Diffusion Equations and Inverse Problems Regularization

Davide Bianchi

Diffusion Equations and Inverse Problems Regularization

Davide Bianchi

PhD Thesis
in
Computer Science and Computational Mathematics (XXIX^{o} ciclo)
Università degli Studi dell'Insubria
Dipartimento di Scienze e Alta Tecnologia
Como - Italy

Author
Davide Bianchi

First Supervisor: Prof. Marco Donatelli . . . . . .
Second Supervisor: Prof. Alberto Giulio Setti . . . . . .

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I express my extreme gratitude to Professor Marco Donatelli and Professor Alberto Giulio Setti for having accepted me as their student from the very beginning of my university career and having passed on me the love for mathematical research. Without their wonderful guidance, patience and never ending support nothing of this work would have been accomplished.

My sincere thanks to the referees of this thesis, Professor Gabriele Grillo and Professor Martin Hanke-Bourgeois, for their comments and suggestions.

Nevertheless, I would like to thank all the Professors who I encountered in my years as a student in the Università dell'Insubria, since everyone of them contributed to my scientific growth. In particular, let me mention: Professor Sergio Luigi Cacciatori, a great friend and a magnificent teacher who was able to let a mathematician like me appreciate Physics; Professor Stefano Pigola, gifted by a keen wit and always prompt to help greatly sharing his knowledge; Professor Stefano Serra Capizzano, for his precious teachings and help as my tutor.

Many thanks go to all my fellow colleagues, co-authors, the computer technicians, in particular to Emanuele Corti, and the Secretary's office Department staff, in particular to Alessandra and Francesca Parassole and to Carmen Tripodi. Their work and help are priceless for the University.

A special thanks to my dearest friends, for their wonderful friendship and for having always expected nothing less than excellence from me.

Finally, I will never be enough grateful to my family for their love and support. This work is dedicated to them: to my mother Claudia for having believed in me from the start, to my father Giorgio, to my sisters Katiaelena and Stefania, and to my nephew Cristian with the wish that it could inspire him to surpass me one day.

Introduction

Inverse problems play an essential role in every applicative field, whether it be experimental physics, biology or chemistry, whenever one wants to recover the original state of an evolved system from its final state. Without attempting to give a precise definition of "inverse problem", indeed, we believe that a direct example can be more clarifying and we will describe it later (see Chapter 4). In this context, it is appropriate to remark how difficult it can be to solve such problems from the point of view of a numerical analyst. Not dealing with exact data, calculus complexity and numerical instability are just some of the issues that could adversely affect an approximated solution, causing it to be very different from what it would be the real solution and therefore useless. Depending on the model problem, dozens of different methods and techniques were introduced over the years and new are constantly developed, making it a grueling task even just trying to enumerate them. In the early stage of our studies, we began to work on some inverse problems arising from the modeling of certain signals by the use of convolution operators and in that context we focused on inverse problem regularization techniques of Tikhonov filter type. As a second crucial step, we searched for applications of our new methods and we came across to some interesting problems concerning changelling nonlinear diffusion equations. From the very start we were confronted by some highly nontrivial, but fascinating theoretical problems which made us temporarily shift from our original goal. It was indeed clear that in order to satisfactorily solve the inverse problems connected to them we had to deal with those technical issues beforehand. For example, when dealing with nonlinear equations arising from groundwater filtration problems, cf. [97], our first attempt to exploit deeper the geometry of the space led us to find in some sense a lack of theoretical tools. It is still quite a novelty the study of nonlinear/porous type equations, cf. [109], in a Riemannian setting. Indeed, we were a little surprised to realize that the Riemannian version of many Euclidean results was still missing, due to the lack of cut-off functions with a suitable control on the decay rates of their derivatives, whose existence is well known in Euclidean space.

Because of the aforementioned reasons, the present thesis can be split into two different parts:

- The first part, completely theoretical, mainly deals with the porous and fast diffusion equations. Chapter 2 presents the porous and fast diffusion equations in the Euclidean setting highlighting the technical issues that arise when trying to extend results in a Riemannian setting. Chapter 3 is devoted to the construction of exhaustion and cut-off functions with controlled gradient and Laplacian, on manifolds with Ricci curvature bounded from below by a (possibly unbounded) nonpositive function of the distance from a fixed reference point. We stress that we realized it without any assumptions on the topology or the injectivity radius. Along the way we prove a generalization of the Li-Yau gradient estimate which is of independent interest. Then we apply our cut-offs to the study of the fast and porous media diffusion, of L^q-properties of the gradient and of the self-adjointness of Schrödinger-type operators. Most of the results presented in this first part come from [12].

- The second part is concerned with inverse problems regularization, mainly applied to image deblurring. In Chapter 5, we present new variants of the Tikhonov filter method, called fractional Tikhonov and weighted Tikhonov. Those regularization methods have been recently proposed to reduce the oversmoothing property of the Tikhonov regularization in

standard form, in order to preserve the details of the approximated solution. Their regularization and convergence properties have been previously investigated showing that they are of optimal order. In this chapter we provide saturation and converse results on their convergence rates. Using the same iterative refinement strategy of iterated Tikhonov regularization, new iterated fractional Tikhonov regularization methods are introduced. We show that these iterated methods are of optimal order and overcome the previous saturation results. Furthermore, nonstationary iterated fractional Tikhonov regularization methods are investigated, establishing their convergence rate under general conditions on the iteration parameters.
In Chapter 6 we investigate the modified linearized Bregman algorithm (MLBA) used in image deblurring problems, with a proper treatment of the boundary artifacts. We consider two standard approaches: the imposition of boundary conditions and the use of the rectangular blurring matrix. The fast convergence of the MLBA depends on a regularizing preconditioner which could be computationally expensive and hence it is usually chosen as a block circulant circulant block (BCCB) matrix, diagonalized by bidimensional discrete Fourier transform. We show that the standard approach based on the BCCB preconditioner may provide low quality restored images and we propose different preconditioning strategies, which improve the quality of the restoration and save some computational cost at the same time. Motivated by a recent nonstationary preconditioned iteration, we propose a new algorithm which combines such method with the MLBA. We prove that it is a regularizing and convergent method. A variant with a stationary preconditioner is also considered.
Most of the results presented in this second part come from [11] and [23].

Finally, we want to make a remark. Despite the different fields we tried our best to keep an uniform notation throughout the chapters. In the first part we used some capital letters to indicate constants or vector fields while in the second part capital letters indicate only operators and sets. Having said that, we are confident that there is no risk of misunderstandings since the context will be self explanatory.

The Porous and Fast diffusion equations

Porous and Fast diffusion equations on $\mathbb{R}^d$: a brief introduction.

The study of the Porous Medium Equation and the Fast Diffusion Equation (hereafter often referred to as PME and FDE, respectively) goes back to the first half of the last century and has been pursued and deeply investigated by several authors in hundreds of papers. Therefore, the present chapter is not meant to be an exhaustive introduction to the PME/FDE problems, nor to be even a thumbnail summary of all the main features and last developments, on the subject. Rather, its goal is to highlight the origin of these and to point out the critical difficulties that prevented to extend some of the fundamental properties of their solutions to a general Riemannian manifold setting. We refer any interested reader to the excellent book [109] for an exhaustive and all-inclusive summary of the PME and FDE problems.

The easiest way to introduce the PME is to look at the flow of a gas inside a porous medium. Let us consider the following physical equations:

$$\text{Mass Balance:} \quad \varepsilon \partial_t \rho + \operatorname{div}(\rho V) = 0, \tag{2.0.1a}$$

where $\varepsilon \in (0,1)$ is the porosity of the medium, a ratio between the volume of void space inside the material and the total volume occupied by the , $\rho : [0,\infty) \times \mathbb{R}^3 \to (0,\infty)$ is the density and $V : \mathbb{R}^3 \to \mathbb{R}^3$ is the velocity vector;

$$\text{Darcy's Law:} \quad \mu V = -\kappa \nabla P, \tag{2.0.1b}$$

where μ is the viscosity, κ is the permeability of the medium and $P : \mathbb{R}^3 \to [0,\infty)$ is the pressure;

$$\text{State Equation:} \quad P = P_0 \rho^\gamma, \tag{2.0.1c}$$

where P_0 is the reference pressure and γ is the politropic exponent, with $\gamma = 1$ for isothermal transformations and $\gamma > 1$ for adiabatic transformations. Putting together equations (2.0.1a), (2.0.1b) and 2.0.1c we obtain

$$\partial_t \rho = C \cdot \operatorname{div}(\rho^\gamma \nabla \rho) = C\Delta(\rho^m),$$

with $C = C(\varepsilon, \kappa, P_0, \gamma)$ and $m = \gamma + 1$. The constant C can be scaled out by the means of a reparametrization, $\hat{t} := C \cdot t$, and if we allow m to take values in the range $(0,\infty)$, and having renamed the previous unknown function ρ into u for consistency with the literature, we are lead to the following equation:

$$\partial_t u = \Delta u^m = \begin{cases} \text{PME} & \text{if } m > 1, \\ \text{heat diffusion equation} & \text{if } m = 1, \\ \text{FDE} & \text{if } 0 < m < 1, \\ \text{logarithmic diffusion equation} & \text{if } m = 0. \end{cases} \tag{2.0.2}$$

Note that

$$\Delta u^m = m \cdot \operatorname{div}(u^{m-1} \nabla u),$$

and defining

$$a(u) := u^{m-1}$$

the *diffusivity coefficient*, then, for $0 < m < 1$, $a(u)$ has a singularity at $u = 0$, namely $a(u)$ explodes to infinity as the solution u approaches to zero, whence the name *fast diffusion*. In the same way, if we let $m = 0$, then $a(u) = u^{-1}$ and $\operatorname{div}(u^{-1}\nabla u) = \Delta \log u$, which accounts for the name *logarithmic diffusion*.

We recall that the PME is frequently used to model gas flow and groundwater filtration while the FDE appeared first in the Okudo-Dawson plasma's diffusion model. For the sake of simplicity we derived both the PME and the FDE from the gas flow inside a porous medium, but for a more detailed insight of the FDE from a physical point of view we invite the reader to look at [82, 10, 47, 9].

Now, to get an idea of the interplay between the geometry of the ambient space and certain properties of solutions of the PME and FDE equations, and in particular of the technical difficulties which arise trying to extend results valid in Euclidean space to the setting Riemannian manifold, we are going to focus our attention to the Cauchy problem for the FDE on the whole space. Let us fix $0 < m < 1$ and introduce the FDE-Cauchy problem on $\mathbb{R}^d$,

$$\begin{cases} \partial_t u(t,x) = \Delta u^m(t,x) & \text{for } x \in (0,+\infty) \times \mathbb{R}^d \\ u(0,x) = u_0(x) & \text{for } x \in \mathbb{R}^d. \end{cases} \tag{2.0.3}$$

By a solution to the problem we mean a curve $u : (0,+\infty) \to X$ where the functional space X depends on the functional class determined by the initial datum u_0. In the ensuing discussion we assume that u_0 is in $\mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d)$, and we are going to define weak and strong solution to the FDE-Cauchy problem (2.0.3) as follows (see [65]):

Definition 2.0.1 (weak and strong solutions for the FDE)**.**
Let $u(t,x) \in C([0,+\infty) : \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d))$ be such that

(i)

$$u(0,x) = u_0, \tag{2.0.4}$$

(ii)

$$\partial_t u = \Delta u^m, \qquad \text{in } D'((0,+\infty) \times \mathbb{R}^d). \tag{2.0.5}$$

Then u is called a weak solution *to the Cauchy problem of the FDE. If in addition u satisfies*

(iii)

$$\partial_t u \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty) \times \mathbb{R}^d), \tag{2.0.6}$$

then u is called a strong solution*. Note that since $0 < m < 1$ then $u^m \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d)$ as well.*

In the by Herrero mentioned above, global existence in time for solutions of the FDE is proven for any $u_0 \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d)$. We remark that the case of the porous equation $m > 1$ is quite different, since it had been realized quite early that the local integrability of the initial datum does not ensure even short time existence, for which a control of the growth at infinity of the initial datum is required.

Going back to the FDE, the global existence of solutions for locally integrable initial data hinges on a weakened form of the conservation of mass, which we will refer to as the *Weak Conservation of Mass* Theorem and which allows to give a lower bound for the *extinction time* of a solution for the FDE-Cauchy problem in terms of local quantities related to the initial datum u_0. Indeed, if u_0 belong to a suitable Lebesgue space, and for certain exponents $m \in (0,1)$, one shows that there exists a time $T = T(u_0)$ such that for every $t \geq T$ the solution $u(t)$ of the FDE-Cauchy problem vanish almost everywhere. By the Weak Conservation of Mass Theorem there exists a critical exponent $m_c = \frac{d-2}{d} \in (0,1)$ such that $T(u_0) = \infty$ for every $m \in (m_c, 1)$, and for such m there is no extinction time. For a deeper understanding about the critical exponent m_c and the extinction time we refer the reader to [108] and to the aformentioned [109].

Because of the techniques used in the proof and the consequences of the above Weak Conservation of Mass Theorem, we believe that it is an useful and meaningful example to highlight some of the deep differences between the Euclidean and the Riemannian setting. Below, we reproduce the statement and the proof of the Weak Conservation of Mass Theorem pointing out the steps where geometry comes to play a fundamental role. In particular we divide the proof into 2 parts: the first part is essentially analytic and it can be adapted without effort to the general setting of Riemannian manifolds, while the second part uses crucially the Euclidean distance function, thus giving rise to not trivial difficulties and preventing a naive adaptation to a manifold setting. In Section 3.4.2 we will be able to generalize this result on Riemannian manifolds.

Hereafter, we will define

$$u^m = |u|^{m-1}u,$$

allowing u to be not necessarily nonnegative. Indeed, the PME/FME equation can be considered for functions of arbitrary sign. Whenever a result will be valid only for nonnegative solutions $u \geq 0$, it will be highlighted.

Theorem 2.0.2 (Weak conservation of mass, [65, Lemma 3.1])**.**
Let $u, v \in C([0,+\infty) : \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d))$ satisfy condition (2.0.5) *and be such that $u \geq v$. Then, for all $R > 0$, $\gamma > 1$ and $t, s \geq 0$*

$$\left[\int_{B_R(x_0)} [u(t) - v(t)]\, dx\right]^{1-m} \leq \left[\int_{B_{\gamma R}(x_0)} [u(s) - v(s)]\, dx\right]^{1-m} + M_{R,\gamma}|t-s| \tag{2.0.7}$$

where we $B_R(x_0)$ denotes the ball of radius R centered at $x_0 \in \mathbb{R}^d$, and where

$$M_{R,\gamma} = \frac{C_0}{(\gamma-1)}\frac{C_1}{R^2}\, vol(B_{\gamma R}(x_0) \setminus B_R(x_0))^{1-m} > 0$$

with the constants $C_i > 0$ depending only on m and d.

Proof.
Without loss of generality we can fix $x_0 = o$, the origin of axis. Note that this is not generally true in the Riemannian counterpart of this statement, Theorem 3.4.6.

- **Step I**
From (2.0.5), for every nonnegative $\eta \in C_c^\infty(0,\infty)$ and $\psi \in C_c^\infty(\mathbb{R}^d)$ we have that

$$\begin{array}{rcl} \langle \partial_t(u-v), \eta\psi\rangle & = & -\langle u-v, \partial_t\eta\,\psi\rangle \\ \| & & \\ \langle \Delta(u^m-v^m), \eta\psi\rangle & = & \langle u^m-v^m, \eta\Delta\psi\rangle \end{array}$$

in distributions, that is,

$$-\int_0^\infty \int_{\mathbb{R}^d} \partial_t\eta\,\psi(u-v)\,dtdx = \int_0^\infty \int_{\mathbb{R}^d} \eta\Delta\psi(u^m-v^m)\,dtdx.$$

Thus

$$-\int_0^\infty \partial_t\eta \left(\int_{\mathbb{R}^d} \psi(u-v)\,dx\right) dt = \int_0^\infty \eta \left(\int_{\mathbb{R}^d} \Delta\psi(u^m-v^m)\,dx\right) dt$$

and this implies the validity of the equality

$$\frac{d}{dt}\int_{\mathbb{R}^d} \psi(u(t)-v(t))\,dx = \int_{\mathbb{R}^d} \Delta\psi(u^m-v^m)\,dx \tag{2.0.8}$$

in $D'(0,\infty)$ and in $\mathrm{L}^1_{\mathrm{loc}}(0,\infty)$ as well for every fixed ψ, as a consequence of (2.0.4). Since, by concavity,

$$(r|r|^{m-1}-s|s|^{m-1}) \le 2^{1-m}(r-s)^m \qquad \text{for all } r \ge s,$$

then (2.0.8) implies

$$\frac{d}{dt}\int_{\mathbb{R}^d} \psi(u(t)-v(t))\,dx \le 2^{1-m}\int_{\mathbb{R}^d} |\Delta\psi|(u-v)^m.$$

We set $g := u-v$. By Holder's inequality, we obtain

$$\frac{d}{dt}\int_{\mathbb{R}^d} \psi g(t) \le C(\psi)\left[\int_{\mathbb{R}^d} \psi g(t)\right]^m, \tag{2.0.9}$$

where

$$C(\psi) = \left[2\int_{\mathbb{R}^d} |\Delta\psi|^{1/(1-m)}\psi^{-m/(1-m)}\right]^{1-m}.$$

Since the function $f_\psi(t) = \int_{\mathbb{R}^d} \psi g(t)$ has weak derivative in $\mathrm{L}^1_{\mathrm{loc}}$, it is a.e. equal to an AC function, and by standard comparison arguments, for all $t_1, t_2 \ge 0$ and every $\psi \in C_c^\infty(M)$,

$$\left[\int \psi g(t_2)\right]^{1-m} \le \left[\int \psi g(t_1)\right]^{1-m} + (1-m)C(\psi)|t_2-t_1|. \tag{2.0.10}$$

This will immediately imply the statement, once we prove that $C(\psi) \le M_{R,\gamma} < \infty$.

- **Step II**
  To this end we consider a function $\psi = \phi_R^b \in C_c(\mathbb{R}^d)$, with

$$0 \le \phi_R \le 1, \qquad \phi_R \equiv 1 \text{ in } B_R(o), \qquad \phi_R \equiv 0 \text{ outside } B_{\gamma R}(o),$$

  where $\gamma > 1$ and $b > 2/(1-m)$. Moreover, we will assume that ϕ_R is radial and

$$\phi_R(x) = \overline{\phi}(r(x)/R)$$

  where $r : \mathbb{R}^d \to [0,\infty)$ is the Euclidean distance function from the origin o, namely $r(x) = \left(\sum_{i=1}^d |x_i|^2\right)^{1/2}$, and $\overline{\phi} : \mathbb{R} \to \mathbb{R}$ is a $C_c^\infty(\mathbb{R})$ function such that:

$$0 \le \overline{\phi} \le 1, \qquad \overline{\phi} \equiv 1 \text{ for } 0 \le s \le 1, \qquad \overline{\phi} \equiv 0 \text{ for } s \ge \gamma.$$

  Then we have

$$\begin{aligned} |\Delta(\psi(x))|^{1/(1-m)} \psi(x)^{-m/(1-m)} &= \phi_R(x)^{-bm/(1-m)} \left| b(b-1)\phi_R^{b-2}|\nabla\phi_R|^2 + b\phi_R^{b-1}\Delta\phi_R \right|^{1/(1-m)} \\ &\le [b(b-1)]^{1/(1-m)} \phi_R^{[(b-2)-bm]/(1-m)} \cdot \left| |\nabla\phi_R|^2 + |\Delta\phi_R| \right|^{1/(1-m)}, \end{aligned} \tag{2.0.11}$$

  where last inequality follows from the fact that we are considering a radial function $0 \le \phi_R(x) = \overline{\phi}(r(x)/R) \le 1$, with $b > 2/(1-m)$. Then we compute

  (a.1)

$$\begin{aligned} |\nabla\phi_R(x)|^2 &= R^{-2}|\overline{\phi}'(r(x)/R)|^2 |\nabla r(x)|^2 \\ &= R^{-2}|\overline{\phi}'(r(x)/R)|^2 \\ &\le \frac{C_0}{(\gamma-1)} R^{-2}; \end{aligned}$$

  (a.2)

$$\begin{aligned} |\Delta\phi_R(x)| &= |R^{-2}\overline{\phi}''(r(x)/R)|^2 |\nabla r(x)|^2 + R^{-1}\overline{\phi}'(r(x)/R)\Delta r(x)| \\ &\le R^{-1}\left(R^{-1}|\overline{\phi}''(r(x)/R)| + |\overline{\phi}'(r(x)/R)| \frac{(d-1)}{r(x)} \right) \\ &\le \frac{C_0}{(\gamma-1)}(d-1)R^{-2}; \end{aligned}$$

  where we used the fact that $\Delta\phi_R$ is supported in $B_{\gamma R}(o) \setminus B_R(o)$ and that the smooth function $\overline{\phi}$ has bounded derivatives in $B_{\gamma R} \setminus B_R$

$$|\overline{\phi}''(r(x)/R)| + |\overline{\phi}'(r(x)/R)| \le \frac{C_0}{(\gamma-1)}.$$

An integration over $B_{\gamma R}(o) \setminus B_R(o)$ gives:

$$\begin{aligned} C(\psi) &= \left[2 \int_{B_{\gamma R} \setminus B_R} |\Delta \psi|^{1/(1-m)} \psi^{-m/(1-m)} \right]^{1-m} \\ &\leq \frac{C_0}{(\gamma-1)} \frac{C_1}{R^2} \mathrm{vol}(B_{\gamma R}(o) \setminus B_R(o)^{1-m}, \end{aligned}$$

with $C_1 = 2^{1-m} b(b-1)$. This concludes the proof.

□

Given $u_0 \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d)$ we define the extinction time $T(u_0) \in (0, +\infty]$ of the solution $u(x,t)$ of the FDE Cauchy problem with an initial datum u_0 as the smallest time such that $u(x,t) \equiv 0$ for every $t \geq T(u_0)$ and for almost every $x \in \mathbb{R}^d$. As immediate consequence of the Weak Conservation of Mass Theorem we have the following:

Corollary 2.0.3. *If $m \in (1-2/d, 1)$, then $T(u_0) = +\infty$ for every $0 \leq u_0 \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d)$.*

Proof. Letting $v \equiv 0$, $t = 0$, $s = T(u_0)$ and $R \geq 1$, it follows from Theorem 2.0.2 that

$$\begin{aligned} T(u_0) &\geq \frac{R^2(\gamma-1) \left[\int_{B_R(x_0)} u_0 \, dx \right]^{1-m}}{C_0 C_1 vol(B_{\gamma R}(x_0) \setminus B_R(x_0))^{1-m}} \\ &\geq \frac{R^2(\gamma-1) \|u_0\|_{\mathrm{L}^1(B_1(x_0))}}{C_0 C_1 vol(B_{\gamma R}(x_0))^{1-m}} \\ &= C(u_0) R^{2-d(1-m)}, \end{aligned}$$

and, since $m > 1-2/d$, the right hand side of the above inequality goes to infinity as $R \to \infty$. □

In **Step II** of the proof of Theorem 2.0.2 was of utmost importance the existence of cut-off functions $\phi_R \in C^2(\mathbb{R}^d)$ of the metric balls with a specific decay rate of the gradient and the Laplacian, namely,

(i) $\phi_R : \mathbb{R}^d \to [0,1]$

(ii) $\phi_R \equiv 1$ on $B_R(x_0)$,

(iii) $\mathrm{supp}(\phi_R) \subset B_{\gamma R}(x_0)$,

(iv) $\sup_{\mathbb{R}^d} |\nabla \phi_R(x)| \leq C R^{-1}$,

(v) $\sup_{\mathbb{R}^d} |\Delta \phi_R(x)| \leq C R^{-2}$.

As we have seen, in an Euclidean space it is possible to define such cut-offs in terms of the distance function r, whose key features are

(i) $r \in C^\infty(\mathbb{R}^d \setminus \{o\})$,

(ii) $|\nabla r(x)| \equiv 1$,

(iii) $\Delta r(x) = \frac{d-1}{r(x)}$.

With the exception of (ii) which holds, at least a.e., on every Riemannian manifold, in general the smoothness condition (i) and the decay rate (iii) are not satisfied by the Riemannian distance. Building such cut-offs in a Riemannian manifold is a difficult task which we take up in Chapter 3. As a result we will be able to obtain suitable versions of some Euclidean results to solutions to the PME/FDE equation to manifolds.

2.1 Introduction to Riemannian manifolds

Before proceeding further, here we give a brief introduction to the Riemannian manifold setting in order to let the remainder of this first part easily understandable and accessible even to readers not expert of these tools. For any reference, all the definitions and statements that we will provide can be found on (or rearranged from) the complete and wonderful books [83, 28, 18].

2.1.1 Topological Manifold

Definition 2.1.1 (**Topological space**). *Let X be a set and let $\mathscr{T} \subseteq \mathscr{P}(X)$, where $\mathscr{P}(X)$ is the collection of all the subsets of X. $\mathscr{T}$ is called a* topology *for X if*

(i) $X \in \mathscr{T}$ *and* $\emptyset \in \mathscr{T}$.

(ii) Let $\{U_\iota\}_{\iota\in A} \subseteq \mathscr{T}$ *be a finite or infinite collection of elements of* $\mathscr{T}$*, where A is a set of indexes. Then* $\bigcup_{\iota\in A} U_\iota \in \mathscr{T}$.

(iii) Let $\{U_j\}_{j=1}^n \subseteq \mathscr{T}$ *be a finite collection of elements of* $\mathscr{T}$*. Then* $\bigcap_{j=1}^n U_j \in \mathscr{T}$.

The elements of $\mathscr{T}$ are called open sets *and the pair $(X,\mathscr{T})$ is a* topological space*. Often we will omit $\mathscr{T}$.*

Definition 2.1.2 (**II-numerable space**). *Let $(X,\mathscr{T})$ be a topological space and let $\{B_\iota\}_{\iota\in A} \subseteq \mathscr{T}$, where A is a set of indexes. $\{B_\iota\}_{\iota\in A}$ is a* base *for $\mathscr{T}$ if*

(i) $\bigcup_{\iota\in A} B_\iota \supseteq X$.

(ii) For every pair B_{ι_1}, B_{ι_2} and for every $x \in B_{\iota_1} \cap B_{\iota_2}$ there exists $B_{\iota_3} \in \{B_\iota\}_{\iota\in A}$ such that $x \in B_{\iota_3} \subseteq B_{\iota_1} \cap B_{\iota_2}$.

If a topological space $(X,\mathscr{T})$ admits a countable base $\{B_\iota\}_{\iota\in A}$, then we say that $(X,\mathscr{T})$ is II-countable*.*

Definition 2.1.3 (**Hausdorff space**). *We say that a topological space $(X,\mathscr{T})$ is* Hausdorff *if for every pair $x_1, x_2 \in X$, $x_1 \neq x_2$, there exist open sets $U_1, U_2 \in \mathscr{T}$ such that*

$$x_1 \in U_1, \qquad x_2 \in U_2 \qquad \text{and} \qquad U_1 \cap U_2 = \emptyset.$$

Definition 2.1.4 (**Chart**). *Given a topological space X, a homeomorphism ξ of an open set U of X onto an open set $\xi(U)$ of $\mathbb{R}^d$, is called* coordinate map *and U is called* coordinate neighbour. *If we write*

$$\xi(x) = \begin{bmatrix} \xi^1(x) & \cdots & \xi^d(x) \end{bmatrix}$$

for each $x \in U$, the resulting functions $\xi^1, \cdots, \xi^d$ are called coordinate functions *of ξ. We call d the dimension of ξ. Finally, the pair (U, ξ) is called a* chart.
If $x_0 \in X$ is fixed and there exists a chart (U_{x_0}, ξ_{x_0}) such that $x_0 \in U_{x_0}$, then (U_{x_0}, ξ_{x_0}) is called local chart *in x_0.*

Definition 2.1.5 (**Locally Euclidean space**). *A topological space X is said* locally Euclidean *of dimension d if for every $x_0 \in X$ there exists a local chart (U_{x_0}, ξ_{x_0}) in x_0 of dimension d.*

Definition 2.1.6 (**Topological manifold**). *A topological space M is called* topological manifold *of dimension* $\dim(M) = d$ *if the following properties are satisfied:*

(1) X is locally Euclidean of dimension d.

(2) X is Hausdorff.

(3) X is II-countable.

2.1.2 Smooth Manifold

Definition 2.1.7 (C^∞ **Atlas**). *Let M be a topological manifold of dimension* $\dim(M) = d$. *A family of local charts*

$$\mathscr{I}_0 = \{(U_\iota, \xi_\iota) : \iota \in A\}$$

with the property

$$\bigcup_{\iota \in A} U_\iota = M,$$

is called topological atlas *of M. The atlas $\mathscr{I}_0$ is said to be of C^∞* class *if for every pair $\iota_1, \iota_2 \in A$ such that $U_{\iota_1} \cap U_{\iota_2} \neq \emptyset$ the map*

$$\xi_{\iota_1} \circ \xi_{\iota_2}^{-1} : \xi_{\iota_2}(U_{\iota_1} \cap U_{\iota_2}) \subseteq \mathbb{R}^d \to \mathbb{R}^d$$

is of C^∞ class in the usual sense of analysis, i.e., $\xi_{\iota_1} \circ \xi_{\iota_2}^{-1}$ has continuous partial derivatives of any order. The map $\xi_{\iota_1} \circ \xi_{\iota_2}^{-1}$ is called transition function.

Definition 2.1.8 (C^∞ **differential structure**). *Let M be a topological manifold of* $\dim(M) = d$ *which has at least one atlas $\mathscr{I}_0$ of C^∞ class. In the set*

$$\{\mathscr{I} : \mathscr{I} \text{ is an atlas for } M \text{ of } C^\infty \text{ class}\} \neq \emptyset,$$

we introduce the following equivalence relationship:

$$\mathscr{I} \sim \mathscr{G} \qquad \Longleftrightarrow \qquad \mathscr{I} \cup \mathscr{G} \text{ is a } C^\infty \text{ atlas.}$$

The equivalence class $[\mathscr{I}]$ which contains the C^∞ atlas $\mathscr{I}$ is called C^∞ differential structure *on M. Hereafter we will avoid the brackets $[\cdot]$ and we will only write $\mathscr{I} = [\mathscr{I}]$.*

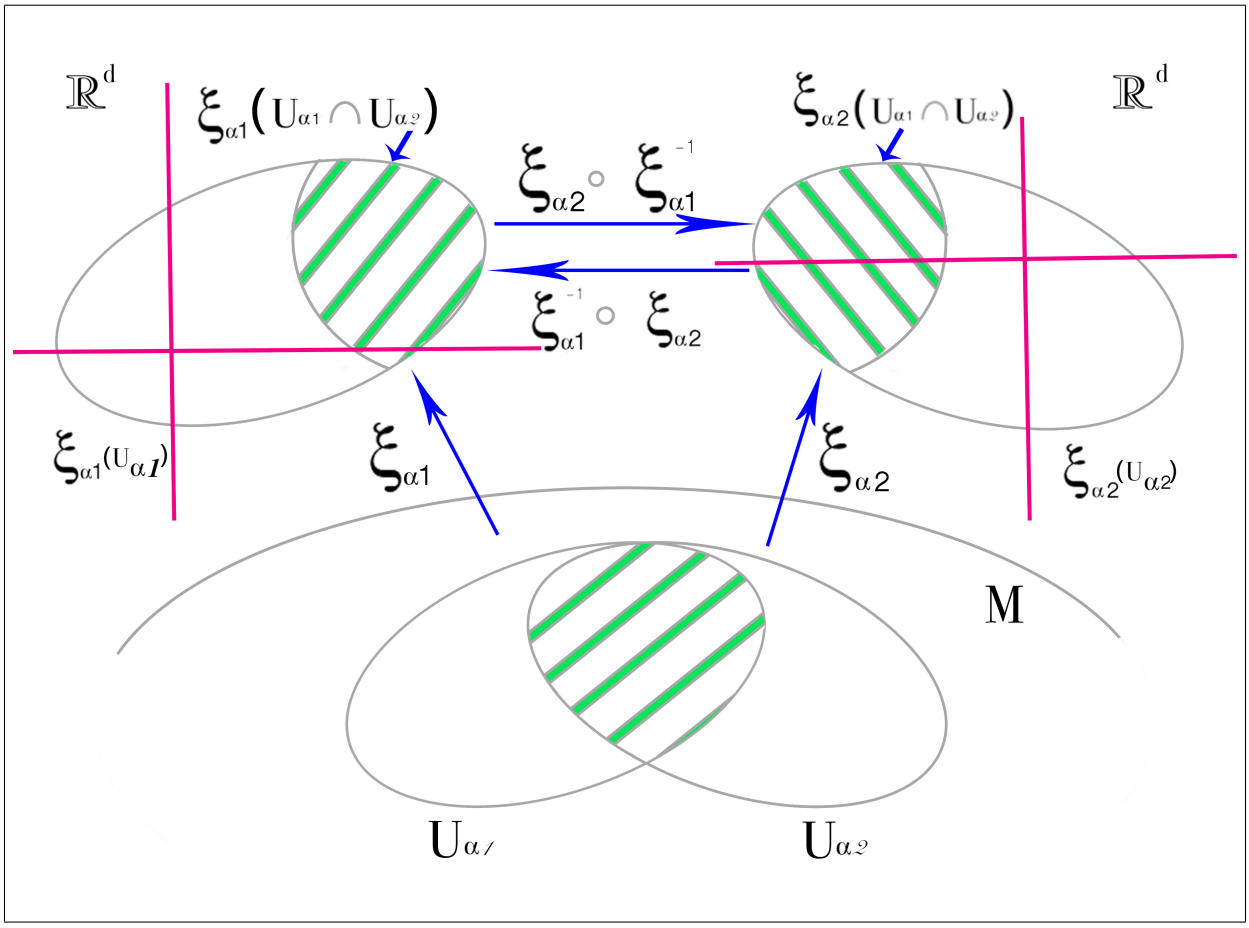


Figure 2.1: Transition functions and charts.

Remark 2.1.9. *$\mathscr{I}$ is a C^∞ differential structure on M if and only if*

(i) *$\mathscr{I}$ is a C^∞ atlas.*

(ii) *$\mathscr{I}$ is maximal, in the sense that if (V,ρ) is a local chart of M such that for every $(U,\xi) \in \mathscr{I}$ with*

(a) *$U \cap V \neq \emptyset$,*

(b) *$\xi \circ \rho^{-1}$ and $\rho \circ \xi^{-1}$ are C^∞,*

then $(V,\rho) \in \mathscr{I}$.

Definition 2.1.10 (Smooth manifold). *We call* smooth manifold *of dimension d the pair $(M,\mathscr{I})$, where M is a topological manifold of* $\dim(M) = d$ *and $\mathscr{I}$ is a C^∞ differential structure on M.*

Example 2.1.11. *Let us consider $S^d \subset \mathbb{R}^{d+1}$, defined as*

$$S^d = \left\{ x \in \mathbb{R}^{d+1} \,:\, \|x\| = 1 \right\},$$

where

$$\|x\| = \left(\sum_{i=1}^{d+1} (x_i)^2 \right)^{1/2} \qquad \text{for every } x = \begin{bmatrix} x_1 & x_2 & \cdots & x_{d+1} \end{bmatrix}.$$

Of course, S^d is connected and compact. Fix

$$\mathscr{I}_0 = \{(U_{\mathbf{N}}, \xi_{\mathbf{N}}), (U_{\mathbf{S}}, \xi_{\mathbf{S}})\},$$

where

$$U_{\mathbf{N}} = S^d \setminus \{\mathbf{N}\} \text{ with } \mathbf{N} = \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}, \qquad U_{\mathbf{S}} = S^d \setminus \{\mathbf{S}\} \text{ with } \mathbf{S} = \begin{bmatrix} 0 & 0 & \cdots & -1 \end{bmatrix},$$

and

$$\xi_{\mathbf{N}} : U_{\mathbf{N}} \to \mathbb{R}^d \quad \text{is such that} \quad \xi_{\mathbf{N}}(x) = \frac{1}{-x_{d+1}+1} \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix},$$
$$\xi_{\mathbf{S}} : U_{\mathbf{S}} \to \mathbb{R}^d \quad \text{is such that} \quad \xi_{\mathbf{S}}(x) = \frac{1}{x_{d+1}+1} \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}.$$

$\mathscr{I}_0$ is a C^∞ atlas and the differential structure $\mathscr{I}$ generated by $\mathscr{I}_0$ is called standard differential structure on S^d. *Finally, $\left(S^d, \mathscr{I}\right)$ is called* sphere.

Definition 2.1.12 (**Locally finiteness**). *A collection $\mathscr{S}$ of subsets of a space X is locally finite provided each point of X has a neighborhood that meets only finitely many elements of $\mathscr{S}$.*

Let $\{f_\iota : \iota \in A\}$ be a collection of smooth functions on a manifold M such that $\{\text{supp}(f_\iota) : \iota \in A\}$ is locally finite. Then the sum $\sum_{\iota \in A} f_\iota$ is a well-defined smooth function on M, since on some neighborhood of each point all but a finite number of f_ι are identically zero.

Definition 2.1.13 (**Partition of unity**). *A smooth partition of unity on a manifold M is a collection $\{f_\iota : \iota \in A\}$ of smooth functions $f_\iota : M \to \mathbb{R}$ such that*

(i) $0 \leq f_\iota \leq 1$ *for every* $\iota \in A$.

(ii) $\{\text{supp}(f_\iota) : \iota \in A\}$ *is locally finite.*

(iii) $\sum_{\iota \in A} f_\iota = 1$.

Proposition 2.1.14. *For every smooth manifold M, given any open covering $\{U_\iota\}_{\iota \in A}$ of M there is a smooth partition of unity $\{f_\iota : \iota \in A\}$ subordinate to $\{U_\iota\}_{\iota \in A}$, i.e., for every ι_1 there exists ι_2 such that* $\text{supp}(f_{\iota_2}) \subset U_{\iota_1}$.

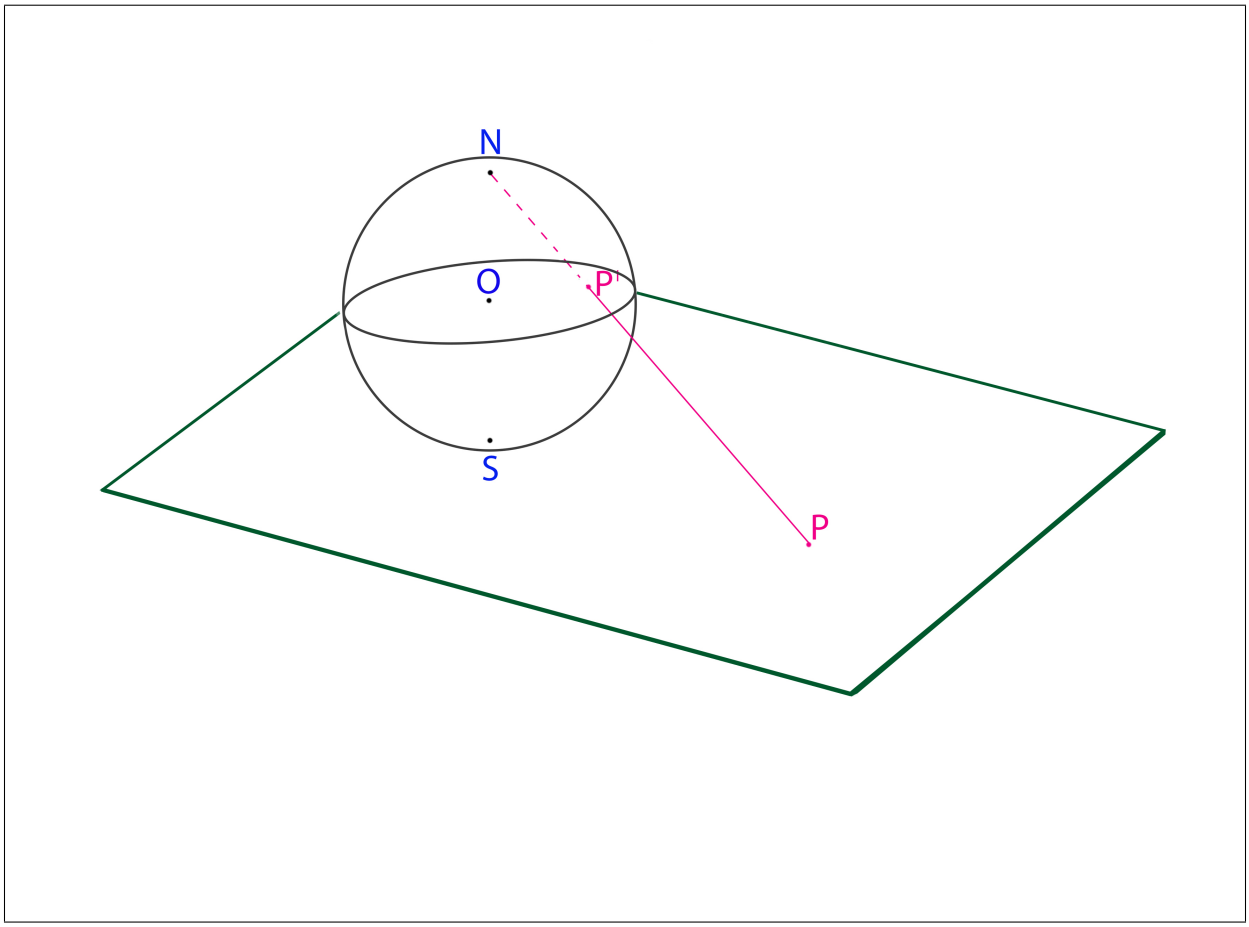


Figure 2.2: Sphere representation.

2.2 Tangent Space

Let $(M,\mathscr{I})$ be a smooth manifold of $\dim(M)=d$. Fix $x\in M$ and let $(U,\xi)\in\mathscr{I}$ be a local chart such that $x\in U$, then we denote

$$\Omega_x(M)=\{\gamma:(-\varepsilon,\varepsilon)\subset\mathbb{R}\to M\,:\,\gamma \text{ is a smooth curve and } \gamma(0)=x,\varepsilon>0\}.$$

We endow $\Omega_x(M)$ with the following equivalence relationship

$$\forall\,\gamma_1,\gamma_2\in\Omega_x(M),\quad \gamma_1\sim\gamma_2 \qquad\Longleftrightarrow\qquad \frac{\partial}{\partial t}(\xi\circ\gamma_1)_{|t=0}=\frac{\partial}{\partial t}(\xi\circ\gamma_2)_{|t=0}.$$

Lemma 2.2.1. *The equivalence relationship introduced above is independent of the local chart* (U,ξ).

Definition 2.2.2 (**Tangent space**). *The quotient space*

$$T_xM=\Omega_x(M)_{/\sim}$$

is called tangent space *of* M *in* x.

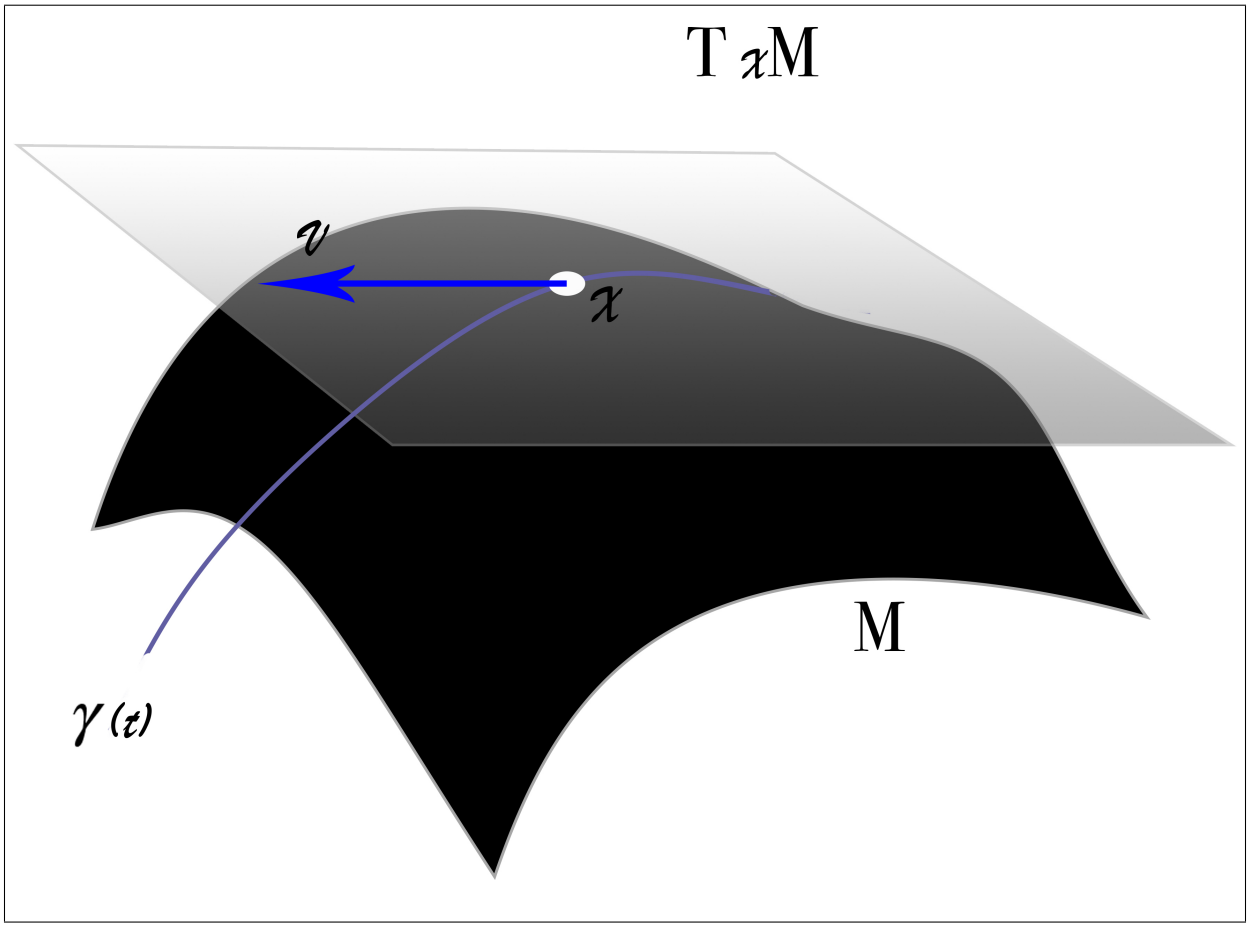


Figure 2.3: Tangent space.

If we fix a local chart $(U,\xi) \in \mathscr{I}$, and define

$$F^{\xi} : T_xM \to \mathbb{R}^d \qquad F^{\xi}(\gamma) = \frac{\partial}{\partial t}(\xi \circ \gamma)_{|t=0},$$

then F^{ξ} endows a vectorial space structure on T_xM. If $v \in T_xM$ we say that v is tangent to M at x.

Lemma 2.2.3. *The vectorial space structure introduced by F^{ξ} is independent of the local chart (U,ξ).*

Definition 2.2.4. *Fixed the local chart $(U,\xi) \in \mathscr{I}$ such that $x \in M$, it is defined a* natural basis *for T_xM induced by the local chart,*

$$\left\{\partial^{\xi}_{1|x}, \cdots, \partial^{\xi}_{d|x}\right\} = \left\{\left(F^{\xi}\right)^{-1}(e_1), \cdots, \left(F^{\xi}\right)^{-1}(e_d)\right\},$$

where $\{e_1, \cdots, e_d\}$ is the canonical base of $\mathbb{R}^d$. We will often drop the apex ξ whenever it will be easy understandable from the context.

Definition 2.2.5 (**Tangent bundle**). *Let*

$$TM = \bigcup_{x\in X} T_xM,$$

that is, the set of all tangent vectors to M. For each $x \in M$ replace $0 \in T_xM$ by 0_x, otherwise the zero tangent vector is in every tangent space. Then each $v \in TM$ is in a unique T_xM, and the projection $\pi : TM \to M$ sends v to x. Thus, $\pi^{-1}(x) = T_xM$.
There is a natural way to make TM a manifold, called the tangent bundle *of M. Let v be tangent to M at some point $x \in M$ and let $(U,\xi) \in \mathscr{I}$ be a local chart such that $x \in U$. Since $v \in T_xM$, there exists a smooth curve γ_v such that*

$$\gamma_v = \left(F^{\xi}\right)^{-1}(v).$$

Then, define $\tilde{U} = \pi^{-1}(U)$, and $\tilde{\xi} : \tilde{U} \subseteq TM \to \mathbb{R}^{2d}$ such that

$$\tilde{\xi} = \left[\xi^1 \circ \pi^{-1} \quad \cdots \quad \xi^d \circ \pi^{-1} \quad \frac{\partial}{\partial t}\left(\gamma_v \circ \xi^1\right)_{t=0} \quad \cdots \quad \frac{\partial}{\partial t}\left(\gamma_v \circ \xi^d\right)_{t=0}\right].$$

Setting

$$\tilde{\mathscr{I}} = \bigcup \left\{ \left(\tilde{U}, \tilde{\xi}\right) : (U,\xi) \in \mathscr{I} \right\},$$

endows the pair $\left(TM, \tilde{\mathscr{I}}\right)$ with a differential structure, $\dim(TM) = 2d$.

Definition 2.2.6 (**Vector field**). *A* vector field *$X \in \mathscr{X}(M)$ is a smooth section of TM, that is, a smooth function $X : M \to TM$ such that $\pi \circ X = \mathrm{id}$. $\mathscr{X}(M)$ is a vector space whose basis is given by $\{\partial_1, \cdots, \partial_d\}$, where $\partial_i : M \to TM$ is the vector field such that $\partial_i(x) = \partial_{i|x}$ for every $x \in M$.*
For every smooth $f : M \to \mathbb{R}$, it holds that

$$X(f)(x) = \frac{\partial}{\partial t}\left[f \circ \left(F^{\xi}\right)^{-1}(X(x))\right]_{t=0}.$$

Moreover, let us define

$$[X,Y](f) := X(Y(f)) - Y(X(f)).$$

Hereafter we will use the notation $\mathscr{F}(M) := \{f : M \to \mathbb{R} : f \in C^{\infty}(M)\}$.

Definition 2.2.7 (**Connection**). *A connection on a smooth manifold M is a function $D : \mathscr{X} \times \mathscr{X} \to \mathscr{X}$ such that*

(i) $D_V W$ is $\mathscr{F}(M)$-linear in V.

(ii) $D_V W$ is $\mathbb{R}$-linear in W.

(iii) $D_V(fW) = (Vf)W + fD_V W$ for every $f \in \mathscr{F}(M)$.

$D_V W$ is called covariant derivative *of W with respect to V for the connection D.*

2.3 Riemannian Manifold

Definition 2.3.1 (Riemannian manifold). *A* Riemannian manifold *is a pair (M,g) where M is a smooth manifold and $g = g_x(\cdot,\cdot)$ is a positive definite inner product on every tangent space T_xM, called* metric tensor*, such that for every vector field $X,Y \in \mathscr{X}(M)$, the map*

$$x \mapsto g_x(X(x),Y(x))$$

is smooth. If we denote $g_{ij}(x) = g_x\left(\partial_i(x),\partial_j(x)\right)$, then the metric tensor g can be expressed by the matrix

$$\{g_{ij}(x)\} = \begin{bmatrix} g_{11}(x) & g_{12}(x) & \cdots & g_{1,d}(x) \\ g_{21}(x) & \cdots & & \\ \vdots & & & \\ g_{d1}(x) & & & g_{dd}(x) \end{bmatrix}.$$

We will denote the inverse of the metric as $g^{-1}(x)$ and its component as $g^{ij}(x)$. Finally, we denote with $\mathbf{g}(x) = \det\{g_{ij}(x)\}$.

From here on, we will use the standard notation of an inner product,

$$g(\cdot,\cdot) = \langle\cdot,\cdot\rangle.$$

Definition 2.3.2 (Riemannian volume). *Let $\mathscr{I} = \{(U_\iota,\xi_\iota) : \iota \in A\}$ be the C^∞ differential structure on the Riemannian manifold M and let $\{\phi_\iota : \iota \in A\}$ be a partition of unity subordinated to $\{U_\iota\}_{\iota\in A}$, as in Proposition 2.1.14. Define the global* Riemannian measure *dx by*

$$dx := \sum_{\iota\in A} \phi_\iota \sqrt{\mathbf{g}}\, d\xi_\iota^1 \cdots d\xi_\iota^d,$$

or, equivalently,

$$\int_M f\,dx = \sum_{\iota\in A} \mathrm{I}(\phi_\iota f; U_\iota),$$

with

$$\mathrm{I}(f;U) := \int_{\xi(U)\subset\mathbb{R}^d} (f\sqrt{\mathbf{g}}) \circ \xi^{-1}\, dx^1 \cdots dx^d.$$

Definition 2.3.3 (Length of a curve). *Let $\gamma\colon [a,b] \subset \mathbb{R} \to M$ be a Lipschitz curve. Let us define the* length *$l(\gamma)$ of the curve γ as*

$$l(\gamma) = \int_a^b \sqrt{g(\dot\gamma(s),\dot\gamma(s))}\, ds,$$

where

$$\dot\gamma(s) = \frac{\partial}{\partial t}\,(\xi\circ\gamma)_{|s=t}.$$

Definition 2.3.4 (Riemannian distance function).
The Riemannian distance function $\mathrm{dist}_M : M \times M \to [0,\infty)$ *is defined as*

$$\mathrm{dist}_M(x,y) = \inf\{l(\gamma) \,:\, \gamma\colon [a,b] \to M \text{ is such that } \gamma(a) = x,\, \gamma(b) = y\}.$$

If M is connected, then (M,g) is a metric space.

Hereafter we will always suppose M to be connected. We define the open metric ball as usual,

$$B_R(x) := \{y \in M \,:\, \mathrm{dist}_M(x,y) < R\}, \qquad R \geq 0,$$

and the volume of the metric ball as

$$vol(B_R(x)) := \int_{B_R(x)} dx.$$

Definition 2.3.5 (Shortest path). *A curve $\gamma\colon [a,b] \subset \mathbb{R} \to M$ is called* shortest path *i if its length is minimal among the curves with the same endpoints, in other words $l(\gamma) \leq l(\hat{\gamma})$ for every curve $\hat{\gamma}\colon [a,b] \subset \mathbb{R} \to M$.*

Definition 2.3.6 (Geodesics and geodesic completeness).
A curve $\gamma\colon [a,b] \subset \mathbb{R} \to M$ is called geodesic *if for every $t \in I$ there is an interval $J \subseteq I$ containing a neighborhood of t such that $\gamma_{|J}$ is a shorter path. In other words, a geodesic is a curve which is locally a distance minimizer.*
A Riemannian manifold (M,g) is said to be geodesically complete *if the metric induced by its distance function* dist_M *is complete, i.e., every Cauchy sequences converge.*

Definition 2.3.7 (Conjugate points). *Two distinct points $x, y \in M$ are said to be conjugate points if there exist two or more distinct geodesic segments having x and y as endpoints.*

Proposition 2.3.8. *The Riemannian distance function is Lipschitz almost everywhere. Moreover, for every pair $x, y \in M$, the function $\mathrm{dist}^2_M(\cdot,\cdot)$ is smooth in a neighborhood of (x,y) if and only if x and y are not conjugate points.*

Proposition 2.3.9. *Let fix $o \in M$ and $r(x) = \mathrm{dist}_M(x,o)$. Then $|\nabla r(x)| \equiv 1$ for almost every $x \in M$.*

Proposition 2.3.10. *Given any tangent vector $v \in T_xM$ there exists a unique geodesic $\gamma_v : I_v \to M$ in M such that*

(i) $\gamma_v'(0) = \frac{\partial}{\partial t}(\xi \circ \gamma)_{|t=0} = v$;

(ii) the domain I_v is the largest possible. Hence, if $\hat{\gamma}\colon J \to M$ is a geodesic such that $\hat{\gamma}'(0) = v$, then $J \subset I$ and $\hat{\gamma} = \gamma_{|J}$.

We call γ_v inextendible.

Proposition 2.3.11. *A Riemannian manifold is geodesically complete if and only if every maximal geodesic is defined on the entire real line.*

Definition 2.3.12 (**Exponential map**). *If $x \in M$, let $\mathfrak{B}_x$ be the set of vectors $v \in T_xM$ such that the inextendible geodesic γ_v, introduced in Proposition 2.3.10, is defined at least on $[0,1]$. The* exponential map *of M at x is the function*

$$\exp_x : \mathfrak{B}_x \subset T_xM \to M$$

such that $\exp_x(v) = \gamma(1)$ *for all* $v \in \mathfrak{B}_x$.

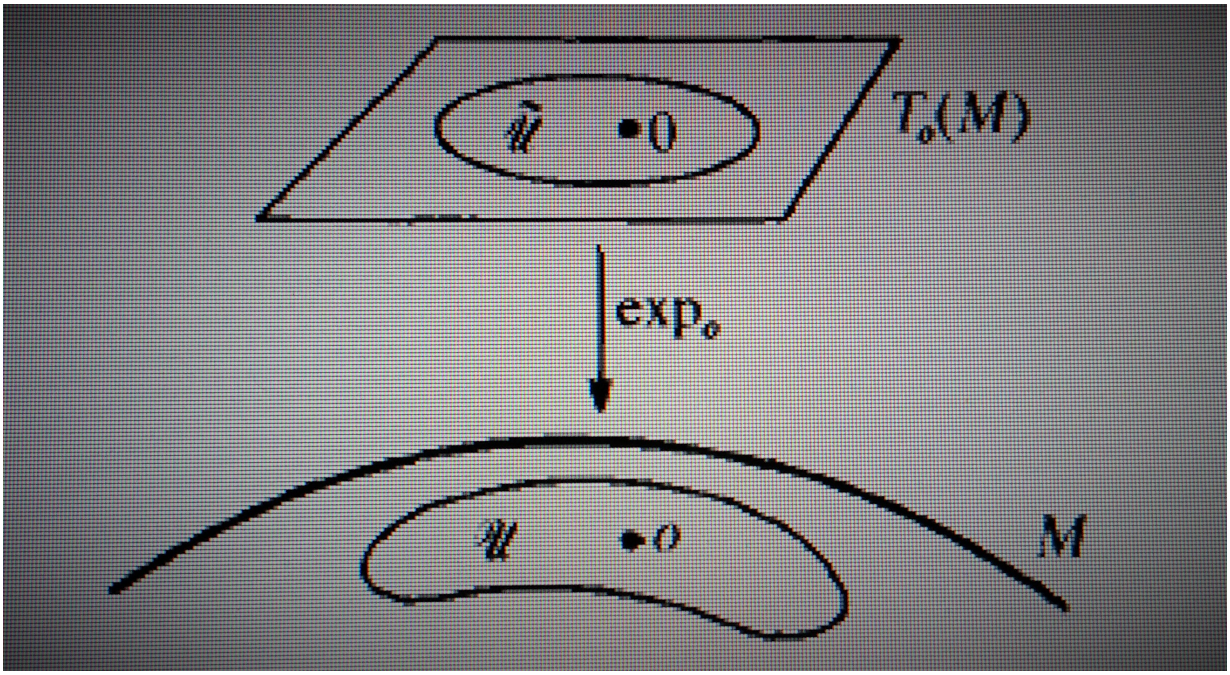


Figure 2.4: Exponential map.

Definition 2.3.13 (**Cut locus and injectivity radius**).
The cut locus *of a tangent space T_xM is defined to be the set of all vectors $v \in T_xM$ such that $t \mapsto \exp_x(tv)$ is a minimizing geodesic for all $t \in [0,1]$ but fails to be minimizing for $t \in [0, 1+\varepsilon)$ for each $\varepsilon > 0$. The* cut locus *of x in M,* $\mathrm{cut}(x)$*, is the image of the cut locus of the tangent space at x under the exponential map.*
The injectivity radius *at x,* inj_x*, is defined as*

$$\mathrm{inj}_x = \inf\{\mathrm{dist}_M(x, \mathrm{cut}(x))\},$$

while the (global) injectivity radius of M is defined as

$$\mathrm{inj}_M = \inf_{x \in M} \mathrm{inj}_x.$$

Proposition 2.3.14 (**Levi-Civita connection**). *With reference to Definition 2.2.7, on a Riemannian manifold (M, g) there exists a unique connection D such that*

(a) $[V,W] = D_V - D_W V$,

(b) $Xg(V,W) = \langle D_X V, W\rangle + \langle V, D_X W\rangle$,

for all $X,V,W \in \mathscr{X}(M)$. *D is called the* Levi-Civita *connection of M.*

Definition 2.3.15 (**Riemannian curvature tensor**). *Let M be a Riemannian manifold with Levi-Civita connection D. The function* $\mathrm{Rm} : \mathscr{X}(M) \times \mathscr{X}(M) \times \mathscr{X}(M) \to \mathscr{X}(M)$ *given by*

$$\mathrm{Rm}_{XY}Z = D_{[X,Y]}Z - [D_X, D_Y]Z$$

is a tensor field on M called the Riemann curvature tensor *of M.*

A two dimensional sub-space Π of the tangent space $T_x(M)$ is called a *tangent plane* to M at x. For tangent vectors v, w define

$$Q(v,w) = \langle v,v\rangle\langle w,w\rangle - \langle v,w\rangle^2.$$

Definition 2.3.16 (**Sectional curvature**). *Let* Π *be a nondegenerate tangent plane to M at x. The number*

$$K(v,w) = \frac{\langle \mathrm{Rm}_{v,w}v, w\rangle}{Q(v,w)}$$

is independent of the choice of the basis v,w *for* Π, *and is called* Sectional curvature $K(\Pi)$ *of M.*

Definition 2.3.17 (**Ricci curvature**). *For any pair* v,w *of tangent vectors in* T_xM, *the* Ricci tensor Ric_M *evaluated at* (v,w) *is defined as the trace of the linear map* $u \mapsto \mathrm{Rm}_{u,v}w$.

Theorem 2.3.18 (**Laplacian comparison Theorem**). *Let M be a geodesically complete Riemannian manifold of dimension* $\dim(M) = d$ *such that*

$$\mathrm{Ric}_M \geq (d-1)\kappa,$$

with $\kappa \in \mathbb{R}$. *Fix a pole* $o \in M$ *and* $r(x) = \mathrm{dist}_M(x,o)$. *Then for every* $x \in M$ *where r is smooth it holds that*

$$\Delta r(x) \leq \begin{cases} (d-1)\sqrt{\kappa}\cot\left(\sqrt{\kappa}r(x)\right) & \text{if } \kappa > 0, \\ \frac{d-1}{r(x)} & \text{if } \kappa = 0, \\ (d-1)\sqrt{|\kappa|}\coth\left(\sqrt{|\kappa|}r(x)\right) & \text{if } \kappa < 0. \end{cases}$$

On the whole manifold, the Laplacian comparison theorem holds in the sense of distributions.

Theorem 2.3.19 (**Bishop-Gromov comparison Theorem**). *Let M be a geodesically complete Riemannian manifold of dimension* $\dim(M) = d$ *such that*

$$\mathrm{Ric}_M \geq (d-1)\kappa,$$

with $\kappa \in \mathbb{R}$. *Let* M_κ *be the complete d-dimensional simply connected space of constant sectional curvature* κ *and* $B_r^\kappa(o)$ *the metric ball centered at the pole o and with radius r. Then, the function*

$$f(r) = \frac{vol(B_r(o))}{vol(B_r^\kappa(o))}$$

is monotone decreasing.

Laplacian cut-offs and applications

As we already observed in Chapter 2, many analytic results in Euclidean setting require the use of compactly supported cut-off functions, essentially to localize differential equations or inequalities or to perform integration by parts arguments. A key feature of d-dimensional Euclidean space is that it is possible to construct cut-offs $\{\phi_R\}$ such that $\phi_R = 1$ on the ball $B_R(o)$, they are supported in the ball $B_{\gamma R}(o)$ and have controlled derivatives up to second order:

$$|\nabla \phi_R| \leq \frac{C}{R}, \qquad |\Delta \phi_R| \leq \frac{C}{R^2}$$

where C is a constant depending only on γ and the dimension. Indeed, such cut-offs can be defined in terms of the distance function r from o, $r(x) = (\sum_i x_i^2)^{1/2}$, as

$$\phi_R(x) = \psi(r(x)/R)$$

where $\psi : \mathbb{R} \to [0,1]$ is smooth, identically 1 in $(-\infty, 1]$ and vanishes in $[\gamma, +\infty)$, and the properties of ϕ_R listed above depend crucially on the fact that the distance function is proper and satisfies

$$|\nabla r(x)| \leq C, \quad |\Delta r(x)| \leq \frac{C}{r(x)} \quad \left(\text{indeed, } |\nabla r(x)| = 1, \quad \Delta r(x) = \frac{d-1}{r(x)} \right).$$

A proper function is often referred to as an exhaustion function, and the existence of Euclidean cut-offs with the above properties is then a consequence of the fact that distance is a well-behaved exhaustion function on $\mathbb{R}^d$.

