(\lambda_l) = \sum_{l=1}^{n} \mu_l N_{i,m}(\lambda_l), \quad i = -m+1, ..., k. \tag{B.7}$$

Denoting with H the matrix of the normalized B-splines on the approximation points:

$$H = \{h_{l,j}\} = \{N_{j,m}(\lambda_l)\}, \quad l = 1, ..., n; \quad j = -m+1, ..., k, \tag{B.8}$$

we can rewrite $(B.7)$ in the form

$$H^* H \mathbf{c} = H^* \boldsymbol{\mu}, \tag{B.9}$$

where $\mathbf{c}$ and $\boldsymbol{\mu}$ are the column vectors of the c_j and of the μ_l respectively. It can be shown that the system has a unique solution if Δ^* satisfies the so called *Schönberg-Whitney conditions*:

Theorem B.2.1. *The matrix* $\mathbf{H}$ *has maximal rank if there exists a sequence of indices* $1 \leq j_1 \leq ... \leq j_{m+k} \leq n$ *such that*

$$\tilde{t}_i < \lambda_{j_i} < \tilde{t}_{i+m}, \quad i = -m+1, ..., k, \tag{B.10}$$

where the $\tilde{t}_i$ *are the knots of the extended partition* Δ^**.*

With equidistant inner knots $t_i = a + i\frac{(b-a)}{k+1}$ and the particular choice $(B.2)$, it is easy to see that the Schönberg-Whitney conditions are satisfied for every $k \leq n - m$: for example, if $k = n - m$, $j_i = i$ for every $i = 1, ..., n$.

Appendix C

The algorithms

In this section of the appendix we present the main algorithms used in the thesis. All numerical experiments have been executed on a Pentium IV PC using Matlab 7.11.0 $R2010b$.

C.1 Test problems from P. C. Hansen's Regularization Tools

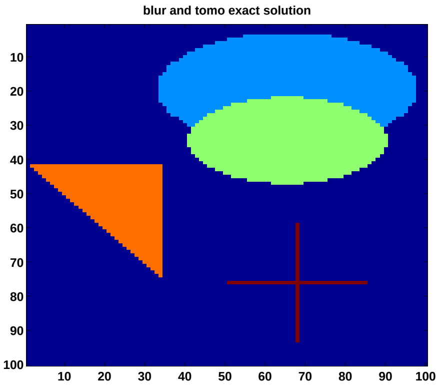


Figure C.1: Exact solution of the test problems tomo and blur from P.C. Hansen's Regularization Tools.

Many test problems used in this thesis are taken from P. C. Hansen's

Regularization Tools. This is a software package that consists of a collection of documented Matlab functions for analysis and solution of discrete ill-posed problems. The package and the underlying theory is published in [35] and the most recent version of the package, which is updated regularly, is described in [41]. The package can be downloaded directly from the page

$$\mathsf{http : //www2.imm.dtu.dk/\ pch/Regutools/}, \tag{C.1}$$

where a complete manual is also available.

All the algorithms of the thesis referring to [35] or [41] are taken from the version 4.1 (for Matlab version 7.3). Below, we describe briefly the files that have been used in the thesis.

More details on these functions such as the synopsis, the input and output arguments, the underlying integral equation and the references can be found in the manual of the Regularization Tools [35] in a pdf format at the web page (C.1).

We consider 10 different very famous test problems:

- baart, deriv2, foxgood, gravity, heat, i_laplace, phillips, shaw generate the square matrix $\mathbf{A}$, the exact solution $\mathbf{x}$ and the exact right-hand side $\mathbf{b}$ of a discrete ill-posed problem, typically arising from a discretization of an integral equation of the first kind. The dimension N of the problem is the main input argument of these functions. In some cases it is possible to choose between 2 or 3 different exact solutions. Of course, $\mathbf{A}$ is always very ill-conditioned, but in some problems the eigenvalues decrease more quickly than in others.

- blur and tomo generate the square matrix $\mathbf{A}$, the exact solution $\mathbf{f}$ and the exact right-hand side $\mathbf{g}$ a 2D image reconstruction problem. In both cases, the vector $\mathbf{f}$ is a columnwise stacked version of a simple test image (cf. Figure C.1) with $J \times J$ pixels and J is the fundamental input argument of the function. In the blur problem, the matrix $\mathbf{A}$ is a symmetric $J^2 \times J^2$ doubly Toeplitz matrix, stored in sparse format associated to an atmospheric turbulence blur and $\mathbf{g} := \mathbf{Af}$ is the blurred image. By modifying the third input argument of the function which is set by default equal to 0.7, it is possible to control the shape of the Gaussian point spread function associated to $\mathbf{A}$. In the tomo problem, the matrix $\mathbf{A}$ arises from the discretization of a simple 2D tomography model. If no additional input arguments are used, $\mathbf{A}$ is a square matrix of dimension $J^2 \times J^2$ as in the case of blur.

