Applied and Numerical Harmonic Analysis

Filippo De Mari
Ernesto De Vito
Editors

Harmonic and Applied Analysis

From Radon Transforms to Machine Learning

Birkhäuser

Applied and Numerical Harmonic Analysis

Filippo De Mari · Ernesto De Vito
Editors

Harmonic and Applied Analysis

From Radon Transforms to Machine Learning

Birkhäuser

Editors
Filippo De Mari
Dipartimento di Matematica
Università di Genova
Genova, Italy

Ernesto De Vito
Dipartimento di Matematica
Università di Genova
Genova, Italy

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-030-86663-1 ISBN 978-3-030-86664-8 (eBook)
https://doi.org/10.1007/978-3-030-86664-8

Mathematics Subject Classification: 53C35, 60B10, 62G05, 90C25

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

ANHA Series Preface

The *Applied and Numerical Harmonic Analysis (ANHA)* book series aims to provide the engineering, mathematical, and scientific communities with significant developments in harmonic analysis, ranging from abstract harmonic analysis to basic applications. The title of the series reflects the importance of applications and numerical implementation, but richness and relevance of applications and implementation depend fundamentally on the structure and depth of theoretical underpinnings. Thus, from our point of view, the interleaving of theory and applications and their creative symbiotic evolution is axiomatic.

Harmonic analysis is a wellspring of ideas and applicability that has flourished, developed, and deepened over time within many disciplines and by means of creative cross-fertilization with diverse areas. The intricate and fundamental relationship between harmonic analysis and fields such as signal processing, partial differential equations (PDEs), and image processing is reflected in our state-of-the-art *ANHA* series.

Our vision of modern harmonic analysis includes mathematical areas such as wavelet theory, Banach algebras, classical Fourier analysis, time–frequency analysis, and fractal geometry, as well as the diverse topics that impinge on them.

For example, wavelet theory can be considered an appropriate tool to deal with some basic problems in digital signal processing, speech and image processing, geophysics, pattern recognition, biomedical engineering, and turbulence. These areas implement the latest technology from sampling methods on surfaces to fast algorithms and computer vision methods. The underlying mathematics of wavelet theory depends not only on classical Fourier analysis but also on ideas from abstract harmonic analysis, including von Neumann algebras and the affine group. This leads to a study of the Heisenberg group and its relationship to Gabor systems, and of the metaplectic group for a meaningful interaction of signal decomposition methods.

The unifying influence of wavelet theory in the aforementioned topics illustrates the justification for providing a means for centralizing and disseminating information from the broader, but still focused, area of harmonic analysis. This will be a

key role of *ANHA*. We intend to publish with the scope and interaction that such a host of issues demands.

Along with our commitment to publish mathematically significant works at the frontiers of harmonic analysis, we have a comparably strong commitment to publish major advances in the following applicable topics in which harmonic analysis plays a substantial role:

Analytic Number theory
Antenna Theory
Artificial Intelligence
Biomedical Signal Processing
Classical Fourier Analysis
Coding Theory
Communications Theory
Compressed Sensing
Crystallography and Quasi-Crystals
Data Mining
Data Science
Deep Learning
Digital Signal Processing
Dimension Reduction and Classification
Fast Algorithms
Frame Theory and Applications
Gabor Theory and Applications
Geophysics
Image Processing
Machine Learning
Manifold Learning
Numerical Partial Differential Equations
Neural Networks
Phaseless Reconstruction
Prediction Theory
Quantum Information Theory
Radar Applications
Sampling Theory (Uniform and Non-uniform) and Applications
Spectral Estimation
Speech Processing
Statistical Signal Processing
Super-resolution
Time Series
Time-Frequency and Time-Scale Analysis
Tomography
Turbulence
*Uncertainty Principles *Waveform design*
Wavelet Theory and Applications

The above point of view for the *ANHA* book series is inspired by the history of Fourier analysis itself, whose tentacles reach into so many fields.

In the last two centuries, Fourier analysis has had a major impact on the development of mathematics, on the understanding of many engineering and scientific phenomena, and on the solution of some of the most important problems in mathematics and the sciences. Historically, Fourier series were developed in the analysis of some of the classical PDEs of mathematical physics; these series were used to solve such equations. In order to understand Fourier series and the kinds of solutions they could represent, some of the most basic notions of analysis were defined, e.g., the concept of "function." Since the coefficients of Fourier series are integrals, it is no surprise that Riemann integrals were conceived to deal with uniqueness properties of trigonometric series. Cantor's set theory was also developed because of such uniqueness questions.

A basic problem in Fourier analysis is to show how complicated phenomena, such as sound waves, can be described in terms of elementary harmonics. There are two aspects of this problem: first, to find, or even define properly, the harmonics or spectrum of a given phenomenon, e.g., the spectroscopy problem in optics; second, to determine which phenomena can be constructed from given classes of harmonics, as done, for example, by the mechanical synthesizers in tidal analysis.

Fourier analysis is also the natural setting for many other problems in engineering, mathematics, and the sciences. For example, Wiener's Tauberian theorem in Fourier analysis not only characterizes the behavior of the prime numbers, but is a fundamental tool for analyzing the ideal structures of Banach algebras. It also provides the proper notion of spectrum for phenomena such as white light. This latter process leads to the Fourier analysis associated with correlation functions in filtering and prediction problems. These problems, in turn, deal naturally with Hardy spaces in complex analysis, as well as inspiring Wiener to consider communications engineering in terms of feed-back and stability, his cybernetics. This latter theory develops concepts to understand complex systems such as learning and cognition and neural networks; and it is arguably a precursor of deep learning and its spectacular interactions with data science and AI.

Nowadays, some of the theory of PDEs has given way to the study of Fourier integral operators. Problems in antenna theory are studied in terms of unimodular trigonometric polynomials. Applications of Fourier analysis abound in signal processing, whether with the fast Fourier transform (FFT), or filter design, or the adaptive modeling inherent in time–frequency-scale methods such as wavelet theory.

The coherent states of mathematical physics are translated and modulated Fourier transforms, and these are used, in conjunction with the uncertainty principle, for dealing with signal reconstruction in communications theory. We are back to the raison d'être of the *ANHA* series!

College Park, MD, USA — John J. Benedetto
College Park, MD, USA — Wojciec Czaja
Boston, MA — Kasso Okoudjou

Preface

As first stated by Galileo Galilei, Mathematics is the language of the Nature and, conversely, our understanding of the Nature gives a fresh impulse to develop new appealing mathematical theories. For example, the first two decades of the last century were characterized by two revolutions in Physics: Relativity and Quantum Mechanics. These new ideas are strongly related to the fast-growing of differential geometry and functional analysis.

A century later a new revolution is on the way: Big Data and Machine Learning are central both in scientific research and in everyday life applications.

Once again Mathematics provides the natural language for a solid understanding of these topics and, conversely, they ask for new sophisticated mathematical tools.

Machine Learning tries to provide a positive answer to the problem of Artificial Intellingence: "Can machines be able to infer new knowledge from their past experience as a human being does ?"[6]. For example, a child is able to recognize a cat provided that his parents have shown him some example of cats. A learning machine would be an algorithm that, starting from a training set of examples of input–output pairs, is able to assign the correct output to a new unlabelled input. Statistical learning theory is the theoretical framework for Machine Learning. However, this kind of problems has a long history outside Machine Learning. For example, in the framework of estimation problems, the first example of learning algorithm goes back to Boscovich and Laplace, who introduced the least absolute regression to fit astronomic data at the end of '700, whereas the most well-known algorithm is the least square regression independently introduced by Legendre and Gauss at the beginning of '800, see [4, 10] for a historical account.

Classical estimation theory is based on some strong *a priori* assumption, as for example that the data follow a normal distribution or that the functional relation between input and output is linear, whereas Machine Learning usually deals with problems where the generating model of the data in largely unknown. In this framework, one of the first examples of learning algorithms is the perceptron introduced by Rosenblatt in the late '50 which is at the root of modern Neural Networks [1].

From a mathematical point of view, Statistical Learning can be seen as a branch of non-parametric estimation theory and of empirical processes, see for example [5] and [11]. As a discipline in its own right its theoretical foundation can be traced back to the work of Vapnik started at the beginning of '70, see [12] and references therein. Around '90 Poggio and Girosi first showed that statistical learning theory can be reformulated as a classical approximation problem [13]. This point of view was further developed by Smale [2] allowing to recast learning theory by using tools of Functional Analisys [3]. This approach makes the connection with the theory of inverse problems very clear. Following this point of view, the chapter *Regularization: from inverse problems to large-scale Machine Learning* provides a brief introduction to Statistical Learning theory, whereas *Ill-posed problems: From linear to non-linear and beyond* is devoted to a review on ill posed inverse problems.

Alhough Statistical Learning can be recast in the framework of Functional Analysis, it naturally asks for advanced concentration inequalities that generalize the classic weak law of large numbers, which, in the present form, is due to Chebychev and Bienaimé in the late '800, see [9] for a historical account. Starting from the seminal work of Talagrand in the '90, see [8] and references therein, in the last decades there is a growing interest on concentration inequalities. A recent method to derive concentration inequalities is the entropy method, which takes inspiration from classical tools in statistical mechanics and quantum field theory, and whose power was first recognized by Ledoux, see [7] and references therein. The chapter *Entropy and Concentration* gives a self-contained introduction to the entropy method close to its original statistics formulation and applies it to derive concentration inequalities that are standard tools to analyze the statistical properties of learning algorithms.

Another feature characterizing Machine Learning is that the data are not uniformly distributed on a bounded subset of some nice Euclidean space, but they live near some unknown submanifold or, even worse, on some discrete graph. Almost by its very definition, a framework in which signals on manifolds and graphs can be treated and analyzed is that of Harmonic Analysis, where several notions of transforms are at the very heart. We focus here on the special role played by the Radon transform, in view of its many applications. More specifically, the pioneering work of Helgason on integral transforms on Riemannian symmetric spaces, which is reviewed in the chapter *Unitarization of the Horocyclic Radon Transform on Symmetric Spaces*, laid the foundations of a large body of problems that span from mathematical issues concerning very general Radon-type transforms to the challenges of up-to-date applied tomographic techniques.

As already pointed out, we are in the Big Data epoch and Machine Learning has to provide efficient algorithms dealing with large databases, as Imagenet (10^6 images, organized according to the WordNet hierarchy), Million Song (10^6 audio features for music tracks), and HIGGS (10^7 Monte Carlo simulations to distinguish between a signal process which produces Higgs bosons and a background process which does not), to name a few. It is crucial to have efficient optimization

techniques to make the learning algorithms computationally efficient, a large class of which are defined as minimization problems of a convex functional on some suitable function space. The chapter *Proximal gradient methods for machine learning and imaging* provides an updated introduction to convex optimization tuned to learning theory.

This volume collects some of the contributions that have been presented during the second and third editions of the Summer Schools that have been held in Genova in 2017 and 2019. In this sense, the book should be thought of as the second volume of what might hopefully become a series, whose first volume is "Harmonic and Applied Analysis", ANHA, 2015 (Dahlke, De Mari, Grohs, Labate Eds.). Most of the chapters appearing here are sets of notes, or adaptations thereof, of the courses that have actually been taught during the summer schools, but a certain degree of expansion has been encouraged. After all, people who attended the schools have developed interests and skills that demand a reasonable continuation in the directions that have been pointed at during the courses. Two of the contributions (Bartolucci–De Mari–Monti and Salzo–Villa) are not directly linked to the actual summer schools, but are indeed topics on which several local students do part of their training.

Genova, Italy

Filippo De Mari
Ernesto De Vito

References

1. Bengio, Y., Goodfellow, I., Courville, A.: Deep learning, vol. 1. MIT press Massachusetts, USA (2017)
2. Cucker, F., Smale, S.: On the mathematical foundation of learning. Am. Math. Soc. **39**(1), 1–49 (2001)
3. Cucker, F., Zhou, D.X.: Learning theory: an approximation theory viewpoint, vol. 24. Cambridge University Press (2007)
4. Farebrother, R.W.: Fitting linear relationships. Springer Series in Statistics. Perspectives in Statistics. Springer-Verlag, New York (1999). DOI 10.1007/978-1-4612-0545-6. URL https://doi.org/10.1007/978-1-4612-0545-6. A history of the calculus of observations 1750–1900
5. Györfi, L., Kohler, M., Krzyżak, A., Walk, H.: A distribution-free theory of nonparametric regression. Springer Series in Statistics. Springer-Verlag, New York (2002). DOI 10.1007/b97848. URL https://doi.org/10.1007/b97848
6. Harnad, S.: The annotation game: On turing (1950) on computing, machinery, and intelligence. In: The Turing test sourcebook: philosophical and methodological issues in the quest for the thinking computer. Kluwer (2006)
7. Ledoux, M.: The concentration of measure phenomenon, *Mathematical Surveys and Monographs*, vol. 89. American Mathematical Society, Providence, RI (2001). DOI 10.1090/surv/089 . URL https://doi.org/10.1090/surv/089
8. Ledoux, M., Talagrand, M.: Probability in Banach spaces. Classics in Mathematics. Springer-Verlag, Berlin (2011). Isoperimetry and processes, Reprint of the 1991 edition
9. Seneta, E.: A tricentenary history of the law of large numbers. Bernoulli **19**(4), 1088–1121 (2013). DOI 10.3150/12-BEJSP12 . URL https://doi.org/10.3150/12-BEJSP12

10. Stigler, S.M.: Gauss and the invention of least squares. Ann. Statist. **9**(3), 465–474 (1981)
11. van der Vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York (1996). DOI 10.1007/978-1-4757-2545-2 . URL https://doi.org/10.1007/978-1-4757-2545-2 . With applications to statistics
12. Vapnik, V.N.: Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York (1998). A Wiley-Interscience Publication
13. Wahba, G.: Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia (1990)

Contents

Contributors

Rima Alaifari ETH Zürich, Zürich, Switzerland

Francesca Bartolucci Seminar for Applied Mathematics, ETH Zurich, Zurich, Switzerland

Filippo De Mari DIMA & MaLGa Center, Università di Genova, Genova, Italy

Ernesto De Vito DIMA & MaLGA, Universita di Genova, Genova, Italy

Andreas Maurer Istituto Italiano di Tecnologia, Genova, Italy

Matteo Monti DIMA & MaLGa Center, Università di Genova, Genova, Italy

Lorenzo Rosasco DIMA & MaLGA, Universita di Genova, Genova, Italy

Alessandro Rudi Inria and Ecole Normale Supérieure, PSL Research University, Paris, France

Saverio Salzo Istituto Italiano di Tecnologia, Via E. Melen 83, Genova, Italy

Silvia Villa DIMA & MaLGa Center, Università degli Studi di Genova, Genova, Italy

Unitarization of the Horocyclic Radon Transform on Symmetric Spaces

Francesca Bartolucci, Filippo De Mari, and Matteo Monti

1 Introduction

The Radon transform has its origin in the problem of recovering a function defined on $\mathbb{R}^d$ from its integrals over hyperplanes. In 1917 Radon proved the reconstruction formula for two- and three-dimensional signals. In $\mathbb{R}^3$ it reads

$$f(x) = -\frac{1}{8\pi^2}\Delta \int_{S^2} \mathcal{R}f(\theta, x\cdot\theta)\mathrm{d}\theta, \tag{1}$$

where Δ is the Laplacian acting on the variable x, S^2 is the sphere in $\mathbb{R}^3$ and for every $\theta \in S^2$ and $t \in \mathbb{R}$ we denote by $\mathcal{R}f(\theta, t)$ the integral of f over the hyperplane $x \cdot \theta = t$. Formula (1) suggests to define two dual transforms $f \mapsto \mathcal{R}f$, $g \mapsto \mathcal{R}^{\#}g$, known as Radon transform and dual Radon transform, or back-projection, respectively. The Radon transform $\mathcal{R}$ maps a function on $\mathbb{R}^d$ into the set of integrals over all hyperplanes, while the dual Radon transform $\mathcal{R}^{\#}$ maps a function defined on the set of hyperplanes of $\mathbb{R}^d$ into its integrals over the sheaves of hyperplanes through a point. Formula (1) can be rewritten

$$f = -\frac{1}{2}\Delta\mathcal{R}^{\#}\mathcal{R}f$$

and solves the inverse problem of recovering f from the measured datum $\mathcal{R}f$.

F. Bartolucci
Seminar for Applied Mathematics, ETH Zurich, Raemistrasse 101, 8092 Zurich, Switzerland
e-mail: francesca.bartolucci@sam.math.ethz.ch

F. De Mari (✉) · M. Monti
DIMA & MaLGa Center, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy
e-mail: demari@dima.unige.it

M. Monti
e-mail: monti.m@dima.unige.it

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8_1

This classical inverse problem is a particular case of the more general issue of recovering an unknown function on a manifold by means of its integrals over a family of submanifolds, already investigated by Gelfand in the 1950s [12]. A natural framework for such general inverse problems was considered by Helgason [17] and is motivated by the group structure hidden in the polar Radon transform setting [19], whereby the signals to be analyzed are in $\mathbb{R}^d$.

In the planar case, $\mathbb{R}^2$ and $[0, 2\pi) \times \mathbb{R}$, which parametrizes the set of lines in the plane by polar coordinates, are both transitive spaces of the rigid motions' group. This is $G = \mathbb{R}^2 \rtimes K$, with $K = \{R_\phi : \phi \in [0, 2\pi)\}$, where

$$R_\phi = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

We write $(b, \phi) \in \mathbb{R}^2 \times [0, 2\pi)$ for the elements in G and define the group law by

$$(b, \phi)(b', \phi') = (b + R_\phi b', \phi + \phi' \bmod 2\pi).$$

The group G acts transitively on $\mathbb{R}^2$ by means of

$$(b, \phi)[x] = R_\phi x + b$$

and the isotropy at the origin $x_0 = (0, 0)$ is the Abelian subgroup

$$K \simeq \{(0, \phi) : \phi \in [0, 2\pi)\}.$$

Therefore $\mathbb{R}^2 \simeq G/K$ under the canonical isomorphism $gK \mapsto g[x_0]$. Clearly, G is a group of affine transformations of the plane and maps lines into lines. A line in the plane is parametrized by the direction ${}^t n(\theta) = (\cos(\theta), \sin(\theta))$, where $\theta \in [0, 2\pi)$, of its normal and by the coordinate t on the oriented normal line[1] which describes its intersection with the given line. The transitive action of G is then given by

$$(b, \phi).(\theta, t) = (\theta + \phi \bmod 2\pi, t + n(\theta) \cdot R_\phi^{-1} b)$$

with isotropy at the y-axis $\xi_0 = (0, 0) \in [0, 2\pi) \times \mathbb{R}$ given by

$$H = \{((0, b_2), \phi) : b_2 \in \mathbb{R}, \phi \in \{0, \pi\}\}.$$

Thus, $[0, 2\pi) \times \mathbb{R} \simeq G/H$ under the canonical isomorphism $gH \mapsto g.\xi_0$. From this group-theoretic point of view, the fact that a point $x \in \mathbb{R}^2$ belongs to the line $(\theta, t) \in [0, 2\pi) \times \mathbb{R}$ is equivalent to requiring that the left cosets $x = g_1 K$ and $(\theta, t) = g_2 H$ intersect. Indeed, $g_1[x_0]$ belongs to the line $g_2.\xi_0$ if and only if there exists $h \in H$ such that $g_1[x_0] = g_2 h[x_0]$, so that $g_1(g_2 h)^{-1} \in K$ and $g_1 K \cap g_2 H \neq \emptyset$. This structure illustrates the following general framework introduced by Helgason.

[1] The orientation is such that the coordinate t is 1 exactly at $(\cos\theta, \sin\theta)$.

Consider two G-spaces X and Ξ, where the actions on $x \in X$ and $\xi \in \Xi$ are

$$(g, x) \mapsto g[x], \quad (g, \xi) \mapsto g.\xi.$$

Both X and Ξ are assumed to be transitive spaces, so that there exist quasi-invariant measures $\mathrm{d}x$ and $\mathrm{d}\xi$. In Helgason's approach, it is assumed that $\mathrm{d}x$ and $\mathrm{d}\xi$ are invariant measures. Fix $x_0 \in X$ and $\xi_0 \in \Xi$ and denote by K and H the corresponding stability subgroups, so that $X \simeq G/K$ and $\Xi \simeq G/H$ under the isomorphisms $gK \mapsto g[x_0]$ and $gH \mapsto g.\xi_0$, respectively. The space X is meant to describe the ambient in which the functions to be analyzed live, for example, the Euclidean plane, or the sphere S^2 or the hyperbolic plane H^2. The second space Ξ parametrizes the set of submanifolds of X over which one wants to integrate functions, for instance, lines in the Euclidean plane, great circles in S^2, geodesics or horocycles in H^2. Motivated by the group structure behind the polar Radon transform, the elements in Ξ can be realized as submanifolds of X introducing the concept of incidence. Two elements $x = g_1 K$ and $\xi = g_2 H$ are said to be incident if they intersect as cosets in G. The concept of incidence translates the fact that a point $x \in X$ belongs to the submanifold parametrized by $\xi \in \Xi$. Any point $\xi \in \Xi$ is realized as a submanifold $\widehat{\xi} \subset X$ by taking all the points $x \in X$ that are incident to ξ. Precisely,

$$\widehat{\xi} = \{x \in X : x \text{ and } \xi \text{ are incident}\} \subset X. \tag{2}$$

Conversely, one builds the "sheaf" of manifolds $\check{x}$ through the point $x \in X$ by taking all the points $\xi \in \Xi$ that are incident to x

$$\check{x} = \{\xi \in \Xi : \xi \text{ and } x \text{ are incident}\} \subset \Xi. \tag{3}$$

By (2) and (3) we have that

$$\widehat{\xi_0} = H[x_0] \subset X, \qquad \check{x}_0 = K.\xi_0 \subset \Xi.$$

Both $\check{x}_0$ and $\widehat{\xi_0}$ are transitive spaces and hence carry quasi-invariant measures. By definition, for any $x = gK$ and $\xi = \gamma H$

$$\check{x} = g.\check{x}_0 \subset \Xi, \qquad \widehat{\xi} = \gamma[\widehat{\xi_0}] \subset X,$$

which are closed subsets by Lemma 1.1 in [19]. If the maps $\xi \mapsto \widehat{\xi}$ and $x \mapsto \check{x}$ are both injective, then the pair of homogeneous spaces (X, Ξ) is called a dual pair. This assumption is called transversality, see Lemma 1.3 in [19] for an equivalent characterization. The transversality condition avoids a redundant parametrization of the submanifolds of X. The reader may consult [19] for numerous examples of dual pairs. It is worth observing that the leading example of the polar Radon transform does not satisfy the transversality condition. Indeed, the points (θ, t) and $(\theta + \pi \mod 2\pi, -t)$ in $[0, 2\pi) \times \mathbb{R}$ both parametrise the line given by the set of points

$$\widehat{(\theta, t)} = (\theta + \pi \widehat{\bmod 2\pi}, -t) = \{x \in \mathbb{R}^2 : x \cdot n(\theta) = t\}.$$

For a deeper study on the injectivity issue, the reader may consider [2].

In Helgason's approach the transitive spaces $\check{x}_0$ and $\widehat{\xi}_0$ are supposed to carry K-invariant and H-invariant measures, respectively, that is

$$\int_{\check{x}_0} g(k^{-1}.\xi)\mathrm{d}\mu_0(\xi) = \int_{\check{x}_0} g(\xi)\mathrm{d}\mu_0(\xi), \qquad g \in L^1(\check{x}_0, \mathrm{d}\mu_0),\ k \in K,$$

$$\int_{\widehat{\xi}_0} f(h^{-1}[x])\mathrm{d}m_0(x) = \int_{\widehat{\xi}_0} f(x)\mathrm{d}m_0(x), \qquad g \in L^1(\widehat{\xi}_0, \mathrm{d}m_0),\ h \in H.$$

In order to define the Radon transform and its dual, one needs to introduce measures on $\widehat{\xi}$ and $\check{x}$. This may be done taking the pushforward of the measure $\mathrm{d}\mu_0$ to $\widehat{\xi} = \widehat{(gH)}$ by the map $\widehat{\xi}_0 \ni x \mapsto g[x] \in \widehat{\xi}$ and of the measure $\mathrm{d}m_0$ to $\check{x} = \check{(gK)}$ by the map $\check{x}_0 \ni \xi \mapsto g.\xi \in \check{x}$, respectively. We denote by $\mathrm{d}\mu_x$ the measure on $\check{x}$ and by $\mathrm{d}m_\xi$ the measure on $\widehat{\xi}$. Since the measures on $\widehat{\xi}_0$ and $\check{x}_0$ are invariant, the measures $\mathrm{d}m_\xi$ and $\mathrm{d}\mu_x$ do not depend on the choice of the representatives of ξ and x and the transversality condition guarantees that they are unique.

Definition 1 The Radon transform of f is the map $\mathcal{R}f : \Xi \to \mathbb{C}$ given by

$$\mathcal{R}f(\xi) = \int_{\widehat{\xi}} f(x)\mathrm{d}m_\xi(x),$$

and the dual Radon transform of g is the map $\mathcal{R}^\# g : X \to \mathbb{C}$ given by

$$\mathcal{R}^\# g(x) = \int_{\check{x}} g(\xi)\mathrm{d}\mu_x(\xi),$$

for any f and g for which the integrals converge.

Observe that, even if the transversality condition is not satisfied for the polar Radon transform, both $\widehat{(\theta, t)}$ and $(\theta + \pi \widehat{\bmod 2\pi}, -t)$ are endowed with the same measure since the arc-length measure is invariant under translations and rotations. For this reason the polar Radon transform satisfies

$$\mathcal{R}^{\mathrm{pol}} f(\theta, t) = \mathcal{R}^{\mathrm{pol}} f(\theta + \pi \bmod 2\pi, -t).$$

In this context, the most relevant issue is to recover f from the values of $\mathcal{R}f$. Another central issue is to prove that the Radon transform, up to a composition with a suitable pseudo-differential operator, can be extended to a unitary map $\mathcal{Q}$ from $L^2(X, \mathrm{d}x)$ to $L^2(\Xi, \mathrm{d}\xi)$ intertwining the quasi-regular representations π and $\hat{\pi}$ of G acting on $L^2(X, \mathrm{d}x)$ and $L^2(\Xi, \mathrm{d}\xi)$, respectively.

In [4], the authors obtain both an intertwining and a unitarization result for the affine Radon transform. The techniques used in [4] mimic the approach followed by Helgason to unitarize the polar Radon transform [19].

Later, inspired by the results in [4] a new approach based on representation theory has been taken in order to treat in a general and unified way the problem of unitarizing and inverting the Radon transform [1] under the assumption that π and $\hat{\pi}$ are irreducible. The approach taken in [1, 4] differs from Helgason's since the assumptions on the measures carried by X and Ξ and by the submanifolds $\hat{\xi} \subset X$ are weaker, namely, their relative invariance instead of (proper) invariance. This allows to consider a wider variety of cases of interest in applications, such as the similitude group studied by Murenzi [3], and the generalized shearlet dilation groups introduced by Führ in [9, 10] for the purpose of generalizing the standard shearlet group introduced in [6, 23]. It is assumed that there exists a non-trivial π-invariant subspace $\mathcal{A}$ of $L^2(X, \mathrm{d}x)$ such that $\mathcal{R}$ is well defined for all $f \in \mathcal{A}$ and the adjoint of the operator $\mathcal{R}\colon \mathcal{A} \to L^2(\Xi, \mathrm{d}\xi)$ has non-trivial domain. Then, it is proved that the Radon transform $\mathcal{R}$ is a closable operator from $\mathcal{A}$ into $L^2(\Xi, \mathrm{d}\xi)$ and that its closure $\overline{\mathcal{R}}$ is independent of the choice of $\mathcal{A}$ and is the unique closed extension of $\mathcal{R}$. The main result states that if the quasi-regular representations π of G on $L^2(X, \mathrm{d}x)$ and $\hat{\pi}$ of G on $L^2(\Xi, \mathrm{d}\xi)$ are irreducible, then the Radon transform $\mathcal{R}$, up to a composition with a suitable pseudo-differential operator, can be extended to a unitary operator $Q : L^2(X, \mathrm{d}x) \to L^2(\Xi, \mathrm{d}\xi)$ which intertwines them, namely,

$$\hat{\pi}(g) Q \pi(g)^{-1} = Q, \qquad g \in G.$$

The proof is based on the extension of Schur's lemma due to Duflo and Moore [7].

A direct consequence of the result above is studied in [1]. Adding the hypothesis of square-integrability of π, the authors derive a new general inversion formula for the Radon transform of the form

$$f = \int_G \chi(g) \langle \mathcal{R} f, \hat{\pi}(g)\Psi \rangle \pi(g)\psi \, \mathrm{d}g,$$

where χ is a character of G and $\psi \in L^2(X, \mathrm{d}x)$ and $\Psi \in L^2(\Xi, \mathrm{d}\xi)$ are suitable mother wavelets and where the Haar integral is weakly convergent. Such formula is obtained by the usual reconstruction formula for square-integrable representations and then by applying the unitary operator Q to both entries of the scalar product $\langle f, \pi(g)\psi \rangle$. We stress that the above formula allows to reconstruct an unknown signal by computing the family of coefficients $\{\langle \mathcal{R} f, \hat{\pi}(g)\Psi \rangle\}_{g \in G}$.

The results achieved in [1, 4] have posed many interesting mathematical challenges. A natural question is to investigate how to generalize these findings to other groups and related representations without the hypothesis of irreducibility, because the techniques used in [1] cannot be transferred directly.

In this direction, we have considered in [5] the case of homogeneous trees. Precisely, we construct the unitarization of the horocyclic Radon transform on a homogeneous tree X and we prove that it intertwines the quasi-regular representations of

the group of isometries of X acting on the space of square-integrable functions on the tree itself and on the space of horocycles, respectively. Since the quasi-regular representation is not irreducible, we adopt a combination of the approach followed by Helgason in the context of symmetric spaces [17] and the techniques that have been developed in [4]. The main observation motivating [5] is that homogeneous trees are the natural discrete counterpart of rank-one symmetric spaces.

This article is devoted to investigate the unitarization problem in the case when X is a symmetric space and Ξ is the set of horocycles of X, which has at large been addressed by Helgason. A remarkable difference from the cases treated in [1] is that the quasi-regular representations π of the group of isometries of the symmetric space X acting on $L^2(X)$ is not irreducible, nor is it the representation $\hat{\pi}$ on $L^2(\Xi)$. We are well aware that the unitarization problem was already considered and essentially solved by Helgason in [17]. Precisely, he constructs a pseudo-differential operator Λ and he proves that the pre-composition with the horocyclic Radon transform yields an isometric operator, see Theorem 3.9 in Chap. II in [17]. Here, we prove that the composition $\Lambda\mathcal{R}$ can actually be extended to a unitary operator $\mathcal{Q} : L^2(X, \mathrm{d}x) \to L^2_\flat(\Xi, \mathrm{d}\xi)$, where $\mathrm{d}x$ and $\mathrm{d}\xi$ are the G-invariant measures and where $L^2_\flat(\Xi, \mathrm{d}\xi)$ is a closed subspace of $L^2(\Xi, \mathrm{d}\xi)$ which accounts for the Weyl symmetries. Furthermore, we are able to show that $\mathcal{Q}$ intertwines the quasi-regular representations π and $\hat{\pi}$.

This work is focused on the horocyclic Radon transform, but another interesting setting could be obtained by considering geodesics. The latter is commonly called X-ray transform and has been introduced and inverted by Helgason on the hyperbolic space $\mathbb{H}^n$, see Theorem 3.12 in Chap. I in [17], and on symmetric spaces of the noncompact type by Rouvière [24]. Although it is not in general true that a horocycle has codimension one in the symmetric space, the horocyclic Radon transform can be seen as the analogue of the Euclidean Radon transform on hyperplanes in $\mathbb{R}^n$, whereas the X-ray transform is the analogue of the Radon on lines in $\mathbb{R}^n$.

The primary reason of the present contribution was to settle the unitarization issue in the setup of noncompact symmetric spaces in all details, in a self-contained and accessible way to the readers that have little experience with the heavy machinery of semisimple groups. We do make use of the basic Lie theoretic notions but avoid as much as possible to make extensive use of the full body of the theory. Rather, we collect all the most relevant results of the theory that may serve as a map.

We are not aware of a general statement such as our Theorem 39 in the literature, though it is quite clear to us that the result comes as no surprise if not for the flexibility of our proof (see once again [4, 5]). We also believe that the material presented here is a readable introduction to a subject that may attract the attention of a wide community of young researchers.

The chapter is organized as it follows. In Sect. 2 we recall the basic facts of the analysis on semisimple Lie groups and we introduce the notation used throughout in the geometric analysis on noncompact Riemannian symmetric spaces. In Sect. 3 we present a brief overview of the general theory of symmetric spaces, illustrating it with the examples of the Euclidean space, the sphere, the upper half plane, the unit disk, and the positive definite symmetric matrices. Of particular interest for our

purposes are Sects. 3.3, 3.4, and 3.5. In Sect. 3.3 we present the notion of boundary of a symmetric space and in Sect. 3.4 we show the infinitely many ways to represent it by changing the reference point in the symmetric space. Finally, in Sect. 3.5 we define the family of horocycles and we prove some technical results needed in Sects. 4 and 5. In Sect. 4 we collect the analytic ingredients that come into play, we endow the symmetric space, its boundary and the family of horocycles with invariant measures and we introduce the Helgason–Fourier transform discussing its main features. Then, we study the horocyclic Radon transform and we discuss its relation with the Helgason–Fourier transform. Finally, in Sect. 5 we prove the unitarization result for the horocyclic Radon transform.

2 Preliminaries

The purpose of this introductory section is to recall the basic facts of the analysis on semisimple Lie groups and to establish the notation used throughout in the geometric analysis on noncompact Riemannian symmetric spaces. For a concise and effective exposition, see [16]. Classical references with a wider scope are [15, 17, 22]. For a detailed introduction to differential geometry and Lie groups, we refer to [27].

A *Lie algebra* $\mathfrak{g}$ is *simple* if it is not Abelian and contains no proper Abelian ideals. A *semisimple* Lie algebra is then the Lie algebra direct sum of (all) its simple ideals. Cartan proved that on every semisimple Lie algebra $\mathfrak{g}$ there exists a *Cartan involution* θ, namely, an involution such that the symmetric bilinear form $B_\theta(X, Y) = -B(X, \theta Y)$ is positive definite, where B is the usual Killing form defined by $B(X, Y) = \mathrm{tr}(\mathrm{ad}X \circ \mathrm{ad}Y)$. Such an involution gives rise to a *Cartan decomposition* of the Lie algebra, namely, a vector space direct sum $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$, where $\mathfrak{k}$ and $\mathfrak{p}$ are the $+1$ and -1 eigenspaces of $\mathfrak{g}$ relative to θ, respectively.

Fix a maximal Abelian subspace $\mathfrak{a}$ of $\mathfrak{p}$. The set $\{\mathrm{ad}H : H \in \mathfrak{a}\}$ is a commuting family of self-adjoint linear maps. Therefore, $\mathfrak{g}$ is the B_θ-orthogonal direct sum of their joint eigenspaces, all of the eigenvalues of which are real and depend linearly on H. For any fixed $\alpha \in \mathfrak{a}^*$, the linear dual of $\mathfrak{a}$, we write

$$\mathfrak{g}_\alpha = \{X \in \mathfrak{g} : (\mathrm{ad}H)X = \alpha(H)X \text{ for all } H \in \mathfrak{a}\}$$

and we say that $\alpha \neq 0$ is a *restricted root*, or simply a *root* of the pair $(\mathfrak{g}, \mathfrak{a})$, whenever $\mathfrak{g}_\alpha \neq \{0\}$. The set of restricted roots is Σ and the spaces $\mathfrak{g}_\alpha$ with $\alpha \in \Sigma$ are called *(restricted) root spaces*.

An element $H \in \mathfrak{a}$ is called *regular* if $\alpha(H) \neq 0$ for all $\alpha \in \Sigma$, otherwise it is *singular*. The set $\mathfrak{a}'$ of regular elements is the complement in $\mathfrak{a}$ of finitely many hyperplanes and its connected components are called the *Weyl chambers*.

We fix a Weyl chamber $\mathfrak{a}^+ \subset \mathfrak{a}$ and we declare a root α to be *positive* if it has positive values on $\mathfrak{a}^+$. A root is *simple* if it cannot be written as the sum of positive

roots. The set Δ of simple roots turns out to be a basis of $\mathfrak{a}^*$. Thus, there are exactly $\ell = \dim \mathfrak{a}$ simple roots. This number is an important invariant and is called the *real rank* of $\mathfrak{g}$. We order the elements in $\mathfrak{a}^*$, hence the roots in Σ, *lexicographically* with respect to an ordering $\delta_1, \ldots, \delta_\ell$ of the simple roots. This means that $\lambda = \sum a_j \delta_j$ is positive (written $\lambda > 0$) if the first non-zero coefficient a_k is positive. Together with $\mathfrak{g}$, θ, and $\mathfrak{a}$ we assume that an ordering ">" has been fixed on $\mathfrak{a}^*$ by choosing a labeling of the simple roots relative to a fixed Weyl chamber $\mathfrak{a}^+$. We consequently denote by Σ^+ and Σ^- the positive and negative roots, respectively. Clearly, $\Sigma = \Sigma^+ \cup \Sigma^-$, a disjoint union.

If G is a Lie group, then it is said to be *semisimple* if such is its Lie algebra. Furthermore, for any Cartan involution θ on its Lie algebra $\mathfrak{g}$ there exists an automorphism Θ of G such that $d\Theta = \theta$ and $\Theta^2 = \mathrm{Id}$.

Theorem 2 (The *Iwasawa decomposition*) *Let G be a connected semisimple Lie group, $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ be a Cartan decomposition of its Lie algebra and fix a maximal Abelian subspace $\mathfrak{a}$ of $\mathfrak{p}$ and an ordering on $\mathfrak{a}^*$. The vector space direct sum*

$$\mathfrak{n} = \sum_{\alpha \in \Sigma^+} \mathfrak{g}_\alpha \tag{4}$$

is a nilpotent Lie algebra and $\mathfrak{g}$ decomposes as the vector space direct sum

$$\mathfrak{g} = \mathfrak{k} + \mathfrak{a} + \mathfrak{n}.$$

Furthermore, let K, A and N be the connected subgroups of G whose Lie algebras are $\mathfrak{k}$, $\mathfrak{a}$ and $\mathfrak{n}$, respectively. The multiplication map $K \times A \times N \to G$ given by $(k, a, n) \mapsto kan$ is a diffeomorphism. The groups A and N are simply connected and AN is solvable.

Observe that AN is in fact a semidirect product. Indeed, A acts on N by conjugation, as is most rapidly seen by observing that $\mathrm{Ad}a(X) \in \mathfrak{g}_\alpha$ if $X \in \mathfrak{g}_\alpha$ for any root $\alpha \in \Sigma$ and for all $a \in A$. Indeed, for any $H \in \mathfrak{a}$, since $\mathfrak{a}$ is Abelian, one has

$$[H, \mathrm{Ad}a(X)] = \mathrm{Ad}a\left([\mathrm{Ad}a^{-1}(H), X]\right) = \mathrm{Ad}a\left([H, X]\right) = \alpha(H)\mathrm{Ad}a(X).$$

Therefore $\mathrm{Ad}a$ preserves root spaces and in particular it preserves $\mathfrak{n}$. Thus A acts on $\mathfrak{n}$ via the adjoint action and, passing to exponentials, it acts on N by conjugation. This is tantamount to saying that A normalizes N inside G. Hence $NA = AN$ is the semidirect product $N \rtimes A$.

Let M and M' denote the *centralizer* and *normalizer* of $\mathfrak{a}$ in K, respectively. This means that

$$M = \left\{m \in K : \mathrm{Ad}m(H) = H \text{ for all } H \in \mathfrak{a}\right\}$$
$$M' = \left\{w \in K : \mathrm{Ad}w(H) \in \mathfrak{a} \text{ for all } H \in \mathfrak{a}\right\}.$$

Passing to exponentials, it follows that if $m \in M$, then $mam^{-1} = a$ for all $a \in A$ and if $w \in M'$, then $waw^{-1} \in A$ for all $a \in A$. The quotient group $W = M'/M$ is called the *Weyl group* of (G, K). The compact Lie groups M and M' have the same Lie algebra, namely, $\mathfrak{m}$, so that W is in fact a finite group. The Weyl group W acts on Σ by

$$(w \cdot \alpha)(H) = \alpha(\mathrm{Ad}w^{-1}H), \qquad H \in \mathfrak{a}. \tag{5}$$

The very same formula defines an action on the whole dual space $\mathfrak{a}^*$. It is worth observing that the action of W on $\mathfrak{a}^*$ maps Weyl chambers in Weyl chambers in a free and transitive way, so that the cardinality of the Weyl chambers is $|W|$. For any $\alpha \in \Sigma$, the vector space dimension of $\mathfrak{g}_\alpha$ is called the *multiplicity* of α and is usually denoted by m_α. The following element of $\mathfrak{a}^*$ plays a crucial role in the theory:

$$\rho = \frac{1}{2} \sum_{\alpha \in \Sigma^+} m_\alpha \alpha. \tag{6}$$

This linear functional on $\mathfrak{a}$ naturally appears in relation with the semidirect product structure of the Iwasawa group AN, see (29).

Example: the decomposition of $\mathrm{SL}(d, \mathbb{R})$. We consider the Lie algebra $\mathfrak{g} = \mathfrak{sl}(d, \mathbb{R})$ of $G = \mathrm{SL}(d, \mathbb{R})$, namely,

$$\mathfrak{sl}(d, \mathbb{R}) = \{X \in \mathfrak{gl}(d, \mathbb{R}) : \mathrm{tr}X = 0\}.$$

The Cartan decomposition associated to the standard involution $\theta(X) = -{}^tX$ reads

$$\mathfrak{sl}(d, \mathbb{R}) = \mathfrak{so}(d, \mathbb{R}) + \mathrm{Sym}_0(d),$$

where $\mathfrak{p} = \mathrm{Sym}_0(d)$ is the space of $d \times d$ symmetric and traceless real matrices. The Cartan involution Θ for $\mathrm{SL}(d, \mathbb{R})$ is then

$$\Theta g = {}^tg^{-1}$$

as for all matrix groups with real entries. Hence $K = \mathrm{SO}(d)$, a maximal compact subgroup of $\mathrm{SL}(d, \mathbb{R})$. The diffeomorphism $(k, X) \mapsto k \exp X$ of $\mathrm{SO}(d) \times \mathrm{Sym}_0(d) \mapsto G$ is just the classical polar decomposition. The center of $\mathrm{SL}(d, \mathbb{R})$ is the identity matrix if d is odd and $\{\pm \mathrm{Id}\}$ if d is even. The natural maximal Abelian subspace of $\mathrm{Sym}_0(d)$ is the $(d-1)$-dimensional vector space consisting of the diagonal matrices $\mathrm{diag}(a_1, \ldots, a_d)$ with $a_1 + \cdots + a_d = 0$. Thus, the real rank of $\mathfrak{sl}(d, \mathbb{R})$ is $d - 1$. Let E_{ij} denote the matrix whose only non-zero entry is 1 at position (i, j). Then, for $H = \mathrm{diag}(a_1, \ldots, a_d)$ and $i \neq j$

$$[H, E_{ij}] = (a_i - a_j)E_{ij}$$

and in fact E_{ij} spans a root space provided that $i \neq j$. It is customary to introduce the linear functionals $e_k(\cdot)$ on $\mathfrak{a}$, with $1 \leq k \leq d$, via $e_k(\mathrm{diag}(a_1, \ldots, a_d)) = a_k$. Thus, for $i \neq j$ the (restricted) root $\alpha_{ij} = e_i - e_j$ acts on $H = \mathrm{diag}(a_1, \ldots, a_d)$ by

$$\alpha_{ij}(H) = a_i - a_j.$$

and we write in the simplified form $\mathfrak{g}_{ij}$ in place of $\mathfrak{g}_{\alpha_{ij}}$ for the root space

$$\mathfrak{g}_{ij} = \mathrm{sp}\{E_{ij}\}, \qquad i \neq j.$$

For $i < j$ the matrix E_{ij} is upper triangular, and for $i > j$ it is lower triangular. A natural choice of Weyl chamber is

$$\mathfrak{a}^+ = \Big\{\mathrm{diag}(a_1, \ldots, a_d) : a_1 > a_2 > \cdots > a_d\Big\}.$$

It is immediate to check that for $j = 1, \ldots, d-1$ the roots $\delta_j = e_j - e_{j+1}$ are the simple ones and that the set of positive roots is

$$\Sigma^+ = \{\alpha_{ij} : i < j\}.$$

It follows that the nilpotent Iwasawa Lie algebra $\mathfrak{n}$ defined in (4) is just the Lie algebra of strictly upper triangular matrices. Notice that $\mathfrak{g}_0 = \mathfrak{a}$, that is, $\mathfrak{m} = \{0\}$ and that $\dim \mathfrak{g}_\alpha = 1$ for every restricted root $\alpha \in \Sigma$. Hence the functional ρ has the form

$$\rho(H) = \frac{1}{2}\sum_{i<j} \alpha_{ij}(H) = \frac{1}{2}\sum_{i<j}(a_i - a_j) = \sum_{j=1}^{d}(\frac{d+1}{2} - j)a_j.$$

Let A be the group of diagonal matrices with positive entries and determinant 1, namely,

$$\mathrm{diag}(\mathrm{e}^{a_1}, \ldots, \mathrm{e}^{a_d}), \qquad a_1 + \cdots + a_d = 0,$$

and let N be the group of unipotent upper triangular matrices, namely, those of the form

$$\begin{bmatrix} 1 & a_{12} & \ldots & \ldots & a_{1,d} \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & a_{d-1,d} \\ 0 & \ldots & \ldots & 0 & 1. \end{bmatrix}.$$

Then $\mathfrak{a}$ and $\mathfrak{n}$ are the Lie algebra of A and N, respectively. Hence $\mathrm{SL}(d, \mathbb{R}) = KAN$ by the Iwasawa decomposition.

3 Symmetric Spaces

Symmetric spaces are very special kinds of homogeneous spaces. The reader is assumed to be familiar with basic Differential Geometry and in particular with the main results on group actions and homogeneous spaces. The natural reference for the material in this section is the celebrated monography [15] by Helgason, of which this is a synthesis with examples. Other sources are, for example, [21, 28].

We very briefly recall the basic facts that we shall use throughout. A homogeneous space X is a transitive G-space. Saying that X is a *G-space* means that we are given a continuous map $G \times X \to X$, written $(g, x) \mapsto gx$ and called an action of G on X, which satisfies

(i) $x \mapsto gx$ is a homeomorphism of X for each $g \in G$,
(ii) $g(hx) = (gh)x$ for all $g, h \in G$ and $x \in X$.