While in many instances a control on the gradient of the cut-off suffices, in many other significant situations it is actually vital to have an explicit uniform decay of $\Delta \phi_R$ in terms of R. We quote, for example, spectral properties of Schrödinger-type operators (see, e.g., [75]), and, most notably from our point of view, the approximation procedures used in the proof of existence, uniqueness and qualitative and quantitative properties of solutions to the Cauchy problem for the porous and fast diffusion equations ([8], [65], [5], [109]), whose properties we have already shown of an example in Chapter 2 with Theorem 2.0.2.

It follows that the extension of such Euclidean results to the setting of Riemannian manifolds will often depend on the existence of good families of cut-offs and well behaved exhaustion functions.

While it is well known that exhaustion functions with a control on the gradient exist under the only assumption of geodesic completeness (see [50, 52, 102]), uniform bounds on the second order derivatives typically require stronger geometric assumptions. For instance, bounded sectional curvature and a uniform strictly positive lower bound on the injectivity radius allows to construct exhaustion functions with controlled Hessian, see [46],[101, pg. 61] and [31, Proposition 26.49]. In a very recent paper, [90], the authors refine the arguments in [31] and show that the conclusions hold assuming only that the Ricci curvature is bounded and the injectivity radius is strictly bounded away from zero (see Definition 2.3.13).

On the other hand, it was proved in [57, Theorem 2.2] that one can construct families of cut-off functions $\{\phi_R\}$ with a Euclidean like behavior of $|\nabla \phi_R|$ and $|\Delta \phi_R|$ in terms of R, provided the Ricci curvature is nonnegative.

Our studies were originally aimed to try to extend the results obtained in [14], by M.Bonforte, G. Grillo and L. Vazquez, where they consider Cartan-Hadamard manifolds with Ricci curvature (and therefore sectional curvature) bounded from below, under relaxed geometric assumption. In doing so it quickly became clear that one of the main tools was indeed the existence of cut-off functions with an explicit decay rate for the $|\nabla\phi_R|$ and $|\Delta\phi_R|$.

We were thus led to investigate the existence of such cut-offs under more general geometric conditions than those considered in [14], in particular, avoiding hypotheses on the injectivity radius. The above mentioned [57, Theorem 2.2] gives a positive answer in the case of nonnegative curvature, and in [95] it is shown that a good exhaustion function exists if the Ricci curvature is bounded below by a negative constant. This suggests that this may be extended to the manifolds case with suitable, not necessarily constant, Ricci curvature lower bounds.

A substantial part of this Chapter is devoted to carry out this program and to produce both exhaustion functions and sequences of cut-offs on manifolds whose Ricci curvature satisfies the lower bound $\mathrm{Ric} \geq -(d-1)G_\alpha(r)$ in the sense of quadratic forms, for a family of possibly unbounded functions G_α of the distance function $r = r(x)$ from a fixed reference point o, and with an explicit dependence on α for the bounds on the gradient and of the Laplacian.

We believe that these cut-off functions will be useful in a number of situation and the second part of the Chapter is devoted to illustrating several instances, mostly coming from fast and porous media diffusions, where this is indeed the case.

The Chapter is organized as follows. In Section 3.1 we set up notation and give the relevant definitions.

Section 3.2 is devoted to the the main technical results of the Chapter, the existence of $C^\infty(M)$ exhaustion functions, Theorem 3.2.1, and of sequences of Laplacian cut-off under generalized Ricci lower bounds, Corollary 3.2.4 and 3.2.3. Their proofs depend on several other additional results, many of independent interest, which we collect in subsection 3.2.1. We mention in particular Theorem 3.2.5, which generalizes the Li-Yau gradient estimate (see [29, Theorem 7.1]) to functions satisfying a Poisson equation with right hand side depending both on the function itself and on $r(x)$ and under quite general Ricci curvature lower bound, and Proposition 3.2.8 which provides a lower bound for the volume of balls with fixed radius in terms the distance of their center from reference fixed point o, as in [95, Proposition 4.3] for manifolds satisfying suitable Ricci variable curvature lower bounds.

The last two sections are devoted to applications.

In Section 3.3 we present a first direct application of the existence of sequences of Laplacian cut-offs to obtain a generalization of the L^q-properties of the gradient and the self-adjointness of Schrödinge-type operators discussed in [102] and [57] to the class of Riemannian manifolds satisfying our more general Ricci curvature conditions.

Section 3.4 is arguably the second main part of the Chapter. We apply the results of Section 3.2 to study uniqueness L^1-contractivity properties and conservation of mass for the porous diffusion equation as well as uniqueness, weak conservation of mass and extinction time properties for solutions of the fast diffusion equation, which we prove under our usual quote general geometric assumptions.

3.1 Basic definitions and assumptions

Throughout the Chapter, $(M, \langle , \rangle)$ is a complete noncompact d-dimensional Riemannian manifold, and we will often simply refer to it as M. We denote by $r(x) := \mathrm{dist}_M(x,o)$ the the Riemannian distance function from a fixed reference point $o \in M$. The gradient and (negative) Laplacian of a function u on M are denoted by ∇u and Δ, respectively. Recall that, in local coordinates ξ^i, they are given by

$$\nabla u = g^{ij}\frac{\partial u}{\partial \xi_i}\frac{\partial}{\partial \xi^j}, \qquad \Delta u = \frac{\partial}{\partial \xi^i}\left(g^{ij}\sqrt{g}\frac{\partial u}{\partial \xi^j}\right),$$

where $\{g_{ij}\}$ is the matrix of the coefficients of the metric in the coordinates $\{\xi^i\}$, $\{g^{ij}\}$ its inverse and $\mathbf{g} = \det\{g_{ij}\}$.

We let $B_R(p)$ be the geodesic ball of radius R centered at $p \in M$, and with $\partial B_R(p)$ and $vol(B_R(p))$ its boundary and Riemannian volume. When $p = o$ we may omit the center.

We will assume that the Ricci curvature of M satisfies the inequality

$$\mathrm{Ric}_M(\cdot,\cdot) \geq -(d-1)G(r),$$

in the sense of quadratic forms where $G(r) \in C^0([0,\infty))$.

We denote with M_G the d-dimensional model manifold with radial Ricci curvature equals to $-(d-1)G(r)$, namely, the manifold which is diffeomorphic to $\mathbb{R}^d$ and whose metric in spherical coordinates is given by

$$\langle \cdot,\cdot \rangle_G = dr^2 + h^2(r)d\xi^2,$$

where $h(r)$is the solution of the problem

$$\begin{cases} h''(r) = G(r)h(r), \\ h(0) = 0, \\ h'(0) = 1. \end{cases} \tag{3.1.1}$$

Let $V_G(r)$ be the volume of the ball of radius r centered at the pole o of M_G so that

$$V_G(r) = C(d)\int_0^r h^{d-1}(t)dt, \tag{3.1.2}$$

so that, by Laplacian comparison,

$$\Delta r \leq (d-1)\frac{h'(r)}{h(r)}$$

pointwise in the complement of the cut locus of o and weakly on M, and by the Bishop-Gromov comparison theorem, for every $0 \leq R_1 \leq R_2$.

$$\frac{vol(B_{R_2})(o)}{V_G(R_2)} \leq \frac{vol(B_{R_1})(o)}{V_G(R_1)}, \tag{3.1.3}$$

Finally, as in [57], we give the following definition

Definition 3.1.1 (Laplacian cut-off). *M admits a sequence of Laplacian cut-off functions, $\{\phi_n\}_{n\in\mathbb{N}} \subset C_c^\infty(M)$, if $\{\phi_n\}_{n\in\mathbb{N}}$ satisfies the following properties:*

1. *$0 \le \phi_n(x) \le 1$ for all $n \in \mathbb{N}$, $x \in M$;*
2. *for all compact $K \subset M$ there exists $n_0(K) \in \mathbb{N}$ such that for every $n \ge n_0(K)$ it holds $\phi_{n|K} \equiv 1$;*
3. *$\sup_{x\in M} |\nabla \phi_n(x)| \to 0$ as $n \to \infty$;*
4. *$\sup_{x\in M} |\Delta \phi_n(x)| \to 0$ as $n \to \infty$.*

To indicate constants we will preferably use capital letters A, C, D, E, possibly with subscripts, which may change from line to line and, whenever necessary, the dependence of the constants on the relevant parameters will be made explicit.

3.2 On the existence of a sequence of Laplacian cut-off

In this section we collect the technical results which will allows us to prove the existence of Laplacian cut-off functions under relaxed curvature bounds. As already mentioned, we will use these cut-offs in Sections 3.3 and 3.4 below in order to extend and further generalize several different results in functional analysis and PDE's. The main result is Theorem 3.2.1, where, following the proof of [95, Theorem 4.2], we construct C^∞ exhaustion function $\mathfrak{r}$ whose gradient and Laplacian are controlled in terms of explicit functions of the distance function r.

The key ingredients for the proof are Theorem 3.2.5, a generalization of Li-Yau gradient estimates which permits to obtain a control on the gradient of solutions of a Poisson equation again in terms of the distance function r and the function G which bounds the curvature from below, and Proposition 3.2.8 which gives a lower bound on the volume of balls with fixed radius in terms of the distance of their center from the reference point o. In Corollary 3.2.4 we use the exhaustion function of Theorem 3.2.1 to construct a sequence $\{\phi_n\}_{n\in\mathbb{N}}$ of Laplacian cut-offs with support contained in a suitable increasing exhaustion of M. Finally, in 3.2.3 we specialize the construction to obtain cut-offs supported in geodesic balls and show that, when $\alpha = 2$, which corresponds to an almost Euclidean situation, it is possible to construct cut-offs for which, as in Euclidean space, are equal to 1 on a ball of radius $R > 0$ and supported in a ball of radius γR with $\gamma > 1$ arbitrarily close to 1. This is obtained using a specific construction modelled on the proof of [30, Theorem 6.33], which basically hinges on the fact that when $\alpha = 2$ the Laplacian of the distance function satisfies $\Delta r \le Cr^{-1}$ weakly on the whole manifold.

Theorem 3.2.1. *Let $\mathrm{Ric}_M(\cdot,\cdot) \ge -(d-1)\frac{\kappa^2}{(1+r^2)^{\alpha/2}}\langle\cdot,\cdot\rangle$ in the sense of quadratic forms, with $\alpha \in [-2,2]$ and $o \in M$ fixed. Then there exists an exhaustion function $\mathfrak{r} : M \to [0,\infty)$, $\mathfrak{r} \in C^\infty(M)$, and positive constants $D_{i,\alpha}$ such that*

- ***Case** $\alpha \in [-2,2)$:*

  (1) $D_{1,\alpha} r^{1-\alpha/2}(x) \le \mathfrak{r}(x) \le D_{2,\alpha} \max\{1; r^{1-\alpha/2}(x)\}$, for every $x \in M$,

(2) $|\nabla \mathfrak{r}| \leq \frac{D_{3,\alpha}}{r^{\alpha/2}}$, *for every* $x \in M \setminus \overline{B}_1(o)$,

(3) $|\Delta \mathfrak{r}| \leq \frac{D_{4,\alpha}}{r^{\alpha}}$, *for every* $x \in M \setminus \overline{B}_1(o)$.

- ***Case*** $\alpha = 2$***:***

(1') $D_{1,2} \max\{1+\log(r(x));0\} \leq \mathfrak{r}(x) \leq D_{2,2} \max\{1+\log(r(x));1\}$, *for every* $x \in M$,

(2') $|\nabla \mathfrak{r}| \leq \frac{D_{3,2}}{r}$, *for every* $x \in M \setminus \overline{B}_1(o)$,

(3') $|\Delta \mathfrak{r}| \leq \frac{D_{4,2}}{r^2}$, *for every* $x \in M \setminus \overline{B}_1(o)$.

Proof. Let us observe that $r(x)$ is not necessarily smooth everywhere but it is Lipschitz on all of M with uniform unitary Lipschitz constant and then it is possible to uniformly approximate $r(x)$ by a smooth function $r_\varepsilon(x)$ such that $|r_\varepsilon(x) - r(x)| \leq \varepsilon$ and $|\nabla r_\varepsilon(x)| \leq 1+\varepsilon$ for every $x \in M$ and $\varepsilon > 0$ fixed, see [52, Section 2], which is enough for our purpose since every ball with respect to the Riemannian distance $r(x)$ contains and is contained by a ball with respect to the approximating function $r_\varepsilon(x)$. Thus, without loss of generality, hereafter we will consider $r(x)$ to be C^∞ on $M \setminus \{p\}$.

We first prove the Theorem for $\alpha \in [0,2)$. Let $\omega_R : B_R(o) \setminus \overline{B}_{1/2}(o) \to [0,1]$ be such that

$$\begin{cases} \Delta \omega_R(x) = \frac{A_1^2 C^2}{r^\alpha(x)} \omega_R(x), \\ \omega_{R|\partial B_{1/2}} \equiv 1, \\ \omega_{R|\partial B_R} \equiv 0, \end{cases}$$

where $A_1 = (1-\alpha/2)/\sqrt{2}$ and $C > 0$ is a constant that is chosen like in Remark 3.2.2. By the maximum principle, $\{\omega_{R_n}\}$ is an increasing and bounded family of functions for every $x \in M \setminus \overline{B}_{1/2}(o)$ as $R_n \to \infty$, and therefore there exists the point-wise limit function

$$\omega(x) := \lim_{R \to \infty} \omega_R(x).$$

By L^p and Schauder estimates, there exists a subsequence which converges in $C^\infty(\overline{B}_{R_n} \setminus B_{1/2})$ for every n, so that $\omega \in C^\infty(M \setminus \overline{B}_{1/2})$ and

$$\begin{cases} \Delta \omega(x) = \frac{A_1^2 C^2}{r^\alpha(x)} \omega(x), \\ \omega_{|\partial B_{1/2}} \equiv 1, \\ 0 < \omega < 1 \qquad\qquad \text{on } M \setminus \overline{B}_{1/2}(o). \end{cases}$$

Integrating by parts and using $\omega_{R|\partial B_R} = 0$, we get

$$\begin{aligned}\int_{B_R\setminus B_{1/2}} \mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R\Delta\omega_R &= \int_{\partial B_{1/2}} \omega_R\frac{\partial\omega_R}{\partial\eta} - \int_{B_R\setminus B_{1/2}} \langle\nabla(\mathrm{e}^{Cr(x)^{1-\alpha/2}}\omega_R),\nabla\omega_R\rangle \\ &= \int_{\partial B_{1/2}} \frac{\partial\omega_R}{\partial\eta} - \int_{B_R\setminus B_{1/2}} \left(\frac{(1-\alpha/2)C}{r^{\alpha/2}(x)}\right)\mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R\langle\nabla r(x),\nabla\omega_R\rangle \\ &\quad - \int_{B_R\setminus B_{1/2}} \mathrm{e}^{Cr^{1-\alpha/2}(x)}|\nabla\omega_R|^2 \\ &\leq A_2 - \int_{B_R\setminus B_{1/2}} \left(\frac{(1-\alpha/2)C}{r^{\alpha/2}(x)}\right)\mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R\langle\nabla r(x),\nabla\omega_R\rangle \\ &\quad - \int_{B_R\setminus B_{1/2}} \mathrm{e}^{Cr^{1-\alpha/2}(x)}|\nabla\omega_R|^2,\end{aligned} \tag{3.2.1}$$

where the constant A_2 is independent of R by elliptic estimates, since ω_R is uniformly bounded for every R and $\omega_{R|\partial B_{1/2}} \equiv 1$. Then,

$$\begin{aligned}\int_{B_R\setminus B_{1/2}} \frac{A_1^2C^2}{r^\alpha(x)}\mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R^2 &\leq A_2 - \int_{B_R\setminus B_{1/2}} \left(\frac{(1-\alpha/2)C}{r^{\alpha/2}(x)}\right)\mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R\langle\nabla r(x),\nabla\omega_R\rangle \\ &\quad - \int_{B_R\setminus B_{1/2}} \mathrm{e}^{Cr^{1-\alpha/2}(x)}|\nabla\omega_R|^2 \\ &\leq A_2 + \int_{B_R\setminus B_{1/2}} \left(\frac{(1-\alpha/2)C}{r^{\alpha/2}(x)}\right)\mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R|\nabla\omega_R| \\ &\quad - \int_{B_R\setminus B_{1/2}} \mathrm{e}^{Cr^{1-\alpha/2}(x)}|\nabla\omega_R|^2 \\ &\leq A_2 + \int_{B_R\setminus B_{1/2}} \left(\frac{(1-\alpha/2)^2C^2}{4r^\alpha(x)}\omega_R^2 + |\nabla\omega_R|^2\right)\mathrm{e}^{Cr^{1-\alpha/2}(x)} \\ &\quad - \int_{B_R\setminus B_{1/2}} \mathrm{e}^{Cr^{1-\alpha/2}(x)}|\nabla\omega_R|^2 \\ &= A_2 + \int_{B_R\setminus B_{1/2}} \frac{(1-\alpha/2)^2C^2}{4r^\alpha(x)}\mathrm{e}^{Cr^{1-\alpha/2}(x)}\omega_R^2.\end{aligned}$$

It follows that

$$\int_{B_R\setminus B_{1/2}} \frac{A_1^2C^2}{2r^\alpha(x)} \mathrm{e}^{Cr^{1-\alpha/2}(x)} \omega_R^2 \leq A_2,$$

and then, by letting $R \to \infty$,

$$\int_{M\setminus B_{1/2}} \frac{A_1^2C^2}{2r^\alpha(x)} \mathrm{e}^{Cr^{1-\alpha/2}(x)} \omega^2 \leq A_2.$$

Let $x \in M \setminus \overline{B}_1$ and $y \in B_{1/4}(x) \subset \left(M \setminus \overline{B}_{3/4}\right) \subset \left(M \setminus \overline{B}_{1/2}\right)$. By the triangle inequality,

$$r(x) - 1/4 \leq r(y) \leq r(x) + 1/4$$

and then

$$\begin{aligned}
&\frac{A_1^2C^2\mathrm{e}^{C(r(x)-1/4)^{1-\alpha/2}}}{2(r(x)+1/4)^\alpha} \int_{B_{1/4}(x)} \omega^2(y) \leq \int_{B_{1/4}(x)} \frac{A_1^2C^2}{2r^\alpha(y)} \mathrm{e}^{Cr^{1-\alpha/2}(y)} \omega^2(y) \\
&\leq \int_{M\setminus B_{1/2}(o)} \frac{A_1^2C^2}{2r^\alpha(x)} \mathrm{e}^{Cr^{1-\alpha/2}(x)} \omega^2 \leq A_2,
\end{aligned}$$

namely

$$\begin{aligned}
\int_{B_{1/4}(x)} \omega^2(y) &\leq \frac{2A_2(r(x)+1/4)^\alpha}{A_1^2C^2} \mathrm{e}^{-C(r(x)-1/4)^{1-\alpha/2}} \\
&\leq \frac{2^{\alpha+1}A_2r^\alpha(x)}{A_1^2C^2} \mathrm{e}^{-2^{-(1-\alpha/2)}Cr^{1-\alpha/2}(x)}.
\end{aligned}$$

By Theorem 3.2.5 and Corollary 3.2.7 applied with

$$\begin{aligned}
&\zeta = r,\ \ f_1(r) = \frac{A_1^2C^2}{r^\alpha},\ \ f_2(\omega) = \omega, \\
&G(r) = \frac{\kappa^2}{(1+r^2)^{\alpha/2}}, \qquad\qquad R_0 = 1/2, R_1 = 3/4, t = 1/8, \gamma = \infty,
\end{aligned}$$

we deduce that

$$|\nabla \log \omega(y)| \leq C_3(d, \kappa, \alpha)$$

on $M \setminus \overline{B}_{3/4} \supset B_{1/4}(x)$, and then, letting σ be a geodesic parametrized by arc length connecting x to y, from the path integral

$$|\log \omega(x) - \log \omega(y)| = \int_0^{\mathrm{dist}(x,y)} |\nabla \log(\sigma(s))| ds \leq \frac{C_3}{4},$$

we infer that $\omega(y) \geq e^{-C_3/4}\omega(x)$, and that implies

$$\mathrm{e}^{-C_3/2} Vol(B_{1/4}(x))\omega^2(x) \leq \frac{2^{\alpha+1}A_2r^\alpha(x)}{A_1^2C^2} \mathrm{e}^{-2^{-(1-\alpha/2)}Cr^{1-\alpha/2}(x)},$$

namely

$$\omega^2(x) \leq \frac{2^{\alpha+1}A_2 r^{\alpha}(x)e^{C_3/2}}{A_1^2C^2}\left(vol(B_{1/4}(x))\right)^{-1} e^{-2^{-(1-\alpha/2)}Cr^{1-\alpha/2}(x)}.$$

By Proposition 3.2.8, and by the fact that $r^{\alpha/2} \leq A_3 e^{A_3 r^{1-\alpha/2}}$, we conclude that

$$\omega(x) \leq C_4 e^{-C_5 r^{1-\alpha/2}(x)}, \qquad \text{on } M \setminus \overline{B}_1(o), \tag{3.2.2}$$

with

$$C_4 = \sqrt{\frac{2^{\alpha+1}A_2A_3e^{C_3/2}}{A_1^2C^2\bar{C}_1}}, \qquad C_5 = \frac{2^{-(1-\alpha/2)}C - \bar{C}_2}{2} - A_3, \tag{3.2.3}$$

and where $\bar{C}_1$ and $\bar{C}_2$ are the constant that appears in the statement of Proposition 3.2.8.

Extend now $\omega(x)$ on all of M fixing $\omega(x) \equiv 1$ for every $x \in B_{1/2}(o)$ and define

$$\mathfrak{r}(x) := (\eta(x) - 1)\log(\omega(x)) + \eta(x),$$

with $\eta \in C^{\infty}$, $\eta(x) \equiv 1$ on $B_{1/2}(o)$ and $\eta(x) = 0$ on $M \setminus \overline{B}_1(o)$. Observe that $0 < E_1 \leq h(x) \leq E_2$ on $\overline{B}_1(o)$ and in particular $\mathfrak{r}(x) \geq E_1 r^{1-\alpha/2}$ on $\overline{B}_1(o)$.

Fix $x \in M \setminus \overline{B}_1(o)$ and let $\sigma_o : [0, r(x)] \to M$ be a geodesic parametrized by arc length joining o and x. Then

$$\begin{aligned}
|\mathfrak{r}(x) - \mathfrak{r}(o)| = \int_0^{r(x)} |\nabla \mathfrak{r}(\sigma_o(s))| ds &= \int_{1/2}^{r(x)} |\nabla \mathfrak{r}(\sigma_o(s))| ds \\
&= \int_{1/2}^{1} |\nabla \mathfrak{r}(\sigma_o(s))| ds + \int_1^{r(x)} |\nabla \mathfrak{r}(\sigma_o(s))| ds \\
&\leq C_6 + \int_1^{r(x)} \frac{C_7}{s^{\alpha/2}} ds \\
&\leq C_8 r^{1-\alpha/2}(x),
\end{aligned} \tag{3.2.4}$$

where we used again Theorem 3.2.5 and Corollary 3.2.7 applied with

$$\zeta = r, \;\; f_1(r) = \frac{A_1^2C^2}{r^{\alpha}}, \;\; f_2(\omega) = \omega, \tag{3.2.5}$$

$$G(r) = \frac{\kappa^2}{(1+r^2)^{\alpha/2}}, \qquad R_0 = 1/2, R_1 = s, t = t(s) \equiv \frac{1}{4}, \gamma = \infty, \tag{3.2.6}$$

with t chosen in such a way that $(1-t)R_1 = (1-t)s > 1/2 = R_0$, uniformly for every $s > 1$. Observe that C_6 can be chosen independent of x. Henceforth,

$$\mathfrak{r}(x) \leq \max\{E_2; (1+C_8)r^{1-\alpha/2}\} \quad \text{on } M. \tag{3.2.7}$$

On the other hand, since $C_5 - \log(C_4) > 0$ by Remark 3.2.2, from inequality (3.2.2) we have that

$$\begin{aligned}
\mathfrak{r}(x) = -\log(\omega(x)) \geq C_5 r^{1-\alpha/2} - \log(C_4) &\geq (C_5 - \log(C_4))r^{1-\alpha/2} \quad \text{on } M \setminus \overline{B}_1(o) \\
&\geq \min\{E_1; C_5 - \log(C_4)\} r^{1-\alpha/2} \quad \text{on } M,
\end{aligned} \tag{3.2.8}$$

and putting together the above inequality (3.2.8) with (3.2.7) we conclude that

$$D_{1,\alpha} r^{1-\alpha/2}(x) \leq \mathfrak{r}(x) \leq D_{2,\alpha} \max\{1; r^{1-\alpha/2}(x)\} \quad \text{for every } x \in M.$$

Finally, for every $x \in M \setminus \overline{B}_1(o)$, we have

(i) $|\nabla \mathfrak{r}| = \left|\frac{\nabla \omega}{\omega}\right|$,

(ii) $|\Delta \mathfrak{r}| \leq \left|\frac{\nabla \omega}{\omega}\right|^2 + \frac{\Delta \omega}{\omega} = \left|\frac{\nabla \omega}{\omega}\right|^2 + \frac{A_1^2 C^2}{r^\alpha(x)}$,

and the last statements of the thesis follow one more time by an application of Theorem 3.2.5 and Corollary 3.2.7 with (3.2.5) and (3.2.6).

To conclude, the cases $\alpha \in [-2,0)$ and $\alpha = 2$ can be proven with suitable modifications of the previous proof. Indeed, observe that for $\alpha \in [0,2)$ we used crucially the lower bound estimate for the volume of ball of fixed radius, $vol\left(B_{1/4}(x)\right)$, that appears in Proposition 3.2.8. Therefore, replacing the exponential function $e^{Cr^{1-\alpha/2}(x)}$ in the integral (3.2.1) with $r^{C[1+(d-1)(1+\sqrt{1+4\kappa^2})]}(x)$, where $C > 0$ is chosen big enough, will do the trick for the case $\alpha = 2$, for example. For the case $\alpha \in [-2,0)$ we need one more remark: the constant C_3 has to be replaced by $\tilde{C}_3(r) = C_3(d,\kappa,\alpha) r^{-\alpha/2}(x)$. The estimates that follow still hold with suitable changes. In 3.2.3 we have

$$\tilde{C}_4(r) = C_4 e^{\frac{r^{-\alpha/2}}{4}} \leq C_4 e^{r^{1-\alpha/2}},$$

and then choosing C big enough such that $C_5 - 1 > 0$ we still recover an upper bound for $\omega(x)$ of the form of (3.2.2). For the lower bound (3.2.7) instead, the estimate comes directly from (3.2.4) where now $\alpha \in [-2,0)$. □

Remark 3.2.2 (On the choice of the constant C in the proof of Theorem 3.2.1.)**.** *If C_4 and C_5 are defined as in* (3.2.3)*, we choose C big enough such that $C_5 > 0$ and $C_5 - \log(C_4) > 0$. We want to stress that all constants that appear in the definition of C_4 and C_5 are independent of the radius R and consequently this independence carries over to C as well.*

Using the exhaustion function of Theorem 3.2.1, it is easy to construct sequences of cut-off functions with explicitly controlled gradient and Laplacian. In the almost Euclidean case where $\alpha = 2$, we actually use a construction inspired by [29] which relies on the fact that the Laplacian of the distance function satisfies the weak inequality $\Delta r \leq Cr^{-1}$ globally on M, and allows to construct cut-offs which are 1 on the ball of radius R and vanish off in a ball of radius γR with γ arbitrarily close to 1.

Corollary 3.2.3. *Let* $\mathrm{Ric}_M(\cdot,\cdot)$ *be as in Theorem 3.2.1. Then, for every $R \geq 1$ when $\alpha \in [-2,2)$, $R > 0$ when $\alpha = 2$, and*

$$\gamma > \Gamma(\alpha,\kappa,d) \geq \begin{cases} \frac{D_{2,\alpha}}{D_{1,\alpha}} \geq 1 & \text{for } \alpha \in [-2,2), \\ 1 & \text{for } \alpha = 2, \end{cases}$$

there exist $\phi : M \to [0,1]$, $\phi \in C_c^\infty(M)$, such that

(i) $\phi_{|B_R(p)} \equiv 1$,

(ii) $supp(\phi) \subset B_{\gamma R}(o)$,

(iii) $|\nabla\phi| \leq \frac{C_1}{R}$,

(iv) $|\Delta\phi| \leq \frac{C_2}{R^{1+\alpha/2}}$,

with C_1, C_2 independent of R. If we choose $R = n$ then we have a sequence of Laplacian cut-offs $\{\phi_n\}_{n\in\mathbb{N}}$ with respect to the metric balls in the sense of Definition 3.1.1.

Proof.

- **Case $\alpha \in [-2, 2)$.**

Let $\mathfrak{r}$, $D_{1,\alpha}$ and $D_{2,\alpha}$ be the function and the constants that appear in the statement of the preceding Theorem, respectively. Define $\Gamma = \frac{D_{2,\alpha}}{D_{1,\alpha}}$, let $\gamma > \Gamma$ be fixed and let $\psi : \mathbb{R} \to [0,1]$ be such that

(i) $\psi(r) \equiv 1$ for $r \leq \frac{D_{2,\alpha}}{D_{1,\alpha}}$, $0 \leq \psi \leq 1$;

(ii) $\sup \psi \subset (-\infty, \theta)$;

(iii) $\psi \in C^\infty$ and $|\psi'| + |\psi''| \leq A_1$.

Then, the function defined by

$$\phi(x) := \psi\left(\frac{\mathfrak{r}(x)}{D_{1,\alpha}R^{1-\alpha/2}}\right),$$

is a cut-off with the desired properties.

- **Case $\alpha = 2$.**

In order to get a better estimate of the constant Γ, for this case we will not use the exhaustion function of Theorem 3.2.1.

Define $a = (d-1)\frac{1+\sqrt{1+4\kappa^2}}{2}$ as in Lemma 3.2.14 and fix $\gamma > 1$. Then there exists a function $u : (0, +\infty) \to \mathbb{R}$ such that

1. $u \in C^\infty((0,+\infty))$ and $u''(r) + \frac{a}{r}u'(r) = \frac{1}{\gamma^{a+1}R^2}$,

2. $u'(r) < 0$ on $[R, \gamma R]$,

3. $u(R) = 1$ and $u(\gamma R)$=0.

Observe that u is precisely the function defined in (5.3.17b) with the constant C_2 given in (3.2.41).

Now let $\omega : \overline{B}_{\gamma R}(o) \setminus B_R(o) \to \mathbb{R}$ satisfy

$$\begin{cases} \Delta\omega = \frac{1}{\gamma^{a+1}R^2}, & \text{on } B_{\gamma R}(o) \setminus B_R(o), \\ \omega_{|\partial B_R} \equiv 1, \\ \omega_{|\partial B_{\gamma R}} \equiv 0, \end{cases}$$

By Proposition 3.2.13 and Lemma 3.2.12, u satisfies the weak inequality

$$\Delta u(r) \geq \frac{1}{\gamma^{a+1}R^2}$$

and, applying the minimum principle to $\omega - u$, we have that

$$\omega \geq u \qquad \text{on } \overline{B}_{\gamma R}(o) \setminus B_R(o). \tag{3.2.9}$$

Next let $x \in B_{\gamma R}(o) \setminus B_R(o)$. Then, for every $y \in B_{\frac{R}{2}}(x)$

$$r(y) \geq r(x) - \frac{R}{2} \geq \frac{R}{2} \geq \mathrm{dist}_M(x,y) = s(y),$$

and for every $y \in B_{\frac{R}{2}}(x)$

$$\mathrm{Ric}_M(\nabla s(y), \nabla s(y)) \geq -\frac{(d-1)\kappa^2}{1+r^2(y)} \geq -\frac{(d-1)\kappa^2}{1+s^2(y)}$$

and therefore

$$\Delta s(y) \leq \frac{a}{s} \qquad \text{in } B_{\frac{R}{2}}(x).$$

Next consider the problem

$$\begin{cases} v''(s) + \frac{a}{s}v'(s) = \frac{1}{\gamma^{a+1}R^2} \\ v'(s) > 0, \\ v(0) = 0, \end{cases}$$

whose solution is

$$v(s) = As^2, \tag{3.2.10}$$

with

$$2A + 2aA = \frac{1}{\gamma^{a+1}R^2},$$

namely

$$A = \frac{1}{2(a+1)\gamma^{a+1}R^2},$$

for which

$$v\left(\frac{R}{2}\right) = \frac{1}{8(a+1)\gamma^{a+1}R^2} < 1. \tag{3.2.11}$$

It follows that $v(s)$ satisfies

(i) $\Delta v(s) \leq \frac{1}{\gamma^{a+1}R^2}$ in $B_{\frac{R}{2}}(x)$,

(ii) $v(s) = \frac{(\gamma-1)^2}{2(a+1)\gamma^{a+1}R^2} < 1$ on $\partial B_{\frac{R}{2}}(x)$.

Let now $\omega : \overline{B}_{\gamma R}(o) \setminus B_R(o) \to \mathbb{R}$ be a function that satisfies

$$\begin{cases} \Delta\omega = \frac{1}{\gamma^{a+1}R^2} & \text{on } B_{\gamma R}(o) \setminus B_R(o), \\ \omega_{|\partial B_R(p)} \equiv 1, \\ \omega_{|\partial B_{\gamma R}(p)} \equiv 0. \end{cases}$$

Similarly, if $x \in B_{\gamma R}(o) \setminus B_R(o)$ then the function $v(y) = v(s(y))$ (where $s(y) = \mathrm{dist}_M(x,y)$) satisfies

$$\Delta v(s) \leq \frac{1}{\gamma^{a+1}R^2} \qquad \text{weakly},$$

and then

$$\Delta\left(\omega(y) - v(s(y))\right) \geq 0 \qquad \text{for every } y \in \Omega = B_{\gamma R}(o) \setminus \overline{B}_R(o) \cap B_{\frac{R}{2}}(x). \tag{3.2.12}$$

Setting

$$\partial\Omega_1 = \overline{B}_{\frac{R}{2}}(x) \cap \partial B_R(o), \qquad \partial\Omega_2 = \overline{B}_{\frac{R}{2}}(x) \cap \partial B_{\gamma R}(o), \qquad \partial\Omega_3 = \partial\Omega \setminus (\partial\Omega_1 \cup \partial\Omega_2),$$

then, by the maximum principle, we have that

$$\omega(y) - v(s(y)) \leq \max\left\{ \left[\omega(y) - v(s(y))\right]_{|\partial\Omega_1}, \left[\omega(y) - v(s(y))\right]_{|\partial\Omega_2}, \left[\omega(y) - v(s(y))\right]_{|\partial\Omega_3} \right\}$$

for every $y \in \Omega$, and using the fact that $s(y) \geq |r(x) - r(y)|$, it follows that

$$\begin{aligned} \omega(y) - v(s(y)) &\leq \max\left\{1 - v(s(y))_{|\partial\Omega_1};\ -v(s(y))_{|\partial\Omega_2};\ \omega(y)_{|\partial\Omega_3} - v(r(x))\right\} \\ &\leq \max\left\{1 - v(|r(x) - r(y)|)_{|\partial\Omega_1};\ 0\ ;\ 1 - v\left(\frac{R}{2}\right)\right\} \\ &\leq \max\left\{1 - v(r(x) - R);\ 0\ ;\ 1 - v\left(\frac{R}{2}\right)\right\} \\ &\leq \max\left\{1 - v\left(\frac{r(x) - R}{2(\gamma-1)}\right);\ 0\ ;\ 1 - v\left(\frac{R}{2}\right)\right\} \\ &= 1 - v\left(\frac{r(x) - R}{2(\gamma-1)}\right). \end{aligned}$$

Since $v(0) = 0$, evaluating at $y = x$, we get

$$\omega(x) \leq 1 - v\left(\frac{r(x) - R}{2(\gamma-1)}\right) \qquad \text{for every } x \in B_{\gamma R}(o) \setminus \overline{B}_R(o). \tag{3.2.13}$$

Combining (3.2.9) with (3.2.13) we have that

$$u(x) \le \omega(x) \le 1 - v\left(\frac{r(x)-R}{2(\gamma-1)}\right) \qquad \text{for every } x \in B_{\gamma R}(o) \setminus \overline{B}_R(o).$$

For $\theta \in [0, \frac{\gamma-1}{2}]$ define

$$\begin{aligned} h_R(\theta) &= u((1+\theta)R) - 1 + v\left(\frac{(\gamma-1-\theta)R}{2(\gamma-1)}\right) \\ &= \frac{1+\frac{\gamma^2-1}{2\gamma^{a+1}(a+1)}}{1-\gamma^{1-a}}\left[(\theta+1)^{1-a}-1\right] + \frac{(\theta+1)^2-1}{2\gamma^{a+1}(a+1)} + \frac{(\gamma-1-\theta)^2}{8\gamma^{a+1}(\gamma-1)^2(a+1)}. \end{aligned}$$

Then $h_R(\theta) = h(\theta)$ is independent of R, monotone decreasing, and, since $h(0) = \frac{(\gamma-1)^2}{2\gamma^{a+1}(2\gamma-1)^2(a+1)} \in (0,1)$, there exists $\theta = \theta(d,\kappa,\gamma) \in (0, \frac{\gamma-1}{2})$ independent of R such that

$$0 < \frac{(\gamma-1)^2}{16\gamma^{a+1}(\gamma-1)^2(a+1)} \le u((1+\theta)R) - 1 + v\left(\frac{r(x)-R}{2(\gamma-1)}\right) < 1. \tag{3.2.14}$$

Finally, let $\psi : [0,1] \to [0,1]$ satisfy

1. $\psi_{|[u((1+\theta)R),1]} \equiv 1$;
2. $\psi_{|[0,1-v((\gamma-1-\theta)R/(\gamma-1))]} \equiv 0$;
3. $\psi \in C^\infty([0,1])$ and $|\psi'| + |\psi''| \le C$, with $C = C(d,\kappa,\gamma)$ independent of R by (3.2.14),

and define

$$\phi = \psi \circ \omega.$$

Recalling that $u(r(x)) \le \omega(x) \le 1 - v\left(\frac{(\gamma-1-\theta)R}{2(\gamma-1)}\right)$, we have that

1. $\phi_{|(\overline{B}_{(1+\theta)R} \setminus B_R)}(x) \equiv 1$,
2. $\phi_{|(\overline{B}_{\gamma R} \setminus B_{(\gamma-\theta)R})}(x) \equiv 0$,
3. $\nabla\phi = \psi'\nabla\omega$,
4. $\Delta\phi = \psi''|\nabla\omega|^2 + \frac{\psi'}{\gamma^{a+1}R^2}$.

We extend ψ to all of M by setting it equal to 1 in B_R, and note that, since

$$0 < \frac{\left(1+\frac{\gamma^2-1}{2(a+1)\gamma^{a+1}}\right)\left[(\gamma-\theta/2)-1\right]}{1-\gamma^{1-a}} + 1 + \frac{(\gamma-\theta/2)^2-1}{2\gamma^{a+1}(a+1)} = u((\gamma-\theta/2)R) \le \omega$$

on $\overline{B}_{((\gamma-\theta/2)R)} \setminus B_{((1+\theta/2)R)}$ independently from R, the required conclusion follows from Theorem 3.2.5, Remark 3.2.6 and Corollary 3.2.7 applied with $\alpha = 2$, $\zeta = r$, $f_1(r) \equiv 1$ and $f_2(\omega) \equiv \frac{1}{\gamma^{a+1}R^2}$. □

In some approximation procedures used in the theory of diffusion, one needs to have sequences of cut-off functions whose zero level sets are compact smooth submanifolds. This is addressed in the next corollary.

Corollary 3.2.4. *Let* $\mathrm{Ric}(\cdot,\cdot)$ *be as in Theorem 3.2.1. Then, for every* $\alpha \in (-2,2]$ *there exists an increasing exhaustion of M by open relatively compact sets* $\{F_n\}_{n\in\subseteq\mathbb{N}}M$ *with smooth boundary and with* $\bar{F}_n \subset F_{n+1}$*, and a sequence of functions,* $\{\phi_n\}_{n\in\mathbb{N}} \subset C_c^\infty(M)$*, such that*

1. $\phi_n \equiv 1$ *on* F_n*;*
2. $0 < \phi_n < 1$ *on* $F_{n+1} \setminus \overline{F}_n$*;*
3. $\phi_n \equiv 0$ *on* ∂F_{n+1} *and* $\mathrm{supp}(\phi_n) = \overline{F}_{n+1}$*;*
4. $\sup_x |\nabla\phi_n(x)| \to 0$*, as* $n \to \infty$*;*
5. $\sup_x |\Delta\phi_n(x)| \to 0$*, as* $n \to \infty$*.*

The sequence $\{\phi_n\}_{n\in\mathbb{N}}$ *is a Laplacian cut-off in the sens of Definition 3.1.1.*

Proof. Let $\mathfrak{r}$ be exhaustion function constructed in Theorem 3.2.1, and let $\alpha \in (-2,2)$. Using (1) in the statement of Theorem 3.2.1, we may write (2) and (3) in the form

$$|\nabla\mathfrak{r}| \le \frac{C_1}{\mathfrak{r}^{\frac{\alpha/2}{1-\alpha/2}}}, \qquad |\Delta\mathfrak{r}| \le \frac{C_2}{\mathfrak{r}^{\frac{\alpha}{1-\alpha/2}}},$$

on $M \setminus \overline{B}_1(o)$. Since $\mathfrak{r} \in C^\infty(M)$, by Sard's theorem we can chose a sequence c_n of regular values of $\mathfrak{r}$ such that $|\frac{c_{n+1}}{c_n} - 2| \le 1/n$. Let $F_n := \{x \in M : \mathfrak{r}(x) < c_n\}$. Then $\{F_n\}_{n\in\mathbb{N}}$ is an exhaustion of M by relatively compact open sets with smooth boundary, such that $\bar{F}_n \subset F_{n+1}$. For every n, let $\psi_n : \mathbb{R} \to [0,1]$ be a smooth real function such that

(a) $\psi_n \equiv 1$ on $(-\infty, c_n]$;

(b) $0 < \psi_n < 1$ on (c_n, c_{n+1});

(c) $\psi_n \equiv 0$ on $[c_{n+1}, +\infty)$;

(d) $|\psi'(s)| \le \frac{A_1}{c_n}$, $|\psi''(s)| \le \frac{A_2}{c_n^2}$.

Then, $\phi_n := \psi_n \circ h$ satisfies the requirements. In particular,

$$|\nabla\psi_n(x)| = |\psi_n'(h(x))||\nabla\mathfrak{r}(x)| \le \frac{D_1}{n^{\frac{1}{1-\alpha/2}}} \xrightarrow{n\to\infty} 0 \text{ for every } \alpha \in [-2,2),$$

$$|\Delta\psi_n(x)| \le |\psi_n''(h(x))||\nabla\mathfrak{r}(x)|^2 + |\psi_n'(h(x))||\Delta h(x)| \le \frac{D_1}{n^{\frac{1+\alpha/2}{1-\alpha/2}}} \xrightarrow{n\to\infty} 0 \text{ for every } \alpha \in (-2,2).$$

The case $\alpha = 2$ is dealt similarly with small changes in the proof. □

3.2.1 Auxiliary results.

In this subsection we collect some results which we used in above constructions. The first one is an extension of the classical gradient Li-Yau estimate which we establish, under rather general Ricci curvature lower bounds, for solutions of Poisson equations with right hand side depending both on the function itself and on the point on the manifold (via an approximate distance function). We belive that this result is of independent interest.

Theorem 3.2.5. *Let* $\mathrm{Ric}_M(\cdot,\cdot) \geq -(d-1)G(r)\langle\cdot,\cdot\rangle$ *on* M *in the sense of quadratic forms, where,* $r = r(x)$ *is the distance function from a fixed point* $o \in M$.
Let $R_1 > R_0 > 0$, $\gamma > 1$ *and let* $\omega : M \setminus \overline{B}_{R_0}(o) \to \mathbb{R}$ *be a* C^2 *function satisfying*

$$\begin{cases} \omega > 0 & \text{on } M \setminus \overline{B}_{R_0}(o), \\ \Delta\omega = f_1(\zeta) f_2(\omega), \end{cases} \tag{3.2.15}$$

where $f_1, f_2 : [0,+\infty) \to \mathbb{R}$ *are* C^1 *functions and* $\zeta : M \to [0,+\infty)$ *is such that* $|\nabla\zeta(x)| \leq L$ *for every* $x \in M$. *Moreover, fix* $t > 0$ *such that* $(1-t)R_1 > R_0$. *Then*

$$\frac{|\nabla\omega|^2}{\omega^2} \leq \max\left\{\Omega_1; \frac{4d\Omega_2 + \sqrt{(4d\Omega_2)^2 + 4\Omega_3}}{2}\right\}, \tag{3.2.16}$$

on $B_{\gamma R_1}(o) \setminus \overline{B}_{R_1}(o)$, *where*

$$\begin{aligned}
\Omega_1 &:= \max\{\omega^{-1} f_1(r) f_2(\omega) : x \in \overline{B}_{(\gamma+t)R_1}(o) \setminus B_{(1-t)R_1}(o)\}; \\
\Omega_2 &:= \frac{A_1}{R_1}\left(\frac{1}{R_1} + 4(d-1)\max\left\{\sqrt{\bar{G}}; \frac{1}{R_1}\right\}\right) + \frac{(2+4d)A_1}{R_1^2} + 2(d-1)\bar{G} \\
&\quad + \max\{2f_1(r)\max\{(\omega^{-1} f_2(\omega) - f_2'(\omega)); 0\} + 2\omega^{-1} L |f_1'(r)|^{2\lambda} |f_2(\omega)| : x \in \mathbf{D}_{\gamma,t,R_1}(o)\}; \\
\Omega_3 &:= \max\left\{\omega^{-1} L |f_1'(r)|^{2(1-\lambda)} |f_2(\omega)| : x \in \mathbf{D}_{\gamma,t,R_1}(o)\right\},
\end{aligned}$$

and

$$\mathbf{D}_{\gamma,t,R_1}(o) := \overline{B}_{(\gamma+t)R_1}(o) \setminus B_{(1-t)R_1}(o), \quad A_1 = A_1(t), \quad \bar{G} := \max\{G(r) : r \in [(1-t)R_1, (\gamma+t)R_1]\}.$$

The parameter $\lambda > 0$ *can be chosen in such a way as to minimize the right hand side of* (3.2.16).

Proof. We adapt some of the ideas in the proof of [29, Theorem 7.1]. Let $t > 0$ be as in the statement, fix $x_i \in \partial B_{\frac{\gamma+1}{2}R_1}(o)$ and consider the ball $B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)$. Since $B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i) \subset B_{(\gamma+t)R_1}(o) \setminus \overline{B}_{(1-t)R_1}(o) \subset M \setminus \overline{B}_{R_0}(o)$, ω satisfies (3.2.15) on $B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)$, so that, defining $v = \log\omega$, we have

$$|\nabla v| = \frac{|\nabla\omega|}{\omega}, \qquad \Delta v = -|\nabla v|^2 + f_1(\zeta) F_2(v), \tag{3.2.17}$$

where $F_2(v) = e^{-v} f_2(e^v) = \omega^{-1} f_2(\omega)$. Set now

$$Q = \vartheta |\nabla v|^2,$$

where the radial function $\vartheta : B_{\frac{\gamma-1}{2}R}(x_i) \to [0,1]$ satisfies

$$\vartheta(y) = \psi(s_i(y)) \quad \text{with } \psi \in C^\infty([0,+\infty)), \quad s_i(y) = \mathrm{dist}_M(y,x_i) \tag{3.2.18}$$

$$\psi_{|[0,\frac{\gamma-1}{2}R_1]}(s_i) \equiv 1, \tag{3.2.19}$$

$$\text{supp } \psi \subset \left[0, \left(\frac{\gamma-1}{2} + t\right) R_1\right), \tag{3.2.20}$$

$$-\frac{A_1(t)}{R_1}\sqrt{\psi} \le \psi' \le 0 \quad \text{on } \left[\frac{\gamma-1}{2}R_1, \left(\frac{\gamma-1}{2} + t\right) R_1\right), \tag{3.2.21}$$

$$|\psi''| \le \frac{A_1(t)}{R_1^2} \quad \text{on } \left[\frac{\gamma-1}{2}R_1, \left(\frac{\gamma-1}{2} + t\right) R_1\right), \tag{3.2.22}$$

and then

$$\vartheta_{|\overline{B}_{\frac{\gamma-1}{2}R_1}(x_i)} \equiv 1,$$
$$\text{supp } \vartheta \subset B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i).$$

The function Q takes on its maximum at some point $q_i \in B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)$. For now, consider q_i not to be a cut point of x_i. Therefore, at q_i we have $\nabla Q = 0$ and $\Delta Q \le 0$. Thus, at q_i,

$$\nabla |\nabla v|^2 = -\vartheta^{-2} Q \nabla \vartheta.$$

and

$$\begin{aligned}
\Delta Q &= \Delta\vartheta |\nabla v|^2 + 2\langle \nabla\vartheta, \nabla|\nabla v|^2\rangle + \vartheta\Delta|\nabla v|^2 \\
&= (\vartheta^{-1}\Delta\vartheta - 2\vartheta^{-2}|\nabla\vartheta|^2)Q + \vartheta\Delta|\nabla v|^2 \\
&= (\vartheta^{-1}\Delta\vartheta - 2\vartheta^{-2}|\nabla\vartheta|^2)Q + 2\vartheta\left(|\mathrm{Hess}(v)|^2 + \langle\nabla\Delta v, \nabla v\rangle + \mathrm{Ric}_M(\nabla v, \nabla v)\right) \\
&\ge (\vartheta^{-1}\Delta\vartheta - 2\vartheta^{-2}|\nabla\vartheta|^2)Q + 2\vartheta\left(|\mathrm{Hess}(v)|^2 + \langle\nabla\Delta v, \nabla v\rangle - (d-1)G(r)|\nabla v|^2\right)
\end{aligned} \tag{3.2.23}$$

where in the last equality we used the Bochner's formula. Note that

$$\begin{aligned}
2\vartheta|\mathrm{Hess}(v)|^2 &\ge \frac{2\vartheta}{d}(\Delta v)^2 \\
&= \frac{2}{d}\vartheta^{-1}(-Q + \vartheta f_1(\zeta) F_2(v))^2.
\end{aligned} \tag{3.2.24}$$

Moreover, for any $\alpha, \beta > 0$,

$$\begin{aligned}
2\vartheta\langle\nabla\Delta v, \nabla v\rangle &= 2\vartheta\langle\nabla(-|\nabla v|^2 + f_1(\zeta)F_2(v)), \nabla v\rangle \\
&= 2\vartheta\langle\nabla(f_1(\zeta)F_2(v)), \nabla v\rangle - 2\vartheta\langle\nabla|\nabla v|^2, \nabla v\rangle \\
&= 2f_1(\zeta)F_2'(v)Q + 2\vartheta^{-1}\langle\nabla\vartheta, \nabla v\rangle Q + 2\vartheta f_1'(\zeta)F_2(v)\langle\nabla\zeta, \nabla v\rangle \\
&\geq 2f_1(\zeta)F_2'(v)Q + 2\vartheta^{-1}\langle\nabla\vartheta, \nabla v\rangle Q - 2\vartheta L|f_1'(\zeta)F_2(v)||\nabla v| \\
&= 2f_1(\zeta)F_2'(v)Q + 2\vartheta^{-1}\langle\nabla\vartheta, \nabla v\rangle Q \\
&\quad - 2\vartheta\sqrt{L}|f_1'(\zeta)|^{(1-\lambda)}|F_2(v)|^{1/2}\left(\sqrt{L}|f_1'(\zeta)|^{\lambda}|F_2(v)|^{1/2}|\nabla v|\right) \\
&\geq 2f_1(\zeta)F_2'(v)Q - \varepsilon^{-1}\vartheta^{-2}|\nabla\vartheta|^2 Q - \varepsilon\vartheta^{-1}Q^2 - \vartheta L|f_1'(\zeta)|^{2(1-\lambda)}|F_2(v)| \\
&\quad - \vartheta L|f_1'(\zeta)|^{2\lambda}|F_2(v)||\nabla v|^2,
\end{aligned}$$

whence, taking $\varepsilon = \frac{1}{4d}$,

$$\begin{aligned}
2\vartheta\langle\nabla\Delta v, \nabla v\rangle \geq 2\left(f_1(\zeta)F_2'(v) - L|f_1'(\zeta)|^{2\lambda}|F_2(v)|\right)Q - 4d\vartheta^{-2}|\nabla\vartheta|^2 Q - \frac{Q^2}{4d}\vartheta^{-1} \\
- L|f_1'(\zeta)|^{2(1-\lambda)}|F_2(v)|\frac{Q}{2}. \qquad (3.2.25)
\end{aligned}$$

Inserting (3.2.24) and (3.2.25) into (3.2.23) and multiplying by ϑ yield

$$\begin{aligned}
\frac{2}{d}(-Q + \vartheta f_1 F_2)^2 - \frac{Q^2}{4d} \leq \big[-\Delta\vartheta + (2+4d)\vartheta^{-1}|\nabla\vartheta|^2 + 2(d-1)G(r)\vartheta \\
-2(f_1F_2' - L|f_1'|^{2\lambda}|F_2|)\vartheta\Big]Q + \vartheta^2 L|f_1'|^{2(1-\lambda)}|F_2|. \qquad (3.2.26)
\end{aligned}$$

If $Q \leq 2\vartheta f_1 F_2$, then $|\nabla v|^2 \leq 2f_1F_2 = 2\omega^{-1}f_1(\zeta)f_2(\omega)$ and (3.2.16) holds. If not,

$$-Q + \vartheta f_1 F_2 \leq -Q/2 \leq 0,$$

and

$$\frac{2}{d}(-Q + \vartheta f_1 F_2)^2 - \frac{Q^2}{4d} \geq \frac{Q^2}{4d}.$$

In this case, using (3.2.26) and the fact that $r(y) \in ((1-t)R_1, (\gamma+t)R_1)$, and setting

$\bar{G} := \max\{G(r) : r \in [(1-t)R_1, (\gamma+t)R_1]\}$ we get

$$\begin{aligned}
Q^2 &\le 4d\left[-\Delta\vartheta + (2+4d)\vartheta^{-1}|\nabla\vartheta|^2 + 2(d-1)G(r(y))\vartheta - 2(f_1F_2' - L|f_1'|^{2\lambda}|F_2|)\vartheta\right]Q \\
&\quad + \vartheta^2 L|f_1'|^{2(1-\lambda)}|F_2| \\
&\le 4d\left[-\Delta\vartheta + (2+4d)\vartheta^{-1}|\nabla\vartheta|^2 + 2(d-1)\bar{G}\vartheta + 2(f_1F_2' - L|f_1'|^{2\lambda}|F_2|)\vartheta\right]Q \\
&\quad + \vartheta^2 L|f_1'|^{2(1-\lambda)}|F_2| \\
&= 4d\left[A_2(d,\kappa,\alpha,t) + 2(d-1)\bar{G}\vartheta + 2f_1(\zeta)(\omega^{-1}f_2(\omega) - f_2'(\omega))\vartheta + 2\omega^{-1}L|f_1'(\zeta)|^{2\lambda}|f_2(\omega)|\vartheta\right]Q \\
&\quad + \vartheta^2 L|f_1'|^{2(1-\lambda)}|F_2| \\
&\le 4d\left[A_2(d,\kappa,\alpha,t) + 2(d-1)\bar{G} + 2f_1(\zeta)\max\left\{\omega^{-1}f_2(\omega) - f_2'(\omega);0\right\} + 2\omega^{-1}L|f_1'(\zeta)|^{2\lambda}|f_2(\omega)|\right]Q \\
&\quad + L|f_1'|^{2(1-\lambda)}|F_2|
\end{aligned} \tag{3.2.27}$$

where

$$A_2(d,\kappa,\alpha,t) = -\Delta\vartheta + (2+4d)\vartheta^{-1}|\nabla\vartheta|^2 \le -\Delta\vartheta + (2+4d)A_1R_1^{-2}, \tag{3.2.28}$$

by (3.2.21). Thus, we have

$$0 \le Q \le \frac{4d\tilde{\Omega}_2 + \sqrt{(4d\tilde{\Omega}_2)^2 + 4\tilde{\Omega}_3}}{2},$$

with

$$\begin{aligned}
\tilde{\Omega}_2 &= A_2(d,\kappa,\alpha,t) + 2(d-1)\bar{G} + 2f_1(\zeta)\max\left\{\omega^{-1}f_2(\omega) - f_2'(\omega);0\right\} + 2\omega^{-1}L|f_1'(\zeta)|^{2\lambda}|f_2(\omega)|, \\
\tilde{\Omega}_3 &= L|f_1'|^{2(1-\lambda)}|F_2|.
\end{aligned}$$

To conclude it remains to show that A_2 is bounded and (3.2.16) will follow. Indeed,

$$\Delta\vartheta = \psi''(s_i) + \psi'(s_i)\Delta s_i,$$

is not identically zero only for $s_i \in \left(\frac{\gamma-1}{2}R_1, (\frac{\gamma-1}{2}+t)R_1\right)$ and since for every $y \in B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)$,

$$\mathrm{Ric}_M(\nabla s_i(y), \nabla s_i(y)) \geq -(d-1)G(r(y)) \geq -(d-1)\bar{G},$$

using Laplacian comparison, $\psi' \leq 0$, (3.2.21) and (3.2.22), we deduce that

$$\begin{aligned}
\Delta\vartheta &\geq \psi''(s_i) + (d-1)\sqrt{\bar{G}}\coth\left(\sqrt{\bar{G}}s_i\right)\psi'(s_i) \\
&\geq \psi''(s_i) + \max\left\{2(d-1)\sqrt{\bar{G}}; \frac{4(d-1)}{(\gamma-1)R_1}\right\}\psi'(s_i) \\
&\geq -\frac{A_1}{R_1^2} - 4(d-1)\max\left\{\sqrt{\bar{G}}; \frac{1}{R_1}\right\}\frac{\sqrt{\psi}A_1}{R_1} \\
&\geq -\frac{A_1}{R_1}\left(\frac{1}{R_1} + 4(d-1)\max\left\{\sqrt{\bar{G}}; \frac{1}{R_1}\right\}\right).
\end{aligned}$$

The above inequality holds pointwise whenever q_i is not a cut point of x_i. If q_i is a cut point, in order to have ϑ smooth in a neighborhood of q_i, we can use a standard argument by Calabi, replacing $s_i(y)$ with its associated upper barrier function $s_{i,\varepsilon,q_i}(y)$ in the definition of ϑ, i.e., $\vartheta(y) = \psi(s_{i,\varepsilon,q_i}(y))$, where $s_{i,\varepsilon,q_i}(y) = \varepsilon + \mathrm{dist}_M(\delta(\varepsilon), y) = \varepsilon + r_{\delta(\varepsilon)}(y)$ and δ is the minimum geodesic joining x_i to q_i. Since ψ is nonincreasing, then q_i is still a maximum for Q and the above estimates hold again. Hence, we proved that on $B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)$

$$\vartheta\frac{|\nabla\omega(x)|^2}{\omega^2(x)} \leq \max\left\{\Omega_{1,i}; \frac{4d\Omega_{2,i} + \sqrt{(4d\Omega_{2,i})^2 + 4\Omega_{3,i}}}{2}\right\}, \tag{3.2.29}$$

where

$$\begin{aligned}
\Omega_{1,i} &:= \max\{\omega^{-1}f_1(\zeta)f_2(\omega) : x \in \overline{B}_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)\}; \\
\Omega_{2,i} &:= A_3(d,\kappa,\gamma,t,\bar{G},R_1) + \max\{2f_1(\zeta)\max\{\omega^{-1}f_2(\omega) - f_2'(\omega); 0\} \\
&\quad + 2\omega^{-1}L|f_1'(\zeta)|^{2\lambda}|f_2(\omega)| : x \in \overline{B}_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)\}, \\
\Omega_{3,i} &:= \max\{\omega^{-1}L|f_1'(\zeta)|^{2(1-\lambda)}|f_2(\omega)| : x \in \overline{B}_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i)\},
\end{aligned}$$

and

$$\begin{aligned}
A_3(d,\kappa,\gamma,t,\bar{G},R_1) :=& \frac{A_1}{R_1}\left(\frac{1}{R_1} + 4(d-1)\max\left\{\sqrt{\bar{G}}; \frac{1}{R_1}\right\}\right) \\
&+ \frac{(2+4d)A_1}{R_1^2} + 2(d-1)\bar{G}.
\end{aligned}$$

Now, by compactness, there exists a finite collection $\{x_i\}_{i=1}^n \subset \partial B_{\frac{\gamma+1}{2}R}(o)$ such that

$$\bigcup_{i=1}^n B_{\left(\frac{\gamma-1}{2}+t\right)R_1}(x_i) \supset \overline{B}_{\gamma R_1}(p) \setminus B_{R_1}(o).$$

Then, choosing $\Omega_1 = \max\{\Omega_{1,i}\}, \Omega_2 = \max\{\Omega_{2,i}\}$ and $\Omega_3 = \max\{\Omega_{3,i}\}$, the thesis follows. □

Remark 3.2.6. *The constant $A_1(t) \to \infty$ as $R_1 \to R_0$. Moreover, the above theorem can be extended easily to the case $\gamma = \infty$ if $\sup G(r) < \infty$ and to the case where ω is defined only on an annulus $B_{\gamma R}(o) \setminus \overline{B}_R(o)$, $R > 1$, namely $\omega : B_{\gamma R}(o) \setminus \overline{B}_R(o) \to \mathbb{R}$ such that*

$$\begin{cases} \omega > 0 & \text{on } B_{\gamma R}(p) \setminus \overline{B}_R(o), \\ \Delta\omega = f_1(\zeta) f_2(\omega). \end{cases}$$

In this latter case, the estimate (3.2.16) *still holds in any inner annulus of the form $B_{(\gamma-\theta)R}(o) \setminus \overline{B}_{(1+\theta)R}(o)$, provided $0 < \theta < \frac{\gamma+1}{2}$, and replacing $\mathbf{D}_{\gamma,t,R_1}(o)$ with $\mathbf{D}_{\gamma,\theta,R}(o) := \overline{B}_{(\gamma-\theta/2)R}(o) \setminus B_{(1+\theta/2)R}(o)$. Note that in this case $\Omega_2 \to \infty$ for $\theta \to 0$, since now the $A_1 = a_1(\theta) \to \infty$ as $\theta \to 0$.*

Corollary 3.2.7. *Let ω as in the previous Theorem 3.2.5 and let $G(r) = \frac{\kappa^2}{(1+r^2)^{\alpha/2}}$ with $\alpha \in [-2,2]$. If*

(i) $\Delta\omega = \frac{\omega}{r^\alpha}$,

or if

(ii) $\Delta\omega \equiv \frac{1}{R_1^\alpha}$ *and* $\omega \geq C > 0$, *with C independent of R_1,*

then

$$\frac{|\nabla\omega|^2}{\omega^2} \leq \frac{A(d,\kappa,\gamma,\alpha,t)}{R_1^\alpha}.$$

Proof. Fix $f_1(\zeta) = f_1(r) = \frac{1}{r^\alpha}$ and $f_2(\omega) = \omega$, and choose $\lambda = \frac{1}{3}$ and $\lambda = \frac{1}{2}$ for $\alpha \in [0,2]$ and for $\alpha \in [-2,0)$, respectively. Then it is just a matter of easy calculations to see that

$$\begin{aligned} \Omega_1 &\leq \frac{A_1}{R_1^\alpha}, \\ \Omega_2 &\leq \frac{A_2}{R_1^{1+\alpha/2}} + \frac{A_3}{R_1^2} + \frac{A_4}{R_1^\alpha} + \frac{A_5}{R_1^{2\lambda(\alpha+1)}} \leq \frac{A_6}{R_1^\alpha}, \\ \Omega_3 &\leq \frac{A_7}{R_1^{2(1-\lambda)(\alpha+1)}}, \end{aligned}$$

from which it follows that

$$\frac{4d\Omega_2 + \sqrt{(4d\Omega_2)^2 + 4\Omega_3}}{2} \leq \frac{A_8}{R_1^\alpha}.$$

If $f_2(\omega) \equiv \frac{1}{R_1^\alpha}$ instead and ω is uniformly bounded from below by a constant C, then

$$\omega^{-1} f_2(\omega) - f_2'(\omega) = \omega^{-1} f_2(\omega) \leq \frac{1}{CR_1^\alpha},$$

and the thesis follows from the same estimates of above. □

We next prove a lower estimate for the volume of ball of a fixed (small) radius in terms of the distance of their center from a fixed point under radial bounds on the Ricci curvature. It generalizes similar estimates known when the Ricci curvature is bounded below by a constant. Note that having a variable lower bound on Ricci makes the geometry no longer homogeneous and therefore requires a significantly more careful analysis.