C.2 Conjugate gradient type methods algorithms

Two different functions for the implementation of CGNE can be found in the Regularization tools: cgls and lsqr_b. cgls is a direct implementation of algorithm 3, lsqr_b is an equivalent implementation of the same algorithm based on Lanczos bidiagonalization (cf. [73] and [36]).

Both routines require the matrix of the system $\mathbf{A}$, a data vector $\mathbf{b}$ and an integer k_{MAX} corresponding to the number of CGNE steps to be performed and return all k_{MAX} solutions, stored as columns of the matrix $\mathbf{X}$. The corresponding solution norms and residual norms are returned in $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$, respectively. If the additional parameter reorth is set equal to 1, then the routines perform a reorthogonalization of the normal equation residual vectors.

To compare CGNE and CGME, a new routine cgne_cgme has been generated based on Algorithm 6. This function is similar to cgls and lsqr_b, but returns also the solutions, the residual norms and the solution norms of the CGME iterates.

A modified version of cgls, cgls_deb has been used for the image deblurring problems to avoid forming the matrix $\mathbf{A}$. In the cgls_deb algorithm, the matrix $\mathbf{A}$ is replaced by the PSF, the data vector $\mathbf{g}$ is replaced by its corresponding image and all matrix-vectors multiplications are replaced by the corresponding 2D convolutions as in Section 3.6, formula (3.41). As a consequence, the synopsis of this function is different from the others:

$$[\mathsf{f}, \mathsf{rho}, \mathsf{eta}] = \mathsf{cgls\_deb}(\mathsf{h}, \mathsf{g}, \mathsf{k}); \tag{C.2}$$

here the input arguments are the <u>matrices</u> of the PSF h and of the blurred image g and the number of iterations k_{MAX}. The output is a 3D matrix f such that for every $k = 1, ..., k_{MAX}$ the Matlab command

$$\mathsf{f}(:, :, \mathsf{k})$$

gives the k-th iterate of the algorithm in the form of an image.

At last, a new routine cgn2 similar to cgls and lsqr_b has been created to implement the conjugate gradient type method with parameter $n = 2$ (cf. Algorithm 7 from Chapter 2).

We emphasize that in the tests where a visualization of the reconstructed solutions was not necessary, all these functions were used without generating the matrix of the reconstructed solutions, but instead overwriting at each step the new iterate of CGNE on the old one, in order to spare memory and time.

C.3 The routine data_approx

In the notations of Appendix B, the routine data_approx generates an approximation $\{(\lambda_1, \tilde{\mu}_1), ..., (\lambda_n, \tilde{\mu}_n)\}$ of the data set $\{(\lambda_1, \mu_1), ..., (\lambda_n, \mu_n)\}$ according to the following scheme (valid for $n \geq 5$):

Step 1: Fix $m = 4$ and the number of inner knots k according to the dimension of the problem: if $n \leq 5000$ then $k = \lfloor n/4 \rfloor$, if $n > 5000$ then $k = \lfloor n/50 \rfloor$. Construct the partition $\Delta = \{t_0 < t_1 < ... < t_k < t_{k+1}\}$ such that $t_0 = \lambda_1$, $t_{k+1} = \lambda_n$ and $t_i = \lambda_1 + i\frac{\lambda_n - \lambda_1}{k+1}$, then choose the extended partition Δ^* according to $(B.2)$.

Step 2: Construct the matrix $\mathbf{H}$ of the normalized B-splines of order m on the approximation points $\lambda_1, ..., \lambda_n$ relative to the extended partition Δ^* according to $(B.3)$, $(B.4)$ and $(B.8)$.

Step 3: Find the unique solution $\mathbf{c}$ of the linear system $(B.9)$.

Step 4: Evaluate the spline $s(t) = \sum_{i=-m+1}^{k} c_i N_{i,m}(t)$ on the approximation points λ_j denoting with $\tilde{\mu}_j = s(\lambda_j)$ the corresponding results.