The G-space X is called *transitive* if for every $x, y \in X$ there exists $g \in G$ such that $gx = y$. In this case X is identified with G/H through the action of G, where H is the isotropy subgroup at some point $x_0 \in X$, namely,

$$H = \{g \in G : gx_0 = x_0\}.$$

This identification depends on the choice of the reference point $x_0 \in X$ and is given by the bijection

$$G/H \to X, \qquad gH \mapsto gx_0.$$

If we choose a different reference point $x_0' = g_0 x_0$ for some $g_0 \in G$, it is sufficient to replace H with $H' = g_0 H g_0^{-1}$. The map $g \mapsto g_0 g g_0^{-1}$ induces a G-equivariant homeomorphism between G/H and G/H'. If the topology on G/H is the quotient topology then the identification map is actually a homeomorphism.

In the present contribution, we often consider different G-spaces of the same group. For clarity, we shall thus adopt notational variations to distinguish among different actions, such as $g[x]$ or $g.x$ or $g \cdot x$ or $g\langle x \rangle$, and so forth.

3.1 Riemannian Globally Symmetric Spaces

Let $\mathcal{M}$ be a Riemannian manifold and let $I(\mathcal{M})$ denote the group of isometries of $\mathcal{M}$. We shall endow $I(\mathcal{M})$ with the compact-open topology, the smallest topology in which all the sets

$$W(C, U) = \{g \in I(\mathcal{M}) : g(C) \subset U\}$$

are open, where C varies in the compacta of $\mathcal{M}$ and U in the open sets.

Theorem 3 (Theorem 2.5, Chap. IV, [15]) *Let $\mathcal{M}$ be a Riemannian manifold.*

(i) The group of isometries $I(\mathcal{M})$ with the compact-open topology is a locally compact topological group acting on $\mathcal{M}$.
(ii) The isotropy subgroup of $I(\mathcal{M})$ at any point of $\mathcal{M}$ is compact.

Definition 4 The Riemannian manifold $\mathcal{M}$ is a Riemannian globally symmetric space if each $p \in \mathcal{M}$ is an isolated fixed point of an isometry σ_p of $\mathcal{M}$ that is involutive ($\sigma_p^2 = \mathrm{Id}$).

It may be shown that each σ_p is in this case unique and that there exists a neighborhood N_p of p in which σ_p is the geodesic symmetry. This means that if $q \in N_p$ and $\gamma(t)$ is the geodesic such that $\gamma(0) = p$ and $\gamma(1) = q$, then $\sigma_p(q) = \gamma(-1)$.

Euclidean space. Let $\mathcal{M} = \mathbb{R}^n$ and fix $p \in \mathbb{R}^n$. The globally defined map $\sigma_p(x) = 2p - x$ is clearly involutive and isometric with respect to the Euclidean distance because $\|\sigma_p(x) - \sigma_p(y)\| = \|y - x\|$. Further, $\sigma_p(x) = x$ if and only if $x = p$, so p is an isolated fixed point.

The sphere. Let $\mathcal{M} = S^{n-1}$ and consider the map defined on $\mathbb{R}^n$ by $x \mapsto \Omega x$ where

$$\Omega = \begin{bmatrix} 1 & \\ & -\mathrm{I}_n \end{bmatrix}.$$

Evidently, it leaves the unit sphere invariant and is an isometry with respect to the natural Riemannian structure on it. It fixes the north pole $e_0 = (1, 0, \ldots, 0)$. Next choose $p \in \mathcal{M}$ and take $R \in \mathrm{SO}(n)$ such that $p = Re_0$. Then $\sigma_p = R\Omega R^{-1}$ is the required involutive isometry, as the reader is invited to check.

The upper half plane. Let $\mathcal{M}$ denote the upper half plane, which we think of as one of the natural models of the 2-dimensional hyperbolic space. We realize it as the complex numbers with positive imaginary part. The Riemannian structure on $\mathcal{M}$ is given by the inner product

$$\langle u, v \rangle_z = \frac{(u, v)}{4y^2},$$

where $u, v \in T_z(\mathcal{M})$ are tangent vectors at $z = x + iy \in \mathcal{M}$. It is important to observe that $G = \mathrm{SL}(2, \mathbb{R})$ acts transitively on $\mathcal{M}$ by means of the *Möbius action*, namely,

$$g[z] = \begin{bmatrix} a & b \\ c & d \end{bmatrix}[z] = \frac{az + b}{cz + d}. \tag{7}$$

The imaginary part of $g[z]$ is positive if such is that of z, so that (7) is indeed an action. To show transitivity, we fix $p = b + ia \in \mathcal{M}$ with $a > 0$ and consider

$$g_p = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{a} & 0 \\ 0 & 1/\sqrt{a} \end{bmatrix} = \begin{bmatrix} \sqrt{a} & b/\sqrt{a} \\ 0 & 1/\sqrt{a} \end{bmatrix},$$

an element of the Iwasawa subgroup NA of $\mathrm{SL}(2, \mathbb{R})$. It is immediate to check that $g_p[i] = p$ and that the isotropy group at i is $K = \mathrm{SO}(2)$, so that $\mathcal{M} \simeq G/K$.

As for the isometric involutions, consider first the Möbius action induced by

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

which is the map $z \mapsto -1/z$, namely, $x + iy \mapsto (-x + iy)/(x^2 + y^2)$, and may also be described in polar coordinates by

$$\rho(\cos\theta + i\sin\theta) \mapsto \frac{1}{\rho^2}(-\cos\theta + i\sin\theta).$$

This fixes only i (for $\rho = 1$ and $\theta = \pi/2$) and is thus a global involution of which i is an isolated fixed point. A global involution fixing only the point p is given by the Möbius action of the $\mathrm{SL}(2,\mathbb{R})$ element $g_p J g_p^{-1}$.

Of course, it needs to be to seen that these maps are indeed isometries relative to the hyperbolic distance. To this end, observe that any differentiable path $\gamma : [a, b] \to \mathcal{M}$, with $\gamma(t) = x(t) + iy(t)$ has length

$$L(\gamma) = \int_a^b \langle \dot\gamma(t), \dot\gamma(t)\rangle^{1/2}\,\mathrm{d}t = \frac{1}{2}\int_a^b \frac{\sqrt{\dot x^2(t) + \dot y^2(t)}}{y(t)}\,\mathrm{d}t.$$

It is then very easy to check that $L(g[\gamma]) = L(\gamma)$ if g is either J or any of the following

$$\begin{bmatrix} \mathrm{e}^s & 0 \\ 0 & \mathrm{e}^{-s} \end{bmatrix} \in A, \qquad \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \in N.$$

We now show that these are enough. Indeed, any lower triangular unipotent matrix in G is of the form JnJ^{-1} for some $n \in N$. Next, any rotation in $\mathrm{SO}(2)$ with $\cos\theta \neq 0$ can be written

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 1 & \tan\theta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\cos\theta & 0 \\ 0 & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\tan\theta & 1 \end{bmatrix}.$$

The rotations with $\cos\theta = 0$ are of the form $\pm J$, and we conclude that any element in K is a finite product of elements[2] chosen in $\{\pm J\} \cup A \cup N$. By the Iwasawa decomposition we conclude that in fact $L(g[\gamma]) = L(\gamma)$ for any $g \in G$. This entails that $\mathrm{SL}(2,\mathbb{R})$ acts by isometries on $\mathcal{M}$. It is worth mentioning that the isometry group of the upper half plane is generated by $\mathrm{SL}(2,\mathbb{R})$ and by the map $z \mapsto 1/\overline{z}$.

The unit disk. A second natural model of the 2-dimensional hyperbolic space is the unit disk $\mathcal{M} = \{z \in \mathbb{C} : |z| < 1\}$, later denoted $\mathbb{D}$. This is the Riemannian manifold with inner product

[2] This argument is nothing else but the Bruhat decomposition of $\mathrm{SL}(2,\mathbb{R})$.

$$\langle u, v\rangle_z = \frac{(u, v)}{(1 - |z|^2)^2},$$

where $u, v \in T_z(\mathcal{M})$ are tangent vectors at $z \in \mathcal{M}$. The group

$$G = \mathrm{SU}(1, 1) := \left\{ \begin{bmatrix} a & b \\ \bar{b} & \bar{a} \end{bmatrix} : a, b \in \mathbb{C},\ |a|^2 - |b|^2 = 1 \right\}$$

acts on $\mathcal{M}$ by the very same Möbius action as given by (7). Following similar reasoning as above, we can prove that the action is transitive using the NA action on the point $0 \in \mathcal{M}$, where the Iwasawa components of G are obtained from that of $\mathrm{SL}(2, \mathbb{R})$ by conjugating within $\mathrm{SL}(2, \mathbb{C})$ first with a $\pi/4$-rotation and then with Λ^{-1}, where

$$\Lambda = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}.$$

The Iwasawa subgroups are explicitly given by

$$K = \left\{ \begin{bmatrix} \mathrm{e}^{i\theta} & 0 \\ 0 & \mathrm{e}^{-i\theta} \end{bmatrix} : \theta \in [0, 2\pi) \right\},$$
$$A = \left\{ \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix} : t \in \mathbb{R} \right\},$$
$$N = \left\{ \begin{bmatrix} 1 + is & -is \\ is & 1 - is \end{bmatrix} : s \in \mathbb{R} \right\}.$$

Of course the isotropy at $o \in \mathcal{M}$ is K and $\mathcal{M} \simeq G/K$. The reader is invited to write the isometric involutions that prove $\mathcal{M}$ to be a symmetric space. We content ourselves with remarking that the *Cayley transform* $c\colon \mathcal{M} \to \mathbb{C}$

$$c(z) = i\frac{z + i}{z - i}$$

is an isometry of the unit disc onto the upper half plane which commutes with the Möbius actions.

The positive definite symmetric matrices. The example of the upper half plane can be generalized in higher dimensions. We have already seen that there exists a diffeomorphism between the upper half plane and G/K where $G = \mathrm{SL}(2, \mathbb{R})$ and $K = \mathrm{SO}(2)$. We are going to investigate the case where $G = \mathrm{SL}(d, \mathbb{R})$, $d \geq 2$. We denote by $(\cdot\,, \cdot)$ the usual scalar product in $\mathbb{R}^d$ and we put

$$\mathrm{P}(d, \mathbb{R}) := \{p \in \mathrm{Sym}(d) : (v, pv) > 0 \text{ for every } v \in \mathbb{R}^d\},$$

the set of $d \times d$ positive definite symmetric matrices. Observe that $\mathrm{P}(d, \mathbb{R})$ is an open subset of $\mathrm{Sym}(d)$ and so it is naturally a smooth manifold. Its dimension is

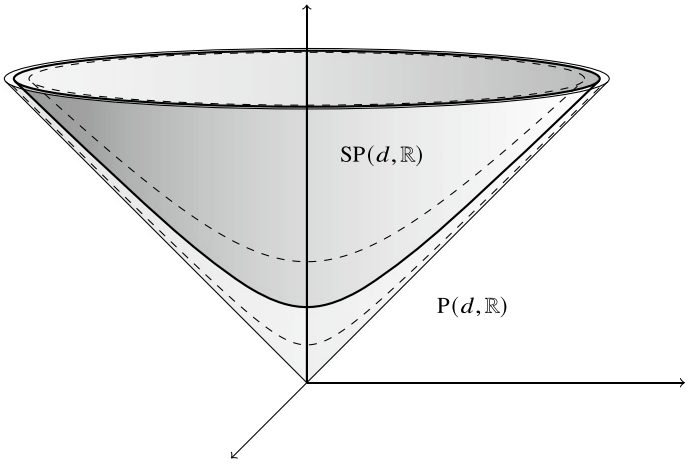


Fig. 1 The foliation of the cone $\mathrm{P}(2,\mathbb{R})$ consists of the connected components of the hyperboloids of two sheets each of which is the preimage under the determinant mapping of a positive number

$$m := \dim(\mathrm{P}(d,\mathbb{R})) = \frac{d(d+1)}{2}.$$

We show that $\mathrm{P}(d,\mathbb{R}) \subseteq \mathbb{R}^m$ is the interior of a convex cone. Let $p,\ q \in \mathrm{P}(d,\mathbb{R})$ and $t > 0$, then $tp \in \mathrm{P}(d,\mathbb{R})$, $(1-t)q \in \mathrm{P}(d,\mathbb{R})$ and also

$$tp + (1-t)q \in \mathrm{P}(d,\mathbb{R}),$$

provided that $0 \leq t \leq 1$. The boundary of $\mathrm{P}(d,\mathbb{R})$ is the set of all singular positive semidefinite matrices. It is easy to see that $\mathrm{P}(d,\mathbb{R})$ is a foliated manifold in which each leaf is the preimage of a positive number through the determinant mapping. The preimage of 1 under the determinant mapping is denoted by

$$\mathrm{SP}(d,\mathbb{R}) := \mathrm{P}(d,\mathbb{R}) \cap \mathrm{SL}(d,\mathbb{R}).$$

The group $\mathrm{GL}(d,\mathbb{R})$ acts on $\mathrm{P}(d,\mathbb{R})$ by the action

$$(g,\, p) \mapsto gp\,{}^t g =: g[p]. \tag{8}$$

We next show that the action is transitive. By the spectral theorem, for every $p \in \mathrm{P}(d,\mathbb{R})$ there exist $O \in \mathrm{SO}(d,\mathbb{R})$ and a diagonal matrix D with positive entries on the diagonal such that $p = O^{-1}DO$. Since D has positive entries on the diagonal, we can take its square root $D^{\frac{1}{2}}$. Let $g = O^{-1}D^{\frac{1}{2}}O$, then $g = {}^t g$ and

$$p = g\,{}^t g = g[\mathrm{I}_d]\ ,$$

which proves that the action is transitive. The stabilizer at $\mathrm{I}_d \in \mathrm{P}(d, \mathbb{R})$ is

$$\mathrm{O}(d, \mathbb{R}) := \{g \in \mathrm{GL}(d, \mathbb{R}) : g\,{}^t g = \mathrm{I}_d\}\ .$$

Hence we have the diffeomorphism

$$\mathrm{P}(d, \mathbb{R}) \simeq \mathrm{GL}(d, \mathbb{R})/\mathrm{O}(d, \mathbb{R}).$$

The submanifold $\mathrm{SP}(d, \mathbb{R})$ is stable under the restriction of the previous action to $\mathrm{SL}(d, \mathbb{R})$, whose action on $\mathrm{SP}(d, \mathbb{R})$ is transitive. The stabilizer of I_d is $\mathrm{SO}(d, \mathbb{R})$, so

$$\mathrm{SP}(d, \mathbb{R}) \simeq \mathrm{SL}(d, \mathbb{R})/\mathrm{SO}(d, \mathbb{R}).$$

Now we analyze the Riemannian structure on $\mathrm{P}(d, \mathbb{R})$, using [25] as main reference. First of all, we observe that if $p \in \mathrm{P}(d, \mathbb{R})$, then $T_p\mathrm{P}(d, \mathbb{R}) \simeq \mathrm{Sym}(d)$. We define

$$\langle X, Y\rangle_p := \mathrm{tr}(p^{-1}Xp^{-1}Y), \tag{9}$$

where $X,\ Y \in T_p(\mathrm{P}(d, \mathbb{R}))$. It is easy to see that $\langle\cdot,\ \cdot\rangle_p$ is an inner product. We check that the $\mathrm{GL}(d, \mathbb{R})$-action preserves this form. Let $g \in \mathrm{GL}(d, \mathbb{R})$. Then by (8) and (9)

$$\begin{aligned}\langle \mathrm{d}g(X), \mathrm{d}g(Y)\rangle_{g.p} &= \langle gX\,{}^t g, gX\,{}^t g\rangle_{g.p}\\ &= \mathrm{tr}({}^t g^{-1}p^{-1}Xp^{-1}Y\,{}^t g)\\ &= \mathrm{tr}(p^{-1}Xp^{-1}Y) = \langle X, Y\rangle_p,\end{aligned}$$

because the trace is invariant under conjugation. Hence the Riemannian structure on $\mathrm{P}(d, \mathbb{R})$ defined in (9) is $\mathrm{GL}(d, \mathbb{R})$-invariant. Now, take $p \in \mathrm{P}(d, \mathbb{R})$ and define the mapping $\sigma_p : \mathrm{P}(d, \mathbb{R}) \to \mathrm{P}(d, \mathbb{R})$ by

$$\sigma_p(q) = pq^{-1}p = pq^{-1}\,{}^t p.$$

Clearly, $\sigma_p(p) = p$ and $\sigma_p^2(q) = q$ for every $q \in \mathrm{P}(d, \mathbb{R})$. It remains to show that p is an isolated fixed point for σ_p. Let $q \in \mathrm{P}(d, \mathbb{R})$ be another nearby fixed point, that is $pq^{-1}p = q$. Thus, there exist $Y \in \mathfrak{g}$ and a small $t > 0$ such that $q = p\exp(tY)$. Hence

$$p(p\exp(tY))^{-1}p = p\exp(tY),$$

that is $\exp(-tY) = \exp(tY)$. If t is smaller than the radius of the ball in which the exponential mapping is injective, this implies $Y = 0$ and so $q = p$. We have proved that $\mathrm{P}(d, \mathbb{R})$ is a symmetric space. Observe that if $p \in \mathrm{SP}(d, \mathbb{R})$, then $\sigma_p(\mathrm{SP}(d, \mathbb{R})) = \mathrm{SP}(d, \mathbb{R})$ and so $\mathrm{SP}(d, \mathbb{R})$ with the Riemannian metric restricted from $\mathrm{P}(d, \mathbb{R})$ is a symmetric space, too.

In the special case $d = 2$, the symmetric space $\mathrm{SP}(2, \mathbb{R})$ is isomorphic to the unit disk, in fact it is one of the possible realizations of the hyperbolic space $\mathbb{H}^1$. It is important to observe that for a general $d > 2$ there are no isometries between $\mathbb{H}^d = \mathrm{SO}(d, 1)/\mathrm{SO}(d)$ and $\mathrm{SP}(d, \mathbb{R})$, because the former has constant curvature while the latter has not.

The next results establish that there the Riemannian globally symmetric spaces are completely described by Lie algebraic data.

Proposition 5 (Lemma 3.2, Chap. IV, [15]) *Let $\mathcal{M}$ be a Riemannian globally symmetric space. Then $I(\mathcal{M})$ has a smooth structure compatible with the compact-open topology which makes it a Lie group.*

Theorem 6 (Theorem 3.3, Chap. IV, [15]) *Let $\mathcal{M}$ be a Riemannian globally symmetric space, $p_0 \in \mathcal{M}$, $G = I_0(\mathcal{M})$, the connected component of the identity of $I(\mathcal{M})$.*

(i) The isotropy subgroup K of G at p_0 is compact, and $\mathcal{M} \simeq G/K$ under the map $gK \mapsto g[p_0]$.

(ii) The map $\sigma : g \mapsto s_{p_0} g s_{p_0}$ is an involutive automorphism of G such that K lies between the closed group K_σ of the fixed points of σ and its identity component. The subgroup K contains no normal subgroups other than $\{e\}$.

(iii) Let $\mathfrak{g}$ be the Lie algebra of G and $\mathfrak{k}$ be the Lie algebra of K. Then

$$\mathfrak{k} = \left\{ X \in \mathfrak{g} : (\mathrm{d}\sigma_e) X = X \right\}$$

and if

$$\mathfrak{p} = \left\{ X \in \mathfrak{g} : (\mathrm{d}\sigma_e) X = -X \right\}$$

then $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ as vector space direct sum. Let π denote the natural projection $G \to G/K$. Then $\mathrm{d}\pi_e$ maps $\mathfrak{k}$ into $\{0\}$ and $\mathfrak{p}$ isomorphically onto $T_{p_0}\mathcal{M}$. If $X \in \mathfrak{p}$, then the geodesic emanating from p_0 with tangent vector $\mathrm{d}\pi_e(X)$ is given by

$$\gamma_{\mathrm{d}\pi_e(X)}(t) = \exp tX \cdot p_0.$$

Moreover, if $Y \in T_{p_0}\mathcal{M}$, then $(\mathrm{d} \exp tX)_{p_0} Y$ is the parallel translate of Y along the geodesic.

Definition 7 Let G be a connected Lie group and H a closed subgroup. The pair (G, H) is called a *symmetric pair* if there exists an involutive analytic automorphism σ of G, briefly called an *involution*, such that

$$(\mathrm{Fix}(\sigma))_0 \subset H \subset \mathrm{Fix}(\sigma).$$

If in addition the group $\mathrm{Ad}_G(H)$ is compact, then (G, H) is called a *Riemannian symmetric pair*.

Proposition 8 (Propositions 3.4 and 3.5, Chap. IV, [15]) *Let (G, K) be a Riemannian symmetric pair, $\pi : G \to G/K$ the projection, $o = \pi(e)$. Let σ be any involution of G such that $(\mathrm{Fix}(\sigma))_0 \subset K \subset \mathrm{Fix}(\sigma)$. In each G-invariant Riemannian structure Q on G/K, and such Q do exist, the manifold G/K is a Riemannian globally symmetric space. The geodesic symmetry σ_o satisfies*

$$\sigma_o \circ \pi = \pi \circ \sigma, \qquad \tau(\sigma(g)) = \sigma_o \tau(g) \sigma_o,$$

where $\tau(g) : G/K \to G/K$ is the natural action of g, namely, $\tau(g)xK = gxK$. In particular σ_o is independent of the choice of Q. Finally, if $\mathfrak{z}$ is the Lie algebra of the center of G and $\mathfrak{k} \cap \mathfrak{z} = \{0\}$, then there exists exactly one involution σ of G such that $(\mathrm{Fix}(\sigma))_0 \subset K \subset \mathrm{Fix}(\sigma)$.

The previous two results may be condensed in the statement that there is a bijective correspondence between Riemannian globally symmetric spaces and Riemannian symmetric pairs.

3.2 Types of Symmetric Spaces

The next step in the general theory of symmetric spaces is to look at the Lie algebra level. This is suggested by Theorem 7, which shows that a Riemannian globally symmetric space gives rise to a pair $(\mathfrak{g}, s)$, where $s = \mathrm{d}\sigma_e$, that satisfies

(i) $\mathfrak{g}$ is a real Lie algebra;
(ii) s is an involutive automorphism of $\mathfrak{g}$;
(iii) the fixed points $\mathfrak{k}$ of s form a Lie algebra compactly contained in $\mathfrak{g}$,

where (iii) holds because K is compact (see Chap. II in [15] for the definition of compactly embedded Lie subalgebra).

A pair $(\mathfrak{g}, s)$ satisfying (i), (ii), and (iii) above is called an *orthogonal symmetric Lie algebra*. If in addition

(iv) $\mathfrak{k} \cap \mathfrak{z} = \{0\}$,

then $(\mathfrak{g}, s)$ is called *effective*. Fix such a pair and consider the decomposition $\mathfrak{g} = \mathfrak{u} + \mathfrak{e}$ into the $+1$ and -1 eigenspaces with respect to s. Motivated by the important decomposition result stated below in Theorem 9, one introduces the following terminology:

(a) if $\mathfrak{g}$ is compact and semisimple, then $(\mathfrak{g}, s)$ is said to be of the *compact type*;
(b) if $\mathfrak{g}$ is noncompact and semisimple and if $\mathfrak{g} = \mathfrak{u} + \mathfrak{e}$ is a Cartan decomposition, then $(\mathfrak{g}, s)$ is said to be of the *noncompact type*;
(c) if $\mathfrak{e}$ is an Abelian ideal in $\mathfrak{g}$, then $(\mathfrak{g}, s)$ is said to be of the *Euclidean type*.

Theorem 9 (Theorem 1.1, Chap. V, [15]) *Suppose that $(\mathfrak{g}, s)$ is an effective orthogonal symmetric Lie algebra. Then there exist ideals $\mathfrak{g}_0$, $\mathfrak{g}_-$ and $\mathfrak{g}_+$ such that*

(i) $\mathfrak{g} = \mathfrak{g}_0 + \mathfrak{g}_- + \mathfrak{g}_+$, *a Lie algebra direct sum;*
(ii) $\mathfrak{g}_0$, $\mathfrak{g}_-$ *and* $\mathfrak{g}_+$ *are invariant under* s *and orthogonal with respect to the Killing form;*
(iii) the pairs $(\mathfrak{g}_0, s_0)$, $(\mathfrak{g}_+, s_+)$ *and* $(\mathfrak{g}_-, s_-)$ *are effective orthogonal symmetric Lie algebras of the Euclidean, compact and noncompact type, respectively.*

The involutions s_0, s_-, and s_+ are those that arise by restricting s to the corresponding ideals. The above result is of course of central importance because it allows to study separately the various cases. Clearly, the decomposition yields a corresponding decomposition of a symmetric space and thus induces the notions of symmetric space of Euclidean, compact and noncompact types. The Euclidean space, the sphere, and the unit disk, introduced in Sect. 3.1, are the prototypical examples of such spaces. There is a remarkable duality between compact and noncompact types in which we are not interested. We content ourselves with mentioning that the compact types have positive sectional curvature and the noncompact ones have negative sectional curvature.

Since we are only interested in noncompact globally symmetric spaces, we focus on the corresponding structural assumptions. To this end, we need yet another piece of terminology and we also slightly change the current notation to tune into the noncompact case. Any pair (G, K) where G is a connected Lie group with Lie algebra $\mathfrak{g}$ and where K is a Lie subgroup of G with Lie algebra $\mathfrak{k}$ is said to be associated to the (effective) orthogonal symmetric Lie algebra $(\mathfrak{g}, \theta)$, and will be called of the noncompact type if such is $(\mathfrak{g}, \theta)$. Thus, from now on we fix an effective orthogonal symmetric Lie algebra $(\mathfrak{g}, \theta)$ of the noncompact type, so that the eigenspace decomposition relative to θ, namely, $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$, is a Cartan decomposition. The next result is a cornerstone in the theory.

Theorem 10 (Theorem 1.1, Chap. VI, [15]) *With the notation above, suppose that* (G, K) *is any pair associated with the effective orthogonal symmetric Lie algebra of the noncompact type* $(\mathfrak{g}, \theta)$. *Then:*

(i) K *is connected, closed and contains the center* Z *of* G. *Moreover,* K *is compact if and only if* Z *is finite. In this case,* K *is a maximal compact subgroup of* G;
(ii) there exists an involutive analytic automorphism Θ *of* G *whose fixed point set is* K *and whose differential at the identity* $e \in G$ *is* θ; *the pair* (G, K) *is a Riemannian symmetric pair;*
(iii) the mapping $\varphi\colon (X, k) \mapsto (\exp X)k$ *is a diffeomorphism of* $\mathfrak{p} \times K$ *onto* G *and the mapping* Exp *is a diffeomorphism of* $\mathfrak{p}$ *onto the globally symmetric space* G/K.

The exponential mapping Exp in item (iii) above, quoted for completeness, is just the Riemannian exponential mapping (see for instance [15]) and will play no explicit role in what follows.

Assumption. From now on, let G be a connected semisimple Lie group with finite center and $X = G/K$ the associated symmetric space of the noncompact type, where K is a maximal compact subgroup of G. We also fix an Iwasawa decomposition $G = KAN$ and we denote by M the centralizer of A in K.

3.3 Boundary of a Symmetric Space

Our basic example of noncompact symmetric space will be the unit disk $\mathbb{D}$, which has a rather obvious (topological) boundary, namely, the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$. The notion of boundary of a symmetric space is highly non-trivial. For a deep study on the matter, the reader is referred to the classical paper of Furstenberg [11] in which a detailed motivation of Definition 11 below may be found. For our purposes, some heuristics and some basic observations will suffice.

Notice first that the Möbius action of $G = \mathrm{SU}(1,1)$ on $\mathbb{C}$ has precisely three orbits, namely, $\mathbb{D}$, S^1 and the complement $\{w \in \mathbb{C} : |w| > 1\}$. We already know that $\mathbb{D}$ is an orbit. Further, AN fixes 1 (easy to check) and K moves it along the unit circle, so that the G-orbit of 1 is S^1. Finally, for $\rho > 1$ the formula

$$k_{\theta/2} \cdot \rho = \begin{bmatrix} \mathrm{e}^{i\theta/2} & 0 \\ 0 & \mathrm{e}^{-i\theta/2} \end{bmatrix} \cdot \rho = \rho \cos\theta + i\rho \sin\theta \tag{10}$$

shows that K maps the point ρ along the circle of radius ρ and any such real point may be reached, say, from 2 by means of A because for $t > 0$ the real numbers

$$a_t[2] = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix} [2] = \frac{2 + \tanh t}{2 \tanh t + 1}$$

span the half-line $(1, +\infty)$. Thus the set $\{w \in \mathbb{C} : |w| > 1\}$ is an orbit.

Let's go back to the unit circle. As already noticed, AN fixes 1 and K moves it along the circle, as can also be deduced from (10) when $\rho = 1$. The very same formula shows also that the elements $k_{\theta/2}$ when θ is any multiple of 2π fix 1. These are $\pm I$, namely, the elements of M, the centralizer of A in K. Therefore, the stabilizer of 1 is the group $P = MAN$ and $S^1 \simeq G/P$. By means of the Iwasawa decomposition we may write

$$S^1 \simeq KAN/MAN$$

and the natural question arises whether this is the same as K/M or not. In the case at hand this is quite clearly so because K acts transitively with isotropy M. This actually holds more generally in the sense that

$$G/P = KAN/MAN \simeq K/M.$$

Indeed, K acts on the coset space G/P in the natural fashion $k \cdot gP = (kg)P$ and by the Iwasawa decomposition $k \in P = MAN$ if and only if $k \in M$. Hence the isotropy at the coset $\{P\}$ is M. Further, again by the Iwasawa decomposition, the action is transitive, and we conclude that $G/P \simeq K/M$. The reverse point of view (that of G acting on K/M with isotropy P) will be illustrated below in (15), where the explicit action of G on K/M is given.

Definition 11 The *boundary* of X is the coset space $B := K/M$.

We remark here *en passant* that M, which will play an important role below, normalizes N, that is

$$mNm^{-1} = N, \qquad m \in M. \tag{11}$$

To see this, look at the Lie algebra level. If α is a positive root and $X \in \mathfrak{g}_\alpha$, then for every $H \in \mathfrak{a}$ it is

$$[H, \mathrm{Ad}mX] = \mathrm{Ad}m[\mathrm{Ad}m^{-1}H, X] = \mathrm{Ad}m[H, X] = \alpha(H)\mathrm{Ad}mX,$$

so that $\mathrm{Ad}m(\mathfrak{g}_\alpha) \subset \mathfrak{g}_\alpha$. An other normalization property that involves N is that for any $\alpha \in A$ and any $\nu \in N$ it holds

$$\alpha\nu aN = aN\alpha\nu. \tag{12}$$

This, in turn, follows from choosing $\nu' \in N$ such that $\nu'\alpha = \alpha\nu$, which gives

$$\alpha\nu aN = \alpha aa^{-1}\nu aN = \alpha aN\alpha^{-1}\alpha = a\alpha N\alpha^{-1}\alpha = aN\alpha = aN\nu'\alpha = aN\alpha\nu.$$

3.4 Changing the Reference Point

In what follows, it will be useful to change the reference point of both the symmetric space X and its boundary. Although conceptually very well known and somehow trivial, the actual explicit determination of what happens when doing so is not to be found in the literature, to the best of our knowledge. In order to see how the various decompositions are affected by changing the origin of our spaces, it is convenient to introduce Borel sections and occasionally adopt a slightly different notation for the (various) G-actions.

The action of G on $X = G/K$ will be written $g[x]$, namely,

$$g[x] = g[hK] = ghK.$$

For any fixed $x_0 \in X = G/K$, a *Borel section* relative to x_0 is a measurable map $s_{x_0}\colon X \to G$ satisfying $s_{x_0}(x)[x_0] = x$ and $s_{x_0}(x_0) = e$, with e the neutral element of G. Borel sections always exist since G is second countable, see Theorem 5.11 in [26].

We next show how, in the present context, a Borel section associated to $o = eK \in G/K$ can be determined quite explicitly. Since K is the isotropy subgroup of G at o, the map $\beta\colon gK \mapsto g[o]$ is a diffeomorphism of G/K onto X. Furthermore, by the Iwasawa decomposition of G (Theorem 2), each element of $g \in G$ can be written as the product $g = nak$ for exactly one triple $(n, a, k) \in N \times A \times K$, and the correspondence $(n, a, k) \leftrightarrow nak$ is a diffeomorphism with G. Hence each class in G/K has a representative of the form naK with unique $a \in A$ and $n \in N$, so that

the mapping $\psi : G/K \to NA$ given by $naK \mapsto na$ is a diffeomorphism. It follows that the measurable, actually smooth, map

$$\psi \circ \beta^{-1} : X \longrightarrow NA$$

is a Borel section. Indeed, $\psi \circ \beta^{-1}(o) = \psi(K) = e$ and, by construction, for every $x \in X$, it holds $\psi \circ \beta^{-1}(x)[o] = x$. From now on, we will denote by s_o the Borel section $\psi \circ \beta^{-1}$ with image $NA \subseteq G$.

Fix now $x \in X$ and let K_x be the isotropy of G at $x \in X$. Evidently,

$$K_x = s_o(x) K s_o(x)^{-1}.$$

It is then possible to write an Iwasawa decomposition w.r.t. the subgroup K_x. In fact,

$$G = s_o(x) G s_o(x)^{-1} = s_o(x) K A N s_o(x)^{-1} = s_o(x) K s_o(x)^{-1} A N = K_x A N,$$

because, as observed earlier, $s_o(x) \in AN$. By using the same approach, one obtains the various versions of the Iwasawa decomposition where the factors appear in a different order. It is worth observing that the subgroups A and N are independent of the maximal compact subgroup K_x, but the individual factors appearing in the decomposition of a fixed element $g \in G$ are not. Given $g \in G$, we denote with $H_x(g)$, $A_x(g)$ the elements of $\mathfrak{a}$ uniquely determined by

$$g \in K_x \exp H_x(g) N, \qquad g \in N \exp A_x(g) K_x \tag{13}$$

and by $\kappa_x(g)$ the unique element in K_x such that $g \in \kappa_x(g) AN$. Clearly,

$$A_x(g^{-1}) = -H_x(g). \tag{14}$$

Once the point $x \in X$ has been fixed, a Borel section $s_x : X \to G$ can also be fixed, so that for every $y \in X$, $s_x(y)[x] = y$ and $s_x(x) = e$. As before, it may be arranged that $s_x(y) \in NA = AN$. Also, we denote by M_x the centralizer of A in K_x, so that $M_x = s_o(x) M s_o(x)^{-1}$. The following technical observation will be useful below.

Lemma 12 *For any $x \in X$ it is*

(i) $\kappa_o \circ \kappa_x\big|_K = id_K$*; in particular, if $k_x = \kappa_x(k)$ for some $k \in K$, then $k = \kappa_o(k_x)$;*
(ii) $\kappa_x \circ \kappa_o\big|_{K_x} = id_{K_x}$.

Proof We start by proving (i). Let $k \in K$. Then according to the Iwasawa decomposition $K_x AN$ it is $k = \kappa_x(k) an$, that is $\kappa_x(k) = k(an)^{-1} \in KAN$. So that $\kappa_o(\kappa_x(k))$ is precisely k, as desired. The proof of (ii) is analogous. □

The action of G on the boundary $B = K/M$ is induced by the decomposition $G/P = KAN/MAN$ in the sense that if $g \in G$ and $kM \in B$ then

$$g\langle kM\rangle := \kappa_o(gk)M. \tag{15}$$

Consider now the action of K_x. By the definition (15) and by item (i) in Lemma 12, for any $k \in K$ it is

$$\kappa_x(k)\langle M\rangle = \kappa_o(\kappa_x(k))M = kM.$$

Thus the action of K_x on the boundary is transitive. Next, observe that an element $k_x = s_o(x)ks_o(x)^{-1}$ stabilizes $M \in K/M$ if and only if $\kappa_o(s_o(x)ks_o(x)^{-1}) \in M$, which means $s_o(x)k \in MAN$. This, together with the fact that M normalizes AN, implies that $k \in M$, hence $k_x \in M_x$. Therefore the isotropy group of K_x at M is M_x. This shows that the map induced by κ_o on K_x/M_x, which we denote $\kappa_{x,o}$, namely,

$$\kappa_{x,o} : K_x/M_x \to K/M, \qquad k_xM_x \mapsto \kappa_{x,o}(k_xM_x) := \kappa_o(k_x)M, \tag{16}$$

is a diffeomorphism. Furthermore, kM and $\kappa_x(k)M_x$ determine the same boundary point, because by (16) $\kappa_o(\kappa_x(k))M = kM$. By Lemma 12 the inverse of $\kappa_{x,o}$ is the map

$$\kappa_{o,x} : K/M \to K_x/M_x, \qquad kM \mapsto \kappa_{o,x}(kM) := \kappa_x(k)M_x.$$

3.5 Horocycles

A hyperplane in $\mathbb{R}^n$ is orthogonal to a family of parallel lines. What is a reasonable analogue of this in, say, Riemannian geometry? Since geodesics are very natural generalizations of lines, a possible answer is given by a manifold that is orthogonal to families of parallel geodesics. In the context of symmetric spaces, such manifolds will be called *horocycles*, sometimes also *horospheres*.

Let us see what this idea leads to in the context of the unit disk, our basic example of noncompact symmetric space. The origin in $\mathbb{D}$ will be denoted o. If $\gamma : [a, b] \to \mathbb{D}$ is a smooth curve with $\gamma(a) = o$ and $\gamma(b) = x \in (-1, 1)$ is a point on the real axis, then the simple inequality

$$\frac{\dot{x}(t)^2}{(1 - x(t)^2)^2} \le \frac{\dot{x}(t)^2 + \dot{y}(t)^2}{(1 - x(t)^2 - y(t)^2)^2}$$

shows that straight real lines through the origin are geodesics. We observe *en passant* that since $\gamma_0(t) = (tx, 0)$ with $t \in [0, 1]$ is such a straight line, then

$$d(o, x) = L(\gamma_0) = \int_0^1 \frac{|x|}{1 - t^2|x|}\,\mathrm{d}t = \frac{1}{2}\log\frac{1 + |x|}{1 - |x|}.$$

As we know, $G = \mathrm{SU}(1, 1)$ acts by isometries via the Möbius action on $\mathbb{D}$. Such maps are conformal and map circles and lines into circles and lines. Hence the

geodesics in $\mathbb{D}$ are circular arcs perpendicular to the boundary $|z| = 1$. All circular arcs perpendicular to the same point at the boundary may be seen as parallel lines and thus a natural notion of horocycle in this context is that of circle tangent to the boundary (except the point on S^1) because such a circle is of course perpendicular to all the above parallel geodesics.

The circle through the origin and tangent to the boundary at $1 \in \mathbb{C}$ is therefore the prototype of horocycle. Observe that

$$n_s[o] = \begin{bmatrix} 1+is & -is \\ is & 1-is \end{bmatrix}[o] = \frac{-is}{1-is} = \frac{s}{s+i} = \frac{s^2}{s^2+1} - i\frac{s}{s^2+1}$$

and an easy calculation shows that these are precisely the points on the circle of radius $1/4$ centered at $1/2 \in \mathbb{C}$ that are contained in $\mathbb{D}$. Furthermore, as $s \to \pm\infty$ one gets the boundary point $b_0 = 1 \in \mathbb{C}$. We have obtained the basic horocycle, which will be denoted ξ_o, as the N-orbit $N[o]$.

Other horocycles tangent to b_0 are the orbits $Na_t[o] = a_t N[o]$ where of course

$$a_t = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}$$

is any member of A (recall that A normalizes N). This is because

$$a_t[o] = \tanh t \in (-1, 1)$$

parametrizes any other point on the geodesic line $(-1, 1) \subset \mathbb{C}$ and an easy calculation shows that its N-orbit is just the circle through that point and tangent to b_0 (see Fig. 2). It is clear that by acting with the rotation group one gets all other horocycles, that is, all the circles in $\mathbb{D}$ tangent to the boundary. Thus, any other horocycle ξ can be written in the form $ka \cdot \xi_0$ with $k \in K$ and $a \in A$. But this means

$$\xi = (ka)N(ka)^{-1}(ka[o]),$$

which exhibits ξ as an orbit of a group conjugate to N, namely, $(ka)N(ka)^{-1}$. This motivates the Definition 13 below.

Definition 13 ([17]) A *horocycle* in X is any orbit of any subgroup of G conjugate to N, that is an orbit $N^g[x]$ where $x \in X$, $g \in G$ and $N^g = gNg^{-1}$. We shall denote by Ξ the set of all horocycles in X.

By Theorem 1.1 in Chap. II in [17], horocycles are closed submanifolds of X, the G-action on X maps horocycles to horocycles and in fact the group G acts transitively on Ξ by

$$(g, N^h[x]) \mapsto g.(N^h[x]) := gN^h[x].$$

We fix $x \in X$ and we consider the horocycle $\xi = N[x]$. By Theorem 1.1 in Chap. II in [17], the isotropy at ξ is $M_x N$ and therefore

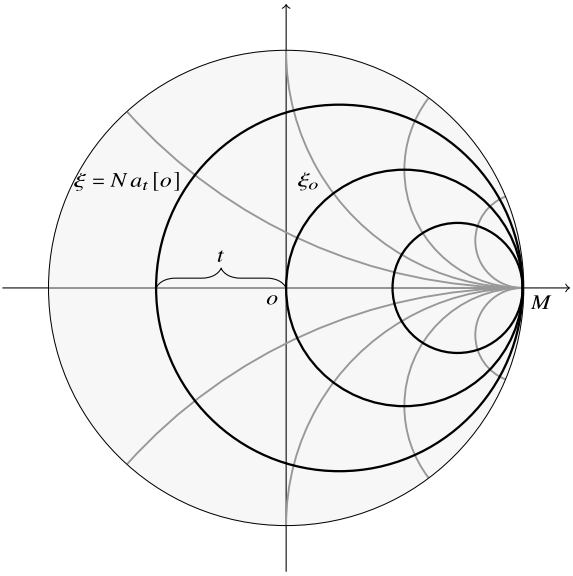


Fig. 2 The basic horocycle ξ_o in the unit disc and the horocycle ξ tangent to the boundary at 1 and with distance $-t$ from the origin o. In gray, the sheaf of parallel geodetics perpendicular to ξ_o and ξ

$$\Xi \simeq G/M_x N$$

under the diffeomorphism $gM_xN \mapsto gN[x]$. Furthermore, by Proposition 1.4 in Chap. II in [17], $(K_x/M_x) \times A$ is diffeomorphic to G/M_xN under the mapping

$$(k_x M_x, a) \mapsto k_x a M_x N. \tag{17}$$

Therefore, for each horocycle $\xi \in \Xi$ there exist unique $k_x M_x \in K_x/M_x$ and $a \in A$ such that

$$\xi = k_x a N[x]. \tag{18}$$

Finally, since K/M is diffeomorphic to K_x/M_x under the mapping $\kappa_{o,x}(kM) = \kappa_x(k)M_x$, we define the diffeomorphism

$$\Psi_x : K/M \times A \longrightarrow \Xi, \quad (kM, a) \mapsto \kappa_x(k) a N[x]. \tag{19}$$

Observe that the boundary point $kM \in K/M$ which identifies the horocycle $\xi = \kappa_x(k)aN[x]$ through (19) is independent of the choice of the reference point $x \in X$. Namely, for every $x, y \in X$

$$\Psi_x(kM, a) = \Psi_y(kM, a')$$

for some $a' \in A$. Indeed, if $\xi = k_x a N[x]$ and if we pick $y \in X$, hence $k_y M_y \in K_y/M_y$ and $a' \in A$ such that $\xi = k_y a' N[y]$, then $k_y M_y = \kappa_y(k_x)M_y$ and this identi-

fies the boundary point $\kappa_o(k_x)M$. Indeed, by the K_yAN and KAN Iwasawa decompositions of k_x, we have that

$$\kappa_y(k_x) \in k_xAN = \kappa_o(k_x)AN,$$

so that

$$\kappa_{y,o}(\kappa_y(k_x)M_y) = \kappa_o(\kappa_y(k_x))M = \kappa_o(k_x)M.$$

We shall say that $\Psi_x(kM, a)$ represents the horocycle with *normal* kM and *composite distance* $\log a$ from x (see below, Definition 15). We stress that the normal of a horocycle is independent of the choice of $x \in X$. The composite distance, however, is different for different reference points.

This parametrization generalizes the geometric picture in $\mathbb{D}$, where a horocycle $\xi = ka_tN[o]$ is identified by the boundary point $kM \in K/M$ to which it is tangent and the "signed distance" t from the reference point, see Fig. 2.

Proposition 14 *Fix a reference point $x \in X$. The horocycle through $y \in X$ with normal kM is $N^{\kappa_x(k)}[y]$.*

Proof An equivalent statement is that, writing $k = \kappa_o(k_x)$ with $k_x \in K_x$, the horocycle through y with normal $\kappa_o(k_x)M$ is $k_xNk_x^{-1}[y]$ because $k_x = \kappa_x(k)$ by item (ii) in Lemma 12.

Since $k = \kappa_o(k_x)$, then kM and k_xM_x identify the same boundary point and a horocycle with normal kM has the form $\xi = k_xaM_xN$ as in (18). If this represents a horocycle through y, then there exists $g \in G$ such that

$$\xi = gNg^{-1}[y] = \kappa_x(g)N\kappa_x(g)^{-1}[y].$$

Now observe that there exist $\alpha \in A$ and $\nu \in N$ such that $\kappa_x(g)^{-1}[y] = \nu\alpha[x]$, then $\xi = \kappa_x(g)\alpha N[x]$. Thus, since $\xi = k_xaN[x]$, we have that

$$\kappa_x(g)\alpha N[x] = k_xaN[x],$$

which by (18) implies $\kappa_x(g)M_x = k_xM_x$. Hence $\kappa_x(g) = k_xm_x$ for some $m_x \in M_x$. However, (11) implies at once that $m_xNm_x^{-1} = N$, and hence $N^{\kappa_x(g)} = N^{k_x}$. □

Definition 15 Fix a reference point $x \in X$ and choose $y \in X$ and $b \in K/M$, so that by Proposition 14 the horocycle $\xi = \xi(y, b)$ passing through y with normal $b = kM$ is uniquely determined, and hence there exists a unique $a \in A$ such that

$$\xi(y, kM) = \kappa_x(k)aN[x].$$

We denote by $A_x(y, b) \in \mathfrak{a}$ the *composite distance* of the horocycle $\xi(y, b)$ from $x \in X$, namely,

$$A_x(y, b) = \log a.$$

The reader is warned not to confuse the composite distance $A_x(y, b)$, which depends on $(y, b) \in X \times B$, with the Abelian component $A_x(g)$ of g in the Iwaswawa decomposition NAK_x, which is a function on G (see (13)). A relation between the two does exist, as pointed out in the next lemma, where we collect several properties of the composite distance which will play a crucial role in our work.