Proposition 3.2.8. *Suppose that*

$$\mathrm{Ric}_M \geq -(d-1)\frac{\kappa^2}{(1+r(x)^2)^{\alpha/2}}, \quad \alpha \in [-2,2].$$

Then, for every $x \in M \setminus \overline{B_1(o)}$, we have

$$vol(B_{1/4}(x)) \geq \begin{cases} \bar{C}_1 e^{-\bar{C}_2 r^{1-\alpha/2}(x)}, & \text{for } \alpha \in [-2,2), \\ \bar{C}_1 r^{-[1+(d-1)(1+\sqrt{1+4\kappa^2})]}, & \text{for } \alpha = 2. \end{cases}$$

Proof. We will give a direct proof for $\alpha \in [0,2)$ while the case for $\alpha = 2$ can be recovered by small modifications of the following considerations.

Let x be fixed and define $s(y) := \mathrm{dist}_M(y,x)$. Then, by hypothesis it holds that

$$\mathrm{Ric}_M(\nabla s(y), \nabla s(y)) \geq -(d-1)\frac{\kappa^2}{(1+r(y)^2)^{\alpha/2}} \geq -(d-1)\frac{\kappa^2}{(1+|r(x)-s(y)|^2)^{\alpha/2}},$$

namely

$$\mathrm{Ric}_M(\nabla s(y), \nabla s(y)) \geq -(d-1)G(s),$$

with $G(s) = \kappa^2/\left(1+|r(x)-s(y)|^2\right)^{\alpha/2}$. Let $h(s) \in C^2([0,r(x)])$ be the solution of the problem

$$\begin{cases} h''(s) = G(s)y(s), \\ h(0) = 0, \\ h'(0) = 1, \end{cases} \tag{3.2.30}$$

on $[0,r(x)]$, and let $\psi(s) \in C^2([0,r(x)])$ be the solution of the problem

$$\begin{cases} \psi''(s) = \frac{\kappa^2}{(r(x)-s)^\alpha}\psi(s), \\ \psi(0) = 0, \\ \psi'(0) = 1, \end{cases}$$

on $[0,r(x)]$. The existence of ψ follows from Lemma 3.2.9, and, since $\kappa^2/(r(x)-s)^\alpha \geq G(s)$, we can apply Lemma 3.2.10 to get

$$0 \leq h(s) \leq \psi(s) \qquad \text{on } [0,r(x)].$$

Since $r(x) \geq 1$, by Corollary 3.2.11 we have that

$$\frac{vol\left(B_{r(x)}(x)\right)}{V_G(r(x))} \leq \frac{vol\left(B_{1/4}(x)\right)}{V_G(1/4)} \leq \hat{C}_1 vol\left(B_{1/4}(x)\right). \tag{3.2.31}$$

Now, let $\beta_p(t)$ be a minimizing geodesic parametrized by arc length connecting x to o and fix $\bar{o} = \beta(r(x)-1)$. Then, $\bar{o} \in S_1(o) = \partial B_1(o)$ and for every $y \in B_1(\bar{o})$ it holds that

$$\mathrm{dist}_M(y,x) \leq \mathrm{dist}_M(y,\bar{o}) + \mathrm{dist}_M(x,\bar{o}) \leq r(x),$$

namely, $B_{r(x)}(x) \supset B_1(\bar{o})$. Since

$$\min_{q \in S_1(p)} vol(B_1(q)) \geq \hat{C}_2 > 0,$$

we have that

$$\frac{vol\left(B_{r(x)}(x)\right)}{V_G(r(x))} \geq \frac{vol(B_1(\bar{o}))}{\hat{C}_3 \int_0^{r(x)} \psi(t)^{d-1} dt} \geq \frac{\hat{C}_2}{\hat{C}_3 r(x)^{1+\frac{(d-1)\alpha}{4}} \mathrm{e}^{\hat{C}_4 r^{1-\frac{\alpha}{2}}(x)}} \geq \frac{\hat{C}_2}{\hat{C}_3 \mathrm{e}^{\hat{C}_5 r^{1-\frac{\alpha}{2}}(x)}}, \tag{3.2.32}$$

where the right hand side inequality comes from Lemma 3.2.9 and the previous observation. Combining (3.2.31) and (3.2.32) we obtain the required concludion. □

Lemma 3.2.9. *Let consider the following ODE problem on* $[0,r)$, $\alpha \in [-2,2]$,

$$\begin{cases} \psi''(s) = G(s)\psi(s), \\ \psi(0) = 0, \\ \psi'(0) = 1, \end{cases} \tag{3.2.33}$$

with $G(s) = \frac{\kappa^2}{(r-s)^\alpha}$. *Then there exists an unique solution* $\psi \in C^2([0,r))$ *such that* $\psi' > 0$ *on* $[0,r)$ *and*

(i) ***Case*** $\alpha \in [-2,0)$

$$\begin{aligned} \psi(s) \leq\ & C_1(r)\frac{2^{\alpha/2}}{\kappa}\sinh\left(\frac{2\kappa}{2-\alpha}\left[(1+(r-s))^{1-\alpha/2}-1\right]\right) \\ & + C_2(r)\frac{2^{\alpha/2}}{\kappa}\cosh\left(\frac{2\kappa}{2-\alpha}\left[(1+(r-s))^{1-\alpha/2}-1\right]\right). \end{aligned} \tag{3.2.34}$$

(ii) ***Case*** $\alpha \in [0,2)$

$$\psi(s) = C_1(r)\sqrt{r-s}\, I_{\frac{1}{2-\alpha}}\left(\frac{\kappa}{1-\frac{\alpha}{2}}(r-s)^{1-\frac{\alpha}{2}}\right) + C_2(r)\sqrt{r-s}\, K_{\frac{1}{2-\alpha}}\left(\frac{\kappa}{1-\frac{\alpha}{2}}(r-s)^{1-\frac{\alpha}{2}}\right), \tag{3.2.35}$$

where $I_\nu(z), K_\nu(z)$ *are the modified Bessel functions.*

(iii) ***Case*** $\alpha = 2$

$$\psi(s) = C_1(r)\,(r-s)^{\frac{1+\sqrt{1+4\kappa^2}}{2}} + C_2(r)\,(r-s)^{\frac{1-\sqrt{1+4\kappa^2}}{2}}. \tag{3.2.36}$$

Moreover, for $r \geq 1$ *it holds that*

$$\begin{cases} \psi(r) \leq C_3 r^{\alpha/2} e^{C_4 r^{1-\alpha/2}}, & \alpha \in [-2,0], \\ \psi(r) \leq C_3 r^{\alpha/4} e^{C_4 r^{1-\frac{\alpha}{2}}}, & \alpha \in [0,2), \\ \psi(r-1) \leq \frac{r^{1+\sqrt{1+4\kappa^2}}}{\sqrt{1+4\kappa^2}}, & \alpha = 2, \end{cases}$$

with C_3 *and* C_4 *constants that depend only on* α *and* κ.

Proof. (i) **Case** $\alpha \in [-2,0]$.

It is not difficult to prove that the right hand side of (3.2.34) is a subsolution of (3.2.33). From the initial conditions we get that

$$C_1(r) = -\left(\frac{1+r}{2}\right)^{\alpha/2} \cosh\left(\frac{2\kappa}{2-\alpha}\left[(1+r)^{1-\alpha/2}-1\right]\right),$$
$$C_2(r) = \left(\frac{1+r}{2}\right)^{\alpha/2} \sinh\left(\frac{2\kappa}{2-\alpha}\left[(1+r)^{1-\alpha/2}-1\right]\right),$$

and then for $r \geq 1$ it follows that

$$\psi(r) \leq C_3 r^{\alpha/2} \mathrm{e}^{C_4 r^{1-\alpha/2}}.$$

(ii) **Case** $\alpha \in [0,2)$.

By a change of variable $x = r - s$, it is easy to check (see [1, pp. 374-379]) that a general solution of the problem (3.2.33) can be expressed in the form of (3.2.35). Imposing $\psi(0) = 0$ it gives

$$C_1(r) = -C_2(r)\frac{K_{\frac{1}{2-\alpha}}\left(\frac{\kappa}{1-\frac{\alpha}{2}} r^{1-\frac{\alpha}{2}}\right)}{I_{\frac{1}{2-\alpha}}\left(\frac{\kappa}{1-\frac{\alpha}{2}} r^{1-\frac{\alpha}{2}}\right)}.$$

Using the following properties

$$\frac{dI_\nu(z)}{dz}(z) = \frac{1}{2}\left(I_{\nu+1}(z) + I_{\nu-1}(z)\right),$$
$$\frac{dK_\nu(z)}{dz}(z) = \frac{1}{2}\left(K_{\nu+1}(z) + K_{\nu-1}(z)\right),$$

and defining $z_r = \frac{\kappa}{1-\frac{\alpha}{2}} r^{1-\frac{\alpha}{2}}$, we get

$$\begin{aligned} \psi'(0) = &-C_1\left\{\frac{1}{2\sqrt{r}} I_{\frac{1}{2-\alpha}}(z_r) + \frac{\sqrt{r}}{2}\kappa r^{-\alpha/2}\left[I_{\frac{1}{2-\alpha}+1}(z_r) + I_{\frac{1}{2-\alpha}-1}(z_r)\right]\right\} \\ &-C_2\left\{\frac{1}{2\sqrt{r}} K_{\frac{1}{2-\alpha}}(z_r) + \frac{\sqrt{r}}{2}\kappa r^{-\alpha/2}\left[K_{\frac{1}{2-\alpha}+1}(z_r) + K_{\frac{1}{2-\alpha}-1}(z_r)\right]\right\}, \end{aligned}$$

and since $\psi'(0)=1$,

$$C_2(r)=\frac{1}{\frac{\kappa}{2}r^{\frac{1-\alpha}{2}}\left\{K_{\frac{1}{2-\alpha}}(z_r)\left[\frac{I_{\frac{1}{2-\alpha}+1}(z_r)+I_{\frac{1}{2-\alpha}-1}(z_r)}{I_{\frac{1}{2-\alpha}}(z_r)}\right]-\left[K_{\frac{1}{2-\alpha}+1}(z_r)+K_{\frac{1}{2-\alpha}-1}(z_r)\right]\right\}}.$$

Making use of the fact that

$$I_\nu(0)=0, \qquad K_\nu(z)\sim\left(\frac{z}{2}\right)^\nu\Gamma(\nu+1) \quad \text{for } z\to 0,$$
$$I_\nu(z_r)\sim A_{\nu,1}\frac{e^{z_r}}{\sqrt{2\pi z_r}} \quad \text{for } z_r\to\infty, \qquad K_\nu(z_r)\sim A_{\nu,2}e^{-z_r}\sqrt{\frac{\pi}{2z_r}} \quad \text{for } z_r\to\infty,$$

where $\sim$ stands for the asymptotic equivalence, see at the end of Section 5.1, we conclude that

$$\psi(r)=C_2(r)C_5(\alpha)\le C_3 r^{\alpha/4}e^{C_4 r^{1-\frac{\alpha}{2}}} \quad \text{for every } r\ge 1,$$

since $C_2(r)$ is of the same order at infinity of the right hand side.

(iii) **Case $\alpha=2$.**

It is just a matter of easy calculations to verify that ψ satisfies (3.2.36) with

$$C_1(r)=\frac{-r^{\frac{1-\sqrt{1+4\kappa^2}}{2}}}{\sqrt{1+4\kappa^2}}, \quad C_2(r)=\frac{r^{\frac{1+\sqrt{1+4\kappa^2}}{2}}}{\sqrt{1+4\kappa^2}}.$$

Finally, since $\psi''(s)\ge 0$ for every s and $\psi'(0)=1$, then $\psi'>0$. □

The following Sturm-Liouville comparison result, which we state without proof, is at the basis of all comparison results valid under Ricci curvature lower bounds.

Lemma 3.2.10. *Let G be a continuous function on $[0,r]$ and let $\phi,\psi\in C^1([0,\infty))$ with $\phi',\psi'\in \mathrm{AC}((0,\infty))$ be solutions of the problems*

$$\begin{cases}\phi''-G\phi\le 0 & \text{a.e. in } (0,r),\\ \phi(0)=0,\end{cases} \qquad \begin{cases}\psi''-G\psi\ge 0 & \text{a.e. in } (0,r),\\ \psi(0)=0,\\ \psi'(0)>0.\end{cases}$$

If $\phi(s)>0$ for $s\in(0,r)$ and $\psi'(0)\ge\phi'(0)$, then $\psi(s)>0$ in $(0,r)$ and

(i) $\frac{\phi'}{\phi}\le\frac{\psi'}{\psi}$,

(ii) $\phi\le\psi$.

Proof. Since $\psi' > 0$, $\psi > 0$ in a neighborhood of 0. We observe in passing that if G is assumed to be nonnegative, then iterating the differential inequality satisfied by ψ we have

$$\psi'(r) = \psi'(0) + \int_0^r G(s)\psi(s)\,ds,$$

so that ψ' is positive in the interval where $\psi \geq 0$, and we conclude that, in fact, $\psi > 0$ on $(0,\infty)$.

In the general case where no assumption is made on the signum of G, we let

$$\beta = \sup\{t : \psi > 0 \text{ in } (0,t)\}, \qquad \tau = \min\{\beta, T\},$$

so that ϕ and ψ are both positive in $(0,\tau)$. The function

$$\psi'\phi - \psi\phi'$$

is continuous on $[0,\infty)$, vanishes in $r = 0$, and satisfies

$$\left(\psi'\phi - \psi\phi'\right)' = \psi''\phi - \psi\phi'' \geq 0,$$

a.e. in $(0,\tau)$. Thus

$$\psi'\phi - \psi\phi' \geq 0,$$

on $[0,\tau)$ and dividing through by $\psi\phi$ we deduce that

$$\frac{\psi'}{\psi} \geq \frac{\phi'}{\phi} \quad \text{in } (0,\tau).$$

Integrating between ε and r $(0 < \varepsilon < r < \tau)$, yields

$$\phi(r) \leq \frac{\phi(\varepsilon)}{\psi(\varepsilon)}\psi(r)$$

and since

$$\lim_{\varepsilon\to 0^+} \frac{\phi(\varepsilon)}{\psi(\varepsilon)} = \frac{\phi(0)}{\psi(0)} \leq 1,$$

we conclude that in fact

$$\phi(r) \leq \psi(r) \quad \text{in } [0,\tau).$$

Since $\phi > 0$ in $(0,T)$ by assumption, this in turn forces $\tau = T$, for otherwise, $\tau = \beta < T$, and we would have, $\phi(\beta) > 0$, while by continuity, $\psi(\beta) = 0$, which is a contradiction. □

Corollary 3.2.11. *Assume that*

$$\mathrm{Ric}_M \geq -(d-1)G(r(x))$$

in the sense of quadratic forms with G positive and C^1 on $[0,\infty)$ and let h be a solution of the differential inequality

$$\begin{cases} h'' - Gh \geq 0 \\ h(0) = 0, \\ h'(0) = 1. \end{cases}$$

Then

$$\Delta r \leq (d-1)\frac{h'(r(x)}{h(r(x))}$$

pointwise in the complement of the cut-locus of M and weakly on all of M. Moreover, for every $0 \leq R_1 \leq R_2$,

$$\frac{vol(B_{R_2})(o)}{V_G(R_2)} \leq \frac{vol(B_{R_1})(o)}{V_G(R_1)}, \tag{3.2.37}$$

where $V_G(R)$ *is the volume of the ball of radius R centered at o in the model manifold with radial Ricci curvature equal to G, namely,*

$$V_G(R) = c_d \int_0^R h(r)^{d-1} ds.$$

Proof. See [86, Theorems 2.4 and 2.14]. □

Lemma 3.2.12. *Set* $\Omega = M \setminus (\{o\} \cup cut(o))$*, and suppose that*

$$\Delta r(x) \leq \phi(r) \qquad \text{pointwise on } \Omega$$

for some $\phi \in C^0([0,+\infty))$*. Let* $f \in C^2(\mathbb{R})$ *be non-negative and set* $F(x) = F(r(x))$ *on M. Suppose either*

i) $f' \leq 0$, *or*

ii) $f' \geq 0$.

Then, we respectively have

i) $\Delta F \geq f''(r) + \phi(r) f'(r)$;

ii) $\Delta F \leq f''(r) + \phi(r) f'(r)$,

weakly on M.

Proof. See [86, Lemma 2.5]. □

Proposition 3.2.13. *Let* $\mathrm{Ric}_M(\nabla r, \nabla r) \geq -(d-1)\frac{\kappa^2}{1+r^2}$*, then*

$$\Delta r(x) \leq (d-1) C_\kappa r^{-1} \qquad \text{for every } r > 0,$$

in the sense of distributions on all of M, and with $C_\kappa = \frac{1+\sqrt{1+\kappa^2}}{2}$.

Proof. See [86, Theorem 2.4 and Proposition 2.11]. □

Lemma 3.2.14. *For every fixed* $R \geq 1$ *and for every* $\gamma > 1$*, there exists a function* $u : (0,+\infty) \to \mathbb{R}$ *such that*

(i) $u \in C^\infty((0,+\infty))$ *and* $u''(r) + \frac{a}{r} u'(r) = \frac{1}{\gamma^{a+1} r^2}$*, where*

(ii) $u'(r) < 0$ *on* $[R, \gamma R]$,

(iii) $u(R) = 1$ *and* $u(\gamma R) = 0$.

Proof. A general solution of (i) can be written in the form

$$u(r) = C_1 + C_2 r^{1-a} + \frac{r^2}{2\gamma^{a+1}R^2(a+1)}. \tag{3.2.38}$$

Since $u(R) = 1$, then

$$u(r) = C_2(r^{1-a} - R^{1-a}) + 1 + \frac{r^2 - R^2}{2\gamma^{a+1}R^2(a+1)}. \tag{3.2.39}$$

In order to have $u'(r) < 0$ on $[R, \gamma R]$, C_2 has to satisfy

$$C_2 > \frac{1}{a^2-1} R^{a-1}. \tag{3.2.40}$$

But condition $u(\gamma R) = 0$ is achieved if and only if

$$C_2 = \frac{1 + \frac{\gamma^2-1}{2(a+1)\gamma^{a+1}}}{(1-\gamma^{1-a})R^{1-a}}, \tag{3.2.41}$$

and putting together equations (3.2.40) and (3.2.41), we get

$$\frac{\gamma^{a-1}}{\gamma^{a-1}-1} + \frac{\gamma^2-1}{2\gamma^2(a+1)(1-\gamma^{a-1})} > \frac{1}{a^2-1},$$

that is satisfied for every $R \geq 1$ and every $\gamma > 1$. Hence, choosing C_2 as in (3.2.41), the thesis follows. □

3.3 Applications. Gagliardo-Niremberg-type L^q-estimates for the gradient and essential self-adjointness of Schroedinger-type operators

As previously mentioned, in [57, Theorem 2.2], B. Güneysu established the existence of a sequence of Laplacian cut-off assuming that the Ricci curvature is nonnegative, and then deduced a number of deep results using the cut-offs he constructed. All the results in that paper which depend only on the existence of sequences of cut-off functions can be generalized to the geometric setting we consider. By way of example, [57, Theorem 2.3] on L^q properties of the gradient, can be extended as follows.

Let us introduce the space

$$\mathrm{L}^2_\alpha(M) := \left\{ f : M \to \mathbb{R} \ : \ \int_M \frac{|f(x)|^2}{(1+r^2(x))^{\alpha/2}} dx < \infty \right\},$$
$$\|f\|_{2,\alpha} := \left(\int_M \frac{|f(x)|^2}{(1+r^2(x))^{\alpha/2}} dx \right)^{1/2}.$$

Theorem 3.3.1. *Let M be like in Theorem 3.2.1, $\bar\alpha := \min\{\alpha; 0\}$ and let*

$$\mathrm{F}_{\bar\alpha}(M) := \left\{ f | f \in C^2(M) \cap \mathrm{L}^\infty(M) \cap \mathrm{L}^2(M),\ |\nabla f| \in \mathrm{L}^2_{\bar\alpha}(M),\ \Delta f \in \mathrm{L}^2(M) \right\}.$$

Then one has

$$|\nabla f| \in \bigcap_{q\in[2,4]} \mathrm{L}^q(M) \quad \text{for any } f \in \mathrm{F}_{\bar\alpha}(M).$$

More precisely, for all of $f \in \mathrm{F}_{\bar\alpha}(M)$ one has

$$\|\nabla f\|_2^2 = \langle f, -\Delta f \rangle, \qquad \|\nabla f\|_4^4 \le (2+\sqrt{d})^2 \|f\|_\infty^2 \left(\|\Delta\|_2^2 + (d-1)\kappa \|\nabla f\|_{2,\bar\alpha}^2 \right).$$

Proof. We give only a sketch of the proof since it can be adapted easily from the arguments presented in [57]. We also remark that the condition $|\nabla f| \in \mathrm{L}^2_{\bar\alpha}(M)$ is necessary only for $\alpha \in [-2,0)$, since $\mathrm{L}^2(M) \subset \mathrm{L}^2_\alpha(M)$ for every $\alpha \in [0,2]$, and if $f \in \mathrm{L}^2(M)$ and $\Delta f \in \mathrm{L}^2(M)$ then $|\nabla f| \in \mathrm{L}^2(M)$, see [104], from which it can be derived either the global integration by part identity in the thesis's statements.

From [56, Lemma 2] we have the inequality

$$\int_M |\nabla f|^4 \, dx \le (2+\sqrt{d})^2 \|f\|_\infty^2 \left(\int_M |\Delta f|^2 \, dx - \int_M \mathrm{Ric}_M(\nabla f, \nabla f) \, dx \right).$$

Inserting into the above inequality the Laplacian cut-offs $\{\phi_R\}$ of Corollary 3.2.3 and taking into account the Ricci lower bound, we get

$$\int_M |\nabla(\phi_R f)|^4 \, dx \le (2+\sqrt{d})^2 \|\phi_R f\|_\infty^2 \left(\int_M |\Delta(\phi_R f)|^2 \, dx + (d-1)\kappa \int_M \frac{|\nabla(\phi_R f)|^2}{(1+r^2)^{\alpha/2}} \, dx \right).$$

Properties 3. and 4. in the definition of the Laplacian cut-offs and by dominated convergence imply that

$$\lim_{R\to\infty} \int_M |\Delta(\phi_R f)|^2 \, dx = \int_M |\Delta f|^2 \, dx, \qquad \lim_{R\to\infty} \int_M \frac{|\nabla(\phi_R f)|^2}{(1+r^2)^{\alpha/2}} \, dx = \int_M \frac{|\nabla f|^2}{(1+r^2)^{\alpha/2}} \, dx,$$

and the required conclusion follows. □

In another direction, one can investigate the positivity preserving property of Schrödinger operators considered by M. Braverman, O. Milatovic and M. Shubin [15, equation (B.4)], and recently addressed in [57, Section 2.4], namely, assuming that $u \in \mathrm{L}^2(M)$ satisfies

$$(b-\Delta)\,u = \nu \geq 0 \qquad \text{in } D'(M), \tag{3.3.1}$$

with $b > 0$ a positive real number, can one conclude that $u \geq 0$ a.e.? Here the inequality $\nu \geq 0$ means that $\langle \nu, \phi \rangle \geq 0$ for every $\phi \in C_c^\infty(M)$, and is equivalent to the fact that ν is a positive measure. As shown in [15], there is a connection between the positivity preserving property of Schrödinger operators for certain functional classes and the essential self-adjointness of the operator, in particular, the essential self-adjointness of $b-\Delta$ on $C_c^\infty(M)$ can be proved using the fact that the operator is positivity preserving for $L^2(M)$ functions. Since it is well know that Δ is essentially self-adjoint on $C_c^\infty(M)$ whenever M is geodesically complete, Braverman Milatovic and M. Shubin made the following conjecture, [15, Conjecture P],

Conjecture 3.3.2 (Conjecture P)**.** *Let M be geodesically complete. Then*

$$u \in L^2(M) \text{ and } (b-\Delta u) = \nu \geq 0 \Rightarrow u \geq 0 \text{ a.e},$$

and proved that a sufficient condition for the above Conjecture to hold is that M supports a sequence of cut-off functions. As mentioned in the introduction they were able to prove the existence of such cut-offs under the assumption of bounded geometry. It is proved in [57, Section 2.4] that this holds for manifolds with nonnegative Ricci curvature (indeed, it is shown that in that case the positivity preserving property actually holds for functions in L^q for every $q \in [1,\infty]$). As a consequence of our results we are able to further enlarge the class of manifolds for which Conjecture *P* holds.

Proposition 3.3.3. *Let M ba a complete Riemannian manifold such that*

$$\mathrm{Ric}_M(\cdot,\cdot) \geq -(d-1)\frac{\kappa^2}{(1+r^2)^{\alpha/2}},$$

for some $\alpha > -2$. Then Conjecture P holds on M.

3.4 Applications. The Porous Medium Equation (PME) and the Fast Diffusion Equation (FDE) for the Cauchy problem on Riemannian manifolds

Hereafter we consider M to be a geodesically complete manifold of dimension d with

$$\mathrm{Ric}_M(\cdot,\cdot) \geq -(d-1)\frac{\kappa^2}{(1+r^2)^{\alpha/2}}\langle \cdot,\cdot \rangle \tag{3.4.1}$$

in the sense of quadratic forms and with respect to a fixed reference point $o \in M$, with $\kappa \geq 0$ and $\alpha \in [-2,2]$. Moreover, γ will be a fixed positive real value such that $\gamma > \Gamma(\alpha,\kappa,d)$ as in Corollary 3.2.3, and we will use the notation $u^m := |u|^{m-1}u$.

The Cauchy problem on M

$$\begin{cases} \partial_t u(t,x) = \Delta u^m(t,x) & \text{for } x \in (0,+\infty)\times M \\ u(0,x) = u_0(x) & \text{for } x \in M, \end{cases} \tag{3.4.2}$$

which is called Porous Medium Equation (PME) when the exponent $m > 1$ and Fast Diffusion Equation (FDE) when $0 < m < 1$, has been widely studied in the Euclidean setting (see [108] and [109] for detailed surveys), and, in recent years, several papers studied the properties of the solutions of those equations in the Riemannian setting, see for example [14], [53], [78], [110] and [54].

This Section is devoted to extensions and refinements of some results concerning solutions to the PME and the FDE of the Cauchy problem is the setting of a Riemannian manifold satisfying condition (3.4.1), mainly through the use of Laplacian cut-offs. The proofs that we propose here are often adaptations of the original proofs. For example, this is the case, [109, Proposition 9.1] compared to Proposition 3.4.2 and [65, Lemma 3.1] compared to Theorem 3.4.6, but in order to make this Chapter reasonably self contained we will reproduce the more relevant details, whenever appropriate.

In Subsection 3.4.1 we focus on the so called strong solutions of the PME proving L^1-contractivity and conservation of mass properties. In Subsection 3.4.2 we consider instead the FDE equation and generalize a weak-conservation of mass property, first proved in [65] and then extended in [14] to the setting of Cartan-Hadamard manifolds with bounded sectional curvature. We obtain an interesting lower bound on the extinction time $T(u_0)$ which depends explicitly on the lower bound on the Ricci curvature. In particular, when (3.4.1) holds with $\alpha = 2$ and $\kappa \geq 0$ in the Ricci inequality (3.4.1) we get a generalization of the critical exponent m_c (see [108, Section 5]) below which finite time extinsion occurs, which reduces to the Euclidean value for $\kappa = 0$, i.e., for $\mathrm{Ric} \geq 0$. See Remark 3.4.8 below.

It is worth to point out again that the only geometric assumption we make is geodesic completeness and the Ricci curvature lower bound (3.4.1). In particular we do not need hypotheses of topological nature nor to impose conditions on the injectivity radius. In this sense, our results appear a genuine generalizations of previous results obtained on the PME/FDE-Cauchy problem posed in a Riemannian setting.

3.4.1 L^1 contractivity and uniqueness of the strong solution of the PME.

Consider the Cauchy problem (3.4.2) with $m > 1$ and with initial datum u_0 which belongs to $\mathrm{L}^1(M)$.

Definition 3.4.1 (Strong solutions for PME)**.** *Let $u \in C([0,\infty) : \mathrm{L}^1(M))$ be such that*

(i)

$$u(0,x) = u_0; \tag{3.4.3}$$

(ii)

$$u^m \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty) : \mathrm{L}^1(M)) \text{ and } \partial_t u, \Delta u^m \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty)\times M); \tag{3.4.4}$$

(iii)

$$\partial_t u = \Delta(u^m) \text{ a.e. in } (0,+\infty)\times M. \tag{3.4.5}$$

Then u is called strong solution for the Cauchy problem (3.4.2) *of the PME, see [109, Definition 9.1]. In view of the next Proposition we will relax the request on u^m in* (3.4.4) *asking only that*

(ii') $\partial_t u, \Delta u^m \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty)\times M)$ *and*

$$\int_{t_1}^{t_2}\int_{\{x:\, n\le r(x)\le \gamma n\}} |u^m(t,x)|\,dtdx = o(n^{1+\alpha/2}) \text{ as } n\to\infty, \tag{3.4.4'}$$

for every $0<t_1<t_2$, with a fixed $\gamma>\Gamma$, see [109, Remark p.197].

In accordance with [109, Proposition 9.1], we have the following result.

Proposition 3.4.2. *Let u, v be two strong solutions. For every $0<t_1<t_2$ we have*

$$\int_M |u(t_2,x)-v(t_2,x)|dx \le \int_M |u(t_1,x)-v(t_1,x)|dx. \tag{3.4.6}$$

Proof. By (ii'), $\Delta u^m, \Delta v^m \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty)\times M)$ and then it can be applied Kato's inequality [72, Lemma A]

$$-\Delta|u^m-v^m| \le -\mathrm{sgn}(u-v)\Delta(u^m-v^m),$$

and by (3.4.5) we get

$$\frac{d}{dt}|u-v| \le \Delta|u^m-v^m| \qquad \text{in } D'((0,+\infty)\times M),$$

namely,

$$\frac{d}{dt}\int_M \phi(x)|u(t)-v(t)|\,dx \le \int_M \Delta\phi(x)\,|u^m(t)-v^m(t)|\,dx$$

for every $\phi\in C_c^\infty(M)$. Then, integrating with respect to time and choosing $\phi=\phi_n$ a Laplacian cut-off functions as in Corollary 3.2.3, we get

$$\begin{aligned}\int_M \phi_n|u(t_2)-v(t_2)|\,dx &\le \int_M \phi_n|u(t_1)-v(t_1)|\,dx + \int_{t_1}^{t_2}\int_M \Delta\phi_n(x)\,|u^m(t)-v^m(t)|\,dx\\ &\le \int_M \phi_n|u(t_1)-v(t_1)|\,dx\\ &\quad + \|\Delta\phi_n(x)\|_\infty \int_{t_1}^{t_2}\int_{\{x:\, n\le r(x)\le\gamma n\}} |u^m(t)-v^m(t)|\,dx.\end{aligned}$$

Letting $n\to\infty$, the required conclusion follows using (3.4.4') and the estimate

$$\|\Delta_x\phi_n\|_\infty \le C/n^{1+\alpha/2}.$$

□

We have an immediate Corollary.

Corollary 3.4.3. *Let u, v be strong solutions of the Cauchy problem 3.4.2 with the same initial data, $u_0 = v_0$. Then $u = v$ almost everywhere. Moreover, the map $u_0 \mapsto u(t)$ is an ordered contraction in* $\mathrm{L}^1(M)$.

Proposition 3.4.4. *For every $t > 0$ we have*

$$\int_M u(t,x)\,dx = \int_M u_0\,dx.$$

Proof. We have that

$$\frac{d}{dt}\int_M \phi u(t)\,dx = \int_M \Delta\phi u^m\,dx,$$

in $D'(M)$ for every $\phi \in C_c^\infty(M)$. Then, taking Laplacian cut-offs $\phi = \phi_R$ and integrating in time the above equation in $[0,t]$, we get

$$\begin{aligned}\int_M \phi_R u(t)\,dx - \int_M \phi_R u(0)\,dx &= \int_0^t \int_M \Delta\phi_R u^m(s)\,dxdt \\ &\leq \|\Delta\phi_R\|_\infty \int_0^t \int_M |u^m(s)|\,dxdt \\ &\leq \frac{tC}{R^{1+\alpha/2}}\|u^m\|_1.\end{aligned}$$

We conclude letting R going to infinity. □

3.4.2 Weak conservation of mass of the FDE

Consider the Cauchy problem (3.4.2) with $0 < m < 1$ and with initial datum u_0 in $\mathrm{L}^1_{\mathrm{loc}}(M)$. We are finally ready to prove a general extension of Theorem 2.0.2 to Riemannian manifolds which we used as example to highlight the crucial role of the existence of cut-off functions characterized by a well-behaved control on the modulus of their Laplacian. Below we report again the definition of *weak and strong solutions* for the FDE-Cauchy problem adapted to the Riemannian manifold setting.

Definition 3.4.5 (weak and strong solutions for the FDE)**.**
Let $u(t,x) \in C([0,+\infty) : \mathrm{L}^1_{\mathrm{loc}}(M))$ be such that

(i)

$$u(0,x) = u_0, \tag{3.4.7}$$

(ii)

$$\partial_t u = \Delta u^m, \qquad \text{in } D'((0,+\infty)\times M). \tag{3.4.8}$$

Then u is called a weak solution *for the Cauchy problem of the FDE. If moreover u satisfies*

(iii)

$$\partial_t u \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty)\times M), \tag{3.4.9}$$

then u is called a strong solution *(see, [65]). Note that since $0<m<1$ then $u^m \in \mathrm{L}^1_{\mathrm{loc}}(M)$ as well.*

Even if the first part of the proof of the following theorem will be almost identical to **Part I**'s proof of Theorem 2.0.2, for the reader's convenience and since the statement which is going to be proved is more general and differs in several details, we will report all the passages.

Theorem 3.4.6 (Weak conservation of mass)**.** *Let $u(t,x), v(t,x) \in \mathrm{L}^1_{\mathrm{loc}}(M)$. If $u(t,x) \geq v(t,x)$ are weak solutions of* (3.4.2) *for the FDE, then for every $R \geq 1$ if $\alpha \in [-2,2)$, $R>0$ if $\alpha = 2$, and for every $\gamma > \Gamma_\alpha \geq 1$, it holds*

$$\left[\int_{B_R(o)} \big(u(t_2,x)-v(t_2,x)\big)\,dx\right]^{1-m} \leq \left[\int_{B_{\gamma R(o)}} \big(u(t_1,x)-v(t_1,x)\big)\,dx\right]^{1-m} + \mathscr{M}_{R,\gamma}(t_2-t_1), \tag{3.4.10}$$

for every $0 \leq t_1 \leq t_2$, where

$$\mathscr{M}_{R,\gamma} = \frac{C}{R^{1+\alpha/2}} Vol(B_{\gamma R}(o) \setminus B_R(o))^{1-m} > 0, \tag{3.4.11}$$

and where the constant C is independent of u and v but depends only on m, d, κ and γ.

If $u(t,x), v(t,x)$ are strong solutions of (3.4.2) *for the FDE, then it holds*

$$\left[\int_{B_R(o)} \big|u(t_2,x)-v(t_2,x)\big|\,dx\right]^{1-m} \leq \left[\int_{B_{\gamma R(o)}} \big|u(t_1,x)-v(t_1,x)\big|\,dx\right]^{1-m} + \mathscr{M}_{R,\gamma}(t_2-t_1), \tag{3.4.10'}$$

where $\mathscr{M}_{R,\gamma}$ is exactly again (3.4.11).

Proof. In the following, the constant C can change from line to line and let us focus now on the first case, namely $u(t,x) \geq v(t,x)$ being weak solutions.

From (3.4.8), for every nonnegative $\eta \in C_c^\infty(0,\infty)$ and $\psi \in C_c^\infty(M)$ we have that

$$\begin{array}{rcl} \langle \partial_t (u-v), \eta\psi\rangle & = & -\langle u-v, \partial_t \eta\psi\rangle \\ \langle \Delta(u^m - v^m), \eta\psi\rangle & = & \langle u^m - v^m, \eta\Delta\psi\rangle \end{array}$$

in distributions, that is,

$$-\int_0^\infty \int_M \partial_t \eta\,\psi (u-v)\,dtdx = \int_0^\infty \int_M \eta\Delta\psi (u^m - v^m)\,dtdx,$$

namely

$$-\int_0^\infty \partial_t \eta \left(\int_M \psi(u-v)\,dx \right) dt = \int_0^\infty \eta \left(\int_M \Delta\psi(u^m - v^m)\,dx \right) dt$$

and which implies

$$\frac{d}{dt}\int_M \psi(u(t)-v(t))\,dx = \int_M \Delta\psi(u^m - v^m)\,dx \tag{3.4.12}$$

in $D'(0,\infty)$ and in $\mathrm{L}^1_{\mathrm{loc}}(0,\infty)$ as well for every fixed ψ, as a consequence of (3.4.7). Since by concavity

$$(r|r|^{m-1} - s|s|^{m-1}) \le 2^{1-m}(r-s)^m \qquad \text{for all } r \ge s,$$

then (3.4.12) implies

$$\frac{d}{dt}\int_M \psi(u(t)-v(t))\,dx \le 2^{1-m}\int_M |\Delta\psi|(u-v)^m\,dx.$$

We set $g := u - v$. By Holder's inequality, we obtain

$$\frac{d}{dt}\int_M \psi g(t) \le C(\psi)\left[\int_M \psi g(t)\right]^m, \tag{3.4.13}$$

where

$$C(\psi) = \left[2\int_M |\Delta\psi|^{1/(1-m)}\psi^{-m/(1-m)}\right]^{1-m}.$$

Since the function $f_\psi(t) = \int_M \psi g(t)$ has weak derivative in $\mathrm{L}^1_{\mathrm{loc}}$, it is a.e. equal to an AC function, and by standard comparison arguments, for all $t_1, t_2 \ge 0$ and every $\psi \in C_c^\infty(M)$,

$$\left[\int \psi g(t_2)\right]^{1-m} \le \left[\int \psi g(t_1)\right]^{1-m} + (1-m)C(\psi)|t_2 - t_1|. \tag{3.4.14}$$

This will immediately imply the statement, once we prove that $C(\psi) \le M_{R,\gamma} < \infty$.

Consider a function $\psi = \phi_R^b \in C_c^2(m)$, with $b > 2/(1-m)$ and ϕ_R as in Corollary 3.2.3, namely $\phi_R : M \to [0,1]$ is such that

(i) $\phi_{R|B_R(p)} \equiv 1$,

(ii) $\mathrm{supp}(\phi_R) \subset B_{\gamma R}(o)$,

(iii) $|\nabla\phi_R| \le \frac{C}{R}$,

(iv) $|\Delta\phi_R| \le \frac{C}{R^{1+\alpha/2}}$,

where $C = C(d,\kappa,\alpha)$ is independent of R.

We then have,

$$
\begin{aligned}
|\Delta(\psi(x))|^{1/(1-m)}\psi(x)^{-m/(1-m)} &= \phi_R(x)^{-bm/(1-m)}\left|b(b-1)\phi_R^{b-2}|\nabla\phi_R|^2+b\phi_R^{b-1}\Delta\phi_R\right|^{1/(1-m)}\\
&\le [b(b-1)]^{1/(1-m)}\phi_R^{[(b-2)-bm]/(1-m)}\cdot\left||\nabla\phi_R|^2+|\Delta\phi_R|\right|^{1/(1-m)}\\
&\le [b(b-1)]^{1/(1-m)}\phi_R^{[(b-2)-bm]/(1-m)}\cdot CR^{-\frac{1+\alpha/2}{1-m}}.
\end{aligned}
\tag{3.4.15}
$$

An integration over $B_{\gamma R}(o)\setminus B_R(o)$, which contains the support of $|\nabla\phi_R|$ and $\Delta\phi_R$, gives

$$
\begin{aligned}
C(\psi) &= \left[2\int_{B_{\gamma R}(p)\setminus B_R(o)}|\Delta\psi|^{1/(1-m)}\psi^{-m/(1-m)}\right]^{1-m}\\
&\le \frac{C}{R^{1+\alpha/2}}(\mathrm{Vol}(B_{\gamma R}(o)\setminus B_R(o))^{1-m}.
\end{aligned}
$$

Let now $u(t,x), v(t,x)$ be strong solutions instead. According to (3.4.9), $\Delta u^m, \Delta v^m \in \mathrm{L}^1_{\mathrm{loc}}((0,+\infty)\times M)$ so that we can apply Kato's inequality [72, Lemma A] to get

$$
-\Delta|u^m-v^m| \le -\mathrm{sgn}(u-v)\Delta(u^m-v^m), \tag{3.4.16}
$$

and then, using (3.4.8) and arguing as in [65, Theorem 2.3]

$$
\frac{d}{dt}|u-v| \le \Delta|u^m-v^m| \qquad \text{in } D'((0,+\infty)\times M).
$$

The conclusion follows from the same arguments used in the previous steps, and, in particular, from equality (3.4.12). □

Definition 3.4.7 (Extinction Time). *Given an initial condition u_0 for the FDE Problem 3.4.2, we call* extinction time $T = T(u_0)$ *the time $T \in [0,\infty)$, if it exists, such that*

$$
u(t,x) \equiv 0 \qquad \text{for almost every } x \in M
$$

and for every $t \ge T(u_0)$. If there is not such an extinction time T, we set $T = \infty$. See [108] and [109]. The same time T can be called blow-up time *of the diffusivity coefficient $a(t,x) = u^{m-1}(t,x)$, since*

$$
a(t,x) \to \infty \qquad \text{as } u(t,x) \to 0.
$$

Remark 3.4.8. *Let $T(u_0)$ be the extinction time of the solution $u(t,x)$ with initial condition $u_0(x)$, as in the above Definition 3.4.7. Let $v(t,x) \equiv 0$ and $s = 0$. Then, if $\alpha = 2$ in* (3.4.1)*, we have*

$$
T(u_0) \ge \frac{R^2}{C(vol(B_{\gamma R}(o))^{1-m}}\left(\int_{B_R(o)}u_0(x)dx\right)^{1-m}.
$$

Now, from the Bishop-Gromov inequality (3.1.3) *and* (3.1.2) *applied with* $r_1 = \gamma R$, $r_2 = 1$, *we have*

$$vol(B_{\gamma R}(o)) \le CV_G(\gamma R) = C\int_0^{\gamma R} h^{d-1}(t)dt,$$

but since $\tilde{h}(t) = t^{\frac{1+\sqrt{1+4\kappa^2}}{2}}$ *is solution of* (3.2.33) *for* $G(t) = \kappa^2/t^2 \ge \kappa^2/(1+t^2)$, *then by Lemma 3.2.9 and Lemma 3.2.10 we can deduce that* $h(t) \le \tilde{h}(t)$ *and get*

$$T(u_0) \ge \bar{C}\frac{R^2}{R^{\left[1+\left(\frac{1+\sqrt{1+4\kappa^2}}{2}\right)(d-1)\right](1-m)}},$$

whence, letting $R \to \infty$, *we deduce that* $T(u_0) = \infty$ *if*

$$2 - \left[1 + \left(\frac{1+\sqrt{1+4\kappa^2}}{2}\right)(d-1)\right](1-m) > 0,$$

that is, rearranging, provided

$$m > m_c = 1 - \frac{2}{\left[1+\left(\frac{1+\sqrt{1+4\kappa^2}}{2}\right)(d-1)\right]}. \tag{3.4.17}$$

Note that, if $\mathrm{Ric} \ge 0$, *so that we can take* $\kappa = 0$, *we recover the Euclidean constant* $m_c = \frac{d-2}{d}$. *On the other hand, if* $\alpha \in [-2,2)$, $vol(B_R(o))$ *may grow super-polynomially, and, in general we can not deduce a non-extinction property. Observe that, as stated in [54, section 3 - examples 3.1], in a model manifold with radial Ricci curvature* $\mathrm{Ric}(\nabla r, \nabla r) = -(d-1)\frac{\kappa^2}{(1+r^2(x))^{\alpha/2}}$, $\alpha \in (0,2)$, *radial functions satisfy a Sobolev inequality of the form*

$$\|f\|_{2\sigma} \le C\|\nabla f\|_2, \qquad \sigma \in (1, d/(d-2)], \tag{3.4.18}$$

which is a key ingredient for a proof of finite extinction time. According to [14, Theorem 6.1], radial strong solutions of the FDE in such model manifolds which satisfy moreover the Poincaré inequality, i.e., inequality (3.4.18) *for* $\sigma = 1$, *vanish in a finite time* $T(u_0)$ *for every* $m \in (0,1)$, *provided that* $u_0 \in \mathrm{L}^q(M)$ *with* $q \ge d(1-m)/2$.

From Theorem 3.4.6 and Remark 3.4.8, we get

Theorem 3.4.9. *Let* $u(t,x) \in \mathrm{L}^1_{\mathrm{loc}}(M)$ *be a weak solution of* (3.4.2) *for the FDE, then for every* $R \ge 2$ *if* $\alpha \in [-2,2)$, $R \ge 1$ *if* $\alpha = 2$, *and for every* $\gamma > \Gamma \ge 1$, *it holds*

$$\int_{B_R(o)} u(t,x)\,dx \le 2^{1/(1-m)}\left\{\int_{B_{\gamma R}(o)} u(s,x)\,dx + \left(\mathscr{M}_{R,\gamma}|t-s|\right)^{1/(1-m)}\right\},$$

for any $t,s \ge 0$ *and where* $\mathscr{M}_{R,\gamma}$ *is like in* (3.4.11). *If there exists an extinction time* $T(u_0)$, *then it is lower bounded by*

$$T(u_0) \ge \frac{R^{1+\alpha/2}}{C(vol(B_{\gamma R}\setminus B_R))^{1-m}}\left(\int_{B_R} u_0(x)dx\right)^{1-m}.$$

Finally, let us observe that inequality (3.4.10) depends on chosen reference point o, in sharp contrast to the result of Theorem 2.0.2. Thus, in order to prove uniqueness of strong solutions for every $m \in (0,1)$ with the method of [65, Theorem 2.3], the first task is to get rid of that dependency. But this alone is not enough, since a key tool there is the Mean Value Theorem for subharmonic functions. Keeping this into consideration, we can prove the following result.

Theorem 3.4.10. *Let M be a geodesically complete manifold and let u, v be strong solutions for the FDE problem* (3.4.2) *with same initial data, $u_0 = v_0$. If*

(i) $\mathrm{Ric}_M(\cdot,\cdot)$ *satisfies* (3.4.1) *with $\alpha = 2$, then $u \equiv v$ for every $m > m_c$, where m_c is defined as in* (3.4.17)*;*

(ii) $\mathrm{Ric}_M(\cdot,\cdot) \geq 0$*, so that* (3.4.1) *holds with $\kappa = 0$, then $u \equiv v$ for every $m \in (0,1)$.*

Proof. From inequality (3.4.10'), we have

$$\begin{aligned}\int_{B_R(o)} |u(t) - v(t)|\,dx &\leq C\left[\int_{B_{\gamma R}(o)} |u(0) - v(0)|\,dx + \frac{vol(B_{\gamma R}(o))}{R^{\frac{2}{1-m}}} t^{\frac{1}{1-m}}\right] \\ &= C\frac{vol(B_{\gamma R}(o))}{R^{\frac{2}{1-m}}} t^{\frac{1}{1-m}}, \end{aligned} \tag{3.4.19}$$

and observe that the above inequality is valid for both the cases (i) and (ii). From Remark 3.4.8

$$vol(B_{\gamma R}(o)) \leq C(o) R^{1+\left(\frac{1+\sqrt{1+4\kappa^2}}{2}\right)(d-1)},$$

and letting $R \to \infty$ in (3.4.19), the right hand side converges to 0 provided $m > m_c$ and the thesis follows for case (i).

Let us now be in case (ii), namely $\kappa = 0$. Then inequality (3.4.19) is true for every $o \in M$. Set

$$f(t,x) = \int_0^t |u^m - v^m|(s,x)\,ds.$$

By integrating in time in (3.4.16) we get $|u(t) - v(t)| \leq \Delta f(t,x)$ in $D'(M)$ for every $t > 0$. Therefore, f is subharmonic and from [76, Theorem 2.1] it holds that

$$f(t,p) \leq C vol(B_R(p))^{-1} \int_{B_R(o)} f(t,x)\,dx, \tag{3.4.20}$$

for every $R > 0$ and for every $o \in M$, with $C = C(d)$. Moreover, from Hölder inequality and (3.4.19) we deduce that

$$\begin{aligned}\int_{B_R(o)} f(t,x)\,dx &\leq C\int_0^t \int_{B_R(o)} |u(t) - v(t)|^m\,dx \\ &\leq C\int_0^t vol(B_R(o))^{1-m}\left(\int_{B_R(o)} |u(s) - v(s)|\right)^m ds \\ &\leq C vol(B_R(o))^{1-m}\int_0^t \frac{vol(B_{\gamma R}(o))^m}{R^{\frac{2m}{1-m}}} s^{m/(1-m)}\,ds \\ &\leq C(\gamma)\frac{vol(B_R(o))}{R^{\frac{2m}{1-m}}} t^{1/(1-m)}, \end{aligned}$$

and inserting the last inequality into (3.4.20) and letting $R \to \infty$ we get the required conclusion. □

3.4.3 PME with growing initial data.

Finally, we examine now the case when $u_0 \in \mathrm{L}^1_{\mathrm{loc}}(M)$. In [5] and [8], the authors provided necessary and sufficient conditions in $\mathbb{R}^d$ on the growth at infinity of the initial data u_0 for the existence and uniqueness of nonnegative solutions of the Cauchy problem (3.4.2) for $m > 1$. The issue of finding the optimal class of existence and uniqueness for nonnegative solutions in the Riemannian setting is still an open problem, see [110, Section 11]. In this Subsection, even if we will still not provide a full treatment of the problem, which would require much more attention and time on its own, we will begin to give some preliminary results adapting arguments used in [8]. In particular, we will extend the validity of inequalities [8, (1.7), (1.8), (1.9) and (1.10)]. Beside the usual assumption (3.4.1) on the curvature we will request M to satisfy here the Sobolev inequality

$$\begin{aligned} &\|f\|_{2^*} \le C\|\nabla f\|_2, \qquad 2^* = \frac{2d}{d-2} \text{ for } d \ge 3, \\ &\|f\|_2 \le C\|\nabla f\|_1, \qquad\qquad\quad \text{for } d = 2, \end{aligned} \tag{3.4.21}$$

for every $f \in \mathrm{H}^1(M)$. Before proceeding we need several technical definitions.

Let $\mathfrak{r}$ and $D_{1,\alpha}$, $D_{2,\alpha}$ be the exhaustion function and the constants which appear in Theorem 3.2.1, respectively, and let $\{\phi_R\}_{R\ge 1}$ be a family of Laplacian cut-off like in Corollary 3.2.3. Let us define

$$\gamma = \Gamma_\alpha + 1, \qquad \Gamma_\alpha = \begin{cases} \frac{D_{2,\alpha}}{D_{1,\alpha}} & \text{for } \alpha \in [-2,2), \\ 1 & \text{for } \alpha = 2, \end{cases}$$

and let us specify $\psi(s) \in C_c^\infty(\mathbb{R})$, $0 \le \psi \le 1$, such that

$$\psi(s) = \begin{cases} 1 & \text{for } s \in (-\infty, \Gamma_\alpha], \\ e^{-\frac{1}{\Gamma_\alpha + 1 - s}} & \text{for } s \in \left(\frac{2\Gamma_\alpha+1}{2}, \Gamma_\alpha + 1\right), \\ 0 & \text{for } s \ge \Gamma_\alpha + 1. \end{cases} \tag{3.4.22}$$

Then, $\phi_R(x) = \psi\left(\frac{h(x)}{D_{1,\alpha}R^{1-\alpha/2}}\right)$ as in 3.2.3.

Definition 3.4.11 (The Banach spaces X and $\mathrm{L}(\rho_\beta)$)**.** *For every $r \ge 1$, set the norms*

$$|f|_r := \sup_{R\ge r} R^{-\left(d+\frac{1+\alpha/2}{m-1}\right)} \int_M \phi_R(x)|f(x)|\,dx; \tag{3.4.23}$$

$$|||f|||_r := \sup_{R\ge r} R^{-\left(d+\frac{1+\alpha/2}{m-1}\right)} \int_{B_R} |f(x)|\,dx. \tag{3.4.24}$$

Observe that $|||f|||_r \le |f|_r \le \Gamma_\alpha^{d+\frac{1+\alpha/2}{m-1}} |||f|||_r$*, that is they are equivalent norms. We define the space X as*

$$X := \left\{ f \in \mathrm{L}^1_{\mathrm{loc}}(M) \,:\, |||f|||_1 < \infty \right\},$$

equipped with the norm $|||\cdot|||_1$.

Fix

$$\rho_{\alpha,\beta}(x) = \frac{1}{(1+\mathfrak{r}^{1+\alpha/2}(x))^\beta},$$

where α *is like in* (3.4.1) *and* $\beta \in \mathbb{R}$*. Then we define the weighted space* $\mathrm{L}^1(\rho_{\alpha,\beta})$ *as*

$$\mathrm{L}^1(\rho_{\alpha,\beta}) := \left\{ f \in \mathrm{L}^1_{\mathrm{loc}}(M) \,:\, \int_M |f|\rho_{\alpha,\beta} < \infty \right\},$$

equipped with the norm $\|f\|_{\mathrm{L}^1(\rho_{\alpha,\beta})} = \int_M |f|\rho_{\alpha,\beta}$.

Definition 3.4.12 (Class of solutions $\mathscr{S}$)**.** *Let* $u_0, v_0 \in \mathrm{L}^1(M) \cap \mathrm{L}^\infty(M)$*. Suppose that the solution map* $S(u_0,t) \mapsto u(t)$*, which associate at every initial datum* u_0 *an unique strong solution* $u(t)$ *for the PME Cauchy problem* (3.4.2)*, is well defined and that satisfies:*

- $S(u_0,\cdot) \in C([0,\infty) : \mathrm{L}^1(M))$*;*
- $\|S(u_0,t) - S(v_0,t)\|_1 \le \|u_0 - v_0\|_1$*;*
- *if* $u_0 \le v_0$ *then* $S(u_0,t) \le S(v_0,t)$*;*
- $-\|\max\{-u_0;0\}\|_\infty \le S(u_0,t) \le \|\max\{u_0;0\}\|_\infty$*;*
- $S(-u_0,t) = -S(u_0,t)$*;*
- $\Delta S(u_0,t)^{m-1} \ge -\frac{C}{t}$ *in* $D'(M)$*.*

We call this class of solutions $\mathscr{S}$*.*

We have the corresponding version of [8, Proposition 1.3] for Riemannian manifolds.

Proposition 3.4.13. *Let* $f \in \mathrm{L}^\infty(M)$ *be nonnegative,* $\Lambda \in (0,\infty)$ *and*

$$\Delta f^{m-1} \ge -\Lambda \qquad \text{in } D'(M). \tag{3.4.25}$$

Then, there exists a constant C depending only on d and $m > 1$ *such that for* $1 \le r \le R$

$$\frac{1}{R^{1+\alpha/2}} \|f\|^{m-1}_{\mathrm{L}^\infty(B_R(p))} \le C\left(\Lambda^{\lambda(m-1)} |f|_r^{2\lambda(m-1)/d} + |f|_r^{m-1} \right), \tag{3.4.26}$$

where

$$\lambda = d/((m-1)d+2). \tag{3.4.27}$$

Proof. Let f satisfies the above hypothesis and let us assume moreover f to be smooth and strictly positive, so that, in particular, f^{m-1} is smooth. To drop the smoothness and strictly positivity assumptions, let us remark that f can be approximated by a smooth positive sequence $\{f_i\}_{i\in\mathbb{N}}$ such that $\Delta f_i^{m-1} \geq -\Lambda$ and $f_i \to f$ in $\mathrm{L}^\infty_{\mathrm{loc}}(M)$. Indeed, noting that every Riemannian manifold has locally bounded geometry, this can be done, for example, by means of a smooth partition of unity and localized standard mollification techniques.

Let ϕ_R be the cut-off that appears in equation (3.4.23), then

$$\Delta(\phi_R f)^{m-1} = \phi_R^{m-1}\Delta f^{m-1} + 2\langle \nabla\phi_R^{m-1}, \nabla f^{m-1}\rangle + f^{m-1}\Delta\phi_R^{m-1},$$

and by (3.4.25)

$$\Delta(\phi_R f)^{m-1} \geq -\Lambda\phi_R^{m-1} + 2\langle \nabla\phi_R^{m-1}, \nabla f^{m-1}\rangle + f^{m-1}\Delta\phi_R^{m-1}. \tag{3.4.28}$$

Observe that $\phi_R^\theta \in C_c^\infty(M)$ for every $\theta > 0$ and that, by Corollary 3.2.3 and (3.4.22), for all $a > 1$

$$\|\phi_R^{a-2}|\Delta\phi_R| + \phi_R^{a-3}|\nabla\phi_R|^2\|_\infty \leq \frac{C(a,\alpha)}{R^{1+\alpha/2}}, \tag{3.4.29}$$

in particular $\Delta\phi_R^{m-1}$ is bounded. Multiplying (3.4.28) by $(\phi_R f)^p$ where $p > 1$ and integrating we obtain

$$\begin{aligned}\int_M \langle \nabla(\phi_R f)^p, \nabla(\phi_R f)^{m-1}\rangle\, dx \leq \Lambda\int_M (\phi_R)^{m-1+p} f^p\, dx - 2\int_M (\phi_R f)^p \langle \nabla\phi_R^{m-1}, \nabla f^{m-1}\rangle\, dx \\ - \int_M (\phi_R f)^p f^{m-1}\Delta\phi_R^{m-1}\, dx.\end{aligned} \tag{3.4.30}$$

We have that

$$\int_M \langle \nabla(\phi_R f)^p, \nabla(\phi_R f)^{m-1}\rangle\, dx = \frac{4(m-1)p}{(m-1+p)^2}\int_M |\nabla(\phi_R f)^{\frac{p+m-1}{2}}|^2\, dx, \tag{3.4.31}$$

$$\begin{aligned}\int_M (\phi_R f)^p \langle \nabla\phi_R^{m-1}, \nabla f^{m-1}\rangle\, dx &= \frac{(m-1)^2}{(m-1+p)^2}\int_M \langle \nabla\phi_R^{m-1+p}, \nabla f^{m-1+p}\rangle\, dx \\ &= -\frac{(m-1)^2}{(m-1+p)^2}\int_M f^{m-1+p}\Delta\phi_R^{m-1+p}\, dx \\ &= -\frac{(m-1)^2}{(m-1+a)^2}\int_M f^{m-1}(f\phi_R)^p \\ &\quad \cdot \left[(m-2+p)\phi_R^{m-3}|\nabla\phi_R|^2 + \phi_R^{m-2}\Delta\phi_R\right] dx.\end{aligned} \tag{3.4.32}$$

Using (3.4.31), (3.4.32) and (3.4.29) in (3.4.30) we get

$$\int_M \left|\nabla(\phi_R f)^{\frac{p+m-1}{2}}\right|^2 dx \leq C\left[\Lambda\int_M (\phi_R f)^p\, dx + \frac{1}{R^{1+\alpha/2}}\int_M (\phi_R f)^p f^{m-1}\, dx\right]. \tag{3.4.33}$$

Fix now $r \geq 1$ and define

$$A = \sup_{R \geq r} \frac{\|f\|^{m-1}_{\mathrm{L}^\infty(B_R(p))}}{R^{1+\alpha/2}}.$$

Then, by the Sobolev inequality (3.4.21) with $d \geq 3$ and by (3.4.33), we have that

$$\left[\int_M (\phi_R f)^{sp+b}\right]^{1/s} dx \leq C(\Lambda + A) \int_M (\phi_R f)^p \, dx, \tag{3.4.34}$$

where

$$s = \frac{d}{d-2}, \qquad b = s(m-1) = \frac{(m-1)d}{d-2},$$

which is inequality [8, (1.46)] where we used d in place N for consistency of our notation. From this point onward, the arguments will be exactly the same as in the original proof [8, Proposition 1.3] with the only difference of defining $\theta_0 = \frac{1+\alpha/2}{m-1} + d$ in equation [8, (1.49)]. Same remark for the case $d = 2$. □

We want to point out that the left-hand side of inequality (3.4.26) depends on the decay rate estimates of the Laplacian of the cut-offs while the λ constant that appears in right-hand side is correlated to the Sobolev constant. Manifolds which satisfy different kind of functional inequalities may give different estimates in (3.4.26).

Lemma 3.4.14. *Let $0 \leq u_0 \in \mathrm{L}^1(M) \cap \mathrm{L}^\infty(M)$ and $u \in \mathscr{S}$. There are constants C_1, C_2 and $C_3 > 0$ depending only on d and $m > 1$ such that if $r \geq 1$ and $0 \leq t \leq T_r(u_0) = C_1/|u_0|_r^{m-1}$ then*

(i) $|u(t)|_r \leq C_2 |u_0|_r$;

(ii) $\frac{\|u(t)\|^{m-1}_{\mathrm{L}^\infty(B_R)}}{R^{1+\alpha/2}} \leq \frac{C_3}{t^{\lambda(m-1)}} |u_0|_r^{2\lambda(m-1)/d}$ *for $R \geq r$,*

with λ as in (3.4.27)*.*

Proof. The proof is a direct adaptation of [8, Lemma 1.4] using Proposition 3.4.13 combined again with Corollary 3.2.3. We will skip the details this time. □

Lemma 3.4.15. *In the same assumptions and notations of Lemma 3.4.14, let $u, v \in \mathscr{S}$ with initial data u_0 and v_0, respectively. Let $r \geq 1$, α as in* (3.4.1)*, $\beta \in \mathbb{R}$ and $t \in [0, \min\{T_r(u_0); T_r(v_0)\}]$. Then*

$$|||u(t) - v(t)|||_r \leq e^{C_1 t^{2\lambda/d}} |||u_0 - v_0|||_r, \tag{3.4.35}$$

and

$$\|u(t) - v(t)\|_{\mathrm{L}^1(\rho_{\alpha,\beta})} \leq e^{C_2 t^{2\lambda/d}} \|u_0 - v_0\|_{\mathrm{L}^1(\rho_{\alpha,\beta})}, \tag{3.4.36}$$

where C_1 depends only on $\max\{|||u_0|||_r; |||v_0|||_r\}$, C_2 depends only on $\max\{|||u_0|||_r; |||v_0|||_r; \beta; r\}$ and λ is like in (3.4.27)*.*

Proof. Defining the usual family of Laplacian cut-offs $\{\phi_R\}_{R\geq 1}$ and using the arguments in the proof of Proposition 3.4.2, we are lead to

$$\begin{aligned}\frac{d}{dt}\int_M \phi_R|u(t)-v(t)| &\leq \int_M \Delta\phi_R\,|u^m(t)-v^m(t)|\\ &\leq \frac{C}{R^{1+\alpha/2}}\int_M \max\{m|u(t)|^{m-1};m|v(t)|^{m-1}\}|u-v|\\ &\leq C\max\{R^{-1-\alpha/2}\|u\|^{m-1}_{\mathrm{L}^\infty(B_{\Gamma_\alpha R})};R^{-1-\alpha/2}\|v\|^{m-1}_{\mathrm{L}^\infty(B_{\Gamma_\alpha R})}\}\int_{B_{\Gamma_\alpha R}}|u-v|\\ &\leq C\max\{R^{-1-\alpha/2}\|u\|^{m-1}_{\mathrm{L}^\infty(B_{\Gamma_\alpha R})};R^{-1-\alpha/2}\|v\|^{m-1}_{\mathrm{L}^\infty(B_{\Gamma_\alpha R})}\}\int_M \phi_{\Gamma_\alpha R}|u-v|.\end{aligned} \tag{3.4.37}$$

By multiplying both members of the above inequality by $R^{-[(1+\alpha/2)/(m-1)+d]}$, using (ii) of Lemma 3.4.14 and integrating in time, we get

$$|u(t)-v(t)|_r \leq |u_0-v_0|_r + C\max\{|u_0|_r;|v_0|_r\}^{(2\lambda/d)(m-1)}\int_0^t \frac{|u(\tau)-v(\tau)|_r}{\tau^{\lambda(m-1)}}\,d\tau,$$

for $0\leq t\leq C_1\min\{|u_0|_r^{1-m};|v_0|_r^{1-m}\}$. We conclude now by comparing $t\mapsto |u(t)-v(t)|_r$ with

$$h(t)=|u_0-v_0|_r e^{(D/(1-\lambda(m-1)))t^{1-\lambda(m-1)}},$$

which is solution of

$$\begin{cases} h'(t)=Dh(t)t^{-\lambda(m-1)},\\ h(0)=|u_0-v_0|_r,\end{cases}$$

where $D=C\max\{|u_0|_r;|v_0|_r\}^{(2\lambda/d)(m-1)}$ and $1-\lambda(m-1)=2\lambda/d$.