C.4 The routine mod_min_max

This section describes the Matlab function implemented for the computation of the index p that divides the vector of the SVD (Fourier) coefficients $|\mathbf{u}_i^* \mathbf{b}^\delta|$, $i = 1, ..., m$, associated to the (perturbed) linear system $\mathbf{Ax} = \mathbf{b}^\delta$, into a vector of *low frequency components*, constituted by its first p entries and a vector of *high frequency components*, constituted by its last $m - p$ entries.

The routine, denoted with the name mod_min_max, is a variation of the *Min_Max Rule* proposed in [101] and requires the singular values $\lambda_1, ..., \lambda_N$ of $\mathbf{A}$.

Suppose for simplicity $m = N$ and let φ_i denote the ratios $|\mathbf{u}_i^* \mathbf{b}^\delta| / \lambda_i$ for $i = 1, ..., N$. Separate the set $\Psi = \{\varphi_1, ..., \varphi_N\}$ into 2 sets

$$\Psi_1 := \{\varphi_i \mid \varphi_i > \lambda_i\} \qquad \Psi_2 := \{\varphi_i \mid \varphi_i \leq \lambda_i\} \tag{C.3}$$

and let N_1 and N_2 be the number of elements in Ψ_1 and Ψ_2 respectively. Then:

(i) If $N_2 = 0$ or $\lambda_N < 10^{-13}$, calculate an approximation $\tilde{\Psi}$ of Ψ by means of cubic B-splines with the routine data_approx of the Section $C.3$ of the Appendix and choose p as the index corresponding to the minimal value in $\tilde{\Psi}$.

(ii) Otherwise, consider the first of the last 5 elements in Ψ_2 such that $\varphi_{i_j+1} \notin \Psi_2$ and choose p as the corresponding index.

When the smallest singular value is close to the machine epsilon or the set Ψ_2 is empty, then Ψ can be used to determine the regularization index. In this case the data noise is assumed to be predominant with respect to the model errors, so the minimum of the sequence φ_i should correspond to the index $i_{\mathbf{b}^\delta}$. Moreover, the data approximation is used to avoid the presence of possible outliers. A typical case is shown in Figure 3.5 with the shaw test problem.

In the second case of the modified Min_Max Rule the model errors are predominant and the greatest indices in Ψ_2 are included in the TSVD provided that they are contiguous (i.e. the successive element does not belong to Ψ_1). This situation is shown in the picture on the right of Figure 3.5 obtained with the phillips test problem.

C.5 Data and files for image deblurring

The experiments on image deblurring performed in Section 3.6 make use of the following files.

- The file psfgauss is taken the *HNO Functions*, a small Matlab package that implements the image deblurring algorithms presented in [38]. The package is available at the web page

  http : //www2.imm.dtu.dk/ pch/HNO/.

  For a given integer J and a fixed number stdev, representing the deviations of the Gaussian along the vertical and horizontal directions, the Matlab command

  $$[\mathsf{h}, \mathsf{center}] = \mathsf{psfGauss}(\mathsf{J}, \mathsf{stdev}); \tag{C.4}$$

  generates the PSF matrix h of dimension $J \times J$ and the center of the PSF.

- The file im_blurring generates a test problem for image deblurring. A gray-scale image is read with the Matlab function im_read. Then a Gaussian PSF is generated by means of the function psfgauss and the image is blurred according to the forumla (3.41) from Section 3.6. At last, Gaussian white noise is added to the blurred image to obtain the perturbed data of the problem.

- The function fou_coeff plots the Fourier coefficients of an image deblurring test problem. Given the (perturbed) image g and the PSF h, it returns the Fourier coefficients, the singular values of the BCCB matrix $\mathbf{A}$ corresponding to h and the index p computed by the function mod_min_max of Section C.4 of the Appendix.

C.6 Data and files for the tomographic problems

The numerical experiments on the tomographic problems described in Section 4.6 make use of the files paralleltomo, fanbeamtomo and seismictomo from P.C. Hansen's *Air Tools*. This is a Matlab software package for tomographic reconstruction (and other imaging problems) consisting of a number of algebraic iterative reconstruction methods. The package, described in the paper [40], can be downloaded at the web page

http : //www2.imm.dtu.dk/ pcha/AIRtools/.