Lemma 16 *Fix a reference point $x \in X$. Then:*

(i) for any $k_x \in K_x$ and $g \in G$ we have

$$A_x(g[x], \kappa_o(k_x)M) = A_x(k_x^{-1}g), \tag{20}$$

where the right-hand side is defined by (13);

(ii) for any $y \in X$, $kM \in K/M$ and $g \in G$ we have

$$A_x(y, kM) = A_{g[x]}(g[y], g\langle kM\rangle); \tag{21}$$

(iii) for any $y, z \in X$ and $kM \in K/M$ we have

$$A_x(y, kM) = A_x(z, kM) + A_z(y, kM). \tag{22}$$

Proof (i) Let $k_x \in K_x$ and $g \in G$. By Proposition 14 and (ii) of Lemma 12, the horocycle passing through $g[x]$ with normal $\kappa_o(k_x)M$ is $k_xNk_x^{-1}g[x]$. By Definition 15, we have that

$$k_xNk_x^{-1}g[x] = k_x \exp(A_x(g[x], \kappa_o(k_x)M))N[x],$$

and so $k_x^{-1}g \in N \exp(A_x(g[x], \kappa_o(k_x)M))K_x$. This proves (i).

(ii) For simplicity, we first prove the statement in the case $x = o$. Let $y \in X$, $kM \in K/M$ and $g \in G$. By Proposition 14, and the fact that A normalizes N, the horocycle passing through $g[y]$ with normal $g\langle kM\rangle = \kappa_o(gk)M$ (see (15)) is

$$N^{\kappa_o(gk)}g[y] = \kappa_o(gk)N\kappa_o(gk)^{-1}g[y] = gkN(gk)^{-1}g[y].$$

By the diffeomorphism given in (18), there exist $h \in K_{g[o]}$ and $a \in A$ such that

$$gkNk^{-1}[y] = haNg[o], \tag{23}$$

and thus, by definition

$$a = \exp(A_{g[o]}(g[y], g\langle kM\rangle)).$$

We need to show that $a = \exp(A_o(y, kM))$. Since $K_{g[o]} = gKg^{-1}$, we have $h = gk_1g^{-1}$ for some $k_1 \in K$ and we claim that

$$k_1\kappa_o(g^{-1})M = kM. \tag{24}$$

By (23) we have that

$$k_1 g^{-1} a N s_o(g[o])[o] = k_1 g^{-1} a N g[o] = kNk^{-1}[y] = kNs_o(k^{-1}[y])[o].$$

Since s_o takes values in AN and writing the NAK decomposition of g^{-1}, there exist $a', a'' \in A$ such that

$$k_1 \kappa_o(g^{-1}) a' N[o] = k a'' N[o].$$

Hence, by (18) we have that $k_1 \kappa_o(g^{-1}) M = kM$, that is the claim (24). Therefore, for some $m \in M$ the right-hand side of (23) is

$$\begin{aligned} haNg[o] &= gkm\kappa_o(g^{-1})^{-1} g^{-1} a N g[o] \\ &= gkmaN \left(\kappa_o(g^{-1})^{-1} g^{-1}\right) g[o] \\ &= gkmaN\kappa_o(g^{-1})^{-1}[o] \\ &= gkmaN[o] = gkaN[o], \end{aligned}$$

where in the second line we have used that $\kappa_o(g^{-1})^{-1} g^{-1} \in AN$ and then (12). Summarizing, we have shown that

$$gkNk^{-1} s_o(y)[o] = gkaN[o].$$

By taking $e \in N$ on the left, there must be $n \in N$ such that $s_o(y)[o] = kan[o]$, so that $(kan)^{-1} s_o(y) \in K$, whence $k^{-1} s_o(y) \in Kan$. This shows that

$$a = \exp(A_o(k^{-1} s_o(y))) = \exp(A_o(y, kM)),$$

where the second equality follows by item (i). This concludes (ii) in the case $x = o$. The general case follows from the latter. Indeed, by applying it with $s_o(x)$ and $g s_o(x)$, respectively, in the first and the second equalities, we obtain

$$A_x(y, kM) = A_o(s_o(x)^{-1}[y], s_o(x)^{-1}\langle kM\rangle) = A_{g[x]}(g[y], g\langle kM\rangle) \ .$$

(iii) For simplicity we start by proving the statement for $x = o$, the general case follows. Fix $y, z \in X$ and $kM \in K/M$. By the definition of s_z, we have that $s_z(o)^{-1} = s_o(z)$ and $K = s_z(o) K_z s_z(o)^{-1}$. Observe that, by the $K_z AN$ Iwasawa decomposition of k

$$s_z(o)k \in s_z(o)\kappa_z(k)AN = s_z(o)\kappa_z(k)s_z(o)^{-1}AN,$$

and then

$$\kappa_o(s_z(o)k) = s_z(o)\kappa_z(k)s_z(o)^{-1}.$$

Furthermore, $s_y(o)k \in K \exp(H_o(s_y(o)k))N$, so that

$$s_z(o)kk^{-1}s_y(o)^{-1} \in s_z(o)\kappa_z(k)s_z(o)^{-1}N\exp(H_o(s_z(o)k) - H_o(s_y(o)k))K. \tag{25}$$

Now, observe that by (14) and (i) it is possible to rewrite

$$\begin{aligned} H_o(s_z(o)k) - H_o(s_y(o)k) &= A_o(k^{-1}s_y(o)^{-1}) - A_o(k^{-1}s_z(o)^{-1}) \\ &= A_o(s_y(o)^{-1}[o], kM) - A_o(s_z(o)^{-1}[o], kM) \\ &= A_o(y, kM) - A_o(z, kM). \end{aligned}$$

Hence, (25) becomes

$$s_z(o)s_y(o)^{-1} \in s_z(o)\kappa_z(k)s_z(o)^{-1}N\exp(A_o(y, kM) - A_o(z, kM))K\ ,$$

and by conjugating by $s_z(o)^{-1} \in AN$

$$\begin{aligned} s_y(o)^{-1}s_z(o) \in& \kappa_z(k)s_z(o)^{-1}N\exp(A_o(y, kM) - A_o(z, kM))Ks_z(o) \\ =& \kappa_z(k)N\exp(A_o(y, kM) - A_o(z, kM))s_z(o)^{-1}Ks_z(o) \\ =& \kappa_z(k)N\exp(A_o(y, kM) - A_o(z, kM))K_z, \end{aligned}$$

where in the first equality we use (12). Finally, we observe that $s_y(o)^{-1}s_z(o) = s_o(y)s_z(o) = s_z(y)$ and then

$$\kappa_z(k)^{-1}s_z(y) \in N\exp(A_o(y, kM) - A_o(z, kM))K_z.$$

Therefore, by item (i) of Lemma 12 and item (i) above

$$A_o(y, kM) - A_o(z, kM) = A_z(\kappa_z(k)^{-1}s_z(y)) = A_z(y, kM).$$

This proves the case $x = o$. The general case trivially follows:

$$\begin{aligned} A_x(z, kM) + A_z(y, kM) &= A_o(z, kM) - A_o(x, kM) + A_o(y, kM) - A_o(z, kM) \\ &= A_x(y, kM). \end{aligned}$$

This finishes the proof of the lemma.

□

Let $x \in X$. By Definition 15, for every $(kM, a) \in K/M \times A$ and $z \in X$

$$z \in \Psi_x(kM, a) \iff A_x(z, kM) = \log a. \tag{26}$$

Then, by (26) together with (21) it follows that

$$\begin{aligned} z \in g.\Psi_x(kM, a) &\iff g^{-1}[z] \in \Psi_x(kM, a) \\ &\iff \log a = A_x(g^{-1}[z], kM) \\ &\iff \log a = A_{g[x]}(z, g\langle kM\rangle) \\ &\iff z \in \Psi_{g[x]}(g\langle kM\rangle, a). \end{aligned}$$

Therefore

$$g.\Psi_x(kM, a) = \Psi_{g[x]}(g\langle kM\rangle, a). \tag{27}$$

Furthermore, if $y \in X$, then by (26) and (22) we have that

$$\begin{aligned} z \in \Psi_x(kM, a) &\iff \log a = A_x(z, kM) \\ &\iff \log a = A_x(y, kM) + A_y(z, kM) \\ &\iff \log(a \exp(-A_x(y, kM))) = A_y(z, kM) \\ &\iff z \in \Psi_y(kM, a \exp(A_y(x, kM))), \end{aligned}$$

where in the last equivalence we use the equality $A_y(x, kM) = -A_x(y, kM)$, which follows immediately from (22). Hence, we have

$$(\Psi_y^{-1} \circ \Psi_x)(kM, a) = (kM, a \exp(A_y(x, kM))). \tag{28}$$

Positive definite symmetric matrices. We return to positive definite symmetric matrices to describe the horocycles explicitly. Recall that the semisimple group associated to the symmetric space is in this case $G = \mathrm{SL}(d, \mathbb{R})$. As we have already seen, the Iwasawa decomposition of G is formed by $K = \mathrm{SO}(d)$, the subgroup A of diagonal matrices with positive entries on the diagonal and the subgroup N of the unit upper triangular matrix. Hence the principal horocycle is

$$\xi_0 = N[\mathrm{I}_d] = \{n\,{}^t n : n \in N\}\ .$$

Let $a = \mathrm{diag}(e^{a_1}, \ldots, e^{a_d}) \in A$, then the horocycle obtained as the N-orbit of $aK \in \mathrm{SP}(d, \mathbb{R})$ is the subset of $\mathrm{SP}(d, \mathbb{R})$ of matrices of the form

$$\left\{ e^{a_i + a_j} \sum_{k=\max(i,j)}^{d} n_{i,k} n_{j,k} \right\}_{i,j},$$

for every choice of $d(d-1)/2$ values $n_{i,j} \in \mathbb{R}$ with $j > i$, where $n_{i,i} = 1$. The subgroup $\overline{N} = \Theta(N)$ coincides with the lower unit triangular matrices. The $\overline{N}$-orbit of a positive definite diagonal matrix aK is the set of all the symmetric positive definite matrices having a^2 as diagonal matrix in the usual LD^tL decomposition. Furthermore, for every $a \in A$, we have $\overline{N}[a] = (N[a^{-1}])^{-1}$. It follows that the horocycle $N[a]$ is the subset of $\mathrm{SP}(d, \mathbb{R})$ of matrices having a^2 as diagonal matrix in the UD^tU decomposition (Fig. 3).

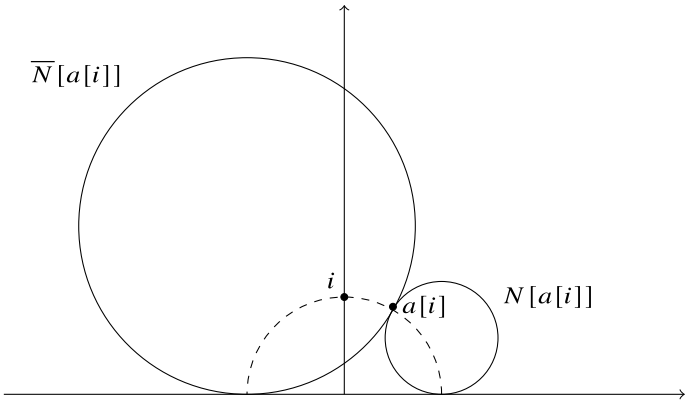


Fig. 3 In the special case $d = 2$, the N-orbit and the $\overline{N}$-orbit are tangent circles. More in general, for any $w \in W$ the intersection between the N-orbit and the (wNw^{-1})-orbit of a point coincides with the point itself, see Proposition 1.7 in Chap. II in [17]

Let $p \in \mathrm{SP}(d, \mathbb{R})$ and let $p = OD^tO$ be the spectral decomposition of p, with $O \in \mathrm{SO}(d)$ and D diagonal, and let $k \in K$. Then we have

$$k[p] = kp^tk = kOD^tO^tk,$$

and since $kO \in \mathrm{SO}(d)$ then $k[p]$ has the same eigenvalues as p. In fact, the K-orbit of $p \in \mathrm{SP}(d, \mathbb{R})$ is the subset of all the matrices in $\mathrm{SP}(d, \mathbb{R})$ with the same eigenvalues of p and if $a \in A$, then the columns of k are the eigenvectors of $k[a]$. Furthermore in each K-orbit there exists a diagonal matrix with entries ordered decreasingly on the diagonal, that is, a matrix that lies on $A_+[\mathrm{I}_d]$.

Finally, by (18), any horocycle $\xi \in \Xi$ can be written as $\xi = kaN[\mathrm{I}_d]$ for some $k \in K$ and $a \in A$. This is thus the subset of $\mathrm{SP}(d, \mathbb{R})$ of matrices having a^2 as diagonal matrix in the UD^tU decomposition w.r.t. the $\mathbb{R}^d$-basis $\{ke_i\}_{i=1,\dots,d}$, where $\{e_i\}_{i=1,\dots,d}$ is the canonical basis of $\mathbb{R}^d$.

4 Analysis on Symmetric Spaces

We collect in this section the analytic ingredients that come into play. Apart from the basic measures and function spaces, we introduce the Helgason–Fourier transform and the Radon transform and recall the results that we use throughout. The main references are [16, 17].

4.1 Measures

This section is devoted to the measures that will be involved in what follows. We first present the Haar measure and then introduce the measures on the spaces X, B, and Ξ. These are necessary in order to define the function spaces that we are interested in, among which the L^2-spaces that carry the regular representations. General references are [8] for the first part, and [16, 17] for the second.

4.1.1 Haar Measures and Modular Functions

We recall some basic definitions and results of Analysis on locally compact groups. We shall use them in the more specific context of Lie groups. A standard reference is Chap. 2 in [8].

A *topological group* is a group G endowed with a topology relative to which the group operations

$$(g, h) \mapsto gh, \qquad g \mapsto g^{-1}$$

are continuous as maps $G \times G \to G$ and $G \to G$, respectively. G is locally compact if every point has a compact neighborhood. We shall also assume our groups to be Hausdorff. In particular, all Lie groups are locally compact topological groups.

A Borel measure μ on the topological space X, that is, a measure on the σ-algebra $\mathcal{B}(X)$ of the Borel sets of X, is called a *Radon measure* if:

(i) it is finite on compact sets;
(ii) it is outer regular on the Borel sets, that is for every Borel set E

$$\mu(E) = \inf\{\mu(U) : U \supseteq E,\ U \text{ open}\}$$

(iii) it is an inner regular on the open sets, that is for every open set U

$$\mu(U) = \sup\{\mu(K) : K \subseteq E,\ K \text{ compact}\}.$$

Definition 17 A *left Haar measure* on the topological group G is a non-zero Radon measure μ such that $\mu(xE) = \mu(E)$ for every Borel set $E \subseteq G$ and every $x \in G$. Similarly for *right Haar measures*.

Of course, the prototype of Haar measure is the Lebesgue measure on the additive group $\mathbb{R}^d$, which is invariant under left (and right) translations. Compactly supported continuous functions on a topological space Y are denoted $C_c(Y)$. An equivalent definition for the left Haar measure μ is to require that for every $f \in C_c(G)$ and $h \in G$,

$$\int_G f(hg)\mathrm{d}\mu(g) = \int_G f(g)\mathrm{d}\mu(g)\ .$$

A fundamental result on Haar measures is the following theorem due to A. Weil.

Theorem 18 (Theorem 2.10, [8]) *Every locally compact group G has a left Haar measure λ, which is essentially unique in the sense that if μ is any other left Haar measure, then there exists a positive constant C such that $\mu = C\lambda$.*

If we fix a left Haar measure μ on G, then for any $g \in G$ the measure μ_g defined by

$$\mu_g(E) = \mu(Eg)$$

is again a left Haar measure. Therefore there must exist a positive real number, denoted $\Delta(g)$ such that

$$\mu_g = \Delta(g)\mu.$$

The function $\Delta : G \to \mathbb{R}_+$ is called the *modular function*. From now on, the choice of a left Haar measure μ is considered as implicitly made, and hence we write

$$\mathrm{d}g := \mathrm{d}\mu(g).$$

Proposition 19 (Proposition 2.24, [8]) *Let G be a locally compact group. The modular function $\Delta : G \to \mathbb{R}_+$ is a continuous homomorphism into the multiplicative group $\mathbb{R}_+$. Furthermore, for every $f \in L^1(G, \mu)$ we have*

$$\int_G f(gh)\mathrm{d}g = \Delta(h)^{-1} \int_G f(g)\mathrm{d}g.$$

A group for which every left Haar measure is also a right Haar measure, hence for which $\Delta \equiv 1$, is called unimodular. Large classes of groups are *unimodular*, such as the Abelian, compact, nilpotent, semisimple, and reductive groups. Many solvable groups, however, are not. Prototypical examples of non-unimodular groups are the Iwasawa NA groups, such as the affine "$ax + b$" group. A practical recipe for the computation of modular functions is given by the following proposition.

Proposition 20 (Proposition 2.30, [8]) *If G is a connected Lie group and* Ad *denotes the adjoint action of G on its Lie algebra, then $\Delta(g) = \det(\mathrm{Ad}(g^{-1}))$.*

The basic spaces X and Ξ in which we are interested are homogeneous spaces of the same group G. From the point of view of Analysis, the natural question arises whether the homogeneous space G/H admits a G-invariant Radon measure or not. The answer to this question is contained in Theorem 21 below, which relates integration on G to an iterated integral, first on H and then on G/H. These formulae are achieved by means of the natural projection operator $P : C_c(G) \to C_c(G/H)$, also known as Weil's mean operator, defined by

$$Pf(gH) = \int_H f(gh)\mathrm{d}h,$$

which is well defined by the left invariance of $\mathrm{d}h$, the Haar measure on H. Furthermore, it is possible to see that P is continuous and surjective. We are now in a position to state this classical result, also known as Weil's decomposition theorem. Here Δ_G and Δ_H are the modular functions of G and H, respectively.

Theorem 21 (Theorem 2.51, [8]) *Let G be a locally compact group and H a closed subgroup. There is a G-invariant Radon measure μ on G/H if and only if $\Delta_G|_H = \Delta_H$. In this case, μ is unique up to a constant factor, and if the factor is suitably chosen then*

$$\int_G f(g)\mathrm{d}g = \int_{G/H} Pf(gH)\mathrm{d}\mu(gH) = \int_{G/H}\int_H f(gh)\mathrm{d}h\mathrm{d}\mu(gH)\ ,$$

for every $f \in C_c(G)$.

Hence, there always exists a G-invariant Radon measure on G/H whenever H is compact, since $\Delta_G|_H = \Delta_H \equiv 1$. Indeed, the image of H under both modular functions is a compact subgroup of the multiplicative group of positive reals, namely, $\{1\}$.

Although many homogeneous spaces do not admit invariant measures (for example, $\mathbb{R}$ as a homogeneous space of the "$ax + b$" group), all of them admit strongly quasi-invariant measures. If μ is a measure on $X = G/H$ and we write $\mu^g(E) = \mu(gE)$ for $E \in \mathcal{B}(X)$, we say that μ is a *quasi-invariant measure* if all the μ^g are equivalent, that is, mutually absolutely continuous. We say that μ is *strongly quasi-invariant* if there exists a continuous function $\lambda : G \times G/H \to (0, +\infty)$ such that

$$\mathrm{d}\mu^g(x) = \lambda(g, x)\mathrm{d}\mu(x), \qquad x \in X, g \in G.$$

In other words, the requirement is that the Radon–Nikodym derivative $(\mathrm{d}\mu^g/\mathrm{d}\mu)(x)$ is jointly continuous in g and x. As mentioned, all homogeneous spaces admit strongly quasi-invariant measures (see Proposition 2.56 and Theorem 2.58 in [8]).

4.1.2 Measures on Semisimple Lie Groups of the Noncompact Type

Let G be a semisimple Lie group. By Theorem 18, there exists a (left) Haar measure on G, unique up to multiplication by a positive constant. We recall that by Theorem 2 there exist subgroups K, A, and N of G such that $G = KAN = NAK$. Since each subgroup carries a Haar measure, the natural question arises whether it is possible to write the Haar measure of G using the Haar measures of the three subgroups involved, which are all, individually, unimodular.

Since K is compact, we normalize its Haar measure in such a way that the total measure is 1. The Haar measure on A is obtained by starting from the (positive) measure that any Riemannian manifold inherits from its metric, see, e.g., Chap. I in [16]. The invariant metric is obtained by taking the restriction to $\mathfrak{a} \times \mathfrak{a}$ of the Killing form, which is positive definite on $\mathfrak{p} \times \mathfrak{p} \supset \mathfrak{a} \times \mathfrak{a}$, whereby $\mathfrak{a}$ is identified

with the tangent space to A at the identity. The standard normalization is to multiply the Riemannian measure by $(2\pi)^{-\ell/2}$, where $\ell = \dim A$. As for N, we normalize its Haar measure $\mathrm{d}n$ so that

$$\int_{\overline{N}} e^{-2\rho(H(\overline{n}))}\mathrm{d}\overline{n} = 1,$$

where $\overline{N} = \Theta(N)$ and $\mathrm{d}\overline{n}$ is the pushforward of $\mathrm{d}n$ under Θ. The convergence of the above integral is no trivial matter and is discussed in detail in [16].

Proposition 22 (Proposition 5.1, Chap. I, [16]) *Let* $\mathrm{d}k$, $\mathrm{d}a$, *and* $\mathrm{d}n$ *be left-invariant Haar measures on* K, A, *and* N, *respectively. Then the left Haar measure* $\mathrm{d}g$ *on* G *can be normalized so that*

$$\begin{aligned}\int_G f(g)\mathrm{d}g &= \int_{K\times A\times N} f(kan)e^{2\rho\log a}\mathrm{d}k\mathrm{d}a\mathrm{d}n\\ &= \int_{N\times A\times K} f(nak)e^{-2\rho(\log a)}\mathrm{d}n\mathrm{d}a\mathrm{d}k\\ &= \int_{A\times N\times K} f(ank)\mathrm{d}a\mathrm{d}n\mathrm{d}k\end{aligned}$$

for every $f \in C_c(G)$.

The case of the group AN deserves a separate comment. We recall by Sect. 2 that AN is in fact a semidirect product since A acts on N by conjugation. Furthermore, for any $H \in \mathfrak{a}$ and any root vector $X_\alpha \in \mathfrak{g}_\alpha$ it holds

$$\mathrm{Ad}(\exp H)(X_\alpha) = \mathrm{e}^{\mathrm{ad}H}(X_\alpha) = \sum_0^\infty \frac{(\mathrm{ad}H)^k}{k!}X_\alpha = \mathrm{e}^{\alpha(H)}X_\alpha.$$

It follows that, upon choosing a basis of m_α root vectors for each positive root α, it is

$$\det \mathrm{Ad}(\exp H)|_{\mathfrak{n}} = \prod_{\alpha>0}\mathrm{e}^{m_\alpha\alpha(H)}$$

or, using (6),

$$\det \mathrm{Ad}a|_{\mathfrak{n}} = \mathrm{e}^{2\rho(\log a)}.$$

Proposition 20 now entails that the modular function of the AN Iwasawa group is

$$\Delta(na) = \mathrm{e}^{-2\rho(\log a)}. \tag{29}$$

Indeed, in the computation of $\det \mathrm{Ad}(na)$ on $\mathfrak{n}+\mathfrak{a}$, all that is relevant is the action of $\mathrm{Ad}a$ on $\mathfrak{n}$ because the action of $\mathrm{Ad}a$ is unimodular on $\mathfrak{a}$ since A is Abelian, the action of $\mathrm{Ad}n$ is unimodular on $\mathfrak{n}$ because N is nilpotent, and that of $\mathrm{Ad}n$ on $\mathfrak{a}$ is again unimodular because its projection on $\mathfrak{a}$ is the identity (see also Cor. 5.2 in Chap. I in [16]).

4.1.3 Measures on X

In order to do an analysis on the symmetric space X it is important to introduce some basic functions spaces and differential operators. The reader is referred to Chap. II in [16].

A quick way to introduce differential operators on X is to say that D is such an operator if it is a linear mapping of $C_c^\infty(X)$ that decreases supports. Such operators have local nature, in the sense that it is possible to find for any coordinate patch $(\mathcal{U}, \phi)$ in X and any open set $\mathcal{W}$ with compact closure in $\mathcal{U}$ a finite number of smooth functions a_α on $\mathcal{W}$ such that

$$Df = \sum_\alpha a_\alpha (D^\alpha (f \circ \phi^{-1})) \circ \phi$$

for any $f \in C^\infty(\mathcal{W})$, where

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_d^{\alpha_d}}$$

is the standard partial derivative operator in $\mathbb{R}^d$ associated with the multi-index $\alpha \in \mathbb{Z}_+^d$. Because of this local nature, it is then possible to extend any differential operator D to $C^\infty(X)$.

On any differentiable manifold, hence on a symmetric space X, two are the most relevant spaces to consider if distribution theory is among the desirable targets. These are the space of smooth complex valued functions $\mathcal{E}(X)$ on X and the space $\mathcal{D}(X)$ of smooth complex valued functions with compact support on X. When this notation, due to Schwartz, is adopted, it is meant that these vector spaces are endowed with suitable topologies, see Chap. II in [16] for the details. We stress that in our analysis the topologies on $\mathcal{E}(X)$ and $\mathcal{D}(X)$ do not enter into play.

Now, our purpose is to determine an explicit G-invariant measure on the symmetric space $X = G/K$, whose existence is guaranteed by the fact that K is compact (see the comment after Theorem 21). Recall that, by Proposition 22, if $g = nak$, then the Haar measure of G can be normalized so that

$$\mathrm{d}g = e^{-2\rho(\log a)} \mathrm{d}n \mathrm{d}a \mathrm{d}k,$$

where $\mathrm{d}k$, $\mathrm{d}a$, and $\mathrm{d}n$ are the Haar measures on K, A and N that have been fixed in the previous paragraph.

We endow X with the G-invariant measure $\mathrm{d}x$ obtained as the pushforward of $\mathrm{d}g$ under the canonical projection $G \to G/K$. Thus, for any smooth compactly supported function $f \in \mathcal{D}(X)$

$$\int_X f(x)\mathrm{d}x = \int_G f(g[o])\mathrm{d}g = \int_{NA} f(na[o]) e^{-2\rho(\log a)} \mathrm{d}n \mathrm{d}a.$$

We henceforth denote by $L^2(X)$ the Lebesgue space of square-integrable (equivalence classes of) functions with respect to this measure. The *quasi-regular representation* π of G on $L^2(X)$ is then defined in the usual way, namely,

$$\pi(g)f(x) := f(g^{-1}[x]), \qquad f \in L^2(X),\ g \in G.$$

It is a unitary non-irreducible representation. Actually, it is possible to construct a family of Hilbert spaces in which $L^2(X)$ can be decomposed as a direct integral, whereby the restriction of π to each of them is irreducible. These are the spherical principal series representations, discussed in Chap. VI in [17]. It is also well known that π is not square-integrable.

4.1.4 Measures on the Boundary

We shall now define positive measures on the boundary B using its various possible parametrizations. Since K and M are compact subgroups of G, there exists a probability K-invariant measure μ^o on $B = K/M$, see the comment below Theorem 21. The choice of this measure is such that the Weil's decomposition holds, assuming that we normalize the Haar measure of M in such a way that the total measure is 1. For every other choice of the reference point $x \in X$ the analogous objects K_x, M_x, and μ^x can be introduced. The relation between μ^o and μ^x can be determined explicitly. We consider the diffeomorphism $T_x\colon K \to K_x$ defined by $k \mapsto s_o(x)ks_o(x)^{-1}$. Its restriction to M is a diffeomorphism between M and M_x. Hence, T_x induces the diffeomorphism $\tilde{T}_x\colon K/M \to K_x/M_x$ defined by

$$\tilde{T}_x(kM) = T_x(k)M_x = s_o(x)ks_o(x)^{-1}M_x = s_o(x)kMs_o(x)^{-1}\ .$$

Let $(\tilde{T}_x)_*(\mu^o)$ be the pushfoward of the measure μ^o under $\tilde{T}_x$. Clearly, $(\tilde{T}_x)_*(\mu^o)$ is a K_x-invariant probability measure on K_x/M_x and therefore $\mu^x = (\tilde{T}_x)_*(\mu^o)$. As we saw in (16), K_x/M_x is diffeomorphic to the boundary K/M through the map $\kappa_{x,o}\colon K_x/M_x \to K/M$. Therefore, we can consider the following K_x-invariant probability measure on the boundary $B = K/M$

$$\nu^x := (\kappa_{x,o})_*(\mu^x).$$

It is worth observing that $\nu^o = \mu^o$ and the following relation follows

$$\nu^x = (\kappa_{x,o} \circ \tilde{T}_x)_*(\nu^o)\ .$$

Lemma 23 *The measure ν^o is G-quasi-invariant. For any $F \in C(K/M)$ and $g \in G$*

$$\int_{K/M} F(g^{-1}\langle kM\rangle)\mathrm{d}\nu^o(kM) = \int_{K/M} F(kM)e^{-2\rho(H_o(gk))}\mathrm{d}\nu^o(kM). \tag{30}$$

Proof By Lemma 5.19 in Chap. I in [17], for every $H \in C(K)$ and $g \in G$,

$$\int_K H(\kappa_o(g^{-1}k))\mathrm{d}k = \int_K H(k)e^{-2\rho(H_o(gk))}\mathrm{d}k. \tag{31}$$

A function $F \in C(K/M)$ will now be regarded as an M-right invariant continuous function on K. By our choice of ν^o, Theorem 21 holds and hence

$$\begin{aligned}\int_K F(k)\mathrm{d}k &= \int_{K/M}\int_M F(kMm)\mathrm{d}m\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} F(kM)\int_M \mathrm{d}m\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} F(kM)\mathrm{d}\nu^o(kM),\end{aligned}$$

where we have used the normalization of the Haar measure of M. The function $k \mapsto F(g^{-1}\langle k\rangle) = F(\kappa_o(g^{-1}k))$ is M-invariant by $\kappa_o(g^{-1}km) = \kappa_o(g^{-1}k)m$. Since $m \in M$ commutes with A and N,

$$gkm \in \kappa_o(gk)m\exp(H_o(gk))N$$

and so $k \mapsto H_o(gk)$ is M-invariant. It follows that $k \mapsto F(k)e^{-2\rho(H_o(gk))}$ is also M-invariant. The assertion follows by applying (31) to F in place of H and then rewriting the integrals over K of the M-invariant functions as integrals over K/M w.r.t. ν^o as before. □

Now we investigate the relation between the different boundary measures introduced above. If $F \in C(K/M)$ and $x \in X$, then

$$\begin{aligned}\int_{K/M} F(kM)\mathrm{d}\nu^x(kM) &= \int_{K/M} F(\kappa_o(\tilde{T}_x(kM)))\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} F(\kappa_o(s_o(x)k)M)\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} F(kM)e^{-2\rho(H_o(s_o(x)^{-1}k))}\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} F(kM)e^{2\rho(A_o(x,kM))}\mathrm{d}\nu^o(kM)\end{aligned}$$

by Lemma 23 and then applying item (i) of Lemma 16 together with (14), since

$$-H_o(s_o(x)^{-1}k) = A_o(k^{-1}s_o(x)) = A_o(s_o(x)[o], kM) = A_o(x, kM).$$

By expressing the integral of a function on K/M with respect to either ν^x or ν^y as above and then using (22) in the form

$$A_o(x, kM) = A_o(y, kM) + A_y(x, kM),$$

the Radon–Nikodym derivative between the measures ν^x and ν^y is then

$$\frac{\mathrm{d}\nu^x}{\mathrm{d}\nu^y}(kM) = e^{2\rho(A_y(x,kM))}. \tag{32}$$

Let $x \in X$, $g \in G$ and $F \in C(K/M)$. Using first (32) with $y = o$ and then (30)

$$\begin{aligned}\int_{K/M} F(g^{-1}\langle kM\rangle)\mathrm{d}\nu^x(kM) &= \int_{K/M} F(g^{-1}\langle kM\rangle)e^{2\rho(A_o(x,kM))}\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} F(kM)e^{2\rho(A_o(x,g\langle kM\rangle))}e^{-2\rho(H_o(gk))}\mathrm{d}\nu^o(kM).\end{aligned}$$

Now observe that, by (20) and (21),

$$\begin{aligned}A_o(x, g\langle kM\rangle) - H_o(gk) &= A_{g^{-1}[o]}(g^{-1}[x], kM) + A_o(k^{-1}g^{-1})\\ &= A_{g^{-1}[o]}(g^{-1}[x], kM) + A_o(g^{-1}[o], kM)\\ &= A_o(g^{-1}[x], kM),\end{aligned}$$

the latter equality being just (22) from Lemma 16. Hence, we obtain a sort of dual relation between the G-action on the boundary and that on the reference points of the boundary measures, namely,

$$\int_{K/M} F(g^{-1}\langle kM\rangle)\mathrm{d}\nu^x(kM) = \int_{K/M} F(kM)\mathrm{d}\nu^{g^{-1}[x]}(kM). \tag{33}$$

4.1.5 Measures on Ξ

Finally, in order to develop the theory in which we are interested, we need to introduce a G-invariant measure on Ξ. We denote by σ the measure on A with density $e^{2\rho(\log a)}$ with respect to the Haar measure $\mathrm{d}a$. For every $x \in X$, we can endow Ξ with the measure $\mathrm{d}\xi$ obtained as the pushforward of the measure $\nu^x \otimes \sigma$ on $K/M \times A$ by means of the map $\Psi_x \colon K/M \times A \to \Xi$ defined in (19), i.e.,

$$\mathrm{d}\xi = \Psi_{x*}(\nu^x \otimes \sigma).$$

It turns out that $\mathrm{d}\xi$ is independent of the choice of $x \in X$. We denote by $L^1(\Xi)$ and $L^2(\Xi)$ the spaces of absolutely integrable functions and square-integrable functions with respect to the measure $\mathrm{d}\xi$, respectively. By definition, for every $F \in L^1(\Xi)$

$$\int_{\Xi} F(\xi)\mathrm{d}\xi = \int_{K/M\times A} (F\circ\Psi_x)(kM,a)\mathrm{d}(\nu^x\otimes\sigma)(kM,a)$$
$$= \int_{K/M\times A} (F\circ\Psi_x)(kM,a)e^{2\rho(\log a)}\mathrm{d}\nu^x(kM)\mathrm{d}a\ .$$

It is easy to verify that $\mathrm{d}\xi$ is G-invariant. We point out that Helgason introduced this measure w.r.t. $o\in X$, see Lemma 3.1 in Chap. II in [17]. Since in our treatment it is important to change the reference point the expression above is useful.

The group G acts on $L^2(\Xi)$ by the quasi-regular representation $\hat{\pi}\colon G\to\mathcal{U}(L^2(\Xi))$ defined by

$$\hat{\pi}(g)F(\xi) := F(g^{-1}.\xi),\qquad F\in L^2(\Xi),\ g\in G.$$

Equivalently, given $x\in X$, by (27)

$$(\hat{\pi}(g)F)\circ\Psi_x(kM,a) = F\circ\Psi_{g^{-1}[x]}(g^{-1}\langle kM\rangle,a), \tag{34}$$

for every $(kM,a)\in K/M\times A$ and $g\in G$.

We denote by $\Delta^{-\frac{1}{2}}$ the map on $K/M\times A$ defined by

$$\Delta^{-\frac{1}{2}}(kM,a) = \mathrm{e}^{\rho(\log a)}.$$

The reason for such notation resides in the fact that this function has the same expression of the inverse of the square root of the modular function of the AN Iwasawa group, see (29).

Finally, for every $x\in X$, we introduce the space $L^2_x(K/M\times A)$ of square-integrable functions on $K/M\times A$ w.r.t. the measure $\nu^x\otimes\mathrm{d}a$. For every $F\in L^2(\Xi)$, we denote by Ψ_x^*F the $(L^2(\Xi),L^2_x(K/M\times A))$-pull-back of F by Ψ_x, that is, we introduce the unitary operator $\Psi_x^*\colon L^2(\Xi)\to L^2_x(K/M\times A)$ given by

$$\Psi_x^*F(kM,a) = (\Delta^{-\frac{1}{2}}\cdot(F\circ\Psi_x))(kM,a)$$

for almost every $(kM,a)\in K/M\times A$. In order to see that Ψ_x^* is unitary, observe that for every $F\in L^2(\Xi)$ we have that

$$\int_{K/M\times A}|\Psi_x^*F(kM,a)|^2\mathrm{d}\nu^x(kM)\mathrm{d}a$$
$$= \int_{K/M\times A}|(\Delta^{-\frac{1}{2}}\cdot(F\circ\Psi_x))(kM,a)|^2\mathrm{d}\nu^x(kM)\mathrm{d}a$$
$$= \int_{K/M\times A}|(F\circ\Psi_x)(kM,a)|^2e^{2\rho(\log a)}\mathrm{d}\nu^x(kM)\mathrm{d}a$$
$$= \int_{\Xi}|F(\xi)|^2\mathrm{d}\xi = \|F\|^2_{L^2(\Xi)},$$

so that Ψ_x^* is an isometry from $L^2(\Xi)$ into $L^2_x(K/M\times A)$. Surjectivity is also clear.

4.2 The Helgason–Fourier Transform

The Helgason–Fourier transform was defined by Helgason in analogy with the Fourier transform on Euclidean spaces in polar coordinates. We briefly recall its definition and its main features.

Definition 24 (Sect. 1, Chap. III, [17]) The *Helgason–Fourier transform* of $f \in \mathcal{D}(X)$ is the function $\mathcal{H}f : K/M \times \mathfrak{a}^* \longrightarrow \mathbb{C}$ defined by

$$\mathcal{H}f(kM, \lambda) = \int_X f(x) e^{(-i\lambda+\rho)(A_o(x,kM))} \mathrm{d}x.$$

As the Euclidean Fourier transform, the Helgason–Fourier transform extends to a unitary operator on $L^2(X)$. The Plancherel measure involves the *Harish-Chandra* $\mathbf{c}$ *function*, a cornerstone in the analysis on symmetric spaces [13, 14]. It is a meromorphic function $\mathbf{c}\colon \mathfrak{a}_c^* \to \mathbb{C}$ defined on the complexified dual space $\mathfrak{a}_c^*$ for which various formulae are available (see, e.g., [18]). It may thus be restricted to the real space $\mathfrak{a}^*$. As an example, in the case of the unit disk, if $\Re(i\lambda) > 0$, then

$$\mathbf{c}(\lambda) = \pi^{-1/2} \frac{\Gamma(\frac{1}{2} i\lambda)}{\Gamma(\frac{1}{2}(i\lambda + 1))},$$

so that

$$|\mathbf{c}(\lambda)|^{-2} = \frac{\pi\lambda}{2} \tanh\left(\frac{\pi\lambda}{2}\right).$$

We denote by $L^2_{o,\mathbf{c}}(K/M \times \mathfrak{a}^*)$ the space of the functions on $K/M \times \mathfrak{a}^*$ that are square-integrable w.r.t. the measure $w^{-1} |\mathbf{c}(\lambda)|^{-2} \mathrm{d}\nu^o \mathrm{d}\lambda$, where w stands for the cardinality of the Weyl group W.

Proposition 25 (Sect. 1, Chap. III, [17]) *For every* $f_1, f_2 \in \mathcal{D}(X)$

$$\int_x f_1(x)\overline{f_2(x)}\mathrm{d}x = \int_{\mathfrak{a}^* \times K/M} \mathcal{H}f_1(kM, \lambda)\overline{\mathcal{H}f_2(kM, \lambda)}\mathrm{d}\nu^o(kM) \frac{\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2}. \quad (35)$$

The rest of the paragraph is devoted to state the Plancherel theorem for the Helgason–Fourier transform.

Property $\sharp$. We say that a function $F \in L^2_{o,\mathbf{c}}(K/M \times \mathfrak{a}^*)$ satisfies Property $\sharp$ if for every $x \in X$ the function

$$\mathfrak{a}^* \ni \lambda \longmapsto \int_{K/M} e^{(\rho+i\lambda)(A_o(x,kM))} F(kM, \lambda)\mathrm{d}\nu^o(kM) \quad (36)$$

is W-invariant almost everywhere (see the comments after (5) for the W-action on $\mathfrak{a}^*$).

We denote by $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\sharp$ the space of functions F in $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)$ satisfying Property $\sharp$. We observe that the integral in (36) is absolutely convergent for almost every $\lambda\in\mathfrak{a}^*$. By Fubini theorem, for every $F\in L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)$ we have that

$$\|F\|^2_{L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)}=\int_{\mathfrak{a}^*}\int_{K/M}|F(kM,\lambda)|^2\mathrm{d}\nu^o(kM)\frac{\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2}<+\infty.$$

Thus, the function $F(\cdot,\lambda)$ is in $L^2(K/M,\nu^o)\subseteq L^1(K/M,\nu^o)$ for almost every $\lambda\in\mathfrak{a}^*$ and, since $\rho(A_o(x,\cdot))$ is bounded on K/M, the integrability properties of $F(\cdot,\lambda)$ continue to hold for the function $e^{(\rho+i\lambda)(A_o(x,\cdot))}F(\cdot,\lambda)$.

Every function $F\in L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\sharp$ is uniquely determined by its restriction on $K/M\times\mathfrak{a}^*_+$. Here $\mathfrak{a}^*_+$ denotes the *positive Weyl chamber*

$$\mathfrak{a}^*_+=\{\lambda\in\mathfrak{a}^*:A_\lambda\in\mathfrak{a}^+\},$$

where A_λ represents λ via the Killing form, in the sense that $\lambda(H)=B(A_\lambda,H)$. If we suppose that $F,G\in L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\sharp$ are such that $F_1|_{K/M\times\mathfrak{a}^*_+}=F_2|_{K/M\times\mathfrak{a}^*_+}$, then

$$\begin{aligned}&\int_{K/M}e^{(\rho+is\lambda)(A_o(x,kM))}(F_1-F_2)(kM,s\lambda)\mathrm{d}\nu^o(kM)\\&=\int_{K/M}e^{(\rho+i\lambda)(A_o(x,kM))}(F_1-F_2)(kM,\lambda)\mathrm{d}\nu^o(kM)=0\end{aligned}$$

for a. e. $\lambda\in\mathfrak{a}^*_+$ and for every $s\in W$. Therefore, by Lemma 5.3 in Chap. II in [17], we can conclude that $F_1-F_2=0$ in $L^2_{o,\boldsymbol{c}}(K/M\times\mathfrak{a}^*)$.

By the Paley–Wiener theorem for the Helgason–Fourier transform (Theorem 5.1 in Chap. III in [17]), $\mathcal{H}f\in L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\sharp$ for every $f\in\mathcal{D}(X)$, so that $\mathcal{H}f$ is uniquely determined by its restriction on $K/M\times\mathfrak{a}^*_+$. We denote by $L^2_{o,\boldsymbol{c}}\left(K/M\times\mathfrak{a}^*_+\right)$ the space of the functions on $K/M\times\mathfrak{a}^*_+$ that are square-integrable w.r.t. the measure $|\mathbf{c}(\lambda)|^{-2}\,\mathrm{d}\nu^o\mathrm{d}\lambda$ and the Plancherel theorem for the Helgason–Fourier transform reads:

Theorem 26 (Theorem 1.5, Chap. III, [17]) *The restricted Helgason–Fourier transform $f\mapsto\mathcal{H}f|_{K/M\times\mathfrak{a}^*_+}$ extends to a unitary operator $\mathscr{H}$ from $L^2(X)$ onto $L^2_{o,\mathbf{c}}\left(K/M\times\mathfrak{a}^*_+\right)$.*

By the Plancherel formula (35), $\mathcal{H}$ is an isometry from $\mathcal{D}(X)$ into $L^2_{o,\boldsymbol{c}}(K/M\times\mathfrak{a}^*)$. Next we show that, by Theorem 26, $\mathcal{H}(\mathcal{D}(X))$ embeds densely in $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\sharp$. Let $F\in L^2_{o,\boldsymbol{c}}(K/M\times\mathfrak{a}^*)^\sharp$ be such that $\langle F,\mathcal{H}f\rangle_{L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)}=0$ for every $f\in\mathcal{D}(X)$. By Fubini theorem it follows that

$$
\begin{aligned}
0 &= \frac{1}{w}\int_{\mathfrak{a}^*}\int_{K/M} F(kM,\lambda)\overline{\int_X f(x)e^{(-i\lambda+\rho)(A_o(x,kM))}\mathrm{d}x}\mathrm{d}\nu^o(kM)\frac{\mathrm{d}\lambda}{|\mathbf{c}(\lambda)|^2}\\
&= \frac{1}{w}\int_{\mathfrak{a}^*}\int_X\int_{K/M} F(kM,\lambda)e^{(i\lambda+\rho)(A_o(x,kM))}\mathrm{d}\nu^o(kM)\overline{f(x)}\mathrm{d}x\frac{\mathrm{d}\lambda}{|\mathbf{c}(\lambda)|^2}\\
&= \int_{\mathfrak{a}_+^*}\int_X\int_{K/M} F(kM,\lambda)e^{(i\lambda+\rho)(A_o(x,kM))}\mathrm{d}\nu^o(kM)\overline{f(x)}\mathrm{d}x\frac{\mathrm{d}\lambda}{|\mathbf{c}(\lambda)|^2}\\
&= \int_{\mathfrak{a}_+^*}\int_{K/M} F(kM,\lambda)\overline{\mathcal{H}f(kM,\lambda)}\mathrm{d}\nu^o(kM)\frac{\mathrm{d}\lambda}{|\mathbf{c}(\lambda)|^2},
\end{aligned}
\tag{37}
$$

where we use that F satisfies Property $\natural$ and $|\mathbf{c}|^2$ is W-invariant. Hence, (37) yields

$$
\langle F|_{K/M\times\mathfrak{a}_+^*}, \mathcal{H}f|_{K/M\times\mathfrak{a}_+^*}\rangle_{L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}_+^*)} = \langle F|_{K/M\times\mathfrak{a}_+^*}, \mathscr{H}f\rangle_{L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}_+^*)} = 0,
$$

for every $f\in\mathcal{D}(X)$, and Theorem 26 implies that $F\equiv 0$ a.e. on $K/M\times\mathfrak{a}_+^*$. Hence, $F=0$ in $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)$ and $\mathcal{H}(\mathcal{D}(X))$ embeds densely in $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\natural$, as claimed.

The following formulation of Theorem 26 suits our needs.

Theorem 27 *The Helgason–Fourier transform $\mathcal{H}$ extends to a unitary operator $\mathscr{H}$ from $L^2(X)$ onto $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\natural$.*

In what follows, we always consider $\mathscr{H}$ as taking values in $L^2_{o,\mathbf{c}}(K/M\times\mathfrak{a}^*)^\natural$.

4.3 The Horocyclic Radon Transform

We next introduce the horocyclic Radon transform, study its range, and we investigate its intertwining properties with the quasi-regular representations π and $\hat{\pi}$ of G.

Because horocycles admit several explicit parametrizations, we define the horocyclic Radon transform appealing directly to the basic parametrization Ψ_o, as clarified in the definition that follows.