To prove inequality (3.4.36) instead, let us introduce the following weight-function

$$\tilde\rho_{\alpha,\beta}(x):=\begin{cases}1 & \text{if } x\in B_1(p),\\ \dfrac{1}{\left[1+\left(\frac{\mathfrak{r}(x)}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}\right]^\beta} & \text{if } x\in M\setminus\overline{B}_1(p),\end{cases}$$

for $\alpha\in[-2,2)$ and with $\beta\geq 0$. From (1) of Theorem 3.2.1, it is not difficult to check that

$$\rho_{\alpha,1}\leq\tilde\rho_{\alpha,1}\leq\max\left\{2;(D_{2,\alpha}/D_{1,\alpha})^{\frac{1+\alpha/2}{1-\alpha/2}}\right\}\rho_{\alpha,1}\qquad\text{on all over } M,$$

that is the norm $\|\cdot\|_{\mathrm{L}^1(\rho_{\alpha,1})}$ is equivalent to the norm $\|\cdot\|_{\mathrm{L}^1(\tilde{\rho}_{\alpha,1})}$ induced by the weight-function $\tilde{\rho}_{\alpha,1}$. Noticing now that the product function $\phi_R\tilde{\rho}_{\alpha,1} \in C_c^\infty(M)$, we can then argue like above to get

$$\frac{d}{dt}\int_M \phi_R\tilde{\rho}_{\alpha,1}|u(t)-v(t)| \leq \int_M \Delta(\phi_R\tilde{\rho}_{\alpha,1})\,|u^m(t)-v^m(t)|\,.$$

Using the estimates in Theorem 3.2.1, the definition of $\tilde{\rho}_{\alpha,1}$ and that $r \geq 1$, $\alpha \in [-2,2)$ and $\mathrm{supp}(|\nabla\phi_R|+|\Delta\phi_R|) \subseteq B_{\Gamma_\alpha R}\setminus\overline{B}_R$, we have

$$\begin{aligned}
\left|\Delta(\phi_R\tilde{\rho}_{\alpha,1})\right| &\leq \tilde{\rho}_{\alpha,1}|\Delta\phi_R| + 2|\langle\nabla\tilde{\rho}_{\alpha,1},\nabla\phi_R\rangle| + \phi_R|\Delta\tilde{\rho}_{\alpha,1}| \\
&\leq \frac{C_1\tilde{\rho}_{\alpha,1}}{R^{1+\alpha/2}} + 2|\nabla\tilde{\rho}_{\alpha,1}||\nabla\phi_R| \\
&\quad + \frac{C_2\tilde{\rho}_{\alpha,1}}{1+\left(\frac{\mathfrak{r}}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}}\left\{\frac{A_1\mathfrak{r}^{\frac{2\alpha}{1-\alpha/2}}}{1+\left(\frac{\mathfrak{r}}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}} + A_2\left(\mathfrak{r}^{-\frac{1-3\alpha/2}{1-\alpha/2}} + \mathfrak{r}^{\frac{\alpha}{1-\alpha/2}}\right)\right\} r^{-\alpha} \\
&\leq C_1\frac{\tilde{\rho}_{\alpha,1}}{R^{1+\alpha/2}} + C_3\frac{\tilde{\rho}_{\alpha,1}}{R^{1+\alpha/2}} + C_2\frac{\tilde{\rho}_{\alpha,1}}{1+\left(\frac{\mathfrak{r}}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}}.
\end{aligned}$$

For $R \leq r(x) \leq \Gamma_\alpha R$, we have

$$\frac{|u(t,x)|^{m-1}}{1+\left(\frac{\mathfrak{r}(x)}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}} \leq \frac{(\Gamma_\alpha R)^{1+\alpha/2}}{1+(\Gamma_\alpha)^{-\frac{1+\alpha/2}{1-\alpha/2}}R^{1+\alpha/2}}\frac{\|u(t,x)\|^{m-1}_{\mathrm{L}^\infty(B_{\Gamma_\alpha R})}}{(\Gamma_\alpha R)^{1+\alpha/2}},$$

and for $r(x) \leq r$

$$\frac{|u(t,x)|^{m-1}}{1+\left(\frac{\mathfrak{r}(x)}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}} \leq r^{1+\alpha/2}\frac{\|u(t,x)\|^{m-1}_{\mathrm{L}^\infty(B_r)}}{r^{1+\alpha/2}},$$

and it follows then that

$$\sup_{x\in M}\frac{|u(t,x)|^{m-1}}{1+\left(\frac{\mathfrak{r}(x)}{D_{2,\alpha}}\right)^{\frac{1+\alpha/2}{1-\alpha/2}}} \leq C(r)\sup_{R\geq r}\frac{\|u(t,x)\|^{m-1}_{\mathrm{L}^\infty(B_R)}}{R^{1+\alpha/2}}.$$

We can then proceed again as from (3.4.37) and conclude thanks to the equivalence of the norms $\|\cdot\|_{\mathrm{L}^1(\rho_{\alpha,1})} \sim \|\cdot\|_{\mathrm{L}^1(\tilde{\rho}_{\alpha,1})}$. The case $\beta < 0$ and the case $\alpha = 2$ are done in the same way with suitable changes. □

3.5 Conclusions, open problems and further comments

We believe that some of the techniques we introduced in this chapter, in particular the sequence of Laplacian cut-offs, will be useful for further applications, such as the extensions of global

properties of different type of PDE's to Riemannian manifolds. Nevertheless, there are still many open problems. For example, it is not clear whether it is possible to obtain a control even on the Hessian of the cut-offs without imposing a strictly positive injectivity radius.
Moreover, in the spirit of the previous section, it would be interesting to extend the theory of PME and FDE with growing initial data, i.e., $u_0 \in \mathrm{L}^1_{\mathrm{loc}}(M)$. In the Riemannian setting there is not even an existence result for the FDE, in that sense.
Finally, regarding concrete applications, thanks to the preliminary results obtained in Theorem 3.4.6 and Remark 3.4.8, we are now in the position to begin a reasonable study concerning the extinction properties of plasma inside toroidal reactors, see [82].

Inverse Problems Regularization

Introduction to inverse problems regularization: an example

When we speak about an *inverse problem* a natural question arises: inverse with respect to what? If we are dealing with a real world application, for example, we could be interested in finding the initial state of the system after we observed its actual evolution, solving a backward problem. Contextualizing this slightly vague example to a concrete one, let us consider an application to image deblurring.

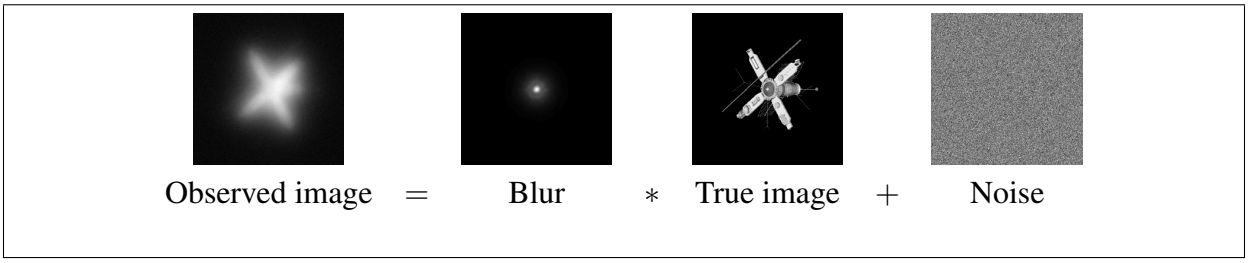


Figure 4.1: Blurring process

Generally speaking, whenever we take a picture of an object the resulting image can be perturbed due to some (atmospheric) blur and background noise, like measurement errors or data approximations, over which we do not have any control. In this case we wish to recover the *true image* from our observation. From a physical point of view, the blurring process can be modeled by a convolution operator. Indeed, if we think of an image as a function f that represent the light and assign at every point of the space (*pixel*) a number (actually, a triple of numbers) which encodes the color in the RGB standard, then the blur process can be represented in the following way,

$$g(s,t) = \iint_{\Omega} h(s-s',t-t')f(s',t')\,ds'dt' + \eta(s,t), \qquad (s,t) \in \Omega \subset \mathbb{R}^2,$$

where $f : \Omega \to \mathbb{R}$ is the true image, $g : \Omega \to \mathbb{R}$ is the observed image and $h : \mathbb{R}^2 \to \mathbb{R}$ and $\eta : \mathbb{R}^2 \to \mathbb{R}$ are called *point spread function* (PSF) and *noise*, respectively. Therefore,

$$g(s,t) = (h * f)(s,t) + \eta(s,t), \qquad (4.0.1)$$

and if we assume that the convolution kernel h is in $L^2(\mathbb{R}^2)$, then the blur operator is a compact integral operator of the first kind, see Proposition 5.1.14. Solving the above problem 4.0.1 is challenging, because even if theoretically a solution f exists, practically it would be impossible to recover it exactly due to the noise affecting the observed data. Indeed, we say that the problem is *ill-posed*. A proper definition of ill-posedness will be given later, see Definition 4.0.2, but to make it clear, small changes on the observed image g greatly affect the recovered solution, making it distant, in some sense, from the true solution f. Since the noise η can not be canceled out of the equation, solving strategies which do not take into account with sufficient care the ill-posedness of the problem would lead to poor approximated solutions.

In order to have a better understanding of what happens when we discretize problem 4.0.1, we consider the one dimensional case and we suppose that $\eta = 0$. Namely,

$$g(s) = \int_a^b h(s-s')f(s')\,ds', \qquad s \in [a,b] \subset \mathbb{R}.$$

If we define $s_i = a + i\xi$, with $\xi = \frac{b-a}{n}$, $n \in \mathbb{N}$ fixed, $i = 0, \cdots, n$, then

$$g(s_j) \approx \xi \sum_{i=0}^{n-1} h(s_j - s_i) f(s_i) = \xi \sum_{i=0}^{n-1} h((j-i)\xi) f(s_i), \quad j = 0, \cdots, n-1.$$

Setting

$$g_j := g(s_j), \qquad h_{j-i} := \xi h((j-i)\xi), \qquad f_i := f(s_i),$$

to simplify notation, the discrete convolution may be rewritten as

$$g_j = \sum_{i=0}^{n-1} h_{j-i} f_i,$$

that is,

$$\mathbf{g} = H\mathbf{f} \qquad \mathbf{g}, \mathbf{f} \in \mathbb{R}^n, \tag{4.0.2}$$

with

$$H = \begin{bmatrix} h_0 & h_{-1} & \cdots & \cdots & h_{-(n-1)} \\ h_1 & h_0 & h_{-1} & \cdots & h_{-(n-2)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & h_{-1} \\ h_{n-1} & \cdots & \cdots & h_1 & h_0 \end{bmatrix}.$$

The matrix M is determined by its first row and column, i.e., by $2n-1$ parameters. Moreover, it is constant on all of its diagonal entries (*Toeplitz* matrix, see Definition 6.1.1). This brings us to define the *stencil* vector $\mathbf{v}_{\text{PSF}}$,

$$\mathbf{v}_{\text{PSF}} = \begin{bmatrix} h_{n-1} & \cdots & h_1 & h_0 & h_{-1} & \cdots & h_{-(n-1)} \end{bmatrix},$$

and

$${}^j\tilde{\mathbf{f}} = \begin{bmatrix} \tilde{f}_{j-(n-1)} & \cdots & \tilde{f}_{j-1} & \tilde{f}_j & \tilde{f}_{j+1} & \cdots & \tilde{f}_{j+n-1} \end{bmatrix}, \quad \text{for } j = 0, \cdots, n-1,$$

where

$$\tilde{f}_i = \begin{cases} f_i & \text{if } i = 0, 1, \cdots, n-1, \\ 0 & \text{otherwise,} \end{cases}$$

from which we can write

$$\begin{aligned} g_j &= \langle \mathbf{v}_{\text{PSF}}, {}^j\tilde{\mathbf{f}} \rangle \\ &= h_{n-1} \cdot \tilde{f}_{j-(n-1)} + \cdots + h_1 \cdot \tilde{f}_{j-1} + h_0 \cdot \tilde{f}_j \\ &\quad + h_{-1} \cdot \tilde{f}_{j+1} + \cdots + h_{-(n-1)} \cdot \tilde{f}_{j+n-1}. \end{aligned}$$

Changing j into $j+1$ (or $j-1$) would produce a shift to the right (or left) of the elements of the stencil $\mathbf{v}_{\text{PSF}}$, since h_0 always acts on $\tilde{f}_j$. To understand the role of $\mathbf{h}$, let us fix n even and

$\mathbf{f} = [0, \cdots, 1, \cdots, 0]$, i.e., $\mathbf{f}$ the unitary vector $\mathbf{e}_{n/2}$ with entry 1 in position $n/2$ and 0 elsewhere. A natural hypothesis is to assume that the entries of $\mathbf{h}$ sum to 1,

$$\sum_{i=-(n-1)}^{n-1} h_i = 1,$$

since the total amount of light which is blurred does not change. A representation of $\mathbf{g}$ under the action of a Gaussian PSF $\mathbf{h}$ can be seen in Figure 4.2. We observe that the light, concentrated in position $n/2$, is redistributed on the entire interval $[a,b]$. In this case we implicitly imposed $f \equiv 0$ outside $[a,b]$ and one of the consequences of this choice is the Toeplitz structure of H. Other choices for the *boundary conditions* would lead to different matrix structures for H as it will be seen in Chapter 6.

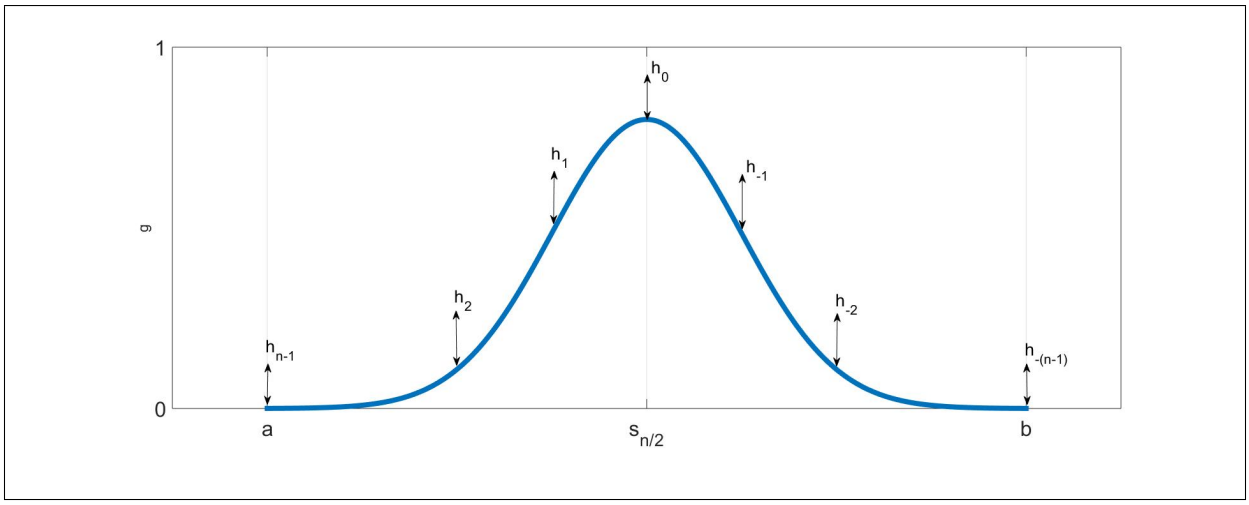


Figure 4.2: Gaussian PSF.

So far we still have not clarified why such a discretized convolution problem (4.0.2) can be hard to solve. We give the following definition.

Definition 4.0.1 (Condition number). *Let $A : \mathbb{C}^n \to \mathbb{C}^n$ be an invertible linear operator. Define the* conditioning number $\kappa(A)$ *of the operator A as*

$$\kappa(A) := \left(\sup_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \right) \cdot \left(\sup_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\|A^{-1}\boldsymbol{u}\|}{\|\boldsymbol{u}\|} \right).$$

Loosely speaking, the condition number measures the continuity of the inverse of an operator with respect to the continuity of the operator itself. The bigger the condition number and the worse will be the reconstructed numerical solution if we directly invert the operator. Indeed, as it has already been said, we do not solve problem (4.0.2) but a modified version of it where the noise comes into play,

$$\mathbf{g}^{\delta} = H\mathbf{f} + \eta, \qquad \|\eta\| = \delta. \tag{4.0.3}$$

If we assume H to be invertible and if we put

$$\mathbf{f}^{\delta} := H^{-1}\mathbf{g}^{\delta},$$

then even for small values of δ, the ratio $\|\mathbf{f}^{\delta} - \mathbf{f}\|/\|\mathbf{f}\|$ is great. Indeed, the matrix H is ill-conditioned, i.e. $\kappa(H)$ is big, and this is an intrinsic property of compact operators, see Chapter 5.1. The Condition number and the consequent sensitiveness of solutions to perturbed data are related to the concept of *discrete ill-posedness* for a problem, see [63]. We present here the (pseudo) definition, due to Hadamard, of *ill-posedness* for a problem in a continuous setting.

Definition 4.0.2 (Ill-posed problem). *We say that a problem is* well-posed *if the following properties hold:*

(i) a solution exists for all admissible data;

(ii) the solution is unique;

(iii) the solution depends continuously on the data.

We say that a problem is ill-posed *if it is not well-posed.*

Of course, the definition lacks of precision. In order to make it mathematically precise several elements should be fixed, such as a definition of what it is considered for solution, what data are admissible and the topology of the space. Nevertheless, this generality makes it flexible in different context.

According to the above definition, Problem 4.0.3 is ill-posed and therefore we need to *regularize* it to compute an approximated and numerically stable solution which could be sufficiently accurate, where regularizing means to substitute H^{-1} with a suitable family of continuous operators $\{R_\alpha\}$ depending on one (or several) parameter α. The choice of the best regularizing parameter α is of utmost importance, and it reflects in some sense the trade-off between accuracy and stability.

Obviously, the image deblurring problem we presented here is just one of the many examples of inverse problems which could be proposed, but it served us with the purpose to briefly motivating Chapter 6, where new strategies of preconditioning as regularizer are introduced to solve inverse problems in the imaging context. Chapter 5 is devoted instead to regularization techniques of filter type which have been recently studied. In this latter case we will deal with inverse problems in the more general setting of Hilbert spaces and for this reason we will use a slightly different notation with respect to the notation used in this chapter and which we will recall in Chapter 6, where we mostly deal with a discretized model. Therefore, we are confident that there will be no risk of misunderstandings.

Fractional and Weighted Tikhonov

5.1 Preliminary definitions: best approximate solution and compact operators theory.

We introduce here the basic theory of compact operators and generalized inverse of operators which we believe it will be useful here and in the next chapter. For a full treatment of the subjects we refer the reader to [48], [94] and [34].

Let $T : X \to Y$ be a continuous linear operator between Hilbert spaces X and Y (over the field $\mathbb{R}$ or $\mathbb{C}$) with inner products $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Y$, and induced norms $\|x\|_X = \sqrt{\langle x, x \rangle_X}$ and $\|y\|_Y = \sqrt{\langle x, x \rangle_Y}$, respectively. We recall that a linear operator T between Hilbert spaces is continuous if and only if is bounded and that the operator norm $\|T\|$ is defined as

$$\|T\| := \sup_{x \in X, x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X}.$$

Hereafter we will omit the subscript for the inner product and the norm as it will be clear from the context. For brevity, we will write $T \in \mathscr{B}(X,Y)$, where

$$\mathscr{B}(X,Y) := \{A : X \to Y : \ A \text{ is a bounded linear operator}\}.$$

For $X = Y$, $\mathscr{B}(X,Y) \equiv \mathscr{B}(X)$.
What we said and most of what we will present in this section can be hold in to the more general setting of Banach spaces, however in this Chapter we will restrict ourselves to the Hilbert case. We have the following definitions.

Definition 5.1.1 (Least square solution). *$x \in X$ is called* least square solution *of the equation $Tx = y$ if*

$$\|Tx - y\| = \inf\{\|Tz - y\| : z \in X\}.$$

Definition 5.1.2 (Best approximate solution). *$x \in X$ is called* best approximate solution *of the equation $Tx = y$ if x is a least square solution of $Tx = y$ and*

$$\|x\| = \inf\{\|z\| : z \text{ is a least square solution of } Tx = y\}$$

holds.

Definition 5.1.3 (Moore-Penrose generalized inverse).
Let $\tilde{T}_{|\mathrm{Ker}(T)^\perp} : \mathrm{Ker}(T)^\perp \to \mathrm{Rg}(T)$. We define the Moore-Penrose generalized inverse *$T^\dagger$ of T the unique linear extension of $\tilde{T}^{-1}$ to*

$$\mathrm{Dom}(T^\dagger) = \mathrm{Rg}(T) \dotplus \mathrm{Rg}(T)^\perp$$

with

$$\mathrm{Ker}(T^\dagger) = \mathrm{Rg}(T)^\perp.$$

We just want to observe that the above definition is well-defined. Indeed, since $\mathrm{Ker}(\tilde{T}) = \{0\}$ and $\mathrm{Rg}(\tilde{T}) = \mathrm{Rg}(T)$, then $\tilde{T}^{-1}$ exists.

Proposition 5.1.4. *Let P and Q be the orthogonal projectors onto* $\mathrm{Ker}(T)$ *and* $\overline{\mathrm{Rg}(T)}$*, respectively. Then* $\mathrm{Rg}(T^\dagger) = \mathrm{Ker}(T)^\perp$*, and the following equalities hold*

$$TT^\dagger T = T, \tag{5.1.1a}$$
$$T^\dagger T T^\dagger = T^\dagger, \tag{5.1.1b}$$
$$T^\dagger T = I - P, \tag{5.1.1c}$$
$$TT^\dagger = Q_{|\mathrm{Dom}(T^\dagger)}. \tag{5.1.1d}$$

Proof. See [48, Proposition 2.3]. □

Proposition 5.1.5. *The Moore-Penrose generalize inverse $T^\dagger$ has a closed graph* $\mathrm{Graph}(T^\dagger)$*. Furthermore $T^\dagger$ is bounded if and only if* $\mathrm{Rg}(T)$ *is closed.*

Proof. See [48, Proposition 2.4]. □

Theorem 5.1.6. *For every $T \in \mathscr{B}(X,Y)$, with X and Y Hilbert spaces, there exists a unique bounded linear operator $T^* \in \mathscr{B}(Y,X)$ such that*

$$\langle Tx, y\rangle = \langle x, T^* y\rangle$$

for every $x \in X$ and $y \in Y$. Moreover, T^* satisfies

$$\|T^*\| = \|T\|$$

and it is called adjoint *of T.*

Proof. See [94, Theorem 4.10]. □

Obviously, in the finite dimensional case T^* corresponds to the Hermitian transpose of the matrix operator T.

Definition 5.1.7 (Self-adjoint operator). *A bounded linear operator $T : X \to X$ is called* self-adjoint *if $T = T^*$.*

Definition 5.1.8 (Compact operator). *Let $K : X \to Y$ be a linear operator and let B_1 be the open unit ball in X, namely, $B_1 := \{x \in X : \|x\| < 1\}$. K is said to be* compact *if it satisfies one of the following equivalent properties:*

(i) $\overline{K(B_1)}$ *is compact in Y;*

(ii) every bounded sequence $\{x_n\}_{n\in\mathbb{N}}$ in X contains a subsequence $\{x_{n_k}\} \subseteq \{x_n\}$ such that $\{Tx_{n_k}\}$ is convergent in Y.

The set of all compact linear operators $K : X \to Y$ is denoted with $\mathscr{B}_0(X,Y)$. For $X = Y$, $\mathscr{B}_0(X,Y) \equiv \mathscr{B}_0(X)$.

A compact operator is obviously bounded and therefore continuous, namely, $\mathscr{B}_0(X,Y) \subset \mathscr{B}(X,Y)$.

Proposition 5.1.9.

(i) Let $\{K_n\}_{n\in\mathbb{N}} \subset \mathscr{B}_0(X,Y)$ be a sequence of compact operators and let $T \in \mathscr{B}(X,Y)$ such that $\|K_n - T\| \to 0$ as $n \to \infty$. Then $T \in \mathscr{B}_0(X,Y)$.

(ii) If $T \in \mathscr{B}(X), A \in \mathscr{B}(Y)$, and $K \in \mathscr{B}_0(X,Y)$, then KT and $AK \in \mathscr{B}_0(X,Y)$.

Proof. See [34, Proposition 4.2]. □

Definition 5.1.10. *An operator T on X has* finite rank *if* $\mathrm{Rg}(T)$ *is finite dimensional.*

Theorem 5.1.11. *If $T \in \mathscr{B}(X,Y)$, the following statements are equivalent.*

(i) T is compact.

(ii) T^* is compact.

(iii) There is a sequence $\{T_n\}_{n\in\mathbb{N}}$ of operators of finite rank such that $\|T_n - T\| \to 0$ as $n \to \infty$.

Proof. See [34, Theorem 4.4]. □

Proposition 5.1.12. *Let $K \in \mathscr{B}_0(X,Y)$. Then* $\mathrm{Rg}(K)$ *is closed if and only if it has finite rank.*

Proof. See [94, Theorem 4.18]. □

Combining now Proposition 5.1.5 and 5.1.12, it follows

Proposition 5.1.13. *Let $K \in \mathscr{B}_0(X,Y)$. If* $\dim \mathrm{Rg}(K) = \infty$ *then $K^\dagger$ is a densely defined unbounded linear operator.*

It follows an example of compact operator which is related to Problem (4.0.1).

Proposition 5.1.14. *If (X,Ω,μ) is a measure space and $k \in \mathrm{L}^2(X \times X)$, then*

$$(Kf)(s) = \int k(s,s')f(s')\,d\mu(s')$$

is a compact operator and $\|K\| \leq \|k\|_{\mathrm{L}^2}$.

Proof. [34, Proposition 4.7]. □

Definition 5.1.15 (Spectrum). *Let $T \in \mathscr{B}(X)$. The* spectrum $\sigma(T)$ *of T is the set of all $\sigma \in \mathbb{C}$ such that the operator $\sigma I - T$ is not invertible on all of X, where I is the identity operator.*

It is trivial to check the set of all the eigenvalues of an operator T is a subset of $\sigma(T)$.

Theorem 5.1.16. *Let $T \in \mathscr{B}_0(X)$ and let T be self-adjoint. Then $\sigma(T) \setminus \{0\} \subset \mathbb{R}$ and it consists of the sequence of eigenvalues $\{\sigma_m\}_{m=1}^N$ of T, such that $N \in \mathbb{N}$ or $N = \infty$ and with finitely many $|\sigma_m| > r$ for any $r > 0$. If $N = \infty$ then $\lim_m \sigma_m = 0$ and $0 \in \sigma(T)$. Moreover, there exists a corresponding orthonormal sequence $\{v_m\}_{m=1}^N \subset X$ such that*

$$Tv_m = \sigma_m v_m, \qquad \text{for all } m = 1, \cdots, N; \tag{5.1.2}$$

$$\mathrm{Ker}(T) = \mathrm{Span}\left(\{v_m\}_{m=1}^N\right)^{\perp}; \tag{5.1.3}$$

$$Tx = \sum_{m=1}^{N} \sigma_m \langle x, v_m \rangle v_m \qquad \text{for every } x \in X. \tag{5.1.4}$$

Proof. See [67, Theorem 4.2.4]. □

By Proposition 5.1.9 and Theorem 5.1.11, K^*K and KK^* are compact self-adjoint operators for every $K \in \mathscr{B}_0(X,Y)$, and we have the following corollary.

Corollary 5.1.17. *Let $K \in \mathscr{B}_0(X,Y)$. Then it holds*

$$Kx = \sum_{m=1}^{+\infty} \sigma_m \langle x, v_m \rangle u_m, \qquad \text{for every } x \in X, \tag{5.1.5}$$

$$K^*y = \sum_{m=1}^{+\infty} \sigma_m \langle y, u_m \rangle v_m, \qquad \text{for every } y \in Y, \tag{5.1.6}$$

with

(i) $\{\sigma_m^2\}$ *the non-ascending and nonzero eigenvalues of K^*K and KK^*, and $\sigma_m = \sqrt{\sigma_m^2}$;*

(ii) $\{v_m\}$ *the orthonormal eigenvectors of K^*K satisfying $Kv_m = \sigma_m u_m$;*

(iii) $\{u_m\}$ *the orthonormal eigenvectors of KK^* satisfying $K^*u_m = \sigma_m v_m$,*

and where the series (5.1.5) *and* (5.1.6) *converge in the L^2 norms induced by the scalar products of X and Y, respectively.*

Proof. See [67, Theorem 4.3.1]. □

Definition 5.1.18 (Singular value expansion).
Let $K \in \mathscr{B}_0(X,Y)$ and let $\{\sigma_m\}, \{v_m\}, \{u_m\}$ be like in Corollary 5.1.17. The triple $(\sigma_m; v_m, u_m)_{m\in\mathbb{N}}$ is called singular value expansion *(s.v.e.) of K.*

Proposition 5.1.19. *Let $K \in \mathscr{B}_0(X,Y)$, let $(\sigma_m; v_m, u_m)$ be its singular value expansion and let $K^\dagger$ its generalized inverse, as in Definition 5.1.3. Then, for every $y \in Y$ it holds that*

(i) $y \in \mathrm{Dom}(K^\dagger)$ *if and only if* $\sum_{m=1}^{\infty} \frac{|\langle y, u_m \rangle|^2}{\sigma_m^2} < \infty$;

(ii) For $y \in \mathrm{Dom}(K^\dagger)$,

$$K^\dagger y = \sum_{m=1}^{\infty} \frac{\langle y, u_m \rangle v_m}{\sigma_m}.$$

Proof. See [48, Theorem 2.8]. □

To summarize, if $K : X \to Y$ is a compact linear operator between Hilbert spaces then we indicate with $(\sigma_m; v_m, u_m)_{m\in\mathbb{N}}$ the s.v.e. of K, where $\{v_m\}_{m\in\mathbb{N}}$ and $\{u_m\}_{m\in\mathbb{N}}$ are a complete orthonormal system of eigenvectors for K^*K and KK^*, respectively, and $\sigma_m > 0$ are written in decreasing order, with 0 being the only accumulating point for the sequence $\{\sigma_m\}_{m\in\mathbb{N}}$ when dimension $\mathrm{Rg}(K) = \infty$. If X is not finite dimensional, then $0 \in \sigma(K^*K)$, the spectrum of K^*K, namely $\sigma(K^*K) = \{0\} \cup \bigcup_{m=1}^{\infty}\{\sigma_m^2\}$. Finally, $\sigma(K)$ is the closure of $\bigcup_{m=1}^{\infty}\{\sigma_m\}$, i.e., $\sigma(K) = \{0\} \cup \bigcup_{m=1}^{\infty}\{\sigma_m\}$.
We propose now another example of compact operator which will be useful later in Section 5.5. For any reference and proof we invite the reader to look at [2, Section 7.5] and [92, Chapter 4]. Let $\Omega = [0, 2\pi]$ and let us define the (fractional) Sobolev space $\mathrm{H}^s(\Omega)$,

$$\mathrm{H}^s(\Omega) = \left\{ x \in \mathrm{L}^2(\Omega) \ : \ \sum_{m\in\mathbb{Z}} \left(1+m^2\right)^s |x_m|^2 < \infty \right\}, \qquad s \in (0,\infty),$$

where x_m are the Fourier coefficients of the function $x : [0, 2\pi] \to \mathbb{C}$, i.e.,

$$x_m = \langle x(t), \mathrm{e}^{\mathrm{i}mt} \rangle_{\mathrm{L}^2} = \frac{1}{2\pi} \int_0^{2\pi} x(t) \mathrm{e}^{-\mathrm{i}mt}\, dtm \qquad m \in \mathbb{Z}.$$

$\mathrm{H}^s(\Omega)$ is an Hilbert space provided with the following inner product

$$\langle x_1(t), x_2(t) \rangle_{\mathrm{H}^s} = \sum_{m\geq 0} \left(1+m^2\right)^{s/2} \langle x_1(t), x_2(t) \rangle_{\mathrm{L}^2}.$$

Proposition 5.1.20.
Let $J_s : \mathrm{H}^s(\Omega) \hookrightarrow \mathrm{L}^2(\Omega)$, $x \mapsto J_s(x) \in \mathrm{L}^2(\Omega)$, *be the embedding operator of* $\mathrm{H}^s(\Omega)$, *and let*

$$v_m(t) = \left(1+m^2\right)^{-s/2} \mathrm{e}^{\mathrm{i}mt}, \qquad u_m = \mathrm{e}^{\mathrm{i}mt}, \qquad \sigma_m = \left(1+m^2\right)^{-s/2}.$$

The operator J_s *is compact for every* $s \in (0,\infty)$, *i.e.,* $J_s \in \mathscr{B}_0\left(\mathrm{H}^s(\Omega), \mathrm{L}^2(\Omega)\right)$, *and* $\left(\sigma_m^2; v_m, u_m\right)_{m\in\mathbb{N}}$ *is its s.v.e.*

Using s.v.e. representation, one may define functions of the compact self-adjoint operator $K \in \mathscr{B}_0(X)$ as

$$f(K) = \sum_{m=1}^{+\infty} f(\sigma_m) \langle \cdot, v_m \rangle u_m.$$

For a general and rigorous treatment of functional calculus for (unbounded) self-adjoint operators we refer again to [94]. With this notation we have the following theorem.

Theorem 5.1.21 (Spectral measure). *Let $K \in \mathscr{B}_0(X)$ be self-adjoint and let $x_1, x_2 \in X$ be fixed elements. There exists an unique regular complex Borel measure μ on $\sigma(K)$, depending on K and x_1, x_2, such that*

$$\int_{\sigma(K)} f(\sigma) d\mu(\sigma) = \langle f(K)x_1, x_2 \rangle$$

for all bounded Borel measurable functions f on $\sigma(K)$. In particular, if $x_1 = x_2 = x$, and if $f = 1$ so that $f(K) = Id$, we have

$$\mu(\sigma(K)) = \|x\|^2.$$

This measure is called spectral measure *associated to x_1, x_2 and K.*

Definition 5.1.22 (Spectral decomposition).
Let $K \in \mathscr{B}_0(X,Y)$ and let $(\sigma_m; v_m, u_m)_{m\in\mathbb{N}}$ be the s.v.e. of K. We call $\{E_{\sigma_m^2}\}_{\sigma_m^2 \in \sigma(K^*K)}$ the spectral decomposition *of the self-adjoint operator K^*K, where $E_{\sigma_m^2}$ is the spectral measure associated to v_m and K^*K.*

Let $\{E_{\sigma^2}\}_{\sigma^2\in\sigma(K^*K)}$ be the spectral decomposition of K^*K, where $K \in \mathscr{B}_0(X,Y)$. From what stated above, we can write $f(K^*K) := \int f(\sigma^2) dE_{\sigma^2}$, where $f : \sigma(K^*K) \subset \mathbb{R} \to \mathbb{C}$ is a bounded Borel measurable function and $\langle Ex_1, x_2 \rangle$ is a regular complex Borel measure for every $x_1, x_2 \in X$. The following equalities hold

$$f(K^*K)x := \int_{\sigma(K^*K)} f(\sigma^2) dE_{\sigma^2} x = \sum_{m=1}^{\infty} f(\sigma_m^2) \langle x, v_m \rangle v_m, \tag{5.1.7}$$

$$\langle f(K^*K)x_1, x_2 \rangle = \int_{\sigma(K^*K)} f(\sigma^2) d\langle E_{\sigma^2} x_1, x_2 \rangle = \sum_{m=1}^{\infty} f(\sigma_m^2) \overline{\langle y, v_m \rangle} \langle x, v_m \rangle, \tag{5.1.8}$$

$$\|f(K^*K)\| = \sup\{|f(\sigma^2)| : \sigma^2 \in \sigma(K^*K)\}. \tag{5.1.9}$$

Proposition 5.1.23. *Let A and T be two self-adjoint operators. Then, for every bounded Borel functions f and g, the product $f(A)g(T)$ commutes if A and T commute.*

Proof. See [94, Proposition 12.24]. □

Finally, let us introduce the following notation for *asymptotic equivalence*. Given two sequence $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}}$ we say that

$$a_n \sim b_n \quad \text{for } n \to \infty$$

if $\lim_{n\to\infty} a_n/b_n = 1$. Moreover, we will write

$$a_n = o(b_n)$$

if $\lim_{n\to\infty} a_n/b_n = 0$, and

$$a_n = O(b_n)$$

if $|a_n| \leq c|b_n|$ definitely for every $n \geq N$, with N fixed and $c > 0$.

5.2 Introduction to fractional and weighted Tikhonov variants

We consider linear operator equations of the form

$$Kx = y, \tag{5.2.1}$$

where $K : X \to Y$ is a compact linear operator between Hilbert spaces X and Y. We say y to be attainable if problem (5.2.1) has a solution $x^\dagger = K^\dagger y$ of minimal norm. Since $K^\dagger$ is unbounded when K is compact and has infinite dimensional range by virtue of Proposition 5.1.13, then problem (5.2.1) is ill-posed in the sense of Definition 4.0.2, and has to be regularized in order to compute a numerical solution. Generally speaking, problem (5.2.1) is approximated by a family of neighboring well-posed problems.

Namely, we want to approximate the solution $x^\dagger$ of the equation (5.2.1), when only an approximation y^δ of y is available with

$$\|y^\delta - y\| \leq \delta, \tag{5.2.2}$$

where δ is called the noise level. Since $K^\dagger y^\delta$ is not a good approximation of $x^\dagger$, we approximate $x^\dagger$ with $x_\alpha^\delta := R_\alpha y^\delta$ where $\{R_\alpha\}$ is a family of continuous operators depending on a parameter α that will be defined later. A classical example is the Tikhonov regularization defined by

$$R_\alpha = \underset{x \in X}{\operatorname{argmin}} \left\{ \|Kx - y^\delta\|_2^2 + \alpha \|x\|_2^2 \right\}, \tag{5.2.3}$$

or equivalently,

$$R_\alpha = (K^* K + \alpha I)^{-1} K^*,$$

where I denotes the identity and K^* the adjoint of K, cf. [55].

Using the singular values expansion of K, filter based regularization methods are defined in terms of functions of the singular values, cf. Proposition 5.3.3. This is a useful tool for the analysis of regularization techniques [63], both for direct and iterative regularization methods [59, 64]. Furthermore, new regularization methods can be defined investigating new classes of filters. For instance, one of the contributions in [73] is the proposal and the analysis of the fractional Tikhonov method. The authors obtain a new class of filtering regularization methods adding an exponent, depending on a parameter, to the filter of the standard Tikhonov method. They provide a detailed analysis of the filtering properties and the optimal order of the method in terms of such extra parameter. A different generalization of the Tikhonov method was recently proposed in [66] with a detailed filtering analysis. Both generalizations are called "fractional Tikhonov regularization" in the literature and they are compared in [51], where the optimal order of the method in [66] is provided as well. To distinguish the two proposals in [73] and [66], we will refer in the following as "fractional Tikhonov regularization" and "weighted-I Tikhonov regularization", respectively. These variants of the Tikhonov method have been introduced to compute good approximations of non-smooth solutions, since it is well known that the Tikhonov

method provides over-smoothed solutions. Finally, a third variant appeared in [69] with the purpose of exploiting the information carried by the spectrum of the operator itself in order to refine the tuning of the regularization process. Indeed, that method induces a pseudo-norm on the solution space X acting as a switch to force the regularization on the noise subspace. Due to this interpretation we will call that variant "weighted-II method" and we will analyze how a suitable mixing of weighted-I and weighted-II methods will make use of the good properties of both providing then a better approximate solution.

In this Chapter, we first provide a saturation result similar to the well-known saturation result for Tikhonov regularization [48]: indeed, Tikhonov regularization under suitable a-priori assumption and a-priori choice rule, $\alpha = \alpha(\delta) \sim c(\delta)^{2/3}$, is of optimal order and the best possible convergence rate obtainable is

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

On the other hand, let $\mathrm{Rg}(K)$ be the range of K and denoting by Q the orthogonal projection on the closure of $\overline{\mathrm{Rg}(K)}$ of the range of K, if

$$\sup\left\{\|x_\alpha^\delta - x^\dagger\| : \|Q(y-y^\delta)\| \le \delta\right\} = o(\delta^{\frac{2}{3}}),$$

then $x^\dagger = 0$, as long as $\mathrm{Rg}(K)$ is not closed, and this shows how Tikhonov regularization for an ill-posed problem with compact operator never yields a convergence rate which is faster than $O(\delta^{\frac{2}{3}})$, it saturates at this rate. Such result motivated us to introduce the iterated version of fractional and weighted-I/II Tikhonov in the same spirit of the iterated Tikhonov method. We prove that those iterated methods can overcome the previous saturation results. Afterwards, inspired by the works [17, 58] we introduce the nonstationary variants of our iterated methods. Differently from the nonstationary iterated Tikhonov, we have two nonstationary sequences of parameters. In the noise free case, we give sufficient conditions on these sequences to guarantee the convergence and obtaining the corresponding convergence rates. In the noise case, we show the stability of the proposed iterative schemes proving that they are regularization methods. Finally, few selected examples confirm the previous theoretical analysis, showing that a proper choice of the nonstationary sequences of parameters can yield better restorations compared to the classical iterated Tikhonov with a geometric sequence of regularization parameter according to [58].

The reminder of this Chapter is organized as follows. Section 5.3 recalls the basic definition of filter based regularization methods and of optimal order of a regularization method. Fractional Tikhonov regularization with optimal order and converse results are studied in Section 5.4, whereas smoothing effects of those methods are studied in the following Section 5.5-Section 5.6 is devoted to saturation results for both variants of fractional Tikhonov regularization. New iterated fractional Tikhonov regularization methods are introduced in Section 5.7, where the analysis of their convergence rate shows that they are able to overcome the previous saturation results. A nonstationary iterated weighted-I/II Tikhonov regularization is investigated in detail in Section 5.8, while a similar nonstationary iterated fractional Tikhonov regularization is discussed in Section 5.9. Finally, some numerical examples are reported in Section 5.10.

5.3 Filter method regularization

As described in the previous Section, we consider a compact linear operator $K \in \mathscr{B}_0(X,Y)$, with X and Y Hilbert spaces. Due to the compactness of the operator K we have a spectral representation of the generalized inverse $K^\dagger$.

By virtue of Proposition 5.1.19, we can give an equivalent definition of the generalized (Moore-Penrose) inverse $K^\dagger$ for compact operators.

Definition 5.3.1 (Generalized Inverse). *We define the* generalized inverse $K^\dagger : \mathrm{Dom}(K^\dagger) \subseteq Y \to X$ *of a compact linear operator* $K : X \to Y$ *as*

$$K^\dagger y = \sum_{m:\,\sigma_m>0} \sigma_m^{-1} \langle y, u_m \rangle v_m, \qquad y \in \mathrm{Dom}(K^\dagger), \tag{5.3.1}$$

where

$$\mathrm{Dom}(K^\dagger) = \left\{ y \in Y : \sum_{m:\,\sigma_m>0} \sigma_m^{-2} |\langle y, u_m \rangle|^2 < \infty \right\}.$$

With respect to problem (5.2.1), we consider the case where only an approximation y^δ of y satisfying the condition (5.2.2) is available, where $y^\delta = y + \eta$. Therefore $x^\dagger = K^\dagger y$, $y \in \mathrm{Dom}(K^\dagger)$, cannot be approximated by $K^\dagger y^\delta$, due to the unboundedness of $K^\dagger$ which is a consequence of $\lim_m \sigma_m = 0$, and hence in practice the problem (5.2.1) is approximated by a family of neighbouring well-posed problems [48]. The faster the convergence to 0 of the sequence $\{\sigma_m\}_{m\in\mathbb{N}}$ the worse is the ill-conditioning of the problem.

Definition 5.3.2. *By a* regularization method *for* $K^\dagger$ *we call any family of operators*

$$\{R_\alpha\}_{\alpha\in(0,\alpha_0)} : Y \to X, \qquad \alpha_0 \in (0, +\infty],$$

with the following properties:

(i) $R_\alpha : Y \to X$ *is a bounded operator for every* α.

(ii) For every $y \in \mathrm{Dom}(K^\dagger)$ *there exists a mapping (rule choice)* $\alpha : \mathbb{R}_+ \times Y \to (0, \alpha_0) \in \mathbb{R}$, $\alpha = \alpha(\delta, y^\delta)$, *such that*

$$\lim_{\delta\to 0} \sup \left\{ \alpha(\delta, y^\delta) : y^\delta \in Y, \|y - y^\delta\| \le \delta \right\} = 0,$$

and

$$\lim_{\delta\to 0} \sup \left\{ \|R_{\alpha(\delta,y^\delta)} y^\delta - K^\dagger y\| : y^\delta \in Y, \|y - y^\delta\| \le \delta \right\} = 0.$$

Throughout this Chapter, c is a constant which can change from one instance to the next. For the sake of clarity, if more than one constant will appear in the same line or equation we will distinguish them by means of a subscript.

Proposition 5.3.3. *Let $K : X \to Y$ be a compact linear operator and $K^\dagger$ its generalized inverse. Let $R_\alpha : Y \to X$ be a family of operators defined for every $\alpha \in (0, \alpha_0)$ as*

$$R_\alpha y := \sum_{m:\, \sigma_m > 0} F_\alpha(\sigma_m)\sigma_m^{-1}\langle y, u_m\rangle v_m, \tag{5.3.2}$$

where $F_\alpha : [0, \sigma_1] \supset \sigma(K) \to \mathbb{R}$ is a Borel function such that

$$\sup_{m:\, \sigma_m > 0} |F_\alpha(\sigma_m)\sigma_m^{-1}| = c(\alpha) < \infty, \tag{5.3.3a}$$

$$|F_\alpha(\sigma_m)| \leq c < \infty, \qquad \text{where } c \text{ does not depend on } (\alpha, m), \tag{5.3.3b}$$

$$\lim_{\alpha \to 0} F_\alpha(\sigma_m) = 1 \text{ point-wise in } \sigma_m. \tag{5.3.3c}$$

Then R_α is a regularization method, with $\|R_\alpha\| = c(\alpha)$, and it is called filter based regularization method*, while F_α is called* filter *function.*

Proof. First we observe that (5.3.3a) is sufficient for the well-posedness and continuity of the operator R_α, indeed

$$\begin{aligned} \|R_\alpha y\|^2 &= \sum_{n:\, \sigma_n > 0} |F_\alpha(\sigma_n)\sigma_n^{-1}|^2 |\langle y, u_n\rangle|^2 \\ &\leq \left(\sup_{0 < \sigma \leq \sigma_1} |F_\alpha(\sigma)\sigma^{-1}| \right)^2 \|y\|^2. \end{aligned}$$

To prove point *(ii)* of Definition 5.3.2 it is sufficient to prove that $R_\alpha \to K^\dagger$ point-wise as $\alpha \to 0$, see Proposition 3.4 in [48]. Observe that by (5.3.3b), we have that $|1 - F_\alpha(\sigma)| \leq 1 + c$. Hence, for every fixed $y \in \mathscr{D}(A^\dagger)$ it holds that

$$\begin{aligned} \lim_{\alpha \to 0} \|R_\alpha y - K^\dagger y\|^2 &= \lim_{\alpha \to 0} \sum_{n:\, \sigma_n > 0} \left(|1 - F_\alpha(\sigma_n)|\right)^2 \sigma_n^{-2} |\langle y, u_n\rangle|^2 \\ &= \sum_{n:\, \sigma_n > 0} \lim_{\alpha \to 0} \left(|1 - F_\alpha(\sigma_n)|\right)^2 \sigma_n^{-2} |\langle y, u_n\rangle|^2 \text{ (dominate convergence)}, \end{aligned}$$

and from (5.3.3c) the thesis follows. □

For ease of notation we set the following notations

$$x_\alpha := R_\alpha y, \qquad y \in \mathscr{D}(K^\dagger), \tag{5.3.4}$$

$$x_\alpha^\delta := R_\alpha y^\delta, \qquad y^\delta \in Y. \tag{5.3.5}$$

Remark 5.3.4. *Let $y^\delta = y + \eta$, with $\|\eta\| = \delta$, where η represent the error component in the observed data y^δ and δ is the noise level. Then we can write*

$$\begin{aligned} x_\alpha^\delta = R_\alpha y^\delta &= \sum_{m:\,\sigma_m>0} F_\alpha(\sigma_m)\sigma_m^{-1}\langle y^\delta, u_m\rangle v_m \\ &= \sum_{m:\,\sigma_m>0} F_\alpha(\sigma_m)\sigma_m^{-1}\langle y, u_m\rangle v_m + \sum_{m:\,\sigma_m>0} F_\alpha(\sigma_m)\sigma_m^{-1}\langle \eta, u_m\rangle v_m \\ &= \sum_{m:\,\sigma_m>0} \sigma_m^{-1}\langle y, u_m\rangle v_m - \sum_{m:\,\sigma_m>0} (1-F_\alpha(\sigma_m))\,\sigma_m^{-1}\langle y, u_m\rangle v_m \\ &\quad + \sum_{m:\,\sigma_m>0} F_\alpha(\sigma_m)\sigma_m^{-1}\langle \eta, u_m\rangle v_m \\ &= x^\dagger - e_a + e_n, \end{aligned}$$

where

$$e_a = \sum_{m:\,\sigma_m>0} (1-F_\alpha(\sigma_m))\,\sigma_m^{-1}\langle y, u_m\rangle v_m$$

is the approximation *error and*

$$e_n = \sum_{m:\,\sigma_m>0} F_\alpha(\sigma_m)\sigma_m^{-1}\langle \eta, u_m\rangle v_m$$

is the noise *error. When σ_m approaches 0, the noise error norm $\|e_n\|$ increases and the noise affects with greater impact the approximated solution x_α^δ. It is said tha the eigenvectors v_m belonging to small eigenvalues σ_m^2 generate the* noise subspace *whereas v_m corresponding to δ_m close to σ_1 generate the* signal subspace. *The role of the filter function F_α is then to mediate between the approximation error and the noise error, damping the effects produced by the noise subspace.*

Example 5.3.5 (**Classic Tikhonov filter**). *One of the most studied filter functions is the* classic Tikhonov filter

$$\mathfrak{F}_\alpha(\sigma) = \frac{\sigma^2}{\sigma^2+\alpha}, \tag{5.3.6}$$

with its associated Tikhonov regularization method

$$R_\alpha y := \sum_{m:\,\sigma_m>0} \mathfrak{F}_\alpha(\sigma_m)\sigma_m^{-1}\langle y, u_m\rangle v_m.$$

It is trivial to prove that the Tikhonov filter (5.3.6) *satisfies conditions* (5.3.3a)-(5.3.3c).

We report hereafter the definition of optimal order under a-priori assumption.

Definition 5.3.6 (**Optimal order under a-priori assumption**).
For every given $\nu, \rho > 0$, let

$$X_{\nu,\rho} := \left\{ x \in X : \exists\, \omega \in X, \|\omega\| \le \rho, x = (K^*K)^{\frac{\nu}{2}}\omega \right\} \subset X.$$

A regularization method R_α is called of optimal order *under the a-priori assumption $x^\dagger \in X_{\nu,\rho}$ if*

$$\Delta(\delta, X_{\nu,\rho}, R_\alpha) \leq c \cdot \delta^{\frac{\nu}{\nu+1}} \rho^{\frac{1}{\nu+1}}, \tag{5.3.7}$$

where for any general set $M \subseteq X$, $\delta > 0$ and for a regularization method R_α, we define

$$\Delta(\delta, M, R_\alpha) := \sup \left\{ \|x^\dagger - x_\alpha^\delta\| : x^\dagger \in M, \|y - y^\delta\| \leq \delta \right\}.$$

If ρ is not known, as it will be usually the case, then we relax the definition introducing the set

$$X_\nu := \bigcup_{\rho > 0} X_{\nu,\rho}$$

and we say that a regularization method R_α is called of optimal order *under the a-priori assumption $x^\dagger \in X_\nu$ if*

$$\Delta(\delta, X_\nu, R_\alpha) \leq c \cdot \delta^{\frac{\nu}{\nu+1}}. \tag{5.3.8}$$

Remark 5.3.7. *Since we are concerned with the convergence rate to* 0 *of $\|x^\dagger - x_\alpha^\delta\|$ as $\delta \to 0$, the a-priori assumption $x^\dagger \in X_\nu$ is usually sufficient for the optimal order analysis, requiring that* (5.3.8) *is satisfied.*

We state two preliminary lemmas useful to prove the next Theorem 5.3.10 which will give sufficient conditions for order optimality.

Lemma 5.3.8. *Let K be a compact linear operator and R_α a filter based regularization method for K. Then*

$$\|KR_\alpha y\| \leq c \cdot \|y\|,$$

where c is a constant that is independent of α and y.

Proof. By definition of a filter based regularization method in 5.3.3, it holds

$$\begin{aligned} KR_\alpha y &= \sum_{m:\, \sigma_m > 0} F_\alpha(\sigma_m) \sigma_m^{-1} \langle y, u_m \rangle K v_m \qquad \text{by continuity of } K, \\ &= \sum_{m:\, \sigma_m > 0} F_\alpha(\sigma_m) \langle y, u_m \rangle u_m \end{aligned}$$

and by (5.3.3b) the thesis follows. □

Lemma 5.3.9. *Let K be a compact linear operator and let $f : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Then*

$$f(K^*K)K^* = K^* f(KK^*). \tag{5.3.9}$$

Proof. Let $y \in Y$. Then

$$f(K^*K)K^* y = \sum_n f(\sigma_n^2) \langle K^* y, v_n \rangle v_n = \sum_n f(\sigma_n^2) \langle y, u_n \rangle K^* u_n = K^* f(KK^*) y.$$

□

We are now ready to introduce the following theorem which states sufficient conditions for order optimality, when filtering methods are employed.

Theorem 5.3.10. *[77] Let $K : X \to Y$ be a compact linear operator, ν and $\rho > 0$, and let $R_\alpha : Y \to X$ be a filter based regularization method. If there exists a fixed $\beta > 0$ such that*

$$\sup_{0<\sigma\leq\sigma_1} |F_\alpha(\sigma)\sigma^{-1}| \leq c \cdot \alpha^{-\beta}, \tag{5.3.10a}$$

$$\sup_{0\leq\sigma\leq\sigma_1} |(1-F_\alpha(\sigma))\sigma^{\nu}| \leq c_\nu \cdot \alpha^{\beta\nu}, \tag{5.3.10b}$$

then R_α is of optimal order, under the a-priori assumption $x^\dagger \in X_{\nu,\rho}$, with the choice rule

$$\alpha = \alpha(\delta,\rho) = \hat{c} \cdot \left(\frac{\delta}{\rho}\right)^{\frac{1}{\beta(\nu+1)}}, \qquad 0 < \hat{c} = \left(\frac{c}{\nu c_\nu}\right)^{\frac{1}{\beta(\nu+1)}}.$$

Proof. Using the notation in equations (5.3.4) and (5.3.5)

$$\|x^\dagger - x_\alpha^\delta\| \leq \|x^\dagger - x_\alpha\| + \|x_\alpha - x_\alpha^\delta\|. \tag{5.3.11}$$

We now study separately the two terms of the right-hand side of the previous inequality.

Observe that a filter based regularization method R_α can be restated as follows:

$$R_\alpha := \tilde{F}_\alpha(K^*K)K^*,$$

where

$$\tilde{F}_\alpha(K^*K) = F_\alpha((K^*K)^{\frac{1}{2}})(K^*K)^{-1} = \sum_{n:\,\sigma_n>0} F_\alpha(\sigma_n)\sigma_n^{-2}\langle\cdot, v_n\rangle v_n.$$

The boundedness of the Borel function $\tilde{F}_\alpha$ is a sufficient condition for the boundedness of R_α. Then

$$\begin{aligned}
\|x_\alpha - x_\alpha^\delta\|^2 &= \langle x_\alpha - x_\alpha^\delta, x_\alpha - x_\alpha^\delta\rangle \\
&= \langle x_\alpha - x_\alpha^\delta, R_\alpha(y-y^\delta)\rangle \\
&= \langle x_\alpha - x_\alpha^\delta, \tilde{F}_\alpha(K^*K)K^*(y-y^\delta)\rangle \\
&= \langle x_\alpha - x_\alpha^\delta, K^*\tilde{F}_\alpha(KK^*)(y-y^\delta)\rangle \qquad \text{(by Lemma 5.3.9)} \\
&= \langle Kx_\alpha - Kx_\alpha^\delta, \tilde{F}_\alpha(KK^*)(y-y^\delta)\rangle \\
&\leq \|KR_\alpha(y-y^\delta)\| \cdot \|\tilde{F}_\alpha(KK^*)\| \cdot \|y-y^\delta\| \\
&\leq c_1 \cdot \alpha^{-\beta}\delta^2,
\end{aligned} \tag{5.3.12}$$

where the last inequality follows thanks to Lemma 5.3.8 and (5.3.10a).

Notice that

$$\begin{aligned} x^\dagger - x_\alpha &= \sum_{n:\,\sigma_n>0} (1-F_\alpha(\sigma_n))\,\sigma_n^{-1}\langle y,u_n\rangle v_n \\ &= \sum_{n:\,\sigma_n>0} (1-F_\alpha(\sigma_n))\,\sigma_n^{-1}\langle Kx^\dagger,u_n\rangle v_n \\ &= \sum_{n:\,\sigma_n>0} (1-F_\alpha(\sigma_n))\,\langle x^\dagger,v_n\rangle v_n \\ &= (I-F_\alpha((K^*K)^{\frac{1}{2}}))x^\dagger. \end{aligned}$$

Then, using the fact that $x^\dagger \in X_{\nu,\rho}$ by assumption *(i)*, it follows that

$$\begin{aligned} \|x^\dagger - x_\alpha\| &= \|(I-F_\alpha((K^*K)^{\frac{1}{2}}))x^\dagger\| \\ &= \|(I-F_\alpha((K^*K)^{\frac{1}{2}}))(K^*K)^{\frac{\nu}{2}}\omega\| \\ &\le c_2\cdot\alpha^{\beta\frac{\nu}{2}}\rho, \end{aligned} \tag{5.3.13}$$

where the last inequality is a consequence of (5.3.10b).

Therefore, combining (5.3.13) and (5.3.12) with (5.3.11), we deduce that

$$\|x^\dagger - x_\alpha^\delta\| \le \sqrt{c_1}\cdot\delta\alpha^{-\frac{\beta}{2}} + c_2\cdot\alpha^{\beta\frac{\nu}{2}}\rho,$$

and the optimal order (5.3.7) is obtained by the choice rule $\alpha = \left(\frac{\delta}{\rho}\right)^{\frac{2}{\beta(\nu+1)}}$.

If instead we are in hypothesis *(ii)*, i.e., $x^\dagger \in X_\nu$ without any assumption on ρ, then equation (5.3.13) becomes

$$\|x^\dagger - x_\alpha\| \le c_3\cdot\alpha^{\beta\frac{\nu}{2}}, \tag{5.3.14}$$

since there exists $\tilde\rho > 0$ and $\omega \in X$, $\|\omega\| \le \tilde\rho$, such that $x^\dagger = (K^*K)^{\frac{\nu}{2}}\omega$. Now, combining (5.3.14) and (5.3.12) with (5.3.11) in the same way as before, the optimal order (5.3.8) is obtained by the choice rule $\alpha = \delta^{\frac{2}{\beta(\nu+1)}}$. □

If we are concerned just with the rate of convergence with respect to δ, the preceding theorem can be applied under the a-priori assumption $x^\dagger \in X_\nu$, adapting the proof to the latter case without any effort. On the contrary, below we present a converse result.

Theorem 5.3.11. *Let K be a compact linear operator with infinite dimensional range and let R_α be a filter based regularization method with filter function $F_\alpha : [0,\sigma_1] \supset \sigma(K) \to \mathbb{R}$. If there exist ν and $\beta > 0$ such that*

$$(1-F_\alpha(\sigma))\,\sigma^\nu \ge c\alpha^{\beta\nu} \qquad \text{for } \sigma \in [c'\alpha^\beta, \sigma_1] \tag{5.3.15}$$

and

$$\|x^\dagger - x_\alpha\| = O(\alpha^{\beta\nu}), \tag{5.3.16}$$

then $x^\dagger \in X_\nu$.

Proof. By (5.3.1) and (5.3.2), it holds

$$\begin{aligned}\|x^\dagger - x_\alpha\|^2 &= \sum_{\sigma_m>0} (1-F_\alpha(\sigma_m))^2\,\sigma_m^{-2}|\langle y,u_m\rangle|^2\\ &= \sum_{\sigma_m>0} (1-F_\alpha(\sigma_m))^2\,|\langle x^\dagger,v_m\rangle|^2\\ &= \sum_{\sigma_m>0} [(1-F_\alpha(\sigma_m))\,\sigma_m^\nu]^2\,\sigma_m^{-2\nu}|\langle x^\dagger,v_m\rangle|^2\\ &\geq \left(c\alpha^{\beta\nu}\right)^2 \sum_{\sigma_m\geq c'\alpha^\beta} \sigma_m^{-2\nu}|\langle x^\dagger,v_m\rangle|^2,\end{aligned}$$

thanks to the assumption (5.3.15). From (5.3.16) we deduce that

$$\lim_{\alpha^\beta\to 0} \sum_{\sigma_m\geq c'\alpha^\beta} \sigma_m^{-2\nu}|\langle x^\dagger,v_m\rangle|^2 < +\infty.$$

Finally, if we define $\omega := \sum_{\sigma_m>0}\sigma^{-\nu}\langle x^\dagger,v_m\rangle v_m$, then ω is well defined and $(K^*K)^{\nu/2}\,\omega = x^\dagger$, i.e., $x^\dagger\in X_\nu$. □

Corollary 5.3.12. *Let K be a compact linear operator with infinite dimensional range. The Tikhonov regularization method, R_α, is of optimal order under the a-priori assumption $x^\dagger\in X_{\nu,\rho}$, with $0<\nu\leq 2$. The best possible rate of convergence with respect to δ is $\|x^\dagger - x_\alpha^\delta\| = O\left(\delta^{\frac{2}{3}}\right)$, which is obtained for $\alpha = \left(\frac{\delta}{\rho}\right)^{\frac{2}{3}}$. On the other hand, if $\|x^\dagger - x_\alpha\| = O(\alpha)$ then $x^\dagger\in X_2$.*

Proof. It is not difficult to prove that the function

$$\mathfrak{F}_\alpha(\sigma)\sigma^{-1} = \frac{\sigma}{\sigma^2+\alpha},$$

has a maximum in $\sigma=\sqrt{\alpha}$ and

$$\sup_{0<\sigma\leq\sigma_1} \left|\mathfrak{F}_\alpha(\sigma)\sigma^{-1}\right| = \frac{1}{2\sqrt{\alpha}},$$

namely, condition 5.3.10 is satisfied for $\beta=1/2$. In the same way, the function

$$(1-\mathfrak{F}_\alpha(\sigma))\,\sigma^\nu = \frac{\alpha\sigma^\nu}{\sigma^2+\alpha}$$

has a maximum in $\sigma=\sqrt{\frac{\nu\alpha}{2-\nu}}$ for $0<\nu<2$, and it holds

$$\sup_{0<\sigma\leq\sigma_1} |(1-\mathfrak{F}_\alpha(\sigma))\,\sigma^\nu| = 2^{-1}\left(\frac{\nu}{2-\nu}\right)^{\nu/2}\alpha^{\nu/2},$$

and condition 5.3.10b is satisfied too. On the contrary, for $\nu = 2$ then

$$\frac{\alpha\sigma^2}{\sigma^2+\alpha}$$

is monotone increasing and therefore

$$(1-\mathfrak{F}_\alpha(\sigma))\,\sigma^2 \geq c\alpha \qquad \text{for every } \sigma \in [c\sqrt{\alpha}, \sigma_1].$$

Applying Theorem 5.3.11 we conclude. □

We will see later in Proposition 5.6.1 that $\delta^{\frac{2}{3}}$ is a kind of barrier for the convergence rate of the classic Tikhonov regularization method, namely it never yields a convergence rate which is faster than $O(\delta^{\frac{2}{3}})$. It is said that the Tikhonov method *saturates* at that rate.