For a fixed integer $J > 0$, the Matlab command

$$[\mathsf{A}, \mathsf{g}, \mathsf{f}] = \mathsf{paralleltomo}(\mathsf{J}) \tag{C.5}$$

generates the exact solution $\mathbf{f}$, the matrix $\mathbf{A}$ and the exact data $\mathbf{g} = \mathbf{Af}$ of a two dimensional tomographic test problem with parallel X-rays. The input argument J is the size of the exact solution $\mathbf{f}$ of the system. Therefore, the matrix $\mathbf{A}$ has $N := J^2$ columns. The number of rows of $\mathbf{A}$ is given by the number of total rays for each angle l_0 multiplied by the number of angles j_0. Consequently, the sinogram $\mathbf{G}$ corresponding to the exact data $\mathbf{g}$ is a matrix with l_0 rows and j_0 columns.

The functions fanbeamtomo and seismictomo generate the test problem in a similar way. We emphasize that for each problem the dimensions of the sinogram are different. If these values are not specified, they are set by default. In particular:

- paralleltomo: $l_0 = \mathsf{round}(\sqrt{2}J)$, $j_0 = 180$;
- fanbeamtomo: $l_0 = \mathsf{round}(\sqrt{2}J)$, $j_0 = 360$;
- seismictomo: $l_0 = \mathsf{2J}$, $j_0 = J$.

Appendix D

CGNE and rounding errors

In literature, there are a number of mathematically equivalent implementations of CGNE and of the other methods discussed above. Many authors suggest LSQR (cf. [73] and [36]), which is an equivalent implementation of CGNE based on Lanczos bidiagonalization.

The principal problem with any of these methods is the loss of orthogonality in the residuals due to finite precision arithmetic. The orthogonality can be maintained by reorthogonalization techniques that are significantly more expensive and require a larger number of intermediate vectors (cf. e.g., [18]). In the literature the influence of round-off errors on conjugate gradient type methods has been studied mainly for well-posed problems.

In [27] Hanke comments on the ill-posed case that the reorthogonalization techniques did not improve the optimal accuracy in the case he considered. Our numerical experiments confirm that the sequence of the relative errors $\|\mathbf{x}^\dagger - \mathbf{z}_k^\delta\|$ does not change significantly for k smaller than the optimal stopping index.

However, even small differences in the computation of the residual norms and of the norm of the solutions $\|\mathbf{z}_k^\delta\|$ may affect seriously the results presented in the following sections of Chapter 3, especially in the 1D examples. Therefore, in these sections the routine lsqr_b from [35], with the parameter reorth $= 1$ was preferred to the other routines in the implementation of the

CGNE algorithm.

In the other cases we proceeded as follows:

- In the tests of Chapter 2, to compare the results obtained by CGNE and CGME we implemented Algorithm 6, generating a new routine cgne_cgme specific for this case;
- In the numerical experiments of Chapter 3 on image deblurring we used the routine cgls_deb;
- In the numerical experiments of Chapter 4 we simply used cgls without reorthogonalization.

Bibliography

[1] M. ABRAMOWITZ & I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] N.I. AKHIEZER & I.M. GLAZMAN *Theory of linear operators in Hilbert spaces*, Pitman, 1981.

[3] C. T. H. BAKER, *The Numerical Treatment of Integral Equations*, Clarendon Press, Oxford, UK, 1977.

[4] M. BERTERO & P. BOCCACCI, *Introduction to Inverse Problems in Imaging*, IOP Publishing, Bristol, 1998.

[5] D. CALVETTI, G. LANDI, L. REICHEL & F. SGALLARI: *Nonnegativity and iterative methods for ill-posed problems*, Inverse Problems **20** (2004), 1747-1758.

[6] C. CLASON, B. JIN, K. KUNISCH, *A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration*, SIAM J. Imaging Sci., **3** (2010), 199-231.

[7] F. COLONIUS & K. KUNISCH *Output least squares stability in elliptic systems*, Appl. Math. Opt. **19** (1989), 33-63.

[8] J. W. DANIEL, *The Conjugate Gradient Method for Linear and Nonlinear Operator Equations*, SIAM J. Numer. Anal., **4** (1967), 10-26.

[9] J. L. CASTELLANOS, S. GOMEZ & V. GUERRA *The triangle method for finding the corner of the L-curve* Appl. Numer. Math. **43** (2002) 359-373.

[10] T. F. CHAN & J. SHEN, *Image Processing and Analysis: variational, PDE, wavelet, and stochastic methods*, SIAM, Philadelphia, 2005.