Definition 28 (Sect. 3, Chap. II, [17]) The *horocyclic Radon transform* $\mathcal{R}f$ of a function $f\in\mathcal{D}(X)$ is the map $\mathcal{R}f:\Xi\to\mathbb{C}$ defined by

$$
(\mathcal{R}f\circ\Psi_o)(kM,a) = \int_N f(kan[o])\mathrm{d}n,
$$

for every $(kM,a)\in K/M\times A$.

If we change reference point and pick $x\in X$, we may use equality (28) and obtain the equivalent definition

$$(\mathcal{R}f \circ \Psi_x)(kM, a) = (\mathcal{R}f \circ \Psi_o)(kM, a\exp(A_o(x, kM))) = \int_N f(ka\exp(A_o(x, kM))n[o])\mathrm{d}n. \tag{38}$$

Definition 29 Let $f \in \mathcal{D}(X)$. We denote by $\mathcal{A}f$ the map $\mathcal{A}f : K/M \times A \to \mathbb{C}$ defined by

$$\mathcal{A}f(kM, a) := \Psi_o^*(\mathcal{R}f)(kM, a) = (\Delta^{-\frac{1}{2}} \cdot (\mathcal{R}f \circ \Psi_o))(kM, a).$$

It is worth observing that if the function f is K-bi-invariant, then $\mathcal{A}f$ coincides with the *Abel transform* of f introduced by Helgason in Chap. III in [17].

We need to introduce the Fourier transform on the Abelian group A.

Definition 30 (Sect. 4.2, Chap. 4, [8]) Let $s \in L^1(A)$. The *Fourier transform* $\mathcal{F}s$ of s is defined on $\mathfrak{a}^*$ by

$$\mathcal{F}s(\lambda) = \int_A s(a)e^{-i\lambda(\log a)}\mathrm{d}a\ .$$

We now state a fundamental theorem in the L^2 theory of the Fourier transform.

Theorem 31 (Theorem 4.26, Chap. 4, [8]) *The Fourier transform* $\mathcal{F} : L^1 \cap L^2(A) \to C(\mathfrak{a}^*)$ *extends uniquely to a unitary operator from* $L^2(A)$ *onto* $L^2(\mathfrak{a}^*)$. *In particular,*

$$\|\mathcal{F}s\|_{L^2(\mathfrak{a}^*)} = \|s\|_{L^2(A)}.$$

We denote by R the *regular representation* of A on $L^2(A)$, which is defined for every $s \in L^2(A)$ and for every $\alpha \in A$ by

$$R_\alpha s(a) = s(\alpha^{-1}a), \quad a \in A.$$

Furthermore, we denote by M the representation of A on $L^2(\mathfrak{a}^*)$ defined for every $r \in L^2(\mathfrak{a}^*)$ and for every $\alpha \in A$ by

$$M_\alpha r(\lambda) = e^{-i\lambda(\log\alpha)}r(\lambda), \quad \lambda \in \mathfrak{a}^*.$$

Proposition 32 (Sect. 7.2, Chap. 5, [20]) *The Fourier transform* $\mathcal{F} : L^2(A) \to L^2(\mathfrak{a}^*)$ *intertwines the regular representation R with the representation M, i.e.,*

$$\mathcal{F}R_\alpha = M_\alpha\mathcal{F},$$

for every $\alpha \in A$.

We are now ready to recall the result which relates the Helgason–Fourier transform with the horocyclic Radon transform. We refer to Proposition 33 as the Fourier

Slice Theorem for the horocyclic Radon transform in analogy with the polar Radon transform, see [19] as a classical reference. For the reader's convenience, we include the proof.

Proposition 33 (Sect. 5, Chap. III, [17]) *For every $f \in \mathcal{D}(X)$ and $kM \in K/M$, the function $a \mapsto \mathcal{A}f(kM, a)$ is in $L^1(A)$ and*

$$(I \otimes \mathcal{F})\mathcal{A}f(kM, \lambda) = \mathcal{H}f(kM, \lambda), \tag{39}$$

for almost every $\lambda \in \mathfrak{a}^*$.

Proof If $f \in \mathcal{D}(X)$ and $kM \in K/M$, then by Proposition 22 and (20)

$$\begin{aligned}
\int_A |\mathcal{A}f(kM, a)|\mathrm{d}a &= \int_A e^{\rho(\log a)}|\mathcal{R}f \circ \Psi_o(kM, a)|\mathrm{d}a \\
&\leq \int_A \int_N e^{\rho(\log a)}|f(kan[o])|\mathrm{d}n\mathrm{d}a \\
&= \int_A \int_N \int_K e^{\rho(\log a)}|f(kank_1[o])|\mathrm{d}k_1\mathrm{d}n\mathrm{d}a \\
&= \int_G e^{\rho(A_o(g))}|f(kg[o])|\mathrm{d}g \\
&= \int_G e^{\rho(A_o(k^{-1}g))}|f(g[o])|\mathrm{d}g \\
&= \int_{\mathrm{supp}(f)} e^{\rho(A_o(x,kM))}|f(x)|\mathrm{d}x < +\infty.
\end{aligned}$$

Thus, $\mathcal{A}f(kM, \cdot)$ is in $L^1(A)$ and by similar steps it is easy to prove that

$$(I \otimes \mathcal{F})\mathcal{A}f(kM, \lambda) = \mathcal{H}f(kM, \lambda),$$

for almost every $\lambda \in \mathfrak{a}^*$.

□

Let $f \in \mathcal{D}(X)$. By the Paley–Wiener theorem for the Helgason–Fourier transform (Theorem 5.1 in Chap. III in [17]), $\mathcal{H}f$ is rapidly decreasing in the variable $\lambda \in \mathfrak{a}^*$ uniformly over K/M, that is for every $n \in \mathbb{N}$

$$|||\mathcal{H}f|||_n := \sup_{kM \in K/M,\, \lambda \in \mathfrak{a}^*} (1 + |\lambda|)^n |\mathcal{H}f(kM, \lambda)| < +\infty.$$

By Theorem 31 and Proposition 33, we have that

$$\begin{aligned}\int_{\Xi}|\mathcal{R}f(\xi)|^2\mathrm{d}\xi &= \int_{K/M\times A}|\Psi_o^*(\mathcal{R}f)(kM,a)|^2\mathrm{d}\nu^o(kM)\mathrm{d}a\\ &= \int_{K/M\times\mathfrak{a}^*}|(I\otimes\mathcal{F})(\Psi_o^*(\mathcal{R}f))(kM,\lambda)|^2\mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\ &= \int_{K/M\times\mathfrak{a}^*}|\mathcal{H}f(kM,\lambda)|^2\mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\ &= \int_{K/M\times\mathfrak{a}^*}\frac{(1+|\lambda|)^{2n}|\mathcal{H}f(kM,\lambda)|^2}{(1+|\lambda|)^{2n}}\mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\ &\le |||\mathcal{H}f|||_n^2\int_{\mathfrak{a}^*}\frac{1}{(1+|\lambda|)^{2n}}\mathrm{d}\lambda<+\infty,\end{aligned}$$

for every $n>\dim A/2$. Therefore, $\mathcal{R}f\in L^2(\Xi)$ for every $f\in\mathcal{D}(X)$.

The horocyclic Radon transform intertwines the regular representations π and $\hat{\pi}$ of G.

Proposition 34 *For every $g\in G$ and $f\in\mathcal{D}(X)$*

$$\mathcal{R}(\pi(g)f)=\hat{\pi}(g)(\mathcal{R}f)\ .$$

Proof Let $g\in G$ and $f\in\mathcal{D}(X)$. It is sufficient to show that $\mathcal{R}(\pi(g)f)\circ\Psi_o=\hat{\pi}(g)(\mathcal{R}f)\circ\Psi_o$ on $K/M\times A$. Let $(kM,a)\in K/M\times A$. Then

$$\begin{aligned}\mathcal{R}(\pi(g)f)\circ\Psi_o(kM,a) &= \int_N\pi(g)f(kan[o])\mathrm{d}n\\ &= \int_N f(g^{-1}kan[o])\mathrm{d}n\\ &= \int_N f(\kappa_o(g^{-1}k)\exp(H_o(g^{-1}k))an[o])\mathrm{d}n,\end{aligned}$$

where we used the decomposition $g^{-1}k\in\kappa_o(g^{-1}k)\exp(H_o(g^{-1}k))N$ and the fact that A normalizes N. Now, by (14), (20) and (22), we have

$$H_o(g^{-1}k)=-A_o(k^{-1}g)=-A_o(g[o],kM)=A_{g[o]}(o,kM)\ .$$

Finally, by $g^{-1}(kM)=\kappa_o(g^{-1}k)M$ and (38) we have that

$$\begin{aligned}\mathcal{R}(\pi(g)f)\circ\Psi_o(kM,a) &= \int_N f(\kappa_o(g^{-1}k)\exp(A_{g[o]}(o,kM))an[o])\mathrm{d}n\\ &= \int_N f(\kappa_o(g^{-1}k)\exp(A_o(g^{-1}[o],g^{-1}\langle kM\rangle))an[o])\mathrm{d}n\\ &= \mathcal{R}f\circ\Psi_{g^{-1}[o]}(g^{-1}\langle kM\rangle,a)\\ &= (\hat{\pi}(g)\mathcal{R}f)\circ\Psi_o(kM,a),\end{aligned}$$

where we used the action of G on Ξ given in (34). □

We now introduce a closed subspace of $L^2(\Xi)$ which will play a crucial role because it is the range of the unitarization of the horocyclic Radon transform. By definition, for every $x \in X$ and every $F \in L^2(\Xi)$

$$\|F\|^2_{L^2(\Xi)} = \int_{K/M} \int_A |\Psi_x^* F(kM, a)|^2 \mathrm{d}a \mathrm{d}\nu^x(kM) < +\infty.$$

So that, the function $\Psi_x^* F(kM, \cdot)$ is in $L^2(A)$ for almost every $kM \in K/M$. Then, by Plancherel formula and Fubini theorem

$$\begin{aligned}\|F\|^2_{L^2(\Xi)} &= \int_{K/M \times A} |\Psi_x^* F(kM, a)|^2 \mathrm{d}\nu^x(kM) \mathrm{d}a \\ &= \int_{K/M \times \mathfrak{a}^*} |(I \otimes \mathcal{F})\Psi_x^* F(kM, \lambda)|^2 \mathrm{d}\nu^x(kM) \mathrm{d}\lambda \\ &= \int_{\mathfrak{a}^*} \int_{K/M} |(I \otimes \mathcal{F})\Psi_x^* F(kM, \lambda)|^2 \mathrm{d}\nu^x(kM) \mathrm{d}\lambda < +\infty.\end{aligned}$$

So that, for almost every $\lambda \in \mathfrak{a}^*$ the function $(I \otimes \mathcal{F})\Psi_x^* F(\cdot, \lambda)$ is in $L^2(K/M, \nu^x) \subseteq L^1(K/M, \nu^x)$ and

$$\begin{aligned}&\left| \int_{K/M} (I \otimes \mathcal{F})\Psi_x^* F(kM, \lambda) \mathrm{d}\nu^x(kM) \right| \\ &\leq \int_{K/M} |(I \otimes \mathcal{F})\Psi_x^* F(kM, \lambda)| \mathrm{d}\nu^x(kM) < +\infty.\end{aligned}$$

Property $\flat$. We say that a function $F \in L^2(\Xi)$ satisfies Property $\flat$ if for every $x \in X$ the function

$$\mathfrak{a}^* \ni \lambda \longmapsto \int_{K/M} (I \otimes \mathcal{F})\Psi_x^* F(kM, \lambda) \mathrm{d}\nu^x(kM)$$

is W-invariant almost everywhere.

We denote by $L^2_\flat(\Xi)$ the space of functions $F \in L^2(\Xi)$ satisfying Property $\flat$. Notice that by the considerations above, the integral appearing in Property $\flat$ is finite for almost every $\lambda \in \mathfrak{a}^*$. Our main results in Sect. 5 are based on the characterization of $L^2_\flat(\Xi)$ given in Proposition 35 below. We denote by $L^2_o(K/M \times \mathfrak{a}^*)$ the space of square-integrable functions on $K/M \times \mathfrak{a}^*$ w.r.t. the measure $\nu^o \otimes \mathrm{d}\lambda$.

Proposition 35 *The operator Φ_o defined on $F \in L^2(\Xi)$ by*

$$\Phi_o F(kM, \lambda) = (I \otimes \mathcal{F})\Psi_o^* F(kM, \lambda), \qquad \text{a.e. } (kM, \lambda) \in K/M \times \mathfrak{a}^*$$

is an isometry from $L^2(\Xi)$ into $L^2_o(K/M \times \mathfrak{a}^*)$. Furthermore, a function F belongs to $L^2_\flat(\Xi)$ if and only if $\Phi_o F$ satisfies Property ♯.

Proof By Parseval identity, for every $F \in L^2(\Xi)$ we have that

$$\begin{aligned}&\int_{K/M\times\mathfrak{a}^*} |\Phi_o F(kM,\lambda)|^2 \,\mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\ &= \int_{K/M}\int_{\mathfrak{a}^*} \left|(I\otimes\mathcal{F})\Psi_o^* F(kM,\lambda)\right|^2 \mathrm{d}\lambda\mathrm{d}\nu^o(kM)\\ &= \int_{K/M\times A} \left|\Psi_o^* F(kM,a)\right|^2 \mathrm{d}\nu^o(kM)\mathrm{d}a = \|F\|^2_{L^2(\Xi)},\end{aligned}$$

so that Φ_o is an isometry from $L^2(\Xi)$ into $L^2_o(K/M\times\mathfrak{a}^*)$. Now, let $F\in L^2(\Xi)$. By equation (28) and by the definition of the regular representation R of A, for almost every $kM\in K/M$ and $\lambda\in\mathfrak{a}^*$ we have that

$$\begin{aligned}\Phi_o F(kM,\lambda) &= (I\otimes\mathcal{F})\Psi_o^* F(kM,\lambda) = (I\otimes\mathcal{F})(\Delta^{-\frac{1}{2}}\cdot(F\circ\Psi_o))(kM,\lambda)\\ &= e^{\rho(A_o(x,kM))}(I\otimes\mathcal{F})(I\otimes R_{\exp(A_x(o,kM))^{-1}})(\Delta^{-\frac{1}{2}}\cdot(F\circ\Psi_x))(kM,\lambda).\end{aligned}$$

Therefore, by Proposition 32 we obtain

$$\begin{aligned}\Phi_o F(kM,\lambda) &= e^{\rho(A_o(x,kM))}(I\otimes M_{\exp(A_x(o,kM))^{-1}})(I\otimes\mathcal{F})(\Delta^{-\frac{1}{2}}\cdot(F\circ\Psi_x))(kM,\lambda)\\ &= e^{(\rho-i\lambda)(A_o(x,kM))}(I\otimes\mathcal{F})(\Delta^{-\frac{1}{2}}\cdot(F\circ\Psi_x))(kM,\lambda)\\ &= e^{(\rho-i\lambda)(A_o(x,kM))}(I\otimes\mathcal{F})\Psi_x^* F(kM,\lambda). \qquad (40)\end{aligned}$$

Now, for every $x\in X$ and for almost every $\lambda\in\mathfrak{a}^*$, (40) yields

$$\begin{aligned}&\int_{K/M} e^{(\rho+i\lambda)(A_o(x,kM))}\Phi_o F(kM,\lambda)\mathrm{d}\nu^o(kM)\\ &= \int_{K/M} e^{(\rho+i\lambda)(A_o(x,kM))}e^{(\rho-i\lambda)(A_o(x,kM))}(I\otimes\mathcal{F})\Psi_x^* F(kM,\lambda)\mathrm{d}\nu^o(kM)\\ &= \int_{K/M}(I\otimes\mathcal{F})\Psi_x^* F(kM,\lambda)e^{2\rho(A_o(x,kM))}\mathrm{d}\nu^o(kM)\\ &= \int_{K/M}(I\otimes\mathcal{F})\Psi_x^* F(kM,\lambda)\mathrm{d}\nu^x(kM). \qquad (41)\end{aligned}$$

Equality (41) allows us to conclude that F satisfies Property ♭ if and only if $\Phi_o F$ satisfies Property ♯ and this concludes our proof.

□

Corollary 36 *For every $f\in\mathcal{D}(X)$,*

$$\Phi_o(\mathcal{R}f) = \mathcal{H}f$$

in $L^2_o(K/M \times \mathfrak{a}^*)$ *and* $\mathcal{R}f \in L^2_\flat(\Xi)$.

Proof The proof follows immediately by Proposition 33 and the fact that the Helgason–Fourier transform satisfies Property ♯. □

Some comments are in order. Proposition 35 with Corollary 36 shows the link between the range of the Radon transform with the range of the Helgason–Fourier transform, which is fundamental in our main result. The range $\mathcal{R}(\mathcal{D}(X))$ has already been completely characterized in Chap. IV in [17]. As it will be made clear in the next section, Property ♭ allows us to formulate our findings synthetically.

5 Unitarization and Intertwining

In order to obtain the unitarization for the horocyclic Radon transform that we are after, we need some technicalities. Figure 4 below might help the reader to keep track of all the spaces and operators involved in our construction.

We put

$$\mathcal{D}_o = \{\varphi \in L^2_o(K/M \times A) : (I \otimes \mathcal{F})\varphi \in L^2_{o,\mathfrak{c}}(K/M \times \mathfrak{a}^*)\}$$

and we define the operator $\mathcal{J}_o \colon \mathcal{D}_o \subseteq L^2_o(K/M \times A) \to L^2_o(K/M \times A)$ as the Fourier multiplier

$$(I \otimes \mathcal{F})(\mathcal{J}_o\varphi)(kM, \lambda) = \frac{1}{\sqrt{w}\,|\mathfrak{c}(\lambda)|}(I \otimes \mathcal{F})\varphi(kM, \lambda), \quad \text{a.e. } (kM, \lambda) \in K/M \times \mathfrak{a}^*.$$

We define the set of functions

$$\mathcal{E} = \{F \in L^2(\Xi) : \Phi_o F \in L^2_{o,\mathfrak{c}}(K/M \times \mathfrak{a}^*)\}$$

and we consider the operator $\Lambda \colon \mathcal{E} \subseteq L^2(\Xi) \to L^2(\Xi)$ given by

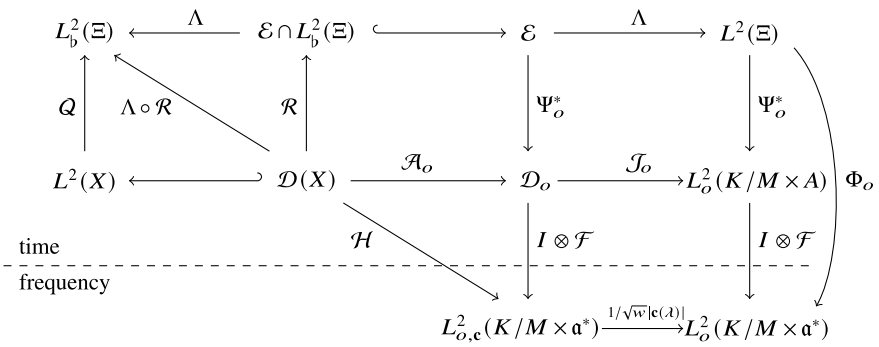


Fig. 4 Spaces and operators that come into play in our construction

$$\Lambda F = \Psi_o^{*-1} \mathcal{J}_o \Psi_o^* F.$$

As a direct consequence of the definition of Λ and $\mathcal{J}_o$, for every $F \in \mathcal{E}$ and for almost every $(kM, \lambda) \in K/M \times \mathfrak{a}^*$ we have (see the rightmost block in Fig. 4)

$$\begin{aligned} \Phi_o(\Lambda F)(kM, \lambda) &= (I \otimes \mathcal{F})(\mathcal{J}_o \Psi_o^* F)(kM, \lambda) \\ &= \frac{1}{\sqrt{w}\,|\mathbf{c}(\lambda)|}(I \otimes \mathcal{F})(\Psi_o^* F)(kM, \lambda) \\ &= \frac{1}{\sqrt{w}\,|\mathbf{c}(\lambda)|}\Phi_o F(kM, \lambda). \end{aligned} \tag{42}$$

The operator Λ intertwines the regular representation $\hat{\pi}$ next.

Proposition 37 *The subspace $\mathcal{E}$ is $\hat{\pi}$-invariant and for all $F \in \mathcal{E}$ and $g \in G$*

$$\hat{\pi}(g)\Lambda F = \Lambda \hat{\pi}(g) F. \tag{43}$$

Proof We consider $F \in \mathcal{E}$, $g \in G$ and we prove that $\hat{\pi}(g)F \in \mathcal{E}$. By (34)

$$\hat{\pi}(g)F \circ \Psi_o(kM, a) = F \circ \Psi_{g^{-1}[o]}(g^{-1}\langle kM\rangle, a)$$

for almost every $(kM, a) \in K/M \times A$. Therefore, we have

$$\Psi_o^*(\hat{\pi}(g)F)(kM, a) = \Psi_{g^{-1}[o]}^* F(g^{-1}\langle kM\rangle, a)$$

and consequently by Eq. (40)

$$\begin{aligned} \Phi_o(\hat{\pi}(g)F)(kM, \lambda) &= (I \otimes \mathcal{F})(\Psi_{g^{-1}[o]}^* F)(g^{-1}\langle kM\rangle, \lambda) \\ &= e^{(\rho - i\lambda)(A_{g^{-1}[o]}(o, g^{-1}\langle kM\rangle))}\Phi_o(F)(g^{-1}\langle kM\rangle, \lambda) \end{aligned} \tag{44}$$

for almost every $(kM, \lambda) \in K/M \times \mathfrak{a}^*$. By Eqs. (44), (33) and (32)

$$\begin{aligned} &\int_{K/M \times \mathfrak{a}^*} |\Phi_o(\hat{\pi}(g)F)(kM, \lambda)|^2 \frac{\mathrm{d}\nu^o(kM)\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2} \\ &= \int_{\mathfrak{a}^*}\int_{K/M} |\Phi_o(F)(g^{-1}\langle kM\rangle, \lambda)|^2 e^{2\rho(A_{g^{-1}[o]}(o, g^{-1}\langle kM\rangle))} \frac{\mathrm{d}\nu^o(kM)\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2} \\ &= \int_{K/M \times \mathfrak{a}^*} |\Phi_o F(kM, \lambda)|^2 e^{2\rho(A_{g^{-1}[o]}(o, kM))} \frac{\mathrm{d}\nu^{g^{-1}[o]}(kM)\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2} \\ &= \int_{K/M \times \mathfrak{a}^*} |\Phi_o F(kM, \lambda)|^2 \frac{\mathrm{d}\nu^o(kM)\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2} < +\infty \end{aligned}$$

and we conclude that $\hat{\pi}(g)F \in \mathcal{E}$. We finally prove the intertwining property (43). We have already observed that, by Proposition 35, it is enough to prove that

$$\Phi_o(\hat{\pi}(g)\Lambda F) = \Phi_o(\Lambda\hat{\pi}(g)F)$$

for every $g \in G$ and $F \in \mathcal{E}$. By Eqs. (44) and (42), for almost every $(kM, \lambda) \in K/M \times \mathfrak{a}^*$, we have the chain of equalities

$$\begin{aligned}\Phi_o(\hat{\pi}(g)\Lambda F)(kM,\lambda) &= e^{(\rho-i\lambda)(A_{g^{-1}[o]}(o,g^{-1}\langle kM\rangle))}\Phi_o(\Lambda F)(g^{-1}\langle kM\rangle,\lambda)\\ &= \frac{1}{\sqrt{w}\,|\mathbf{c}(\lambda)|}e^{(\rho-i\lambda)(A_{g^{-1}[o]}(o,g^{-1}\langle kM\rangle))}\Phi_o(F)(g^{-1}\langle kM\rangle,\lambda)\\ &= \frac{1}{\sqrt{w}\,|\mathbf{c}(\lambda)|}\Phi_o(\hat{\pi}(g)F)(kM,\lambda) = \Phi_o(\Lambda\hat{\pi}(g)F)(kM,\lambda),\end{aligned}$$

which proves the intertwining relation. □

The next result follows directly by Proposition 35 and Eq. (42).

Corollary 38 *For every $F \in \mathcal{E}$, $\Lambda F \in L^2_\natural(\Xi)$ if and only if $F \in L^2_\natural(\Xi)$.*

Proof By Proposition 35, $\Lambda F \in L^2_\natural(\Xi)$ if and only if $\Phi_o(\Lambda F)$ satisfies Property ♮. By (42) and since $\lambda \mapsto |\mathbf{c}(\lambda)|$ is W-invariant, $\Phi_o(\Lambda F)$ satisfies Property ♮ if and only if $\Phi_o(F)$ satisfies Property ♮, which is equivalent to $F \in L^2_\natural(\Xi)$. This concludes the proof. □

We are now in a position to prove our main result.

Theorem 39 *The composite operator $\Lambda\mathcal{R}$ extends to a unitary operator*

$$Q\colon L^2(X) \longrightarrow L^2_\natural(\Xi)$$

which intertwines the representations π and $\hat{\pi}$, i.e.,

$$\hat{\pi}(g)Q = Q\pi(g), \qquad g \in G. \tag{45}$$

Theorem 39 implies that π and the restriction $\hat{\pi}|_{L^2_\natural(\Xi)}$ of $\hat{\pi}$ to $L^2_\natural(\Xi)$ are unitarily equivalent representations. Moreover, $\hat{\pi}|_{L^2_\natural(\Xi)}$ (and then $\hat{\pi}$) is not irreducible, too.

Proof We first show that $\Lambda\mathcal{R}$ extends to a unitary operator Q from $L^2(X)$ onto $L^2(\Xi)$. It might be useful to keep in mind the leftmost block in Fig. 4. Let $f \in \mathcal{D}(X)$, by the Fourier Slice Theorem (39), the Plancherel formula and the definition of $\mathcal{J}_o$ and Λ, we have that

$$\begin{aligned}
\|f\|^2_{L^2(X)} &= \|\mathcal{H}f\|^2_{L^2_{o,\mathfrak{c}}(K/M\times\mathfrak{a}^*)^\sharp}\\
&= \|(I\otimes\mathcal{F})(\Psi_o^*(\mathcal{R}f))\|^2_{L^2_{o,\mathfrak{c}}(K/M\times\mathfrak{a}^*)^\sharp}\\
&= \int_{K/M\times\mathfrak{a}^*} |(I\otimes\mathcal{F})(\mathcal{J}_o\Psi_o^*(\mathcal{R}f))(kM,\lambda)|^2 \mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\
&= \int_{K/M\times\mathfrak{a}^*} |(I\otimes\mathcal{F})(\Psi_o^*(\Lambda\mathcal{R}f))(kM,\lambda)|^2 \mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\
&= \int_{K/M\times A} |\Psi_o^*(\Lambda\mathcal{R}f)(kM,a)|^2 \mathrm{d}\nu^o(kM)\mathrm{d}a\\
&= \|\Lambda\mathcal{R}f\|^2_{L^2(\Xi)}.
\end{aligned}$$

Hence, $\Lambda\mathcal{R}$ is an isometric operator from $\mathcal{D}(X)$ into $L^2(\Xi)$. Since $\mathcal{D}(X)$ is dense in $L^2(X)$, $\Lambda\mathcal{R}$ extends to a unique isometry from $L^2(X)$ onto the closure of $\mathrm{Ran}(\Lambda\mathcal{R})$ in $L^2(\Xi)$. We must show that $\Lambda\mathcal{R}$ has dense image in $L^2_\flat(\Xi)$. The inclusion $\mathrm{Ran}(\Lambda\mathcal{R})\subseteq L^2_\flat(\Xi)$ follows immediately from Corollary 36 and Corollary 38. Let $F\in L^2_\flat(\Xi)$ be such that $\langle F,\Lambda\mathcal{R}f\rangle_{L^2(\Xi)}=0$ for every $\mathcal{D}(X)$. By the Plancherel formula and the Fourier Slice Theorem (39) we have that

$$\begin{aligned}
0 &= \langle F,\Lambda\mathcal{R}f\rangle_{L^2(\Xi)}\\
&= \int_{K/M\times A} (F\circ\Psi_o)(kM,a)\overline{(\Lambda\mathcal{R}f\circ\Psi_o)(kM,a)}e^{2\rho(\log a)}\mathrm{d}\nu^o(kM)\mathrm{d}a\\
&= \int_{K/M\times A} (\Psi_o^*F)(kM,a)\overline{(\mathcal{J}_o\Psi_o^*(\mathcal{R}f))(kM,a)}\mathrm{d}\nu^o(kM)\mathrm{d}a\\
&= \int_{K/M\times\mathfrak{a}^*} \Phi_o(F)(kM,\lambda)\overline{(I\otimes\mathcal{F})(\mathcal{J}_o\Psi_o^*(\mathcal{R}f))(kM,\lambda)}\mathrm{d}\nu^o(kM)\mathrm{d}\lambda\\
&= \int_{K/M\times\mathfrak{a}^*} \Phi_o(F)(kM,\lambda)\overline{(I\otimes\mathcal{F})(\Psi_o^*(\mathcal{R}f))(kM,\lambda)}\frac{\mathrm{d}\nu^o(kM)\mathrm{d}\lambda}{\sqrt{w}|\mathbf{c}(\lambda)|}\\
&= \int_{K/M\times\mathfrak{a}^*} \sqrt{w}|\mathbf{c}(\lambda)|\Phi_o(F)(kM,\lambda)\overline{\mathcal{H}f(kM,\lambda)}\frac{\mathrm{d}\nu^o(kM)\mathrm{d}\lambda}{w|\mathbf{c}(\lambda)|^2}.
\end{aligned}$$

For simplicity, we denote by ΘF the function on $K/M\times\mathfrak{a}^*$ defined as

$$\Theta F(kM,\lambda)=\sqrt{w}|\mathbf{c}(\lambda)|\Phi_o(F)(kM,\lambda),\qquad \text{a.e. } (kM,\lambda)\in K/M\times\mathfrak{a}^*.$$

Hence we have proved that $\langle\Theta F,\mathcal{H}f\rangle=0$ for every $f\in\mathcal{D}(X)$. The next two facts follow immediately by Proposition 35. Since Φ_o is an isometry from $L^2(\Xi)$ into $L^2_o(K/M\times\mathfrak{a}^*)$, the function ΘF belongs to $L^2_{o,\mathfrak{c}}(K/M\times\mathfrak{a}^*)$. Further, since $F\in L^2_\flat(\Xi)$ and since $\lambda\mapsto|\mathbf{c}(\lambda)|$ is W-invariant, then $\Theta F\in L^2_{o,\mathfrak{c}}(K/M\times\mathfrak{a}^*)^\sharp$. By Theorem 26, $\mathcal{H}(\mathcal{D}(X))$ is dense in $L^2_{o,\mathfrak{c}}(K/M\times\mathfrak{a}^*)^\sharp$. Hence, $\Theta F=0$ in $L^2_{o,\mathfrak{c}}(K/M\times\mathfrak{a}^*)^\sharp$ and then $\Phi_o(F)=0$ in $L^2_o(K/M\times\mathfrak{a}^*)$. Since Φ_o is an isometry from $L^2(\Xi)$ into $L^2_o(K/M\times\mathfrak{a}^*)$, then $F=0$ in $L^2(\Xi)$. Therefore, $\overline{\mathrm{Ran}(\Lambda\mathcal{R})}=$

$L^2_\flat(\Xi)$ and $\Lambda\mathcal{R}$ extends uniquely to a surjective isometry

$$Q\colon L^2(X) \longrightarrow L^2_\flat(\Xi).$$

Observe that $Qf = \Lambda\mathcal{R}f$ for every $f \in \mathcal{D}(X)$. The intertwining property (45) follows immediately from Propositions 34 and 37. □

References

1. Alberti, G.S., Bartolucci, F., De Mari, F., De Vito, E.: Unitarization and inversion formulae for the Radon transform between dual pairs. SIAM J. Math. Anal. **51**(6), 4356–4381 (2019)
2. Alberti, G.S., Bartolucci, F., De Mari, F., De Vito, E.: Radon Transform: Dual Pairs and Irreducible Representations, pp. 1–28. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-56005-8_1
3. Antoine, J.P., Murenzi, R.: Two-dimensional directional wavelets and the scale-angle representation. Signal Process. **52**(3), 259–281 (1996)
4. Bartolucci, F., De Mari, F., De Vito, E., Odone, F.: The Radon transform intertwines wavelets and shearlets. Appl. Comput. Harmon. Anal. **47**(3), 822–847 (2019)
5. Bartolucci, F., De Mari, F., Monti, M.: Unitarization of the Radon transform on homogeneous trees (2020). arXiv:2002.06696. Submitted
6. Dahlke, S., Steidl, G., Teschke, G.: The continuous shearlet transform in arbitrary space dimensions. J. Fourier Anal. Appl. **16**(3), 340–364 (2010)
7. Duflo, M., Moore, C.C.: On the regular representation of a nonunimodular locally compact group. J. Funct. Anal. **21**(2), 209–243 (1976)
8. Folland, G.B.: A Course in Abstract Harmonic Analysis. Textbooks in Mathematics, 2nd edn. CRC Press, Boca Raton (2016)
9. Führ, H.: Continuous wavelet transforms with abelian dilation groups. J. Math. Phys. **39**(8), 3974–3986 (1998)
10. Führ, H., Tousi, R.R.: Simplified vanishing moment criteria for wavelets over general dilation groups, with applications to abelian and shearlet dilation groups. Appl. Comput. Harmon. Anal. **43**(3), 449–481 (2017). https://doi.org/10.1016/j.acha.2016.03.003
11. Furstenberg, H.: A Poisson formula for semi-simple Lie groups. Ann. Math., pp. 335–386 (1963)
12. Gel'fand, I.: Integral geometry and its relation to the theory of group representations. Russ. Math. Surv. **15**(2), 143–151 (1960). https://doi.org/10.1070/rm1960v015n02abeh004218
13. Harish-Chandra: Spherical functions on a semisimple Lie group, I. Am. J. Math. **80**(2), 241–310 (1958). http://www.jstor.org/stable/2372786
14. Harish-Chandra: Spherical functions on a semisimple Lie group II. Am. J. Math. **80**(3), 553–613 (1958). http://www.jstor.org/stable/2372772
15. Helgason, S.: Differential Geometry, Lie Groups, and Symmetric Spaces, vol. 80. Academic, Cambridge (1979)
16. Helgason, S.: Groups & Geometric Analysis: Radon Transforms, Invariant Differential Operators and Spherical Functions, vol. 1. Academic, Cambridge (1984)
17. Helgason, S.: Geometric Analysis on Symmetric Spaces (1994)
18. Helgason, S.: Harish-Chandra's c-function. A mathematical jewel, pp. 55–67. Springer, Berlin (1994)
19. Helgason, S.: The Radon Transform. Progress in Mathematics, vol. 5, 2nd edn. Birkhäuser Boston, Inc., Boston (1999)

20. Holschneider, M.: Wavelets: An Analysis Tool. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York (1995)
21. Iozzi, A.: Symmetric spaces, course notes, ETH (2012)
22. Knapp, A.W.: Representation Theory of Semisimple Groups: An Overview Based on Examples, vol. 36. Princeton University Press, Princeton (2001)
23. Labate, D., Lim, W.Q., Kutyniok, G., Weiss, G.: Sparse multidimensional representation using shearlets. In: Optics & Photonics 2005, pp. 59, 140U–59, 140U. International Society for Optics and Photonics (2005)
24. Rouvière, F.: Geodesic Radon transforms on symmetric spaces (2004)
25. Sardar, P.: Geometry of the symmetric space SL(n,$\mathbb{R}$)/SO(n,$\mathbb{R}$) (2017). https://www.youtube.com/watch?v=SnfYvKJIxrg&t=2341s
26. Varadarajan, V.S.: Geometry of Quantum Theory, 2nd edn. Springer, New York (1985)
27. Warner, F.W.: Foundations of Differentiable Manifolds and Lie Groups. Graduate Texts in Mathematics, vol. 94. Springer, New York (1983)
28. Wolf, J.A.: Harmonic Analysis on Commutative Spaces. Mathematical Surveys and Monographs, vol. 142. American Mathematical Society, Providence (2007). https://doi.org/10.1090/surv/142

Entropy and Concentration

Andreas Maurer

1 Introduction

Concentration inequalities bound the probabilities that random quantities deviate from their average, median, or otherwise typical values. If this deviation is small with high probability, then a repeated experiment or observation will likely produce a similar result. In this way concentration inequalities can give quantitative guarantees of reproducibility, a concept at the heart of empirical science [25].

In this chapter we limit ourselves to study quantities whose randomness arises through the dependence on many independent random variables. Suppose that (Ω_i, Σ_i) are measurable spaces for $i \in \{1, ..., n\}$ and that f is real valued function defined on the product space $\Omega = \prod_{i=1}^{n} \Omega_i$,

$$f : \mathbf{x} = (x_1, ..., x_n) \in \Omega \mapsto f(\mathbf{x}) \in \mathbb{R}.$$

Now let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent random variables, where X_i is distributed as μ_i in Ω_i. For $t > 0$ and $\mathbf{X}'$ iid to $\mathbf{X}$ we then want to give bounds on the upwards deviation probability

$$\Pr_{\mathbf{X}} \left\{ f(\mathbf{X}) - E\left[f\left(\mathbf{X}'\right) \right] > t \right\}$$

in terms of the deviation t, the measures μ_i and properties of the function f. Downward deviation bounds are then obtained by replacing f with $-f$. Usually we will just write $\Pr\{f - Ef > t\}$ for the deviation probability above.

The first bounds of this type were given by Chebychev and Bienaimé [11] in the late 19th century for additive functions of the form

A. Maurer (✉)
Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy
e-mail: am@andreas-maurer.eu

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8_2

$$f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i). \tag{1}$$

The subject has since been developed by Bernstein, Chernoff, Bennett, Hoeffding, and many others [4, 16], and results were extended from sums to more general and complicated nonlinear functions. During the past decades research activity has been stimulated by the contributions of Michel Talagrand [27, 28] and by the relevance of concentration phenomena to the rapidly growing field of computer science. Some concentration inequalities, like the well known bounded difference inequality, have become standard tools in the analysis of algorithms [23].

One of the more recent methods to derive concentration inequalities, the so-called *entropy method*, is rooted in the early investigations of Boltzmann [5] and Gibbs [12] into the foundations of statistical mechanics. While the modern entropy method evolved along a complicated historical path via quantum field theory and the logarithmic Sobolev-inequality of Leonard Gross [14], its hidden simplicity was understood and emphasized by Michel Ledoux, who also recognized the key role which the subadditivity of entropy can play in the derivation of concentration inequalities [18]. The method has been refined by Bobkov, Massart [20], Bousquet [9], and Boucheron et al. [7]. Recently Boucheron et al. [8] showed that the entropy method is sufficiently strong to derive a form of Talagrand's convex distance inequality.

In this chapter we present a variation of the entropy method in a compact and simplified form, closely tied to its origins in statistical mechanics. We give an exposition of the method in Sect. 2 and compress it into a toolbox to derive concentration inequalities.

In Sect. 3 we will then use this method to prove two classical concentration inequalities, the bounded difference inequality and a generalization of Bennett's inequality. As example applications we treat vector-valued concentration and generalization in empirical risk minimization, a standard problem in machine learning theory.

In Sect. 4 we address more difficult problems. The bounded difference inequality is used to prove the famous Gaussian concentration inequality. We also give some more recent inequalities which we apply to analyze the concentration of convex Lipschitz functions on $[0, 1]^n$, or of the spectral norm of a random matrix.

In Sect. 5 we describe some of the more advanced techniques, self-boundedness, and decoupling. As examples we give sub-Gaussian lower tail bounds for convex Lipschitz functions and a version of the Hanson-Wright inequality for bounded random variables and we derive an exponential inequality for the suprema of empirical processes. We conclude with another version of Bernstein's inequality and its application to U-statistics.

We limit ourselves to exponential deviation bounds from the mean. For moment bounds and other advanced methods to establish concentration inequalities, such as the transportation method or an in-depth treatment of logarithmic Sobolev inequalities, we recommend the monographs by Ledoux [18] and Boucheron, Lugosi, and Massart [6], and the overview article by McDiarmid [23]. Another important recent

development not covered is the method of exchangeable pairs proposed by Chatterjee [10] .

We fix some conventions and notation:

If (Ω, Σ) is any measurable space $\mathcal{A}(\Omega)$ will denote the algebra of bounded, measurable real valued functions on Ω. When there is no ambiguity we often just write $\mathcal{A}$ for $\mathcal{A}(\Omega)$. Although we give some results for unbounded functions, most functions for which we will prove concentration inequalities are assumed to be measurable and bounded, that is $f \in \mathcal{A}$. This assumption simplifies the statement of our results, because it guarantees the existence of algebraic and exponential moments and makes our arguments more transparent.

If (Ω, Σ, μ) is a probability space we write $\Pr F = \mu(F)$ for $F \in \Sigma$, and $E[f] = \int_\Omega f d\mu$ for $f \in L_1[\mu]$ and $\sigma^2[f] = E\left[(f - E[f])^2\right]$ for $f \in L_2[\mu]$. Wherever we use $\Pr$, E or σ^2, we assume that there is an underlying probability space (Ω, Σ, μ). If we refer to other measures than μ, then we identify them with corresponding subscripts.

If $\mathcal{X}$ is any set and $n \in \mathbb{N}$, then for $y \in \mathcal{X}$ and $k \in \{1, ..., n\}$ the substitution operator $S_y^k : \mathcal{X}^n \to \mathcal{X}^n$ is defined as

$$S_y^k x = (x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n) \text{ for } x = (x_1, ..., x_n) \in \mathcal{X}^n.$$

This and other notation which we introduce along the way is also summarized in a final section in tabular form.

2 The Entropy Method

In this section we develop the entropy method and package it into a toolbox to prove concentration inequalities.

2.1 Markov's Inequality and Exponential Moment Method

The most important tool in the proof of deviation bounds is Markov's inequality, which we now introduce along with two corollaries, Chebychev's inequality and the exponential moment method.

Theorem 1 (Markov inequality) *Let* $f \in L_1[\mu]$, $f \geq 0$ *and* $t > 0$. *Then*

$$\Pr\{f > t\} \leq \frac{E[f]}{t}$$

Proof Since $f \geq 0$ and $t > 0$ we have $1_{f>t} \leq f/t$ and therefore

$$\Pr\{f > t\} = E\left[1_{f>t}\right] \leq E\left[f/t\right] = \frac{E\left[f\right]}{t}.$$

□

Corollary 2 (Chebychev inequality) *Let* $f \in L_2\left[\mu\right]$ *and* $t > 0$. *Then*

$$\Pr\{|f - E\left[f\right]| > t\} = \Pr\left\{(f - E\left[f\right])^2 > t^2\right\} \leq \frac{E\left[(f - E\left[f\right])^2\right]}{t^2} = \frac{\sigma^2\left(f\right)}{t^2}.$$

To use Chebychev's inequality we need to bound the variance $\sigma^2\left(f\right)$. If f is a sum of independent variables, the variance of f is just the sum of the variances of the individual variables, but this doesn't work for general functions. In Sect. 3.1, however, we give the Efron–Stein inequality, which asserts that for functions of independent variables the variance is bounded by the expected sum of conditional variances.

The idea of Chebychev's inequality obviously extends to other even centered moments $E\left[(f - E\left[f\right])^{2p}\right]$. Bounding higher moments of functions of independent variables is an important technique discussed, for example, in [6].

Here the most important corollary of Markov's inequality is the *exponential moment method*, an idea apparently due to Bernstein [4].

Corollary 3 (exponential moment method) *Let* $f \in \mathcal{A}$, $\beta \geq 0$ *and* $t > 0$. *Then*

$$\Pr\{f > t\} = \Pr\left\{e^{\beta f} > e^{\beta t}\right\} \leq e^{-\beta t} E\left[e^{\beta f}\right].$$

To use this we need to bound the quantity $E\left[e^{\beta f}\right]$ and to optimize the right-hand side above over β. We call $E\left[e^{\beta f}\right]$ the *partition function*, denoted $Z_{\beta f} = E\left[e^{\beta f}\right]$. Bounding the partition function (or its logarithm) is the principal problem in the derivation of exponential tail bounds.

If f is a sum of independent components (as in (1)), then the partition function is the product of the partition functions corresponding to these components, and its logarithm (called the *moment generating function*) is additive. This is a convenient basis to obtain deviation bounds for sums, but it does not immediately extend to general non-additive functions. The approach is taken here, the entropy method, balances simplicity, and generality.

2.2 Entropy and Concentration

For the remainder of this section we take the function $f \in \mathcal{A}$ as fixed. We could interpret the points $x \in \Omega$ as possible states of a physical system and f as the negative energy (or Hamiltonian) function, so that $-f\left(x\right)$ is the system's energy in the state x. The measure μ then models an a priori probability distribution of states in the absence of any constraining information. We will define another probability measure on Ω, with specified expected energy, but with otherwise minimal assumptions.

If ρ is a function on Ω, $\rho \geq 0$ and $E[\rho] = 1$, the Kullback–Leibler divergence $KL(\rho d\mu, d\mu)$ of $\rho d\mu$ to $d\mu$ is

$$KL(\rho d\mu, d\mu) = E[\rho \ln \rho].$$

$KL(\rho d\mu, d\mu)$ can be interpreted as the information we gain, if we are told that the probability measure is $\rho d\mu$ instead of the a priori measure $d\mu$.

Theorem 4 *For all $f \in \mathcal{A}$, $\beta \in \mathbb{R}$*

$$\sup_{\rho} \beta E[\rho f] - E[\rho \ln \rho] = \ln E\left[e^{\beta f}\right],$$

where the supremum is over all nonnegative measurable functions ρ on Ω satisfying $E[\rho] = 1$.