Finally, the following proposition provides sufficient conditions under which a filter based method is not of optimal order with the a-priori assumption $x^\dagger \in X_\nu$.

Proposition 5.3.13. *Let K be a compact linear operator with infinite dimensional range and let R_α be a filter based regularization method with filter function $F_\alpha : [0,\sigma_1] \supset \sigma(K) \to \mathbb{R}$. If*

$$\limsup_n F_\alpha(\sigma_n)\sigma_n^{-2} = \infty, \tag{5.3.17a}$$

$$\liminf_n (1-F_\alpha(\sigma_n)) = c(\alpha) > 0, \tag{5.3.17b}$$

then, for any $\nu > 0$, R_α is not of optimal order under the a-priori assumption $x^\dagger \in X_\nu$, i.e., for every choice rule $\alpha = \alpha(\delta)$,

$$\Delta(\delta, X_\nu, R_\alpha) \neq O(\delta^{\frac{\nu}{1+\nu}}). \tag{5.3.18}$$

Proof. By hypothesis, $\sigma_n > 0$ for every n and $\sigma_n \to 0$ as $n \to \infty$. Let $\nu > 0$ and $\delta > 0$ be fixed, and let $\alpha = \alpha(\delta)$ be a generic choice rule. We define

$$\begin{aligned} y_n &:= u_n, \\ y_n^\delta &:= (1+\delta)u_n, \\ x_\alpha^n &:= R_\alpha y_n, \\ x_\alpha^{n,\delta} &:= R_\alpha y_n^\delta, \\ x_n^\dagger &:= K^\dagger y_n. \end{aligned}$$

Then $x_n^\dagger = \sigma_n^{-1} v_n \in X_{\nu,\sigma_n^{-\nu-1}} \subset X_\nu$ for every n, namely $\{x_n^\dagger\}_{n\in\mathbb{N}} \subset X_\nu$. It holds

$$x_\alpha^n - x_\alpha^{n,\delta} = \delta \cdot R_\alpha u_n = \delta \cdot F_\alpha(\sigma_n)\sigma_n^{-1} v_n, \tag{5.3.19}$$

$$x_n^\dagger - x_\alpha^n = \sum_{m:\,\sigma_m>0} (1-F_\alpha(\sigma_m))\,\sigma_m^{-1}\langle u_n, u_m\rangle v_m = (1-F_\alpha(\sigma_n))\,\sigma_n^{-1} v_n. \tag{5.3.20}$$

By hypothesis, for every fixed $\delta > 0$, $\alpha = \alpha(\delta)$, $0 < \varepsilon < 1$ and $k \in \mathbb{N}$, there exists $n_k \in \mathbb{N}$ such that for every $n \geq n_k$ we find

$$(1 - F_\alpha(\sigma_n)) > c(1-\varepsilon), \tag{5.3.21}$$

$$F_\alpha(\sigma_n)\sigma_n^{-2} > \frac{\delta^{\frac{\nu-1}{1+\nu}}}{c(1-\varepsilon)} \cdot k, \tag{5.3.22}$$

where c is the constant $c(\alpha)$ that appears in (5.3.17b). Therefore there exists a subsequence $\{\sigma_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\sigma_n\}_{n\in\mathbb{N}}$ depending on δ for which (5.3.21) and (5.3.22) hold for every n_k. Without loss of generality and for the sake of clarity, we may assume that $\sigma_{n_k} = \sigma_n$ and $k = n$. Thus we obtain

$$\begin{aligned}
\|x_n^\dagger - x_\alpha^{n,\delta}\|^2 &= \|(x_n^\dagger - x_\alpha^n) - (x_\alpha^n - x_\alpha^{n,\delta})\|^2 \\
&= \|x_n^\dagger - x_\alpha^n\|^2 + 2\langle x_n^\dagger - x_\alpha^n, x_\alpha^n - x_\alpha^{n,\delta}\rangle + \|x_\alpha^n - x_\alpha^{n,\delta}\|^2 \\
&\geq 2\langle x_n^\dagger - x_\alpha^n, x_\alpha^n - x_\alpha^{n,\delta}\rangle \\
&= 2\delta \cdot (1 - F_\alpha(\sigma_n))\, F_\alpha(\sigma_n)\sigma_n^{-2} \qquad \text{(by (5.3.19) and (5.3.20))} \\
&> 2n\delta^{\frac{2\nu}{1+\nu}}.
\end{aligned}$$

Hence $\Delta(\delta, X_\nu, R_\alpha) > \sqrt{2n} \cdot \delta^{\frac{\nu}{1+\nu}}$, and the thesis follows. □

5.4 Fractional variants of Tikhonov regularization

In this section we discuss three recent types of regularization methods that generalize the classical Tikhonov method and that were first introduced and studied in [66], [73] and [69]. We will use the notation $F_{\alpha,\cdot}$ to indicate the new filters, where $\cdot$ will be replaced by the extra parameter introduced by the respective method. Every method will be studied separately to avoid confusion and misunderstandings.

5.4.1 Weighted-I and Weighted-II Tikhonov regularization

Definition 5.4.1 ([66])**.** *We call* weighted-I Tikhonov *method the filter based method*

$$R_{\alpha,r}y := \sum_{m:\,\sigma_m>0} F_{\alpha,r}(\sigma_m)\sigma_m^{-1}\langle y, u_m\rangle v_m,$$

where the filter function is

$$F_{\alpha,r}(\sigma) = \frac{\sigma^{r+1}}{\sigma^{r+1} + \alpha}, \tag{5.4.1}$$

or equivalently

$$F_{\alpha,r}(\sigma) = \frac{\sigma^2}{\sigma^2 + \alpha\sigma^{1-r}}, \tag{5.4.2}$$

for $\alpha > 0$ and $r \geq 0$. For $r = 1$ the classic Tikhonov filter is recovered.

According to (5.3.4) and (5.3.5), we fix the following notation

$$x_{\alpha,r} := R_{\alpha,r} y, \qquad y \in \mathrm{Dom}(K^\dagger), \tag{5.4.3}$$

$$x^\delta_{\alpha,r} := R_{\alpha,r} y^\delta, \qquad y^\delta \in Y. \tag{5.4.4}$$

Definition 5.4.2 ([69])**.** *We call* weighted-II Tikhonov *method the filter based method*

$$R_{\alpha,j} y := \sum_{m:\, \sigma_m > 0} F_{\alpha,j}(\sigma_m)\sigma_m^{-1}\langle y, u_m\rangle v_m,$$

where the filter function is

$$F_{\alpha,j}(\sigma) = \frac{\sigma^2}{\sigma^2 + \alpha\left[1 - \left(\frac{\sigma}{\sigma_1}\right)^2\right]^j}, \tag{5.4.5}$$

for $\alpha > 0$ and $j \in \mathbb{N}$. For $j = 0$ the classic Tikhonov filter is recovered.

With reference to Remark 5.3.4, let us observe that the weighted-II filter $F_{\alpha,j}(\sigma)$ is almost 1 when σ belongs to the signal subspace and is almost the standard Tikhonov filter $F_\alpha(\sigma) = \frac{\sigma^2}{\sigma^2+\alpha}$ when σ belongs to the noise subspace. The idea is that in the signal subspace, i.e. when $\sigma \sim \sigma_1$, where the noise error norm is controlled, the regularization is minimal, bringing to 0 the approximation error, while the action of the filter function is focused on the noise subspace, i.e. when $\sigma \sim 0$. Generally speaking, the weighted-II filter acts like a switch for the regularization to take place. Like above, we fix the following notation

$$x_{\alpha,j} := R_{\alpha,j} y, \qquad y \in \mathrm{Dom}(K^\dagger), \tag{5.4.6}$$

$$x^\delta_{\alpha,j} := R_{\alpha,j} y^\delta, \qquad y^\delta \in Y. \tag{5.4.7}$$

Given an operator W on any Hilbert space, if we consider the semi-norm $\|\cdot\|_W$ induced by W, i.e. $\|x\|_W := \langle Wx, Wx\rangle$, then the weighted-I Tikhonov method can also be defined as the unique minimizer of the following functional,

$$R_{\alpha,r} y := \mathrm{argmin}_{x\in X}\left\{\|Kx - y\|_W^2 + \alpha\|x\|^2\right\}, \tag{5.4.8}$$

or, equivalently,

$$R_{\alpha,r} y := \mathrm{argmin}_{x\in X}\left\{\|Kx - y\|^2 + \alpha\|x\|_{W'}^2\right\}, \tag{5.4.9}$$

where the semi-norms $\|\cdot\|_W$ and $\|\cdot\|_{W'}$ are induced by the operators
$W := (KK^*)^{\frac{r-1}{4}} : Y \to Y$ and $W' := (K^*K)^{\frac{1-r}{4}} : X \to X$, respectively. For $0 \le r < 1$, W is to be intended as the Moore-Penrose (pseudo) inverse and that as well applies to W' in the case $r > 1$.

Looking for a stationary point in equation (5.4.8), we have that

$$\begin{aligned}
0 &= \nabla\left(\|Kx-y\|_W^2+\alpha\|x\|^2\right)\\
&= \nabla\left(\langle Kx,Kx\rangle_W - 2\langle Kx,y\rangle_W + \langle y,y\rangle_W + \alpha\langle x,x\rangle\right)\\
&= \nabla\left(\langle (KK^*)^{\frac{r-1}{4}}Kx,(KK^*)^{\frac{r-1}{4}}Kx\rangle - 2\langle (KK^*)^{\frac{r-1}{4}}Kx,(KK^*)^{\frac{r-1}{4}}y\rangle + \alpha\langle x,x\rangle\right)\\
&= \nabla\left(\langle x,K^*(KK^*)^{\frac{r-1}{2}}Kx\rangle - 2\langle x,K^*(KK^*)^{\frac{r-1}{2}}y\rangle + \alpha\langle x,x\rangle\right)\\
&= \nabla\left(\langle x,(K^*K)^{\frac{r-1}{2}}K^*Kx\rangle - 2\langle x,(K^*K)^{\frac{r-1}{2}}K^*y\rangle + \alpha\langle x,x\rangle\right)\\
&= \nabla\left(\langle x,(K^*K)^{\frac{r+1}{2}}x\rangle - 2\langle x,(K^*K)^{\frac{r-1}{2}}K^*y\rangle + \alpha\langle x,x\rangle\right)\\
&= 2(K^*K)^{\frac{r+1}{2}}x - 2(K^*K)^{\frac{r-1}{2}}K^*y + 2\alpha x,
\end{aligned}$$

from which we deduce the following expression for the operator $R_{\alpha,r}$,

$$R_{\alpha,r}y = \left[(K^*K)^{\frac{r+1}{2}}+\alpha I\right]^{-1}(K^*K)^{\frac{r-1}{2}}K^*y, \tag{5.4.10}$$

$$= \left[K^*K+\alpha(K^*K)^{\frac{1-r}{2}}\right]^{-1}K^*y. \tag{5.4.11}$$

In the same way, the weighted-II Tikhonov method can be defined as the unique minimizer of the functional

$$R_{\alpha,j}y := \operatorname{argmin}_{x\in X}\left\{\|Kx-y\|^2+\alpha\|x\|_B^2\right\}, \tag{5.4.12}$$

where the semi-norm $\|\cdot\|_B$ is induced by the operator $B := \left(I-\frac{K^*K}{\|K^*K\|}\right)^{j/2} : X\to X$, and developing calculations as above, with the only difference that now the weighted norm $\|\cdot\|_B$ of X applies to the second addendum, we can deduce that

$$R_{\alpha,j}y = \left[K^*K+\alpha\left(I-\frac{K^*K}{\|K^*K\|}\right)^j\right]^{-1}K^*y. \tag{5.4.13}$$

Both the methods can be classified then in the more general contest of weighted generalized inverse methods, namely

$$R_\alpha y := \operatorname{argmin}_{x\in X}\left\{\|Kx-y\|^2+\alpha\|x\|_\Lambda^2\right\}, \tag{5.4.14}$$

or again

$$R_\alpha y = \left[K^*K+\alpha\Lambda^*\Lambda\right]^{-1}K^*y, \tag{5.4.15}$$

where Λ is a suitable operator. We will not get into details, for references see [48, Chapter 8]. We just want to observe that if $\Lambda^*\Lambda$ and K^*K commute, then indicating with $(\lambda_n;v_n,u_n)_{n\in\mathbb{N}}$ the s.v.e. of Λ, the operator (5.4.15) can be expressed in the following way

$$R_\alpha y := \sum_{m:\,\sigma_m>0} F_\alpha(\sigma_m,\lambda_m)\sigma_m^{-1}\langle y,u_m\rangle v_m, \qquad \text{with } F_\alpha(\sigma,\lambda)=\frac{\sigma^2}{\sigma^2+\alpha\lambda^2}. \tag{5.4.16}$$

Now, let $f : [0, \sigma_1] \to \mathbb{R}$ be a continuous function and consider the operator $f(K^*K)$, that commutes with K^*K. From equations (5.4.16), (5.4.2) and (5.4.5), it is clear that both the weighted-I and weighted-II filter methods are of the form (5.4.14) with $\Lambda = f(K^*K)$ and where $f(\sigma^2) = \sigma^{1-r}$ and $f(\sigma^2) = \left(1 - \left(\frac{\sigma}{\sigma_1}\right)^2\right)^j$, respectively for the weighted-I and weighted-II case. That is the reason that motivated us to rename the original method of Hochstenbach and Reichel, that appeared in [66], into *weighted-I Tikhonov method* and subsequently to rename the method of Huckle, that appeared in [69], into *weighted-II Tikhonov method*. In this way it would be easier to distinguish from the *fractional Tikhonov method* introduced by Klann and Ramlau in [73].

The optimal order of the weighted-I Tikhonov regularization was proved in [51]. The following proposition restates such result, putting in evidence the dependence on r of ν, and provides a converse result.

Proposition 5.4.3. *Let K be a compact linear operator with infinite dimensional range. For every given $r \geq 0$ the weighted-I Tikhonov method, $R_{\alpha,r}$, is a regularization method of optimal order, under the a-priori assumption $x^\dagger \in X_{\nu,\rho}$, with $0 < \nu \leq r+1$. The best possible rate of convergence with respect to δ is $\|x^\dagger - x^\delta_{\alpha,r}\| = O\left(\delta^{\frac{r+1}{r+2}}\right)$, which is obtained for $\alpha = \left(\frac{\delta}{\rho}\right)^{\frac{r+1}{\nu+1}}$ with $\nu = r+1$. On the other hand, if $\|x^\dagger - x_{\alpha,r}\| = O(\alpha)$ then $x^\dagger \in X_{r+1}$.*

Proof. First, it is easy to check the validity of (5.3.3a), (5.3.3b) and (5.3.3c). Therefore, weighted-I Tikhonov is a regularization filter method and it remains to prove the optimal order property. The left-hand side of condition (5.3.10a) becomes

$$\sup_{0<\sigma\leq\sigma_1} \left| \frac{\sigma^r}{\sigma^{r+1} + \alpha} \right|.$$

By derivation, if $r > 0$ then it is straightforward to see that the quantity above is bounded by $\alpha^{-\beta}$, with $\beta = 1/(r+1)$. Similarly, the left-hand side of condition (5.3.10b) takes the form

$$\sup_{0\leq\sigma\leq\sigma_1} \left| \frac{\alpha\sigma^\nu}{\sigma^{r+1} + \alpha} \right|,$$

and it is easy to check that it is bounded by $\alpha^{\beta\nu}$ if and only if $0 < \nu \leq r+1$. From Theorem 5.3.10, as long as $0 < \nu \leq r+1$, with $r > 0$, if $x^\dagger \in X_{\nu,\rho}$ then we find order optimality (5.3.7) and the best possible rate of convergence obtainable with respect to δ is $O(\delta^{\frac{\nu}{\nu+1}})$, for $\nu = r+1$.

On the contrary, with $\beta = 1/(r+1)$ and $\nu = r+1$, we deduce that

$$|(1 - F_{\alpha,r}(\sigma))\,\sigma^\nu| = \frac{\alpha\sigma^\nu}{\sigma^{r+1} + \alpha} \geq \frac{1}{2}\alpha, \qquad \text{for } \sigma \in [\alpha^\beta, \sigma_1].$$

Therefore, if $\|x^\dagger - x_{\alpha,r}\| = O(\alpha)$ then $x^\dagger \in X_\nu$ by Theorem 5.3.11. □

The next proposition provides a proof for the optimal order of weighted-II Tikhonov regularization which instead fails to have a converse result.

Proposition 5.4.4. *Let K be a compact linear operator with infinite dimensional range. For every given integer $j \geq 0$ the weighted-II Tikhonov method, $R_{\alpha,j}$, is a regularization method of optimal order, under the a-priori assumption $x^\dagger \in X_{\nu,\rho}$, with $0 < \nu \leq 2$. The best possible rate of convergence with respect to δ is $\|x^\dagger - x^\delta_{\alpha,j}\| = O\left(\delta^{\frac{2}{3}}\right)$, that is obtained for $\alpha = \left(\frac{\delta}{\rho}\right)^{\frac{2}{\nu+1}}$ with $\nu = 2$.*

Proof. Even in this case, the validity of (5.3.3a), (5.3.3b) and (5.3.3c) are trivial to check having as a consequence that the weighted-II Tikhonov is a regularization filter method. Without loss of generality, we suppose $\sigma_1 = 1$. The left-hand side of condition (5.3.10b) takes the form

$$\sup_{0\leq\sigma\leq 1}\left|\frac{\alpha\left(1-\sigma^2\right)^j\sigma^\nu}{\sigma^2+\alpha\left(1-\sigma^2\right)^j}\right|,$$

which is bounded above by

$$\sup_{0\leq\sigma\leq 1}|\alpha g(\sigma)|, \qquad g(\sigma)=\frac{\sigma^\nu}{\sigma^2+\alpha\left(1-\sigma^2\right)^j}.$$

Let us study the function $g(\sigma)$. By deriving it, we get that g has a maximum at every point σ_* which satisfy the following equation

$$(1-\sigma^2)^{j-1}=\frac{(2-\nu)\sigma^2}{\alpha\left[2j\sigma^2+\nu\left(1-\sigma^2\right)\right]},$$

if and only if $\nu \in [0,2]$. It is not difficult to see that there exist only one $\sigma_* \in [0,1]$ which satisfies it, see the following graphic example, Figure 5.1,

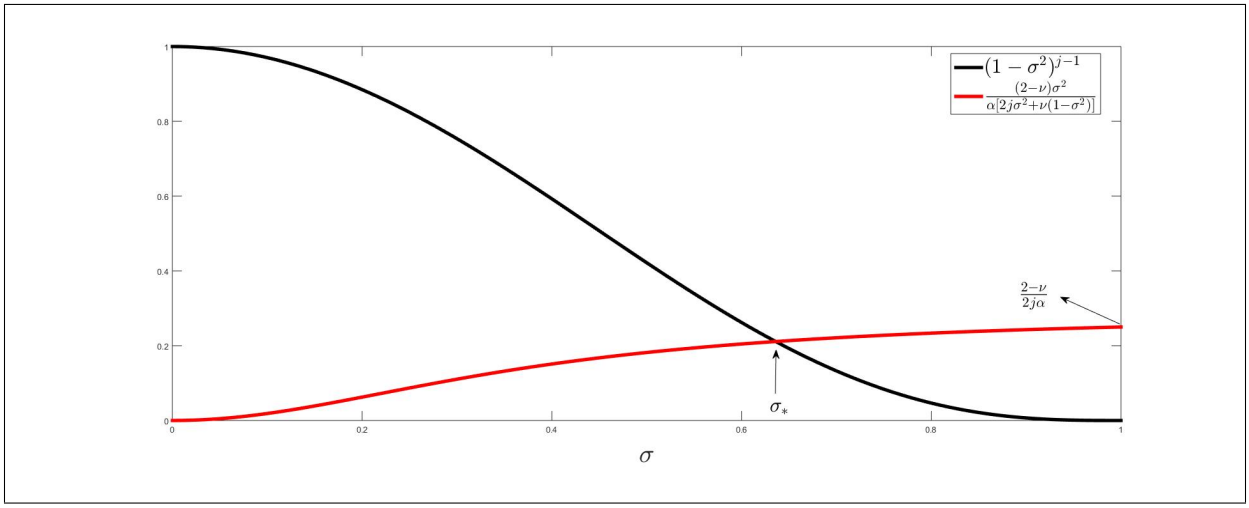


Figure 5.1: Graphic solution for σ_*.

and that

$$\sigma_*=\sqrt{\alpha}\left\{\frac{(1-\sigma_*^2)^{j-1}\left[2j\sigma_*^2+\nu\left(1-\sigma_*^2\right)\right]}{2-\nu}\right\}^{1/2}=\sqrt{\alpha}h(\sigma_*),$$

with

$$h(\sigma_*) = \left\{ \frac{(1-\sigma_*^2)^{j-1}\left[2j\sigma_*^2 + \nu\left(1-\sigma_*^2\right)\right]}{2-\nu} \right\}^{1/2}.$$

If we fix $\alpha \in (0, \alpha_0)$, with $\alpha_0 < \infty$, then necessarily $\sigma_* \in (0, \lambda_{\alpha_0,j})$, where $\lambda_{\alpha_0,j} < 1$. Since $h(\sigma_*) = 0$ if and only if $\sigma_* = 1$, then $h(\sigma_*)$ is uniformly bounded away from 0, i.e., $h(\sigma_*) \in [\hat{c}, 1]$, with $\hat{c} = \hat{c}(\alpha_0, j, \nu) > 0$ and independent of α. Henceforth, we can write

$$\sigma_* = c_{\alpha_0,j,\nu}\sqrt{\alpha} = c\sqrt{\alpha},$$

and then we have

$$\begin{aligned}
\sup_{0\le\sigma\le 1} \left|\left(1 - F_{\alpha,j}(\sigma)\right)\sigma^\nu\right| &= \sup_{0\le\sigma\le 1} \left| \frac{\alpha\left(1-\sigma^2\right)^j \sigma^\nu}{\sigma^2 + \alpha\left(1-\sigma^2\right)^j} \right| \\
&\le \sup_{0\le\sigma\le 1} |\alpha g(\sigma)| \\
&= \alpha g(\sigma_*) \\
&= \alpha g(c\sqrt{\alpha}) \\
&= \frac{\alpha^{\nu/2}}{c^2 + (1-c^2\alpha)^j} \\
&\le c^{-2}\alpha^{\nu/2}, \qquad (5.4.17)
\end{aligned}$$

which is (5.3.10b) with $\beta = 1/2$. Instead, the validity of (5.3.10a) comes again from studying the function g, having fixed $\nu = 1$. Therefore, from Theorem 5.3.10, as long as $0 < \nu \le 2$, if $x^\dagger \in X_{\nu,\rho}$ then we find order optimality (5.3.7) and the best possible rate of convergence obtainable with respect to δ is $O(\delta^{\frac{\nu}{\nu+1}})$, for $\nu = 2$. □

Observe that the optimal order for the weighted-II Tikhonov is independent of the auxiliary parameter j.

5.4.2 Fractional Tikhonov regularization

Here we introduce the *fractional Tikhonov* method defined and discussed in [73].

Definition 5.4.5 ([73])**.** *We call* Fractional Tikhonov *method the filter based method*

$$R_{\alpha,\gamma} y := \sum_{m:\,\sigma_m>0} F_{\alpha,\gamma}(\sigma_m)\sigma_m^{-1}\langle y, u_m\rangle v_m,$$

where the filter function is

$$F_{\alpha,\gamma}(\sigma) = \frac{\sigma^{2\gamma}}{(\sigma^2+\alpha)^\gamma}, \qquad (5.4.18)$$

for $\alpha > 0$ and $\gamma \ge 1/2$. For $\gamma = 1$ the classic Tikhonov filter is recovered.

Note that $F_{\alpha,\gamma}$ is well-defined also for $0<\gamma<1/2$, but the condition (5.3.3a) requires $\gamma\geq 1/2$ to guarantee that $F_{\alpha,\gamma}$ is a filter function.

We use the notation for $x_{\alpha,\gamma}$ and $x^{\delta}_{\alpha,\gamma}$ like in equations (5.4.3) and (5.4.4), respectively. The optimal order of the fractional Tikhonov regularization was proved in [73, Proposition 3.2]. The following proposition restates such result including also $\gamma=1/2$ and provides a converse result.

Proposition 5.4.6. *The extended fractional Tikhonov filter method is a regularization method of optimal order, under the a-priori assumption $x^{\dagger}\in X_{\nu,\rho}$, for every $\gamma\geq 1/2$ and $0<\nu\leq 2$. The best possible rate of convergence with respect to δ is $\|x^{\dagger}-x^{\delta}_{\alpha,\gamma}\|=O\left(\delta^{\frac{2}{3}}\right)$, that is obtained for $\alpha=\left(\frac{\delta}{\rho}\right)^{\frac{2}{\nu+1}}$ with $\nu=2$. On the other hand, if $\|x^{\dagger}-x_{\alpha,\gamma}\|=O(\alpha)$ then $x^{\dagger}\in X_2$.*

Proof. Condition (5.3.3a) is verified for $\gamma\geq 1/2$ and the same holds for conditions (5.3.3b) and (5.3.3c). Deriving the filter function, it is immediate to see that equation (5.3.10a) is verified for $\gamma\geq 1/2$, with $\beta=1/2$. It remains to check equation (5.3.10b):

$$\begin{aligned}\left(1-F_{\alpha,\gamma}(\sigma)\right)\sigma^{\nu}&=\frac{\left(\sigma^2+\alpha\right)^{\gamma}-\sigma^{2\gamma}}{\left(\sigma^2+\alpha\right)^{\gamma}}\sigma^{\nu}\\&=\frac{\left(\frac{\sigma^2}{\alpha}+1\right)^{\gamma}-\left(\frac{\sigma^2}{\alpha}\right)^{\gamma}}{\left(\frac{\sigma^2}{\alpha}+1\right)^{\gamma-1}}\cdot\frac{\alpha\sigma^{\nu}}{\sigma^2+\alpha}\\&=h\left(\frac{\sigma^2}{\alpha}\right)\cdot\left(1-\mathfrak{F}_{\alpha}(\sigma)\right)\sigma^{\nu},\end{aligned}$$

where $h(x)=\frac{(x+1)^{\gamma}-x^{\gamma}}{(x+1)^{\gamma-1}}$ is monotone, $h(0)=1$ for every γ, and $\lim_{x\to\infty}h(x)=\gamma$. Namely $h(x)\in(\gamma,1]$ for $0\leq\gamma\leq 1$ and $h(x)\in[1,\gamma)$ for $\gamma\geq 1$. Therefore we deduce that

$$\gamma(1-\mathfrak{F}_{\alpha}(\sigma))\leq\left(1-F_{\alpha,\gamma}(\sigma)\right)\leq(1-\mathfrak{F}_{\alpha}(\sigma)),\qquad\text{for } 0\leq\gamma\leq 1,\tag{5.4.19}$$

$$(1-\mathfrak{F}_{\alpha}(\sigma))\leq\left(1-F_{\alpha,\gamma}(\sigma)\right)\leq\gamma(1-\mathfrak{F}_{\alpha}(\sigma)),\qquad\text{for } \gamma\geq 1,\tag{5.4.20}$$

from which we infer that

$$\sup_{\sigma\in[0,\sigma_1]}\left|\left(1-F_{\alpha,\gamma}(\sigma)\right)\sigma^{\nu}\right|\leq\max\{1,\gamma\}\sup_{\sigma\in[0,\sigma_1]}|(1-\mathfrak{F}_{\alpha}(\sigma))\,\sigma^{\nu}|\leq c\alpha^{\frac{\nu}{2}},\tag{5.4.21}$$

since $\mathfrak{F}_{\alpha}(\sigma)$ is standard Tikhonov, that is of optimal order, with $\beta=1/2$ and for every $0<\nu\leq 2$, see Corollary 5.3.12. On the contrary, with $\beta=1/2$ and $\nu=2$, and by equations (5.4.19) and (5.4.20), we deduce that

$$\left(1-F_{\alpha,\gamma}(\sigma)\right)\sigma^2\geq\min\{1,\gamma\}\,(1-\mathfrak{F}_{\alpha}(\sigma))\,\sigma^2\geq\frac{1}{2}\alpha,\qquad\text{for }\sigma\in[\alpha^{\frac{1}{2}},\sigma_1].\tag{5.4.22}$$

Therefore, if $\|x^{\dagger}-x_{\alpha,r}\|=O(\alpha)$ then $x^{\dagger}\in X_2$ by Theorem 5.3.11. □

5.5 Smoothing effect

In this section we deal with the *oversmoothing* property that affects the classical Tikhonov regularization method. Indeed, it was observed that the approximated solution is smoother than the true solution, i.e., it lives in a space of higher regularity. We will see that the weighted-I and fractional filters can overcome the oversmoothing effect for a proper choice of their extra regularization parameters. In order to easily understand this kind of behavior we are going to restrict our study to the fractional Sobolev spaces H^s of one dimensional functions.

Let $\Omega = [0, 2\pi]$ and let $J_s : \mathrm{H}^s(\Omega) \hookrightarrow \mathrm{L}^2(\Omega)$ be the embedding operator of the Hilbert space $\mathrm{H}^s(\Omega)$, with $s \in (0, \infty)$. It was seen in Proposition 5.1.20 that J_s is compact with s.v.e. given by

$$v_m(t) = \left(1+m^2\right)^{-s/2} \mathrm{e}^{\mathrm{i}mt}, \qquad u_m = \mathrm{e}^{\mathrm{i}mt}, \qquad \sigma_m = \left(1+m^2\right)^{-s/2}.$$

Let us consider the following problem

$$J_s x = y. \tag{5.5.1}$$

Since J_s is compact, and therefore ill-conditioned, we regularize the above problem introducing a family of filter functions,

$$x_\alpha^\delta(t) = \sum_{m>0} F_\alpha\left(\sigma_m\right) \sigma_m^{-1} \langle \cdot, u_m \rangle v_m(t).$$

The true solution $x^\dagger$, i.e., $J_s x^\dagger = y$, belongs to H^s. We are concerned about the regularity of the approximated solution x_α^δ when we deal with general approximated data $y^\delta \in \mathrm{L}^2$. The next propositions state the H^p spaces in which the approximated solutions live depending on the filter method.

Proposition 5.5.1. *For data $y^\delta \in \mathrm{L}^2(\Omega)$, the approximated solution $x_{\alpha,r}^\delta$ of the weighted-I Tikhonov filter for Problem* (5.5.1) *belongs to* $\mathrm{H}^{s(r+1)}(\Omega)$.

Proof. We have that

$$\begin{aligned} x_{\alpha,r}^\delta(\cdot) &= \sum_{m>0} \left(\frac{\sigma_m^{r+1}}{\sigma_m^{r+1} + \alpha} \right) \sigma_m^{-1} \langle y^\delta, u_m \rangle v_m(\cdot) \\ &= \sum_{m>0} \left(\frac{(1+m^2)^{-(r+1)s/2}}{(1+m^2)^{-(r+1)s/2} + \alpha} \right) \langle y^\delta, u_m \rangle \mathrm{e}^{\mathrm{i}m\cdot}. \end{aligned}$$

Then, the Fourier coefficients of $x_{\alpha,r}^\delta$ are given by

$$\left(x_{\alpha,r}^\delta\right)_m = \left(\frac{(1+m^2)^{-(r+1)s/2}}{(1+m^2)^{-(r+1)s/2} + \alpha} \right) \langle y^\delta, u_m \rangle,$$

from which we can calculate the H^p norm,

$$\begin{aligned}\|x_{\alpha,r}^{\delta}\|_{\mathrm{H}^p}^2 &= \sum_{m>0} \left(1+m^2\right)^p \left(\frac{(1+m^2)^{-(r+1)s/2}}{(1+m^2)^{-(r+1)s/2}+\alpha}\right)^2 \left|\langle y^{\delta}, u_m\rangle\right|^2 \\ &\leq c_{\alpha} \sum_{m>0} \left(1+m^2\right)^{p-s(r+1)} \left|\langle y^{\delta}, u_m\rangle\right|^2. \end{aligned} \tag{5.5.2}$$

The right hand-side of the above inequality is bounded for every data $y^{\delta} \in \mathrm{L}^2(\Omega)$ if $p \leq s(1+r)$. □

Proposition 5.5.2. *For data $y^{\delta} \in \mathrm{L}^2(\Omega)$, the approximated solution $x_{\alpha,j}^{\delta}$ of the weighted-II Tikhonov filter for Problem* (5.5.1) *belongs to $\mathrm{H}^{2s}(\Omega)$, for any $j \in \mathbb{N}$.*

Proof. The proof is almost identical to the previous one, with the only difference in the filter function which is applied. Equation (5.5.2) becomes

$$\begin{aligned}\|x_{\alpha,r}^{\delta}\|_{\mathrm{H}^p}^2 &= \sum_{m>0} \left(1+m^2\right)^p \left(\frac{(1+m^2)^{-s}}{(1+m^2)^{-s}+\alpha\left[1-(1+m^2)^{-s}\right]^j}\right)^2 \left|\langle y^{\delta}, u_m\rangle\right|^2 \\ &\leq c_{\alpha} \sum_{m>0} \left(1+m^2\right)^{p-2s} \left|\langle y^{\delta}, u_m\rangle\right|^2, \end{aligned}$$

and again, the right hand-side of the above inequality is bounded for every data $y^{\delta} \in \mathrm{L}^2(\Omega)$ if $p \leq 2s$. □

Proposition 5.5.3. *For data $y^{\delta} \in \mathrm{L}^2(\Omega)$, the approximated solution $x_{\alpha,\gamma}^{\delta}$ of the Tikhonov fractional filter for Problem* (5.5.1) *belongs to $\mathrm{H}^{2s\gamma}(\Omega)$.*

Proof. Even in this case, the strategy of the proof is the same. Equation (5.5.2) becomes

$$\begin{aligned}\|x_{\alpha,r}^{\delta}\|_{\mathrm{H}^p}^2 &= \sum_{m>0} \left(1+m^2\right)^p \left(\frac{(1+m^2)^{-s}}{(1+m^2)^{-s}+\alpha}\right)^{2\gamma} \left|\langle y^{\delta}, u_m\rangle\right|^2 \\ &\leq c_{\alpha} \sum_{m>0} \left(1+m^2\right)^{p-2s\gamma} \left|\langle y^{\delta} u_m\rangle\right|^2. \end{aligned}$$

The right hand-side is bounded for every data $y^{\delta} \in \mathrm{L}^2(\Omega)$ if $p \leq 2s\gamma$. □

In Proposition 5.5.1, for $r = 1$ we recover the classical Tikhonov filter and its *oversmoothing* property, i.e., if the true solution $x^{\dagger} \in \mathrm{H}^s$, then the approximated solution $x_{\alpha}^{\delta} \in \mathrm{H}^{2s}$. Therefore, the weighted-I Tikhonov filter *undersmooth* the approximated solution for every $0 < r < 1$, compared to the classical Tikhonov. The same remark goes for the fractional Tikhonov filter with $0 < \gamma < 1/2$, while the weighted-II Tikhonov does not provide any *undersmoothing* effect for any $j \in \mathbb{N}$.

5.6 Saturation results

The following proposition deals with a saturation property similar to a well known result for classic Tikhonov, cf. [48, Proposition 5.3]. We generalize it to regularization methods of the form

$$R_{\alpha,f}y := \operatorname{argmin}_{x\in X}\left\{\|Kx-y\|^2+\alpha\|x\|^2_{f(K^*K)}\right\},$$

or equivalently

$$R_{\alpha,f}y = (K^*K+\alpha f(K^*K))^{-1}K^*y, \tag{5.6.1}$$

where $f:[0,\sigma_1]\to\mathbb{R}$ is a bounded measurable function such that the corresponding filter function

$$F_{\alpha,f}(\sigma) = \frac{\sigma^2}{\sigma^2+\alpha f(\sigma^2)}$$

satisfies properties (5.3.3a), (5.3.3b) and (5.3.3c), see the preceding discussion for (5.4.14). Again, we fix the following notation,

$$\begin{aligned} x_{\alpha,f} &:= R_{\alpha,f}y, \qquad & y\in \mathrm{Dom}(K^\dagger),\\ x^\delta_{\alpha,f} &:= R_{\alpha,f}y^\delta, \qquad & y^\delta\in Y. \end{aligned}$$

Proposition 5.6.1 (**Saturation for weighted Tikhonov regularization**).
Let $K:X\to Y$ be a compact linear operator with infinite dimensional range and $R_{\alpha,f}$ be the corresponding family of regularization operators as in equation 5.6.1. Let $\alpha=\alpha(\delta,y^\delta)$ be any parameter choice rule and let

$$\frac{\sigma^2}{f(\sigma^2)}\sim c\sigma^s \qquad \text{as } \sigma\to 0, \tag{5.6.2}$$

with $c,s>0$. If

$$\sup\left\{\|x^\delta_{\alpha,f}-x^\dagger\| : \|Q(y-y^\delta)\|\le\delta\right\} = o(\delta^{\frac{s}{s+1}}), \tag{5.6.3}$$

then $x^\dagger=0$, where we indicated with Q the orthogonal projector onto $\overline{R(K)}$.

Proof. Define

$$\begin{aligned} &\delta_m := \sigma_m^{s+1}, && y^\delta_m := y+\delta_m u_m && \text{so that } \|y-y^\delta_m\|\le\delta_m,\\ &\alpha_m := \alpha(\delta_m,y^\delta_m), && x_m := x_{\alpha_m,f}, && x^\delta_m := x^{\delta_m}_{\alpha_m,f}. \end{aligned}$$

By the assumption that K has not finite dimensional range, then $\lim_{m\to\infty}\sigma_m=0$. According to 5.6.1 we have

$$\begin{aligned} x^\delta_m - x^\dagger &= R_{\alpha_m,f}y^\delta_m - x^\dagger\\ &= R_{\alpha_m,f}y+\delta_m R_{\alpha_m,f}u_m - x^\dagger\\ &= x_m - x^\dagger + \delta_m F_{\alpha_m,f}(\sigma_m)\sigma_m^{-1}v_m \end{aligned}$$

and hence by (5.4.1)

$$\|x_m^\delta - x^\dagger\|^2 = \|x_m - x^\dagger\|^2 + 2\frac{\delta_m\sigma_m}{\sigma_m^2+\alpha_m f(\sigma_m^2)}\operatorname{Re}\langle x_m - x^\dagger, v_m\rangle + \left(\frac{\delta_m\sigma_m}{\sigma_m^2+\alpha_m f(\sigma_m^2)}\right)^2.$$

Since $F_{\alpha,f}$ satisfies (5.3.3a), we can deduce that f can not be identically zero in any interval of the form $[0,\lambda]$ and therefore it is possible to divide by $f(\sigma_m^2)$ if we took a suitable subsequence $\{\sigma_{m_n}\}\subseteq\{\sigma_m\}$. Without loss of generality, we assume $\{\sigma_m\}=\{\sigma_{m_n}\}$. Then, we have that

$$\begin{aligned}\left[\left(\frac{\delta_m\sigma_m}{f(\sigma_m^2)}\right)^{-\frac{1}{2}}\|x_m^\delta - x^\dagger\|\right]^2 &\geq \frac{2}{\frac{\sigma_m^2}{f(\sigma_m^2)}+\alpha_m}\operatorname{Re}\langle x_m - x^\dagger, v_m\rangle + \frac{\delta_m\sigma_m f(\sigma_m^2)}{[\sigma_m^2+\alpha_m f(\sigma_m^2)]^2}\\ &= \left[\frac{\frac{2f(\sigma_m^2)}{\sigma_m^2}}{1+\alpha_m\frac{f(\sigma_m^2)}{\sigma_m^2}}\right]\operatorname{Re}\langle x_m - x^\dagger, v_m\rangle + \frac{\delta_m\sigma_m^{-3}f(\sigma_m^2)}{\left[1+\alpha_m\frac{f(\sigma_m^2)}{\sigma_m^2}\right]^2},\end{aligned}$$

and passing to the lim sup, recalling assumption (5.6.2) and that $\delta_m=\sigma_m^{s+1}$, we get

$$\begin{aligned}\limsup_{m\to\infty}\left(\delta_m^{-\frac{s}{s+1}}\|x_m^\delta - x^\dagger\|\right)^2 \geq c\Bigg\{&\limsup_{\sigma_m\to\infty}\left[\frac{2\delta_m^{-\frac{s}{s+1}}}{1+\alpha_m\delta_m^{-\frac{s}{s+1}}}\right]\operatorname{Re}\langle x_m - x^\dagger, v_m\rangle\\ &+\liminf_{m\to\infty}\frac{1}{\left[1+\alpha_m\delta_m^{-\frac{s}{s+1}}\right]^2}\Bigg\},\end{aligned}\tag{5.6.4}$$

where c is a positive constant.

By 5.6.1,

$$\begin{aligned}(K^*K+\alpha_m f(K^*K))(x^\dagger - x_m^\delta) &= K^*Kx^\dagger + \alpha_m f(K^*K)x^\dagger - K^*y_m^\delta\\ &= \alpha_m f(K^*K)x^\dagger - \delta_m K^*u_m,\end{aligned}\tag{5.6.5}$$

so that

$$\begin{aligned}\alpha_m x^\dagger &= \delta_m[f(K^*K)]^{-1}K^*u_m + \left([f(K^*K)]^{-1}K^*K+\alpha_m I\right)(x^\dagger - x_m^\delta)\\ &= \frac{\delta_m\sigma_m}{f(\sigma_m^2)}v_m + \left([f(K^*K)]^{-1}K^*K+\alpha_m I\right)(x^\dagger - x_m^\delta).\end{aligned}$$

By hypothesis, $\lim_{\sigma\to0}\sigma^2/f(\sigma^2)\sim\lim_{\sigma\to0}\sigma^s=0$, and by (ii) from Definition 5.3.2

$$\alpha_m\leq\sup\left\{\alpha(\delta_m,y^{\delta_m}) : y^{\delta_m}\in Y, \|y-y^{\delta_m}\|\leq\delta_m\right\}\longrightarrow 0 \quad \text{as } \delta_m\to 0,$$

namely, $\{\alpha_m\}$ is uniformly bounded. Henceforth,

$$\|[f(K^*K)]^{-1}K^*K+\alpha_m I\|\leq c \quad \text{for every } m\in\mathbb{N},$$

and then

$$\alpha_m\|x^\dagger\| = O\left(\frac{\delta_m\sigma_m}{f(\sigma_m^2)} + \|x^\dagger - x_m^\delta\|\right). \tag{5.6.6}$$

Since $\delta_m = \sigma_m^{s+1}$ and, again, $\lim_{m\to\infty}\sigma_m^2/f(\sigma_m^2) = c\sigma_m^s$, it follows from (5.6.6) that

$$\alpha_m\|x^\dagger\| \le c\left(\delta_m^{\frac{2s}{s+1}} + \|x^\dagger - x_m^\delta\|\right).$$

Then, if $x^\dagger \neq 0$,

$$\lim_{m\to\infty}\alpha_m\delta_m^{-\frac{s}{s+1}} = 0, \tag{5.6.7}$$

because by assumption, $\|x^\dagger - x_m^\delta\| = o\left(\delta_m^{\frac{s}{s+1}}\right)$.

Hence, the second term in the right-hand side of (5.6.4) tends to 1. Since, by assumption, the left-hand side of (5.6.4) tends to 0, we obtain

$$0 \ge c\left\{\limsup_{m\to\infty}\frac{2}{1+\delta_m^{-\frac{s}{s+1}}\alpha_m}\delta_m^{-\frac{s}{s+1}}\,\mathrm{Re}\langle x_m - x^\dagger, v_m\rangle + 1\right\}.$$

Now, from (5.6.3) we have that $\|x_m - x^\dagger\| = o\left(\delta^{\frac{s}{s+1}}\right)$ as well, so that, if $x^\dagger \neq 0$, we obtain, from (5.6.7) applied to the preceding inequality, the contradiction $0 \ge c > 0$. Hence, $x^\dagger = 0$. □

Note that for $f(\sigma^2) \equiv 1$ (classical Tikhonov) the previous proposition gives exactly Proposition 5.3 in [48] with $s = 2$.

For $f(\sigma^2) = \sigma^{1-r}$ and $f(\sigma^2) = \left(1-\sigma^2\right)^j$ we recover instead saturation results for weighted-I and weighted-II regularization methods, respectively. Indeed,

$$\frac{\sigma^2}{\sigma^{1-r}} \sim \sigma^{r+1}, \qquad \frac{\sigma^2}{(1-\sigma^2)^j} \sim \sigma^2 \quad \text{for } \sigma \to 0.$$

We can state then the following corollaries.

Corollary 5.6.2. *With the same notation of the preceding Proposition 5.6.1, let $R_{\alpha,r}$ be the family of regularization operators as in Definition 5.4.1. If*

$$\sup\left\{\|x_{\alpha,r}^\delta - x^\dagger\| : \|Q(y-y^\delta)\| \le \delta\right\} = o(\delta^{\frac{r+1}{r+2}}),$$

then $x^\dagger = 0$.

Observe that taking a large r, it is possible to overcome the saturation result of classical Tikhonov obtaining a convergence rate arbitrarily close to $O(\delta)$.

Corollary 5.6.3. *With the same notation of the preceding Proposition 5.6.1, let $R_{\alpha,j}$ be the family of regularization operators as in Definition 5.4.2. If*

$$\sup\left\{\|x_{\alpha,j}^\delta - x^\dagger\| : \|Q(y-y^\delta)\| \le \delta\right\} = o(\delta^{\frac{2}{3}}),$$

then $x^\dagger = 0$.

In this case instead, weighted-II Tikhonov saturates at the same level of classical Tikhonov, independently on the choice of the parameter j.

A similar saturation result can be proved also for the fractional Tikhonov regularization in Definition 5.4.5.

Proposition 5.6.4 (Saturation for fractional Tikhonov regularization)**.** *Let $K : X \to Y$ be a compact linear operator with infinite dimensional range and let $R_{\alpha,\gamma}$ be the corresponding family of fractional Tikhonov regularization operators in Definition 5.4.5, with fixed $\gamma \geq 1/2$. Let $\alpha = \alpha(\delta, y^\delta)$ be any parameter choice rule. If*

$$\sup\left\{\|x^\delta_{\alpha,\gamma} - x^\dagger\| : \|Q(y-y^\delta)\| \leq \delta\right\} = o(\delta^{\frac{2}{3}}), \tag{5.6.8}$$

then $x^\dagger = 0$, where we indicated with Q the orthogonal projector onto $\overline{R(K)}$.

Proof. If $\gamma = 1$, the thesis follows from the saturation result for standard Tikhonov [48, Proposition 5.3] or Proposition 5.6.1. For $\gamma \neq 1$, recalling that

$$x_{\alpha,\gamma} - x^\dagger = \sum_{\sigma_m > 0} \left(F_{\alpha,\gamma}(\sigma_m) - 1\right) \sigma_m^{-1} \langle y, u_m\rangle v_m,$$

by equations (5.4.19) and (5.4.20), we obtain

$$\|x_{\alpha,\gamma} - x^\dagger\| > c\|x_{\alpha,1} - x^\dagger\|, \tag{5.6.9}$$

where $c = \min\{1, \gamma\}$ and $x_{\alpha,1}$ is standard Tikhonov. Let us define

$$\phi_\gamma(y) := \|x_{\alpha,\gamma} - x^\dagger\|.$$

Then, by the continuity of ϕ_γ, there exists $\delta > 0$ such that, for every $y^\delta \in \overline{B}_\delta(y)$, we find

$$\phi_\gamma(y^\delta) > c \cdot \phi_1(y^\delta),$$

with $\overline{B}_\delta(y)$ being the closure of the ball of center y and radius δ. Passing to the sup we obtain that

$$\sup\left\{\|x^\delta_{\alpha,\gamma} - x^\dagger\| : \|Q(y-y^\delta)\| \leq \delta\right\} \geq c \cdot \sup\left\{\|x^\delta_{\alpha,1} - x^\dagger\| : \|Q(y-y^\delta)\| \leq \delta\right\}. \tag{5.6.10}$$

Therefore, using relation (5.6.8), we deduce

$$\sup\left\{\|x^\delta_{\alpha,1} - x^\dagger\| : \|y-y^\delta\| \leq \delta\right\} = o(\delta^{\frac{2}{3}}), \tag{5.6.11}$$

and the thesis follows again from the saturation result for standard Tikhonov, see Proposition 5.6.1. □

Even in this case, differently from the weighted-I Tikhonov regularization, for the fractional Tikhonov method it is not possible to overcome the saturation result of classical Tikhonov, even for a large γ.

5.7 Stationary iterated regularization

We define new iterated regularization methods based on weighed and fractional Tikhonov regularization using the same iterative refinement strategy of iterated Tikhonov regularization [17, 48]. We will show that the iterated methods go beyond the saturation results proved in the previous section. In this section the regularization parameter will still be α with the iteration step, n, assumed to be fixed. On the contrary, in Section 5.8, we will analyze the nonstationary counterpart of this iterative method, in which α will be replaced by a pre-fixed sequence $\{\alpha_n\}$ and we will be concerned on the rate of convergence with respect to the index n.

5.7.1 Iterated weighted Tikhonov regularization

We propose now an iterated regularization methods based on weighted-I/II Tikhonov.

Definition 5.7.1 (**Stationary iterated weighted-I Tikhonov**). *We define the* stationary iterated weighted-I Tikhonov method *(*SIWT-I*) as*

$$\begin{cases} x^0_{\alpha,r} := 0; \\ \left((K^*K)^{\frac{r+1}{2}} + \alpha I\right) x^n_{\alpha,r} := (K^*K)^{\frac{r-1}{2}} K^* y + \alpha x^{n-1}_{\alpha,r}, \end{cases} \tag{5.7.1}$$

with $\alpha > 0$ and $r \geq 0$, or equivalently

$$\begin{cases} x^0_{\alpha,r} := 0 \\ x^n_{\alpha,r} := \ \operatorname{argmin}_{x\in X} \left\{ \|Kx - y\|^2_W + \alpha \|x - x^{n-1}_{\alpha,r}\|^2 \right\}, \end{cases} \tag{5.7.2}$$

where $\|\cdot\|_W$ is the semi-norm introduced in (5.4.8). *We define $x^{n,\delta}_{\alpha,r}$ as the n-th iteration of weighted-I Tikhonov if $y = y^\delta$.*

Proposition 5.7.2. *For any given $n \in \mathbb{N}$ and $r > 0$, the SIWT in* (5.7.1) *is a filter based regularization method, with filter function*

$$F^{(n)}_{\alpha,r}(\sigma) = \frac{(\sigma^{r+1} + \alpha)^n - \alpha^n}{(\sigma^{r+1} + \alpha)^n}. \tag{5.7.3}$$

Moreover, the method is of optimal order, under the a-priori assumption $x^\dagger \in X_{\nu,\rho}$, for $r > 0$ and $0 < \nu \leq n(r+1)$, with best convergence rate $\|x^\dagger - x^{n,\delta}_{\alpha,r}\| = O(\delta^{\frac{n(r+1)}{1+n(r+1)}})$, that is obtained for $\alpha = (\frac{\delta}{\rho})^{\frac{n(r+1)}{1+\nu}}$, with $\nu = n(r+1)$. On the other hand, if $\|x^\dagger - x^n_{\alpha,r}\| = O(\alpha^n)$, then $x^\dagger \in X_{n(r+1)}$.

Proof. Multiplying both sides of (5.7.1) by $\left((K^*K)^{\frac{r+1}{2}} + \alpha I\right)^{n-1}$ and iterating the process, we get

$$\begin{aligned} \left((K^*K)^{\frac{r+1}{2}} + \alpha I\right)^n x^n_{\alpha,r} &= \left\{ \sum_{j=0}^{n-1} \alpha^j \left((K^*K)^{\frac{r+1}{2}} + \alpha I\right)^{n-1-j} \right\} (K^*K)^{\frac{r-1}{2}} K^* y \\ &= \left[\left((K^*K)^{\frac{r+1}{2}} + \alpha I\right)^n - \alpha^n I \right] (K^*K)^{-1} K^* y. \end{aligned}$$

Therefore, the filter function in (5.3.2) is equal to

$$F_{\alpha,r}^{(n)}(\sigma) = \frac{(\sigma^{r+1}+\alpha)^n - \alpha^n}{(\sigma^{r+1}+\alpha)^n},$$

as we stated. Condition (5.3.3c) is straightforward to verify. Moreover, note that

$$\begin{aligned} F_{\alpha,r}^{(n)}(\sigma) &= \frac{(\sigma^{r+1}+\alpha)^n - \alpha^n}{(\sigma^{r+1}+\alpha)^n} \\ &= \frac{\sigma^{r+1}}{\sigma^{r+1}+\alpha} \cdot \frac{\left(\sum_{j=0}^{n-1} \alpha^j (\sigma^{r+1}+\alpha)^{n-1-j}\right)}{(\sigma^{r+1}+\alpha)^{n-1}} \\ &= F_{\alpha,r}(\sigma) \cdot \left(1 + \left(\frac{\alpha}{\sigma^{r+1}+\alpha}\right) + \cdots + \left(\frac{\alpha}{\sigma^{r+1}+\alpha}\right)^{n-1}\right), \end{aligned}$$

from which it follows that

$$F_{\alpha,r}(\sigma) \le F_{\alpha,r}^{(n)}(\sigma) \le n F_{\alpha,r}(\sigma). \tag{5.7.4}$$

Therefore, conditions (5.3.3a), (5.3.3b) and (5.3.10a) follow immediately by the regularity of the weighted-I Tikhonov filter method for $r > 0$ and by the order optimality for $r > 0$. Finally, condition (5.3.10b) becomes

$$\sup_{\sigma\in[0,\sigma_1]} \left| \frac{\alpha^n \sigma^\nu}{(\sigma^{r+1}+\alpha)^n} \right|,$$

and deriving one checks that it is bounded by $\alpha^{\beta\nu}$, with $\beta = 1/(r+1)$, if and only if $0 < \nu \le n(r+1)$. Applying now Proposition 5.3.10 the rest of the thesis follows.

On the contrary, if we define $\beta = 1/(r+1)$ and $\nu = n(r+1)$, then we deduce that

$$\left(1 - F_{\alpha,r}^{(n)}(\sigma)\right) \sigma^\nu = \frac{\alpha^n \sigma^\nu}{(\sigma^{r+1}+\alpha)^n} \ge \frac{1}{2^n}\alpha^n \qquad \text{for } \sigma \in [\alpha^\beta, \sigma_1].$$

Therefore, if $\|x^\dagger - x_{\alpha,r}^n\| = O(\alpha^n)$, then by Theorem 5.3.11 it follows that $x^\dagger \in X_{n(r+1)}$. □

If n is large, then we note that the convergence rate approaches $O(\delta)$ also for a fixed small r. The study of the convergence for increasing n and fixed α will be dealt with in Section 5.8.

Definition 5.7.3 (Stationary iterated weighted-II Tikhonov). *We define the* stationary iterated weighted-II Tikhonov method *(*SIWT-II*) as*

$$\begin{cases} x_{\alpha,j}^0 := 0; \\ \left(K^*K + \alpha\left[I - \frac{K^*K}{\|K^*K\|}\right]^j\right) x_{\alpha,j}^n := K^*y + \alpha\left[I - \frac{K^*K}{\|K^*K\|}\right]^j x_{\alpha,j}^{n-1}, \end{cases} \tag{5.7.5}$$

with $\alpha > 0$ and $j \in \mathbb{N}$, or equivalently

$$\begin{cases} x_{\alpha,j}^0 := 0 \\ x_{\alpha,j}^n := \ \operatorname{argmin}_{x\in X} \left\{ \|Kx - y\|_B^2 + \alpha \|x - x_{\alpha,j}^{n-1}\|^2 \right\}, \end{cases} \tag{5.7.6}$$

where $\|\cdot\|_B$ *is the semi-norm introduced in* (5.4.12). *We define* $x_{\alpha,j}^{n,\delta}$ *as the n-th iteration of weighted-II Tikhonov if* $y = y^\delta$.

Proposition 5.7.4. *For any given* $n, j \in \mathbb{N}$, *the SIWT-II in* (5.7.1) *is a filter based regularization method, with filter function*

$$F_{\alpha,j}^{(n)}(\sigma) = \frac{\left(\sigma^2 + \alpha\left[1 - \left(\frac{\sigma^2}{\sigma_1}\right)^2\right]^j\right)^n - \left(\alpha\left[1 - \left(\frac{\sigma^2}{\sigma_1}\right)^2\right]^j\right)^n}{\left(\sigma^2 + \alpha\left[1 - \left(\frac{\sigma^2}{\sigma_1}\right)^2\right]^j\right)^n}. \tag{5.7.7}$$

Moreover, the method is of optimal order, under the a-priori assumption $x^\dagger \in X_{\nu,\rho}$, *for* $j \in \mathbb{N}$ *and* $0 < \nu \leq 2n$, *with best convergence rate* $\|x^\dagger - x_{\alpha,j}^{n,\delta}\| = O(\delta^{\frac{2n}{1+2n}})$, *that is obtained for* $\alpha = (\frac{\delta}{\rho})^{\frac{2n}{1+\nu}}$, *with* $\nu = 2n$.

Proof. The first part of the proof mimics Proposition 5.7.2's proof, but we generalize it for a wider class of regularization methods. Indeed, let us consider a bounded measurable function $f : [0, \sigma_1] \to \mathbb{R}$ such satisfies conditions (5.3.3a), (5.3.3b) and (5.3.3c), and let us introduce the following iterated stationary method,

$$\begin{cases} x_{\alpha_0,f}^0 := 0, \\ [K^*K + \alpha_n f(K^*K)]\, x_{\alpha_n,f}^n := K^*y + \alpha_n f(K^*K)\, x_{\alpha_{n-1},f}^{n-1}, \end{cases} \tag{5.7.8}$$

or equivalently

$$\begin{cases} x_{\alpha_0,f}^0 := 0, \\ x_{\alpha_n,f}^n := \operatorname{argmin}_{x\in X}\left\{\|Kx - y\|^2 + \alpha_n \|x - x_{\alpha_{n-1},f}^{n-1}\|_{f(K^*K)}^2\right\}. \end{cases} \tag{5.7.9}$$

Multiplying both sides of (5.7.8) by $(K^*K + \alpha f(K^*K))^{n-1}$ and iterating the process, and using Proposition 5.1.23, we get

$$\begin{aligned} (K^*K + \alpha f(K^*K))^n x_{\alpha,f}^n &= \left\{\sum_{j=0}^{n-1} [\alpha f(K^*K)]^j (K^*K + \alpha f(K^*K))^{n-1-j}\right\} K^*y \\ &= [(K^*K + \alpha f(K^*K))^n - (\alpha f(K^*K))^n](K^*K)^{-1}K^*y. \end{aligned}$$

Therefore, the filter function in (5.3.2) is equal to

$$F_{\alpha,f}^{(n)}(\sigma) = \frac{\left(\sigma^2 + \alpha f\left(\sigma^2\right)\right)^n - \left(\alpha f\left(\sigma^2\right)\right)^n}{\left(\sigma^2 + \alpha f\left(\sigma^2\right)\right)^n}.$$

If we fix

$$f\left(\sigma^2\right) = \left[1 - \left(\frac{\sigma}{\sigma_1}\right)^2\right]^j,$$

then we get the filter function for the SIWT-II, i.e.,

$$F_{\alpha,j}^{(n)}(\sigma)=\frac{\left(\sigma^2+\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j\right)^n-\left(\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j\right)^n}{\left(\sigma^2+\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j\right)^n}.$$

Again, condition (5.3.3c) is straightforward to verify. Moreover, we recover the following relation

$$\begin{aligned}F_{\alpha,f}^{(n)}(\sigma)&=\frac{\left(\sigma^2+\alpha f\left(\sigma^2\right)\right)^n-\left(\alpha f\left(\sigma^2\right)\right)^n}{\left(\sigma^2+\alpha f\left(\sigma^2\right)\right)^n}\\&=\frac{\sigma^2}{\sigma^2+\alpha f\left(\sigma^2\right)}\cdot\frac{\left(\sum_{j=0}^{n-1}\left(\alpha f\left(\sigma^2\right)\right)^j\left(\sigma^2+\alpha f\left(\sigma^2\right)\right)^{n-1-j}\right)}{\left(\sigma^2+\alpha f\left(\sigma^2\right)\right)^{n-1}}\\&=F_{\alpha,f}(\sigma)\cdot\left(1+\left(\frac{\alpha f\left(\sigma^2\right)}{\sigma^2+\alpha f\left(\sigma^2\right)}\right)+\cdots+\left(\frac{\alpha f\left(\sigma^2\right)}{\sigma^2+\alpha f\left(\sigma^2\right)}\right)^{n-1}\right),\end{aligned}$$

from which it follows that

$$F_{\alpha,f}(\sigma)\le F_{\alpha,f}^{(n)}(\sigma)\le nF_{\alpha,f}(\sigma).$$

Therefore, once we substitute in the above inequalities the general function f for $\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j$, conditions (5.3.3a), (5.3.3b) and (5.3.10a) follow immediately by the regularity and by the order optimality of the weighted-II Tikhonov filter method for every $j\in\mathbb{N}$. Finally, condition (5.3.10b) becomes

$$\begin{aligned}\sup_{\sigma\in[0,\sigma_1]}\left|\left(\frac{\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j}{\sigma^2+\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j}\right)^n\sigma^\nu\right|&\le\sup_{\sigma\in[0,\sigma_1]}\left|\left(\frac{\alpha}{\sigma^2+\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j}\right)^n\sigma^\nu\right|\\&=\sup_{\sigma\in[0,\sigma_1]}\left|\left(\frac{\alpha\sigma^{\nu/n}}{\sigma^2+\alpha\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^j}\right)^n\right|,\end{aligned}$$

and then, using the same approach like in (5.4.17), it is easy to check that the last term in the above inequality is bounded by $\alpha^{\beta\nu}$, with $\beta=1/2$, if and only if $0<\nu\le 2n$. Applying now Proposition 5.3.10 the rest of the thesis follows. □

5.7.2 Iterated fractional Tikhonov regularization

With the same path as in the previous subsection, we propose here the stationary iterated version of the fractional Tikhonov method.

Definition 5.7.5 (Stationary iterated fractional Tikhonov)**.** *We define the* stationary iterated fractional Tikhonov method *(*SIFT*) as*

$$\begin{cases} x^0_{\alpha,\gamma} := 0; \\ (K^*K+\alpha I)^{\gamma} x^n_{\alpha,\gamma} := (K^*K)^{\gamma-1}K^*y + \left[(K^*K+\alpha I)^{\gamma} - (K^*K)^{\gamma}\right] x^{n-1}_{\alpha,\gamma}, \end{cases} \tag{5.7.10}$$

with $\gamma \geq 1/2$. We define $x^{n,\delta}_{\alpha,\gamma}$ for the n-th iteration of fractional Tikhonov if $y = y^{\delta}$.

Proposition 5.7.6. *For any given $n \in \mathbb{N}$ and $\gamma \geq 1/2$, the SIFT in* (5.7.10) *is a filter based regularization method, with filter function*

$$F^{(n)}_{\alpha,\gamma}(\sigma) = \frac{\left(\sigma^2+\alpha\right)^{\gamma n} - \left[\left(\sigma^2+\alpha\right)^{\gamma} - \sigma^{2\gamma}\right]^n}{(\sigma^2+\alpha)^{\gamma n}}. \tag{5.7.11}$$

Moreover, the method is of optimal order, under the a-priori assumption $x^{\dagger} \in X_{\nu,\rho}$, for $\gamma \geq 1/2$ and $0 < \nu \leq 2n$, with best convergence rate $\|x^{\dagger} - x^{n,\delta}_{\alpha,\gamma}\| = O(\delta^{\frac{2n}{2n+1}})$, that is obtained for $\alpha = (\frac{\delta}{\rho})^{\frac{2n}{\nu+1}}$, with $\nu = 2n$. On the other hand, if $\|x^{\dagger} - x^{n}_{\alpha,\gamma}\| = O(\alpha^n)$, then $x^{\dagger} \in X_{2n}$.