[11] P. E. DANIELSSON, P. EDHOLM & M. SEGER, *Towards exact 3D-reconstruction for helical cone-beam scanning of long objects. A new detector arrangement and a new completeness condition*, in Proceedings of the 1997 International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, edited by D. W. Townsend & P. E. Kinahan, Pittsburgh, 1997.

[12] P. J. DAVIS, *Circulant matrices*, Wiley, New York, 1979.

[13] M. K. DAVISON *The ill-conditioned nature of the limited angle tomography problem*, SIAM J. Appl. Math. **43**, 428-448.

[14] L. M. DELVES & J. L. MOHAMED, *Computational Methods for Integral Equations*, Cambridge University Press, Cambridge, UK, 1985.

[15] H. W. ENGL & H. GFRERER, *A Posteriori Parameter Choice for General Regularization Methods for Solving Linear Ill-posed Problems*, Appl. Numer. Math. **7** (1988) 395-417.

[16] H. W. ENGL, K. KUNISCH & A. NEUBAUER, *Convergence rates for Tikhonov regularization of non-linear ill-posed problems*, Inverse Problems, **5** (1989), 523-540.

[17] H. W. ENGL, M. HANKE & A. NEUBAUER, *Regularization of Inverse Problems*, Kluwer Academic Publishers, 1996.

[18] G. H. GOLUB & C. F. VAN LOAN, *Matrix Computations*, The John Hopkins University Press, Baltimore, London, 1989.

[19] F. GONZALES, *Notes on Integral Geometry and Harmonic Analysis*, COE Lecture Note Vol. 24, Kyushu University 2010.

[20] C. W. GROETSCH, *Elements of Applicable and Functional Analysis*, Dekker, New York, 1980.

[21] C. W. GROETSCH, *Generalized Inverses of Linear Operators: Representation and Approximation*, Dekker, New York, 1977.

[22] C. W. GROETSCH, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Boston, 1984.

[23] U. HAEMARIK & U. TAUTENHAHN, *On the Monotone Error Rule for Parameter Choice in Iterative and Continuous Regularization Methods*, BIT Numer. Math. **41**,5 (2001), 1029-1038.

[24] U. HAEMARIK & R. PALM, *On Rules for Stopping the Conjugate Gradient Type Methods in Ill-posed Problems*, Math. Model. Anal. **12**,1 (2007) 61-70.

[25] U. HAEMARIK & R. PALM, *Comparison of Stopping Rules in Conjugate Gradient Type Methods for Solving Ill-posed problems*, in Proceedings of the 10-th International Conference MMA 2005 & CMAM 2, Trakai. Technika (2005) 285-291.

[26] U. HAEMARIK, R. PALM & T. RAUS *Comparison of Parameter Choices in Regularization Algorithms in Case of Different Information about Noise Level*, Calcolo **48**,1 (2011) 47-59.

[27] M. HANKE, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Longman House, Harlow, 1995.

[28] M. HANKE, *Limitations of the L-curve method in ill-posed problems,* BIT **36** (1996), 287-301.

[29] M. HANKE & J. G. NAGY, *Restoration of Atmospherically Blurred Images by Symmetric Indefinite Conjugate Gradient Techniques*, Inverse Problems, **12** (1996), 157-173.

[30] M. HANKE, A. NEUBAUER, & O. SCHERZER, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., **72** (1995) 21-37.

[31] M. HANKE, *A regularization Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Problems, **13** (1997), 79-95.

[32] P. C. HANSEN, *The discrete Picard condition for discrete ill-posed problems*, BIT **30** (1990), 658-672.

[33] P. C. HANSEN, *Analysis of the discrete ill-posed problems by means of the L-curve*, SIAM Review, **34** (1992), 561-580.

[34] P. C. HANSEN & D. P. O'LEARY, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., **14** (1993), 1487-1503.

[35] P. C. HANSEN, *Regularization Tools: A Matlab Package for Analysis and Solution of Discrete Ill-posed Problems* (version 4.1), Numerical Algorithms, **6** (1994) 1-35.

[36] P.C. HANSEN, *Rank Deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.

[37] P. C. HANSEN, T. K. JENSEN & G. RODRIGUEZ, *An adaptive pruning algorithm for the discrete L-curve criterion*, J. Comput. Appl. Math. **198** (2006), 483-492.

[38] P. C. HANSEN, J. G. NAGY, D. P. O'LEARY, *Deblurring Images, Matrices, Spectra and Filtering*, SIAM, Philadelphia, 2006.