The supremum is attained for the density

$$\rho_{\beta f} = e^{\beta f} / E\left[e^{\beta f}\right].$$

Proof We can assume $\beta = 1$ by absorbing it in f. Let $\rho \geq 0$ satisfy $E[\rho] = 1$, so that $\rho d\mu$ is a probability measure and $g \in \mathcal{A} \mapsto E_\rho[g] := E[\rho g]$ an expectation functional. Let $\phi(x) = 1/\rho(x)$ if $\rho(x) > 0$ and $\phi(x) = 0$ if $\rho(x) = 0$. Then $E[\rho \ln \rho] = -E[\rho \ln \phi] = -E_\rho[\ln \phi]$ and with Jensen's inequality

$$\begin{aligned} E[\rho f] - E[\rho \ln \rho] &= E_\rho[f + \ln \phi] = \ln \exp\left(E_\rho[f + \ln \phi]\right) \\ &\leq \ln E_\rho\left[\exp(f + \ln \phi)\right] = \ln E_\rho\left[\phi e^f\right] \\ &= \ln E\left[\rho \phi e^f\right] = \ln E\left[1_{\rho>0} e^f\right] \\ &\leq \ln E\left[e^f\right]. \end{aligned}$$

On the other hand

$$E\left[\rho_f f\right] - E\left[\rho_f \ln \rho_f\right] = \frac{E\left[f e^f\right]}{E\left[e^f\right]} - \frac{E\left[e^f \ln\left(e^f / E\left[e^f\right]\right)\right]}{E\left[e^f\right]} = \ln E\left[e^f\right].$$

□

In statistical physics the maximizing probability measure $d\mu_{\beta f} = \rho_{\beta f} d\mu = e^{\beta f} d\mu / E\left[e^{\beta f}\right]$ is called the *thermal measure*, sometimes also the *canonical ensemble*. It is used to describe a system in thermal equilibrium with a heat reservoir at temperature $T \approx 1/\beta$. The corresponding expectation functional

$$E_{\beta f}[g] = \frac{E\left[g e^{\beta f}\right]}{E\left[e^{\beta f}\right]} = Z_{\beta f}^{-1} E\left[g e^{\beta f}\right], \text{ for } g \in \mathcal{A}$$

is called the *thermal expectation*. The normalizing quantity $Z_{\beta f} = E\left[e^{\beta f}\right]$ is the *partition function* already introduced above. Notice that for any constant c we have $E_{\beta(f+c)}[g] = E_{\beta f}[g]$.

The value of the function $\rho \mapsto E[\rho \ln \rho]$ at the thermal density $\rho_{\beta f} = Z_{\beta f}^{-1} e^{\beta f}$ is called the *canonical entropy* or simply entropy,

$$\mathrm{Ent}_f(\beta) = E\left[\rho_{\beta f} \ln \rho_{\beta f}\right] = \beta E_{\beta f}[f] - \ln Z_{\beta f}. \tag{2}$$

Note that $\mathrm{Ent}_{-f}(\beta) = \mathrm{Ent}_f(-\beta)$, a simple but very useful fact.

Suppose that ρ is any probability density on Ω, which gives the same expected value for the energy as $\rho_{\beta f}$, so that $E[\rho f] = E_{\beta f}[f]$. Then

$$\begin{aligned} 0 &\leq KL\left(\rho d\mu, Z_{\beta f}^{-1} e^{\beta f} d\mu\right) \\ &= E[\rho \ln \rho] - \beta E[\rho f] + \ln Z_{\beta f} \\ &= E[\rho \ln \rho] - \beta E_{\beta f}[f] + \ln Z_{\beta f} \\ &= KL(\rho d\mu, d\mu) - KL\left(\rho_{\beta f} d\mu, d\mu\right). \end{aligned}$$

The thermal measure $d\mu_{\beta f} = \rho_{\beta f} d\mu$ therefore minimizes the information gain relative to the a priori measure $d\mu$, given the expected value $-E_{\beta f}[f]$ of the internal energy.

For $g \in \mathcal{A}$ and $\rho = Z_{\beta f}^{-1} e^{\beta f}$ Theorem 4 gives

$$E_{\beta f}[g] \leq \mathrm{Ent}_f(\beta) + \ln E\left[e^g\right],$$

which allows to decouple g from f. This plays an important role later on in this chapter.

For $\beta \neq 0$ define a function

$$A_f(\beta) = \frac{1}{\beta} \ln Z_{\beta f} = \frac{1}{\beta} \ln E\left[e^{\beta f}\right]. \tag{3}$$

By l'Hospital's rule we have $\lim_{\beta \to 0} A_f(\beta) = E[f]$, so A_f extends continuously to $\mathbb{R}$ by setting $A_f(0) = E[f]$. In statistical physics the quantity $A_f(\beta)$ so defined is called the *free energy* corresponding to the Hamiltonian (energy function) $H = -f$ and temperature $T \approx \beta^{-1}$. Theorem 4 exhibits the free energy and the canonical entropy as a pair of convex conjugates. Dividing (2) by β and writing $U = E_{\beta f}[f]$, we recover the classical thermodynamic relation

$$A = U - T\,\mathrm{Ent},$$

which describes the macroscopically available energy A as the difference between the internal energy U and an energy portion T Ent, which is inaccessible due to ignorance of the microscopic state.

The following theorem establishes the connection of entropy, the exponential moment method and concentration inequalities.

Theorem 5 *For $f \in \mathcal{A}$ and any $\beta \geq 0$ we have*

$$\ln E\left[e^{\beta(f-Ef)}\right] = \beta \int_0^\beta \frac{Ent_f(\gamma)}{\gamma^2} d\gamma$$

and, for $t \geq 0$,

$$\Pr\{f - Ef > t\} \leq \inf_{\beta \geq 0} \exp\left(\beta \int_0^\beta \frac{Ent_f(\gamma)}{\gamma^2} d\gamma - \beta t\right).$$

Proof Differentiating the free energy with respect to β we find

$$A'_f(\beta) = \frac{1}{\beta} E_{\beta f}[f] - \frac{1}{\beta^2} \ln Z_{\beta f} = \beta^{-2} \mathrm{Ent}_f(\beta).$$

By the fundamental theorem of calculus

$$\begin{aligned} \ln E\left[e^{\beta(f-Ef)}\right] &= \ln Z_{\beta f} - \beta E[f] = \beta\left(A_f(\beta) - A_f(0)\right) \\ &= \beta \int_0^\beta A'_f(\gamma)\, d\gamma = \beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma, \end{aligned}$$

which is the first inequality. Then by Markov's inequality

$$\begin{aligned} \Pr\{f - Ef > t\} &\leq e^{-\beta t} E\left[e^{\beta(f-Ef)}\right] \\ &\leq \exp\left(\beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma - \beta t\right). \end{aligned}$$

□

Our strategy to establish concentration results will therefore be the search for appropriate bounds on the entropy.

2.3 Entropy and Energy Fluctuations

The *thermal variance* of a function $g \in \mathcal{A}$ is just the variance of g relative to the thermal expectation. It is denoted $\sigma^2_{\beta f}(g)$ and defined by

$$\sigma^2_{\beta f}(g) = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right] = E_{\beta f}\left[g^2\right] - \left(E_{\beta f}[g]\right)^2.$$

For constant c we have $\sigma^2_{\beta(f+c)}[g] = \sigma^2_{\beta f}[g]$.

The proof of the following lemma consists of straightforward calculations, an easy exercise to familiarize oneself with thermal measure, expectation and variance.

Lemma 6 *The following formulas hold for* $f \in \mathcal{A}$
1. $\frac{d}{d\beta}\left(\ln Z_{\beta f}\right) = E_{\beta f}[f]$.
2. *If* $h : \beta \mapsto h(\beta) \in \mathcal{A}$ *is differentiable and* $(d/d\beta)\, h(\beta) \in \mathcal{A}$ *then*

$$\frac{d}{d\beta} E_{\beta f}[h(\beta)] = E_{\beta f}[h(\beta)\, f] - E_{\beta f}[h(\beta)]\, E_{\beta f}[f] + E_{\beta f}\left[\frac{d}{d\beta} h(\beta)\right].$$

3. $\frac{d}{d\beta} E_{\beta f}\left[f^k\right] = E_{\beta f}\left[f^{k+1}\right] - E_{\beta f}\left[f^k\right] E_{\beta f}[f]$.
4. $\frac{d^2}{d\beta^2}\left(\ln Z_{\beta f}\right) = \frac{d}{d\beta} E_{\beta f}[f] = \sigma^2_{\beta f}[f]$.

Proof 1. is immediate and 2. a straightforward computation. 3. and 4. are immediate consequences of 1. and 2. □

Since the members of $\mathcal{A}$ are bounded it follows from 2. that for $f, g \in \mathcal{A}$ the functions $\beta \mapsto E_{\beta f}[g]$ and $\beta \mapsto \sigma^2_{\beta f}[g]$ are C_∞.

The thermal variance of f itself corresponds to energy fluctuations. The next theorem represents entropy as a double integral of such fluctuations. The utility of this representation to derive concentration results has been noted by David McAllester [22].

Theorem 7 *We have for* $\beta > 0$

$$Ent_f(\beta) = \int_0^\beta \int_t^\beta \sigma^2_{sf}[f]\, ds\, dt.$$

Proof Using the previous lemma and the fundamental theorem of calculus we obtain the formulas

$$\beta E_{\beta f}[f] = \int_0^\beta E_{\beta f}[f]\, dt = \int_0^\beta \left(\int_0^\beta \sigma^2_{sf}[f]\, ds + E[f]\right) dt$$

and

$$\ln Z_{\beta f} = \int_0^\beta E_{tf}[f]\, dt = \int_0^\beta \left(\int_0^t \sigma^2_{sf}[f]\, ds + E[f]\right) dt,$$

which we subtract to obtain

$$\begin{aligned}\mathrm{Ent}_f(\beta) &= \beta E_{\beta f}[f] - \ln Z_{\beta f} = \int_0^\beta \left(\int_0^\beta \sigma^2_{sf}[f]\, ds - \int_0^t \sigma^2_{sf}[f]\, ds\right) dt \\ &= \int_0^\beta \left(\int_t^\beta \sigma^2_{sf}[f]\, ds\right) dt.\end{aligned}$$

□

Since bounding $\sigma^2_{\beta f}[f]$ is central to our method, it is worth mentioning an interpretation in terms of heat capacity, or specific heat. Recall that $-E_{\beta f}[f]$ is the expected internal energy. The rate of change of this quantity with temperature T is the heat capacity. By conclusion 4 of Lemma 6 we have

$$\frac{d}{dT}\left(-E_{\beta f}[f]\right) = \frac{1}{T^2}\sigma^2_{\beta f}[f],$$

which exhibits the proportionality of heat capacity and energy fluctuations.

2.4 Product Spaces and Conditional Operations

We now set $\Omega = \prod_{k=1}^n \Omega_k$ and $d\mu = \prod_{k=1}^n d\mu_k$, where each μ_k is the probability measure representing the distribution of some variable X_k in the space Ω_k, so that the X_k are assumed to be independent.

With $\mathcal{A}_k$ we denote the subalgebra of those functions $f \in \mathcal{A}$, which are independent of the k-th argument. To efficiently deal with operations performed on individual arguments of functions in $\mathcal{A}$ we need some special notation.

Now let $k \in \{1, ..., n\}$ and $y \in \Omega_k$. If Ξ is any set and F is any function $F : \Omega \to \Xi$, we extend the definition of the *substitution operator* S^k_y to F by $S^k_y(F) = F \circ S^k_y$. This means

$$S^k_y(F)(x_1, ..., x_n) = F(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n),$$

so the k-th argument is simply replaced by y. Since for $f \in \mathcal{A}$ the function $S^k_y f$ is independent of x_k (which had been replaced by y) we see that S^k_y is a homomorphic (linear and multiplication-preserving) projection of $\mathcal{A}$ onto $\mathcal{A}_k$.

For $k \in \{1, ..., n\}$ and $y, y' \in \Omega_k$ we define the difference operator $D^k_{y,y'} : \mathcal{A} \to \mathcal{A}_k$ by

$$D^k_{y,y'} f = S^k_y f - S^k_{y'} f \text{ for } f \in \mathcal{A}.$$

Clearly $D^k_{y,y'}$ annihilates $\mathcal{A}_k$. The operator $r_k : \mathcal{A} \to \mathcal{A}_k$, defined by $r_k f = \sup_{y,y' \in \Omega_k} D^k_{y,y'} f$ is called the k-th *conditional range*. We also use the abbreviations $\inf_k f = \inf_{y \in \Omega_k} S^k_y f$ and $\sup_k f = \sup_{y \in \Omega_k} S^k_y f$ for the conditional infimum and supremum.

Given the measures μ_k and $k \in \{1, ..., n\}$ we the operator $E_k : \mathcal{A} \to \mathcal{A}_k$ by

$$E_k f = E_{y \sim \mu_k}\left[S^k_y f\right] = \int_{\Omega_k} S^k_y f \, d\mu_k(y).$$

The operator $E_k[.] = E\left[.|X_1, ..., X_{k-1}, X_{k+1}, ..., X_n\right]$ is the expectation conditional to all variables with indices different to k. E_k is a linear projection onto $\mathcal{A}_k$. Also the E_k commute among each other, and for $h \in \mathcal{A}$ and $g \in \mathcal{A}_k$ we have

$$E\left[\left[E_k h\right] g\right] = E\left[E_k\left[hg\right]\right] = E\left[hg\right]. \tag{4}$$

Replacing the operator E by E_k leads to the definition of conditional thermodynamic quantities, all of which are now members of the algebra $\mathcal{A}_k$:

- The conditional partition function $Z_{k,\beta f} = E_k\left[e^{\beta f}\right]$,
- The conditional thermal expectation $E_{k,\beta f}[g] = Z_{k,\beta f}^{-1} E_k\left[g e^{\beta f}\right]$ for $g \in \mathcal{A}$,
- The conditional entropy $\mathrm{Ent}_{k,f}(\beta) = \beta E_{k,\beta f}[f] - \ln Z_{k,\beta f}$,
- The conditional thermal variance $\sigma_{k,\beta f}^2[g] = E_{k,\beta f}\left[\left(g - E_{k,\beta f}[g]\right)^2\right]$ for $g \in \mathcal{A}$. As $\beta \to 0$ this becomes
- The conditional variance $\sigma_k^2[g] = E_k\left[(g - E_k[g])^2\right]$ for $g \in \mathcal{A}$.

The previously established relations hold also for the corresponding conditional quantities. Of particular importance for our method is the conditional version of Theorem 7

$$\mathrm{Ent}_{k,f}(\beta) = \int_0^\beta \int_t^\beta \sigma_{k,sf}^2[f]\, ds\, dt.$$

The following lemma, which states that the conditional thermal expectation just behaves like a conditional expectation, will also be used frequently.

Lemma 8 *For any $f, g \in \mathcal{A}$, $k \in \{1, ..., n\}$, $\beta \in \mathbb{R}$*

$$E_{\beta f}\left[E_{k,\beta f}[g]\right] = E_{\beta f}[g].$$

Proof Using $E\left[E_k[h]\, g\right] = E\left[h E_k[g]\right]$

$$\begin{aligned} E_{\beta f}\left[E_{k,\beta f}[g]\right] &= Z_{\beta f}^{-1} E\left[E_k\left[g e^{\beta f}\right] \frac{e^{\beta f}}{E_k\left[e^{\beta f}\right]}\right] \\ &= Z_{\beta f}^{-1} E\left[g e^{\beta f} E_k\left[\left(\frac{e^{\beta f}}{E_k\left[e^{\beta f}\right]}\right)\right]\right] \\ &= Z_{\beta f}^{-1} E\left[g e^{\beta f}\right] \\ &= E_{\beta f}[g]. \end{aligned}$$

□

2.5 The Subadditivity of Entropy

In the non-interacting case, when the energy function f is a sum, $f = \sum f_k$, it is easily verified that $\mathrm{Ent}_{k,f}(\beta)(\mathbf{x}) = \mathrm{Ent}_{k,f}(\beta)$ is independent of $\mathbf{x}$ and that

$$\mathrm{Ent}_f(\beta) = \sum_{k=1}^{n} \mathrm{Ent}_{k,f}(\beta).$$

In statistical physics it is said that entropy is an extensive quantity: the entropy of non-interacting systems is equal to the sum of the individual entropies.

Equality no longer holds in the interacting, nonlinear case, but there is a subadditivity property which is sufficient for the purpose of concentration inequalities:

The total entropy is no greater than the thermal average of the sum of the conditional entropies.

Theorem 9 *For $f \in \mathcal{A}$ and $\beta > 0$*

$$Ent_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n} Ent_{k,f}(\beta)\right] \tag{5}$$

In 1975 Elliott Lieb [19] gave a proof of this result, which was probably known some time before, at least in the classical setting relevant to our arguments. Together with Theorem 5 and Theorem 7 it completes our basic toolbox to prove concentration inequalities. For the proof we need two auxiliary results.

Lemma 10 *Let $h, g > 0$ be bounded measurable functions on Ω. Then for any expectation E*

$$E[h] \ln \frac{E[h]}{E[g]} \leq E\left[h \ln \frac{h}{g}\right].$$

Proof Define an expectation functional E_g by $E_g[h] = E[gh]/E[g]$. The function $\Phi(t) = t \ln t$ is convex for positive t, since $\Phi'' = 1/t > 0$. Then

$$\Phi\left(E_g\left[\frac{h}{g}\right]\right) = \frac{E[h]}{E[g]} \ln \frac{E[h]}{E[g]}.$$

Thus, by Jensen's inequality,

$$\begin{aligned} E[h] \ln \frac{E[h]}{E[g]} &= E[g]\, E_g\left[\frac{h}{g}\right] \ln E_g\left[\frac{h}{g}\right] = E[g]\, \Phi\left(E_g\left[\frac{h}{g}\right]\right) \\ &\leq E[g]\, E_g\left[\Phi\left(\frac{h}{g}\right)\right] = E\left[h \ln \frac{h}{g}\right]. \end{aligned}$$

□

Next we prove (5) for general positive functions.

Lemma 11 *Let $\rho \in \mathcal{A}$, $\rho > 0$. Then*

$$E\left[\rho \ln \frac{\rho}{E[\rho]}\right] \leq \sum_k E\left[\rho \ln \frac{\rho}{E_k[\rho]}\right].$$

Proof Write $E^k[.] = E_1 E_2 ... E_k[.]$ with E^0 being the identity map on $\mathcal{A}$. The innocuous looking identity $E\left[E^k[.]\right] = E[.]$ is an obvious consequence of the fact that we work with product probabilities. Without independence it would not hold, and the following simple argument would break down. Note that $E^n = E$. We expand

$$\frac{\rho}{E[\rho]} = \frac{E^0[\rho]}{E^1[\rho]} \frac{E^1[\rho]}{E_2[\rho]} \cdots \frac{E^{n-1}[\rho]}{E^n[\rho]} = \prod_{k=1}^n \frac{E^{k-1}[\rho]}{E^{k-1}[E_k[\rho]]}.$$

We get from Lemma 10, using $E\left[E^{k-1}[.]\right] = E[.]$,

$$\begin{aligned} E\left[\rho \ln \frac{\rho}{E[\rho]}\right] &= \sum_k E\left[E^{k-1}[\rho] \ln \frac{E^{k-1}[\rho]}{E^{k-1}[E_k[\rho]]}\right] \\ &\leq \sum_k E\left[E^{k-1}\left[\rho \ln \frac{\rho}{E_k[\rho]}\right]\right] = \sum_k E\left[\rho \ln \frac{\rho}{E_k[\rho]}\right]. \end{aligned}$$

□

Finally we specialize to the canonical entropy.

Proof of Theorem 9 9 Set $\rho = e^{\beta f}$ in Lemma 11 to get

$$\begin{aligned} \mathrm{Ent}_f(\beta) &= Z_{\beta f}^{-1} E\left[e^{\beta f} \ln \frac{e^{\beta f}}{E\left[e^{\beta f}\right]}\right] \\ &\leq Z_{\beta f}^{-1} \sum_k E\left[e^{\beta f} \ln \frac{e^{\beta f}}{E_k\left[e^{\beta f}\right]}\right] \\ &= \sum_k E_{\beta f}\left[\beta f - \ln E_k\left[e^{\beta f}\right]\right] \\ &= E_{\beta f}\left[\sum_k \mathrm{Ent}_{k,f}(\beta)\right], \end{aligned}$$

where we used Lemma 8 in the last identity. □

2.6 Summary of Results

The exponential moment method, Corollary 3, and Theorems 5, 7, and 9 provide us with the tools to prove several useful concentration inequalities. Here is a summary:

Theorem 12 *For $f \in A$ and $\beta > 0$ we have*

$$\Pr\{f - Ef > t\} \leq E\left[e^{\beta(f-Ef)}\right] e^{-\beta t} \tag{6}$$

$$\ln E\left[e^{\beta(f-Ef)}\right] = \beta \int_0^\beta \frac{Ent_f(\gamma)}{\gamma^2} d\gamma \tag{7}$$

$$Ent_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^n Ent_{k,f}(\beta)\right] \tag{8}$$

$$Ent_f(\beta) = \int_0^\beta \int_t^\beta \sigma_{sf}^2[f]\, ds\, dt \tag{9}$$

$$Ent_{k,f}(\beta) = \int_0^\beta \int_t^\beta \sigma_{k,sf}^2[f]\, ds\, dt \tag{10}$$

Concatenating the exponential moment bound (6), the entropy representation of the moment generating function (7), the subadditivity of entropy (8) and the fluctuation representation of the conditional entropy (10), we obtain the following generic concentration inequality.

$$\Pr\{f - Ef > t\} \leq \inf_{\beta>0} \exp\left(\beta \int_0^\beta \gamma^{-2} E_{\gamma f}\left[\sum_{k=1}^n \int_0^\gamma \int_t^\gamma \sigma_{k,sf}^2[f]\, ds\, dt\right] d\gamma - \beta t\right).$$

This is the template for the results given in the next section.

3 First Applications of the Entropy Method

We now develop some first consequences of the method, beginning with the Efron–Stein inequality, a general bound on the variance. Then we continue with the derivation of the bounded difference inequality, a simple and perhaps the most useful concentration inequality, for which we give two illustrating applications. Then we give a Bennett-Bernstein type inequality which we apply to the concentration of vector-valued random variables.

3.1 The Efron–Stein Inequality

Combining the fluctuation representations (9) and (10) with the subadditivity (8) of entropy and dividing by β^2 we obtain

$$\frac{1}{\beta^2}\int_0^\beta \int_t^\beta \sigma_{sf}^2[f]\,ds\,dt \leq E_{\beta f}\left[\sum_{k=1}^n \frac{1}{\beta^2}\int_0^\beta \int_t^\beta \sigma_{k,sf}^2[f]\,ds\,dt.\right]$$

Using the continuity properties of $\beta \mapsto E_{\beta f}[g]$ and $\beta \mapsto \sigma_{\beta f}^2[f]$, which follow from Lemma 6 we can take the limit as $\beta \to 0$ and multiply by 2 to obtain

$$\sigma^2[f] \leq E\left[\sum_k \sigma_k^2[f]\right] = E\left[\Sigma^2(f)\right], \tag{11}$$

where we introduced the notation $\Sigma^2(f) = \sum_k \sigma_k^2[f]$ for the sum of conditional variances.

Equation (11) is the famous Efron–Stein–Steele inequality [26]. It is an easy exercise to provide the details of the above limit process and to extend the inequality to general functions $f \in L_2[\mu]$ by approximation with a sequence of truncations.

3.2 The Bounded Difference Inequality

The variance of a bounded real random variable is never greater than a quarter of the square of its range.

Lemma 13 *If $f \in \mathcal{A}$ satisfies $a \leq f \leq b$ then $\sigma^2[f] \leq (b-a)^2/4$.*

Proof

$$\begin{aligned}\sigma^2(f) &= E[(f - E[f])\,f] = E[(f - E[f])\,(f-a)]\\ &\leq E[(b - E[f])\,(f-a)] = (b - E[f])\,(E[f]-a)\\ &\leq \frac{(b-a)^2}{4}.\end{aligned}$$

To see the last inequality use calculus to find the maximal value of the function $t \to (b-t)\,(t-a)$. □.

The bounded difference inequality bounds the deviation of a function from its mean in terms of the *sum of squared conditional ranges*, which is the operator $R^2 : \mathcal{A} \to \mathcal{A}$ defined by

$$R^2(f) = \sum_{k=1}^{n} r_k(f)^2 = \sum_{k=1}^{n} \sup_{y,y' \in \Omega_k} \left(D^k_{y,y'} f\right)^2 .$$

Theorem 14 (Bounded difference inequality) *For $f \in \mathcal{A}$ and $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{\sup_{\mathbf{x}\in\Omega} R^2(f)(\mathbf{x})}\right).$$

Proof Applied to the conditional thermal variance Lemma 13 gives

$$\sigma^2_{k,sf}[f] \leq \frac{1}{4} \sup_{y,y' \in \Omega_k} \left(D^k_{y,y'} f\right)^2 = \frac{1}{4} r_k(f)^2 ,$$

so combining the subadditivity of entropy (8) and the fluctuation representation (10) gives

$$\begin{aligned} \mathrm{Ent}_f(\gamma) &\leq E_{\gamma f}\left[\sum_{k=1}^{n} \mathrm{Ent}_{k,f}(\gamma)\right] = E_{\gamma f}\left[\sum_{k=1}^{n} \int_0^\gamma \int_t^\gamma \sigma^2_{k,sf}[f]\, ds\, dt\right] \\ &\leq \frac{1}{4} E_{\gamma f}\left[\int_0^\gamma \int_t^\gamma \sum_{k=1}^{n} r_k(f)^2\right] ds\, dt \\ &= \frac{\gamma^2}{8} E_{\gamma f}\left[R^2(f)\right]. \end{aligned} \tag{12}$$

Bounding the thermal expectation $E_{\gamma f}$ by the supremum over $\mathbf{x} \in \Omega$ we obtain from Theorem 12 (7)

$$\ln E\left[e^{\beta(f-Ef)}\right] = \beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma \leq \frac{\beta^2}{8} \sup_{\mathbf{x}\in\Omega} R^2(f)(\mathbf{x}) ,$$

and the tail bound (6) gives for all $\beta > 0$

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{\beta^2}{8} \sup_{\mathbf{x}\in\Omega} R^2(f)(\mathbf{x}) - \beta t\right).$$

Substitution of the minimizing value $\beta = 4t / \left(\sup_{\mathbf{x}\in\Omega} R^2(f)(\mathbf{x})\right)$ completes the proof. □

Notice that the conditional range $r_k(f)$ is a function in $\mathcal{A}_k$ and may depend on all x_i except x_k. The sum $R^2(f) = \sum_{k=1}^{n} r_k(f)^2$ may thus depend on all the x_i. It is therefore a very pleasant feature that the supremum over $\mathbf{x}$ is taken *outside the sum*. In the sequel this will allow us to derive the Gaussian concentration inequality from Theorem 14. The bound (12) will be re-used in Sect. 5.4 to prove a version of the Hanson-Wright inequality for quadratic forms.

In the literature one often sees the following weaker version of Theorem 14.

Corollary 15 *For $f \in \mathcal{A}$ and $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{\sum_{k=1}^{n} \sup_{\mathbf{x}\in\Omega} r_k(f)^2(\mathbf{x})}\right).$$

If f is a sum $f = \sum_k X_k$, then r_k^2 is independent of $\mathbf{x}$ and the two results are equivalent. In this case we obtain the well known Hoeffding inequality [16].

Corollary 16 (Hoeffding's inequality) *Let X_k be real random variables $a_k \leq X_k \leq b_k$. Then*

$$\Pr\left\{\sum_k (X_k - E[X_k]) > t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right).$$

In returning to the general case of non-additive functions, it is remarkable that for many applications the following "little bounded difference inequality", which is yet weaker than Corollary 15, seems to be sufficient.

Corollary 17 *For $f \in \mathcal{A}$ and $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{nc^2}\right),$$

where

$$c = \max_k \sup_{\mathbf{x}\in\Omega, y, y'\in\Omega_k} D^k_{y,y'} f(\mathbf{x}).$$

3.3 Vector-Valued Concentration

Suppose the X_i are independent random variables with values in a normed space $\mathcal{B}$ such that $EX_i = 0$ and $\|X_i\| \leq c_i$. Let $\Omega_i = \{y \in \mathcal{B} : \|y\| \leq c_i\}$ and define $f : \prod_{i=1}^{n} \Omega_i \to \mathbb{R}$ by

$$f(\mathbf{x}) = \left\|\sum_i x_i\right\|.$$

Then by the triangle inequality, for y, y' with $\|y\|, \|y'\| \leq c_k$

$$\begin{aligned} D^k_{y,y'} f(\mathbf{x}) &= \left\| \sum_i S^k_y(\mathbf{x})_i \right\| - \left\| \sum_i S^k_{y'}(\mathbf{x})_i \right\| \\ &\leq \left\| \sum_i S^k_y(x)_i - \sum_i S^k_{y'}(x)_i \right\| = \left\| y - y' \right\| \\ &\leq 2c_k, \end{aligned}$$

so $R^2(f)(\mathbf{x}) \leq 4\sum_i c_i^2$. It follows from Corollary 15 that

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sum_i c_i^2}\right),$$

or that for $\delta > 0$ with probability at least $1 - \delta$ in $(X_1, ..., X_n)$

$$\left\| \sum_i X_i \right\| \leq E\left\| \sum_i X_i \right\| + \sqrt{2\sum_i c_i^2 \ln(1/\delta)}. \tag{13}$$

If $\mathcal{B}$ is a Hilbert space we can bound $E\left\|\sum_i X_i\right\| \leq \sqrt{\sum_i E\left[\|X_i\|^2\right]}$ by Jensen's inequality and if all the X_i are iid we get with probability at least $1 - \delta$

$$\left\| \frac{1}{n}\sum_i X_i \right\| \leq \sqrt{\frac{E\left[\|X_1\|^2\right]}{n}} + c_1\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{14}$$

3.4 Rademacher Complexities and Generalization

Now let $\mathcal{X}$ be any measurable space and $\mathcal{F}$ a countable class of functions $f : \mathcal{X} \to [0, 1]$ and $\mathbf{X} = (X_1, ..., X_n)$ be a vector of iid random variables with values in $\mathcal{X}$.

Empirical risk minimization really wants to find $f \in \mathcal{F}$ with minimal risk $E[f(X)]$, but, as the true distribution of X is unknown, it has to be content with minimizing the empirical surrogate

$$\frac{1}{n}\sum_i f(X_i).$$

One way to justify this method is by giving a bound on the uniform estimation error

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\left| \sum_i f(X_i) - E[f(X)] \right|.$$

The vector space

$$\mathcal{B} = \left\{ g : \mathcal{F} \to \mathbb{R} : \sup_{f \in \mathcal{F}} |g(f)| < \infty \right\}$$

becomes a normed space with norm $\|g\| = \sup_{f \in \mathcal{F}} |g(f)|$. For each X_i define $\hat{X}_i \in \mathcal{B}$ by $\hat{X}_i(f) = f(X_i) - E[f(X_i)]$. Then the $\hat{X}_i$ are zero mean random variables in $\mathcal{B}$ satisfying $\left\| \hat{X}_i \right\| \leq 1$, and (13) of the preceding section gives with probability at least $1 - \delta$

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i f(X_i) - E[f(X_i)] \right| \leq \frac{1}{n} E \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - E[f(X_i)] \right| + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

The expectation term on the right-hand side can be bounded in terms of *Rademacher complexity* [3]. This is the function $\mathcal{R} : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$ defined as

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) = \frac{2}{n} E_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(x_i) \right|,$$

where the $\epsilon = (\epsilon_1, ..., \epsilon_n)$ are vectors of independent Rademacher variables which are uniformly distributed on $\{-1, 1\}$. We have, with X_i' iid to X_i

$$\begin{aligned} \frac{1}{n} E \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - E[f(X_i)] \right| &\leq \frac{1}{n} E_{\mathbf{X}\mathbf{X}'} \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - f\left(X_i'\right) \right| \\ &= \frac{1}{n} E_{\mathbf{X}\mathbf{X}'} \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i \left(f(X_i) - f\left(X_i'\right) \right) \right|, \end{aligned}$$

for any $\epsilon \in \{-1, 1\}^n$, because the expectation is invariant under the interchange of X_i and X_i' on an arbitrary subset of indices. Passing to the expectation in ϵ and using the triangle inequality gives

$$\begin{aligned} \frac{1}{n} E \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - E[f(X_i)] \right| &\leq \frac{1}{n} E_{\mathbf{X}\mathbf{X}'} E_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i \left(f(X_i) - f\left(X_i'\right) \right) \right| \\ &\leq \frac{2}{n} E_{\mathbf{X}} E_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(X_i) \right| \\ &= E_{\mathbf{X}} \mathcal{R}(\mathcal{F}, \mathbf{X}). \end{aligned}$$

Now we use the bounded difference inequality again to bound the deviation of $\mathcal{R}(\mathcal{F}, .)$ from its expectation. We have, again using the triangle inequality,

$$D^k_{y,y'}\mathcal{R}(\mathcal{F},\mathbf{x}) = \frac{2}{n}E_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_i \epsilon_i S^k_y f(x_i)\right| - \sup_{f\in\mathcal{F}}\left|\sum_i \epsilon_i S^k_{y'} f(x_i)\right|\right]$$
$$\leq \frac{2}{n}E_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\epsilon_i\left(f(y) - f\left(y'\right)\right)\right|\right] \leq \frac{2}{n}$$

and thus obtain

$$\Pr\left\{E\left[\mathcal{R}(\mathcal{F},.)\right] > \mathcal{R}(\mathcal{F},.) + t\right\} \leq e^{-nt^2/2},$$

or, for every $\delta > 0$ with probability at least $1-\delta$

$$E\left[\mathcal{R}(\mathcal{F},\mathbf{X})\right] \leq \mathcal{R}(\mathcal{F},\mathbf{X}) + \sqrt{\frac{2\ln(1/\delta)}{n}}. \tag{15}$$

By a union bound we conclude that with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_i f(X_i) - E[f(X_i)]\right| \leq \mathcal{R}(\mathcal{F},\mathbf{X}) + 2\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

3.5 The Bennett and Bernstein Inequalities

The proof of the bounded difference inequality relied on bounding the thermal variance $\sigma^2_{k,\beta f}(f)$ uniformly in β, using the constraints on the conditional ranges of f. We now consider the case, where we only use one constraint on the ranges, say $f - E_k[f] \leq 1$, but we use information on the conditional variances. This leads to a Bennett type inequality as in [23]. Recall the notation for the sum of conditional variances $\Sigma^2(f) := \sum \sigma^2_k(f)$. Again we start with a bound on the thermal variance.

Lemma 18 *Assume $f - Ef \leq 1$. Then for $\beta > 0$*

$$\sigma^2_{\beta f}(f) \leq e^\beta \sigma^2(f)$$

Proof

$$\sigma^2_{\beta f}(f) = \sigma^2_{\beta(f-Ef)}(f - Ef) = E_{\beta(f-Ef)}\left[(f-Ef)^2\right] - \left(E_{\beta(f-Ef)}[f-Ef]\right)^2$$
$$\leq E_{\beta(f-Ef)}\left[(f-Ef)^2\right] = \frac{E\left[(f-Ef)^2 e^{\beta(f-Ef)}\right]}{E\left[e^{\beta(f-Ef)}\right]}$$
$$\leq E\left[(f-Ef)^2 e^{\beta(f-Ef)}\right] \text{ (by Jensen's inequality)}$$
$$\leq e^\beta E\left[(f-Ef)^2\right] \text{ (now using } f - Ef \leq 1).$$

□

Next we bound the entropy $\mathrm{Ent}_f(\beta)$.

Lemma 19 *Assume that $f - E_k f \leq 1$ for all $k \in \{1, ..., n\}$. Then for $\beta > 0$*

$$Ent_f(\beta) \leq \left(\beta e^{\beta} - e^{\beta} + 1\right) E_{\beta f}\left[\Sigma^2(f)\right].$$

Proof From Theorem 12 and the previous lemma we get

$$\mathrm{Ent}_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n} \int_0^{\beta} \int_t^{\beta} \sigma_{k,sf}^2[f]\, ds\, dt\right] \leq \int_0^{\beta} \int_t^{\beta} e^s ds\, dt\, E_{\beta f}\left[\Sigma^2(f)\right].$$

The conclusion follows from the formula

$$\int_0^{\beta} \int_t^{\beta} e^s ds\, dt = \int_0^{\beta} \left(e^{\beta} - e^t\right) dt = \beta e^{\beta} - e^{\beta} + 1.$$

□

We need one more technical Lemma.

Lemma 20 *For $x \geq 0$*

$$(1+x)\ln(1+x) - x \geq 3x^2/(6+2x).$$

Proof We have to show that

$$f_1(x) := \left(6 + 8x + 2x^2\right)\ln(1+x) - 6x - 5x^2 \geq 0.$$

Since $f_1(0) = 0$ and $f_1'(x) = 4 f_2(x)$ with $f_2(x) := (2+x)\ln(1+x) - 2x$, it is enough to show that $f_2(x) \geq 0$. But $f_2(0) = 0$ and $f_2'(x) = (1+x)^{-1} + \ln(1+x) - 1$, so $f_2'(0) = 0$, but $f_2''(x) = x(1+x)^{-2} \geq 0$, so $f_2(x) \geq 0$. □

Now we can prove our version of Bennett's inequality.

Theorem 21 *Assume $f - E_k f \leq 1, \forall k$. Let $t > 0$ and denote $V = \sup_{\mathbf{x}\in\Omega} \Sigma^2(f)(\mathbf{x})$. Then*

$$\begin{aligned}\Pr\{f - E[f] > t\} &\leq \exp\left(-V\left(\left(1 + tV^{-1}\right)\ln\left(1 + tV^{-1}\right) - tV^{-1}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2V + 2t/3}\right).\end{aligned}$$

Proof Fix $\beta > 0$. We define the real function

$$\psi(t) = e^t - t - 1, \tag{16}$$

which arises from deleting the first two terms in the power series expansion of the exponential function and observe that

$$\int_0^\beta \frac{\gamma e^\gamma - e^\gamma + 1}{\gamma^2} d\gamma = \beta^{-1}\left(e^\beta - \beta - 1\right) = \beta^{-1}\psi(\beta),$$

because $(d/d\gamma)\left(\gamma^{-1}(e^\gamma - 1)\right) = \gamma^{-2}(\gamma e^\gamma - e^\gamma + 1)$ and $\lim_{\gamma\to 0}\gamma^{-1}(e^\gamma - 1) = 1$. Theorem 12 and Lemma 19 combined with a uniform bound then give

$$\begin{aligned}\ln E e^{\beta(f-Ef)} &= \beta\int_0^\beta \frac{\mathrm{Ent}_f(\gamma)\, d\gamma}{\gamma^2} \\ &\leq \beta\left(\int_0^\beta \frac{\gamma e^\gamma - e^\gamma + 1}{\gamma^2} d\gamma\right)\sup_{\mathbf{x}\in\Omega}\Sigma^2(f)(\mathbf{x}) = \psi(\beta)V.\end{aligned}$$

It now follows from Theorem 12 that $\Pr\{f - E[f] > t\} \leq \exp(\psi(\beta)V - \beta t)$ for any $\beta > 0$. Substitution of $\beta = \ln\left(1 + tV^{-1}\right)$ gives the first inequality, the second follows from Lemma 20. □

Observe that f is assumed bounded above by the assumptions of the theorem. The existence of exponential moments $E\left[e^{\beta f}\right]$ is needed only for $\beta \geq 0$, so the assumption $f \in \mathcal{A}$ can be dropped in this case.

If f is additive the theorem reduces to the familiar Bennett and Bernstein inequalities [16].

Corollary 22 *Let X_k be real random variables $X_k \leq E[X_k] + 1$ and let $V = \sum_k \sigma^2(X_k)$. Then*

$$\begin{aligned}\Pr\left\{\sum_k (X_k - E[X_k]) > t\right\} &\leq \exp\left(-V\left(\left(1 + tV^{-1}\right)\ln\left(1 + tV^{-1}\right) - tV^{-1}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2V + 2t/3}\right).\end{aligned}$$

Theorem 21 and its corollary can be applied to functions satisfying $f - E_k[f] < b$ by a simple rescaling argument. Then Bernstein's inequality becomes

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega}\Sigma^2(f)(\mathbf{x}) + 2bt/3}\right).$$

Inequalities of this kind exhibit two types of tails, depending in which of the two terms in the denominator $A + Bt$ of the exponent is dominant. In the sub-Gaussian regime $A >> Bt$ the tail decays as $e^{-t^2/A}$. This is the way the bounded difference inequality behaves globally, but with a very crude approximation for A, while Bernstein's inequality uses variance information. But for larger deviations, when $A << Bt$, the tail only decays as $e^{-t/A}$. This subexponential behavior is absent in the bounded difference inequality and the price paid for the fine-tuning in Bernstein's inequality.

3.6 Vector-Valued Concentration Revisited

We look again at the situation of Sect. 3.3. Suppose again that the X_i are independent zero mean random variables with values in normed space, which we now assume to be a Hilbert space H, but that now we have a uniform bound $\|X_i\| \leq c$. Again we define $f : \{y \in H : \|y\| \leq c\}^n \to \mathbb{R}$ by $f(\mathbf{x}) = \left\|\sum_i x_i\right\|$ and observe that for $y, y' \in H$, $D^k_{y,y'} f(\mathbf{x}) \leq \left\|y - y'\right\|$. This implies that $f - E_k[f] \leq 2c$ and also

$$\sigma_k^2(f) = \frac{1}{2} E_{(y,y')\sim\mu_k^2} \left(D^k_{y,y'} f(\mathbf{x})\right)^2 \leq \frac{1}{2} E_{(y,y')\sim\mu_k^2} \left\|y - y'\right\|^2 = E\|X_k\|^2 .$$

Thus $\Sigma^2(f) \leq \sum_i E\|X_i\|^2$ and by Bernstein's inequality, Theorem 21,

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sum_i E\|X_i\|^2 + 4ct/3}\right),$$

or that for $\delta > 0$ with probability at least $1 - \delta$ in $(X_1, ..., X_n)$

$$\left\|\sum_i X_i\right\| \leq \sqrt{\sum_i E\left[\|X_i\|^2\right]} + \sqrt{2\sum_i E\|X_i\|^2 \ln(1/\delta)} + 4c\ln(1/\delta)/3,$$

where we again used Jensen's inequality to bound $E\left\|\sum_i X_i\right\|$. If all the X_i are iid we get with probability at least $1 - \delta$

$$\left\|\frac{1}{n}\sum_i X_i\right\| \leq \sqrt{\frac{E\left[\|X_1\|^2\right]}{n}}\left(1 + \sqrt{2\ln(1/\delta)}\right) + \frac{4c\ln(1/\delta)}{2n}.$$

If the variance $E\left[\|X_1\|^2\right]$ is small and n is large, this is much better than the bound (14), which we got from the bounded difference inequality.

4 Inequalities for Lipschitz Functions and Dimension Free Bounds

We now prove some more advanced concentration inequalities. First we will use the bounded difference inequality to prove a famous sub-gaussian bound for Lipschitz functions of independent standard normal variables. We then derive an exponential Efron–Stein inequality which allows to prove a similar result for convex Lipschitz functions on $[0, 1]^n$. We also obtain a concentration inequality for the operator norm of a random matrix, with deviations independent of the size of the matrix.

4.1 Gaussian Concentration

The advantage of the bounded difference inequality, Theorem 14, over its simplified Corollary 15 is the supremum over $\mathbf{x}$ outside the sum over k. This allows us to prove the following powerful Gaussian concentration inequality (Tsirelson-Ibragimov–Sudakov inequality, Theorem 5.6 in [6]). We assume $\Omega_k = \mathbb{R}$ and μ_k to be the distribution of a standard normal variable, and we require f to be an L-Lipschitz function, which means that for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$

$$f(\mathbf{x}) - f\left(\mathbf{x}'\right) \leq L \left\|\mathbf{x} - \mathbf{x}'\right\|,$$

where $\|.\|$ is the Euclidean norm on $\mathbb{R}^n$.

Theorem 23 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz and let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent standard normal variables. Then for any $s > 0$*

$$\Pr\{f(\mathbf{X}) > \mathbb{E}f(\mathbf{X}) + s\} \leq e^{-s^2/2L^2}.$$

Note that the function f is not assumed to be bounded on $\mathbb{R}^n$.

Proof The idea of the proof is to use the central limit theorem to approximate the X_i by appropriately scaled Rademacher sums $h_K(\epsilon_i)$ and to apply the bounded difference inequality to $f(h_K(\epsilon_1), ..., h_K(\epsilon_n))$.

By an approximation argument using convolution with Gaussian kernels of decreasing width it suffices to prove the result if f is C^2 with $\left|\left(\partial^2/x_i^2\right) f(\mathbf{x})\right| \leq B$ for all $\mathbf{x} \in \mathbb{R}^n$ and $i \in \{1, ..., n\}$, where B is a finite, but arbitrarily large constant. For $K \in \mathbb{N}$ define a function $h_K : \{-1, 1\}^K \to \mathbb{R}$, a vector-valued function $\mathbf{h}_K : \{-1, 1\}^{Kn} \to \mathbb{R}^n$ and a function $G_K : \{-1, 1\}^{Kn} \to \mathbb{R}$ by

$$\begin{aligned} h_K(\epsilon) &= \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \epsilon_k, \text{ for } \epsilon \in \{-1, 1\}^K \\ \mathbf{h}_K(\epsilon) &= (h_K(\epsilon_1), ..., h_K(\epsilon_n)) \text{ for } \epsilon = (\epsilon_1, ..., \epsilon_n) \in \{-1, 1\}^{Kn} \\ G_K &= f(\mathbf{h}_K(\epsilon)) \text{ for } \epsilon \in \{-1, 1\}^{Kn}. \end{aligned}$$

We will use Theorem 14 on the function G_K applied to independent Rademacher variables ϵ.