Proof. Multiplying both sides of (5.7.11) by $(K^*K+\alpha I)^{(n-1)\gamma}$ and iterating the process, we get

$$\begin{aligned} (K^*K+\alpha I)^{n\gamma} x^n_{\alpha,\gamma} &= \left\{ \sum_{j=0}^{n-1} (K^*K+\alpha I)^{j\gamma} \left[(K^*K+\alpha I)^{\gamma} - (K^*K)^{\gamma}\right]^{n-1-j} \right\} (K^*K)^{\gamma-1}K^*y \\ &= \left\{ (K^*K+\alpha I)^{\gamma n} - \left[(K^*K+\alpha I)^{\gamma} - (K^*K)^{\gamma}\right]^n \right\} (K^*K)^{-1}K^*y, \end{aligned}$$

where we used the fact that $(K^*K+\alpha I)^{-\gamma}$ and $\left[(K^*K+\alpha I)^{\gamma} - (K^*K)^{\gamma}\right]$ commute. Therefore, the filter function in (5.3.2) is given by

$$F^{n}_{\alpha,\gamma}(\sigma) = \frac{(\sigma^2+\alpha)^{\gamma n} - \left[\left(\sigma^2+\alpha\right)^{\gamma} - \sigma^{2\gamma}\right]^n}{(\sigma^2+\alpha)^{\gamma n}},$$

as we stated. We observe that

$$\begin{aligned} F^{(n)}_{\alpha,\gamma}(\sigma) &= \frac{(\sigma^2+\alpha)^{\gamma n} - \left[\left(\sigma^2+\alpha\right)^{\gamma} - \sigma^{2\gamma}\right]^n}{(\sigma^2+\alpha)^{\gamma n}} \\ &= \frac{\sigma^{2\gamma}}{(\sigma^2+\alpha)^{\gamma}} \cdot \frac{1}{(\sigma^2+\alpha)^{\gamma(n-1)}} \cdot \sum_{j=0}^{n-1} (\sigma^2+\alpha)^{\gamma j} \left[(\sigma^2+\alpha)^{\gamma} - \sigma^{2\gamma}\right]^{n-1-j} \\ &= \frac{\sigma^{2\gamma}}{(\sigma^2+\alpha)^{\gamma}} \cdot \left\{ 1 + \left[1 - \left(\frac{\sigma^2}{\sigma^2+\alpha}\right)^{\gamma}\right] + \cdots + \left[1 - \left(\frac{\sigma^2}{\sigma^2+\alpha}\right)^{\gamma}\right]^{n-1} \right\}, \end{aligned}$$

from which we deduce that

$$F^{(n)}_{\alpha,\gamma}(\sigma) \leq n F_{\alpha,\gamma}(\sigma). \tag{5.7.12}$$

Therefore, since $F_{\alpha,\gamma}$ is a regularization method of optimal order, conditions (5.3.3a), (5.3.3b) and (5.3.10a) are satisfied. Moreover, it is easy to check condition (5.3.3c) and so we get the regularity for the method. It remains to check condition (5.3.10b) for the order optimality.

From equations (5.4.19) and (5.4.20) we deduce that

$$\begin{aligned} 1-F_{\alpha,\gamma}^{(n)}(\sigma) &= \left[\frac{(\sigma^2+\alpha)^\gamma-\sigma^{2\gamma}}{(\sigma^2+\alpha)^\gamma}\right]^n \\ &= \left[1-\frac{\sigma^{2\gamma}}{(\sigma^2+\alpha)^\gamma}\right]^n \\ &= \left(1-F_{\alpha,\gamma}(\sigma)\right)^n \\ &\leq (\max\{1,\gamma\})^n\,(1-\mathfrak{F}_\alpha(\sigma))^n \\ &= c\,(1-\mathfrak{F}_\alpha^n(\sigma))\,, \end{aligned} \tag{5.7.13}$$

where $\mathfrak{F}_\alpha(\sigma)$ is the standard Tikhonov filter and $\mathfrak{F}_\alpha^{(n)}(\sigma)$ is the filter function of the stationary iterated Tikhonov, i.e., $\mathfrak{F}_\alpha^{(n)}(\sigma)=\frac{(\sigma^2+\alpha)^n-\alpha^n}{(\sigma^2+\alpha)^n}$. Now condition (5.3.10b) follows from the properties of stationary iterated Tikhonov, with $\beta=1/2$ and $0<\nu\leq 2n$, see [59, p. 124]. By applying Proposition 5.3.10 we get the best convergence rate, $O(\delta^{\frac{2n}{2n+1}})$.

On the contrary, set $\beta=1/2$ and $\nu=2n$. First, let us observe that from equations (5.7.13) and (5.4.19), (5.4.20), we infer that

$$1-F_{\alpha,\gamma}^{(n)}(\sigma)\geq(\min\{1,\gamma\})^n\left(1-\mathfrak{F}_\alpha^{(n)}(\sigma)\right).$$

Then, we deduce that

$$\begin{aligned} \left(1-F_{\alpha,\gamma}^{(n)}(\sigma)\right)\sigma^\nu &\geq c\frac{\alpha^n\sigma^{2n}}{(\sigma^2+\alpha)^n} \\ &\geq c\alpha^n \qquad \text{for } \sigma\in[\alpha^\beta,\sigma_1]. \end{aligned}$$

Therefore, if $\|x^\dagger-x_{\alpha,\gamma}^n\|=O(\alpha^n)$, then $x^\dagger\in X_{2n}$ by Theorem 5.3.11. □

The previous proposition shows that, similarly to SIWT, a large n allows to overcome the saturation result in Proposition 5.6.4. The study of the convergence for increasing n and fixed α will be dealt with in Section 5.9.

5.8 Nonstationary iterated weighted Tikhonov regularization

We introduce a nonstationary version of the iteration (5.7.1). We study the convergence and we prove that the new iteration is a regularization method.

Definition 5.8.1 (NSWIT-I). *Let $\{\alpha_n\}_{n\in\mathbb{N}},\{r_n\}_{n\in\mathbb{N}}\subset\mathbb{R}_{>0}$ be sequences of positive real numbers. We define a* nonstationary iterated weighted-I Tikhonov method *(NSIWT-I) as follows*

$$\begin{cases} x_{\alpha_0,r_0}^0:=0, \\ \left[(K^*K)^{\frac{r_n+1}{2}}+\alpha_n I\right]x_{\alpha_n,r_n}^n:=(K^*K)^{\frac{r_n-1}{2}}K^*y+\alpha_n x_{\alpha_{n-1},r_{n-1}}^{n-1}, \end{cases} \tag{5.8.1}$$

or equivalently

$$\begin{cases} x^0_{\alpha_0,r_0} := 0, \\ x^n_{\alpha_n,r_n} := \operatorname{argmin}_{x\in X} \left\{ \|Kx-y\|^2_{W_n} + \alpha_n \|x - x^{n-1}_{\alpha_{n-1},r_{n-1}}\|^2 \right\}, \end{cases} \tag{5.8.2}$$

where $\|\cdot\|_{W_n}$ is the semi-norm introduced by the operator $W_n := (KK^*)^{\frac{r_n-1}{2}}$ and depending on n, due to the non stationary character of r_n.

Definition 5.8.2 (NSWIT-II). *Let $\{\alpha_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$ and $\{j_n\}_{n\in\mathbb{N}} \subset \mathbb{N}$ be sequences of positive real numbers and integers, respectively. We define a* nonstationary iterated weighted-II Tikhonov method *(NSIWT-II) as follows*

$$\begin{cases} x^0_{\alpha_0,j_0} := 0, \\ \left[K^*K + \alpha_n \left(I - \frac{K^*K}{\|K^*K\|} \right)^{j_n} \right] x^n_{\alpha_n,j_n} := K^*y + \alpha_n \left(I - \frac{K^*K}{\|K^*K\|} \right)^{j_n} x^{n-1}_{\alpha_{n-1},j_{n-1}}, \end{cases} \tag{5.8.3}$$

or equivalently

$$\begin{cases} x^0_{\alpha_0,j_0} := 0, \\ x^n_{\alpha_n,j_n} := \operatorname{argmin}_{x\in X} \left\{ \|Kx-y\|^2 + \alpha_n \|x - x^{n-1}_{\alpha_{n-1},j_{n-1}}\|^2_{B_n} \right\}, \end{cases} \tag{5.8.4}$$

where $\|\cdot\|_{B_n}$ is the semi-norm introduced by the operator $B_n := \left(I - \frac{K^*K}{\|K^*K\|} \right)^{j_n}$ and depending on n, due to the non stationary character of j_n.

5.8.1 Convergence analysis

We are concerned about the properties of the sequence $\{\alpha_n\}$ such that the iteration (5.8.1) shall converge. To this aim we need some preliminary lemmas.

Remark 5.8.3. *Hereafter, without loss of generality, we will consider $\sigma_1 = 1$, namely $\|K\| = 1$.*

Lemma 5.8.4. *Let $\{t_n\}_{n\in\mathbb{N}}$ be a sequence of real numbers such that $0 \le t_n < 1$ for every n. Then*

$$\prod_{n=1}^{\infty}(1-t_n) > 0 \qquad \text{if and only if} \qquad \sum_{n=1}^{\infty} t_n < \infty. \tag{5.8.5}$$

Proof. See [92, Theorem 15.5] □

Lemma 5.8.5. *Let $\{t_k\}_{k\in\mathbb{N}}$ be a sequence of positive real numbers and let $N > 0$. Then*

$$\sum_{k=1}^{n} t_k \sim c \sum_{k=N}^{n} t_k \quad \text{for } n \to \infty.$$

with $c > 0$ (in particular, $c = 1$ when $\sum_{k=N}^{\infty} t_k = \sum_{k=1}^{\infty} t_k = \infty$).

Proof. Obviously, both the series converge or diverge simultaneously due to the Asymptotic Comparison test. If they converge, the thesis follows trivially. On the contrary, if they both diverge then we conclude by observing that $\sum_{k=N}^{n} t_k / \sum_{k=1}^{n} t_k$ is a monotonic increasing sequence bounded from above by 1. Indeed, if we set

$$A_n := \sum_{k=N}^{n} t_k, \qquad B_n := \sum_{k=1}^{n} t_k,$$

then $A_{n+1}/B_{n+1} \geq A_n/B_n$ for every n and it is easy to see that $\sup_n\{A_n/B_n\} = 1$. □

Lemma 5.8.6. *For every sequence* $\{t_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ *such that* $\lim_{k\to\infty} t_k = t \in (0,\infty]$*, we find*

$$\sum_{k=1}^{n} \frac{1}{t_k} \sim c \sum_{k=1}^{n} \frac{1}{1+t_k}, \quad c > 0.$$

Proof. If $\lim_{k\to\infty} t_k = t \in (0,\infty]$, then

$$\frac{1}{t_k} \sim \left(1 + \frac{1}{t}\right) \frac{1}{1+t_k}, \tag{5.8.6}$$

where $1/t = 0$ if $t = \infty$. Therefore, from the Asymptotic Comparison test for series, both series converge or diverge simultaneously. When they converge the thesis follows trivially. Assume then that the series diverge. If we set

$$X_n := \frac{\sum_{k=1}^{n} \frac{1}{t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}},$$

we want to show that the limit of X_n exists finite and, moreover, that is $\lim_{n\to\infty} X_n = 1 + 1/t$. Indeed, for any fixed $\varepsilon > 0$ there exists N_ε^1 such that for any $k \geq N_\varepsilon^1$ it holds that

$$\frac{1}{t_k} < \left(1 + \frac{1}{t} + \frac{\varepsilon}{2}\right) \frac{1}{1+t_k}, \tag{5.8.7}$$

and for any fixed ε and N_ε^1, there exists N_ε^2 such that for every $n \geq N_\varepsilon^2$ it holds that

$$\frac{\sum_{k=1}^{N_\varepsilon^1} \frac{1}{t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} < \frac{\varepsilon}{2}. \tag{5.8.8}$$

Hence, for any $n \geq \max\{N_\varepsilon^1, N_\varepsilon^2\}$, thanks to (5.8.7) and (5.8.8), we have that

$$X_n = \frac{\sum_{k=1}^{n} \frac{1}{t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} < \frac{\sum_{k=1}^{N_\varepsilon^1} \frac{1}{t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} + \left(1 + \frac{1}{t} + \frac{\varepsilon}{2}\right) \frac{\sum_{k=N_\varepsilon^1+1}^{n} \frac{1}{1+t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} < \frac{\varepsilon}{2} + 1 + \frac{1}{t} + \frac{\varepsilon}{2} = 1 + \frac{1}{t} + \varepsilon.$$

On the other hand, there exists N_ε^3 such that for every $k \geq N_\varepsilon^3$ it holds

$$\frac{1}{t_k} > \left(1 + \frac{1}{t} - \frac{\varepsilon}{2}\right) \frac{1}{1+t_k}, \tag{5.8.9}$$

and, by Lemma 5.8.5, for any fixed N_ε^3 and for any fixed $\delta < \frac{\varepsilon}{2}(1 + \frac{1}{t} - \frac{\varepsilon}{2})^{-1}$, there exists N_ε^4 such that for every $n \geq N_\varepsilon^4$ it holds

$$\frac{\sum_{k=N_\varepsilon^3+1}^{n} \frac{1}{1+t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} > (1-\delta). \tag{5.8.10}$$

Hence, fo any $n \geq \max\{N_\varepsilon^3, N_\varepsilon^4\}$, thanks to (5.8.9) and (5.8.10), we have that

$$X_n = \frac{\sum_{k=1}^{n} \frac{1}{t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} > \frac{\sum_{k=1}^{N_\varepsilon^3} \frac{1}{t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} + \left(1 + \frac{1}{t} - \frac{\varepsilon}{2}\right) \frac{\sum_{k=N_\varepsilon^3+1}^{n} \frac{1}{1+t_k}}{\sum_{k=1}^{n} \frac{1}{1+t_k}} > \left(1 + \frac{1}{t} - \frac{\varepsilon}{2}\right)(1-\delta) > 1 + \frac{1}{t} - \varepsilon.$$

Then, choosing $n \geq \max\{N_\varepsilon^i : i = 1,2,3,4\}$, the proof is concluded. □

We can now prove a necessary and sufficient condition on the sequence $\{\alpha_n\}$ for the convergence of NSIWT-I and NSWIT-II. That will be a consequence of the following, more general, theorem.

Theorem 5.8.7. *Let $f_n : [0, \sigma_1] \to \mathbb{R}$ be a sequence of bounded measurable functions such satisfy* (5.3.3a)*,* (5.3.3b) *and* (5.3.3c)*, and let us introduce the following nonstationary method,*

$$\begin{cases} x^0_{\alpha_0, f_0} := 0, \\ [K^*K + \alpha_n f_n(K^*K)]\, x^n_{\alpha_n, f_n} := K^*y + \alpha_n f_n(K^*K)\, x^{n-1}_{\alpha_{n-1}, f_{n-1}}, \end{cases} \tag{5.8.11}$$

or equivalently

$$\begin{cases} x^0_{\alpha_0, f_0} := 0, \\ x^n_{\alpha_n, f_n} := \operatorname{argmin}_{x \in X} \left\{ \|Kx - y\|^2 + \alpha_n \|x - x^{n-1}_{\alpha_{n-1}, f_{n-1}}\|^2_{f_n(K^*K)} \right\}, \end{cases} \tag{5.8.12}$$

For every $x^\dagger \in X$, the above method (5.8.11) *converges to $x^\dagger$ as $n \to \infty$ if and only if*

$$\sum_{k=1}^{n} \frac{\sigma^2}{\sigma^2 + \alpha_k f_k(\sigma^2)}$$

diverges for every $\sigma \in \sigma(K) \setminus \{0\}$.

Proof. Rewriting equation (5.8.1) and reminding that $y = Kx^\dagger$, we have

$$\begin{aligned} x^n_{\alpha_n, f_n} &= [K^*K + \alpha_n f_n(K^*K)]^{-1} K^*Kx^\dagger + \alpha_n [K^*K + \alpha_n f_n(K^*K)]^{-1} f_n(K^*K)\, x^{n-1}_{\alpha_{n-1}, f_{n-1}} \\ &= \left\{ I - \alpha_n [K^*K + \alpha_n f_n(K^*K)]^{-1} f_n(K^*K) \right\} x^\dagger \\ &\quad + \alpha_n [K^*K + \alpha_n f_n(K^*K)]^{-1} f_n(K^*K)\, x^{n-1}_{\alpha_{n-1}, f_{n-1}}, \end{aligned}$$

from which it follows that

$$\begin{aligned} x^\dagger - x^n_{\alpha_n,f_n} &= \alpha_n \left[K^*K + \alpha_n f_n\left(K^*K\right)\right]^{-1} f_n\left(K^*K\right)\left(x^\dagger - x^{n-1}_{\alpha_{n-1},f_{n-1}}\right) \\ &= (\cdots) \text{ iterating the process } n-1 \text{ times} \\ &= \prod_{k=1}^{n} \alpha_k \left[K^*K + \alpha_k f_k\left(K^*K\right)\right]^{-1} f_k\left(K^*K\right) x^\dagger \end{aligned} \tag{5.8.13}$$

since for convenience we put $x^0_{\alpha_0,f_0} := 0$. As a consequence, the method shall converge for every $x^\dagger$ if and only if

$$\lim_{n\to\infty} \left\| \prod_{k=1}^{n} \alpha_k \left[K^*K + \alpha_k f_k\left(K^*K\right)\right]^{-1} f_k\left(K^*K\right) x^\dagger \right\| = 0 \tag{5.8.14}$$

for every $x^\dagger \in X$, namely, if and only if

$$\lim_{n\to\infty} \int_{\sigma(K^*K)} \left| \prod_{k=1}^{n} \frac{\alpha_k f_k\left(\sigma^2\right)}{\sigma^2 + \alpha_k f_k\left(\sigma^2\right)} \right|^2 d\langle E_{\sigma^2} x^\dagger, x^\dagger \rangle = 0 \tag{5.8.15}$$

for every Borel-measure $\langle Ex^\dagger, x^\dagger\rangle$ induced by $x^\dagger \in X$. Since

$$\left| \prod_{k=1}^{n} \frac{\alpha_k f_k\left(\sigma^2\right)}{\sigma^2 + \alpha_k f_k\left(\sigma^2\right)} \right|^2 \leq 1$$

for every n, and since

$$\int_{\sigma(K^*K)} d\langle E_{\sigma^2} x^\dagger, x^\dagger\rangle = \|x^\dagger\|^2,$$

the Dominated Convergence Theorem [92, Theorem 1.34, pag. 26] implies the following equality

$$\begin{array}{c} \lim_{n\to\infty} \int_{\sigma(K^*K)} \left| \prod_{k=1}^{n} \frac{\alpha_k f_k\left(\sigma^2\right)}{\sigma^2 + \alpha_k f_k\left(\sigma^2\right)} \right|^2 d\langle E_{\sigma^2} x^\dagger, x^\dagger\rangle \\ \| \\ \int_{\sigma(K^*K)} \lim_{n\to\infty} \left| \prod_{k=1}^{n} \frac{\alpha_k f_k\left(\sigma^2\right)}{\sigma^2 + \alpha_k f_k\left(\sigma^2\right)} \right|^2 d\langle E_{\sigma^2} x^\dagger, x^\dagger\rangle. \end{array} \tag{5.8.16}$$

Hence, the method is convergent for every $x^\dagger \in X$ if and only if

$$\prod_{k=1}^{\infty} \frac{\alpha_k f_k\left(\sigma^2\right)}{\sigma^2 + \alpha_k f_k\left(\sigma^2\right)} = \prod_{k=1}^{\infty} \left(1 - \frac{\sigma^2}{\sigma^2 + \alpha_k f_k\left(\sigma^2\right)}\right) = 0, \tag{5.8.17}$$

for $\langle Ex^\dagger, x^\dagger\rangle$-a.e. σ^2 and every induced Borel measure $\langle Ex^\dagger, x^\dagger\rangle$, i.e., for every $\sigma \in \sigma(K) \setminus \{0\}$. Applying now Lemma 5.8.4 the thesis follows. □

Corollary 5.8.8. *For every $x^\dagger \in X$, the method* (5.8.1) *converges to $x^\dagger$ as $n \to \infty$ if and only if*

$$\sum_{k=1}^{n} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}$$

diverges for every $\sigma \in \sigma(K) \setminus \{0\}$.

Corollary 5.8.9. *For every $x^\dagger \in X$, the method* (5.8.3) *converges to $x^\dagger$ as $n \to \infty$ if and only if*

$$\sum_{k=1}^{n} \frac{\sigma^2}{\sigma^2+\alpha_k\left[1-\left(\frac{\sigma}{\sigma_1}\right)^2\right]^{j_k}}$$

diverges for every $\sigma \in \sigma(K) \setminus \{0\}$.

Corollary 5.8.10.

(1) If $\sup_{k\in\mathbb{N}}\{r_k\} = r \in [0,\infty)$, then the NSIWT-I method converges if and only if $\sum_{k=1}^{n} \alpha_k^{-1}$ diverges.

(2) Let $\lim_{k\to\infty} r_k = \infty$ monotonically and let us set $\beta_n = \sum_{k=1}^{n} \alpha_k^{-1}$. If $\lim_{n\to\infty} \beta_n^{1/r_n} = \infty$, then the NSIWT-I method converges.

Proof.
(1) For every $\sigma \in \sigma(K) \setminus \{0\}$, we observe that

$$\sum_{k=1}^{\infty} \frac{\sigma^{r+1}}{\sigma^{r+1}+\alpha_k} \le \sum_{k=1}^{\infty} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k} \le \sum_{k=1}^{\infty} \frac{1}{1+\alpha_k} \le \sum_{k=1}^{\infty} \frac{1}{\alpha_k}. \tag{5.8.18}$$

If the NSIWT-I method converges then, by Theorem 5.8.7 and by (5.8.18), $\sum_{k=1}^{\infty} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}$ diverges and hence $\sum_{k=1}^{\infty} \frac{1}{\alpha_k} = \infty$. On the other hand, if $\sum_{k=1}^{\infty} \alpha_k^{-1} = \infty$, then we can possibly have three different cases: $\lim_{k\to\infty} \alpha_k \in [0,\infty)$, $\nexists \lim_{k\to\infty} \alpha_k$ or $\lim_{k\to\infty} \alpha_k = \infty$. In the first two cases, $\frac{\sigma^{r+1}}{\sigma^{r+1}+\alpha_k} \nrightarrow 0$ for every $\sigma > 0$, and then the corresponding series diverges. In the latter case instead $\alpha_k^{-1} \sim c_{\sigma,r} \frac{\sigma^{r+1}}{\sigma^{r+1}+\alpha_k}$ for every $\sigma > 0$, and hence the series $\sum_{k=1}^{n} \alpha_k^{-1}$ and $\sum_{k=1}^{n} \frac{\sigma^{r+1}}{\sigma^{r+1}+\alpha_k}$ converges or diverge simultaneously by the Asymptotic Comparison test. Then, by $\sum_{k=1}^{\infty} \alpha_k^{-1} = \infty$, we deduce that $\sum_{k=1}^{\infty} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}$ diverges for every $\sigma > 0$ and the NSIWT-I method converges.

(2) Note that

$$\lim_{n\to\infty} \beta_n^{1/r_n} = \infty \qquad \Longleftrightarrow \qquad \lim_{n\to\infty} \sigma^{r_n} \left(\sum_{k=1}^{n} \alpha_k^{-1} \right) = \infty \quad \forall \sigma \in \sigma(K) \setminus \{0\},$$

namely,

$$\lim_{n\to\infty} \beta_n^{1/r_n} = \infty \qquad \Longleftrightarrow \qquad \left(\sum_{k=1}^{n} \alpha_k^{-1} \right)^{-1} = o(\sigma^{r_n}) \quad \forall \sigma \in \sigma(K) \setminus \{0\}. \tag{5.8.19}$$

We can assume that $0 < \sigma < 1$. For $\sigma = 1$ the result is indeed trivial owing to the equivalence

$$\sum_{k=1}^{\infty} \frac{1}{1+\alpha_k} = \infty \Longleftrightarrow \sum_{k=1}^{\infty} \alpha_k^{-1} = \infty \qquad \text{(see the previous point).}$$

Let us fix $\sigma \in (0,1)$ and for the sake of simplicity let suppose that $\{\alpha_k\}$ admits limit, i.e., $\lim_{k\to\infty} \alpha_k \in [0,\infty]$. We have two cases:

$$\lim_{k\to\infty} \frac{\alpha_k}{\sigma^{r_k+1}} = 0 \quad \text{or} \quad \lim_{k\to\infty} \frac{\alpha_k}{\sigma^{r_k+1}} \in (0,\infty].$$

In the first case, $\frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k} \nrightarrow 0$ for $k \to \infty$, then the corresponding series $\sum_{k=1}^{n} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}$ diverges. In this case we did not use (5.8.19), but note that

$$\sigma^{n+1}\alpha_n^{-1} \leq \sigma^{n+1} \sum_{k=1}^{n} \alpha_k^{-1}$$

and then, if $\lim_{k\to\infty} \alpha_k/\sigma^{r_k+1} = 0$, it holds $\left(\sum_{k=1}^{n} \alpha_k^{-1}\right)^{-1} = o(\sigma^{r_n+1})$. In the second case, we have $\frac{1}{\sigma^{r_k+1}+\alpha_k} \sim c\alpha_k^{-1}$, for $k \to \infty$. Therefore, there exists $N = N(\sigma)$ such that $\frac{1}{\sigma^{r_k+1}+\alpha_k} \geq \frac{c}{2}\alpha_k^{-1}$ for every $k \geq N$. Hence, fixed $n > N$, we have

$$\frac{c}{2}\sigma^{r_n+1} \sum_{k=N}^{n} \alpha_k^{-1} \leq \sigma^{r_n+1} \left(\sum_{k=1}^{N-1} \frac{1}{\sigma^{r_k+1}+\alpha_k} + \frac{c}{2} \sum_{k=N}^{n} \alpha_k^{-1} \right) \leq \sum_{k=1}^{n} \frac{\sigma^{r_n+1}}{\sigma^{r_k+1}+\alpha_k} \leq \sum_{k=1}^{n} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k},$$

where the last inequality stands in virtue of the monotonicity of $\{r_k\}$. Since, by Lemma 5.8.5, $\sum_{k=N}^{n} \alpha_k^{-1} \sim c \sum_{k=1}^{n} \alpha_k^{-1}$ then, by the preceding inequalities, the equivalence (5.8.19) implies that $\sum_{k=1}^{n} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k} = \infty$. Finally, due to the arbitrarily choice of σ, we can conclude that $\sum_{k=1}^{n} \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}$ diverges for every $\sigma \in \sigma(K) \setminus \{0\}$, and therefore the NSIWT-I method converges. If $\{\alpha_k\}$ does not have limit, then the proof can be carried out identically but handling with more care the different cases

$$\liminf_{k\to\infty} \frac{\alpha_k}{\sigma^{r_k+1}} = 0 \quad \text{or} \quad \liminf_{k\to\infty} \frac{\alpha_k}{\sigma^{r_k+1}} \in (0,\infty].$$

□

Corollary 5.8.10 applies immediately to the stationary case, where $\alpha_k = \alpha$ and $r_k = r$ for every $k \in \mathbb{N}$, showing that SIWT-I converges. On the other hand, from point *(2)* of Corollary 5.8.10, given a monotone divergent sequence $r_k \to \infty$ we need a sequence α_k such that $\left(\sum_{k=1}^{n} \alpha_k^{-1}\right)^{1/r_n} \to \infty$ for $n \to \infty$ in order to preserve the convergence of NSIWT-I.

Corollary 5.8.11. *If $\sum_{k=1}^{n} \alpha_k^{-1}$ diverges then the NSIWT-II method converges.*

Proof. It is just an easy adaptation of the proof of (1) in Corollary 5.8.10, when now we have to study the series

$$\sum_{k=1}^{\infty} \frac{\sigma^2}{\sigma^2 + \alpha_k \left[1 - \left(\frac{\sigma}{\sigma_1}\right)^2\right]^{j_k}}.$$

We leave the details. □

Now, we investigate the convergence rate of NSIWT-I and NSIWT-II.

Theorem 5.8.12. *Let $\{x^n_{\alpha_n,f_n}\}_{n\in\mathbb{N}}$ be a convergent sequence of the general method* (5.8.11)*, with $x^\dagger \in X_\nu$ for some $\nu > 0$, and let $\{\vartheta_n\}_{n\in\mathbb{N}}$ be a divergent sequence of positive real numbers. If*

$$\lim_{n\to\infty} \vartheta_n \sigma^\nu \prod_{k=1}^{n} \left(1 - \frac{\sigma^2}{\sigma^2 + \alpha_k f_k(\sigma^2)}\right) = 0 \qquad \text{for every } \sigma \in \sigma(K)\setminus\{0\}; \tag{5.8.20a}$$

$$\sup_{\sigma\in\sigma(K)\setminus\{0\}} \vartheta_n \sigma^\nu \prod_{k=1}^{n} \left(1 - \frac{\sigma^2}{\sigma^2 + \alpha_k f_k(\sigma^2)}\right) \le c < \infty \qquad \text{uniformly with respect to } n, \tag{5.8.20b}$$

then

$$\|x^\dagger - x^n_{\alpha_n,f_n}\| = o(\vartheta_n^{-1}). \tag{5.8.21}$$

Proof. From equation (5.8.13), for $x^\dagger \in X_\nu$, we have

$$\begin{aligned}\lim_{n\to\infty} \vartheta_n \|x^\dagger - x^n_{\alpha_n,f_n}\| &= \lim_{n\to\infty} \left[\int_{\sigma(K^*K)} \left|\vartheta_n \sigma^\nu \prod_{k=1}^{n} \left(1 - \frac{\sigma^2}{\sigma^2 + \alpha_k f_k(\sigma^2)}\right)\right|^2 d\langle E_{\sigma^2}\omega,\omega\rangle\right]^{1/2} \\ &= \left[\int_{\sigma(K^*K)} \left|\lim_{n\to\infty} \vartheta_n \sigma^\nu \prod_{k=1}^{n} \left(1 - \frac{\sigma^2}{\sigma^2 + \alpha_k f_k(\sigma^2)}\right)\right|^2 d\langle E_{\sigma^2}\omega,\omega\rangle\right]^{1/2},\end{aligned}$$

by (5.8.20b) and the Dominated Convergence Theorem. Now, from hypothesis (5.8.20a), the thesis follows. □

Contextualizing the above theorem to the cases of NSIWT-I and NSIWT-II, we have the next corollary.

Corollary 5.8.13. *If conditions* (5.8.20a) *and* (5.8.20b) *are satisfied for*

$$f_k(\sigma^2) = \sigma^{1-r_k} \qquad \text{and} \qquad f_k(\sigma^2) = \left[1 - \left(\frac{\sigma}{\sigma_1}\right)^2\right]^{j_k},$$

then we have, respectively,

$$\|x^\dagger - x^n_{\alpha_n,r_n}\| = o(\vartheta_n^{-1}) \qquad \text{and} \qquad \|x^\dagger - x^n_{\alpha_n,j_n}\| = o(\vartheta_n^{-1}).$$

The following Corollary 5.8.14 and Proposition 5.8.17 keep investigating deeper the convergence rate of the NSIWT-I method.

Corollary 5.8.14. *We define*

$$\beta_n = \sum_{k=1}^{n} \alpha_k^{-1}, \qquad \tilde{\beta}_n = \sum_{k=1}^{n} \frac{1}{1+\alpha_k}.$$

Let $\{r_k\}_{k\in\mathbb{N}}$ be a sequence of positive real numbers, and let $x^\dagger \in X_\nu$ for some $\nu > 0$. If

(i) $\sup_{k\in\mathbb{N}}\{r_k\} = r \in (0,\infty)$,

(ii) $\lim_{n\to\infty} \beta_n = \infty$,

then

$$\|x^\dagger - x^n_{\alpha_n,r_n}\| = \begin{cases} o(\beta_n^{-\frac{\nu}{r+1}}) & \text{if } \lim\limits_{n\to\infty} \alpha_n = \alpha \in (0,\infty] \qquad (5.8.22a) \\ O(\beta_n^{-\frac{\nu}{r+1}}) & \text{if } \lim\limits_{n\to\infty} \alpha_n = 0 \text{ and } \alpha_n^{-1} \le c\beta_{n-1},\ c>0 \qquad (5.8.22b) \\ o(\tilde{\beta}_n^{-\frac{\nu}{r+1}}) & \text{otherwise.} \qquad (5.8.22c) \end{cases}$$

Proof. For the sake of simplicity, let us assume that the sequences $\{\alpha_k\}$, $\{r_k\}$ admit limits. First, note that from (i), (ii) and Corollary 5.8.10 it follows that the NSIWT-I method is convergent. Now, since $1-x \le \mathrm{e}^{-x} \le c_{\nu,r}x^{-\nu/r+1}$, and using (i.2), we have

$$\begin{aligned} \sigma^\nu \prod_{k=1}^{n}\left(1 - \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}\right) &\le \sigma^\nu \mathrm{e}^{-\sum_{k=1}^n \frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}} \\ &\le \sigma^\nu \mathrm{e}^{-\sigma^{r+1}\sum_{k=1}^n \frac{1}{\sigma^{r+1}+\alpha_k}} \\ &\le c_{\nu,r}\sigma^\nu \left(\frac{1}{\sigma^{r+1}\sum_{k=1}^n \frac{1}{\sigma^{r+1}+\alpha_k}}\right)^{\frac{\nu}{r+1}} \\ &\le c_{\nu,r}\left(\sum_{k=1}^{n}\frac{1}{1+\alpha_k}\right)^{-\frac{\nu}{r+1}}. \end{aligned}$$

Moreover, note that $\frac{1}{1+\alpha_k} \sim \frac{c}{1+\alpha_k/\sigma^{r_k+1}}$. Therefore, conditions (5.8.20a) and (5.8.20b) of Theorem 5.8.12 are satisfied with

$$\vartheta_n = \left(\sum_{k=1}^{n}\frac{1}{1+\alpha_k}\right)^{\frac{\nu}{r+1}},$$

indeed

$$\sup_{\sigma\in[0,1]}\left\{\sigma^\nu\left(\sum_{k=1}^{n}\frac{1}{1+\alpha_k}\right)^{\frac{\nu}{r+1}}\prod_{k=1}^{n}\left(1-\frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}\right)\right\} \le c_{\nu,r},$$

and

$$\sigma^{\nu}\left(\sum_{k=1}^{n}\frac{1}{1+\alpha_k}\right)^{\frac{\nu}{r+1}}\prod_{k=1}^{n}\left(1-\frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}\right)\leq\left(\sum_{k=1}^{n}\frac{1}{1+\alpha_k}\right)^{\frac{\nu}{r+1}}\mathrm{e}^{-\sum_{k=1}^{n}\frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}}$$
$$=\left(\sum_{k=1}^{n}\frac{1}{1+\alpha_k}\right)^{\frac{\nu}{r+1}}\mathrm{e}^{-\sum_{k=1}^{n}\frac{1}{1+\alpha_k/\sigma^{r_k+1}}}$$
$$\leq c\left(\sum_{k=N(\sigma)}^{n}\frac{1}{1+\alpha_k/\sigma^{r_k+1}}\right)^{\frac{\nu}{r+1}}\mathrm{e}^{-\sum_{k=N(\sigma)}^{n}\frac{1}{1+\alpha_k/\sigma^{r_k+1}}},$$

where $N(\sigma)$ is chosen such that $\frac{1}{1+\alpha_k}\leq\frac{c/2}{1+\alpha_k/\sigma^{r_k+1}}$ for every $k\geq N(\sigma)$, and the right hand side of the last inequality tends to 0 as $n\to\infty$ for every fixed σ. If $\lim_{k\to\infty}\alpha_k=\alpha\in(0,\infty]$, then $\beta_n\sim c\sum_{k=1}^{n}\frac{1}{1+\alpha_k}$ for $n\to\infty$ by Lemma 5.8.6. Equations (5.8.22a) and (5.8.22c) follow. Eventually, observing that $1-\frac{\sigma^{r_k+1}}{\sigma^{r_k+1}+\alpha_k}\leq 1-\frac{\sigma^{r+1}}{\sigma^{r+1}+\alpha_k}$, equation (5.8.22b) follows instead by a straightforward application of [Lemma 1,2,3 and Theorem 1][58].

In the general case where no assumptions are made on the existence of the limits for the sequences $\{\alpha_k\}$ and $\{r_k\}$, we can apply the same arguments being careful to study the lim inf and lim sup of these sequences. □

When $r=1$ (classical iterated Tikhonov), equation (5.8.22b) is the result in [58, Theorem 1]. On the other hand, if $\lim_{n\to\infty}\alpha_n=\alpha\in(0,\infty]$, then the convergence rate is improved by the small "o".

Remark 5.8.15. *As we stated in* (5.8.22b)*, when* $\lim_{n\to\infty}\alpha_n=0$*, to obtain a convergence rate of order* $O(\beta_n^{-\nu/(r+1)})$ *the sequence* $\{\alpha_n\}$ *has to satisfy the condition* $\alpha_n^{-1}\leq c\beta_{n-1}$ *for a positive real number* $c>0$*. Then,* $\sum_{k=1}^{n}\alpha_k^{-1}=\beta_n=O(q^n)$*, where* $q=(1+c)>1$*. To overcome this bound, choosing sequences* $\{\hat{r}_n\}$ *and* $\{\hat{\alpha}_n\}$ *such that* $\hat{r}_n$ *diverges monotonically and* $\beta_n^{1/r_n}\to\infty$ *as* $n\to\infty$*, we are able to obtain a faster convergence rate, in a sense that has still to be defined. In the following Proposition 5.8.17 we will give the proof for a specific case.*

Definition 5.8.16. *Following the approach in* [17, (2.3), (2.4) pag. 26]*, we say that the sequence* $\{\hat{x}_n\}$ converges uniformly faster *than the sequence* $\{x_n\}$ *if*

$$x^{\dagger}-\hat{x}_n=R_n(x^{\dagger}-x_n),\tag{5.8.23}$$

where $\{R_n\}$ *is a sequence of operators such that* $\|R_n\|\to 0$ *as* $n\to\infty$*.*
We say instead that $\{\hat{x}_n\}$ converges non-uniformly faster *than* $\{x_n\}$ *if* (5.8.23) *holds and*

$$\inf_{n\in\mathbb{N}}\|R_n\|>0,\qquad\lim_{n\to\infty}\|R_nx\|=0\ \textit{for every}\ x\in X.$$

We are ready to state the following comparison result.

Proposition 5.8.17. *Let $\{x^n_{\alpha_n}\}$ be the sequence generated by the nonstationary iterated Tikhonov with $\alpha_n = \alpha_0 q^n$, where $\alpha_0 \in (0,\infty), q \in (0,1)$, and let $\{x^n_{\hat{\alpha}_n,\hat{r}_n}\}$ be the sequence generated by NSIWT-I, where $\hat{\alpha}_n = 1/n!$ and $\hat{r}_n = n$, both applied to the same compact operator $K : X \to Y$. Then, $\{x^n_{\hat{\alpha}_n,\hat{r}_n}\}$ converges, non uniformly, faster than $\{x^n_{\alpha_n}\}$.*

Proof. Observe that the sequence $\{x^n_{\alpha_n}\}$ corresponds to a NSIWT-I method $\{x^n_{\alpha_n,r_n}\}$ with $r_n = 1$ for every n. Moreover, both the sequences $\{x^n_{\alpha_n}\}$ and $\{x^n_{\hat{\alpha}_n,\hat{r}_n}\}$ converge, indeed they satisfy conditions (1) and (2) of Corollary 5.8.10, respectively. Assuming that $x_0 = 0$ and applying the same strategy used in Theorem 5.8.7, without any effort it is possible to show that

$$x^\dagger - x^n_{\hat{\alpha}_n,\hat{r}_n} = \prod_{k=1}^{n} \hat{\alpha}_k \left((K^*K)^{\frac{\hat{r}_k+1}{2}} + \hat{\alpha}_k I \right)^{-1} x^\dagger,$$
$$x^\dagger = \prod_{k=1}^{n} \alpha_k^{-1} \left(K^*K + \alpha_k I\right)(x^\dagger - x^n_{\alpha_n}).$$

Therefore we find

$$x^\dagger - x^n_{\hat{\alpha}_n,\hat{r}_n} = \left[\prod_{k=1}^{n} \hat{\alpha}_k \alpha_k^{-1} \left((K^*K)^{\frac{\hat{r}_k+1}{2}} + \hat{\alpha}_k I \right)^{-1} (K^*K + \alpha_k I) \right] (x^\dagger - x^n_{\alpha_n}) = R_n (x^\dagger - x^n_{\alpha_n}).$$

Since $0 \in \sigma(K^*K)$, we infer $\|R_n\| > 1$ for every n, and hence $\inf_{n\in\mathbb{N}} \|R_n\| \geq 1$. If we prove that

$$\lim_{n\to\infty} \|R_n x\| = 0,$$

for every $x \in X$, then the thesis follows. Since

$$\lim_{n\to\infty} \|R_n x\| = 0 \Longleftrightarrow \lim_{n\to\infty} \prod_{k=1}^{n} \frac{\hat{\alpha}_k(\sigma^2 + \alpha_k)}{\alpha_k(\sigma^{\hat{r}_k+1} + \hat{\alpha}_k)} = 0 \Longleftrightarrow \sum_{k=1}^{\infty} \frac{\alpha_k \sigma^{\hat{r}_k+1} - \hat{\alpha}_k \sigma^2}{\alpha_k \sigma^{\hat{r}_k+1} + \alpha_k \hat{\alpha}_k} = \infty \ \ \forall \sigma > 0,$$

if we substitute the values $\alpha_n = \alpha_0 q^n$, then $\hat{\alpha}_n = 1/n!$ and $\hat{r}_n = n$, we obtain

$$\sum_{k=1}^{\infty} \frac{\alpha_k \sigma^{\hat{r}_k+1} - \hat{\alpha}_k \sigma^2}{\alpha_k \sigma^{\hat{r}_k+1} + \alpha_k \hat{\alpha}_k} = \sum_{k=1}^{\infty} \frac{1 - \frac{\sigma}{\alpha_0 n! (q\sigma)^n}}{1 + \frac{1/n!}{\sigma^{n+1}}},$$

and the right hand side of the above equality diverges: indeed

$$\frac{1 - \frac{\sigma}{\alpha_0 n! (q\sigma)^n}}{1 + \frac{1/n!}{\sigma^{n+1}}} \longrightarrow 1 \ \text{ for every fixed } q, \sigma \in (0,1) \text{ and } \alpha_0 \in (0,\infty).$$

□

5.8.2 Analysis of convergence for perturbed data

Let now consider $y^\delta = y + \delta\eta$, with $y \in \mathrm{Rg}(K)$ and $\|\eta\| = 1$, i.e., $\|y^\delta - y\| = \delta$. We are concerned about the convergence of the NSIWT-I/II methods when the initial datum y is perturbed. Again, we will initially prove a more general statement involving the method (5.8.11), with initial datum y^δ, and then the convergence results for perturbed data in the NSIWT-I/II cases will follow as corollary. We use the notation $x^{n,\delta}_{\alpha_n,f_n}$ for the solution of the method (5.8.11) .

Theorem 5.8.18. *Under the assumptions of Theorem 5.8.7, let $f_j : [0,\sigma_1] \to \mathbb{R}$ be such that*

$$\max_{\sigma\in[0,1]} \left| \prod_{j=1}^{n} \frac{\alpha_j \sigma}{\sigma^2 + \alpha_j f_j(\sigma^2)} \right| \leq c, \qquad \text{for every } n. \tag{5.8.24}$$

If $\{\delta_n\}$ is a sequence convergent to 0 with $\delta_n \geq 0$ and such that

$$\lim_{n\to\infty} \delta_n \cdot \sum_{k=1}^{n} \alpha_k^{-1} = 0, \tag{5.8.25}$$

then,

$$\lim_{n\to\infty} \|x^\dagger - x^{n,\delta_n}_{\alpha_n,f_n}\| = 0.$$

Proof. From the definition of the method (5.8.11), for every given j,n, we find that

$$\begin{aligned} x^{j,\delta_n}_{\alpha_j,f_j} &= \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} \left(K^* y^{\delta_n} + \alpha_j f_j(K^*K)\, x^{j-1,\delta_n}_{\alpha_{j-1},f_{j-1}}\right) \\ &= \left\{ I - \alpha_j \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} f_j(K^*K) \right\} x^\dagger \\ &\quad + \alpha_j \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} f_j(K^*K)\, x^{j-1,\delta_n}_{\alpha_{j-1},f_{j-1}} \\ &\quad + \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} K^*(y^{\delta_n} - y), \end{aligned}$$

namely,

$$\begin{aligned} x^\dagger - x^{j,\delta_n}_{\alpha_j,f_j} &= \alpha_j \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} f_j(K^*K)\,(x^\dagger - x^{j-1,\delta_n}_{\alpha_{j-1},f_{j-1}}) \\ &\quad - \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} K^*(y^{\delta_n} - y). \end{aligned}$$

Hence, by induction, for every fixed n we have

$$\begin{aligned} x^\dagger - x^{n,\delta_n}_{\alpha_n,f_n} &= \prod_{k=1}^{n} \alpha_k \left[K^*K + \alpha_k f_k(K^*K)\right]^{-1} f_k(K^*K)\, x^\dagger \\ &\quad - \sum_{k=1}^{n} \alpha_k^{-1} \prod_{i=k}^{n} \alpha_i \left[K^*K + \alpha_i f_i(K^*K)\right]^{-1} K^*(y^{\delta_n} - y). \end{aligned}$$

If we set $g_{k,n}(K^*K) = \prod_{i=k}^{n} \alpha_i \left[K^*K + \alpha_i f_i(K^*K)\right]^{-1}$, then we have

$$\begin{aligned}\|g_{k,n}(K^*K)K^*y\|^2 &= \langle g_{k,n}(K^*K)K^*y, g_{k,n}(K^*K)K^*y\rangle \\ &= \langle g_{k,n}(KK^*)KK^*y, g_{k,n}(KK^*)y\rangle \\ &= \langle g_{k,n}(KK^*)(KK^*)^{1/2}y, g_{k,n}(K^*K)(KK^*)^{1/2}y\rangle \\ &= \|g_{k,n}(KK^*)(KK^*)^{1/2}y\|^2,\end{aligned}$$

where we used the fact that $g_{k,n}(K^*K)K^* = K^*g_{k,n}(KK^*)$ and Proposition 5.1.23. Therefore,

$$\begin{aligned}\left\|\prod_{j=k}^{n} \alpha_j \left[K^*K + \alpha_j f_j(K^*K)\right]^{-1} K^*\right\| &= \left\|\prod_{j=k}^{n} \alpha_j \left[KK^* + \alpha_j f_j(KK^*)\right]^{-1} (KK^*)^{\frac{1}{2}}\right\| \\ &= \max_{\sigma\in[0,1]} \left|\prod_{j=k}^{n} \frac{\alpha_j \sigma}{\sigma^2 + \alpha_j f_j(\sigma^2)}\right| \leq c,\end{aligned}$$

by hypothesis (5.8.24). It follows that

$$\begin{aligned}\|x^\dagger - x^{n,\delta_n}_{\alpha_n,f_n}\| &\leq \|\prod_{k=1}^{n} \alpha_k \left[K^*K + \alpha_k f_k(K^*K)\right]^{-1} f_k(K^*K)x^\dagger\| + c\sum_{k=1}^{n} \alpha_k^{-1}\|y^{\delta_n} - y\| \\ &= \|x^\dagger - x^n_{\alpha_n,f_n}\| + c\delta_n \sum_{k=1}^{n} \alpha_k^{-1},\end{aligned}$$

and by Corollary 5.8.14 and (5.8.25), $\|x^\dagger - x^{n,\delta_n}_{\alpha_n,r_n}\| \to 0$ for $n \to \infty$. □

Corollary 5.8.19. *Under the assumptions of Corollary 5.8.10, if $\{\delta_n\}$ is a sequence convergent to 0 with $\delta_n \geq 0$ and such that*

$$\lim_{n\to\infty} \delta_n \cdot \sum_{k=1}^{n} \alpha_k^{-1} = 0,$$

then the NSIWT-I method with perturbed data is convergent,

$$\lim_{n\to\infty} \|x^\dagger - x^{n,\delta_n}_{\alpha_n,r_n}\| = 0.$$

Proof. We apply Theorem 5.8.18 with

$$f_j(\sigma^2) = \sigma^{1-r_j}.$$

Then hypothesis (5.8.24) is satisfied and the thesis follows at once. □

Corollary 5.8.20. *Under the assumptions of Corollary 5.8.11, let $0 \leq \alpha_n \leq 1$ for every n. If $\{\delta_n\}$ is a sequence convergent to 0 with $\delta_n \geq 0$ and such that*

$$\lim_{n\to\infty} \delta_n \cdot \sum_{k=1}^{n} \alpha_k^{-1} = 0,$$

then the NSIWT-II method with perturbed data is convergent,

$$\lim_{n\to\infty} \|x^\dagger - x^{n,\delta_n}_{\alpha_n,j_n}\| = 0.$$

Proof. Even in this case, we apply Theorem 5.8.18 with

$$f_l(\sigma^2) = \left(1-\sigma^2\right)^{j_l}.$$

From the proof of Proposition 5.4.4 we know that

$$\sup_{0\le\sigma\le 1}\left|\frac{\alpha\sigma}{\sigma^2+\alpha\left(1-\sigma^2\right)^j}\right| \le c\sqrt{\alpha}.$$

Henceforth, if $0 \le \alpha_n \le 1$ for every n, then hypothesis (5.8.24) is satisfied and again the thesis follows at once. □

5.9 Nonstationary iterated fractional Tikhonov

Definition 5.9.1 (Nonstationary iterated fractional Tikhonov)**.** *Let $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{\gamma_n\}_{n\in\mathbb{N}}$ be sequences of real numbers such that $\alpha_n > 0$ and $\gamma_n \ge 1/2$ for every n. We define the* nonstationary iterated fractional Tikhonov method (NSIFT) *as*

$$\begin{cases} x^0_{\alpha_0,\gamma_0} := 0; \\ \left(K^*K+\alpha_n I\right)^{\gamma_n} x^n_{\alpha_n,\gamma_n} := \left(K^*K\right)^{\gamma_n-1}K^*y + \left[\left(K^*K+\alpha_n I\right)^{\gamma_n} - \left(K^*K\right)^{\gamma_n}\right] x^{n-1}_{\alpha_{n-1},\gamma_{n-1}}. \end{cases} \tag{5.9.1}$$

We denote by $x^{n,\delta}_{\alpha_n,\gamma_n}$ the n-th iteration of NSIFT if $y = y^\delta$.

Theorem 5.9.2. *For every $x^\dagger \in X$, the* NSIFT *method* (5.9.1) *converges to $x^\dagger \in X$ as $n \to \infty$ if and only if $\sum_n \left(\frac{\sigma^2}{\sigma^2+\alpha_n}\right)^{\gamma_n}$ diverges for every $\sigma \in \sigma(K)\setminus\{0\}$.*

Proof. The proof follows the same steps as in Theorem 5.8.7. Therefore we will omit details. What follows is that

$$x^\dagger - x^n_{\alpha_n,\gamma_n} = \prod_{k=1}^n \left(K^*K+\alpha_k I\right)^{-\gamma_k}\left[\left(K^*K+\alpha_k I\right)^{\gamma_k} - \left(K^*K\right)^{\gamma_k}\right] x^\dagger,$$

and hence

$$\|x^\dagger - x^n_{\alpha_n,\gamma_n}\|^2 = \int_{\sigma(K^*K)} \left|\prod_{k=1}^n \frac{(\sigma^2+\alpha_k)^{\gamma_k} - \sigma^{2\gamma_k}}{(\sigma^2+\alpha_k)^{\gamma_k}}\right|^2 d\langle E_{\sigma^2}x^\dagger, x^\dagger\rangle.$$

Then, the method converges for every $x^\dagger \in X$ if and only if

$$\lim_{n\to\infty}\prod_{k=1}^n\left[1-\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{\gamma_k}\right] = 0$$

for every $\sigma \in \sigma(K)\setminus\{0\}$. The thesis follows by Lemma 5.8.4. □

Corollary 5.9.3.

(1) Let $\lim_{k\to\infty}\gamma_k=\gamma\in[1/2,\infty)$. *Then the NSIFT method converges if and only if*

$$\sum_{k=1}^{n}\alpha_k^{-\gamma}=\infty.$$

More in general, if $\sup_{k\in\mathbb{N}}\{\gamma_k\}=s\in[1/2,\infty)$ *and* $\sum_{k=1}^{\infty}\alpha_k^{-s}=\infty$, *then the NSIFT method converges.*

(2) Let $\lim_{k\to\infty}\gamma_k=\infty$. *If* $\lim_{k\to\infty}\alpha_k=0$ *and* $\lim_{k\to\infty}\alpha_k\gamma_k=l\in[0,\infty)$, *then the NSIFT method converges.*

Proof. (1) It is immediate noticing that

$$\sum_{k=1}^{n}\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{\gamma_k}\sim c\sum_{k=1}^{n}\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{\gamma}$$
$$\sum_{k=1}^{n}\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{\gamma_k}\geq\sum_{k=1}^{n}\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{s}.$$

(2) We observe that

$$\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{\gamma_k}=\left(1-\frac{\alpha_k}{\sigma^2+\alpha_k}\right)^{\gamma_k}\sim \mathrm{e}^{-\frac{\alpha_k\gamma_k}{\sigma^2+\alpha_k}}\to \mathrm{e}^{-l/\sigma^2}\neq 0$$

for $k\to\infty$. Then $\sum_{k=1}^{n}\left(\frac{\sigma^2}{\sigma^2+\alpha_k}\right)^{\gamma_k}$ diverges for every $\sigma>0$ and the NSIFT method converges. □

Theorem 5.9.4. *Let* $\{x^n_{\alpha_n,\gamma_n}\}_{n\in\mathbb{N}}$ *be a convergent sequence of the NSIFT method, with* $x^\dagger\in X_\nu$ *for some* $\nu>0$, *and let* $\{\vartheta_n\}_{n\in\mathbb{N}}$ *be a divergent sequence of positive real numbers. If*

$$\lim_{n\to\infty}\vartheta_n\sigma^\nu\prod_{k=1}^{n}\left(1-\frac{\sigma^{2\gamma_k}}{(\sigma^2+\alpha_k)^{\gamma_k}}\right)=0\qquad\text{for every }\sigma\in\sigma(K)\setminus\{0\};\tag{5.9.2a}$$

$$\sup_{\sigma\in\sigma(K)\setminus\{0\}}\vartheta_n\sigma^\nu\prod_{k=1}^{n}\left(1-\frac{\sigma^{2\gamma_k}}{(\sigma^2+\alpha_k)^{\gamma_k}}\right)\leq c<\infty\qquad\text{uniformly with respect to } n,\tag{5.9.2b}$$

then

$$\|x^\dagger-x^n_{\alpha_n,\gamma_n}\|=o(\vartheta_n^{-1}).\tag{5.9.3}$$

Proof. As seen in Theorem 5.8.12, the thesis follows easily from the Dominated Convergence Theorem. □

Corollary 5.9.5. *Let* $\{\gamma_k\}_{k\in\mathbb{N}}$ *be a sequence of positive real numbers,* $\gamma_k\geq 1/2$, *and let* $x^\dagger\in X_\nu$ *for some* $\nu>0$. *If*

(i) $\sup_{k\in\mathbb{N}}\{\gamma_k\} = s \in [1/2,\infty)$,

(ii) $\lim_{n\to\infty}\beta_n = \infty$,

then

$$\|x^\dagger - x^n_{\alpha_n,\gamma_n}\| = o(\beta_n^{-\frac{\nu}{2s}}) \qquad \text{if } \exists \lim_{k\to\infty}\alpha_k = \alpha \in (0,\infty], \tag{5.9.4}$$

$$\|x^\dagger - x^n_{\alpha_n,\gamma_n}\| = o(\tilde{\beta}_n^{-\frac{\nu}{2s}}) \qquad \text{otherwise}, \tag{5.9.5}$$

where we defined

$$\beta_n = \sum_{k=1}^n \alpha_k^{-s}, \qquad \tilde{\beta}_n = \sum_{k=1}^n \frac{1}{1+\alpha_k^s}.$$

Proof. See Corollary 5.8.14. □

Theorem 5.9.6. *Under the assumptions of Corollary 5.9.3, if* $\{\delta_n\}$ *is a sequence convergent to* 0 *with* $\delta_n \geq 0$ *and such that*

$$\lim_{n\to\infty}\delta_n \cdot \sum_{k=1}^n \alpha_k^{-\gamma_k} = 0, \tag{5.9.6}$$

then, $\lim_{n\to\infty}\|x^\dagger - x^{n,\delta_n}_{\alpha_n,\gamma_n}\| = 0$.

Proof. Here is a sketch of the proof, since it follows step by step from the proof of Theorem 5.8.18. If we set

$$\psi_k(K^*K) := [(K^*K+\alpha_k I)^{\gamma_k} - (K^*K)^{\gamma_k}]$$
$$\phi_k(K^*K) := \psi_k(K^*K)\,[K^*K+\alpha_k I]^{-\gamma_k},$$

then from (5.9.1) it is possible to show that

$$x^\dagger - x^{n,\delta_n}_{\alpha_n,\gamma_n} = \prod_{k=1}^n \phi_k(K^*K)x^\dagger - \sum_{k=1}^n \psi_k(K^*K)^{-1}\prod_{i=k}^n \phi_i(K^*K)\,(K^*K)^{\gamma_k-1}K^*(y^{\delta_n}-y),$$

for every integer n and for every perturbed data $y^{\delta_n} = y+\delta_n\eta$. Owing to the equality

$$\left\|\prod_{i=k}^n \phi_i(K^*K)\,(K^*K)^{\gamma_k-1}K^*\right\| = \left\|\prod_{i=k}^n \phi_i(KK^*)\,(KK^*)^{\gamma_k-1}\,(KK^*)^{1/2}\right\|,$$

we deduce

$$\begin{aligned}\|x^\dagger - x^{n,\delta_n}_{\alpha_n,\gamma_n}\| &\leq \|x^\dagger - x^n_{\alpha_n,\gamma_n}\| + \delta_n\sum_{k=1}^n \left\|\psi_k(K^*K)^{-1}\right\| \\ &= \|x^\dagger - x^n_{\alpha_n,\gamma_n}\| + \delta_n\sum_{k=1}^n \alpha_k^{-\gamma_k}.\end{aligned}$$

□

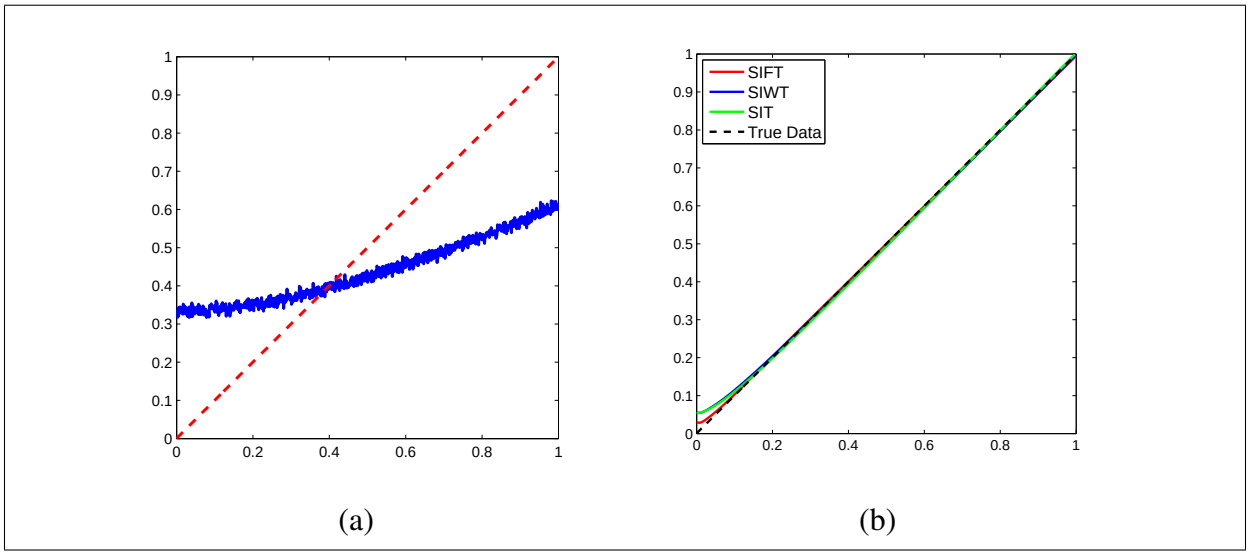


Figure 5.2: Example 1 – "Foxgood" test case: (a) the true solution (dashed curve) and the observed data (solid curve), (b) approximated solutions by SIFT with $\gamma = 0.8$ and $\alpha = 10^{-3}$, SIWT with $r = 0.6$ and $\alpha = 10^{-2}$, and SIWT with $r = 1$ and $\alpha = 10^{-3}$.

5.10 Numerical results

We now give few selected examples with a special focus on the nonstationary iterations proposed in this paper. For a larger comparison between fractional and classical Tikhonov refer to [73, 66, 51]. To produce our results we used Matlab 8.1.0.604 using a laptop pc with processor Intel iCore i5-3337U with 6 GB of RAM running Windows 8.1.

We add to the noise-free right-hand side vector y, the "noise-vector" e that has in all examples normally distributed pseudorandom entries with mean zero, and is normalized to correspond to a chosen noise-level

$$\xi = \frac{\|e\|}{\|y\|}.$$

As a stopping criterion for the methods we used the Discrepancy Principle [59], that terminates the iterative method at the iteration

$$\hat{k} = \min_k \{k : \|y^\delta - Kx_k\| \leq \tau\delta\},$$

where $\tau = 1.01$. This criterion stops the iterations when the norm of the residual reaches the norm of the noise so that the latter is not reconstructed.

To compare the restorations with the different methods, we consider both the visual representation and the relative restoration error that is $\|\hat{x} - x^\dagger\|/\|x^\dagger\|$ for the computed approximation $\hat{x}$. We will focus only on the weighted-I and the fractional Tikhonov methods.

α	Method	r/γ				
		0.4	0.6	0.8	1	1.2
5×10^{-2}	SIFT	337.09(7)	0.02498(13)	0.03481(19)	0.03752(29)	0.03838(43)
	SIWT	0.02589(9)	0.03202(13)	0.03609(19)	0.03752(29)	0.03932(43)
10^{-2}	SIFT	320.85(3)	0.02048(5)	0.02633(7)	0.03731(7)	0.03783(9)
	SIWT	0.01697(3)	0.01818(5)	0.03361(5)	0.03731(7)	0.03672(11)
5×10^{-3}	SIFT	423.37(3)	0.02216(3)	0.02190(5)	0.03102(5)	0.03723(5)
	SIWT	0.02421(3)	0.01573(3)	0.03186(3)	0.03103(5)	0.03347(7)
10^{-3}	SIFT	402.97(1)	0.02299(1)	**0.00698(3)**	0.01756(3)	0.02443(3)
	SIWT	0.06403(1)	0.02210(1)	0.02528(1)	0.01756(3)	0.02736(3)
5×10^{-4}	SIFT	531.72(1)	0.02119(1)	0.01729(1)	0.02507(1)	0.03119(1)
	SIWT	0.10518(1)	0.04506(1)	0.01482(1)	0.02507(1)	0.02086(3)
10^{-4}	SIFT	1012.2(1)	0.07246(1)	0.04229(1)	0.02704(1)	0.01675(1)
	SIWT	0.25927(1)	0.13000(1)	0.07213(1)	0.02704(1)	0.01154(1)

Table 5.1: Example 1: relative errors and iteration numbers between brackets for SIWT and SIFT for different choices of α, r, and γ.

5.10.1 Example 1

This test case is the so-called *Foxgood* in the toolbox REGULARIZATION TOOL by P. Hansen [62] using 1024 points. We have added a noise vector with $\xi = 0.02$ to the observed signal. In Figure 5.2(a) the true signal and the measured data can be seen.

In Table 5.1 we show the relative errors with different choices of α, r and γ. In brackets we report the iteration at which the discrepancy principle stopped the method. Note that SIFT with $\gamma = 1$ and SIWT with $r = 1$ are exactly the classical Tikhonov method and hence produce the same result. Figure 5.2(b) shows the reconstruction for SIFT with $\gamma = 0.8$ and $\alpha = 10^{-3}$, SIWT with $r = 0.6$ and $\alpha = 10^{-2}$, and SIWT with $r = 1$ (classical Iterated Tikhonov) with $\alpha = 10^{-3}$.

From these results, using both fractional and weighted-I iterated Tikhonov, we can see that we can obtain better restorations than with the classical version. However, in order to obtain such results, one has to evaluate α very carefully. Indeed α does not only affects the convergence speed, but also the quality of the restoration: a small perturbation in α can lead to quite different restoration errors. The nonstationary version of the methods can help also to avoid such a careful and often difficult estimation.