[39] P. C. HANSEN, M. KILMER, R. H. KJELDSEN, *Exploiting residual information in the parameter choice for discrete ill-posed problems*, BIT Numerical Mathematics, **46**,1 (2006), 41-59.

[40] P. C. HANSEN & M. SAXILD-HANSEN, *AIR Tools - A MATLAB Package of Algebraic Iterative Reconstruction Methods*, Journal of Computational and Applied Mathematics, **236** (2012), 2167-2178.

[41] P. C. HANSEN, *Regularization Tools Version 4.0 for Matlab 7.3, Numerical Algorithms*, **46** (2007), 189-194.

[42] T. HEIN & B. HOFMANN, *Approximate source conditions for nonlinear ill-posed problems – chances and limitations*, Inverse Problems, **25** (2009), 035003 (16pp).

[43] S. HELGASON, *The Radon Transform*, 2nd. Ed., Birkhäuser Progress in Math., 1999.

[44] G. HELMBERG, *Introduction to Spectral Theory in Hilbert Spaces*, North Holland, Amsterdam, 1969.

[45] M. R. HESTENES & E. STIEFEL, *Methods of Conjugate Gradients for Solving Linear Systems*, J. Research Nat. Bur. Standards **49** (1952), 409-436.

[46] B. HOFMANN, B. KALTENBACHER, C. PÖSCHL & O. SCHERZER, *A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators*, Inverse Problems, **23**,3 (2007), 987-1010.

[47] T. HOHAGE, *Iterative methods in inverse obstacle scattering: regularization theory of linear and nonlinear exponentially ill-posed problems*, PhD Thesis University of Linz, Austria, 1999.

[48] P. A. JANSSON *Deconvolution of Images and Spectra*, San Diego, Academic 1997.

[49] Q. JIN, *Inexact Newton-Landweber iteration for solving nonlinear inverse problems in Banach spaces*, Inverse Problems **28** (2012) 065002, 14pp.

[50] H. HU & J. ZHANG, *Exact Weighted-FBP Algorithm for Three-Orthogonal-Circular Scanning Reconstruction*, Sensors, **9** (2009), 4606-4614.

[51] B. KALTENBACHER & B. HOFMANN, *Convergence rates for the iteratively regularized gauss-newton method in Banach spaces*, Inverse Problems, **26**,3 (2010) 035007 21pp.

[52] B. KALTENBACHER & A. NEUBAUER, *Convergence of projected iterative regularization methods for nonlinear problems with smooth solutions*, Inverse Problems, **22** (2006), 1105-1119.

[53] B. KALTENBACHER, A. NEUBAUER, & O. SCHERZER, *Iterative Regularization Methods for Nonlinear Ill-posed Problems*, de Gruyter, 2007.

[54] B. KALTENBACHER & I. TOMBA, *Convergence rates for an iteratively regularized Newton-Landweber iteration in Banach space*, Inverse Problems **29** (2013) 025010.

[55] B. KALTENBACHER, *Convergence rates for the iteratively regularized Landweber iteration in Banach space*, Proceedings of the 25th IFIP TC7 Conference on System Modeling and Optimization, Springer, 2013, to appear.

[56] A. KATSEVICH, *Analysis of an exact inversion formula for spiral cone-beam CT*, Physics in Medicine and Biology, **47** (2002) 2583-2598.

[57] A. KATSEVICH, *Theoretically exact filtered backprojection-type inversion algorithm for spiral CT*, SIAM Journal of Applied Mathemathics, **62** (2002), 2012-2026.

[58] A. KATSEVICH, *An improved exact filtered backprojection algorithm for spiral computed tomography.* Advances in Applied Mathematics, **32** (2004), 681-697.

[59] C. T. KELLEY, *Iterative Methods for Linear and Nonlinear Equations*, Society for Industrial and Applied Mathematics, Philadelphia 1995.

[60] M. KILMER & G. W. STEWART, *Iterative Regularization and MINRES*, SIAM J. Matrix Anal. Appl., **21**,2 (1999) 613-628.

[61] A. KIRSCH, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer-Verlag, New York, 1996.

[62] R. KRESS, *Linear Integral Equations*, Applied Mathematical Sciences vol. 82, Second Edition, Springer-Verlag, New York, 1999.

[63] L. LANDWEBER, *An Iteration Formula for Fredholm Integral Equation of the First Kind*, Amer. J. Math. **73** (1951), 615-624.