Fix an arbitrary configuration $\epsilon \in \{-1, 1\}^{Kn}$ and let $\mathbf{x} = (x_1, ..., x_n) = \mathbf{h}_K(\epsilon)$. For each $i \in \{1, ..., n\}$ we introduce the real function $f_i(t) = S_t^i f(\mathbf{x})$, so that we replace the i-th argument x_i by t, leaving all the other x_j fixed. Since f is C^2 we have for any $t \in \mathbb{R}$

$$f_i(x + t) - f_i(x) = t f_i'(x) + \frac{t^2}{2} f_i''(s)$$

for some $s \in \mathbb{R}$, and by the Lipschitz condition and the bound on $\left|f_i''\right|$

$$\begin{aligned}\left(f_i\left(x+t\right)-f_i\left(x\right)\right)^2 &= t^2\left(f_i'\left(x\right)\right)^2 + t^3 f_i'\left(x\right) f_i''\left(s\right) + \frac{t^4}{4}\left(f_i''\left(s\right)\right)^2 \\ &\leq t^2\left(f_i'\left(x\right)\right)^2 + \left|t\right|^3 LB + \frac{t^4}{4}B^2.\end{aligned}$$

Now fix a pair of indices (i,k) with $i \in \{1,...,n\}$ and $k \in \{1,...,K\}$ and arbitrary values $y, y' \in \{-1,1\}$ with $y \neq y'$. We want to bound $\left(D_{y,y'}^{(i,k)} G_K\left(\epsilon\right)\right)^2$. Now either one of y or y' is equal to ϵ_{ik}, so either $S_y^{(i,k)} G_K\left(\epsilon\right)$ or $S_{y'}^{(i,k)} G_K\left(\epsilon\right)$ is equal to $G_K\left(\epsilon\right)$. Without loss of generality we assume the second. Furthermore $S_y^k h_K\left(\epsilon_i\right)$ and $h_K\left(\epsilon_i\right)$ differ by at most $2/\sqrt{K}$, so

$$\begin{aligned}\left(D_{y,y'}^{(i,k)} G_K\left(\epsilon\right)\right)^2 &= \left(f\left(x_1,...,S_y^k h_K\left(\epsilon_i\right),...,x_n\right) - f\left(x_1,...,h_K\left(\epsilon_i\right),...,x_n\right)\right)^2 \\ &= \left(f_i\left(h_K\left(\epsilon_i\right) \pm \frac{2}{\sqrt{K}}\right) - f_i\left(h_K\left(\epsilon_i\right)\right)\right)^2 \\ &\leq \frac{4 f_i'\left(h_K\left(\epsilon_i\right)\right)^2}{K} + \frac{8LB}{K^{3/2}} + \frac{4B^2}{K^2}.\end{aligned}$$

Now $f_i'\left(h_K\left(\epsilon_i\right)\right)$ is just equal to $\left(\partial/\partial x_i\right) f\left(\mathbf{x}\right)$, so

$$\sum_i f_i'\left(h_K\left(\epsilon_i\right)\right)^2 \leq \sup_{\mathbf{x}\in\mathbb{R}^n} \left\|\nabla f\left(\mathbf{x}\right)\right\|^2 \leq L^2.$$

Since ϵ was arbitrary we have

$$\sup_{\epsilon} \sum_{k,i} \sup_{y,y'} \left(D_{y,y'}^{(i,k)} G_K\left(\epsilon\right)\right)^2 \leq 4L^2 + \frac{8nLB}{K^{1/2}} + \frac{4nB^2}{K}.$$

From Theorem 14 we conclude from $f\left(\mathbf{h}_K\left(\epsilon\right)\right) = G_K\left(\epsilon\right)$ that

$$\Pr\left\{f\left(\mathbf{h}_K\left(\epsilon\right)\right) - \mathbb{E} f\left(\mathbf{h}_K\left(\epsilon'\right)\right) > s\right\} \leq \exp\left(\frac{-s^2}{2L^2 + 4nLB/K^{1/2} + 2nB^2/K}\right).$$

The conclusion now follows from the central limit theorem since $h_K\left(\epsilon\right) \to \mathbf{X}$ weakly as $K \to \infty$. □

4.2 Exponential Efron Stein Inequalities

We will now use the entropy method to derive some other "dimension free" bounds of this type. We need the following very useful result.

Lemma 24 (Chebychev's association inequality) *Let g and h be real functions, X a real random variable.*

If g and h are either both nondecreasing or both nonincreasing then

$$E\left[g\left(X\right)h\left(X\right)\right] \geq E\left[g\left(X\right)\right]E\left[h\left(X\right)\right].$$

If either one of g or h is nondecreasing and the other nonincreasing then

$$E\left[g\left(X\right)h\left(X\right)\right] \leq E\left[g\left(X\right)\right]E\left[h\left(X\right)\right].$$

Proof Let X' be a random variable iid to X. Then

$$E\left[g\left(X\right)h\left(X\right)\right] - E\left[g\left(X\right)\right]E\left[h\left(X\right)\right] = \frac{1}{2}E\left[\left(g\left(X\right) - g\left(X'\right)\right)\left(h\left(X\right) - h\left(X'\right)\right)\right].$$

Now if g and h are either both nondecreasing or both nonincreasing then

$$\left(g\left(X\right) - g\left(X'\right)\right)\left(h\left(X\right) - h\left(X'\right)\right)$$

is always nonnegative, because both factors always have the same sign, in the other case it is always nonpositive. □

We use this inequality to prove a bound on the thermal variance. First recall that for two iid random variables X and X' we have

$$\begin{aligned}\sigma^2\left(X\right) &= \frac{1}{2}E_{XX'}\left[\left(X - X'\right)^2\right] \\ &= \frac{1}{2}E_{XX'}\left[\left(X - X'\right)^2 1_{X>X'}\right] + \frac{1}{2}E_{XX'}\left[\left(X - X'\right)^2 1_{X<X'}\right] \\ &= E_{XX'}\left[\left(X - X'\right)_+^2\right].\end{aligned}$$

Lemma 25 *Let $0 \leq s \leq \beta$. Then*

$$\sigma_{sf}^2\left(f\right) \leq E_{x\sim\mu_{\beta f}}\left[E_{x'\sim\mu}\left[\left(f\left(x\right) - f\left(x'\right)\right)_+^2\right]\right].$$

Proof Let ψ be any real function. Lemma 6 (2) gives

$$\frac{d}{d\beta}E_{\beta f}\left[\psi\left(f\right)\right] = E_{\beta f}\left[\psi\left(f\right)f\right] - E_{\beta f}\left[\psi\left(f\right)\right]E_{\beta f}\left[f\right]. \tag{17}$$

By Chebychev's association inequality $E_{\beta f}[\psi(f)]$ is nonincreasing (nondecreasing) in β if ψ is nonincreasing (nondecreasing). Now define $g : \mathbb{R}^2 \to \mathbb{R}$ by

$$g(s,t) = E_{x\sim\mu_{sf}}\left[E_{x'\sim\mu_{tf}}\left[\left(f(x) - f\left(x'\right)\right)^2 1_{f(x)\geq f(x')}\right]\right],$$

so that

$$\sigma_{sf}^2(f) = \frac{1}{2}E_{x\sim\mu_{sf}}\left[E_{x'\sim\mu_{sf}}\left[\left(f(x) - f\left(x'\right)\right)^2\right]\right] = g(s,s).$$

Now for fixed x the function $\left(f(x) - f\left(x'\right)\right)^2 1_{f(x)\geq f(x')}$ is nonincreasing in $f\left(x'\right)$, so $g(s,t)$ is nonincreasing in t. On the other hand, for fixed x', $\left(f(x) - f\left(x'\right)\right)^2 1_{f(x)\geq f(x')}$ is nondecreasing in $f(x)$, so $g(s,t)$ is nondecreasing in s (this involves exchanging the two expectations in the definition of $g(s,t)$). So, since $\mu_{0f} = \mu$, we get from $0 \leq s \leq \beta$ that

$$\sigma_{sf}^2(f) = g(s,s) \leq g(\beta,0) = E_{x\sim\mu_{\beta f}}\left[E_{x'\sim\mu}\left[\left(f(x) - f\left(x'\right)\right)_+^2\right]\right].$$

□

Here is another way to write the conclusion: let $h \in \mathcal{A}$ be defined by $h(x) = E_{x'\sim\mu}\left[\left(f(x) - f\left(x'\right)\right)_+^2\right]$. Then $\sigma_{sf}^2(f) \leq E_{\beta f}[h]$.

Define two operators $D^2 : \mathcal{A} \to \mathcal{A}$ and $V_+^2 : \mathcal{A} \to \mathcal{A}$ by

$$D^2 f = \sum_k \left(f - \inf_{y\in\Omega_k} S_y^k f\right)^2$$

$$\text{and } V_+^2 f = \sum_k E_{y\sim\mu_k}\left[\left(\left(f - S_y^k f\right)_+\right)^2\right].$$

Clearly $V_+^2 f \leq D^2 f$ as $D^2 f$ is obtained by bounding the expectations in the definition of V_+^2 by their suprema.

Lemma 26 *For $\beta > 0$ and $f \in \mathcal{A}$*

$$Ent_f(\beta) \leq \frac{\beta^2}{2}E_{\beta f}\left[V^+(f)\right].$$

Proof For $k \in \{1, ..., n\}$ write $h_k = E_{y\sim\mu_k}\left[\left(f - S_y^k f\right)_+^2\right]$, so that $V^+(f) = \sum_k h_k$. The conditional version of Lemma 25 then reads for $0 \leq s \leq \beta$ and $k \in \{1, ..., n\}$

$$\sigma_{k,sf}^2(f) \leq E_{k,\beta f}[h_k].$$

Theorem 12 gives

$$\begin{aligned}\mathrm{Ent}_f(\beta) &\leq \int_0^\beta \int_t^\beta \sum_k E_{\beta f}\left[\sigma_{k,sf}^2(f)\right] ds dt \\ &\leq \int_0^\beta \int_t^\beta \sum_k E_{\beta f}\left[E_{k,\beta f}[h_k]\right] ds dt \\ &= \int_0^\beta \int_t^\beta \sum_k E_{\beta f}[h_k]\, ds dt \\ &= \frac{\beta^2}{2} E_{\beta f}\left[V^+(f)\right],\end{aligned}$$

where we used the identity $E_{\beta f}\left[E_{k,\beta f}[h]\right] = E_{\beta f}[h]$ for $h \in \mathcal{A}$. □

The usual arguments involving Theorem 12 and an optimization in β now immediately lead to

Theorem 27 *With $t > 0$*

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} V_+^2 f(\mathbf{x})}\right) \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} D^2 f(\mathbf{x})}\right).$$

We get a corresponding lower tail bound only for D^2 and we have to use an estimate similar to what was used in the proof of Bennett's inequality.

Lemma 28 *If $f - \inf_k f \leq 1, \forall k$ then for $\beta > 0$*

$$Ent_{-f}(\beta) \leq \psi(\beta) E_{-\beta f}\left[D^2 f\right],$$

with $\psi(t) = e^t - t - 1$ defined as in (16).

Proof Let $k \in \{1, ..., n\}$. We write $h_k := f - \inf_k f$. Then $h_k \in [0, 1]$ and for $s \leq \beta$ we have $1 \leq e^{(\beta-s)h_k} \leq e^{\beta-s}$, so

$$E_{k,-sh_k}\left[h_k^2\right] = \frac{E_k\left[h_k^2 e^{-\beta h_k} e^{(\beta-s)h_k}\right]}{E_k\left[e^{-\beta h_k} e^{(\beta-s)h_k}\right]} \leq e^{(\beta-s)} \frac{E_k\left[h_k^2 e^{-\beta h_k}\right]}{E_k\left[e^{-\beta h_k}\right]} = e^{(\beta-s)} E_{k,-\beta h_k}\left[h_k^2\right].$$

We therefore have

$$\begin{aligned}\int_0^\beta \int_t^\beta E_{k,-sf}\left[h_k^2\right] ds\, dt &= \int_0^\beta \int_t^\beta E_{k,-sh_k}\left[h_k^2\right] ds\, dt \\ &\leq \left(\int_0^\beta \int_t^\beta e^{\beta-s} ds\, dt\right) E_{k,-\beta h_k}\left[h_k^2\right] = \psi(\beta) E_{k,-\beta f}\left[h_k^2\right],\end{aligned}$$

where we used the formula

$$\int_0^\beta \int_t^\beta e^{-s} ds\, dt = 1 - e^{-\beta} - \beta e^{-\beta}.$$

Thus, using Theorem 12 and the identity $E_{-\beta f} E_{k,-\beta f} = E_{-\beta f}$,

$$\begin{aligned}\mathrm{Ent}_{-f}(\beta) &\leq E_{-\beta f}\left[\sum_k \int_0^\beta \int_t^\beta \sigma_{k,-sf}^2[f]\,ds\,dt\right] \leq E_{-\beta f}\left[\sum_k \int_0^\beta \int_t^\beta E_{k,-sf}\left[h_k^2\right] ds\,dt\right] \\ &\leq \psi(\beta) E_{-\beta f}\left[\sum_k E_{k,-\beta f}\left[h_k^2\right]\right] = \psi(\beta) E_{-\beta f}\left[D^2 f\right].\end{aligned}$$

□

Lemmas 26 and 28 together with (7) imply the inequalities

$$\ln E\left[e^{\beta(f-E[f])}\right] \leq \frac{\beta}{2}\int_0^\beta E_{\gamma f}\left[V_+^2 f\right] d\gamma. \tag{18}$$

and, if $f - \inf_k f \leq 1$ for all k, then

$$\ln E\left[e^{\beta(E[f]-f)}\right] \leq \frac{\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}\left[D^2 f\right] d\gamma, \tag{19}$$

where in the last inequality we also used the fact that $\gamma \mapsto \psi(\gamma)/\gamma^2$ is nondecreasing. Bounding the thermal expectation with the uniform norm and substitution of $\beta = \ln\left(1 + t\left\|D^2 f\right\|_\infty^{-1}\right)$ gives the following lower tail bound as in the proof of the Bennett-Bernstein inequalities.

Theorem 29 *If $f - \inf_k f \leq 1$ for all k, then for $t > 0$ and with $\Delta := \sup_{\mathbf{x}\in\Omega} D^2 f(\mathbf{x})$*

$$\begin{aligned}\Pr\{Ef - f > t\} &\leq \exp\left(-\Delta\left(\left(1 + \frac{t}{\Delta}\right)\ln\left(1 + \frac{t}{\Delta}\right) - \frac{t}{\Delta}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} D^2 f(\mathbf{x}) + 2t/3}\right).\end{aligned}$$

4.3 Convex Lipschitz Functions

In Sect. 4.1 we gave a sub-gaussian bound for Lipschitz functions of independent standard normal variables. Now we prove the same upper tail bound under different hypotheses. Instead of assuming μ_k to be standard normal we require $\Omega_k = [0, 1]$ and let μ_k be perfectly arbitrary. On the other hand, in addition to being an L-Lipschitz function, we require f to be convex (actually only separately convex in each argument).

Theorem 30 *Let $\Omega_k = \mathcal{I}$, an interval of unit diameter, and let $f \in \mathcal{A}$ be C^1, L-Lipschitz and such that $y \in [0, 1] \mapsto S_y^k f(\mathbf{x})$ is convex for all k and all $\mathbf{x}$. Then*

$$\Pr\{f - Ef > t\} \leq e^{-t^2/2L^2}.$$

Proof By an approximation argument we can assume f to be differentiable. Let $\mathbf{x} \in [0,1]^n$, $k \in \{1, ..., n\}$ and $y \in [0,1]$ such that $S_y^k f(\mathbf{x}) \leq f(\mathbf{x})$. Then, using separate convexity,

$$f(\mathbf{x}) - S_y^k f(\mathbf{x}) \leq \left\langle \mathbf{x} - S_y^k \mathbf{x}, \partial f(\mathbf{x}) \right\rangle_{\mathbb{R}^n} = (x_k - y) \frac{\partial}{\partial x_k} f(\mathbf{x}) \leq \left| \frac{\partial}{\partial x_k} f(\mathbf{x}) \right|.$$

We therefore have $f(\mathbf{x}) - \inf_y S_y^k f(\mathbf{x}) \leq |(\partial/\partial x_k) f(\mathbf{x})|$ and

$$D^2 f(\mathbf{x}) = \sum_{k=1}^{n} \left(f(\mathbf{x}) - \inf_y S_y^k f(\mathbf{x}) \right)^2 \leq \|\nabla f(\mathbf{x})\|_{\mathbb{R}^n}^2 \leq L^2.$$

Theorem 27 then gives the conclusion. □

For future reference we record the following fact: if Ω_k is an interval of unit diameter and A an $m \times n$-matrix then $\mathbf{x} \mapsto \|A\mathbf{x}\|$ is a convex function with Lipschitz constant $\|A\|$ and thus

$$D^2(\|A\mathbf{x}\|) \leq \|A\|^2. \tag{20}$$

4.4 The Operator Norm of a Random Matrix

For $\mathbf{x} \in [-1,1]^{n^2}$ let $M(\mathbf{x})$ be the $n \times n$ matrix whose entries are given by the components of $\mathbf{x}$. We are interested in the concentration properties of the operator norm of $M(\mathbf{X})$, when $\mathbf{X}$ is a vector with independent, but possibly not identically distributed components chosen from $[-1,1]$. The function in question is then $f : [-1,1]^{n^2} \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \|M(\mathbf{x})\| = \sup_{\|w\|, \|v\| = 1} \langle M(\mathbf{x}) v, w \rangle,$$

where $\langle ., . \rangle$ and $\|.\|$ refer to inner product and norm in $\mathbb{R}^n$.

To bound $D^2 f(\mathbf{x})$ first let $\mathbf{x} \in [-1,1]^{n^2}$ be arbitrary but fixed, and let v and w be unit vectors witnessing the supremum in the definition of $f(\mathbf{x})$.

Now let (k,l) be any index to a matrix entry and choose any $y \in [-1,1]$ such that $S_y^{(k,l)} f(\mathbf{x}) \leq f(\mathbf{x})$. Then

$$\begin{aligned} f(\mathbf{x}) - S_y^{(k,l)} f(\mathbf{x}) &= \langle M(\mathbf{x}) v, w \rangle - \sup_{\|w'\|, \|v'\| = 1} \left\langle S_y^{(k,l)} M(\mathbf{x}) v', w' \right\rangle \\ &\leq \left\langle \left(M(\mathbf{x}) - S_y^{(k,l)} M(\mathbf{x}) \right) v, w \right\rangle = (x_{kl} - y) v_k w_l \\ &\leq 2 |v_k| |w_l|. \end{aligned}$$

Note that $f - \inf_k f \leq 2$. Also

$$D^2 f(\mathbf{x}) = \sum_{k,l} \left(f(\mathbf{x}) - \inf_{y \in [-1,1]} S_y^{(k,l)} f(\mathbf{x}) \right)^2$$
$$\leq 4 \sum_{k,l} |v_k|^2 |w_l|^2 = 4.$$

The results of the previous section (rescaling for the lower tail to get $f - \inf_k f \leq 1$) then lead to a concentration inequality independent of the size of the random matrix.

Theorem 31 *Let $\mathbf{X} = \left(X_{ij}\right)_{1 \leq i,j \leq n}$ be a vector of n^2 independent random variables with values in $[-1, 1]$, and $\mathbf{X}'$ iid to $\mathbf{X}$. Then for $t > 0$.*

$$\Pr\left\{ \|M(\mathbf{X})\| - E\left[\left\| M\left(\mathbf{X}'\right) \right\| \right] \geq t \right\} \leq \exp\left(\frac{-t^2}{8} \right)$$

and

$$\Pr\left\{ E\left[\left\| M\left(\mathbf{X}'\right) \right\| \right] - \|M(\mathbf{X})\| \geq t \right\} \leq \exp\left(\frac{-t^2}{8 + 4t/3} \right).$$

Observe that the argument depends on the fact that the unit vectors v and w could be fixed independent of k and l. This would not have been possible with the bounded difference inequality. Also note that square matrices were chosen for notational convenience only. The same proof would work for rectangular matrices.

5 Beyond Uniform Bounds

All of the above applications of the entropy method to derive upper tail bounds involved an inequality of the form

$$\mathrm{Ent}_f(\gamma) \leq \xi(\gamma) E_{\gamma f}[G(f)],$$

where ξ is some nonnegative real function and G is some operator $G : \mathcal{A} \to \mathcal{A}$, which is positively homogeneous of order two. For the bounded difference inequality $\xi(\gamma) = \gamma^2/8$ and $G = R^2$, for the Bennett inequality $\xi(\gamma) = \gamma e^{\gamma} - e^{\gamma} + 1$ and $G = \Sigma^2$, for Theorem 27 we had $\xi(\gamma) = \gamma^2/2$ and $G = V_+^2$. Theorem 12 is then used to conclude that

$$\ln E e^{\beta(f - Ef)} \leq \beta \int_0^{\beta} \frac{\xi(\gamma)}{\gamma^2} E_{\gamma f}[G(f)] \, d\gamma \leq \beta \sup_{\mathbf{x}} G(f)(\mathbf{x}) \int_0^{\beta} \frac{\xi(\gamma) \, d\gamma}{\gamma^2}. \tag{21}$$

An analogous strategy was employed for the various lower tail bounds.

The uniform estimate $E_{\gamma f}[G(f)] \le \sup_{\mathbf{x}} G(f)(\mathbf{x})$ in (21), while being very simple, is somewhat loose and can sometimes be avoided by exploiting special properties of the thermal expectation and the function in question.

5.1 Self-boundedness

The first possibility we consider is that the function $G(f)$ can be bounded in terms of the function f itself, a property referred to as *self-boundedness* [8]. For example, if simply $G(f) \le f$, then $E_{\gamma f}[G(f)] \le E_{\gamma f}[f] = (d/d\gamma) \ln Z_{\gamma f}$, and if the function ξ has some reasonable behavior, then the first integral in (21) above can be bounded by partial integration or even more easily. As an example we apply this idea in the setting of Theorems 27 and 29.

Lemma 32 *Suppose that for $f \in \mathcal{A}$ there are nonnegative numbers a, b such that*
(i) $V_+^2 f \le af + b$. Then for $0 \le \beta < 2/a$

$$\ln E\left[e^{\beta(f-E[f])}\right] \le \frac{\beta^2 (aEf + b)}{2 - a\beta},$$

(ii) $D^2 f \le af + b$. If in addition $f - \inf_k f \le 1$ for all k, then for $\beta < 0$ and $a \ge 1$

$$\ln E\left[e^{\beta(E[f]-f)}\right] \le \frac{\beta^2 (aE[f] + b)}{2}.$$

Proof (i) We use (18) and get

$$\begin{aligned}\ln E\left[e^{\beta(f-E[f])}\right] &\le \frac{\beta}{2}\int_0^\beta E_{\gamma f}\left[V_+^2 f\right] d\gamma \le \frac{a\beta}{2}\int_0^\beta E_{\gamma f}[f]\, d\gamma + \frac{b\beta^2}{2} \\ &= \frac{a\beta}{2} \ln Z_{\beta f} + \frac{b\beta^2}{2},\end{aligned}$$

where the last identity follows from the fact that $E_{\gamma f}[f] = (d/d\gamma) \ln Z_{\gamma f}$. Thus

$$\ln E\left[e^{\beta(f-E[f])}\right] \le \frac{a\beta}{2} \ln E e^{\beta(f-E[f])} + \frac{a\beta^2}{2} Ef + \frac{b\beta^2}{2},$$

and rearranging this inequality for $\beta \in (0, 2/a)$ establishes the claim.

(ii) We use (19)

$$\begin{aligned}\ln E\left[e^{\beta(E[f]-f)}\right] &\leq \frac{\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}\left[D^2 f\right]d\gamma \\ &\leq \frac{a\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}[f]\,d\gamma + b\psi(\beta) = \frac{-a\psi(\beta)}{\beta}\ln Z_{-\beta f} + b\psi(\beta) \\ &= \frac{-a\psi(\beta)}{\beta}\ln E\left[e^{\beta(E[f]-f)}\right] + \psi(\beta)\left(aE[f]+b\right).\end{aligned}$$

Rearranging gives

$$\ln E\left[e^{\beta(E[f]-f)}\right] \leq \frac{\psi(\beta)}{1+a\beta^{-1}\psi(\beta)}\left(aE[f]+b\right) \leq \frac{\beta^2\left(aE[f]+b\right)}{2},$$

where one verifies that for $\beta > 0$ and $a \geq 1$ we have $\psi(\beta)\left(1+a\beta^{-1}\psi(\beta)\right)^{-1} \leq \beta^2/2$. □

The bound in part (i) requires an upper bound on β. To proceed we need the following optimization lemma, which will be used several times in the sequel and leads to tail bounds with both sub-Gaussian and subexponential regimes, similar to Bernstein's inequality.

Lemma 33 *Let C and b denote two positive real numbers, $t > 0$. Then*

$$\inf_{\beta\in[0,1/b)}\left(-\beta t + \frac{C\beta^2}{1-b\beta}\right) \leq \frac{-t^2}{2\left(2C+bt\right)}. \tag{22}$$

Proof Let $h(t) = 1 + t - \sqrt{1+2t}$. Then use

$$\begin{aligned}2h(t)(1+t) &= 2(1+t)^2 - 2(1+t)\sqrt{1+2t} \\ &= (1+t)^2 - 2(1+t)\sqrt{1+2t} + (1+2t) + t^2 \\ &= \left(1+t-\sqrt{1+2t}\right)^2 + t^2 \\ &\geq t^2,\end{aligned}$$

so that

$$h(t) \geq \frac{t^2}{2(1+t)}. \tag{23}$$

Substituting

$$\beta = \frac{1}{b}\left(1 - \left(1+\frac{bt}{C}\right)^{-1/2}\right)$$

in the left side of (22) we obtain

$$\inf_{\beta\in[0,1/b)}\left(-\beta t+\frac{C\beta^2}{1-b\beta}\right)\leq-\frac{2C}{b^2}h\left(\frac{bt}{2C}\right)\leq\frac{-t^2}{2\,(2C+bt)},$$

where we have used (23). □

Theorem 34 *Suppose for $f\in A$ there are nonnegative numbers a, b such that*
(i) $V_+^2 f\leq af+b$. Then for $t>0$ we have

$$\Pr\{f-E[f]>t\}\leq\exp\left(\frac{-t^2}{2\,(aE[f]+b+at/2)}\right).$$

(ii) $D^2 f\leq af+b$. If in addition, $a\geq 1$ and $f-\inf_k f\leq 1, \forall k\in\{1,\ldots,n\}$, then

$$\Pr\{E[f]-f>t\}\leq\exp\left(\frac{-t^2}{2\,(aE[f]+b)}\right).$$

Proof Part (i) follows from Lemmas 32 (i) and Lemma 33). Part (ii) is immediate from Lemma 32 (ii). □

Boucheron et al. [8] have given a refined version for the lower tail, where the condition $a\geq 1$ is relaxed to $a\geq 1/3$ for the lower tail. There they also show that Theorems 34 and 27 together suffice to derive a version of the convex distance inequality which differs from Talagrand's original result only in that it has an inferior constant in the exponent.

5.2 *Convex Lipschitz Functions Revisited*

In Sect. 4.3 we gave a sub-Gaussian bound for the upper tail of separately convex Lipschitz functions on $[0,1]^n$. Now we use self-boundedness to complement this with a sub-Gaussian lower bound, using an elegant trick of Boucheron et al. [6] where the lower bound in Theorem 34 is applied to the square of the Lipschitz function f. The essence of the trick is the following simple lemma.

Lemma 35 *If $f\geq 0$ then $D^2\left(f^2\right)\leq 4D^2(f)\,f^2$.*

Proof Since $f\geq 0$ we have $\inf_k\left(f^2\right)=(\inf_k f)^2$, so, using $(a+b)^2\leq 2a^2+2b^2$,

$$\begin{aligned}D^2\left(f^2\right)&=\sum_k\left(f^2-\inf_k f^2\right)^2=\sum_k\left(f-\inf_k f\right)^2\left(f+\inf_k f\right)^2\\&\leq 4f^2\sum_k\left(f-\inf_k f\right)^2=4D^2(f)\,f^2.\end{aligned}$$

□

For the sub-Gaussian lower bound we need the additional assumption that f^2 takes values in an interval of length at most one.

Theorem 36 *Let $\Omega_k = [0,1]$ and let $f \in \mathcal{A}$ be L-Lipschitz, nonnegative and such that $y \in [0,1] \mapsto S_y^k f(\mathbf{x})$ is convex for all k and all $\mathbf{x}$, and suppose in addition, that f^2 takes values in an interval of length at most one. Then for all $t \in [0, E[f]]$*

$$\Pr\{E[f] - f > t\} \leq e^{-t^2/8L^2}.$$

Proof The trick is to study the function f^2 instead of f. Let $\mathbf{x} \in [0,1]^n$. Using separate convexity as in the proof of Theorem 30 we have $D^2 f \leq L^2$, so by the previous lemma $D^2(f)^2 \leq 4L^2 f^2$. For any k we have $f^2(\mathbf{x}) - \inf f_k^2(\mathbf{x}) \leq 1$, so by the lower tail bound of Theorem 34 we get a lower tail bound for f^2

$$\Pr\left\{E\left[f^2\right] - f^2 > t\right\} \leq \exp\left(\frac{-t^2}{8L^2 E\left[f^2\right]}\right).$$

Thus

$$\begin{aligned}
\Pr\{E[f] - f > t\} &= \Pr\left\{\sqrt{E\left[f^2\right]}(E[f] - f) > \sqrt{E\left[f^2\right]}t\right\} \\
&\leq \Pr\left\{\left(\sqrt{E\left[f^2\right]} + f\right)\left(\sqrt{E\left[f^2\right]} - f\right) > \sqrt{E\left[f^2\right]}t\right\} \\
&= \Pr\left\{E\left[f^2\right] - f^2 > \sqrt{E\left[f^2\right]}t\right\} \\
&\leq \exp\left(\frac{-t^2}{8L^2}\right).
\end{aligned}$$

Here we used $E[f] \leq \sqrt{E\left[f^2\right]}$ and the assumption that f is nonnegative in the first inequality. □

5.3 Decoupling

A second method to avoid the uniform bound on the thermal expectation uses decoupling. By the duality formula of Theorem 4 we have for any $f, g \in \mathcal{A}$ and $\beta \in \mathbb{R}$

$$E_{\beta f}[g] \leq \mathrm{Ent}_f(\beta) + \ln E\left[e^g\right]. \tag{24}$$

Recall the discussion at the beginning of Sect. 5, where we had a general bound of the form $\mathrm{Ent}_f(\beta) \leq \xi(\beta) E_{\beta f}[G(f)]$. Using (24) we can now obtain for any $\lambda > 0$

$$\mathrm{Ent}_f(\beta) \le \xi(\beta)\lambda^{-1} E_{\beta f}[\lambda G(f)] \le \xi(\beta)\lambda^{-1}\left(\mathrm{Ent}_f(\beta) + \ln E\left[\exp(\lambda G(f))\right]\right),$$

and for values of β and λ where $\lambda > \xi(\beta)$ we obtain

$$\begin{aligned}\mathrm{Ent}_f(\beta) &\le \frac{\xi(\beta)}{\lambda - \xi(\beta)} \ln E\left[\exp(\lambda G(f))\right] \\ &= \frac{\xi(\beta)}{\lambda - \xi(\beta)}\left(\ln E\left[e^{\lambda(G(f) - E[G(f)])}\right] + \lambda E[G(f)]\right).\end{aligned} \tag{25}$$

Hence, if we can control the moment generating function of $G(f)$ (or some suitable bound thereof), we obtain concentration inequalities for f, effectively passing from the thermal measure $\mu_{\beta f}$ to the thermal measure $\mu_{\lambda G(f)}$. The second line shows that in this way the supremum of $G(f)$ can possibly be replaced by an expectation. The $\lambda - \xi(\beta)$ in the denominator makes some constraint on β necessary, so the improvement comes at the price of an extra or enlarged subexponential term in the resulting concentration inequality. We conclude this chapter with three applications of this trick, which has been proposed in [7].

5.4 *Quadratic Forms*

As a first illustration we give a version of the Hanson-Wright inequality (Theorem 6.2.1 in [29]) for bounded variables. Let A be a symmetric $n \times n$-matrix, which is zero on the diagonal, that is $A_{ii} = 0$ for all i, and suppose that $X_1, \ldots, X_n$ are independent random variables with values in an interval $\mathcal{I}$ of unit diameter. We study the random variable $f(\mathbf{X})$, where

$$f(\mathbf{x}) = \sum_{i,j} x_i A_{ij} x_j.$$

As operator G we use R^2, the sum of squared conditional ranges which appears in the bounded difference inequality. For the function in question we have

$$D^k_{y,y'} f(\mathbf{x}) = 2\left(y - y'\right)\sum_i A_{ki} x_i = 2\left(y - y'\right)(A\mathbf{x})_k,$$

and, since $\mathcal{I}$ has unit diameter

$$R^2(f)(\mathbf{x}) = \sum_k \sup_{y,y' \in \mathcal{I}} \left(D^k_{y,y'} f(\mathbf{x})\right)^2 \le 4\sum_k (A\mathbf{x})_k^2 = 4\|A\mathbf{x}\|^2.$$

We can therefore conclude from (12) in the proof of the bounded difference inequality (Theorem 14), that $\mathrm{Ent}_f(\gamma) \le \left(\gamma^2/8\right) E_{\gamma f}\left[R^2(f)\right] \le \left(\gamma^2/2\right) E_{\gamma f}\left[\|A\mathbf{X}\|^2\right]$. But instead of bounding the last thermal expectation by a supremum, as we did before, we now look for concentration properties of the function $\mathbf{x} \mapsto \|A\mathbf{x}\|^2$.

By (20) and Lemma 35 we have the self-bounding inequality $D^2\left(\|A\mathbf{x}\|^2\right) \leq 4\|A\|^2\|A\mathbf{x}\|^2$ and Lemma 32 gives for $0 \leq \lambda < 1/\left(2\|A\|^2\right)$

$$\ln E\left[e^{\lambda\|A\mathbf{x}\|^2}\right] \leq \frac{\lambda E\left[\|A\mathbf{x}\|^2\right]}{1-2\|A\|^2\lambda}.$$

Now Let $0 < \gamma < 1/\|A\|$ and set $\lambda := \gamma/(2\|A\|) < 1/\left(2\|A\|^2\right)$. Using the above bound on $\mathrm{Ent}_f(\gamma)$ and the decoupling inequality (24) we get

$$\begin{aligned}\lambda\mathrm{Ent}_f(\gamma) &\leq \frac{\gamma^2}{2}E_{\gamma f}\left[\lambda\|Ax\|^2\right] \leq \frac{\gamma^2}{2}\left(\mathrm{Ent}_f(\gamma) + \ln E\left[e^{\lambda\|Ax\|^2}\right]\right)\\ &\leq \frac{\gamma^2}{2}\mathrm{Ent}_f(\gamma) + \frac{\gamma^2}{2}\frac{\lambda E\left[\|Ax\|^2\right]}{1-2\|A\|^2\lambda}.\end{aligned}$$

Collect terms in $\mathrm{Ent}_f(\gamma)$, divide by $\lambda - \gamma^2/2$ (which is positive by the constraint on γ and the choice of λ) and substitute the value of λ to get

$$\mathrm{Ent}_f(\gamma) \leq \frac{\gamma^2}{(1-\|A\|\gamma)^2}\frac{E\left[\|Ax\|^2\right]}{2}.$$

From Theorem 12 we conclude that for $\beta < 1/\|A\|$

$$\begin{aligned}\Pr\{f - Ef\} &\leq \exp\left(\beta\int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2}d\gamma - \beta t\right)\\ &\leq \exp\left(\frac{\beta^2}{1-\|A\|\beta}\frac{E\left[\|Ax\|^2\right]}{2} - \beta t\right),\end{aligned}$$

and using Lemma 33 to minimize the last expression in $\beta \in (0, 1/\|A\|)$ gives our version of the Hanson-Wright inequality for bounded variables.

Theorem 37 *Let A be a symmetric $n \times n$-matrix, zero on the diagonal, and $\mathbf{X} = (X_1, ..., X_n)$ a vector of independent random variables with values in an interval $\mathcal{I}$ of unit diameter. Let $f : \mathcal{X}^n \to \mathbb{R}$ be defined by $f(x) = \sum_{ij} x_i A_{ij} x_j$. Then for $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2E\left[\|A\mathbf{X}\|^2\right] + 2\|A\|t}\right).$$

5.5 The Supremum of an Empirical Process

We will now apply the decoupling trick to the upwards tail of the supremum of an empirical process, sharpening the bound obtained in Sect. 3.4.

Theorem 38 *Let $X_1, ..., X_n$ be independent with values in some space $\mathcal{X}$ with X_i distributed as μ_i, and let $\mathcal{F}$ be an at most countable class of functions $f : \mathcal{X} \to [-1, 1]$ with $E[f(X_i)] = 0$. Define $F : \mathcal{X}^n \to \mathbb{R}$ and $W : \mathcal{X}^n \to \mathbb{R}$ by*

$$F(\mathbf{x}) = \sup_{f \in \mathcal{F}} \sum_i f(x_i) \text{ and}$$
$$W(\mathbf{x}) = \sup_{f \in \mathcal{F}} \sum_i \left(f^2(x_i) + E\left[f^2(X_i)\right] \right).$$

Then for $t > 0$

$$\Pr\{F - E[F] > t\} \leq \exp\left(\frac{-t^2}{2E[W] + t} \right).$$

This inequality improves over Theorem 12.2 in [6], since by the triangle inequality $E[W] \leq \Sigma^2 + \sigma^2$ and the constants in the denominator of the exponent are better by a factor of two, and optimal for the variance term.

Proof Let $0 < \gamma \leq \beta < 2$. Using Theorem 26 and (24) we get

$$\mathrm{Ent}_F(\gamma) \leq \frac{\gamma}{2} E_{\gamma F}\left[\gamma V_+^2(F)\right] \leq \frac{\gamma}{2}\left(\mathrm{Ent}_F(\gamma) + \ln E e^{\gamma V_+^2(F)}\right).$$

Rearranging gives

$$\mathrm{Ent}_F(\gamma) \leq \frac{\gamma}{2-\gamma} \ln E e^{\gamma V_+^2(F)}. \tag{26}$$

Fix some $\mathbf{x} \in \mathcal{X}^n$ and let $\hat{f} \in \mathcal{F}$ witness the maximum in the definition of $F(\mathbf{x})$. For $y \in \mathcal{X}$ we have $\left(F - S_y^k F\right)_+ \leq \left(\hat{f}(x_i) - \hat{f}(y)\right)_+$ and by the zero mean assumption

$$\begin{aligned} V_+^2(F)(\mathbf{x}) &= \sum_k E_{y \sim \mu_k}\left[\left(F(\mathbf{x}) - S_y^k F(\mathbf{x})\right)_+^2\right] \\ &\leq \sum_k E_{y \sim \mu_k}\left(\hat{f}(x_k) - \hat{f}(y)\right)_+^2 \\ &\leq \sum_k E_{y \sim \mu_k}\left(\hat{f}(x_k) - \hat{f}(y)\right)^2 \\ &= \sum_k \left(\hat{f}^2(x_k) + E\left[\hat{f}^2(X_k)\right]\right) \\ &\leq W(\mathbf{x}). \end{aligned}$$

So $V_+^2(F) \leq W$. It follows from (26) that

$$\mathrm{Ent}_F(\gamma) \leq \frac{\gamma}{2-\gamma} \ln E e^{\gamma V^+(F)} \leq \frac{\gamma}{2-\gamma} \ln E\left[e^{\gamma W}\right]. \tag{27}$$

Next we establish self-boundedness of W. Let $\hat{f} \in \mathcal{F}$ (different from the previous $\hat{f}$, which we don't need any more) witness the maximum in the definition of $W(\mathbf{x})$. Then

$$\begin{aligned} V_+^2(W)(\mathbf{x}) &= \sum_k E_{y\sim\mu_k}\left(W(\mathbf{x}) - S_y^k W(\mathbf{x})\right)_+^2 \\ &\leq \sum_k E_{y\sim\mu_k}\left[\left(\hat{f}^2(x_k) - \hat{f}^2(y)\right)_+^2\right] \\ &\leq \sum_k \hat{f}^2(x_k) \\ &\leq W. \end{aligned}$$

It therefore follows from the self-bounding lemma, Lemma 32, that

$$\ln E\left[e^{\gamma W}\right] \leq \frac{\gamma^2 E[W]}{2-\gamma} + \gamma E[W] = \frac{\gamma E[W]}{1-\gamma/2}.$$

Combining this with (27) gives

$$\mathrm{Ent}_F(\gamma) \leq \frac{\gamma}{2-\gamma}\left(\frac{\gamma E[W]}{1-\gamma/2}\right) = \frac{\gamma^2}{(1-\gamma/2)^2}\frac{E[W]}{2}.$$

From (6) in Theorem 12 we conclude that

$$\begin{aligned} \ln E e^{\beta(F-EF)} &= \beta\int_0^\beta \frac{\mathrm{Ent}_F(\gamma)}{\gamma^2}d\gamma \leq \beta\int_0^\beta \frac{1}{(1-\gamma/2)^2}d\gamma\ \frac{E[W]}{2} \\ &= \frac{\beta^2}{1-\beta/2}\frac{E[W]}{2}. \end{aligned}$$

Using Lemma 33 it follows that

$$\begin{aligned} \Pr\{F - E[F] > t\} &\leq \inf_{\beta\in(0,2)} \exp\left(-\beta t + \frac{\beta^2}{1-\beta/2}\frac{E[W]}{2}\right) \\ &\leq \exp\left(\frac{-t^2}{2E[W]+t}\right). \end{aligned}$$

□

5.6 *Another Version of Bernstein's Inequality*

A potential weakness of Theorem 21 is the occurrence of the supremum in the definition of the variance parameter $V = \sup_{\mathbf{x}\in\Omega}\Sigma^2(f)(\mathbf{x})$. If the supremum could be

replaced by an expectation, the variance parameter would become the Efron–Stein upper bound $E\left[\Sigma^2(f)\right]$ on the variance $\sigma^2(f)$, making the inequality considerably stronger. Such a modification is possible at the expense of enlarging the subexponential term in Bernstein's inequality. Define the interaction functional

$$J(f) = 2\left(\sup_{\mathbf{x},\mathbf{z}\in\Omega}\sum_{k,l:k\neq l}\sigma_k^2\left(f - S_{z_l}^l f\right)(\mathbf{x})\right)^{1/2}.$$

The following theorem is given in [21]

Theorem 39 *Suppose $f \in \mathcal{A}(\Omega)$ satisfies $f - E_k f \leq b$ for all k. Then for all $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2(f)\right] + (2b/3 + J(f))\,t}\right).$$

Here we will use the tools introduced above to prove a slight strengthening of this result, removing the boundedness conditions above.

Let $f : \Omega = \prod_{i=1}^n \Omega_i \to \mathbb{R}$ and consider the three conditions

$$(A) = ((f - E_k f) \leq b \text{ for all } k)$$
$$(B) = \left(E_k\left[(f - E_k f)^m\right] \leq \frac{1}{2}m!\sigma_k^2(f)\,b^{m-2} \text{ for } m \geq 2 \text{ and all } k\right)$$
$$(C) = \left(\sum_{k=1}^n E_k\left[(f - E_k f)^m\right] \leq \frac{\Sigma^2(f)}{2}m!b^{m-2} \text{ for } m \geq 2\right).$$

Then $(A) \implies (B) \implies (C)$. The last condition (sometimes called "Bernstein condition" in the literature) is sufficient for the following version of Bernstein's inequality, which extends Theorem 2.10 in [6] from sums to general functions and replaces the one-sided boundedness requirement of Theorem 39 by the Bernstein condition.

Theorem 40 *Let $f : \Omega = \prod_{i=1}^n \Omega_i \to \mathbb{R}$ be measurable and suppose that (C) holds. Then for $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2(f)\right] + (2b + J(f))\,t}\right).$$

The first step is to bound the entropy of f under the condition (C), thus replacing Lemma 19 in the proof of Theorem 21.

Lemma 41 *Suppose (C) holds with $b = 1$. Then for all $\beta \in [0, 1)$*

$$Ent_f(\beta) \leq \frac{\beta^2 E_{\beta f}\left[\Sigma^2(f)\right]}{2(1-\beta)^2}.$$

Proof First we get from the variational property of variance, that

$$\sigma_{k,\beta f}^2(f) \le E_{k,\beta f}\left[(f - E_k(f))^2\right] = \frac{E_k\left[(f - E_k(f))^2 e^{\beta(f-E_k f)}\right]}{E_k\left[e^{\beta(f-E_k f)}\right]}$$
$$\le E_k\left[(f - E_k(f))^2 e^{\beta(f-E_k f)}\right],$$

where we used Jensen's inequality to get $E_k\left[\exp\left(\beta\left(f - E_k f\right)\right)\right] \ge 1$ for the second inequality. From monotone convergence and (C) we then get

$$\sum_{k=1}^n \sigma_{k,\beta f}^2(f) \le \sum_{k=1}^n E_k\left[(f - E_k f)^2 e^{\beta(f-E_k f)}\right] = \sum_{m=0}^\infty \sum_{k=1}^n \frac{\beta^m}{m!} E_k\left[(f - E_k f)^{m+2}\right]$$
$$\le \frac{\Sigma^2(f)}{2} \sum_{m=0}^\infty (m+1)(m+2)\beta^m.$$

Thus from Theorem 12

$$\mathrm{Ent}_f(\beta) \le E_{\beta f}\left[\int_0^\beta \int_t^\beta \sum_{k=1}^n \sigma_{k,sf}^2(f)\, ds\, dt\right]$$
$$\le \frac{E_{\beta f}\left[\Sigma^2(f)\right]}{2} \sum_{m=0}^\infty (m+1)(m+2) \int_0^\beta \int_t^\beta s^m ds dt$$
$$= \frac{E_{\beta f}\left[\Sigma^2(f)\right]}{2} \beta^2 \sum_{m=0}^\infty (m+1)\beta^m = \frac{\beta^2 E_{\beta f}\left[\Sigma^2(f)\right]}{2(1-\beta)^2}.$$

□

At this point we could bound the thermal expectation $E_{\beta f}\left[\Sigma^2(f)\right]$ by a supremum and proceed along the usual path to obtain a version of Theorem 21 under condition (C), which, for sums of independent variables, would reduce to Theorem 2.10 in [6]. Instead we wish to exploit the decoupling idea and look for concentration properties of $\Sigma^2(f)$.

The crucial property of the interaction functional J is, that J^2 is a self-bound for $\Sigma^2(f)$. The following Lemma is also the key to the proof of Theorem 39.

Lemma 42 *We have* $D^2\left(\Sigma^2(f)\right) \le J(f)^2\ \Sigma^2(f)$ *for any* $f \in \mathcal{A}(\Omega)$.

Proof Fix $\mathbf{x} \in \Omega$. Below all members of $\mathcal{A}$ are understood as evaluated on $\mathbf{x}$. For $l \in \{1, ..., n\}$ let $z_l \in \Omega_l$ be a minimizer in z of $S_z^l \Sigma^2(f)$. Then

$$D^2\left(\Sigma^2(f)\right) = \sum_l \left(\sum_{k:k\neq l} \left(\sigma_k^2(f) - S_{z_l}^l \sigma_k^2(f)\right)\right)^2.$$

The sum over $k \neq l$, since $\sigma_k^2(f) \in \mathcal{A}_k$, so $S_{z_l}^l \sigma_k^2(f) = \sigma_k^2(f)$. Then, using $2\sigma_k^2(f) = E_{(y,y')\sim\mu_k^2}\left(D_{y,y'}^k f\right)^2$, we get

$$\begin{aligned}
4D^2\left(\Sigma^2(f)\right) &= \sum_l \left(\sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left(D_{y,y'}^k f\right)^2 - S_{z_l}^l E_{(y,y')\sim\mu_k^2}\left(D_{y,y'}^k f\right)^2 \right)^2 \\
&= \sum_l \left(\sum_{k\neq l} E_{(y,y')\sim\mu_k^2}\left[\left(D_{y,y'}^k f\right)^2 - \left(D_{y,y'}^k S_{z_l}^l f\right)^2\right] \right)^2 \\
&= \sum_l \left(\sum_{k\neq l} E_{(y,y')\sim\mu_k^2}\left[\left(D_{y,y'}^k f - D_{y,y'}^k S_{z_l}^l f\right)\left(D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right)\right] \right)^2 \\
&\leq \sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k\left(f - S_{z_l}^l f\right)\right]^2 \times \\
&\qquad \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right]^2
\end{aligned}$$

by an application of Cauchy–Schwarz. Now, using $(a+b)^2 \leq 2a^2 + 2b^2$, we can bound the last sum independent of l by

$$\begin{aligned}
&\sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[2\left(D_{y,y'}^k f\right)^2 + 2\left(D_{y,y'}^k S_{z_l}^l f\right)^2\right] \\
&= 4\sum_{k:k\neq l} \sigma_k^2(f) + 4S_{z_l}^l \sum_{k:k\neq l} \sigma_k^2(f) \\
&\leq 4\left(\Sigma^2(f) + S_{z_l}^l \Sigma^2(f)\right) = 4\left(\Sigma^2(f) + \inf_{z\in\Omega_l} S_z^l \Sigma^2(f)\right) \leq 8\Sigma^2(f),
\end{aligned}$$

so that

$$\begin{aligned}
D^2\left(\Sigma^2(f)\right) &\leq 2\sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k\left(f - S_{z_l}^l f\right)\right]^2 \Sigma^2(f) \\
&\leq 4\sup_{\mathbf{x},\mathbf{z}\in\Omega} \sum_{k,l:k\neq l} \sigma_k^2\left(f - S_z^l f\right)(\mathbf{x})\, \Sigma^2(f) = J^2(f)\,\Sigma^2(f).
\end{aligned}$$

□

Now we can use decoupling to put these pieces together.