For the nonstationary iterations we assume the regularization parameter α_n at each iteration be given according to the geometric sequence

$$\alpha_n = \alpha_0 q^n, \qquad q \in (0,1), \qquad n = 1,2,\ldots. \tag{5.10.1}$$

Setting $r_n = 0.6$ and $\gamma_n = 0.8$, Table 5.2 shows that NSIFT and NSIWT-I provide a relative error lower than the classical nonstationary iterated Tikhonov (NSIT). Finally, since NSIFT and NSIWT-I allow a nonstationary choice also for r_n and γ_n, in Table 5.2 we report the results for

α_0	Method	q		
		0.7	0.8	0.9
10^{-1}	NSIFT ($\gamma_n = 0.8$)	0.024453(9)	0.030868(11)	0.028849(17)
	NSIWT-I ($r_n = 0.6$)	0.025223(7)	0.027628(9)	0.028534(13)
	NSIT	0.035162(9)	0.031627(13)	0.036472(19)
	NSIFT (γ_n in (5.10.2))	0.032489(9)	0.027974(13)	0.037199(17)
	NSIWT-I (r_n in (5.10.2))	0.031493(9)	0.027436(13)	0.036059(17)
10^{-2}	NSIFT ($\gamma_n = 0.8$)	**0.014781(5)**	0.021687(5)	0.028709(5)
	NSIWT-I ($r_n = 0.6$)	**0.014503(3)**	0.021501(3)	0.028396(3)
	NSIT	0.024838(5)	0.030866(5)	0.028835(7)
	NSIFT (γ_n in (5.10.2))	0.023848(5)	0.030002(5)	0.027636(7)
	NSIWT-I (r_n in (5.10.2))	0.023482(5)	0.029638(5)	0.027366(7)

Table 5.2: Example 1: relative errors and iteration numbers between brackets for NSIWT-I and NSIFT with the nonstationary α_n in (5.10.1) and different choices of r_n and γ_n (NSIT is $r_n = \gamma_n = 1$).

the following nonincreasing sequences

$$r_n = \gamma_n = \begin{cases} 1 - \frac{n-1}{100} & n < 50, \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{5.10.2}$$

Again both NSIWT-I and NSIFT are able to get better results than NSIT. Even tough the errors are not as good as those for the best choices $r_n = 0.6$ and $\gamma_n = 0.8$, the choice (5.10.2) stresses the robustness of our nonstationary iterations.

5.10.2 Example 2

We consider the test problem *deriv2(·,3)* in the toolbox REGULARIZATION TOOL by P. Hansen [62] using 1024 points. For the noise vector it holds $\xi = 0.05$. In Figure 5.3(a) we can see the measured data and the true signal. We compare NSIWT-I and NSIFT with the NSIT.

Firstly, α_n is defined by the classical choice in (5.10.1). Table 5.3 shows the results for different choices of r_n and γ_n. Note that NSIWT-I and NSIFT usually outperform NSIT. Nevertheless, our nonstationary iterations allow also unbounded sequences of r_n and γ_n. Therefore, according to Proposition 5.8.17, we set

$$\alpha_n = \frac{1}{n!}, \qquad r_n = \frac{n}{10}, \qquad \gamma_n = \frac{n}{2}. \tag{5.10.3}$$

Table 5.4 shows that the relative restoration error obtained with the unbounded sequences r_n and γ_n in (5.10.3) is lower than the best one (according to Table 5.3), obtained by NSIT by employing the geometric sequence (5.10.1) for α_n. The computed approximations are also compared in Figure 5.3(b), where we note a better restoration of the corner for NSIWT-I and NSIFT.

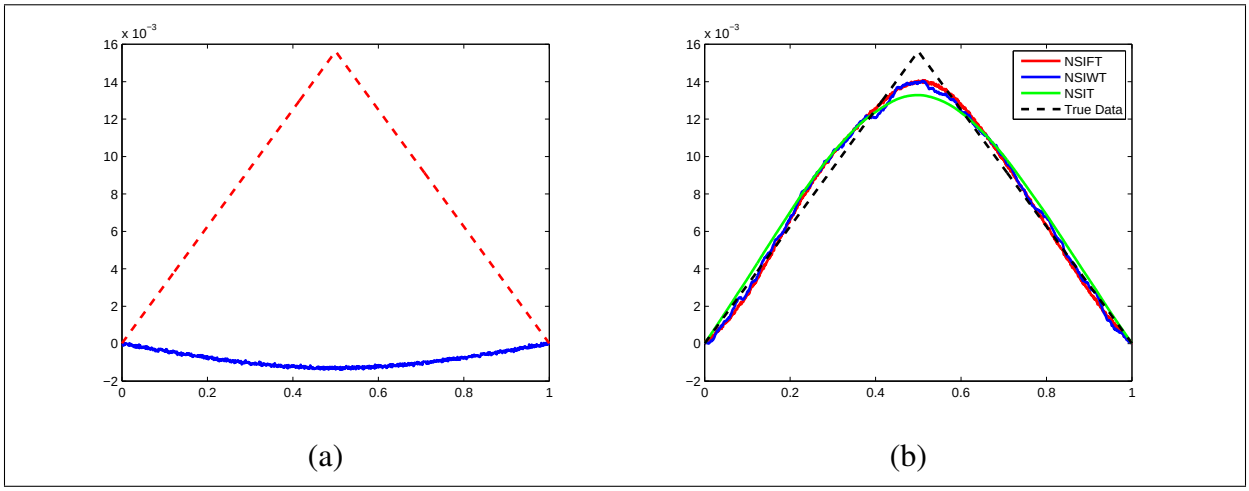


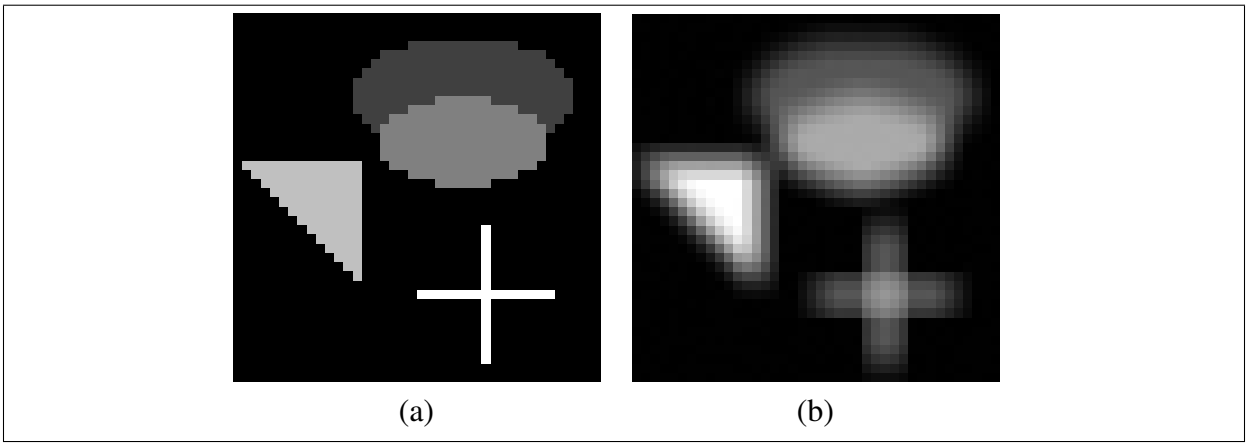
Figure 5.3: Example 2 – "deriv2" test case: (a) the true solution (dashed curve) and the observed data (solid curve), (b) approximated solutions.

α_0	Method	q		
		0.7	0.8	0.9
10^{-1}	NSIFT ($\gamma_n = 0.8$)	0.08981(11)	0.09394(13)	0.09445(19)
	NSIWT-I ($r_n = 0.6$)	0.08051(13)	0.09181(17)	0.09401(29)
	NSIT	0.08502(15)	0.09175(21)	0.09466(37)
	NSIFT (γ_n in (5.10.2))	0.09428(13)	0.09089(19)	0.09327(29)
	NSIWT-I (r_n in (5.10.2))	0.09073(13)	0.08648(19)	0.09199(29)
10^{-2}	NSIFT ($\gamma_n = 0.8$)	0.09114(5)	0.08953(7)	0.08998(9)
	NSIWT-I ($r_n = 0.6$)	**0.07807(7)**	0.09411(7)	0.09183(11)
	NSIT	0.08183(9)	0.09174(11)	0.09379(17)
	NSIFT (γ_n in (5.10.2))	**0.07839(9)**	0.08721(11)	0.09246(15)
	NSIWT-I (r_n in (5.10.2))	0.09399(7)	0.08389(11)	0.08990(15)

Table 5.3: Example 2: relative errors and iteration numbers between brackets for NSIWT-I and NSIFT with the nonstationary α_n in (5.10.1) and different choices of r_n and γ_n (NSIT is $r_n = \gamma_n = 1$).

	NSIFT	NSIWT-I	NSIT
Error	0.054831(9)	0.059211(7)	0.081835(9)

Table 5.4: Example 2: relative restoration errors and iteration numbers between brackets for NSIFT and NSIWT-I with parameters in (5.10.3) and NSIT with $\alpha_n = 0.01 \cdot 0.7^n$.



(a) (b)

Figure 5.4: Example 3 – "blur" test case: (a) the true image, (b) the measured data.

5.10.3 Example 3

We consider the test problem *blur(·,·,·)* in the toolbox REGULARIZATION TOOL by P. Hansen [62]. This is a two dimensional deblurring problem, the true solution is a 40×40 image, the blurring operator is a symmetric BTTB (block Toeplitz with Toeplitz block) with bandwidth 6. This blur is created by a truncated Gaussian point spread function with variance 2. For the noise vector it holds $\nu = 0.005$. Figure 5.4(a) shows the true image while the observed image is in Figure 5.4(b).

Firstly, α_n is defined by the classical choice in (5.10.1). Table 5.5 provides the results for a good stationary choice of r_n and γ_n. Note that NSIWT-I and NSIFT usually outperform NSIT. Table 5.6 shows that the relative restoration error obtained with the unbounded sequences r_n and γ_n in (5.10.3) is lower than the best one (according to Table 5.5), obtained by the stationary choice of r_n and γ_n. We note that NSIWT-I and NSIFT are less sensitive than NSIT to an appropriate choice of α_0 and q. In particular using r_n and γ_n in (5.10.3), NSIWT-I and NSIFT do not need any parameter estimation and the computed solutions have a relative restoration error lower than NSIT with the best parameter setting (see Table 5.5) and they provide also a better reconstruction, in particular of the edges, see Figure 5.5.

Finally, note that for the NSIT a nondecreasing sequence of α_n could be considered instead of the geometric sequence (5.10.1), see [38]. Nevertheless, this strategy requires a proper choice of α_0 and this is out of the scope of this paper, but it could be investigated in the future in connection with our fractional and weighted-I variants. A further development of our iterative schemes is in the direction of the nonstationary preconditioning strategy in [41], which is inspired by an approximated solution of the NSIT and hence could be investigated also in a fractional framework.

5.11 Conclusions, open problems and further comments

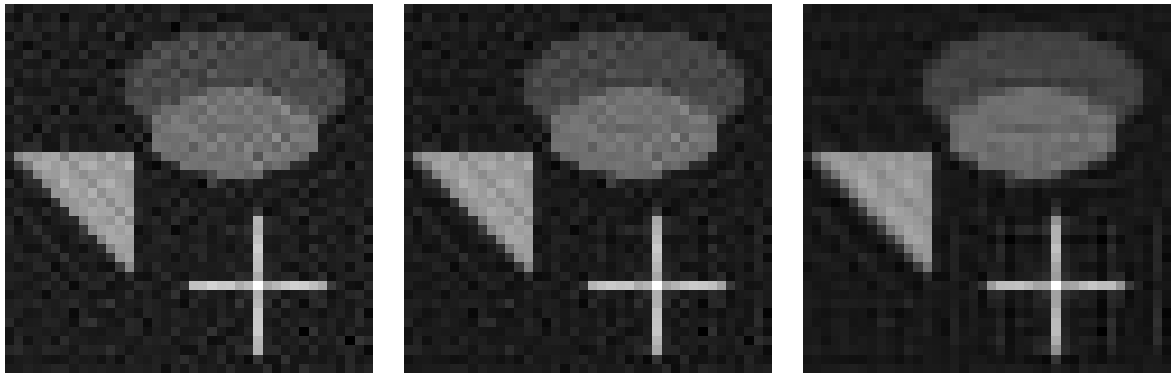
We extended recently proposed weighted and fractional versions of Tikhonov regularization to iterative regularization methods in the spirit of classical iterated Tikhonov regularization. The

α_0	Method	q		
		0.7	0.8	0.9
10^{-1}	NSIFT ($\gamma_n = 0.5$)	0.19970(9)	0.19526(13)	0.19847(17)
	NSIWT-I ($r_n = 0.2$)	0.18936(7)	**0.18920(9)**	0.19732(11)
	NSIT	0.19816(15)	0.21786(20)	0.28703(20)
10^{-2}	NSIFT ($\gamma_n = 0.5$)	0.19398(5)	0.19962(5)	0.19595(7)
	NSIWT-I ($r_n = 0.2$)	0.20822(3)	0.19547(3)	0.19109(3)
	NSIT	0.19518(9)	0.20531(11)	0.20747(17)

Table 5.5: Example 3: relative errors for NSIWT-I and NSIFT with the nonstationary α_n in (5.10.1).

	NSIFT	NSIWT-I	NSIT
Error	0.19335(10)	0.18765(8)	0.19518(9)

Table 5.6: Example 3: relative restorations errors for NSIFT and NSIWT-I with parameters in (5.10.3) and NSIT with $\alpha_n = 0.01 \cdot 0.7^n$.



(a) (b) (c)

Figure 5.5: Example 3 – "blur" reconstructions: (a) NSIFT and (b) NSIWT-I with parameters in (5.10.3), (c) NSIT with $\alpha_n = 0.01 \cdot 0.7^n$.

analysis uses the well-known technique of filter functions and contains all types of desired results: the proposed methods are regularizing, they converge, one can prove convergence rates, and the rates saturate at a known level. Furthermore, numerical examples have been provided showing that the weighted or fractional variants can be superior with respect to classical iterated Tikhonov regularization.

It would be interesting to investigate experimentally further the action of the weighted-II Tikhonov filter as a switch for the regularization in combination with further types of filters. Moreover, extensions of those filters to generic Banach spaces are still unknown.

Regularization Preconditioners

6.1 Preliminary definitions

We begin giving some useful definitions that will be used later in the Chapter.

Definition 6.1.1 (Toeplitz matrix). *Let T be an $n \times n$ matrix. We say that T is* Toeplitz *if it takes the form*

$$T = \begin{bmatrix} t_0 & t_{-1} & t_{-2} & \cdots & \cdots & t_{-(n-1)} \\ t_1 & t_0 & t_{-1} & \ddots & & \vdots \\ t_2 & t_1 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & t_{-1} & t_{-2} \\ \vdots & & \ddots & t_1 & t_0 & t_{-1} \\ t_{n-1} & \cdots & \cdots & t_2 & t_1 & t_0 \end{bmatrix}.$$

It is fully specified by the vector $\boldsymbol{v} = \begin{bmatrix} t_{n-1} & \cdots & t_1 & t_0 & t_{-1} & \cdots & t_{-(n-1)} \end{bmatrix}$.
If the (i,j) element of the matrix T is denoted by $T_{i,j}$, then we have

$$T_{i,j} = T_{i+1,j+1} = t_{i-j}.$$

Definition 6.1.2 (Hankel matrix). *Let H be an $n \times n$ matrix. We say that H is* Hankel *if it takes the form*

$$H = \begin{bmatrix} h_0 & h_1 & h_2 & h_4 & \cdots & h_{n-1} \\ h_1 & h_2 & h_3 & \cdots & \cdots & h_n \\ h_2 & h_3 & \cdots & \cdots & & \vdots \\ \vdots & & & & & \vdots \\ \vdots & & & & h_{2n-4} & h_{2n-3} \\ h_{n-1} & \cdots & \cdots & h_{2n-4} & h_{2n-3} & h_{2n-2} \end{bmatrix}.$$

If the (i,j) element of the matrix H is denoted by $H_{i,j}$, then we have

$$H_{i,j} = T_{i+1,j-1} = h_{i+j-2}.$$

Definition 6.1.3 (Discrete Fourier transform). *Let $\omega_n = e^{-\frac{2\pi \mathrm{i}}{n}}$ be the n^{th} root of unity, where $\mathrm{i}^2 = -1$. The* discrete Fourier transform *(*DFT*) is then expressed in the following way*

$$F_n = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-2} & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \cdots & \omega_n^{2(n-1)} \\ \vdots & & & & & \vdots \\ 1 & \omega_n^{n-2} & \cdots & & \cdots & \omega_n^{(n-2)(n-1)} \\ 1 & \omega_n^{n-1} & \cdots & & \cdots & \omega_n^{(n-1)(n-1)} \end{bmatrix},$$

with $F_n^{-1} = \frac{1}{n} F_n^*$.

Definition 6.1.4 (**Circulant matrix**). *Let C be an $n \times n$ matrix. We say that C is* circulant *if it takes the form*

$$C = \begin{bmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & & & c_2 \\ c_2 & c_1 & c_0 & c_{n-1} & & \vdots \\ \vdots & & & & & \\ c_{n-2} & & & & \ddots & c_{n-1} \\ c_{n-1} & c_{n-2} & \cdots & \cdots & c_1 & c_0 \end{bmatrix}.$$

It is fully specified by the first column vector $\boldsymbol{c} = \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \end{bmatrix}^t$.
If we define

$$U_n^* = \frac{1}{\sqrt{n}} F_n, \qquad U_n = \sqrt{n} F_n^{-1},$$

where F_n is the DFT, then C can be diagonalized by U_n. In fact, we have the following relation

$$C = U_n \mathrm{diag}\left(F_n \boldsymbol{c}\right) U_n^*.$$

The above definitions of Toeplitz, Hankel and circulant matrices can be applied even to block matrices. For example, a Toeplitz block matrix T has the form

$$T = \begin{bmatrix} T_0 & T_{-1} & T_{-2} & \cdots & \cdots & T_{-(n-1)} \\ T_1 & T_0 & T_{-1} & \ddots & & \vdots \\ T_2 & T_1 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & T_{-1} & T_{-2} \\ \vdots & & \ddots & T_1 & T_0 & T_{-1} \\ T_{n-1} & \cdots & \cdots & T_2 & T_1 & T_0 \end{bmatrix},$$

where every block T_j is a square $m \times m$ matrix. A *block Toeplitz Toeplitz block* (BTTB) is a block Toeplitz matrix with every block T_j Toeplitz.

Theorem 6.1.5 (**Cauchy's Interlacing Theorem**). *Let $A \in \mathbb{C}^{n \times n}$ be an Hermitian matrix, i.e., $A = A^*$, with eigenvalues $\lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$. Let A be partitioned as*

$$A = \begin{bmatrix} E & B^* \\ B & G \end{bmatrix},$$

where $E \in \mathbb{C}^{m \times m}, B \in \mathbb{C}^{(n-m) \times m}$ and $G \in \mathbb{C}^{(n-m) \times (n-m)}$. Then the eigenvalues $\theta_m \leq \theta_{m-1} \leq \cdots \leq \theta_1$ of E satisfy

$$\lambda_{k+n-m} \leq \theta_k \leq \lambda_k.$$

6.2 Introduction

Image deblurring is the process of reconstructing an approximation of an image from blurred and noisy measurements. By assuming that the point spread function h (PSF) is spatially-invariant, the observed image $g(x,y)$ is related to the true image $f(x,y)$ via the integral equation

$$g(s,t) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(s-s',t-t')f(s',t')\,ds'\,dt' + \eta(s,t), \qquad (s,t) \in \Omega \subset \mathbb{R}^2, \tag{6.2.1}$$

where $\eta(s,t)$ is the noise.

By collocation of the previous integral equation on a uniform grid, we obtain the grayscale images of the observed image, of the true image, and of the PSF, denoted by G, F, and H, respectively. Since collected images are available only in a finite region, the field of view (FOV), the measured intensities near the boundary are affected by data outside the FOV.

Figure 6.1: Field of view. We see what is inside the square box.

Given an $n \times n$ observed image G (for the sake of simplicity we assume square images), and a $p \times p$ PSF with $p \leq n$, then F is $m \times m$ with $m = n + p - 1$. Denoting by $\mathbf{g}$ and $\mathbf{f}$ the stack ordered vectors corresponding to G and F, the discretization of (6.2.1) by a rectangular quadrature rule with uniform grid (for example) leads to the under-determined linear system

$$\mathbf{g} = A\mathbf{f} + \eta, \tag{6.2.2}$$

where the matrix A is of size $n^2 \times m^2$. When imposing proper Boundary Conditions (BCs), the image A becomes square $n^2 \times n^2$ and in some cases, depending on the BCs and the symmetry of the PSF, it can be diagonalized by discrete trigonometric transforms. For example, the matrix A is block circulant circulant block (BCCB) and it is diagonalizzable by Discrete Fourier Transform (DFT), when periodic BCs are imposed. See Figure 6.2.

Due to the ill-posedness of (6.2.1), A is severely ill-conditioned and may be singular. In such case, linear systems of equations (6.2.2) are commonly referred to as linear discrete ill-posed

problems [59]. Therefore a good approximation of $\mathbf{f}$ cannot be obtained from the algebraic solution (e.g., the least-square solution) of (6.2.2), but regularization methods are required. The basic idea of regularization is to replace the original ill-conditioned problem with a nearby well-conditioned problem, whose solution approximates the true solution. One of the popular regularization techniques is the Tikhonov regularization, as we have already seen in the previous Chapter, and it amounts in solving

$$\min_{\mathbf{f}}\{\|A\mathbf{f}-\mathbf{g}\|_2^2+\mu\|\mathbf{f}\|_2^2\}, \tag{6.2.3}$$

where $\|\cdot\|_p$ denotes the vector p-norm, $p\geq 1$, and $\mu>0$ is a regularization parameter to be chosen. Compare it to equation (5.2.3). Hereafter, we use $\|\cdot\|\equiv\|\cdot\|_2$ to denote the ℓ_2-norm. The first term in (6.2.3) is usually refereed as fidelity term and the second as regularization term. This approach is computationally attractive, since it leads to a linear problem and indeed several efficient methods have been developed for computing its solution and for estimating μ [59]. On the other hand, the edges of restored image are usually over-smoothed. To overcome this unpleasant property, nonlinear strategies have been employed, like total variation (TV) [91] and thresholding iterative methods [35, 49]. Anyway, several nonlinear regularization methods have an inner step that apply a least-square regularization and hence can benefit from strategies previously developed for such simpler model, as we will show in the following.

In this Chapter we consider a regularization strategy based on wavelet decomposition that has been recently largely investigated [21, 22, 26, 49, 36, 35]. This approach is motivated by the fact that most real images usually have sparse approximations under some wavelet basis. In particular, in this Chapter we consider the tight frame systems previously used in [19, 21, 22]. Solving (6.2.2) in a tight frame domain, the redundancy of system leads to robust signal representation in which partial loss of the data can be tolerated without adverse effects. In order to obtain the sparse approximation, we minimize the weighted ℓ_1-norm of the tight frame coefficients. Let W^* be a wavelet or tight-frame synthesis operator ($W^*W=I$), the wavelets or tight-frame coefficients of the original image $\mathbf{f}$ are $\mathbf{x}$ such that

$$\mathbf{f}=W^*\mathbf{x}. \tag{6.2.4}$$

In the following, we will investigate the synthesis approach, but our proposal can be applied also to the analysis and to the balanced approach described in [22, 99]. Reformulating the deblurring problem (6.2.2) in terms of frame coefficients

$$\min_{\mathbf{x}}\{\mu\|\mathbf{x}\|_1+\|\mathbf{x}\|^2 : AW^*\mathbf{x}=\mathbf{g}\}, \tag{6.2.5}$$

a regularized solution of this problem can be obtained by the Bregman splitting [115]. Indeed, as we will see in Section 6.4, the regularization is made upon equation (6.4.3) (which is equivalent to the above equation (6.2.5) if K is surjective) in the same spirit of iterative regularization methods like iterated Tikhonov, introduced in the preceding chapter. As for the iterative soft-thresholding [35, 49] for the unconstrained version of (6.2.5) and the Landweber method for the least-square solution of (6.2.2), the Bregman splitting converges very slowly for image deblurring problems. Hence a preconditioning strategy is usually employed, obtaining the Modified

Linearized Bregman Algorithm (MLBA) [21]. The preconditioner is usually chosen as a BCCB approximation of $(AA^* + \alpha I)^{-1}$, $\alpha > 0$, see [21, 22, 99], which is the simplest regularized version of the inverse of AA^* which, also when theoretically available, cannot be computed due to the severe ill-conditioning of A.

In this Chapter, we show that the BCCB preconditioner used in the literature leads often to poor restorations when the matrix A has the rectangular $n^2 \times m^2$ structure or is obtained by imposing accurate BCs, like antireflective BCs [98]. Note that in real applications we have to take into account the boundary effects to obtain high quality restorations, otherwise the restored image is severely affected by ringing effects [64, 81]. This topic has been recently investigated in connection with nonlinear models based on wavelets or TV in [103, 3, 6], but, at our knowledge, this is the first time that it is considered in connection with the MLBA. In this context we propose and discuss other preconditioning strategies for the MLBA with the synthesis approach. Our preconditioners are inspired by the experience with least-square regularization where the regularization preconditioning is studied since a long time, see the seminal paper [61]. In particular a nonstationary preconditioned iteration suggested by [41] leads to a new algorithm that is no longer a Bregman iteration.

We investigate the following two strategies to define accurate and computationally cheap preconditioners:

(1) an approximation of the blurring operator in a small Krylov subspace;

(2) a symmetrization of the original PSF H;

The choice (1) is quite natural and already considered for similar problems (see e.g. [3]), but we will show that a properly chosen Krylov subspace of small size (say spanned by at most five vectors), with a proper choice of the initial guess, is usually enough to obtain a good approximation. The choice (2) can be very useful in many applications where the PSF is obtained experimentally by measurements and is a perturbation of a symmetric kernel. In this case the approximated quadrantally symmetric (i.e., symmetric with respect to each quadrant) PSF leads to a matrix diagonalizable by Discrete Cosine Transform (DCT).

Following the idea to combine preconditioned regularizing iterative methods for least-square ill-posed problems with the MLBA, we propose a new algorithm based on the recent proposal in [37, 41]. The nonstationary preconditioner is defined by a parameter computed by solving a nonlinear problem with a computational cost of $O(n^2)$. We observe that this method can be applied only with square matrices and so only when A is obtained by imposing BCs. The new algorithm is no longer a Bregman iteration and we cannot apply the convergence analysis developed in [21] for the MLBA. Therefore, here we prove its convergence and its regularization character. Furthermore, when a good value for the parameter in the preconditioner is available, we provide a variant of the algorithm with a stationary preconditioner which can improve the quality of the restorations even if the previous convergence analysis does not hold any longer.

A large number of numerical experiments in Section 6.7 shows that our proposals not only outperform the standard MLBA with BCCB preconditioning, but are also good competitors for other recent methods dealing with boundary artifacts proposed in [3, 6].

Besides, we mention that the two deblurring models based on the rectangular matrix A and the imposition of BCs, we have tested also a third strategy based on the enlargement of the domain to reduce the ringing effects like in [89, 103], but, according to the results in [3], the quality of the restored images were not better than those obtained with the other two models, while the CPU time was higher. Hence, we do not discuss further this strategy here.

The Chapter is organized as follows. In Section 6.3, we describe the structure of the blurring matrix A explaining how fast trigonometric transforms can be used in the computations both for the rectangular matrix and the square matrices arising from the imposition of classical BCs. Section 6.4 reviews briefly the synthesis approach and the MLBA for solving the corresponding minimization problem. In Section 6.5 we propose possible regularization preconditioners, combining accurate restorations and a low computational cost. A new algorithm is proposed in Section 6.6 combining the MLBA with the method in [41]. Section 6.7 contains a large number of numerical experiments, comparing our proposal with some state of the art algorithms, for the restoration of images with unknown boundaries. Concluding remarks are provided in Section 6.8.

6.3 The structure of the blurring matrix

We count mainly three strategies in order to obtain both accurate and fast restorations with reduced boundary artifacts. In this Chapter, we just consider two of them: the use of the original rectangular matrix and the imposition of BCs. As mentioned in the Introduction, we do not consider the third strategy introduced in [89], since from one side it is equivalent to the reflective (or Neumann) BCs in the case of quadrantally PSF, cf. [45], and from the other side, according to several tests that we have performed and the numerical results in [3], it does not provide a better restoration than the other two strategies, while it usually requires a larger CPU time.

In this section we describe the structure of the matrix A and how fast computations with such matrix, like matrix-vector product or least-square solutions, can be implemented. Firstly we introduce the rectangular $n^2 \times m^2$ matrix and then the square $n^2 \times n^2$ matrix obtained when imposing proper BCs.

6.3.1 The rectangular matrix

The fact that this matrix is not square prevents the use of Fast Fourier Transform (FFT). To cope with this difficulty, one can construct an $m^2 \times m^2$ blurring matrix A_{big} that is BCCB, and hence, the FFT can be used. Let $M \in \mathbb{R}^{n^2 \times m^2}$ be a masking matrix which, when applied to a vector in $\mathbb{R}^{m^2}$, selects only the entries in the FOV, i.e., their rows are a subset of the rows of an identity matrix of order m^2. The rectangular blurring matrix can be written as

$$A = MA_{\text{big}}. \tag{6.3.1}$$

Hence the matrix vector $A\mathbf{x}$ can be easily computed by two bi-dimensional FFTs of order m^2, followed by a selection of the pixels inside the FOV. Similarly $A^*\mathbf{x} = A_{\text{big}}^* M^*\mathbf{x}$ and thus two FFTs

are applied to the zero-padded version of $\mathbf{x}$ of size m^2. This approach was used in [3] and it is numerically equivalent to that adopted in [112].

We observe that the matrix M in (6.3.1) leads to a matrix A independent of the BCs used in the definition of A_{big}. Thus, when the PSF is quadrantally symmetric, we suggest to use of the DCT instead of the DFT (see the discussion on reflective BCs in the next subsection).

The structure of the matrix A and its representation in (6.3.1) allow fast computations of the matrix vector product with A and A^*, but they prevent the use of fast transforms for solving linear systems with polynomials of AA^* as coefficient matrix even when A is full-rank.

6.3.2 Boundary conditions

The BC approach forces a functional dependency between the elements of $\mathbf{f}$ external to the FOV and those internal to this area. This has the effect of extending F outside of the FOV without adding any unknowns to the associated image deblurring problem. Therefore, the matrix A can be written as an $n^2 \times n^2$ square matrix, whose structure can be exploited by fast algorithms. If the BC model is not a good approximation of the real world outside the FOV, the reconstructed image can be severely affected by some unwanted artifacts near the boundary, called ringing effects [64].

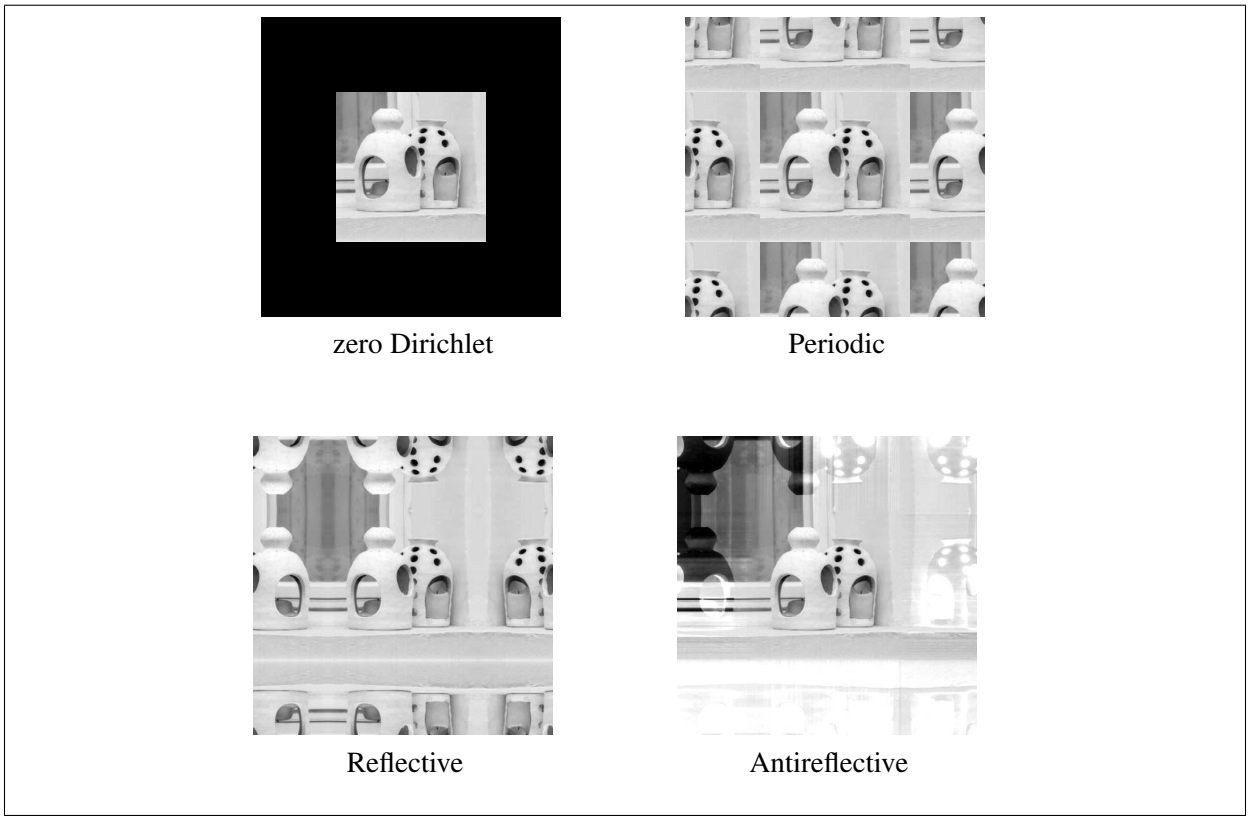


Figure 6.2: Examples of boundary conditions.

The use of different BCs can be motivated from information on the true image and/or from the

Zero	Periodic	Reflective
$0\ \ 0\ \ 0$	$F\ \ F\ \ F$	$F_{rc}\ \ F_r\ \ F_{rc}$
$0\ \ F\ \ 0$	$F\ \ F\ \ F$	$F_c\ \ F\ \ F_c$
$0\ \ 0\ \ 0$	$F\ \ F\ \ F$	$F_{rc}\ \ F_r\ \ F_{rc}$

Table 6.1: Pad of the original image F obtained by imposing the classical BCs considered in [64], with $F_c = \text{fliplr}(F)$, $F_r = \text{flipud}(F)$, and $F_{rc} = \text{flipud}(\text{fliplr}(F))$, where $\text{fliplr}(\cdot)$ and $\text{flipud}(\cdot)$ are the MATLAB functions that perform the left-right and up-down flip, respectively.

availability of fast transforms to diagonalize the matrix A within $O(n^2 \log(n))$ arithmetic operations. Indeed, the matrix-vector product can be always computed by the 2D FFT, after a proper padding of the image to convolve (the resulting image is the inner $n \times n$ part of the convolution), cf. [80], while the availability of fast transforms to diagonalize the matrix A depends on the BCs. Anyway, the shift-invariant property of the blur leads to a matrix A that can be well approximated by a BCCB matrix C, which is diagonalized by DFT, because usually in the applications

$$A - C = R + N, \tag{6.3.2}$$

where R is a matrix of small rank and N is a matrix of small norm. More precisely, for any $\varepsilon > 0$ there is a constant $c_\varepsilon > 0$ independent of n and depending only on ε and on the PSF, such that the splitting (6.3.2) holds with

$$\text{rank}(R) \leq c_\varepsilon \cdot n, \qquad \|N\| \leq \varepsilon, \tag{6.3.3}$$

where $\text{rank}(R)$ denotes the rank of R (see [60]). Note that n^2 is the size of the matrix A.

In the following we recall common BCs that will be used in the numerical tests. For a detailed description of zero, periodic, and reflective, refer to [64], while for antireflective BC see the review paper [44] and the original proposal in [98].

Zero (i.e., Dirichlet) BCs assume that the object is zero outside of the FOV. That is, one assumes that F has been extracted from a larger array padded by zeros (see Table6.1). This is a good choice when the true image is mostly zero outside the FOV, as is the case for many astronomical or medical images with a black background. Unfortunately, these BCs have a bad effect on reconstructions of images that are nonzero outside the border, leading to reconstructed image with severe "ringing" near the boundary. The corresponding matrix A has a block-Toeplitz-Toeplitz-block (BTTB) structure which is not diagonalizable by fast trigonometric transforms.

Periodic BCs assume that the observed image repeats in all directions. More specifically, one assumes that F has been extracted from a larger array of the form in Table 6.1. The corresponding matrix A is BCCB and so is always diagonalized by 2D DFT. Clearly, if the true image is not periodic outside the FOV the reconstructed image will be affected by severe ringing effects.

Reflective (i.e., Neumann or symmetric) BCs assume that outside the FOV the image is a mirror image of F [81]. That is, one assumes that F has been extracted from a larger array symmetrically padded like in Table 6.1. The matrix A has a block structure that combines Toeplitz and Hankel structures, but it can be easily diagonalized by the DCT, when the PSF is quadrantally symmetric [81].

AntiReflective BCs have a more elaborate definition, but have a simple motivation: the antisymmetric pad yields an extension that preserves the continuity of the normal derivative [98]. They are given by

$$\begin{aligned}
F(1-i,j) &= 2F(1,j) - F(i+1,j), && 1 \le i \le p,\ 1 \le j \le n;\\
F(i,1-j) &= 2F(i,1) - F(i,j+1), && 1 \le i \le n,\ 1 \le j \le p;\\
F(n+i,j) &= 2F(n,j) - F(n-i,j), && 1 \le i \le p,\ 1 \le j \le n;\\
F(i,n+j) &= 2F(i,n) - F(i,n-j), && 1 \le i \le n,\ 1 \le j \le p;
\end{aligned}$$

for the edges, while for the corners the more computationally attractive choice is to antireflect first in one direction and then in the other [40]. This yields

$$\begin{aligned}
F(1-i,1-j) &= 4F(1,1) - 2F(1,j+1) - 2F(i+1,1) + F(i+1,j+1),\\
F(1-i,n+j) &= 4F(1,n) - 2F(1,n-j) - 2F(i+1,n) + F(i+1,n-j),\\
F(n+i,1-j) &= 4F(n,1) - 2F(n,j+1) - 2F(n-i,1) + F(n-i,j+1),\\
F(n+i,n+j) &= 4F(n,n) - 2F(n,n-j) - 2F(n-i,n) + F(n-i,n-j),
\end{aligned}$$

for $1 \le i,j \le p$. The structure of the matrix A is quite involved, but it can be diagonalized by the antireflective transform, when the PSF is quadrantally symmetric [4]. Since A is not normal the antireflective transform is not unitary, but it can be represented as a modification of the discrete sine transform, formed by adding a uniform sampling of constant and linear functions to the eigenvector basis preserving an "almost" unitary behaviour [44].[1]

Due to the structure of A, the application of A^* could generate artifacts at the boundary [42]; consequently it was proposed to replace A^* by a reblurring matrix A' obtained by imposing the same BCs to the PSF rotated by 180° [39]. Note that in the case of zero and periodic BCs $A' = A^*$. Furthermore, the MATLAB Toolbox RestoreTools [79] (that we will use in the numerical results) implements the reblurring approach to overload the matrix-vector product with A^*, see [39]. Therefore, for the sake of notational simplicity and uniformity, in the following the symbol A^* has to be intended as the reblurring matrix A' in the case of antireflective BCs.

The reflective BCs will not be considered in the numerical results in Section 6.7 since they have the same computational properties of the antireflective BCs, e.g., for quadrantally PSF simply replace the antireflective transform with the DCT, and usually provide restorations slightly

[1] MATLAB functions for working with antireflective BC (antireflective transform, antisymmetric pad, ecc.) can be download at http://scienze-como.uninsubria.it/mdonatelli/Software/software.html

worser or at least comparable with the antireflective BCs, as we have numerically observed according to similar results with other regularization strategies [39, 84, 32]. On the other hand, we have recalled also the reflective BCs to motivate a preconditioner based on the DCT, instead of the DFT, when the PSF is quadrantally symmetric.

6.4 MLBA for the synthesis approach

For the synthesis approach [21, 49] the coefficient matrix is

$$K = AW^* \in \mathbb{R}^{n^2 \times s}, \tag{6.4.1}$$

where $n^2 \leq s$, A is the blurring matrix and W^* is a tight-frame or wavelet synthesis operator. Note that using tight-frames $W^*W = I$ but $WW^* \neq I$ [36]. The use of tight-frames instead of wavelets is motivated by the fact that the redundancy of tight-frame systems leads to robust signal representations in which partial loss of the data can be tolerated, without adverse effects, see e.g. [26].

Denote by $\mathbf{x}$ the frame coefficients of the image $\mathbf{f}$ according to (6.2.4). Let the nonlinear operator $\mathbf{S}_\mu$ be defined component-wise as

$$[\mathbf{S}_\mu(\mathbf{x})]_i = S_\mu(x_i), \tag{6.4.2}$$

with S_μ the soft-thresholding function

$$S_\mu(x_i) = \operatorname{sgn}(x_i)\max\{|x_i| - \mu, 0\}.$$

Note that for image deblurring problems the singular values of A, and so those of K, decay exponentially to zero and we cannot assume that K is surjective. Therefore, the deblurring problem can be reformulated in terms of the frame coefficients $\mathbf{x}$ as

$$\min_{\mathbf{x}\in\mathbb{R}^s}\left\{\mu\|\mathbf{x}\|_1 + \frac{1}{2\lambda}\|\mathbf{x}\|^2 : \arg\min_{\mathbf{x}\in\mathbb{R}^s}\|K\mathbf{x} - \mathbf{g}\|^2\right\}. \tag{6.4.3}$$

which is equivalent to (6.2.5) if K is surjective.

The following linearized Bregman iteration

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + K^*(\mathbf{g} - K\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \lambda\mathbf{S}_\mu(\mathbf{z}^{n+1}), \end{cases} \tag{6.4.4}$$

where $\mathbf{z}^0 = \mathbf{x}^0 = \mathbf{0}$, was introduced in [115] to solve problem (6.4.3) and later applied to image deblurring in [21]. A detailed convergence analysis of the linearized Bregman iteration (6.4.4) was given in [20] when K is surjective, but we report here Theorem 3.1 in [21] that does not require such assumption.

Theorem 6.4.1 ([21]). *Let $K \in \mathbb{R}^{n^2\times s}$, $n^2 < s$ and let $0 < \lambda < \frac{1}{\|K^*K\|}$. Then the sequence $\{\mathbf{x}^{n+1}\}$ generated by* (6.4.4) *converges to the unique solution of* (6.4.3).

As observed in [20] the convergence speed of (6.4.4) depends of the condition number of K which, as observed before, is very large for image deblurring and hence the method results to be very slow. To accelerate its convergence in the case of $KK^* \neq I$, in [21] the authors modified iteration (6.4.4) by replacing K^* with $K^\dagger$, where $K^\dagger$ denotes the pseudo-inverse of K. If K is surjective $K^\dagger = K^*(KK^*)^{-1}$ since $n^2 \leq s$. For image deblurring problems they suggested to replace $(KK^*)^\dagger$ with a symmetric positive definite matrix P such that

$$P \approx (KK^*)^\dagger = (AA^*)^\dagger. \tag{6.4.5}$$

Then the MLBA for frame-based image deblurring becomes [21]

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + K^*P(\mathbf{g} - K\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \lambda \mathbf{S}_\mu(\mathbf{z}^{n+1}), \end{cases} \tag{6.4.6}$$

where $\mathbf{z}^0 = \mathbf{x}^0 = \mathbf{0}$.

Remark 6.4.2. *The MLBA* (6.4.6) *is the linearized Bregman iteration* (6.4.4) *for the linear system*

$$P^{1/2}K\mathbf{x} = P^{1/2}\mathbf{g},$$

which is a preconditioned version of original linear system $K\mathbf{x} = \mathbf{g}$ by the preconditioner $P^{1/2}$. In fact, by replacing K and $\mathbf{g}$ in (6.4.4) *by $P^{1/2}K$ and $P^{1/2}\mathbf{g}$, respectively, we obtain* (6.4.6).

The previous remark is the key observation used in [21] to prove that the MLBA algorithm converges to a minimizer of

$$\min_{\mathbf{x}\in\mathbb{R}^s} \left\{ \mu\|\mathbf{x}\|_1 + \frac{1}{2\lambda}\|\mathbf{x}\|^2 : \mathbf{x} = \arg\min_{\mathbf{x}\in\mathbb{R}^s} \|K\mathbf{x} - \mathbf{g}\|_P^2 \right\}, \tag{6.4.7}$$

where $\|\mathbf{x}\|_P = \langle P^{1/2}\mathbf{x}, P^{1/2}\mathbf{x}\rangle$.

Theorem 6.4.3 ([21])**.** *Assume P is a symmetric positive definite matrix and let $0 < \lambda < \frac{1}{\|K^*PK\|}$. Then the sequence $\{\mathbf{x}^{n+1}\}$ generated by the MLBA* (6.4.6) *converges to the unique solution of* (6.4.7).

The standard choice for P is

$$P = (KK^* + \alpha I)^{-1} = (AA^* + \alpha I)^{-1}. \tag{6.4.8}$$

In such case

$$\|K^*PK\| < 1$$

and hence $\lambda = 1$ is a good choice according to Theorem 6.4.3. When $AA^* + \alpha I$ is not easily invertible other choices as P can be explored, but usually the quantity $\|K^*PK\|$ becomes hard to estimate. Therefore, assuming that P is a good approximation of $(AA^* + \alpha I)^{-1}$, we set $\lambda = 1$ in the algorithm. Anyway, the validity of the assumption $\|K^*PK\| < 1$ can be guaranteed choosing α large enough.

In conclusion, we consider the following version of the MLBA for the synthesis approach

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + WA^*P(\mathbf{g} - AW^*\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}), \end{cases} \tag{6.4.9}$$

stopped by the *discrepancy principle* as in [21], i.e., at the first iteration $n = \tilde{n} > 0$ such that

$$\|\mathbf{r}^{\tilde{n}}\| \leq \gamma\delta < \|\mathbf{r}^n\|, \qquad n = 0, 1, \ldots, \tilde{n} - 1, \tag{6.4.10}$$

where $\gamma > 1$, $\delta = \|\eta\|$, and $\mathbf{r}^n = \mathbf{g} - AW^*\mathbf{x}^n$ is the residual at the n-th iteration. Here $\mathbf{z}^0 = \mathbf{x}^0 = \mathbf{0}$ and we assume that the noise level δ is explicitly known.

6.5 On the choice of the preconditioner P

In this section we explore possible choices of $P \neq (AA^* + \alpha I)^{-1}$ which are computationally attractive. Let A be the rectangular, anti-reflective or BTTB matrix depending on the chosen treatment of the boundary of the image, and let C be the BCCB obtained from the same PSF. Since the matrix P in Theorem 6.4.3 serves as a preconditioner to accelerate the convergence, in this section we describe some preconditioning strategies, in order to combine fast computations with accurate restorations achievable when $P \approx (AA^* + \alpha I)^{-1}$. The first proposal in Section 6.5.1 is the classical approach already used in the literature, cf. [21, 22, 99]. The second proposal in Section 6.5.2 is an approximation strategy considered for similar methods, c.f. [3], but this is the first time that it is explored with MLBA. The third proposal is inspired by a similar approach used with numerical methods that require symmetric matrices, see [60, 81].

6.5.1 BCCB preconditioner

Let C be the matrix obtained imposing periodic BCs. As described in Section 6.3.2, the matrix C is diagonalizable by DFT. Hence, the matrix-vector product with the matrix

$$P = (CC^* + \alpha I)^{-1}$$

can be efficiently computed by FFT and its use was previously proposed in [21].

Algorithm 1.

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + WA^*(CC^* + \alpha I)^{-1}(\mathbf{g} - AW^*\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}). \end{cases} \tag{6.5.1}$$

We suggest to replace the DFT with the DCT, in the case of a quadrantally symmetric PSF like the Gaussian blur. The latter choice not only is motivated by computational considerations, since the complex DFT is replaced by a real transform (DCT), but also by the quality of the computed approximation. Indeed, using the DCT the matrix C can be seen as an approximation of A by imposing reflective BCs instead of periodic BCs, which results usually in a better approximation and so provides better restorations. The same expedient will be used for the following preconditioners as well.

Note that $\|(CC^* + \alpha I)^{-1}\| < \|(AA^* + \alpha I)^{-1}\|$ is a sufficient condition to apply the Theorem 6.4.3 with $\lambda = 1$. This assumption could be hard to be satisfied in practice. Nevertheless, it is expected that

$$\|K^*(CC^* + \alpha I)^{-1}K\| < 1 \tag{6.5.2}$$

if α is large enough.

In the literature regarding preconditioning of Toeplitz matrices by circulant matrices, several strategies have been proposed to compute the matrix C, cf. [25]. In this Chapter we simply consider the matrix obtained imposing periodic BCs, which corresponds to the natural Strang preconditioner, since we have not observed numerical differences, when using other strategies like the optimal Frobenius norm approximation preconditioner [27]. Roughly speaking, this follows from the fact that the entries of A depend on the value of the pixels of the PSF according to the shift invariance structure. In particular the central coefficient of the PSF belongs to the main diagonal of A and the pixels near the center of the PSF belongs to the central diagonals of the central blocks. Finally, the PSF is almost centered in the middle of a $n \times n$ image and hence every pixel is distant at most $n/2$ pixels in every direction (usually much less due to the compact support). Hence the block band and the band of each block of A are at most $n/2$.

6.5.2 Krylov subspace approximation

The preconditioner in the previous section is essentially defined as an approximation of the operator A in the Fourier domain. Another strategy, useful also for more general matrices, is to employ orthogonal or oblique projections into subspaces of small dimension. A common choice is a proper Krylov subspace.

The matrix vector product $\mathbf{t}^n = (AA^* + \alpha I)^{-1}\mathbf{r}^n$ can be computed solving the linear system

$$(AA^* + \alpha I)\mathbf{t}^n = \mathbf{r}^n, \tag{6.5.3}$$

whose solution can be approximated by few iterations of conjugate gradient (CG) since $AA^* + \alpha I$ is symmetric and positive definite. One or few steps of CG to approximate the vector $(AA^* + \alpha I)^{-1}\mathbf{r}^n$ is a common strategy, see e.g. [3]. Here we explore the use of a good preconditioner associated with a proper choice of the initial guess and the stopping criteria. We solve (6.5.3) by preconditioned CG (PCG) with preconditioner the matrix $(CC^* + \alpha I)^{-1}$ introduced in Section 6.5.1. This is equivalent to solve the linear system

$$(CC^* + \alpha I)^{-1/2}(AA^* + \alpha I)(CC^* + \alpha I)^{-1/2}\mathbf{y}^n = (CC^* + \alpha I)^{-1/2}\mathbf{r}^n, \tag{6.5.4}$$

with $\mathbf{y}^n = (CC^* + \alpha I)^{1/2}\mathbf{t}^n$.

The Krylov subspace of size j generated by the matrix B and the vector $\mathbf{v}$ is defined by

$$\mathscr{K}_j(B, \mathbf{v}) = \text{span}\{\mathbf{v}, B\mathbf{v}, \ldots, B^{j-1}\mathbf{v}\}, \qquad j \in \mathbb{N}.$$

We denote by $\mathbf{y}^n_{\beta_n}$ the vector that minimizes the energy norm of the error of the linear system (6.5.4) into the Krylov subspace

$$\mathscr{K}_{\beta_n} := \mathscr{K}_{\beta_n}\left((CC^* + \alpha I)^{-1/2}(AA^* + \alpha I)(CC^* + \alpha I)^{-1/2},\ (CC^* + \alpha I)^{-1/2}\mathbf{r}^n\right).$$

Therefore, defining

$$\mathbf{t}^n_{\beta_n} = (CC^* + \alpha I)^{-1/2}\mathbf{y}^n_{\beta_n}$$

the following algorithm can be sketched.

Algorithm 2.

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + WA^*\mathbf{t}^n_{\beta_n}, \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}). \end{cases} \tag{6.5.5}$$

Of course a large β_n is not practical and also the convergence of the Algorithm 2 could fail if $\mathbf{t}^n_{\beta_n}$ is not a good approximation of $\mathbf{t}^n$ in (6.5.3). However, in practice β_n can be taken very small and the PCG converges very rapidly assuring also the convergence of the Algorithm 2 as numerically confirmed by the results in Section 6.7. This follows from discussion at the beginning of Section 6.3.2 and the well-conditioning of the coefficient matrix of the linear system (6.5.3). Indeed, all the eigenvalues of $AA^* + \alpha I$ are in $[\alpha, c]$, with c constant independent of n and usually $c = 1 + \alpha$, because A arises from the discretization of (6.2.1) and, thanks to the physical properties of the PSF (nonnegative entries and sum of all pixels equal to one), its largest singular value is bounded by one. If follows that the BCCB preconditioner $CC^* + \alpha I$ is very effective since the spectrum of $(CC^* + \alpha I)^{-1/2}(AA^* + \alpha I)(CC^* + \alpha I)^{-1/2}$ is clustered at 1, with $O(n)$ outliers according to equations (6.3.2) and (6.3.3), while n^2 is the total number of eigenvalues (see [60, 24]).

Moreover, we observe that if the Algorithm 2 is converging then, in the noise free case, the residual is going to zero (otherwise stagnates around δ) and hence also the solution of the linear system (6.5.3) approaches the zero vector. This has two interesting consequences. First, a good initial guess for the PCG is the zero vector since it is a good approximation of the solution of the linear system (6.5.3), at least for n large enough. Second, the size of the Krylov subspace $\mathscr{K}_{\beta_n}$ should decreases when n increases reaching the same fixed accuracy in the approximation of $\mathbf{t}^n$ (see the following discussion on β_n).

Note that the computation of $\mathbf{t}^n_{\beta_n}$ requires β_n matrix-vector products with $AA^* + \alpha I$ and with $CC^* + \alpha I$. Nevertheless, according to the previous discussion, a small β_n, e.g., $\beta_n \leq 5$, is enough. In the numerical results in Section 6.7 we fix

$$\beta_n = \min\{5, \beta_{\text{tol}}\}, \tag{6.5.6}$$

where β_{tol} is the number of PCG iterations required for reaching the tolerance 10^{-3} in terms of the norm of the relative residual in the linear system (6.5.4). We observe that in our numerical results, β_n decreases quickly obtaining $\beta_n = 1$ for all $n > \bar{n}$, with $\bar{n}$ small.

Finally, we note that the preconditioner P obtained by the PCG approximation is not stationary and changes at each iteration of the MLBA. Therefore, Theorem 6.4.3 cannot be applied. Nevertheless, Algorithm 2 with the condition (6.5.6) has been convergent in all our numerical experiments confirming that $\mathbf{t}^n_{\beta_n}$ is a good approximation of $\mathbf{t}^n$, at least for n large enough.

6.5.3 Preconditioning by symmetrization

The preconditioner in Section 6.5.1 is related to a different boundary model, namely periodic BCs, but the deblurring problem and in particular the PSF are the same. Unfortunately, periodic BCs lead to poor restorations for generic images and for more accurate models, like reflective or antireflective BCs, the matrix A cannot be diagonalized by fast trigonometric transforms when the PSF is not quadrantally symmetric. Therefore, in this section we use a different strategy: the preconditioner is defined by a different PSF that leads to fast computations with an accurate deblurring model.

We consider a simple implementation of this strategy that can be useful when the PSF is experimentally measured. Indeed, in some applications, the PSF is nonsymmetric even if it is just a numerical perturbation of a Gaussian-like blur, cf. Example 1 and [60]. Recalling that for the reflective and antireflective BCs fast transforms (cosine and antireflective, respectivelly) can be implemented only in the quadrantally symmetric case, a quadrantally symmetric PSF $\widetilde{H}$ can be obtained from the original PSF H by defining

$$\widetilde{H}(i,j) = \frac{H(i,j)+H(-i,j)+H(i,-j)+H(-i,-j)}{4}, \qquad i,j = 1,\dots,n.$$

Note that $\widetilde{H}$ is the optimal Frobenius norm approximation of H in the set of quadrantally symmetric PSFs, see[81]. Therefore, we consider the matrix Q obtained imposing reflective BC to $\widetilde{H}$ when A is the BTTB or the rectangular matrix, while for A antireflective, Q is defined imposing antireflective BCs as well. In this way

$$P = (QQ^* + \alpha I)^{-1}$$

can be diagonalized by DCT or by antireflective transform and the MLBA becomes

Algorithm 3.

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + WA^*(QQ^* + \alpha I)^{-1}(\mathbf{g} - AW^*\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}). \end{cases} \tag{6.5.7}$$

In analogy to Algorithm 1, $\|(QQ^* + \alpha I)^{-1}\| < \|(AA^* + \alpha I)^{-1}\|$ is a sufficient condition to apply Theorem 6.4.3 with $\lambda = 1$. It is expected that

$$\|K^*(QQ^* + \alpha I)^{-1}K\| < 1$$

if α is large enough.

6.6 Approximated Tikhonov regularization instead of preconditioning

In this section we propose an approach to approximate $K^\dagger$ different from the use of the matrix P in (6.4.5) as suggested in [21]. Motivated by a very recent preconditioning proposal in [37, 41], we replace the whole matrix $K^\dagger$ by a regularized approximation obtained by C.

In [37] the authors suggest to solve the preconditioned linear system

$$ZA\mathbf{f} = Z\mathbf{g},$$

where Z is a regularized approximation of $K^\dagger$, by a Van Cittert iteration [33], instead to solve a preconditioned Landweber iteration. Unfortunately, ZA is not symmetric and the convergence analysis, based on the complex eigenvalues of ZA, is hard to generalize. Differently, the non-stationary preconditioned iteration proposed in [41] results in a similar iteration, but an elegant convergence analysis is provided under a minor approximation assumption.

For $P = (AA^* + \alpha I)^{-1}$, the correction term $K^*P(\mathbf{g} - K\mathbf{x}^n)$ in the MLBA (6.4.6) can be seen as the Tikhonov solution, with parameter α, of the error equation. In detail, $\mathbf{z}^{n+1}$ in the iteration (6.4.6) can be rewritten as

$$\mathbf{z}^{n+1} = \mathbf{z}^n + \mathbf{p}^n, \tag{6.6.1}$$

where

$$\begin{aligned}\mathbf{p}^n &= K^*P(\mathbf{g} - K\mathbf{x}^n)\\ &= K^*(KK^* + \alpha I)^{-1}(\mathbf{g} - K\mathbf{x}^n)\\ &= (K^*K + \alpha I)^{-1}K^*(\mathbf{g} - K\mathbf{x}^n),\end{aligned}$$

since $(AA^* + \alpha I)^{-1} = (KK^* + \alpha I)^{-1}$ and $K^*(KK^* + \alpha I)^{-1} = (K^*K + \alpha I)^{-1}K^*$. Note that the correction $\mathbf{p}^n$ is the solution of the Tikhonov problem

$$\min_{\mathbf{p}\in\mathbb{R}^s}\{\|K\mathbf{p} - \mathbf{r}^n\|^2 + \alpha\|\mathbf{p}\|^2\},$$

which is a regularized approximation of the error equation

$$K\mathbf{e}^n = \mathbf{r}^n \tag{6.6.2}$$

in the noise free case, i.e., $\delta = 0$, where $\mathbf{e}^n = \mathbf{x} - \mathbf{x}^n$ denotes the error at the current iteration.

In real applications $\delta \neq 0$ and so equation (6.6.2) is (only) correct up to the perturbation in the data. Taking this into account, one may as well consider instead of the error equation (6.6.2) the "model equation"

$$L\mathbf{e}^n = \mathbf{r}^n, \tag{6.6.3}$$

where L is an approximation of K, possibly tolerating a slightly larger misfit. Solving (6.6.3) by means of Tikhonov regularization, we find

$$\begin{aligned}\widetilde{\mathbf{p}}^n &= (L^*L + \alpha I)^{-1}L^*\mathbf{r}^n\\ &= WC^*(CC^* + \alpha I)^{-1}\mathbf{r}^n,\end{aligned}$$

where we have choosen

$$L = CW^*. \tag{6.6.4}$$

Using $\widetilde{\mathbf{p}}^n$ in (6.6.1) to replace $\mathbf{p}^n$ we obtain a new algorithm.

Algorithm 4.

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + WC^*(CC^* + \alpha I)^{-1}(\mathbf{g} - AW^*\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}). \end{cases} \tag{6.6.5}$$

As before, the matrix C is chosen as a BCCB in general or diagonalizable by DCT (i.e., the reflective BC matrix), when the PSF is quadrantally symmetric. Unfortunately, this preconditioning strategy cannot be applied to the rectangular matrix approach because, in such case, C should have the same size of A, but this condition prevents the possibility of computing $\widetilde{\mathbf{p}}^n$ by fast trigonometric transforms.

Remark 6.6.1. *The iteration* (6.6.5) *uses the preconditioned linear system*

$$WC^*(CC^* + \alpha I)^{-1}K\mathbf{x} = WC^*(CC^* + \alpha I)^{-1}\mathbf{g}, \tag{6.6.6}$$

to update an aproximation inspired by the linearized Bregman iteration (6.4.4)*, but without resorting to the normal equations.*

Clearly, Algorithm 4 is no longer a MLBA and so a different convergence analysis is required. Unfortunately, classical results for convex optimization cannot be applied since the coefficient matrix $WC^*(CC^* + \alpha I)^{-1}K$ in (6.6.6) is not symmetric positive definite. An alternative convergence proof could be very hard because also the complex analysis convergence in [37] cannot be easily combined with the Bregman splitting and soft-thresholding.

Therefore, accordingly to [41], we consider a nonstationary choice of α that allows to provide a convergence analysis of the resulting algorithm and avoid the a-priori choice of α. On the other hand, if a good estimation of α is available, then Algorithm 4 can provides better restorations and hence it is also considered in the numerical results in Section 6.7.

Assumption 6.6.2. *Let $A, C \in \mathbb{R}^{n^2 \times n^2}$ and $W \in \mathbb{R}^{n^2 \times s}$, $n^2 \leq s$, such that*

$$\|(C - A)\mathbf{v}\| \leq \rho \|A\mathbf{v}\|, \qquad \forall \mathbf{v} \in \mathbb{R}^{n^2}, \tag{6.6.7a}$$

and

$$\|CW^*(\mathbf{u} - S_\mu(\mathbf{u}))\| \leq \rho\delta, \qquad \forall \mathbf{u} \in \mathbb{R}^s, \tag{6.6.7b}$$

with a fixed $0 < \rho < 1/2$, where $\delta = \|\eta\|$ is the noise level.

The Assumption (6.6.7a) is the same spectral equivalence required in [41]. Let L be defined in (6.6.4), then equation (6.6.7a) translates into

$$\|(L - K)\mathbf{u}\| \leq \rho \|K\mathbf{u}\|, \qquad \forall \mathbf{u} \in \mathbb{R}^s. \tag{6.6.8}$$

Instead, the Assumption (6.6.7b) was not present in [41] and it is equivalent to consider the soft-threshold parameter μ as a continuous function with respect to the noise level δ, i.e., $\mu = \mu(\delta)$, and such that $\mu(\delta) \to 0$ as $\delta \to 0$. This is a common request in many soft-thresholding based methods, see for instance Theorem 4.1 in [35]. Nevertheless, in this Chapter we will not

concentrate on μ and will not give any specific δ-dependent rule to compute it. Let us just observe that Assumption (6.6.7b) can be restated in the equivalent way

$$\mu \leq \frac{\rho\delta}{|||CW^*|||},$$

where here we defined

$$|||K||| = \sup_{\|\mathbf{x}\|_\infty \equiv 1} \|K\mathbf{x}\|_2, \quad \text{with } K : (\mathbb{R}^s, \|\cdot\|_\infty) \to \left(\mathbb{R}^{n^2}, \|\cdot\|_2\right) \quad \text{and } \|\mathbf{x}\|_\infty = \max_{j=1,\cdots,s} \{|x_j|\}.$$

Indeed, it follows easily from the fact that $\|\mathbf{u} - S_\mu(\mathbf{u})\|_\infty \leq \mu$ for every $\mathbf{u} \in \mathbb{R}^s$.