[64] A. K. LOUIS, *Orthogonal Function Series Expansion and the Null Space of the Radon Transform*, SIAM J. Math. Anal. **15** (1984), 621-633.

[65] P. MAASS, *The X-Ray Transform: Singular Value Decomposition and Resolution*, Inverse Problems **3** (1987), 729-741.

[66] T. M. MACROBERT, *Spherical Harmonics: An Elementary Treatise on Harmonic Functions with Applications*, Pergamon Press, 1967.

[67] V. A. MOROZOV, *On the Solution of Functional Equations by the Method of Regularization*, Soviet Math. Dokl., **7** (1966) 414-417.

[68] F. NATTERER, *The Mathematics of Computerized Tomography*, J. Wiley, B.G. Teubner, New York, Leipzig, 1986.

[69] F. NATTERER, and F. WÜBBELING *Mathematical Methods in Image Reconstruction*, Cambridge University Press, SIAM 2001.

[70] A. S. NEMIROVSKII, *The Regularization Properties of the Adjoint Gradient Method in Ill-posed Problems*, USSR Comput. Math. and Math. Phys., **26**,2 (1986) 7-16.

[71] A. NEUBAUER, *Tikhonov-Regularization of Ill-Posed Linear Operator Equations on Closed Convex Sets*, PhD thesis, Johannes Kepler Universität Linz November 1985, appeared in Verlag der Wissenschaftlichen Gesellschaften Österreich, Wien, 1986.

[72] A. V. OPPENHEIM & R. W. SCHAFER, *Discrete-Time Signal Processing*, Prentice Hall Inc., New Jersey, 1989.

[73] C. C. PAIGE & M. A. SAUNDERS, *LSQR: an algorithm for sparse linear equations and sparse least squares*, ACM trans. Math. Software, **8** (1982), 43-71.

[74] R. PLATO, *Optimal algorithms for linear ill-posed problems yield regularization methods*, Numer. Funct. Anal. Optim., **11** (1990), 111-118.

[75] J. RADON, *Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten*, Berichte Sächsische Akademie der Wissenschaften, Math.-Phys., Kl, **69** (1917), 262-267.

[76] L. REICHEL & H. SADOK, *A New L-Curve for Ill-Posed Problems*, J. Comput. Appl. Math., **219** (2008) 493-508.

[77] L. REICHEL & G. RODRIGUEZ, *Old and new parameter choice rules for discrete ill-posed problems*, Num. Alg., 2012, DOI 10.1007/s11075-012-9612-8.

[78] A. RIEDER, *On convergence rates of inexact Newton regularizations*, Numer. Math. **88** (2001), 347-365.

[79] A. RIEDER, *Inexact Newton regularization using conjugate gradients as inner iteration*, SIAM J. Numer. Anal. **43** (2005) 604-622.

[80] W. RUDIN, *Functional Analysis*, Mc Graw-Hill Book Company, 1973.

[81] O. SCHERZER, *A modified Landweber iteration for solving parameter estimation problems*, Appl. Math. Opt., **68** (1998), 38-45.

[82] T. SCHUSTER, B. KALTENBACHER, B. HOFMANN, & K. KAZIMIERSKI, *Regularization Methods in Banach Spaces* De Gruyter, Berlin, New York, 2012.

[83] R. T. SEELEY, *Spherical Harmonics*, Amer. Math. Monthly **73** (1966), 115-121.

[84] L. A. SHEPP & B.F. LOGAN *The Fourier Reconstruction of a Head Section*, IEEE Trans. Nuclear Sci. **NS-21** (1974), 21-43.

[85] M. A. SHUBIN, *Pseudodifferential Operators and Spectral Theory*, Second Edition, Springer-Verlag 2001.

[86] L. L. SHUMAKER, *Spline functions basic theory*, John Wiley and Sons 1981.

[87] T. STEIHAUG, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Num. Anal., **20** (1983), 626-637.

[88] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., Vol. 23, Amer. Math. Soc., Providence, Rhode Island, 1975.

[89] K. C. TAM, S. SAMARASEKERA & F. SAUER, *Exact cone-beam CT with a spiral scan*, Physics in Medicine and Biology, **43** (1998), 1015-1024.

[90] A. N. TIKHONOV, *Regularization of Incorrectly Posed Problems*, Soviet Math. Dokl., **4** (1963) 1624-1627.