Proof of Theorem 40 By rescaling it suffices to prove the result for $b = 1$. We can also assume $J := J(f) > 0$. Let $0 < \gamma \leq \beta < 1/(1 + J/2)$ and set $\theta = \gamma/(J(1-\gamma))$. Then $\gamma^2/\left(2(1-\gamma)^2\right) < \theta < 2/J^2$. By the Lemma 41

$$\theta \mathrm{Ent}_f(\gamma) \leq \frac{\gamma^2}{2(1-\gamma)^2} E_{\gamma f}\left[\theta \Sigma^2(f)\right] \leq \frac{\gamma^2}{2(1-\gamma)^2}\left(\mathrm{Ent}_f(\gamma) + \ln E\left[e^{\theta \Sigma^2(f)}\right]\right),$$

where the second inequality follows from the decoupling inequality (24). Subtract $\gamma^2/\left(2(1-\gamma)^2\right)\mathrm{Ent}_f(\gamma)$ to get

$$\mathrm{Ent}_f(\gamma)\left(\theta - \frac{\gamma^2}{2(1-\gamma)^2}\right) \leq \frac{\gamma^2}{2(1-\gamma)^2} \ln E\left[e^{\theta \Sigma^2(f)}\right].$$

Since $\gamma^2/\left(2(1-\gamma)^2\right) < \theta$ this simplifies, using the value of θ, to

$$\mathrm{Ent}_f(\gamma) \leq \frac{\gamma J}{2(1-(1+J/2)\gamma)} \ln E\left[e^{\theta \Sigma^2(f)}\right]. \tag{28}$$

On the other hand $\theta < 2/J^2$, so by the self-boundedness of $\Sigma^2(f)$ (Lemma 42) and part (i) of Lemma 32 give

$$\ln E\left[e^{\theta \Sigma^2(f)}\right] \leq \frac{\theta}{1 - J^2\theta/2} E\left[\Sigma^2(f)\right] = \frac{\gamma/J}{1-(1+J/2)\gamma} E\left[\Sigma^2(f)\right]. \tag{29}$$

Combining (28) and (29) to get a bound on $S_f(\gamma)$ gives

$$\mathrm{Ent}_f(\gamma) \leq \frac{\gamma^2}{2(1-(1+J/2)\gamma)^2} E\left[\Sigma^2(f)\right]$$

and from Theorem 12 and Lemma 33

$$\begin{aligned}\Pr\{f - Ef > t\} &\leq \inf_{\beta \in (0, 1/(1+J/2))} \exp\left(\frac{E\left[\Sigma^2(f)\right]}{2} \frac{\beta^2}{1-(1+J/2)\beta} - \beta t\right) \\ &\leq \exp\left(\frac{-t^2}{2\left(E\left[\Sigma^2(f)\right] + (1+J/2)t\right)}\right).\end{aligned}$$

□

To use Theorem 40 one has to bound b and J. For the latter it is often sufficient to use the simple bound

$$J(f) \leq n \max_{k \neq l} \sup_{\mathbf{x} \in \Omega} \sup_{z,z',y,y' \in \Omega_l} D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}). \tag{30}$$

which can be obtained from Lemma 13.

We conclude with an application to U-statistics. Let $m < n$ be integers, $\Omega_i = \mathcal{X}$ and $\kappa : \mathcal{X}^m \to \mathbb{R}$ a symmetric kernel. For a subset of indices with cardinality m, $S = \{j_1, ..., j_m\} \subseteq \{1, ..., n\}$ define $\kappa_S : \mathcal{X}^n \to \mathbb{R}$ by $\kappa_S(\mathbf{x}) = \kappa\left(x_{j_1}, ..., x_{j_m}\right)$. The U-statistic of order m induced by κ is then the function $U : \mathcal{X}^n \to \mathbb{R}$ given by

$$U(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{S \subseteq \{1,\ldots,n\}} \kappa_S(\mathbf{x}).$$

U-statistics were introduced by Hoeffding [15]. Their importance stems from the fact that for iid $\mathbf{X} = (X_1, \ldots, X_n)$ the random variable $U(\mathbf{X})$ is an unbiased estimator for $E[\kappa(X_1, \ldots, X_m)]$. Starting with the work of Hoeffding there has been a lot of work on concentration inequalities for U-statistics. To simplify the presentation we will not use the advantage of Theorem 40 over Theorem 39 and assume the kernel κ to be bounded, $\kappa : \mathcal{X}^m \to [0, 1]$ for simplicity.

Notice that,if $k \notin S$, then $\kappa_S \in \mathcal{A}_k$, so $\kappa_S(\mathbf{x}) - E_k[\kappa_S(\mathbf{x})] = 0$ and thus

$$\begin{aligned} U(\mathbf{x}) - E_k[U(\mathbf{x})] &= \binom{n}{m}^{-1} \sum_{\substack{S \subseteq \{1,\ldots,n\} \\ k \in S}} (\kappa_S(\mathbf{x}) - E_k[\kappa_S(\mathbf{x})]) \\ &\leq \binom{n}{m}^{-1} |\{S \subseteq \{1, \ldots, n\} : k \in S\}| \\ &= \frac{\binom{n-1}{m-1}}{\binom{n}{m}}^{-1} = \frac{m!\,(n-1)!}{n!\,(m-1)!} = \frac{m}{n}, \end{aligned}$$

so we can set the quantity b in Theorem 40 to m/n. To bound J use (30) to get

$$\begin{aligned} J(U) &\leq n \max_{k \neq l} \sup_{\mathbf{x} \in \Omega} \sup_{z,z',y,y' \in \Omega_l} D^l_{z,z'} D^k_{y,y'} U(\mathbf{x}) \\ &\leq n \binom{n}{m}^{-1} \sum_{\substack{S \subseteq \{1,\ldots,n\} \\ k,l \in S : k \neq l}} D^l_{z,z'} D^k_{y,y'} \kappa_S(\mathbf{x}) \\ &= 2n \binom{n}{m}^{-1} |\{S \subseteq \{1, \ldots, n\} : k, l \in S, k \neq l\}| \\ &= \frac{2n\binom{n-2}{m-2}}{\binom{n}{m}} \leq \frac{2m^2}{n}. \end{aligned}$$

Substitution in Theorem 40 gives for $t > 0$

$$\Pr\{U - EU > t\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2(U)\right] + 2\left(m + m^2\right)t/n}\right).$$

It can be shown (see, e.g., [21], Houdré [17]) that in general $E\left[\Sigma^2(f)\right] \leq \sigma^2(f) + J^2(f)/4$, so that for U-statistics the Efron–Stein inequality is tight in the sense that $E\left[\Sigma^2(U)\right] \leq \sigma^2(U) + m^4/n^2$. It follows that for deviations $t > 1/n$

$$\Pr\{U - EU > t\} \leq \exp\left(\frac{-t^2}{2\sigma^2(U) + 2\left(m + 2m^2\right)t/n}\right).$$

This inequality can be compared to the classical work of Hoeffding [15] and more recent results of Arcones [2], which both consider undecoupled, nondegenerate U-statistics of arbitrary order. Hoeffding [15] does not have the correct variance term, while [2] gives the correct variance term but severely overestimates the subexponential coefficient in Bernstein's inequality to be exponential in the degree m of the U-statistic (above it is only of order m^2). This exponential dependence on m results from the use of the decoupling inequalities in [24] and seems to beset most works on U-statistics of higher order (e.g., [1, 13]), which in many other ways improve over our simple inequality above.

6 Appendix I. Table of Notation

General notation

$\Omega = \prod_{k=1}^{n} \Omega_k$	underlying (product-) probability space
$\mathcal{A}$	bounded measurable functions on Ω
$\mu = \otimes_{k=1}^{n} \mu_k$	(product-) probability measure on Ω
X_k	random variable distributed as μ_k in Ω_k
$f \in \mathcal{A}$	fixed function under investigation
$g \in \mathcal{A}$	generic function
$E[g] = \int_\Omega g d\mu$	expectation of g in μ
$\sigma^2[g] = E\left[(g - E[g])^2\right]$	variance of g in μ

Notation for the entropy method

$\beta = 1/T$	inverse temperature
$E_{\beta f}[g] = E\left[ge^{\beta f}\right]/E\left[e^{\beta f}\right]$	thermal expectation of g
$Z_{\beta f} = E\left[e^{\beta f}\right]$	partition function
$d\mu_{\beta f} = Z_{\beta f}^{-1} e^{\beta f} d\mu$	thermal measure (canonical ensemble)
$\mathrm{Ent}_f(\beta) = \beta E_{\beta f}[f] - \ln Z_{\beta f}.$	(canonical) entropy
$A_f(\beta) = \frac{1}{\beta} \ln Z_{\beta f}$	free energy
$\sigma_{\beta f}^2(g) = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right]$	thermal variance of g
$\psi(t) = e^t - t - 1$	
$S_y^k F(\mathbf{x}) = F\left(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n\right)$	substitution operator
$E_k[g](\mathbf{x}) = \int_{\Omega_k} S_y^k g \, d\mu_k(y)$	conditional expectation
$\mathcal{A}_k \subset \mathcal{A}$	functions independent of k-th variable
$Z_{k,\beta f} = E_k\left[e^{\beta f}\right]$	conditional partition function
$E_{k,\beta f}[g] = Z_{k,\beta f}^{-1} E_k\left[ge^{\beta f}\right]$	conditional thermal expectation
$\mathrm{Ent}_{k,f}(\beta) = \beta E_{k,\beta f}[g] - \ln Z_{k,\beta f}$	conditional entropy
$\sigma_{k,\beta f}^2[g] = E_{k,\beta f}\left[\left(g - E_{k,\beta f}[g]\right)^2\right]$	conditional thermal variance
$\sigma_k^2[g] = E_k\left[(g - E_k[g])^2\right]$	conditional variance

Operators on $\mathcal{A}$

$D^k_{y,y'} g = S^k_y g - S^k_{y'} g$	difference operator
$r_k(g) = \sup_{y,y' \in \Omega_k} D^k_{y,y'} f$	conditional range operator
$R^2(g) = \sum_k r_k^2(g)$	sum of conditional square ranges
$\Sigma^2(g) = \sum_k \sigma_k^2[g]$	sum of conditional variances
$(\inf_k g)(\mathbf{x}) = \inf_{y \in \Omega_k} S^k_y g(\mathbf{x})$	conditional infimum operator
$V^2_+ g = \sum_k E_{y \sim \mu_k} \left[\left(\left(g - S^k_y \right)_+ \right)^2 \right]$	Efron–Stein variance proxy
$D^2 g = \sum_k (g - \inf_k g)^2 .$	worst case variance proxy

References

1. Adamczak, R., et al.: Moment inequalities for u-statistics. Ann. Probab. **34**(6), 2288–2314 (2006)
2. Arcones, M.A.: A Bernstein-type inequality for u-statistics and u-processes. Stat. Probab. Lett. **22**(3), 239–247 (1995)
3. Bartlett, P., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
4. Bernstein, S.: Theory of Probability. Moscow (1927)
5. Boltzmann, L.: Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht. Kk Hof-und Staatsdruckerei (1877)
6. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities. Oxford University Press, Oxford (2013)
7. Boucheron, S., Lugosi, G., Massart, P., et al.: Concentration inequalities using the entropy method. Ann. Probab. **31**(3), 1583–1614 (2003)
8. Boucheron, S., Lugosi, G., Massart, P., et al.: On concentration of self-bounding functions. Electron. J. Probab. **14**, 1884–1899 (2009)
9. Bousquet, O.: A Bennett concentration inequality and its application to suprema of empirical processes. C.R. Math. **334**(6), 495–500 (2002)
10. Chatterjee, S.: Concentration inequalities with exchangeable pairs. Ph.D. thesis, Citeseer (2005)
11. Chebyshev, P.L.: Sur les valeurs limites des intégrales. Imprimerie de Gauthier-Villars (1874)
12. Gibbs, J.W.: Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics. C. Scribner's Sons, New York (1902)
13. Giné, E., Latała, R., Zinn, J.: Exponential and moment inequalities for u-statistics. High Dimensional Probability II, pp. 13–38. Springer, Berlin (2000)
14. Gross, L.: Logarithmic sobolev inequalities. Am. J. Math. **97**(4), 1061–1083 (1975)
15. Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat., pp. 293–325 (1948)
16. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 301 (1963)
17. Houdré, C.: The iterated jackknife estimate of variance. Statist. Probab. Lett. **35**(2), 197–201 (1997)
18. Ledoux, M.: The Concentration of Measure Phenomenon, vol. 89. American Mathematical Society, Providence (2001)
19. Lieb, E.H.: Some convexity and subadditivity properties of entropy. Inequalities, pp. 67–79. Springer, Berlin (2002)

20. Massart, P., et al.: About the constants in Talagrand's concentration inequalities for empirical processes. Ann. Probab. **28**(2), 863–884 (2000)
21. Maurer, A., et al.: A Bernstein-type inequality for functions of bounded interaction. Bernoulli **25**(2), 1451–1471 (2019)
22. McAllester, D., Ortiz, L.: Concentration inequalities for the missing mass and for histogram rule error. J. Mach. Learn. Res. **4**(Oct), 895–911 (2003)
23. McDiarmid, C.: Concentration. Probabilistic Methods of Algorithmic Discrete Mathematics, pp. 195–248. Springer, Berlin (1998)
24. de la Peña, V.H.: Decoupling and khintchine's inequalities for u-statistics. Ann. Probab., pp. 1877–1892 (1992)
25. Popper, K.R.: Logik der forschung (1934). The Logic of Scientific Discovery. [Google Scholar] (1968)
26. Steele, J.M.: An Efron-Stein inequality for nonsymmetric statistics. Ann. Probab., pp. 753–758 (1986)
27. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques **81**(1), 73–205 (1995)
28. Talagrand, M.: A new look at independence. Ann. Probab., pp. 1–34 (1996)
29. Vershynin, R.: High-dimensional Probability: An Introduction with Applications in Data Science, vol. 47. Cambridge University Press, Cambridge (2018)

Ill-Posed Problems: From Linear to Nonlinear and Beyond

Rima Alaifari

1 Introduction

In many applications in science and engineering, one seeks to recover an object from some acquired measurements. Such reconstruction problems are often vulnerable to small changes in the measured data, in the sense that slight perturbations in the data can lead to large deviations in the reconstructed objects. Such a property is of course highly unpleasant, as it makes the reconstruction unreliable. This observation has led to studying so-called *inverse problems* and to developing the theory of *regularization* that aims at introducing reconstruction algorithms that are sensitive to these instabilities and that aim at extracting information as stably as possible from such unstable systems.

For problems that can be modeled with a *linear* forward operator, the theory of *linear inverse problems* is quite developed. The unboundedness of the inverse (or generalized inverse) can be nicely characterized by the closedness of the range of the forward operator. When the forward operator has a singular value decomposition, the instabilities can often be quantified, as we show in the example of the truncated Hilbert transform (arising in limited data computerized tomography) in Sect. 3. Moreover, regularization theory is in many cases very effective for linear problems. It can give guarantees for convergence (and convergence rates) of algorithms, when they are suitably chosen so that regularization can be guaranteed. Regularization is, in principle, tied to assuming prior knowledge on the solution and, therefore, to searching for solutions only in a restricted set. This way, we also demonstrate how regularization can be guaranteed for the truncated Hilbert transform in Sect. 4.2.

When the underlying problem is nonlinear, one can still (to some extent) describe unboundedness of the inverse problem and suggest regularization methods. However, as the example of phase retrieval demonstrates (see Sect. 6), even if the inverse is

R. Alaifari (✉)
ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
e-mail: rima.alaifari@math.ethz.ch

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8_3

bounded, the nonlinear case is very different: if the inverse is not *uniformly* bounded, one will in practice still witness instabilities in the reconstruction. The problem of phase retrieval is very particular in that regularization is not effective for stable recovery from phaseless measurements. However, one can give up on the strong notion of unique reconstruction to achieve some sort of stable recovery that we call *atoll function reconstruction.*

The last problem we visit in Sect. 7 is the one of image classification with deep neural networks (DNNs). The network is obtained by training it on a set of test images for which the correct labels are provided. The DNN should then perform the task of correctly classifying objects in images it has not seen during training. For this problem, we do not even have a rigorous mathematical formulation of the forward operator at hand, at least not one that allows for a useful analysis. However, we witness instabilities in the form of so-called *adversarial examples.* One can construct algorithms that find, for most correctly classified images, a slightly perturbed version that the neural network will no longer classify correctly. While some remedies have been suggested in the form of altering the training process in form of *adversarial training,* these methods still do not give rise to robust classifiers, as one can think of other ways to "fool" the network, such as slightly deforming the images.

This chapter is organized as follows: in Sect. 2 we provide an introduction to linear inverse problems. In Sect. 3 we introduce the problem of limited data computerized tomography and describe the truncated Hilbert transform as the resulting linear forward operator. This includes the derivation of the SVD of the truncated Hilbert transform as well as the asymptotic behavior of its singular values. The theory of regularization of linear problems is the subject of Sect. 5, in which results on the regularization of the truncated Hilbert transform are given as well. Nonlinear problems are then discussed in Sect. 6, with Gabor phase retrieval as an example. Finally, we present the problem of image classification in Sect. 7.

2 Linear Inverse Problems

A study of ill-posed problems naturally begins with linear operators as the theory for the linear case is rather complete. We will therefore follow this path and refer to the classical book of Engl et al. [21] for further details. In Sect. 3, we will then analyze a specific linear problem using the theory of ill-posed problems combined with Sturm–Liouville theory and Wentzel–Kramers–Brillouin approximations. When the underlying operator is nonlinear, the theory is by far less straightforward. However, the main results for nonlinear inverse problems are inspired by the linear case.

In what follows, $T : X \to Y$ will always denote a bounded linear operator from X to Y, where X and Y are separable Hilbert spaces. The range and nullspace of an operator will be denoted by $\mathcal{R}$ and $\mathcal{N}$, respectively. We take T to be the forward operator of the inverse problem in question and follow the classical definition by Hadamard: a problem is *well-posed* if it satisfies the following criteria.

Definition 1 (*Hadamard's well-posedness*) The inversion of $T : X \to Y$ is said to be well-posed if the following properties hold:

- **Existence:** For all $g \in Y$, there exists $f \in X$ such that $Tf = g$, i.e.,

$$\mathcal{R}(T) = Y. \tag{1}$$

- **Uniqueness:** For all $g \in Y$, the solution is unique, i.e.,

$$\mathcal{N}(T) = \{0\}. \tag{2}$$

- **Stability:** The solution depends continuously on the data, i.e.,

$$T^{-1} \in \mathcal{L}(Y, X). \tag{3}$$

Here, $\mathcal{L}(Y, X)$ denotes the set of all bounded linear transforms mapping from Y to X.

If at least one of the three conditions is violated, the problem is said to be *ill-posed*.

Non-uniqueness can be undesirable, for instance, in reconstruction problems where one is interested in recovering a signal, an image, an electron density, etc. from some acquired measurements. If the same set of measurements can be created by two different objects, then this is problematic for the reconstruction problem. We will address the question of uniqueness in the reconstruction problems that will be discussed in Sects. 3 and 6. However, if the inverse problem in question is the identification of system parameters in a PDE so that a given state is reached, then having more than one candidate parameter set is of course less of an obstruction (though some notion of uniqueness might still be desirable for the numerical solution that involves an optimization problem).

The lack of stability creates serious numerical issues. Acquired measurements are never exact and can only be given up to a certain accuracy. No stable dependence of the solution on the data means that straightforward calculations of the reconstruction will, in principle, be unreliable, as they might be arbitrarily far from the ground truth. The theory of *regularization* aims at tackling this issue: the idea is to extract information as stably as possible, meaning that the guarantees that one will be able to give, will depend on the *noise level*, i.e., on the accuracy of the measurements.

In case the first property (existence) is violated, i.e., if $Tf = g$ is not solvable for a given right-hand side $g \notin \mathcal{R}(T)$, one can relax the notion of a solution. Instead of a classical solution, one can search for a *generalized solution.* To introduce this concept, we first start with ensuring existence by simply searching for the "best fit".

Definition 2 (*Least-squares solution*) An element $f \in X$ is called a least-squares solution for $Tf = g$ if

$$\|Tf - g\|_Y \leq \|Th - g\|_Y, \quad \forall h \in X.$$

To further characterize least-squares solutions, we let Q be the orthogonal projection of Y on $\overline{\mathcal{R}(T)}$, i.e.,

$$\langle Qg, u\rangle_Y = \langle g, u\rangle_Y, \quad \forall\, g \in Y,\ \forall\, u \in \overline{\mathcal{R}(T)}.$$

The following facts can then be stated.

$$\|Qg - g\|_Y \leq \|u - g\|_Y, \quad \forall\, u \in \overline{\mathcal{R}(T)}, \qquad \text{(minimality property)}$$

and

$$Qg - g \in \mathcal{R}(T)^{\perp}. \tag{4}$$

In addition, we also recall a standard result.

Theorem 3 *Let $T : X \to Y$ be a bounded linear operator between Hilbert spaces X and Y. Then,*

$$\mathcal{N}(T) = \mathcal{R}(T^*)^{\perp}, \tag{5}$$
$$\overline{\mathcal{R}(T)} = \mathcal{N}(T^*)^{\perp}. \tag{6}$$

The above properties enable us to make the following link between least-squares solutions, orthogonal projections, and solutions to the *normal equation.*

Theorem 4 *Let $g \in Y$, $f \in X$ and $T \in \mathcal{L}(X, Y)$. The following are equivalent:*

$$Tf = Qg, \tag{7}$$
$$\|Tf - g\|_Y \leq \|Th - g\|_Y, \quad \forall\, h \in X, \tag{8}$$
$$T^*Tf = T^*g. \tag{9}$$

Proof We show that (7) $\Rightarrow$ (8) $\Rightarrow$ (9) $\Rightarrow$ (7). The first implication is established by employing (4):

$$\|Th - g\|_Y^2 = \|Th - Qg\|_Y^2 + \|Qg - g\|_Y^2, \tag{10}$$
$$= \|Th - Qg\|_Y^2 + \|Tf - g\|_Y^2, \tag{11}$$
$$\geq \|Tf - g\|_Y^2. \tag{12}$$

The second implication can be established as follows: first, we note that since $Qg \in \overline{\mathcal{R}(T)}$, there exists a sequence $(f_n)_{n\in\mathbb{N}} \subset X$ s.t. $Tf_n \to Qg$ as $n \to \infty$. Hence, by assumption, Eq. (8) yields

$$\|Qg - g\|_Y^2 = \lim_{n\to\infty} \|Tf_n - g\|_Y^2 \geq \|Tf - g\|_Y^2.$$

Together with (4), this then implies

$$\begin{aligned}\|Tf - g\|_Y^2 &= \|Tf - Qg\|_Y^2 + \|Qg - g\|_Y^2, \\ &\geq \|Tf - Qg\|_Y^2 + \|Tf - g\|_Y^2.\end{aligned}$$

Thus, $Tf = Qg$. Moreover,

$$Tf - g = Qg - g \in \mathcal{R}(T)^\perp = \mathcal{N}(T^*).$$

We can thus conclude that

$$T^*(Tf - g) = 0.$$

Part (9) $\Rightarrow$ (7) can be seen as follows: The assumption that $Tf - g \in \mathcal{N}(T^*) = \mathcal{R}(T)^\perp$ implies

$$Q(Tf - g) = 0.$$

Consequently,

$$0 = Q(Tf - g) = QTf - Qg = Tf - Qg,$$

and, hence, $Tf = Qg$. □

The above theorem is extremely useful, as it states that least-squares solutions are nothing else than solutions to the normal equation $T^*Tf = T^*g$. Also, they are equivalent to solutions of $Tf = Qg$, i.e., the original equation with right-hand side projected on $\overline{\mathcal{R}(T)}$. With this at hand, one can make a statement about the existence of least-squares solutions:

Corollary 5 *Let $L(g)$ denote the set of least-squares solutions to right-hand side g, i.e.,*

$$L(g) := \{f \in X : T^*Tf = T^*g\}.$$

Then, the following hold:

(i) $L(g)$ is nonempty if and only if $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$.
(ii) If $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$, then $L(g)$ is a nonempty, closed, and convex subset of X.

Proof Part (i): First, if $f \in L(g)$, then $T^*Tf = T^*g$, and hence

$$Tf - g \in \mathcal{N}(T^*) = \mathcal{R}(T)^\perp.$$

Decomposing g as $g = Tf + (g - Tf)$, it then follows that $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$. To show the other direction, we remark that for $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$, a splitting $g = g_1 + g_2$ with $g_1 \in \mathcal{R}(T)$, $g_2 \in \mathcal{R}(T)^\perp$ is unique with the only option for g_1 to be $g_1 = Qg$. The property that $g_1 \in \mathcal{R}(T)$ means the existence of an element $f \in X$ such that $g_1 = Tf$, and hence

$$Qg = Tf.$$

By Theorem 4, we can thus conclude that this element f is a least-squares solution, i.e., $f \in L(g)$.

Part (ii): Nonemptiness is already established in Part (i). To show convexity, we proceed in the standard way of considering $f, f' \in L(g)$ and

$$h := tf' + (1-t)f$$

for arbitrary $t \in [0, 1]$. By Theorem 4, we have

$$T^*Tf = T^*g,$$
$$T^*Tf' = T^*g.$$

Linearity of T then yields

$$T^*Th = tT^*Tf' + (1-t)T^*Tf = tT^*g + (1-t)T^*g = T^*g,$$

i.e., $h \in L(g)$.

To show that $L(g)$ is closed, we consider a sequence $(f_n)_{n\in\mathbb{N}} \subset L(g)$ with $f_n \to f$ as $n \to \infty$. Closure of X implies $f \in X$. Moreover, since each f_n is in $L(g)$,

$$\|Tf - g\|_Y = \lim_{n\to\infty} \|Tf_n - g\|_Y \leq \|Th - g\|_Y, \quad \forall\, h \in X.$$

Thus, $f \in L(g)$. □

By its definition, the least-squares solution is not necessarily unique. For if $\mathcal{N}(T) \neq \{0\}$ and $\overline{f}$ is a least-squares solution, then also any $\overline{f} + f_0$ with $f_0 \in \mathcal{N}(T)$ is a least-squares solution as well. This implies that if $\mathcal{N}(T) \neq \{0\}$, then there are infinitely many least-squares solutions. We will later characterize the set of least-squares solutions. For now, we are interested in a notion of solution that enjoys uniqueness. One popular choice is to pick the least-squares solution with minimal norm and this is sometimes referred to as *best approximate solution.*

Definition 6 (*Best approximate solution*) An element $f \in X$ is called best approximate solution if it is a least-squares solution, i.e., $f \in L(g)$, and it has minimal norm, i.e.,

$$\|f\|_X = \inf\{\|h\|_X : h \in L(g)\}.$$

Corollary 7 *If $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$, then the best approximate solution to $Tf = g$ is unique.*

This is a direct consequence of the convexity of $L(g)$ stated in Corollary 5. Note also that $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$ is the only interesting case, since otherwise $L(g)$ is empty.

2.1 The Moore–Penrose Generalized Inverse

So far, the previous discussion has been centered around restoring the notions of existence and uniqueness and the best approximate solution, or minimum-norm least-squares solution, is a candidate that fulfills these requirements. We intend to continue this discussion in an operator-theoretic way. To do so requires the introduction of the *Moore–Penrose generalized inverse*, which is the map from $g \in \mathcal{R}(T) \oplus \mathcal{R}(T)^{\perp}$ to the unique best approximate solution, i.e., to the element in $L(g)$ with minimal norm. One way to define the Moore–Penrose generalized inverse $T^{\dagger}$ of T is to introduce the invertible restriction $\widetilde{T} := T|_{\mathcal{N}(T)^{\perp}}$ of T.

Definition 8 (*Moore–Penrose generalized inverse*) For $T \in \mathcal{L}(X, Y)$, its Moore–Penrose generalized inverse is defined as the unique linear extension of $\widetilde{T}^{-1}$ to

$$\mathcal{D}(T^{\dagger}) := \mathcal{R}(T) \oplus \mathcal{R}(T)^{\perp}$$

with

$$\mathcal{N}(T^{\dagger}) = \mathcal{R}(T)^{\perp}.$$

Thus, the Moore–Penrose generalized inverse maps each $g \in \mathcal{D}(T^{\dagger})$ to $f \in L(g)$ with minimal norm, $f := T^{\dagger}g$. We record a few basic properties of $T^{\dagger}$.

Corollary 9 *Let $T^{\dagger}$ be the Moore–Penrose generalized inverse of $T \in \mathcal{L}(X, Y)$. Then,*

(i) the domain $\mathcal{D}(T^{\dagger})$ is dense in Y. Moreover, $\mathcal{R}(T)$ is closed if and only if $\mathcal{D}(T^{\dagger}) = Y$.

(ii) If $\mathcal{R}(T)$ is closed and T^{-1} exists, it follows that

$$T^{\dagger}|_{\mathcal{R}(T)} = T^{-1}.$$

(iii) $\mathcal{R}(T^{\dagger}) = \mathcal{N}(T)^{\perp}$ $(= \overline{\mathcal{R}(T^{*})})$.

(iv) $T^{\dagger}$ is linear.

(v) For $g \in \mathcal{D}(T^{\dagger})$, $T^{\dagger}g$ is the unique element in $L(g) \cap \mathcal{N}(T)^{\perp}$.

Proof Except for item (iii), the statements are straightforward consequences of the definition of $T^{\dagger}$ and the results we have collected so far about least-squares solutions. To show item (iii), let $h \in \mathcal{R}(T^{\dagger})$ so that there exists an element $g \in \mathcal{D}(T^{\dagger})$ with $T^{\dagger}g = h$. By definition, g can be decomposed as $g = g_1 + g_2$ with $g_1 \in \mathcal{R}(T)$ and $g_2 \in \mathcal{R}(T)^{\perp} = \mathcal{N}(T^{\dagger})$. Thus, by linearity of $T^{\dagger}$,

$$h = T^{\dagger}g = T^{\dagger}g_1 + T^{\dagger}g_2 = T^{\dagger}g_1 = \widetilde{T}^{-1}g_1.$$

Hence, $h \in \mathcal{R}(\widetilde{T}^{-1}) = \mathcal{N}(T)^{\perp}$ so that overall, $\mathcal{R}(T^{\dagger}) \subseteq \mathcal{N}(T)^{\perp}$. On the other hand, suppose that $h \in \mathcal{N}(T)^{\perp}$. Then, by definition of $\widetilde{T}$, $\widetilde{T}h = Th$. This implies

$$h = \widetilde{T}^{-1} T h = T^{\dagger} T h,$$

so that $h \in \mathcal{R}(T^{\dagger})$ and hence $\mathcal{N}(T)^{\perp} \subseteq \mathcal{R}(T^{\dagger})$. □

The first item in the above corollary states that $\mathcal{D}(T^{\dagger})$ is dense in Y. Note also, that by item (i) in Corollary 5, $L(g)$ is empty when $g \notin \mathcal{D}(T^{\dagger})$, so that in this case no least-squares solution exists.

We emphasize that in the discussion about generalized solutions and inverses we have not addressed the question of stability at all. In fact, while the Moore–Penrose generalized inverse restores uniqueness, it does not restore stability of the inversion problem. Boundedness (and hence continuity) of $T^{\dagger}$ can, however, be characterized as follows:

Theorem 10 *Let $T^{\dagger}$ be the Moore–Penrose generalized inverse of $T \in \mathcal{L}(X, Y)$. Then, $T^{\dagger}$ is bounded if and only if $\mathcal{R}(T)$ is closed.*

Proof The proof is based on establishing that the graph of $T^{\dagger}$ is closed. See, e.g., Proposition 2.4 and its proof in [21] for the full argument. □

Remark 11 Since the closedness of $\mathcal{R}(T)$ is equivalent to $\mathcal{D}(T^{\dagger}) = Y$ (cf. Corollary 9, item (i)), one further has

$$\mathcal{D}(T^{\dagger}) = Y \iff T^{\dagger} \text{ is bounded.}$$

In terms of Hadamard's well-posedness criteria, this amounts to equivalence of existence and stability in the least-squares solution sense.

2.2 Compact Operators

In inverse problems, an important class of bounded linear transforms $T \in \mathcal{L}(X, Y)$ is compact operators. On one hand, they appear frequently, e.g., integral operators are typically compact under some suitable assumptions on the kernel. On the other hand, they are inherently ill-posed, as we will see in this section.

Definition 12 (*Compact operator*) Let $K : X \to Y$ be a linear operator. Then, K is said to be compact if for any bounded set $B \subset X$, the closure of its image $\overline{K(B)}$ is compact.

Alternatively, compact operators can be characterized as follows.

Theorem 13 *A linear operator $K : X \to Y$ is compact if and only if for any bounded sequence $(f_n)_{n\in\mathbb{N}} \subset X$, the sequence $(Kf_n)_{n\in\mathbb{N}}$ has a convergent subsequence.*

We remark that it is easily shown that compact operators are always bounded, i.e., $K \in \mathcal{L}(X, Y)$. Some useful properties of compact operators are the following.

Lemma 14 *Let $K_1 \in \mathcal{L}(X, Y)$ and $K_2 \in \mathcal{L}(Y, Z)$. If at least one of the two operators is compact, then $K_1 K_2$ is also compact.*

Lemma 15 (Identity operator) *The identity operator* $\mathrm{id}: X \to X$ is compact if and only if X is finite dimensional.

The range of a compact operator has the following special property.

Theorem 16 (Range of compact operators) *Let $K \in \mathcal{L}(X, Y)$ be a compact operator. Then, its range $\mathcal{R}(K)$ is closed if and only if it is finite dimensional.*

The proof can be found in Proposition 2.7 of [21] but we still provide it for the sake of completeness and illustration:

Proof Clearly, if $\mathcal{R}(K)$ is finite dimensional, then it is closed. For the other direction, note that the operator

$$\widetilde{K} := K|_{\mathcal{N}(K)^\perp} : \mathcal{N}(K)^\perp \to \mathcal{R}(K)$$

is bijective, linear, and compact. If $\mathcal{R}(K)$ is closed, then the inverse mapping theorem implies that $\widetilde{K}^{-1}$ is bounded. Hence, by Lemma 14, also $\widetilde{K}\widetilde{K}^{-1} = \mathrm{Id}|_{\mathcal{R}(K)}$ is compact. Thus, by Lemma 15, $\mathcal{R}(K)$ is finite dimensional. □

In view of Theorem 10, the above result implies that for compact operators, the Moore–Penrose generalized inverse $K^\dagger$ is unbounded when the range of K is infinite dimensional:

Corollary 17 *Let $K \in \mathcal{L}(X, Y)$ be compact and let $K^\dagger$ be its generalized inverse. Then, $K^\dagger$ is bounded if and only if $\mathcal{R}(K)$ is finite dimensional.*

This property of inherent instability for the inversion of compact operators can be seen more explicitly when the singular value decomposition (SVD) is utilized. Since it will be useful in Sect. 3, we will start with a reminder on the spectrum of an operator $T \in \mathcal{L}(X)$ (cf. Sect. VI.3 in [42]).

Definition 18 (*Resolvent set and spectrum*) For $T \in \mathcal{L}(X)$, the resolvent set $\rho(T)$ is defined as

$$\rho(T) := \{\lambda \in \mathbb{C} : \lambda\mathrm{Id} - T \text{ is bijective, } R_\lambda(T) := (\lambda\mathrm{Id} - T)^{-1} \text{ is bounded}\}.$$

The operator $R_\lambda(T)$ is called the resolvent of T at λ. The spectrum $\sigma(T)$ of T is defined as $\sigma(T) := \mathbb{C}\backslash\rho(T)$.

In fact, the boundedness of the inverse does not have to be asked for explicitly: if $\lambda\mathrm{Id} - T$ is bijective, then by the inverse mapping theorem its inverse is bounded (note that $\lambda\mathrm{Id} - T$ is linear). For compact operators $K : X \to X$, the spectrum can be rather elegantly described as the following properties hold (cf. Theorems VI.15, VI.16, and VI.17 in [42]):

Theorem 19 (Riesz–Schauder theorem) *The spectrum $\sigma(K)$ of a compact operator $K \in \mathcal{L}(X)$ is at most countable with 0 as the only possible accumulation point. Every $\lambda \in \sigma(K)\backslash\{0\}$ is an eigenvalue of K.*

Remark 20 If X is infinite dimensional, then $0 \in \sigma(K)$. For if X is infinite dimensional, then K being bijective implies $\mathcal{R}(K) = X$ and hence $\mathcal{R}(K)$ is closed. From this, however, and Theorem 16 we can deduce that $\mathcal{R}(K)$ and, hence X, is finite dimensional, which is a contradiction. Therefore, K cannot be a bijection, and thus $0 \in \sigma(K)$.

Theorem 21 (Hilbert–Schmidt theorem) *Let $K : X \to X$ be a compact, self-adjoint operator. Then, there exist a sequence of nonzero eigenvalues $(\lambda_n)_{n=1}^N \subset \mathbb{R}$ and a sequence $(u_n)_{n=1}^N \subset X$ such that*

$$Ku_n = \lambda_n u_n,$$

where $(u_n)_{n=1}^N$ is an orthonormal basis of $\overline{\mathcal{R}(K)}$ and $N = rank(K)$. If $N = \infty$, then $\lambda_n \to 0$ as $n \to \infty$.

Theorem 22 *Let $K : X \to Y$ be a compact operator. Then, there exist sequences $(\sigma_n)_{n=1}^N \subset \mathbb{R}_+$, $(u_n)_{n=1}^N \subset X$ and $(v_n)_{n=1}^N \subset Y$, with $N \in \mathbb{N} \cup \infty$, such that*

$$\begin{aligned} Ku_n &= \sigma_n v_n, \\ K^* v_n &= \sigma_n u_n. \end{aligned}$$

The system $(u_n)_{n=1}^N$ is an orthonormal basis of $\overline{\mathcal{R}(K^*)} = \mathcal{N}(K)^\perp$, and $(v_n)_{n=1}^N$ is an orthonormal basis of $\overline{\mathcal{R}(K)}$. If $N = \infty$, then $\sigma_n \to 0$ as $n \to \infty$. The operator K takes on the representation

$$K = \sum_{n=1}^{N} \sigma_n \langle \cdot, u_n \rangle_X v_n.$$

The numbers $\sigma_n > 0$ are called singular values *of K and the system (σ_n, u_n, v_n) is said to form the* singular value decomposition (SVD) *of K.*

As we have seen before, the generalized solution may not always exist. More precisely, it only exists when $g \in \mathcal{D}(K^\dagger)$ (cf. Corollary 5). For compact operators, the existence of the generalized solution can be characterized with the *Picard criterion* (we will always assume $N = \infty$ in what follows, as this is the interesting case in which the generalized inverse is unbounded (cf. Corollary 17)).

Theorem 23 (Picard criterion). *Let $K \in \mathcal{L}(X, Y)$ be a compact operator. Then,*

$$g \in \mathcal{D}(K^\dagger) \iff \sum_{n=1}^{\infty} \frac{|\langle g, v_n \rangle_Y|^2}{\sigma_n^2} < \infty. \tag{13}$$

Proof • "$\Longrightarrow$": Since $g \in \mathcal{D}(K^\dagger)$, we can decompose it into $g = g_1 + g_2$ with $g_1 = Kf_1 \in \mathcal{R}(K)$, $g_2 \in \mathcal{R}(K)^\perp$, for some $f_1 \in X$. Then, noting that $v_n \in \mathcal{R}(K)$, one can deduce

$$\langle g, v_n\rangle_Y = \langle g_1 + g_2, v_n\rangle_Y = \langle g_1, v_n\rangle_Y = \langle f_1, K^* v_n\rangle_X = \sigma_n \langle f_1, u_n\rangle_X.$$

Thus,

$$\sum_{n=1}^{\infty} \frac{|\langle g, v_n\rangle_Y|^2}{\sigma_n^2} = \sum_{n=1}^{\infty} |\langle f_1, u_n\rangle_X|^2 \leq \|f_1\|_X^2 < \infty,$$

where we have used that $(u_n)_{n\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{N}(K)^\perp$, a closed subspace of X.

• "$\Longleftarrow$": By assumption, the sequence $\left(\sigma_n^{-1}\langle g, v_n\rangle_Y\right)_{n\in\mathbb{N}}$ is in $\ell^2(\mathbb{N})$. Thus, the Riesz–Fischer theorem yields that f defined as

$$f := \sum_{n=1}^{\infty} \sigma_n^{-1}\langle g, v_n\rangle_Y u_n$$

is an element in X.

The orthogonal projection Q onto $\overline{\mathcal{R}(K)}$ can be expressed with the orthonormal basis $(v_n)_{n\in\mathbb{N}}$ of $\overline{\mathcal{R}(K)}$:

$$Q = \sum_{n=1}^{\infty} \langle \cdot, v_n\rangle_Y v_n.$$

With this, one can deduce

$$Kf = \sum_{n=1}^{\infty} \sigma_n^{-1}\langle g, v_n\rangle_Y K u_n = \sum_{n=1}^{\infty} \langle g, v_n\rangle_Y v_n = Qg.$$

Thus, $Qg \in \mathcal{R}(K)$, and hence $g \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp = \mathcal{D}(K^\dagger)$.

□

With the singular value decomposition of K at hand, one can even give an explicit representation for the Moore–Penrose generalized inverse $K^\dagger$.

Theorem 24 *Let $K \in \mathcal{L}(X, Y)$ be compact with SVD $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$. Then, the operator $K^\dagger : \mathcal{D}(K^\dagger) \to \mathcal{N}(K)^\perp$ can be represented as*

$$K^\dagger = \sum_{n=1}^{\infty} \frac{\langle \cdot, v_n\rangle_Y}{\sigma_n} u_n.$$

Proof Let $g \in \mathcal{D}(K^\dagger)$. Then, the Picard criterion holds and

$$f := \sum_{n=1}^{\infty} \sigma_n^{-1} \langle g, v_n \rangle_Y u_n \in X,$$

with $Kf = Qg$ as derived in the previous proof. Thus, by Theorem 4, $f \in L(g)$. In view of item 5 of Corollary 9, it remains to show that $f \in \mathcal{N}(K)^{\perp}$. This follows immediately from the above definition of f and $(u_n)_{n\in\mathbb{N}}$ spanning $\mathcal{N}(K)^{\perp}$. □

Remark 25 (i) For compact operators, the fact that a generalized solution does not always exist is characterized by Picard's criterion: it only exists if $(\langle g, v_n \rangle_Y / \sigma_n)_{n\in\mathbb{N}}$ decays fast enough. Note that here $\sigma_n \to 0$ as $n \to \infty$. Thus, Theorem 23 links the existence of the generalized solution to the decay of the coefficients $\langle g, v_n \rangle_Y$ with respect to the singular values σ_n. The faster the singular values decay, the more rapid the decay of the coefficients of g has to be in order for a solution to exist.

(ii) The Picard criterion also reveals that while error components corresponding to large singular values σ_n are harmless, error components corresponding to small σ_n get amplified and cause severe numerical issues.

(iii) Typically, the ill-posedness of compact operator equations is characterized by the decay rate of $(\sigma_n)_{n\in\mathbb{N}}$. A problem is said to be *mildly ill-posed* if $(\sigma_n)_{n\in\mathbb{N}}$ decays algebraically.
When the singular values decay exponentially, the problem is said to be *severely ill-posed*.

Example 26 We conclude the treatment of compact operators by featuring a concrete example that we borrow from Sect. 1.5 in [21]. The *backward heat equation* can be described as assuming a final temperature at time $T = 1$:

$$h(x) := u(x, 1), \quad x \in [0, \pi], \; h(0) = h(\pi) = 0,$$

where the temperature $u(x, t)$ at position $x \in [0, \pi]$ and time $t \geq 0$ is governed by

$$\frac{\partial u}{\partial t}(x, t) = \frac{\partial^2 u}{\partial x^2}(x, t),$$

with homogenous Dirichlet boundary conditions

$$u(0, t) = u(\pi, t) = 0, \quad t \geq 0.$$

For the backward heat equation, we then aim at determining the initial temperature $u(x, 0)$, $x \in [0, \pi]$. One can find that the corresponding forward operator is an integral operator of the first kind:

$$h(x) = \frac{2}{\pi} \sum_{n=1}^{\infty} \int_0^{\pi} u(\tau, 0) \sin(n\tau) d\tau e^{-n^2} \sin(nx).$$

With the kernel of the integral operator thus given explicitly, one can derive the SVD of the operator which is $(e^{-n^2}, \sqrt{2/\pi}\sin(nx), \sqrt{2/\pi}\sin(nx))_{n\in\mathbb{N}}$. Thus, the inverse problem is severely ill-posed (since $\sigma_n = e^{-n^2}$).

We further note that the singular functions $(\sqrt{2/\pi}\sin(nx))_{n\in\mathbb{N}}$ form an orthonormal basis of $L^2([0,\pi])$ so that $\mathcal{R}(K)$ is dense in $L^2([0,\pi])$ and $\mathcal{D}(K^\dagger) = \mathcal{R}(K)$. Applying the Picard criterion then yields that the backward heat equation is (uniquely) solvable if and only if

$$\sum_{n=1}^{\infty} e^{2n^2}|h_n|^2 < \infty,$$

where $h_n := \sqrt{2/\pi}\int_0^\pi f(\tau)\sin(n\tau)d\tau$. In other words, a solution exists if and only if the Fourier coefficients of the final temperature h decay rapidly (much faster than e^{-n^2}).