Algorithm. 4–NS. *Let $\mathbf{z}^0$ be given and set $\mathbf{r}^0 = \mathbf{g} - KS_\mu(\mathbf{z}^0)$. Choose $\tau = \frac{1+2\rho}{1-2\rho}$ with ρ from* (6.6.7a)*, and fix $q \in (2\rho, 1)$.*

While $\|\mathbf{r}^n\| > \tau\delta$, let $\tau_n = \|\mathbf{r}^n\|/\delta$ and let α_n be such that

$$\alpha_n \|(CC^* + \alpha_n I)^{-1}\mathbf{r}^n\| = q_n \|\mathbf{r}^n\|, \qquad q_n = \max\{q, 2\rho + (1+\rho)/\tau_n\}, \tag{6.6.9a}$$

compute

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + WC^*(CC^* + \alpha_n I)^{-1}(\mathbf{g} - AW^*\mathbf{x}^n), \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}). \end{cases} \tag{6.6.9b}$$

Note that the iteration (6.6.9b) is the same of Algorithm 4 where a nonstationary α is chosen at every iteration according to (6.6.9a). In Corollary 6.6.6 we will prove that, if $\delta > 0$, then Algorithm 4–NS will terminate after $n = n_\delta \geq 0$ iterations with

$$\|\mathbf{r}^{n_\delta}\| \leq \tau\delta < \|\mathbf{r}^n\|, \qquad n = 0, 1, \cdots, n_\delta - 1, \tag{6.6.10}$$

which is the discrepancy principle (6.4.10) with $\tilde{n} = n_\delta$ and $\gamma = \tau = (1+2\rho)/(1-2\rho)$.

The parameter q in Algorithm 4, like in [41], is meant as a safeguard to prevent that the residual decreases too rapidly. Our theoretical results do not utilize this parameter.

Remark 6.6.3. *It is not difficult to see that there is a unique positive parameter α_n that satisfies* (6.6.9a)*. This parameter can be computed with a few step of an appropriate Newton scheme [48]. Accordingly, parameter α_n, and therefore Algorithm 4–NS, are well defined.*

We define

$$\mathbf{h}^n = L^*(LL^* + \alpha_n I)^{-1}(\mathbf{g} - K\mathbf{S}_\mu(\mathbf{z}^n)), \tag{6.6.11}$$

such that (6.6.9b) can be compactly rewritten as

$$\begin{cases} \mathbf{z}^{n+1} = \mathbf{z}^n + \mathbf{h}^n, \\ \mathbf{x}^{n+1} = \mathbf{S}_\mu(\mathbf{z}^{n+1}). \end{cases}$$

For the purpose of the subsequent convergence and regularization results, when $\delta > 0$, even if it will be always the case, we will highlight by the subscript δ (for instance $\{\mathbf{x}_\delta^n\}$) the sequences

generated by Algorithm 4–NS, starting from initial data $\mathbf{g}^\delta = A\mathbf{f} + \eta$ affected by noise, whereas we avoid the subscript (for instance $\{\mathbf{x}^n\}$) for the sequences generated starting from exact initial data $\mathbf{g} = A\mathbf{f}$, i.e., $\delta = 0$.

For the following analysis, instead of working with the error $\mathbf{e}_\delta^n = \mathbf{x} - \mathbf{x}_\delta^n$, it is useful to consider the partial error with respect to $\mathbf{z}_\delta^n$, namely

$$\tilde{\mathbf{e}}_\delta^n = \mathbf{x} - \mathbf{z}_\delta^n. \tag{6.6.12}$$

Proposition 6.6.4. *Assume that the assumptions* (6.6.7) *are satisfied for some* $0 < \rho < 1/2$. *If* $\|\mathbf{r}_\delta^n\| > \tau\delta$ *and we define* $\tau_n = \|\mathbf{r}_\delta^n\|/\delta$, *then it follows that*

$$\|\mathbf{r}_\delta^n - L\tilde{\mathbf{e}}_\delta^n\| \le \left(\rho + \frac{1+2\rho}{\tau_n}\right)\|\mathbf{r}_\delta^n\| < (1-\rho)\|\mathbf{r}_\delta^n\|, \tag{6.6.13}$$

where $\tilde{\mathbf{e}}^n$ *is defined in* (6.6.12).

Proof. In the free noise case we have $\mathbf{g} = K\mathbf{x}$. As a consequence

$$\begin{aligned}\mathbf{r}_\delta^n - L\mathbf{e}_\delta^n &= \mathbf{g}^\delta - K\mathbf{x}_\delta^n - L(\mathbf{x} - \mathbf{z}_\delta^n) + L\mathbf{x}_\delta^n - LS_\mu(\mathbf{z}_\delta^n)\\ &= \mathbf{g}^\delta - \mathbf{g} + (K-L)\mathbf{e}_\delta^n + L(\mathbf{z}_\delta^n - S_\mu(\mathbf{z}_\delta^n)).\end{aligned}$$

Using now assumptions (6.6.7), in particular (6.6.8), and $\|\mathbf{g}^\delta - \mathbf{g}\| \le \delta$, we derive the following estimate

$$\begin{aligned}\|\mathbf{r}_\delta^n - L\mathbf{e}_\delta^n\| &\le \|\mathbf{g}^\delta - \mathbf{g}\| + \|(K-L)\mathbf{e}_\delta^n\| + \|L(\mathbf{z}_\delta^n - S_\mu(\mathbf{z}_\delta^n))\|\\ &\le \|\mathbf{g}^\delta - \mathbf{g}\| + \rho\|K\mathbf{e}_\delta^n\| + \rho\delta\\ &\le \|\mathbf{g}^\delta - \mathbf{g}\| + \rho(\|\mathbf{r}_\delta^n\| + \|\mathbf{g}^\delta - \mathbf{g}\| + \delta)\\ &\le (1+2\rho)\delta + \rho\|\mathbf{r}_\delta^n\|.\end{aligned}$$

The first inequality in (6.6.13) now follows from the hypothesis $\delta = \|\mathbf{r}_\delta^n\|/\tau_n$. The second inequality follows from $\rho + \frac{1+2\rho}{\tau_n} < \rho + \frac{1+2\rho}{\tau}$. □

We are going to show that the sequence $\{\mathbf{x}_\delta^n\}$ approaches $\mathbf{x}$ as $\delta \to 0$. The proof combines Proposition 6.6.4 with suitable modifications of the results in [41].

Proposition 6.6.5. *Let* $\tilde{\mathbf{e}}_\delta^n$ *be defined in* (6.6.12). *If the assumptions* (6.6.7) *are satisfied, then* $\|\tilde{\mathbf{e}}_\delta^n\|$ *of Algorithm 4–NS decreases monotonically for* $n = 0, 1, \ldots, n_\delta - 1$. *In particular, we deduce*

$$\|\tilde{\mathbf{e}}_\delta^n\|^2 - \|\tilde{\mathbf{e}}_\delta^{n+1}\|^2 \ge \frac{8\rho^2}{1+2\rho}\|(CC^* + \alpha_n)^{-1}\mathbf{r}_\delta^n\|\|\mathbf{r}_\delta^n\|. \tag{6.6.14}$$

Proof. We have

$$\begin{aligned}
\|\tilde{\mathbf{e}}_\delta^n\|^2 - \|\tilde{\mathbf{e}}_\delta^{n+1}\|^2 &= 2\langle \tilde{\mathbf{e}}_\delta^n, \mathbf{h}^n\rangle - \|\mathbf{h}^n\|^2 \\
&= 2\langle L\tilde{\mathbf{e}}_\delta^n, (CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\rangle - \langle \mathbf{r}_\delta^n, CC^*(CC^* + \alpha_n I)^{-2}\mathbf{r}_\delta^n\rangle \\
&= 2\langle \mathbf{r}_\delta^n, (CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\rangle - \langle \mathbf{r}_\delta^n, CC^*(CC^* + \alpha_n I)^{-2}\mathbf{r}_\delta^n\rangle \\
&\quad - 2\langle \mathbf{r}_\delta^n - L\tilde{\mathbf{e}}_\delta^n, (CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\rangle \\
&\geq 2\langle \mathbf{r}_\delta^n, (CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\rangle - 2\langle \mathbf{r}_\delta^n, CC^*(CC^* + \alpha_n I)^{-2}\mathbf{r}_\delta^n\rangle \\
&\quad - 2\langle \mathbf{r}_\delta^n - L\tilde{\mathbf{e}}_\delta^n, (CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\rangle \\
&= 2\alpha_n\langle \mathbf{r}_\delta^n, (CC^* + \alpha_n I)^{-2}\mathbf{r}_\delta^n\rangle - 2\langle \mathbf{r}_\delta^n - L\tilde{\mathbf{e}}_\delta^n, (CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\rangle \\
&\geq 2\alpha_n\langle \mathbf{r}_\delta^n, (CC^* + \alpha_n I)^{-2}\mathbf{r}_\delta^n\rangle - 2\|\mathbf{r}_\delta^n - L\tilde{\mathbf{e}}_\delta^n\|\|(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\| \\
&= 2\|(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\|\left(\|\alpha_n(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\| - \|\mathbf{r}_\delta^n - L\tilde{\mathbf{e}}_\delta^n\|\right) \\
&\geq 2\|(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\|\left(q_n\|\mathbf{r}_\delta^n\| - \left(\rho + \frac{1+2\rho}{\tau_n}\right)\|\mathbf{r}_\delta^n\|\right) \\
&\geq \frac{8\rho^2}{1+2\rho}\|(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\|\|\mathbf{r}_\delta^n\|,
\end{aligned}$$

where the relevant inequalities are a consequence of equation (6.6.9a) and Proposition 6.6.4. The last inequality follows from (6.6.9a) and $\tau_n > \tau = (1+2\rho)/(1-2\rho)$ for $\|\mathbf{r}_\delta^n\| > \tau\delta$. □

Corollary 6.6.6. *Under the assumptions* (6.6.7)*, there holds*

$$\|\tilde{\mathbf{e}}_\delta^0\| \geq \frac{8\rho^2}{1+2\rho}\sum_{n=0}^{n_\delta-1}\|(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\|\|\mathbf{r}_\delta^n\| \geq c\sum_{n=0}^{n_\delta-1}\|\mathbf{r}_\delta^n\|^2 \tag{6.6.15}$$

for some constant $c > 0$, depending only on ρ and q in (6.6.9a).

Proof. The following proof is almost the same as Corollary 3 in [41], but we include it to make the Chapter self contained.

The first inequality follows by taking the sum of the quantities in (6.6.14) from $n = 0$ up to $n = n_\delta - 1$.

For the second inequality, note that for every $\alpha > \frac{q_n\|C\|^2}{1-q_n}$ and every $\sigma \in \sigma(C) \subset [0, \|C\|^2]$, with $\sigma(C)$ being the spectrum of C, we have

$$\frac{\alpha}{\sigma^2 + \alpha} \geq \frac{\alpha}{\|C\|^2 + \alpha} = (1 + \|C\|^2/\alpha)^{-1} > q_n,$$

and hence,

$$\alpha\|(CC^* + \alpha I)^{-1}\mathbf{r}_\delta^n\| > q_n\|\mathbf{r}_\delta^n\|,$$

as $\|\mathbf{r}_\delta^n\| > 0$ for $n < n_\delta$. This implies that α_n in (6.6.9a) satisfies $0 < \alpha_n \leq \frac{q_n\|C\|^2}{1-q_n}$, thus

$$\|(CC^* + \alpha_n I)^{-1}\mathbf{r}_\delta^n\| = \frac{q_n}{\alpha_n}\|\mathbf{r}_\delta^n\| \geq \frac{(1-q_n)}{\|C\|^2}\|\mathbf{r}_\delta^n\|.$$

Now, according to the choice of parameters in Algorithm 4-NS, we deduce
$1-q_n = \min\{1-q, 1-2\rho-(1+\rho)/\tau_n\}$, and

$$1-2\rho-(1+\rho)/\tau_n = \frac{1+2\rho}{\tau} - \frac{1+\rho}{\tau_n} > \frac{1+2\rho}{\tau} - \frac{1+\rho}{\tau} = \frac{\rho}{\tau}.$$

Therefore, there exists $c>0$, depending only on ρ and q such that $1-q_n \geq c\|C\|^2\left(\frac{8\rho^2}{1+2\rho}\right)^{-1}$, and

$$\|(CC^*+\alpha_n I)^{-1}\mathbf{r}_\delta^n\| \geq c\left(\frac{8\rho^2}{1+2\rho}\right)^{-1}\|\mathbf{r}_\delta^n\| \qquad \text{for } n=0,1,\cdots,n_\delta-1.$$

Now the second inequality follows immediately. □

From the outer inequality of (6.6.15) it can be seen that the sum of the squares of the residual norms is bounded, and hence, if $\delta>0$, there must be a first integer $n_\delta<\infty$ such that (6.6.10) is fulfilled, i.e., Algorithm 4-NS terminates after finitely many iterations.

Remark 6.6.7. *Recalling that the soft-threshold parameter μ is taken as a continuous function with respect to the noise level δ such that $\mu(\delta)\to 0$ as $\delta\to 0$, then the operator $\mathbf{g}\mapsto\mathbf{z}^n$ is continuous for every fixed n.*

In the next theorem we are going to give a convergence and regularity result.

Theorem 6.6.8. *Assume that $\mathbf{z}^0$ is not a solution of the linear system*

$$\mathbf{g} = AW^*\mathbf{x}, \tag{6.6.16}$$

and that δ_m is a sequence of positive real numbers such that $\delta_m\to 0$ as $m\to\infty$. Then, if Assumption 6.6.2 is valid, the sequence $\{\mathbf{x}_{\delta_m}^{n(\delta_m)}\}_{m\in\mathbb{N}}$, generated by the discrepancy principle rule (6.6.10)*, converges as $m\to\infty$ to the solution of* (6.6.16) *which is closest to $\mathbf{z}^0$ in Euclidean norm.*

Proof. We are going to show convergence for the sequence $\{\mathbf{z}_{\delta_m}^{n(\delta_m)}\}_{m\in\mathbb{N}}$ and then the thesis will follow easily from the continuity of $S_{\mu(\delta)}$ and Remark 6.6.7, i.e.,

$$\lim_{m\to\infty}\mathbf{x}_{\delta_m}^{n(\delta_m)} = \lim_{m\to\infty} S_{\mu(\delta_m)}(\mathbf{z}_{\delta_m}^{n(\delta_m)}) = S_{\lim_{m\to\infty}\mu(\delta_m)}(\lim_{m\to\infty}\mathbf{z}_{\delta_m}^{n(\delta_m)}) = \lim_{m\to\infty}\mathbf{z}_{\delta_m}^{n(\delta_m)}.$$

The proof of the convergence for the sequence $\{\mathbf{z}_{\delta_m}^{n(\delta_m)}\}$ can be divided into two steps: at step one, we show the convergence in the free noise case $\delta=0$. In particular, the sequence $\{\mathbf{z}^n\}$ converges to a solution of (6.6.16) that is the closest to $\mathbf{z}^0$. At the second step, we show that given a sequence of positive real numbers $\delta_m\to 0$ as $m\to\infty$, then we get a corresponding sequence $\{\mathbf{z}_{\delta_m}^{n(\delta_m)}\}$ converging as $m\to\infty$.

Concerning the first step of the proof, we will not give details since it can be just copied from [41][Theorem 4]. Indeed, if $\delta=0$, from Remark 6.6.7 it follows that $\mathbf{r}_\delta^n=\mathbf{r}^n$, and the sequence $\{\mathbf{z}^n\}$ coincides with the one generated by algorithm 1 in [41]. We just say that the

main ingredients are the convergence of the sequence $\|\mathbf{e}^n\|$ granted by Proposition 6.6.5 and the convergence to 0 of the sequence $\|\mathbf{r}^n\|\|(CC^* + \alpha_n I)^{-1}\mathbf{r}^n\|$, since general term of a converging series from Corollary 6.6.6. Moreover, in the free noise case the sequence $\{\mathbf{z}^n\}$ will not stop, i.e., $n \to \infty$, since the discrepancy principle will not be satisfied by any n, in particular $n_\delta \to \infty$ for $\delta \to 0$.

Hence, let $\mathbf{x}$ be the converging point of the sequence $\{\mathbf{z}^n\}$ and let $\delta_m > 0$ be a sequence of positive real numbers converging to 0. For every δ_m, let $n = n(\delta_m)$ be the first positive integer such that (6.6.10) is satisfied, whose existence is granted by Corollary 6.6.6, and let $\{\mathbf{z}_{\delta_m}^{n(\delta_m)}\}$ be the corresponding sequence. For every fixed $\varepsilon > 0$, there exists $\overline{n} = \overline{n}(\varepsilon)$ such that

$$\|\mathbf{x} - \mathbf{z}^n\| \leq \varepsilon/2 \qquad \text{for every } n > \overline{n}(\varepsilon), \tag{6.6.17}$$

and there exists $\overline{\delta} = \overline{\delta}(\varepsilon)$ for which

$$\|\mathbf{z}^{\overline{n}} - \mathbf{z}_\delta^{\overline{n}}\| \leq \varepsilon/2 \qquad \text{for every } 0 < \delta < \overline{\delta}, \tag{6.6.18}$$

due to the continuity of the operator $\mathbf{g} \mapsto \mathbf{z}^n$ for every fixed n, see Remark 6.6.7. Therefore, let us choose $\overline{m} = \overline{m}(\varepsilon)$ large enough such that $\delta_m < \overline{\delta}$ and such that $n(\delta_m) > \overline{n}$ for every $m > \overline{m}$. Such $\overline{m}$ does exists since $\delta_m \to 0$ and $n_\delta \to \infty$ for $\delta \to 0$. Hence, for every $m > \overline{m}$, we have

$$\begin{aligned}
\|\mathbf{x} - \mathbf{z}_{\delta_m}^{n(\delta_m)}\| &= \|\tilde{\mathbf{e}}_{\delta_m}^{n(\delta_m)}\| \\
&\leq \|\tilde{\mathbf{e}}_{\delta_m}^{\overline{n}}\| \\
&= \|\mathbf{x} - \mathbf{z}_{\delta_m}^{\overline{n}}\| \\
&\leq \|\mathbf{x} - \mathbf{z}^{\overline{n}}\| + \|\mathbf{z}^{\overline{n}} - \mathbf{z}_{\delta_m}^{\overline{n}}\| \leq \varepsilon,
\end{aligned}$$

where the first inequality comes from Proposition 6.6.5 and the last one from (6.6.17) and (6.6.18). □

6.7 Numerical results

In this section, we will show the numerical results for image deblurring using our proposed Algorithms 1–4. We compare them with some available deblurring algorithms, which implement a proper treatment of the boundary artifacts. In particular we consider two of the algorithms proposed in [3], namely FA-MD for the Frame-based analysis model and TV-MD for the Total Variation model, and the Algorithm [6] called here FTVd since, in the case of nonsymmetric PSF, it reduces to an implementation of the algorithm in [114] with the trick described in [89]. The codes of the previous algorithms are available at the web-page of the authors and we use the default parameters and stop conditions. The regularization parameter is chosen by hand in order to provide the best restoration (see the following discussion). All the images are in grayscale intensity with values range in $[0,1]$, where 0 is black and 1 is white.

Our tests were done by using MATLAB 7.11.0 (R2010b) with floating-point precision about $2.22 \cdot 10^{-16}$ on a Lenovo laptop with Intel(R) Core(TM) i2 CPU 2.20 GHz and 2 GB memory.

Assuming that the noise level is available or easily estimated, we stop all Algorithms 1–4 using the discrepancy principle (6.4.10) with $\gamma = 10^{-15}$. Algorithm 4–NS is stopped according to the modified discrepancy principle (6.6.10). Moreover, for the Algorithm 4–NS we set

$$q = 0.5 \qquad \text{and} \qquad \rho = 10^{-4}.$$

Therefore, q and ρ do not need to be estimated.

The accuracy of the solution is measured by the PSNR value, which is defined as

$$\text{PSNR} = 20\log_{10}\frac{255\cdot n}{\|\mathbf{f}-\tilde{\mathbf{f}}\|},$$

with $\mathbf{f}$ and $\tilde{\mathbf{f}}$ being the original and the restored images in the FOV, respectively. The initial guess of each algorithm is set to be the zero vector.

To estimate μ and α, since they are mutually dependent and they are related to the preconditioner, we fix a possible μ (usually the results are not very sensible varying μ if α is properly chosen) and then the optimum α, which gives the largest PSNR, is chosen by trial and error. Possible strategies to estimate α will be investigated in future works. Only for Algorithm4–NS we pay a slightly more attention in the choice of μ since this is the only parameter of the method. Similarly, for all the other methods considered for comparison, the regularization parameter is chosen by trial and error, as the one leading to the largest PSNR.

We take only the more appropriate BCs for each example. In particular, if the image has a black background, like in astronomical imaging, we consider zero BCs, while when the image is a generic picture we use antireflective BCs. In the following the "Algorithm x" is denoted by "Alg-BCx" and "Alg-Rect$x$", when A is obtained imposing BCs or is the rectangular matrix, respectively. We recall that Algorithm 4 is available only for the BC approach.

Finally, the last remark. Some of the PSFs look small (e.g. Figure 6.5 (b) and Figure 6.6 (b)), but we note that both PSFs have not a small support. Pixels far from the center of the PSF are very small, hence they look black in the image, but greater than zero.

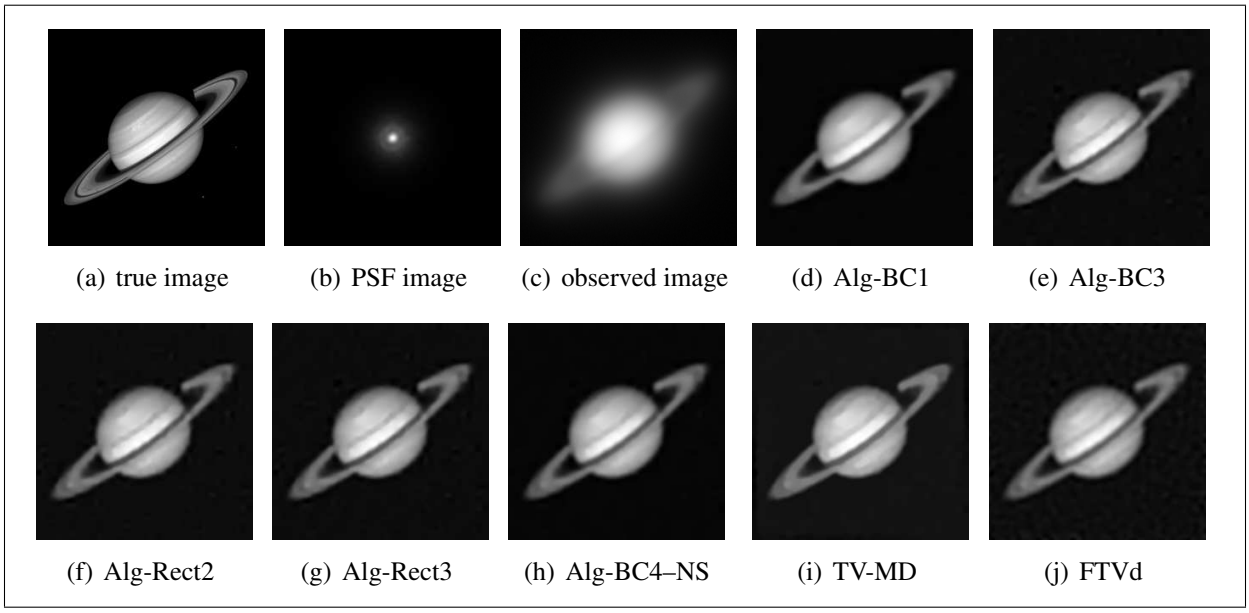
6.7.1 Linear B-spline framelets

The tight-frame used in our tests is the piecewise linear B-spline framelets given in [21]. Namely, given the masks

$$b_0 = \frac{1}{4}\begin{bmatrix}1 & 2 & 1\end{bmatrix}, \quad b_1 = \frac{\sqrt{2}}{4}\begin{bmatrix}1 & 0 & -1\end{bmatrix}, \qquad b_2 = \frac{1}{4}\begin{bmatrix}-1 & 2 & -1\end{bmatrix},$$

we define the 1D filters of size $n\times n$ by imposing reflective BCs

$$B_0 = \frac{1}{4}\begin{bmatrix} 3 & 1 & 0 & \dots & 0 \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ 0 & \dots & 0 & 1 & 3 \end{bmatrix}, \quad B_1 = \frac{1}{4}\begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix},$$



(a) true image (b) PSF image (c) observed image (d) Alg-BC1 (e) Alg-BC3

(f) Alg-Rect2 (g) Alg-Rect3 (h) Alg-BC4–NS (i) TV-MD (j) FTVd

Figure 6.3: Example 1: true image, PSF, observed image and restored images.

and

$$B_2 = \frac{1}{4}\begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

The nine 2D filters are obtained by

$$B_{i,j} = B_i \otimes B_j, \qquad i,j = 0,1,2,$$

where $\otimes$ denotes the tensor product operator. Finally, the corresponding tight-frame analysis operator is

$$W = \begin{bmatrix} B_{0,0} \\ B_{0,1} \\ \vdots \\ B_{2,2} \end{bmatrix}.$$

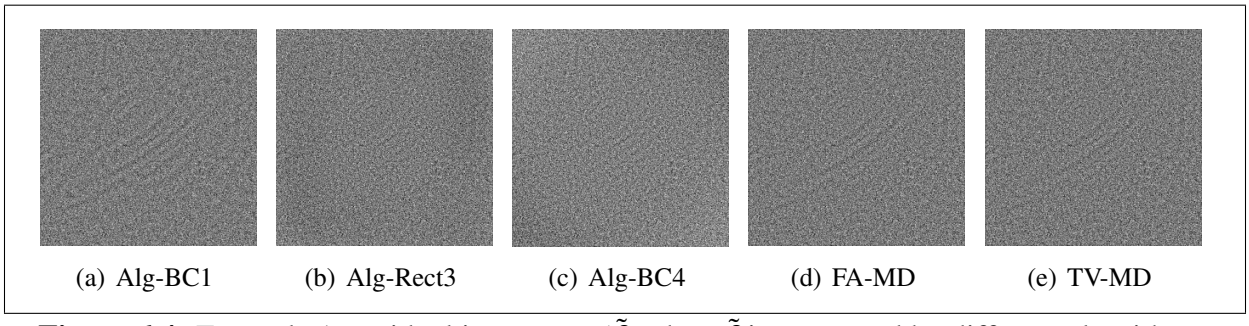
Throughout the experiments, the level of the framelet decomposition is 4 like in [21] and the level of wavelet decomposition is the one used in FA–MD.

6.7.2 Example 1: Saturn image

The first example is 256×256 Saturn image in Figure 6.3 (a) while the astronomical PSF is taken from the "satellite" test problem in [80] Figure 6.3 (b). We add a 1% of Gaussian white noise to obtain the observed image in Figure 6.3 (c). We assume zero BCs.

Algorithm	PSNR	Iter.	CPU time(s)	Regular. parameter
Alg-BC1	30.97	322	200.99	$\alpha = 0.045$
Alg-BC2	31.60	9	18.18	$\alpha = 0.0004$
Alg-BC3	31.56	10	7.07	$\alpha = 0.0005$
Alg-Rect1	30.95	493	948.94	$\alpha = 0.07$
Alg-Rect2	31.62	8	22.20	$\alpha = 0.0003$
Alg-Rect3	**31.61**	7	**13.50**	$\alpha = 0.0003$
Alg-BC4	31.49	29	16.56	$\alpha = 0.0018$
Alg-BC4–NS	31.25	15	10.32	$\mu = 6$
FA-MD	30.87		90.85	$\lambda = 0.001$
TV-MD	31.17		47.61	$\lambda = 0.01$
FTVd	30.50		1.75	$1/\alpha = 0.0013$

Table 6.2: Example 1: PSNR, number of iterations, and CPU time in seconds for the best regularization parameter (maximum PSNR) reported in the last column. For our algorithms $\mu = 10$ except for Alg-BC4–NS.



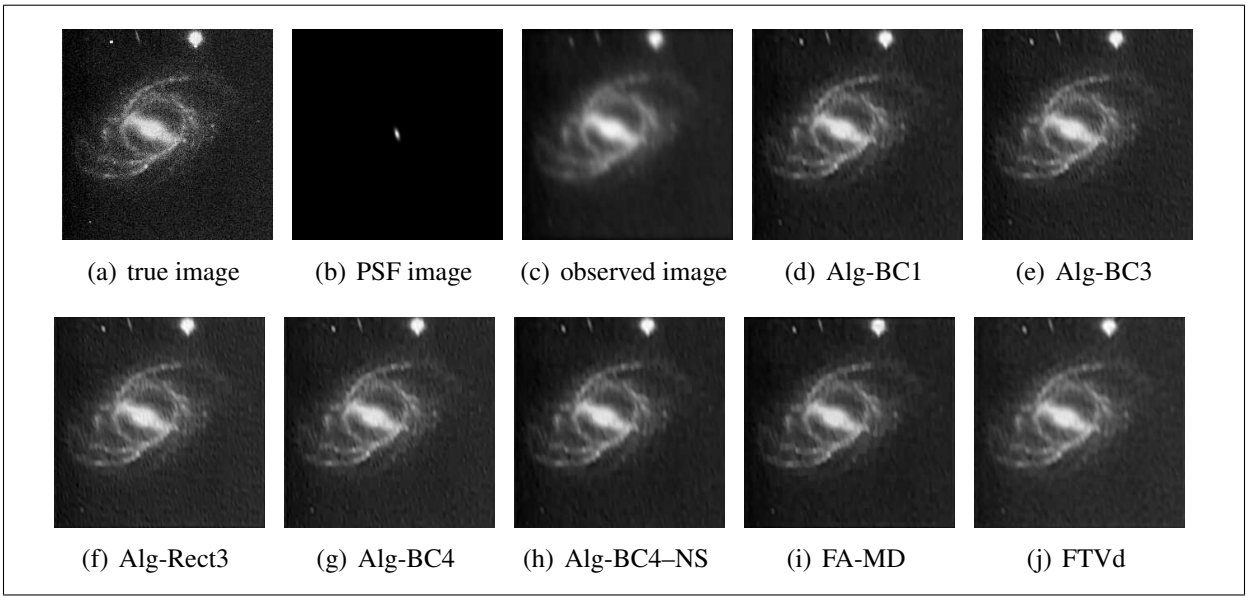
(a) Alg-BC1 (b) Alg-Rect3 (c) Alg-BC4 (d) FA-MD (e) TV-MD

Figure 6.4: Example 1: residual image $\mathbf{g} - A\tilde{\mathbf{f}}$, where $\tilde{\mathbf{f}}$ is computed by different algorithms.

Note that Alg-BC3 and Alg-Rect3 use the DCT for the preconditioner, while Alg-BC4 like Alg-BC1 and Alg-Rect1 use FFT, since the PSF is not quadrantally symmetric.

Table 6.2 reports the PSNR and the CPU time for the different algorithms. Note that Algorithm 1 provides a poor and time consuming restoration. Moreover, it requires a larger value of the parameter α with respect to the algorithms 2–4, which is necessary to satisfy condition (6.5.2) and assure the convergence. The algortihms 2 and 3 have the largest PSNR with reasonable CPU time, in particular Alg-Rect3 seems to be a good choice and Alg-BC3 gives a comparable restoration in about half time. Algorithm 4 gives a slightly lower PSNR even if the computed restorations are better than Algorithm 1 and the other algorithms from the literature, keeping also a low CPU time.

The algorithms in [3] (FA-MD and TV-MD) in this example lead to a larger CPU time, while the FTVd is very fast but the computed restoration is the worst. Figure 6.3 shows the corresponding restored images. To test the quality of the restorations, Figure 6.4 shows the residual images defined as $\mathbf{g} - A\tilde{\mathbf{f}}$, where $\tilde{\mathbf{f}}$ is the restored image.



(a) true image (b) PSF image (c) observed image (d) Alg-BC1 (e) Alg-BC3

(f) Alg-Rect3 (g) Alg-BC4 (h) Alg-BC4–NS (i) FA-MD (j) FTVd

Figure 6.5: Example 2: true image, PSF, observed image and restored images.

6.7.3 Example 2: Galaxy image

We consider another astronomical example with the 256×256 image in Figure 6.5 (a) corrupted by oblique Gaussian blur taken from the "GaussianBlur422" test problem in [80], see Figure 6.5 (b). We add a 2% of Gaussian white noise to obtain the observed image in Figure 6.5 (c). We impose zero BCs and the computational properties of the different algorithms are the same as in Example 1.

Table 6.3 shows that Alg-BC4 is the best algorithm, since it obtains about the same PSNR of Algorithm 2, but with about 1/4 of the CPU time. The variant Alg-BC4–NS with a nonstationary choice of the preconditioner results to be very effective and comparable with the other algorithms based on the BC model avoiding the choice of parameter α. Differently, the rectangular approach gives a slightly lower PSNR with a larger CPU time. Concerning the other methods, the same observations reported for Example 1 still apply. Figure 6.5 shows some of the corresponding restored images.

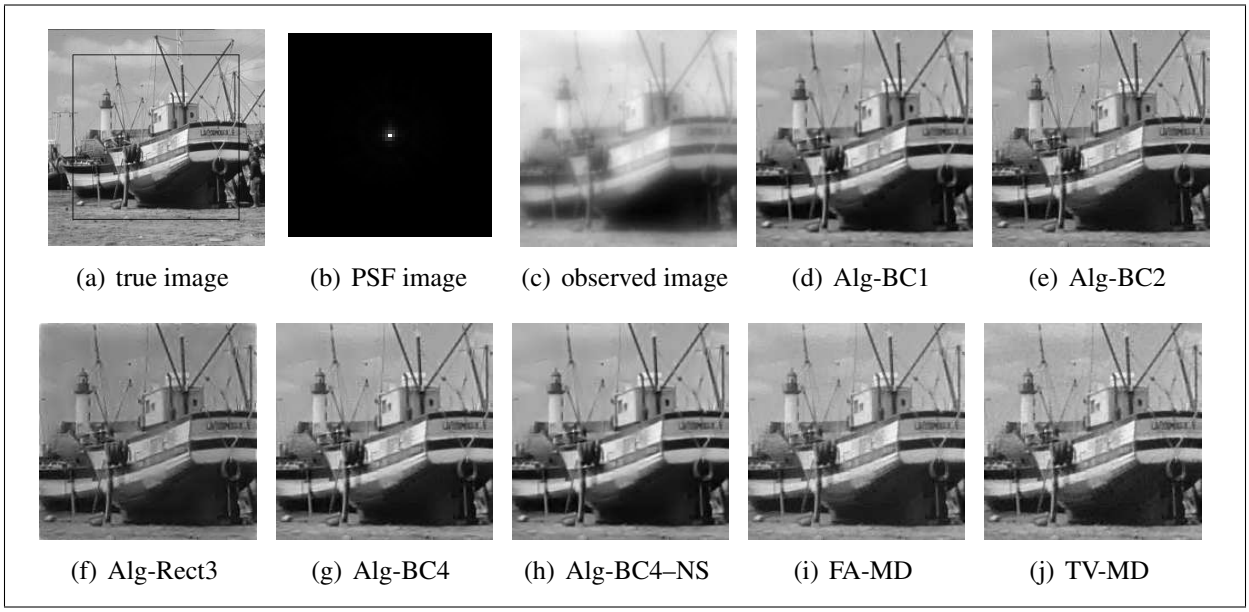
6.7.4 Example 3: Boat image

To set up a scenario of unknown boundaries, the observed image of size 196×196 is obtained convolving the full (256×256) image by the nonsymmetric 61×61 PSF in Figure 6.6(b), using arbitrary BCs (periodic, for computational convenience) and then keeping only the pixels in the FOV 196×196 (i.e., those not depending on the BCs). The FOV is denoted by a black box in the true image in Figure 6.6 (a). 1% of white Gaussian noise is added to the 196×196 blurred image.

We impose antireflective BCs owing to the generic structure of that picture. Hence the preconditioner in Alg-BC3 is diagonalized by the antireflective transform, according to the structure

Algorithm	PSNR	Iter.	CPU time(s)	Regular. parameter
Alg-BC1	25.02	21	16.57	$\alpha = 0.04$
Alg-BC2	25.07	12	22.88	$\alpha = 0.008$
Alg-BC3	25.05	27	17.58	$\alpha = 0.02$
Alg-Rect1	24.91	25	42.14	$\alpha = 0.06$
Alg-Rect2	24.98	12	28.21	$\alpha = 0.007$
Alg-Rect3	24.96	21	36.18	$\alpha = 0.01$
Alg-BC4	**25.06**	12	**6.21**	$\alpha = 0.008$
Alg-BC4–NS	25.01	29	17.10	$\mu = 4$
FA-MD	24.50		77.05	$\lambda = 0.02$
TV-MD	24.55		68.91	$\lambda = 0.09$
FTVd	24.62		1.51	$1/\alpha = 0.027$

Table 6.3: Example 2: PSNR and CPU time for the best regularization parameter (maximum PSNR). For our algorithms $\mu = 10$ except for Alg-BC4–NS.



(a) true image (b) PSF image (c) observed image (d) Alg-BC1 (e) Alg-BC2

(f) Alg-Rect3 (g) Alg-BC4 (h) Alg-BC4–NS (i) FA-MD (j) TV-MD

Figure 6.6: Example 3: true image, PSF, observed image and restored images.

Algorithm	PSNR	Iter.	CPU time(s)	Regular. parameter
Alg-BC1	29.43	97	34.26	$\mu = 20, \quad \alpha = 0.37$
Alg-BC2	30.11	11	17.21	$\mu = 20, \quad \alpha = 0.025$
Alg-BC3	30.09	10	4.03	$\mu = 20, \quad \alpha = 0.022$
Alg-Rect1	27.19	74	32.63	$\mu = 200, \quad \alpha = 0.03$
Alg-Rect2	27.10	74	45.95	$\mu = 200, \quad \alpha = 0.03$
Alg-Rect3	27.22	74	33.14	$\mu = 200, \quad \alpha = 0.03$
Alg-BC4	**30.17**	13	**3.67**	$\mu = 20, \quad \alpha = 0.03$
Alg-BC4–NS	29.77	60	19.57	$\mu = 30$
FA-MD	29.61		15.95	$\lambda = 0.04$
TV-MD	29.87		16.74	$\lambda = 0.1$
FTVd	28.95		0.73	$1/\alpha = 0.0069$

Table 6.4: Example 3: PSNR and CPU time for the best regularization parameter (maximum PSNR).

of the matrix A. Alg-Rect3 uses the DCT as usual, while Alg-BC4 like Alg-BC1 and Alg-Rect1 use FFT since the PSF is not quadrantally symmetric.

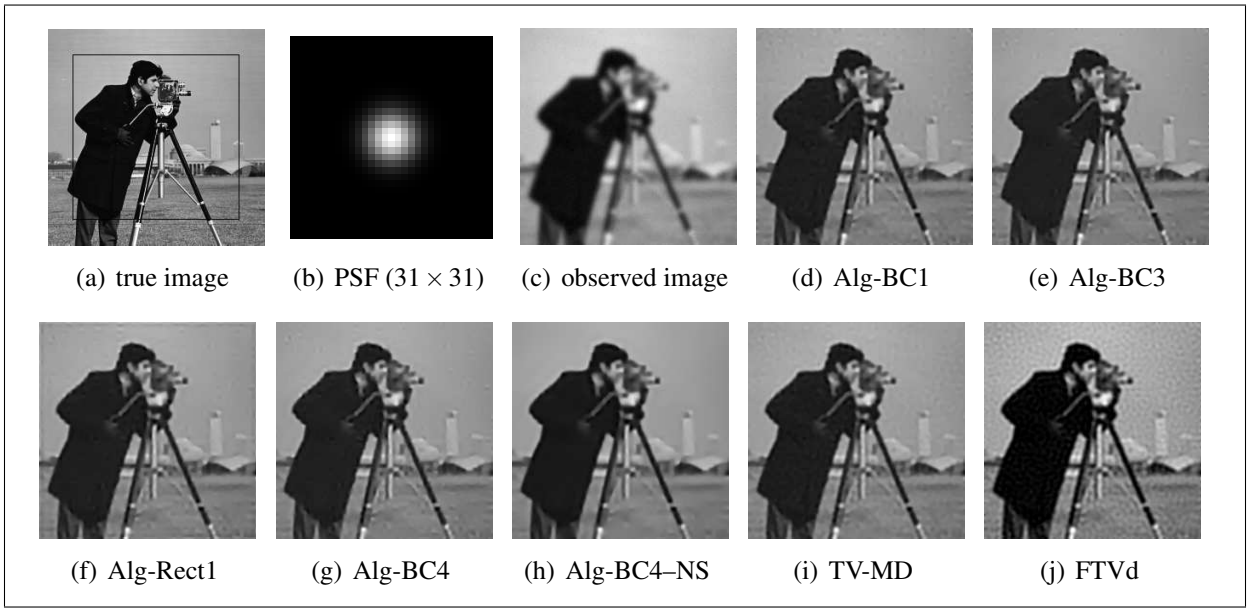
Table 6.4 shows the PSNR and the CPU time for the best restorations shown in Figure 6.6. Note that our algorithms with the rectangular matrix are less effective than the antireflective BC approach, leading to a lower PSNR. To obtain reasonable restorations in the rectangular case we need a large μ and so we take a different μ for the two deblurring models (BC and rectangular matrix). The best algorithm results to be Alg-BC4, since it combines a good restoration with a low CPU time.

6.7.5 Example 4: Cameraman with Gaussian blur

In this example we consider the classical Cameraman image 256×256 distorted by a 31×31 Gaussian blur with standard deviation 2.5 and a 2% of white Gaussian noise (see Figure 6.7). The size of the observed image is 226×226 according to the support of the PSF.

The PSF is quadrantally symmetric and hence, following our discussion in Section 6.5.1, Algorithm 1 is implemented using the DCT instead of FFT. Clearly $\widetilde{H} = H$ and so Alg-Rect3 reduces to Alg-Rect1 (they are really the same algorithm). We impose antireflective BCs and hence Alg-BC3 is the standard MLBA (6.4.9) with $P = (AA^* + \alpha I)^{-1}$, but recalling that we are using the reblurring approach where A^* is replaced by A'. On the other hand Alg-BC2 is no longer useful since the matrix-vector product with the matrix $P = (AA^* + \alpha I)^{-1}$ can be computed by two antireflective transforms without requiring the PCG: in fact, the use of preconditioning would represent an unnecessary increase of the CPU time without increasing the PSNR with respect to Alg-BC3. Finally, Alg-BC4 is implemented by DCT like Alg-BC1.

Table 6.5 shows that all the compared methods in this example provide comparable results. Nevertheless, it is interesting to observe that Alg-BC4 gives a slightly better restoration with a lower CPU time than the standard MLBA, i.e., Alg-BC3. TV-MD computes again a comparable restoration, but with more than a double CPU time. Figure 6.7 shows the restored images.



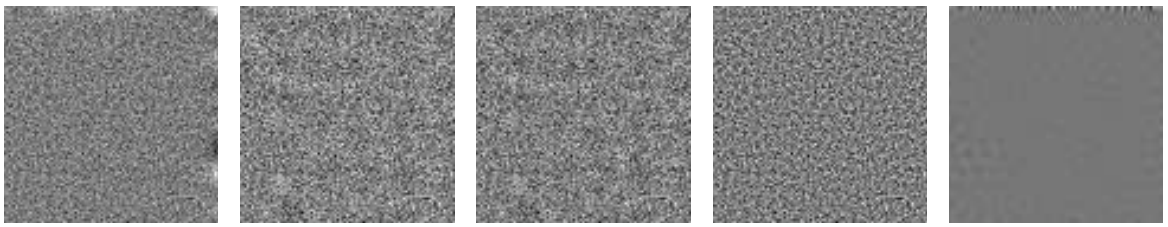
(a) true image (b) PSF (31×31) (c) observed image (d) Alg-BC1 (e) Alg-BC3

(f) Alg-Rect1 (g) Alg-BC4 (h) Alg-BC4–NS (i) TV-MD (j) FTVd

Figure 6.7: Example 4: true image, PSF, observed image and restored images.

Algorithm	PSNR	Iter.	CPU time(s)	Regular. parameter
Alg-BC1	23.67	51	20.19	$\alpha = 0.05$
Alg-BC3	23.74	17	6.03	$\alpha = 0.01$
Alg-Rect1	23.59	24	12.44	$\alpha = 0.02$
Alg-Rect2	23.64	17	16.49	$\alpha = 0.009$
Alg-BC4	**23.76**	17	**5.60**	$\alpha = 0.01$
Alg-BC4–NS	23.53	39	14.78	$\mu = 40$
FA-MD	23.44		14.10	$\lambda = 0.01$
TV-MD	23.55		13.37	$\lambda = 0.11$
FTVd	23.10		0.89	$1/\alpha = 0.0088$

Table 6.5: Example 4: PSNR and CPU time for the best regularization parameter (maximum PSNR). For our algorithms $\mu = 40$.



(a) Alg-BC1 (b) Alg-BC4 (c) Alg-BC4–NS (d) TV-MD (e) FTVd

Figure 6.8: Example 4: North-East corner of the residual images.

To test the ability of the different algorithms in dealing with the boundary effects, Figure 6.8 shows the North-East corner of the residual images. We can see that Alg-BC1 and FTVd have some ringing effects at the boundary (in Figure 6.8 (e) the apparent constant error in the central area is due to the grayscale rescaling), while Algorithm 4 and TV-MD do not show any particular distortion at the boundary.

6.8 Conclusions, open problems and further comments

In this Chapter, we have investigated several regularization preconditioning strategies for the MLBA applied to the synthesis approach with accurate restoration models, for image deblurring and unknown boundaries. Our numerical results show that Alg-BC4, which combines the favorite BCs (depending on the problem) with an approximated Tikhonov regularization preconditioner, represents a robust and effective algorithm. Indeed, it provides accurate restorations in all our examples with a reduced CPU time, also in comparison to the state of the art algorithms [3, 6] and the standard MLBA when available, cf. Example 4.

We have investigated only the synthesis approach, but the same preconditioning strategies can be applied to the analysis and the balanced approach [22, 99] as well. Moreover, possible future investigations could consider the use of a preconditioner obtained by a small rank approximation of the PSF as in [70], strategies for estimating the parameter α, and nonstationary sequences to approximate the best α avoiding its estimation as done in [68].

Conclusions

Proper studies on several different topics were carried on and presented in this work, which from the beginning was not intended to be a wholly comprehensive treatment of just a specific subject. Many results were provided as much many other open problems and possible future developments arose, which we have already extensively talked about at the end of every preceding chapter. Nevertheless, with reference to possible future works, in particular a fascinating interlacing problem is given by some studies in the '70's regarding plasma diffusion inside toroidal reactors, see [82, 10]. The strict interplay between the geometry of the space and the time extinction of FDE solutions, which was highlighted in Chapter 3.4.2, calls out for an in depth study on the natural inverse problems which come from it and that will be investigated in a near future.

Bibliography

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical table*, volume 55. Courier Corporation.

[2] Adams, R. and Fournier, J. (2003). *Sobolev spaces*, volume 140. Academic press.

[3] Almeida, M. S. C. and Figueiredo, M. A. T. (2013). Deconvolving images with unknown boundaries using the alternating direction method of multipliers. *IEEE Trans. Image Process.*, 22:3074–3086.

[4] Aricò, A., Donatelli, M., and Serra-Capizzano, S. (2008). Spectral analysis of the anti-reflective algebra. *Linear Algebra Appl.*, 428:657–675.

[5] Aronson, D. G. and Caffarelli, L. A. (1983). The initial trace of a solution of the porous medium equation. *Trans. Amer. Math. Soc.*, 280(1):351–366.

[6] Bai, Z. J., Cassani, D., Donatelli, M., and Serra-Capizzano, S. (2014). A fast alternating minimization algorithm for total variation deblurring without boundary artifacts. *J. Math. Anal. Appl.*, 415:373–393.

[7] Bénilan, P. (1983). A strong regularity lp for solution of the porous media equation. *Res. Notes Math.*, 89:39–58.

[8] Bénilan, P., Crandall, M. G., and Michel, P. (1982). Solutions of the porous medium equation in r (n) under optimal conditions on initial values. Technical report, DTIC Document.

[9] Berryman, J. G. (1977). Evolution of a stable profile for a class of nonlinear diffusion equations with fixed boundaries. *J. Math. Phys.*, 18(11):2108–2111.

[10] Berryman, J. G. and Holland, C. J. (1978). Nonlinear diffusion problem arising in plasma physics. *Phys. Rev. Lett.*, 40(26):1720.

[11] Bianchi, D., Buccini, A., Donatelli, M., and Serra-Capizzano, S. (2015). Iterated fractional tikhonov regularization. *Inverse Prob.*, 31(5):055005.

[12] Bianchi, D. and Setti, A. G. (2016). Laplacian cut-offs, porous and fast diffusion on manifolds and other applications. *arXiv preprint arXiv:1607.06008*.

[13] Bonforte, M. and Grillo, G. (2005). Asymptotics of the porous media equation via sobolev inequalities. *J. Funct. Anal.*, 225(1):33–62.

[14] Bonforte, M., Grillo, G., and Vazquez, J. L. (2008). Fast diffusion flow on manifolds of nonpositive curvature. *JEE*, 8(1):99–128.

[15] Braverman, M., Milatovic, O., and Shubin, M. (2002). Essential self-adjointness of schrödinger-type operators on manifolds. *Russ. Math. Surv.*, 57(4):641.

[16] Brezis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*. Springer.

[17] Brill, M. and Schock, E. (1987). Iterative solution of ill-posed problems – a survey. *Theory Practice Appl. Geophys.*, 1:13–37.

[18] Burago, D., Burago, Y., and Ivanov, S. (2001). *A course in metric geometry*, volume 33. American Mathematical Society Providence.

[19] Cai, J.-F., Chan, R., Shen, L., and Shen, Z. (2003). Wavelet algorithms for high-resolution image reconstruction. *SIAM J. Sci. Comput.*, 24:1408–1432.

[20] Cai, J. F., Osher, S., and Shen, Z. (2009a). Convergence of the linearized bregman iteration for ℓ_1-norm minimization. *Math. Comput.*

[21] Cai, J. F., Osher, S., and Shen, Z. (2009b). Linearized Bregman iterations for frame-based image deblurring. *SIAM J. Imaging Sci.*, 2(1):226–252.

[22] Cai, J. F., Osher, S., and Shen, Z. (2009c). Split Bregman methods and frame based image restoration. *Multiscale Model. Simul.*, 8(2):337–369.

[23] Cai, Y., Donatelli, M., Bianchi, D., and Huang, T.-Z. (2016). Regularization preconditioners for frame-based image deblurring with reduced boundary artifacts. *SIAM J. Sci. Comput.*, 38(1):B164–B189.

[24] Chan, R. H. and Jin, X. Q. (2007). *An Introduction to Iterative Toeplitz Solvers*. Society for Industrial and Applied Mathematics (SIAM).

[25] Chan, R. H. and Ng, M. K. (1996). Conjugate gradient method for toeplitz systems. *SIAM Review*, 38(3):427–482.

[26] Chan, R. H., Riemenschneider, S. D., Shen, L., and Shen, Z. (2004). Tight frame: an efficient way for high-resolution image reconstruction. *Appl. Comput. Harmon. Anal.*, 17:91–115.

[27] Chan, T. F. (1988). An optimal circulant preconditioner for toeplitz systems. *SIAM J. Sci. Stat. Comp.*, 9:766–771.

[28] Chavel, I. (2006). *Riemannian geometry: a modern introduction*, volume 98. Cambridge university press.

[29] Cheeger, J. (2001). *Degeneration of Riemannian metrics under Ricci curvature bounds*. Accademia Nazionale dei Lincei. Scuola Normale Superiore. Lezione Fermiane.

[30] Cheeger, J. and Colding, T. H. (1996). Lower bounds on ricci curvature and the almost rigidity of warped products. *Ann. of Math.*, 144(1):189–237.

[31] Chow, B. *The Ricci flow: techniques and applications. Part III, Geometric-analytic aspects*. American Mathematical Society.

[32] Christiansen, M. and Hanke, M. (2008). Deblurring methods using antireflective boundary conditions. *SIAM J. Sci. Comput.*, 30(2):855–872.

[33] Cittert, P. H. V. (1931). Zum einfluss der spaltbreite auf die intensitatsverteilung in spektrallinen ii. *Z. Phys. 69*.

[34] Conway, J. B. (1997). *A course in functional analysis*, volume 96. Springer Science & Business Media.

[35] Daubechies, I., Defrise, M., and Mol, C. D. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457.

[36] Daubechies, I., Han, B., Ron, A., and Shen, Z. (2003). Framelets: Mra-based constructions of wavelet frames. *Appl. Comput. Harmon. Anal.*, 14:1–46.

[37] Dell'Acqua, P., Donatelli, M., and Estatico, C. (2014). Preconditioners for image restoration by reblurring techniques. *J. Comput. Appl. Math.*, 272:313–333.

[38] Donatelli, M. (2012). On nondecreasing sequences of regularization parameters for nonstationary iterated Tikhonov. *Numer. Algorithms*, 40(4):651–668.

[39] Donatelli, M., Estatico, C., Martinelli, A., and Serra-Capizzano, S. (2006). Improved image deblurring with anti-reflective boundary conditions and re-blurring. *Inverse Prob.*, 22:2035–2053.

[40] Donatelli, M., Estatico, C., Nagy, J., Perrone, L., and Serra-Capizzano, S. (2003). Anti-reflective boundary conditions and fast 2d deblurring models. In Luk, F. T., editor, *in Advanced Signal Processing Algorithms, Architectures and Implementations XIII*. Proceeding of SPIE 5205.

[41] Donatelli, M. and Hanke, M. (2013). Fast nonstationary preconditioned iterative methods for ill-posed problems, with application to image deblurring. *Inverse Prob.*, 29(9):095008, 16.

[42] Donatelli, M. and Serra-Capizzano, S. (2005). Anti-reflective boundary conditions and re-blurring. *Inverse Prob.*, pages 169–182.

[43] Donatelli, M. and Serra-Capizzano, S. (2007). Filter factor analysis of an iterative multilevel regularizing method. *Electron. Trans. Numer. Anal.*, 29:163–177.

[44] Donatelli, M. and Serra-Capizzano, S. (2010a). Antireflective boundary conditions for deblurring problems. *IJECE*, 2010.

[45] Donatelli, M. and Serra-Capizzano, S. (2010b). On the treatment of boundary artifacts in image restoration by reflection and/or anti-reflection. In *Matrix Methods: Theory, Algorithms and Applications*, pages 227–237. World Scientific.

[46] Donnelly, H. (1997). Exhaustion functions and the spectrum of riemannian manifolds. *Indiana Univ. Math. J.*, 46(2):505–527.

[47] Drake, J. and Berryman, J. G. (1977). Theory of nonlinear diffusion of plasma across the magnetic field of a toroidal multipole. *Phys Fluids*, 20(5):851–857.

[48] Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer.

[49] Figueiredo, M. and Nowak, R. (2003). An em algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916.

[50] Gaffney, M. P. (1959). The conservation property of the heat equation on riemannian manifolds. *Commun. Pure Appl. Math.*, 12(1):1–11.

[51] Gert, D., Klann, E., Ramlau, R., and Reichel, L. (2014). On fractional Tikhonov regularization. *private notes*.

[52] Greene, R. and Wu, H. (1979). C^∞ approximations of convex, subharmonic, and plurisubharmonic functions. *Ann. Sci. cole Norm. Sup.*, 12(1):47–84.

[53] Grillo, G. and Muratori, M. (2014). Radial fast diffusion on the hyperbolic space. *Proc. London Math. Soc.*, 109(2):283–317.

[54] Grillo, G. and Muratori, M. (2016). Smoothing effects for the porous medium equation on cartan–hadamard manifolds. *Nonlinear Anal.*, 131:346–362.

[55] Groetsch, C. W. (1984). *The theory of Tikhonov regularization for Fredholm equations of the first kind*. Pitman, Boston, MA.

[56] Grummt, R. and Kolb, M. (2012). Essential selfadjointness of singular magnetic schrödinger operators on riemannian manifolds. *J. Math. Anal. Appl.*, 388(1):480–489.

[57] Güneysu, B. (2016). Sequences of laplacian cut-off functions. *J. Geom. Anal.*, 26(1):171–184.

[58] Hanke, M. and Groetsch, C. W. (1998). Nonstationary iterated Tikhonov regularization. *J. Optim. Theory Appl.*, 98(1):37–53.

[59] Hanke, M. and Hansen, P. C. (1993). Regularization methods for large-scale problems. *Surveys Math. Indust.*, 3(4):253–315.

[60] Hanke, M. and Nagy, J. (1996). Restoration of atmospherically blurred images by symmetric indefinite conjugate gradient techniques. *Inverse Prob.*, 12:157–173.

[61] Hanke, M., Nagy, J., and Plemmons, R. (1993). Preconditioned iterative regularization. In *Numerical linear algebra and Scientific Computing*. de Gruyter, Berlin.

[62] Hansen, P. C. (1994). Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems. *Numer. Algorithms*, 6(1-2):1–35.

[63] Hansen, P. C. (1998). *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. Siam.

[64] Hansen, P. C., Nagy, J., and O'Leary, D. P. (2005). *Deblurring Images Matrices, Spectra and Filtering*. SIAM Publications.

[65] Herrero, M. A. and Pierre, M. (1985). The cauchy problem for $u_t = \delta u$ when $0 < m < 1$. *Trans. Amer. Math. Soc.*, 291(1):145–158.

[66] Hochstenbach, M. E. and Reichel, L. (2011). Fractional Tikhonov regularization for linear discrete ill-posed problems. *BIT*, 51(1):197–215.

[67] Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

[68] Huang, J., Donatelli, M., and Chan, R. Nonstationary iterated thresholding algorithms for image deblurring. *Inverse Probl. Imaging*, 7(3):717–736.

[69] Huckle, T. H. and Sedlacek, M. (2012). Tikhonov–phillips regularization with operator dependent seminorms. *Numer. Algorithms*, 60(2):339–353.

[70] Kamm, J. and Nagy, J. G. (1998)). Kronecker product and svd approximations in image restoration. *Linear Algebra Appl.*, 284:177–192.

[71] Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Commun Pure Appl Math*, 30(5):509–541.

[72] Kato, T. (1972). Schrödinger operators with singular potentials. *Israel J. Math.*, 13(1-2):135–148.

[73] Klann, E. and Ramlau, R. (2008). Regularization by fractional filter methods and data smoothing. *Inverse Prob.*, 24(2):025018.

[74] Lebedev, N. N., Silverman, R. A., and Livhtenberg, D. B. (1965). Special functions and their applications. *Phys. Today*, 18:70.

[75] Leinfelder, H. and Simader, C. G. (1981). Schrödinger operators with singular magnetic vector potentials. *Math. Z.*, 176(1):1–19.

[76] Li, P. and Schoen, R. (1984). L p and mean value properties of subharmonic functions on riemannian manifolds. *Acta Math.*, 153(1):279–301.

[77] Louis, A. K. (1989). *Inverse und schlecht gestellte Probleme*. Teubner, Stuttgart.

[78] Lu, P., L., L. N. J., Vázquez, and Villani, C. (2009). Local aronson–bénilan estimates and entropy formulae for porous medium and fast diffusion equations on manifolds. *J. Math. Pures Appl.*, 91(1):1–19.

[79] Nagy, J., Palmer, K., and Perrone, L. Restoretools: an object oriented matlab package for image restoration.

[80] Nagy, J. G., Palmer, K., and Perrone, L. (2004). Iterative methods for image deblurring: A matlab object oriented approach. *Numer. Algorithms*, 36:73–93.

[81] Ng, M., Chan, R., and Tang, W. C. (1999). A fast algorithm for deblurring models with neumann boundary conditions. *SIAM J. Sci. Comput.*, 21:851–866.

[82] Okuda, H. and Dawson, J. M. (1973). Theory and numerical simulation on plasma diffusion across a magnetic field. *Physics of Fluids (1958-1988)*, 16(3):408–426.

[83] ONeil, B. (1983). *Semi-Riemannian Geometry*. Academic Press, New York.

[84] Perrone, L. (2006). Kronecker product approximations for image restoration with antireflective boundary conditions. *Numer. Linear Algebra Appl.*, 13:1–22.

[85] Pigola, S., Rigoli, M., and Setti, A. G. (2005). *Maximum principles on Riemannian manifolds and applications*. American Mathematical Soc.

[86] Pigola, S., Rigoli, M., and Setti, A. G. (2008). *Vanishing and finiteness results in geometric analysis: a generalization of the Bochner technique*, volume 266. Springer Science & Business Media.

[87] Reed, M. and Simon, B. (1972). *Methods of Modern Mathematical Physics: Vol.: 1.: Functional Analysis*. Academic press.

[88] Reed, M. and Simon, B. (1980). Functional analysis, volume 1 of methods of modern mathematical physics. *Academic Press, Orlando,*, 4:4.

[89] Reeves, S. J. (2005). Fast image restoration without boundary artifacts. *IEEE Trans. Image Process.*, 14:1448–1453.

[90] Rimoldi, M. and Veronelli, G. (2016). Extremals of log sobolev inequality on non-compact manifolds and ricci soliton structures. *arXiv preprint arXiv:1605.09240*.

[91] Rudin, L., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D*, 60:259–268.

[92] Rudin, W. (1987a). *Real and complex analysis*. Tata McGraw-Hill Education.

[93] Rudin, W. (1987b). *Real and complex analysis*. Tata McGraw-Hill Education.

[94] Rudin, W. (1991). *Functional analysis. International series in pure and applied mathematics*. McGraw-Hill, Inc., New York.

[95] Schoen, R. and Yau, S.-T. (1994a). *Lectures on differential geometry*, volume 1. International press Cambridge.

[96] Schoen, R. and Yau, S.-T. (1994b). *Lectures on differential geometry*, volume 2. International press Cambridge.

[97] Semplice, M. (2010). Preconditioned implicit solvers for nonlinear pdes in monument conservation. *SIAM J. Sci. Comput.*, 32(5):3071–3091.

[98] Serra-Capizzano, S. (2003). A note on anti-reflective boundary conditions and fast deblurring models. *SIAM J. Sci. Comput.*, 25(3):307–325.

[99] Shen, Z., Toh, K. C., and Yun, S. An accelerated proximal gradient algorithm for frame-based image restoration via the balanced approach. *SIAM J. Imaging Sciences*, 4:573–596.

[100] Shishkin, G. (2008). Grid approximation of a parabolic convection-diffusion equation on a priori adapted grids: ε-uniformly convergent schemes. *Zh. Vychisl. Mat. Mat. Fiz.*, 48(6):1014–1033.

[101] Shubin, M. (1992). Spectral theory of elliptic operators on non-compact manifolds. *Astérisque*, 207:37–108.

[102] Shubin, M. (2001). Essential self-adjointness for semi-bounded magnetic schrödinger operators on non-compact manifolds. *J. Funct. Anal.*, 186(1):92–116.

[103] Sorel, M. (2012). Removing boundary artifacts for real-time iterated shrinkage deconvolution. *IEEE Trans. Image Process.*, 21(4):2329–2334.

[104] Strichartz, R. S. (1983). Analysis of the laplacian on the complete riemannian manifold. *J. Funct. Anal.*, 52(1):48–79.

[105] Stynes, M. (2005). Steady-state convection-diffusion problems. *Acta Numer.*, pages 445–508.

[106] Tilli, P. (1998). Locally toeplitz sequences: spectral properties and applications. *Linear Algebra Appl.*, 278:91–120.

[107] Tyrtyshnikov, E. E. and Zamarashkin, N. (1998). Spectra of multilevel toeplitz matrices: advanced theory via simple matrix relationships. *Linear Algebra Appl.*, 278:15–27.

[108] Vázquez, J. L. (2006). Smoothing and decay estimates for nonlinear parabolic equations of porous medium type. *Oxford Lecture Notes in Maths and its Applications*, 33.

[109] Vázquez, J. L. (2007). *The porous medium equation: mathematical theory*. Oxford University Press.

[110] Vázquez, J. L. (2015). Fundamental solution and long time behavior of the porous medium equation in hyperbolic space. *J. Math. Pures Appl.*, 104(3):454–484.

[111] Vichnevetsky, R. (1987). Wave propagation and reflection in irregular grids for hyperbolic equations. *Appl. Numer. Math.*, 3(1-2):133–166.

[112] Vio, R., Bardsley, J., Donatelli, M., and Wamsteker, W. (2005). Dealing with edge effects in least-squares image deconvolution problems. *Astron. Astrophys.*, 442:397–403.

[113] Wang, F. and Zhu, X. (2013). On the structure of spaces with Bakry-\'{E}mery Ricci curvature bounded below. *ArXiv e-prints*.

[114] Wang, Y., Yang, J., Yin, W., and Zhang, Y. (2008). A new alternating minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sci.*, 1(3):248–272.

[115] Yin, W., Osher, S., Goldfarb, D., and Darbon, J. (2008). Bregman iterative algorithms for $l1$-minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143–168.