[91] A. N. TIKHONOV, *Solution of Incorrectly Formulated Problems and the Regularization Method*, Soviet Math. Dokl., **4** (1963), 1035-1038.

[92] F. TREVES, *Basic Linear Partial Differential Equations*, Dover, New York, 2006.

[93] H. K. TUY, *An inversion formula for cone-beam reconstruction*, SIAM J. Appl. Math., **43** (1983), 546-552.

[94] C. F. VAN LOAN, *Computational frameworks for the fast Fourier Transform*, SIAM, Philadelphia, 1992.

[95] C. R. VOGEL, *Non Convergence of the L-curve Regularization Parameter Selection Method*, Inverse Problems, **12** (1996), 535-547.

[96] C. R. VOGEL, *Computational Methods for Inverse Problems*, SIAM Frontiers in Applied Mathematics, 2002.

[97] A. J. WUNDERLICH, *The Katsevich Inversion Formula for Cone-Beam Computed Tomography*, Master Thesis, 2006.

[98] J. YANG, Q. KONG, T. ZHOU & M. JIANG, *Cone-beam cover method: an approach to performing backprojection in Katsevich's exact algorithm for spiral cone-beam CT*, Journal of X-Ray Science and Technology, **12** (2004), 199-214.

[99] Z. B. XU & G. F. ROACH, *Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces*, Journal of Mathematical Analysis and Applications, **157**,1 (1991), 189-210.

[100] S. S. YOUNG, R. G. DRIGGERS & E. L. JACOBS, *Image Deblurring*, Boston, Artech House Publishers, 2008.

[101] F. ZAMA, *Computation of Regularization Parameters using the Fourier Coefficients*, AMS acta, 2009, www.amsacta.unibo.it.

Acknowledgements

First of all, I would like to thank my supervisor Elena Loli Piccolomini, especially for the support and for the patience she showed in the difficult moments and during the numerous discussions we have made about this thesis. Then I would like to thank Professor Germana Landi, whose suggestions and remarks were crucial in the central chapters of the thesis and allowed Elena and me to improve the level of this part significantly. A special thank to Professor Barbara Kaltenbacher, for the very kind hospitality in Klagenfurt and for her illuminating guide in the final part of the thesis. Without her advices, that part would not have been possible.

I am very grateful also to Professor Alberto Parmeggiani, whose support was very important for me during these three years.

I would also like to thank the Department of Mathematics of the University of Bologna, that awarded me with 3 fellowships for the Ph.D and that gave me the opportunity of studying and carrying out my research in such a beautiful environment. I also thank the Alpen Adria Universität of Klagenfurt for the hospitality in April 2012 and October-November 2012 respectively. In particular a special greeting is dedicated to the secretary Anita Wachter, who provided me with a beautiful apartment during my second period in Klagenfurt.

And now, let's turn to italian!

Alla fine, sono arrivato in fondo anche a questa fatica ed ancora non ci credo. Sebbene non sia un amante dei ringraziamenti, questa volta vorrei ringraziare le persone a me più vicine, per la pazienza che mi hanno dedicato in questi anni di ricerca pieni di alti e bassi.

Innanzitutto, la mia famiglia, mia sorella, mio padre e mia madre, che hanno sempre creduto in me, anche quando tornavo a casa arrabbiato, triste o sfiduciato. Inutile dire che il supporto di tutti, compresi i miei nonni e mia zia Carla, è stato e sarà sempre fondamentale per me.

Grazie anche ad Anna e Marco, che posso definire membri onorari della famiglia, ai miei cugini Daniele ed Irene e ai miei zii, Lidia e Davide.

Voglio poi ringraziare i miei tre compagni di viaggio in quest'avventura: Gabriele Barbieri, Luca Ferrari e Giulio Tralli. In particolare Giulio, col quale ho condiviso in pieno quest'esperienza, con discussioni infinite, dalla matematica ai massimi sistemi alle stupidaggini, oltre ad un viaggio memorabile insieme, e tantissime altre esperienze.

Un immenso grazie anche a tutti i miei amici, in particolare quelli più stretti, Luca Fabbri, Barbara Galletti, Laura Melega, Andrea Biondi, Roberto e Laura Costantini, Emanuele Salomoni, Elena Campieri, Anna Cavrini, Luca Cioni e tutti gli altri.

Infine, l'ultimo ringraziamento va a Silvia, che semplicemente mi ha dato quella felicità che cercavo da tanto tempo e che spero di avere finalmente trovato.