2.3 General Bounded Linear Transforms

To conclude this section, we complete the discussion by noting that more generally than for compact operators, the spectrum of T^*T reveals the stability properties of the inverse problem with forward operator $T \in \mathcal{L}(X,Y)$. A *self-adjoint* bounded linear operator $A : X \to X$ is completely characterized by the spectral theorem (see Chap. VII in [42] for a detailed treatment). It can be derived by drawing on the functional calculus for continuous functions f which allows to meaningfully define the operator $f(A) \in \mathcal{L}(X)$ (see Theorem VII.1 in [42]). With the continuous functional calculus introduced, one can deduce that for $\psi \in X$, $\langle \psi, f(A)\psi\rangle_X$ is a continuous linear functional on $C(\sigma(A))$, the set of continuous functions on the spectrum $\sigma(A)$. The Riesz–Markov theorem further implies that there is a unique measure μ_ψ on the compact set $\sigma(A)$ such that

$$\langle \psi, f(A)\psi\rangle_X = \int_{\sigma(A)} f(\lambda)d\mu_\psi. \tag{14}$$

The measure μ_ψ is said to be the *spectral measure associated with* ψ. Introducing spectral measures leads to a natural extension of the functional calculus to general bounded Borel functions: defining $f(A)$ for f a bounded Borel function on $\mathbb{R}$ such that (14) holds, the polarization identity allows for recovery of $\langle \psi, f(A)\phi\rangle$, $\phi \in X$, and hence for $f(A)$ by utilizing the Riesz lemma.

By generalizing to bounded Borel functions, one can take characteristic functions χ_Ω of Borel sets Ω and define *spectral projections* of A as operators

$$P_\Omega := \chi_\Omega(A).$$

Lemma 27 *Let $A \in \mathcal{L}(X)$ be self-adjoint and let (P_Ω) be its family of spectral projections. Then, the following hold:*

(i) *each spectral projection P_Ω is an orthogonal projection;*
(ii) $P_\emptyset = 0$*;*
(iii) *there exists $a > 0$ s.t. $P_{(-a,a)} = \mathrm{id}$;* and
(iv) for any sequence of pairwise disjoint bounded Borel sets $(\Omega_n)_{n\in\mathbb{N}}$ and $\Omega := \bigcup_{n=1}^{\infty} \Omega_n$,

$$P_\Omega = \lim_{N\to\infty} \left(\sum_{n=1}^{N} P_{\Omega_n} \right).$$

Here, as before, Id denotes the identity operator on X. A family of projections fulfilling conditions (i) to (iv) is called a *projection-valued measure*. With this concept at hand, we can state the spectral theorem for bounded self-adjoint operators in its *projection-valued measure form* (cf. Theorem VII.8 in [42]).

Theorem 28 (Spectral theorem). *There is a one-to-one correspondence between bounded linear self-adjoint operators and bounded projection-valued measures:*

$$A \longmapsto (P_\Omega) = (\chi_\Omega(A)),$$
$$(P_\Omega) \longmapsto A = \int_{\sigma(A)} \lambda d P_\lambda.$$

In the next section, we introduce and study an inverse problem with non-compact forward operator and derive its spectral properties to understand the sources of ill-posedness.

3 Limited Data Computerized Tomography

In the remainder of this chapter, we will treat three different ill-posed problems. In this section, we start with the first, which is also the most classical of them: it can be modeled by a linear forward operator and we will show how the spectral properties can be derived and used to prove that certain reconstruction algorithms are *regularization methods*, a concept we introduce in Sect. 4.

The linear inverse problem we aim to discuss stems from medical imaging, more precisely from computerized tomography (CT). In the classical two-dimensional (2D) setup, *full data* is acquired, which means that there is a moving X-ray source that rotates 180° ("short scan") or 360° ("long scan") around the 2D region (i.e., a cross-section of a three-dimensional object). In such a scan, the rotating source shoots X-rays from different directions in a sufficiently dense scanning scheme at the object of interest. The attenuation of the X-ray beams is then recorded on the other side of the object by an array of detectors. The region which is covered by a full angular range (i.e., at least 180°) is called the field-of-view (FOV), see Fig. 1.

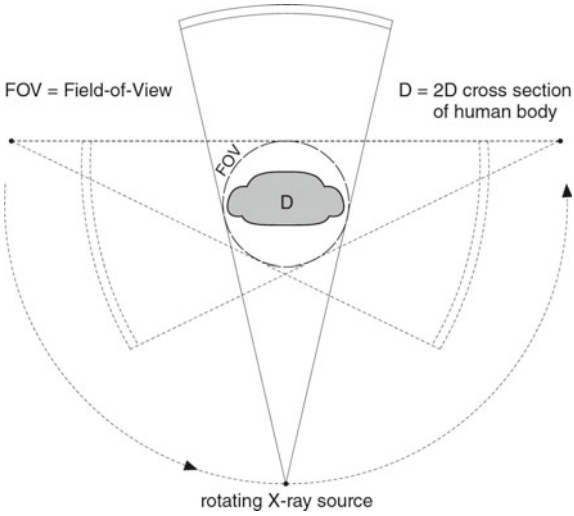


Fig. 1 Classical 2D CT: the field-of-view fully covers the object support D

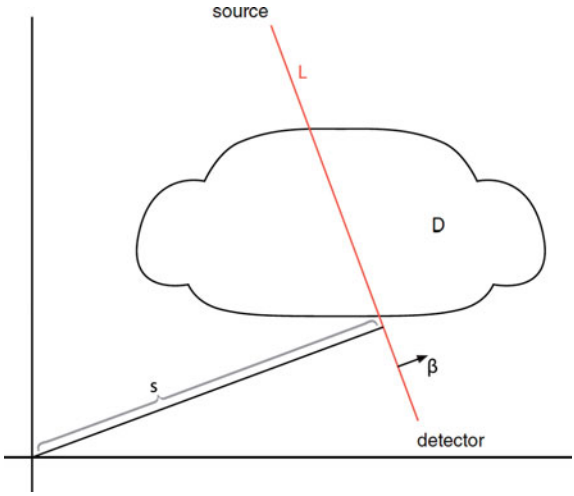


Fig. 2 Line integrals of the 2D cross-section

The attenuation of the X-ray beams, i.e., their intensity loss as they travel through the object, can reveal the structure of the object density. To make this more precise, we introduce the object support $D \subset \mathbb{R}^2$, the object density $f_D \in L^2(D)$ and parametrize a line L on which an X-ray beam travels by its distance $s \geq 0$ from the origin and its orientation $\theta \in S^1$, see Fig. 2. The attenuation can be modeled with the Beer–Lambert law of optics so that the CT scan essentially yields line integrals of f:

$$\int_{\mathbb{R}} f_D(s\theta + t\theta^{\perp})dt = -\ln \frac{I_L(\theta, s)}{I_0(\theta, s)},$$

where $I_0(\theta, s)$ and $I_L(\theta, s)$ denote the intensity of the traveling beam as it exits the source and as it arrives at the detector, respectively. The mapping from functions to their line integrals is well known as the *Radon transform.* Thus, the measurements collected in 2D CT can be modeled as the Radon transform of f_D:

$$(Rf_D)(\theta, s) = \int_{\mathbb{R}} f_D(s\theta + t\theta^{\perp})dt, \quad s \in \mathbb{R}, \theta \in S^1.$$

Reconstruction of f_D from 2D CT data hence amounts to inversion of the Radon transform. In light of the preceding discussion in Sect. 1, a natural question that arises in order to understand this linear inverse problem is whether we can determine the spectral properties of R^*R. This has been done in [39] and we briefly summarize the findings therein. First of all, one can assume w.l.o.g. that $D \subseteq B_2$, where B_2 denotes the unit disk in $\mathbb{R}^2$ and consider the Radon transform as a mapping from functions in $L^2(B_2)$ into a certain weighted L^2-space. Then, this operator is continuous (cf. Chap. 2 in [39]).

Theorem 29 (Continuity of the Radon transform) *Let R be defined as the Radon transform with the following mapping:*

$$R : L^2(B_2) \to L^2(S^1 \times [-1, 1], (1 - s^2)^{-1/2}),$$

where $L^2(S^1 \times [-1, 1], (1 - s^2)^{-1/2})$ is the weighted L^2-space with weight $(1 - s^2)^{-1/2}$ on $[-1, 1]$. Then, R is continuous.

This operator admits a singular value decomposition (cf. Chap. 4 in [39]) with singular values decaying at a rate $\sigma_n \sim 1/\sqrt{n}$. In light of the discussion in the previous section, we can conclude that the reconstruction problem from full Radon transform data is only mildly ill-posed. This also implies that CT reconstruction can be rather easily regularized using standard methods such as *filtered back-projection (FBP),* see, e.g., Sect. V.1 in [39]. Radon inversion becomes much more delicate when no longer full Radon transform data are available, in which case one speaks of a *limited data* problem. We present one such instance in the following.

3.1 Truncated Projections

In this limited data CT problem, one has full angular coverage of the X-ray beams in only a subregion of the support D of the 2D object, see Fig. 3 for an illustration. In this case, one can only hope at recovering the object on the intersection of D with the field-of-view. We will refer to this intersection as the *region-of-interest (ROI).*

In such a setup, the FBP leads to only very poor reconstructions. Instead of performing Radon inversion in two steps (filtering and back-projection), an equivalent formulation can be given involving three steps:

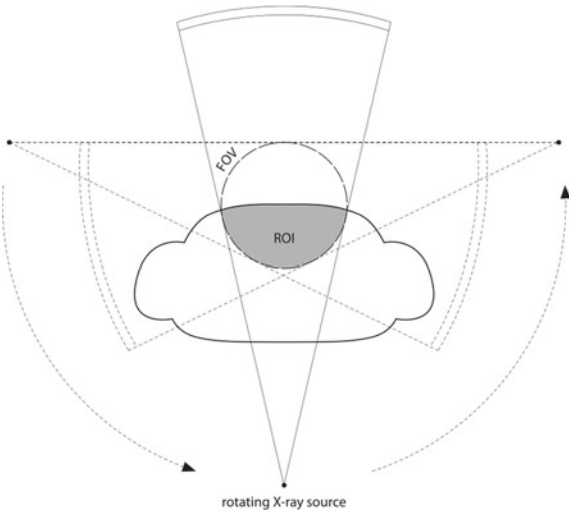


Fig. 3 Limited data 2D CT: the field-of-view does not cover the object support D

Differentiated back-projection (DBP):

(i) Differentiation:

$$r_D(\phi, s) = \frac{\partial}{\partial s}(Rf_D)(\theta, s), \quad \theta^\perp = (\cos\phi, \sin\phi).$$

(ii) Back-projection:

$$b_{\phi_1,\phi_2}(x) = \frac{1}{\pi}\int_{\phi_1}^{\phi_2} r_D(\phi, s)|_{s=x\cdot\theta}\, d\phi.$$

It can be calculated that

$$b_{\phi_1,\phi_2}(x) = (H_{\theta_2^\perp} f_D)(x) - (H_{\theta_1^\perp} f_D)(x),$$

where $H_\theta f_D$ denotes the Hilbert transform of f_D along the line through x with direction θ.

(iii) Hilbert transform inversion: The choice $\phi_2 = \phi_1 + \pi$ yields $\theta_2^\perp = -\theta_1^\perp$ and hence

$$b_{\phi_1,\phi_1+\pi}(x) = 2(H_{\theta_2^\perp} f_D)(x).$$

Thus, the inversion of $H_{\theta_2^\perp}$ recovers f_D on a line, so that the reconstruction of f_D can be obtained by solving a family of one-dimensional (1D) problems.

The first two steps of DBP pose no problem in the case of truncated projections: differentiation is a local process and back-projection is angular averaging, which can be done in the region-of-interest because one has full angular coverage therein.

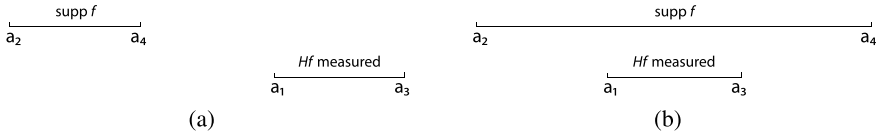
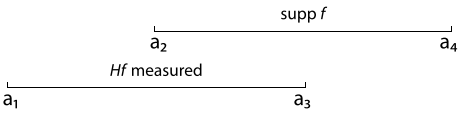


Fig. 4 The truncated Hilbert transform with a gap (**a**) and the interior problem (**b**)

Fig. 5 The truncated Hilbert transform with overlap



The question of inverting the Hilbert transform is more delicate and requires careful analysis.

For the remainder of this section, let a_1, a_2, a_3, a_4 be positive real numbers such that $a_1 < a_3$ and $a_2 < a_4$. We will denote a 1D slice of the object density f_D by f and assume that $f \in L^2([a_2, a_4])$, i.e., $\text{supp} f \subseteq [a_2, a_4]$. Furthermore, the Hilbert transform $H : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is defined by the principal value integral

$$(Hf)(x) = \frac{1}{\pi} \text{p.v.} \int_{\mathbb{R}} \frac{f(y)}{y - x} dy.$$

For any $f \in L^2(\mathbb{R})$, the inversion of Hilbert transform data is simple, if Hf is measured *on all of* $\mathbb{R}$: The inverse of the Hilbert transform is just $H^{-1} = -H$, so that f can be recovered via $f = -HHf$.

In the case of truncated projections, each Hilbert transform of a slice f is only known on a finite interval, which we denote by $[a_1, a_3] \subset \mathbb{R}$. For an interval $I \subset \mathbb{R}$, let $\mathcal{P}_I : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ denote the projection operator on I, i.e., $(\mathcal{P}_I f)(x) = f(x)$ for $x \in I$ and $(\mathcal{P}_I f)(x) = 0$ otherwise. With this, one can define the *truncated Hilbert transform* as the operator

$$H_T := \mathcal{P}_{[a_1,a_3]} H \mathcal{P}_{[a_2,a_4]}.$$

Note that its adjoint is given by $H_T^* = -\mathcal{P}_{[a_2,a_4]} H \mathcal{P}_{[a_1,a_3]}$, which follows from a basic property of the Hilbert transform that its adjoint is simply $H^* = -H$. To understand the stability properties of solving

$$H_T f = g,$$

for given right-hand side g, it is essential to study the spectrum $\sigma(H_T^* H_T)$.

This has been done in [29, 30] for the truncated Hilbert transform *with a gap*, i.e., for the case $[a_1, a_3] \cap [a_2, a_4] = \emptyset$, and the *interior problem,* i.e., when $[a_1, a_3] \subset [a_2, a_4]$, respectively, see Fig. 4. Here, we present a different limited data scenario than in [29, 30]: the truncated Hilbert transform *with overlap*, i.e., $a_1 < a_2 < a_3 < a_4$, see Fig. 5.

The common theme of the spectral analysis in these problems is to relate the operators in question with differential operators through an intertwining property and to exploit the spectrum of the differential operators. This idea goes back to older problems, such as the spectral analysis of the interior Radon transform [34]. The most prominent example is the problem of Landau, Pollak, and Slepian [31, 32, 44] in communication theory. It can be described as follows: a signal (for example, in telecommunication) is naturally time-limited, say, to an interval $[-T, T]$. At the same time, when signals are transmitted through devices, this can only happen up to a certain frequency. Let $[-W, W]$ be the corresponding bandwidth. Then, the process of transmitting a time-limited signal can be described as applying the operator

$$\mathcal{F}_{TW} := \mathcal{P}_{[-W,W]}\mathcal{F}\mathcal{P}_{[-T,T]},$$

where $\mathcal{F}$ denotes the Fourier transform on $L^2(\mathbb{R})$. From the uncertainty principle, it is apparent that for any signal f, taking $\mathcal{F}_{TW}$ means some loss of information, in either time or frequency (or both). Engineers have thus been interested in quantifying this loss which can be measured by the ratio

$$\frac{\|\mathcal{F}_{TW} f\|^2_{L^2(\mathbb{R})}}{\|f\|^2_{L^2(\mathbb{R})}} = \frac{\langle \mathcal{F}^*_{TW}\mathcal{F}_{TW} f, f\rangle_{L^2(\mathbb{R})}}{\|f\|^2_{L^2(\mathbb{R})}}.$$

This value is maximized when f is an eigenvector to the largest eigenvalue of $\mathcal{F}^*_{TW}\mathcal{F}_{TW}$. The fact that the eigenvalues and eigenvectors of $\mathcal{F}^*_{TW}\mathcal{F}_{TW}$ can be determined relies on the nice property that this operator commutes with the second-order differential operator

$$(Df)(x) = \left((T^2 - x^2) f'(x)\right)' - \frac{W^2}{\pi^2} x^2 f(x)$$

and Sturm–Liouville theory can be employed to obtain its eigensystem. The eigenfunctions $(u_n)_{n\in\mathbb{N}}$ of D are known as the *prolate spheroidal wave functions* and, due to the commutation property, they are also the eigenfunctions of $\mathcal{F}^*_{TW}\mathcal{F}_{TW}$. The eigenvalues of $\mathcal{F}^*_{TW}\mathcal{F}_{TW}$ can then be determined as

$$\lambda_n := \|\mathcal{F}^*_{TW}\mathcal{F}_{TW} u_n\|_{L^2(\mathbb{R})} / \|u_n\|_{L^2(\mathbb{R})}.$$

The commutation of D with $\mathcal{F}^*_{TW}\mathcal{F}_{TW}$ appears to be a lucky accident and while it is not the sole instance of a limited data integral operator commuting with a differential operator, there still exists no coherent theory for this phenomenon. However, the results of Landau, Pollak, and Slepian have to some extent been generalized in [26].

We now return to our problem of finding the spectrum $\sigma(H^*_T H_T)$ for H_T the truncated Hilbert transform with overlap. Motivated by the analysis of the Landau–Pollak–Slepian operator, as well as the interior Radon transform and the two truncated Hilbert transforms in [29, 30], we, too, embark on the journey of finding commuting

differential operators of which the spectral properties can be understood. We can formulate our goal more precisely as finding second-order differential operators L_S and $\widetilde{L_S}$ such that

$$H_T L_S = \widetilde{L_S} H_T.$$

In view of the desired aim to analyze the spectral properties of these operators, we require L_S and $\widetilde{L_S}$ to be self-adjoint. If it turns out that L_S, $\widetilde{L_S}$ have simple discrete spectra, then one can work toward obtaining the SVD of H_T. Note that beforehand, there is no guarantee that H_T has an SVD, i.e., that $\sigma(H_T^* H_T)$ is purely discrete.

Since we seek to find differential, and hence unbounded, operators L_S, $\widetilde{L_S}$ that are also self-adjoint, it is worth reviewing a fundamental theorem.

Theorem 30 (Hellinger–Toeplitz theorem) *Let A be a linear everywhere-defined operator on a Hilbert space X with*

$$\langle f_1, Af_2\rangle_X = \langle Af_1, f_2\rangle_X, \quad \forall f_1, f_2 \in X.$$

Then, A is bounded.

In other words, an *unbounded self-adjoint* operator A cannot have its domain agreeing with all of X, i.e., $\mathcal{D}(A) \subsetneq X$. To make this more precise, we give the following definition:

Definition 31 (*Symmetric operator*) A densely defined operator A on X is symmetric if and only if

$$\langle f_1, Af_2\rangle_X = \langle Af_1, f_2\rangle_X, \quad \forall f_1, f_2 \in \mathcal{D}(A).$$

We remark that for a symmetric operator A, $\mathcal{D}(A) \subseteq \mathcal{D}(A^*)$. Furthermore, A is self-adjoint if and only if it is symmetric and $\mathcal{D}(A) = \mathcal{D}(A^*)$. Starting from a symmetric operator, we will thus search for *self-adjoint extensions* by specifying suitable domains. Note that the spectrum of an unbounded operator is very sensitive to the choice of the domain. As in [29, 30], we choose to start with the differential form $L(x, d_x)$ defined as

$$L(x, d_x)\psi(x) := \big(P(x)\psi'(x)\big)' + Q(x)\psi(x), \tag{15}$$

where $P(x) := \prod_{i=1}^4 (x - a_i)$ and $Q(x) := 2\left(x - \frac{1}{4}\sum_{i=1}^4 a_i\right)^2$. The aim is to find self-adjoint operators L_S, $\widetilde{L_S}$ with $\mathcal{D}(L_S) \subset L^2([a_2, a_4])$ and $\mathcal{D}(\widetilde{L_S}) \subset L^2([a_1, a_3])$, respectively. Formally, they are self-adjoint extensions of symmetric operators L_{min}, $\widetilde{L}_{min}$ with $\mathcal{D}(L_S) \supset \mathcal{D}(L_{min})$ and $\mathcal{D}(\widetilde{L_S}) \supset \mathcal{D}(\widetilde{L}_{min})$ and we refer to [9] for a definition of these operators.

For the spectral analysis, the interest lies in solutions ψ to the Sturm–Liouville problem (SLP)

$$L(x, d_x)\psi(x) = \lambda\psi(x), \tag{16}$$

for $\lambda \in \mathbb{C}$. A point $x_0 \in \overline{\mathbb{C}}$ is said to be *ordinary* if the functions

$$\widetilde{P}(x) := \frac{P'(x)}{P(x)},$$
$$\widetilde{Q}(x) := \frac{Q(x) - \lambda}{P(x)}$$

are analytic at x_0. The points a_i, $i = 1, \ldots, 4$ in (15) are not ordinary. More precisely, they are *regular singular*, meaning that at a_i, $\widetilde{P}(x)$ has a pole of up to order 1 and $\widetilde{Q}(x)$ has a pole of up to order 2. A standard result known as Fuchs' theorem roughly states that at regular singular points, solutions to (16) are either bounded or have a logarithmic singularity.

In the analysis of the interior problem and the truncated Hilbert transform with a gap, the same differential form $L(x, d_x)$ as in (15) appears. However, in these cases, one seeks self-adjoint operators on intervals I and J for which a_i are *endpoints,* but no point a_i appears *inside* the intervals I and J. This makes the corresponding eigenvalue problems *standard singular* Sturm–Liouville problems with the only singular points being the endpoints of the interval. In such a case, the spectral properties of a self-adjoint extension are well known (cf. [47]).

The case of the truncated Hilbert transform with overlap is fundamentally different: here, we need to consider self-adjoint extensions on intervals $[a_2, a_4]$ and $[a_1, a_3]$ with a_3 and a_2, respectively, in the interior of the interval. With this, one exits standard Sturm–Liouville theory and is required to work with a singular Sturm–Liouville problem *on two intervals* (meaning that, e.g., when considering L_S with $\mathcal{D}(L_S) \subset L^2([a_2, a_4])$, there are two involved subintervals $[a_2, a_3]$ and $[a_3, a_4]$ as the interval $[a_2, a_4]$ is interrupted by an *interior singular point*, namely a_3). For such *two interval problems*, there is some theory on self-adjoint extensions (cf. [47]) that we will employ. However, the results on the spectral properties are no longer straightforward. For a self-adjoint extension L_S with $\mathcal{D}(L_S) \subset L^2([a_2, a_4])$, where $a_3 \in (a_2, a_4)$, the introduction of two boundary conditions (BCs) and two *transmission conditions (TCs)* is necessary to obtain a self-adjoint realization. The rationale for having four conditions is that now one works with *two* intervals and hence needs double the number of conditions (for a second-order ODE on one interval we need two conditions). The two BCs will act on the endpoints a_2 and a_4, while the two TCs will be there to connect the two subintervals (a_2, a_3) and (a_3, a_4).

Ultimately, we seek solutions to (16) that take on a special form: the (potential) eigenfunctions $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ of L_S and $\widetilde{L_S}$, respectively, should satisfy

$$H_T u_n = \sigma_n v_n,$$
$$H_T^* v_n = \sigma_n u_n,$$

for some real numbers σ_n. We can use this goal, together with some intuition on the Hilbert transform, to find suitable BCs and TCs. For example, when we take the Hilbert transform of a function ψ with a jump discontinuity at x_0, then $H\psi$ will

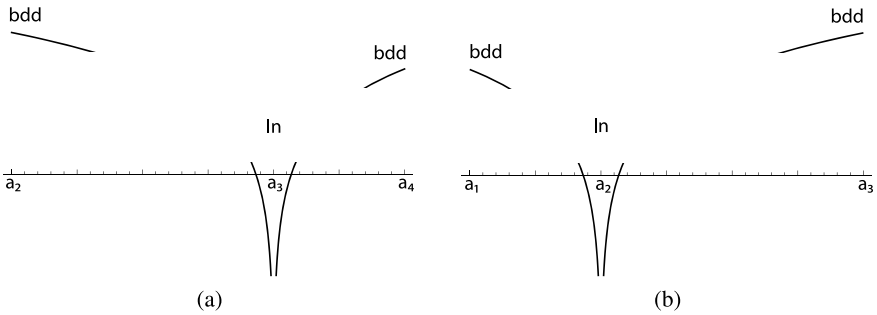


Fig. 6 Sketch of the (potential) singular functions u_n and v_n of H_T. They are bounded at the endpoints and have a logarithmic singularity at the interior singular point

have a logarithmic singularity at x_0. On the other hand, suppose that at a point x_0, a function ψ has a logarithmic singularity, i.e., in a region around x_0,

$$\psi(x) = \phi_{11}(x) + \phi_{12}(x) \ln|x - x_0|, \quad x < x_0, \tag{17}$$

and

$$\psi(x) = \phi_{21}(x) + \phi_{22}(x) \ln|x - x_0|, \quad x > x_0, \tag{18}$$

for some analytic functions ϕ_{ij}. Then, in order for $H\psi$ to be bounded at x_0, one necessarily has

$$\lim_{x \to x_0^-} \phi_{11}(x) = \lim_{x \to x_0^+} \phi_{21}(x), \tag{19}$$

$$\lim_{x \to x_0^-} \phi_{12}(x) = \lim_{x \to x_0^+} \phi_{22}(x). \tag{20}$$

These observations lead to a specific picture of the singular functions of H_T, should they exist. They are bounded at the endpoints and have a logarithmic singularity of type (17)–(20) at the interior singular point. See [9] for full details and Fig. 6 for an illustration.

To present a suitable candidate for L_S, we need a few more preparations. First, for an open interval $I \subset \mathbb{R}$, define the function space

$$AC_{loc}(I) := \{\psi : I \to \mathbb{C} : \psi \text{ is absolutely continuous on all } [\alpha, \beta] \subset I\}.$$

Next, let the *maximal domain* $D_{max} \subset L^2([a_2, a_4])$ be given by

$$D_{max} := \{\psi : (a_2, a_4) \to \mathbb{C} : \psi|_{(a_i, a_{i+1})}, (P\psi')|_{(a_i, a_{i+1})} \in AC_{loc}((a_i, a_{i+1})), i = 2, 3, \\ \psi, L\psi \in L^2([a_2, a_4])\},$$

and recall the notion of the Lagrange sesquilinear form $[\cdot,\cdot]$ of two functions $r, s \in D_{max}$ which is defined as

$$[r, s] := r P \overline{s}' - \overline{s} P r'.$$

One can deduce from Green's formula that for all $r, s \in D_{max}$, the limits $\lim_{\alpha \to a_2^+} [r, s](\alpha)$, $\lim_{\beta \to a_3^-} [r, s](\beta)$, $\lim_{\alpha \to a_3^+} [r, s](\alpha)$, and $\lim_{\beta \to a_4^-} [r, s](\beta)$ exist and are finite.

Theorem 32 *Let $L_S : D(L_S) \to L^2([a_2, a_4])$ be the extension of L_{min} to the domain*

$$\begin{aligned} D(L_S) := \{\psi \in D_{max} : [\psi, r](a_2^+) = [\psi, r](a_4^-) = 0, \\ [\psi, r](a_3^-) = [\psi, r](a_3^+), \\ [\psi, s](a_3^-) = [\psi, s](a_3^+)\}, \end{aligned}$$

with $r, s \in D_{max}$ chosen as

$$\begin{aligned} r(y) &:= 1, \\ s(y) &:= \sum_{i=1}^{4} \prod_{\substack{j \neq i \\ j \in \{1,\dots,4\}}} \frac{1}{a_i - a_j} \ln |y - a_i|. \end{aligned}$$

Then, L_S is a self-adjoint operator.

Proof See Chap. 13 in [47], in which all self-adjoint extensions for two interval problems are given. □

Note that for $\lambda \in \mathbb{C}$, the above two BCs at a_2 and a_4, as well as the two TCs at a_3 simplify for solutions of $L\psi = \lambda\psi$. More precisely, if ψ solves $L_S\psi = \lambda\psi$, then the above BCs mean that ψ is bounded at the endpoints a_2 and a_4. Furthermore, the two TCs translate to (17)–(20). This gives hope that indeed, for this choice of L_S, there is a (much anticipated) relation between H_T and L_S. In main contrast to Sturm–Liouville problems on one interval with no interior singular point, we have no straightforward guarantee that the spectrum of L_S is purely discrete. One of the main findings in [9] is that this is, however, indeed the case. We summarize the result as follows.

Theorem 33 (Spectrum of L_S) *Let L_S be the self-adjoint extension of L_{min} with domain $D(L_S) \subset L^2([a_2, a_4])$ as defined in Theorem 32. Then, its spectrum $\sigma(L_S)$ has the following properties:*

(i) *$\sigma(L_S)$ is purely discrete.*
(ii) *The set of eigenfunctions $(u_n)_{n\in\mathbb{N}}$ of L_S are complete in $L^2([a_2, a_4])$.*
(iii) *$\sigma(L_S)$ is simple, i.e., each eigenvalue has multiplicity* 1.
(iv) *For all eigenfunctions u_n of L_S:*

$$\left(H_T L(y, d_y) u_n\right)(x) = L(x, d_x)\left(H_T u_n\right)(x).$$

One can define a second self-adjoint operator $\widetilde{L_S}$ with $D(\widetilde{L_S}) \subset L^2([a_1, a_3])$ equivalently to the definition of L_S (by simply replacing the points a_2, a_3 and a_4 by a_1, a_2, and a_3, respectively. This allows to write property (iv) in the above more compactly as

$$H_T L_S = \widetilde{L_S} H_T.$$

With that, one finally arrives at the following.

Theorem 34 (SVD of H_T) *The eigenfunctions u_n of L_S, together with*

$$\begin{aligned} v_n &:= H_T u_n / \|H_T u_n\|_{L^2([a_1,a_3])}, \\ \sigma_n &:= \|H_T u_n\|_{L^2([a_1,a_3])}, \end{aligned}$$

form the SVD of H_T:

$$\begin{aligned} H_T u_n &= \sigma_n v_n, \\ H_T^* v_n &= \sigma_n u_n. \end{aligned}$$

The functions $(v_n)_{n\in\mathbb{N}}$ are the eigenfunctions of $\widetilde{L_S}$ and form a complete orthonormal system of $L^2([a_1, a_3])$. In light of Hadamard's well-posedness criteria one can further show that

$$\mathcal{N}(H_T) = \{0\},$$

i.e., the inversion problem enjoys uniqueness, and

$$\mathcal{R}(H_T) \neq L^2([a_2, a_4]),$$

while $\mathcal{R}(H_T)$ is dense in $L^2([a_2, a_4])$. Thus, inverting from truncated Hilbert transform data is ill-posed in the sense that the solution does not depend continuously on the data. Another interesting fact is the following.

Theorem 35 *The values* 0 *and* 1 *are (the only) accumulation points of the singular values of H_T.*

This property implies that H_T is *not* a compact operator. The accumulation of the singular values at 0 causes the instability of inverting the truncated Hilbert transform. As discussed in Remark 25, the decay rate of the singular values reveals the nature of the ill-posedness. To find how severe the ill-posedness of inverting H_T is, we again make use of the differential operators L_S, $\widetilde{L_S}$. We aim at finding the asymptotics of the eigenfunctions ψ_n of L_S as $\lambda_n \to \pm\infty$ in

$$L_S \psi_n = \lambda_n \psi_n.$$

Note that the two accumulation points $+\infty$ and $-\infty$ of λ_n correspond to the accumulation points 0 and 1 of σ_n, respectively. The asymptotic analysis of ψ_n for $n \to \pm\infty$ is based on three ingredients:

(i) Global asymptotics: away from the regular singular points a_i, the solution ψ_n is analytic and its asymptotics can be described with WKB (Wentzel–Kramers–Brillouin) approximations.
(ii) Local asymptotics: close to the regular singular points a_i, the solution ψ_n can be characterized by linear combinations of the Bessel functions J_0 and Y_0.
(iii) Asymptotic matching: global and local asymptotics need to be matched in specified regions in which both are valid.

The above is just a sketch of the recipe and we refer to [5] for the full argument. Since the eigenfunctions of L_S and $\widetilde{L_S}$ correspond to the two sets of singular functions of H_T, one has thus found the asymptotics of the singular functions of H_T. This can be used to further derive the asymptotic behavior of $\sigma_n \to 1$ and $\sigma_{-n} \to 0$ as $n \to \infty$. The result can be stated as follows.

Theorem 36 *Let $(\sigma_n)_{n\in\mathbb{N}}$ and $(\sigma_{-n})_{n\in\mathbb{N}}$ denote the sequences of singular values of H_T accumulating at 1 and 0, respectively. Then, there exist constants $c_1, c_2 > 0$ depending on only P and the points a_i, $i = 1, \ldots, 4$ such that*

$$\begin{aligned}\sigma_n &= 2e^{-c_1 n} \cdot \left(1 + O\left(n^{-1/2+\delta}\right)\right), \\ \sigma_{-n} &= 1 - 2e^{-c_2 n} \cdot \left(1 + O\left(n^{-1/2+\delta}\right)\right), \quad \text{as } n \to \infty,\end{aligned}$$

for some small fixed $\delta > 0$.

Thus, the decay to 0 is exponential, which leads us to classify this inversion problem as severely ill-posed. We remark that this is typical for limited data problems in CT.

4 Regularization

So far, we have discussed detecting the instability of an inverse problem and, if an SVD exists, characterizing the severity of the ill-posedness through the decay rate of the singular values. This, of course, is only of theoretical interest, if it does not lead to new reconstruction methods dealing with these instabilities. In this section, we will see how one can aim at extracting information *as stably as possible* from an unstable system. This is the goal of *regularization.* With the example of the truncated Hilbert transform, we will further demonstrate in Sect. 4.2, how the derived knowledge of the SVD of the underlying operator and its asymptotic properties can enable us to prove rigorous results on the proposed *regularization methods.*

Clearly, solving for $Tf = g$ can be done by applying the Moore–Penrose generalized inverse to the right-hand side, resulting in a best-approximate solution:

$$f^\dagger = T^\dagger g.$$

However, in practice, g is not known exactly, but only some measurement g^δ is acquired up to some *noise level* δ, i.e., one only has a guarantee of the form

$$\|g - g^\delta\|_Y \leq \delta.$$

The main issue is the lack of continuous dependence of the data on the right-hand side: recall that if Hadamard's third property is violated, then $T^\dagger$ is not a continuous operator. Thus, in general, $T^\dagger g^\delta$ is not a good approximation of $T^\dagger g$. Also note that $T^\dagger g^\delta$ might not even exist, since $\mathcal{D}(T^\dagger) \subsetneq Y$ when $T^\dagger$ is not continuous.

In *regularization theory*, one seeks to find an approximation f^δ of $f^\dagger$ such that

- f^δ depends continuously on g^δ,
- $f^\delta \to f^\dagger$ as $\delta \to 0$.

This is achieved by constructing a family of continuous operators $(R_\alpha)_{\alpha \in (0,\overline{\alpha})}$, $\overline{\alpha} \in \mathbb{R}_+ \cup \infty$, that approximate the unbounded operator $T^\dagger$. More precisely, for $\alpha = \alpha(\delta, g^\delta)$, define $f_\alpha^\delta := R_\alpha g^\delta$. The goal is to choose $\alpha(\delta, g^\delta)$ and R_α such that $f_\alpha^\delta \to f^\dagger$ as $\delta \to 0$. In other words, the lower the noise level, the more accurate the approximation f_α^δ is required to be. For high noise levels δ, the reconstruction f_α^δ does not need to be close to $f^\dagger$. This matches the intuition that if the right-hand side is only known up to δ, it is not feasible to aim at an approximation of the true solution $f^\dagger$ that is closer than δ.. A precise definition of a regularization can be given as follows.

Definition 37 Let $T \in \mathcal{L}(X, Y)$, $\overline{\alpha} \in \mathbb{R}_+ \cup \{\infty\}$ and let $R_\alpha : Y \to X$ be a continuous operator for every $\alpha \in (0, \overline{\alpha})$. Suppose that for all $g \in \mathcal{D}(T^\dagger)$ there exists a parameter choice rule $\alpha = \alpha(\delta, g^\delta) : \mathbb{R}_+ \times Y \to (0, \overline{\alpha})$ such that the following hold:

$$\limsup_{\delta \to 0} \left\{ \alpha(\delta, g^\delta) : g^\delta \in Y, \|g - g^\delta\|_Y \leq \delta \right\} = 0, \tag{21}$$

and

$$\limsup_{\delta \to 0} \left\{ \|R_{\alpha(\delta, g^\delta)} g^\delta - T^\dagger g\|_X : g^\delta \in Y, \|g - g^\delta\|_Y \leq \delta \right\} = 0. \tag{22}$$

Then, the family $(R_\alpha)_{\alpha \in (0,\overline{\alpha})}$ is called a *regularization* for $T^\dagger$. For every $g \in \mathcal{D}(T^\dagger)$, a pair (R_α, α) is called a *convergent regularization method* for solving $Tf = g$, if Eqs. (21) and (22) hold.

A regularization method is thus defined by two components:

- the operators R_α and
- the parameter choice rule $\alpha(\delta, g^\delta)$.

A fundamental result by Bakushinsky states that α cannot be chosen independently of δ.

Theorem 38 (A. B. Bakushinsky) *If $\alpha = \alpha(g^\delta)$ yields a convergent regularization method, then $T^\dagger$ is bounded.*

There are two possible choices of dependencies left and they are divided into *a priori* parameter choice rules, i.e., $\alpha = \alpha(\delta)$, and *a posteriori* parameter choice rules, i.e., $\alpha = \alpha(\delta, g^\delta)$.

Existence of a priori parameter choice rules can be guaranteed when $(R_\alpha)_{\alpha\in(0,\overline{\alpha})}$ is a family of continuous operators converging to $T^\dagger$ point-wise.

Theorem 39 *If for all $\alpha > 0$, R_α is a continuous operator and*

$$R_\alpha \to T^\dagger \text{ point-wise on } \mathcal{D}(T^\dagger), \quad \text{as } \alpha \to 0,$$

then $(R_\alpha)_{\alpha\in\mathbb{R}_+}$ is a regularization of $T^\dagger$ and for all $g \in \mathcal{D}(T^\dagger)$ there exists an a priori parameter choice rule $\alpha(\delta)$ for which (R_α, α) is a convergent regularization method for $Tf = g$.

A regularization consisting of linear operators R_α is called a *linear regularization method.* Note that one can also consider a family of nonlinear operators R_α for approximating a linear operator $T^\dagger$. A well-known example is the conjugate gradient method equipped with an early stopping criterion to ensure regularization.

In order to construct regularization methods, the following viewpoint is helpful: Suppose that the operator T^*T was continuously invertible with spectral projections P_λ, so that its inverse could be expressed as

$$\left(T^*T\right)^{-1} = \int_{\sigma(T^*T)} \frac{1}{\lambda} dP_\lambda.$$

Then, in view of (9), the following holds for the best-approximate solution $f^\dagger$:

$$f^\dagger = \int_{\sigma(T^*T)} \frac{1}{\lambda} dP_\lambda T^* g. \tag{23}$$

If, however, $\mathcal{R}(T)$ is not closed, the above integral does not exist because zero belongs to the spectrum of T^*T, and hence the integrand $1/\lambda$ has a pole at zero. The concept of regularization is to replace $1/\lambda$ by a family of functions $(s_\alpha(\lambda))_{\alpha\in\mathbb{R}_+}$ that approximates $1/\lambda$ and satisfies some continuity conditions. Instead of computing $f^\dagger$ one then constructs

$$f_\alpha := \int_{\sigma(T^*T)} s_\alpha(\lambda) dP_\lambda T^* g, \tag{24}$$

and the corresponding regularization operators are given by

$$R_\alpha := \int_{\sigma(T^*T)} s_\alpha(\lambda) dP_\lambda T^*. \tag{25}$$

More precisely, one has the following.

Theorem 40 *For all $\alpha > 0$, let $s_\alpha : [0, \|T\|^2] \to \mathbb{R}$ be piecewise continuous and suppose that there is a constant $C > 0$ such that for all $\lambda \in (0, \|T\|^2]$*

$$|\lambda s_\alpha(\lambda)| \leq C, \tag{26}$$

and

$$\lim_{\alpha \to 0} s_\alpha(\lambda) = \frac{1}{\lambda}. \tag{27}$$

Then, for all $g \in \mathcal{D}(T^\dagger)$,

$$\lim_{\alpha \to 0} f_\alpha = f^\dagger$$

with $f^\dagger = T^\dagger y$ *and*

$$f_\alpha := \int_{\lambda \in \sigma(T^*T)} s_\alpha(\lambda) d P_\lambda T^* g.$$

Note that in view of the discussion in Sect. 2.3, the continuous functional calculus enables us to write

$$f_\alpha = s_\alpha(T^*T) T^* g. \tag{28}$$

Similarly, for reconstruction from noisy data g^δ, the approximation via s_α is expressed as

$$f_\alpha^\delta = s_\alpha(T^*T) T^* g^\delta. \tag{29}$$

Further note that, due to the ill-posedness, α has to be chosen carefully because when $g \notin \mathcal{D}(T^\dagger)$,

$$\lim_{\alpha \to 0} \| f_\alpha^\delta \|_X = \infty.$$

One way to ensure convergence of f_α^δ is to choose $\alpha(\delta, g^\delta)$ via Morozov's discrepancy principle, which is an a posteriori rule.

Theorem 41 *Let* s_α *be as in Theorem 40, fulfilling (26) and (27) and assume that for each* $\lambda > 0$, $\alpha \mapsto s_\alpha(\lambda)$ *is continuous from the left. Also, define*

$$r_\alpha(\lambda) := 1 - \lambda s_\alpha(\lambda).$$

Furthermore, let

$$S_\alpha := \sup \left\{ s_\alpha(\lambda) : \ \lambda \in [0, \|T\|^2] \right\}$$

be such that

$$S_\alpha \leq \frac{\tilde{c}}{\alpha}, \ \ \textit{for } \alpha > 0,$$

for some constant $\tilde{c} > 0$ *and*

$$\tau > \sup \left\{ |r_\alpha(\lambda)| : \ \alpha > 0, \ \lambda \in [0, \|T\|^2] \right\}.$$

Then, the discrepancy principle defined by

$$\alpha(\delta, g^\delta) := \sup\left\{\alpha > 0 : \ \|Tf_\alpha^\delta - g^\delta\|_Y \leq \tau\delta\right\} \tag{30}$$

and R_α defined as in (25) form a convergent regularization method (R_α, α) for all $g \in \mathcal{R}(T)$.

Remark 42 • The requirement that $g \in \mathcal{R}(T)$ is not restrictive: if $g \in \mathcal{D}(T^\dagger)$, but $g \notin \mathcal{R}(T)$, then we can simply solve for

$$T^*Tf = T^*g$$

instead of $Tf = g$. This is then solvable for $g \in \mathcal{D}(T^\dagger)$ and the same result applies for this normal equation.

- The philosophy of the discrepancy principle is very intuitive: one compares the residual with the error bound δ and does not aim at an approximation that achieves a residual below δ as this is not meaningful: for noisy data g^δ, with $\|g - g^\delta\|_Y \leq \delta$, the best one should ask for is a residual of the order of δ. On the other hand, from the viewpoint of regularization, one should aim at a regularization parameter as large as possible to ensure stability. This is a trade-off between accuracy and stability and the discrepancy principle roughly aims at achieving the optimal balance.

Next, we give two simple and well-known examples of regularization methods.

Example 43 One way to regularize the inversion of T^*T is via a simple threshold:

$$s_\alpha(\lambda) := \begin{cases} \frac{1}{\lambda}, & \text{for } \lambda \geq \alpha, \\ 0, & \text{for } \lambda < \alpha. \end{cases}$$

For an operator with an SVD $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$, this amounts to a *truncated SVD:*

$$f_\alpha^\delta := \sum_{\substack{n=1 \\ \sigma_n^2 \geq \alpha}}^{\infty} \frac{1}{\sigma_n} \langle g^\delta, v_n\rangle_Y u_n.$$

Example 44 The most prominent example for regularization is *Tikhonov regularization* which amounts to defining

$$s_\alpha(\lambda) := \frac{1}{\lambda + \alpha},$$

for $\alpha > 0$. Since $\{\lambda + \alpha : \lambda \in \sigma(T^*T)\}$ is the spectrum of $T^*T + \alpha\mathrm{Id}$, Tikhonov regularization can be interpreted as

$$f_\alpha^\delta = \int_{\sigma(T^*T)} s_\alpha(\lambda) dP_\lambda T^*g^\delta = \left(T^*T + \alpha\mathrm{Id}\right)^{-1} T^*g^\delta.$$

In other words, one solves the regularized normal equation

$$\left(T^*T + \alpha \mathrm{Id}\right) f_\alpha^\delta = T^* g^\delta. \tag{31}$$

For an operator with an SVD $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$, this can be written as

$$f_\alpha^\delta := \sum_{n=1}^{\infty} \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle g^\delta, v_n \rangle_Y u_n, \tag{32}$$

i.e., the unbounded term $1/\sigma_n$ is replaced by the bounded term $\sigma_n/(\sigma_n^2 + \alpha)$. One can show that f_α^δ in (32) is the unique minimizer of the *Tikhonov functional*

$$f \mapsto \|Tf - g^\delta\|_Y^2 + \alpha \|f\|_X^2. \tag{33}$$

Formulated this way, Tikhonov regularization exhibits the typical form that instead of simply minimizing the residual, one introduces an additional *penalty term*, in this case the norm of f. As the noise level δ decreases, one can choose smaller α, so that the penalty term becomes less emphasized.

Theorem 45 *If $\alpha(\delta, g^\delta)$ is chosen according to the discrepancy principle (30), Tikhonov regularization converges for all $g \in \mathcal{R}(T)$.*

4.1 Miller's Theory

As suggested in (33), one can alternatively study regularization from an optimization perspective. This has been suggested by Miller [37] and further discussed by Bertero, De Mol, and Viano in [15]. Suppose for simplicity that T^{-1} exists and that a right-hand side g^δ is given with noise level $\|g - g^\delta\|_Y \leq \delta$. In case of ill-posedness, T^{-1} is unbounded and hence the set

$$H(\delta, g^\delta) := \{f \in X : \|Tf - g^\delta\|_Y \leq \delta\}$$

is unbounded. Thus, finding an element in $H(\delta, g^\delta)$ does not give any guarantee on how close it is to the exact solution. To achieve regularization, one introduces a restricted set of *admissible solutions* $\mathcal{S}(\delta, g^\delta) \subset H(\delta, g^\delta)$, i.e., one assumes prior knowledge on the solution and hence only searches in a restricted set $\mathcal{S}(\delta, g^\delta)$. If

$$\operatorname{diam} \mathcal{S}(\delta, g^\delta) \to 0, \quad as\ \delta \to 0, \tag{34}$$

then the problem is said to be regularized and any method (R_α, α) that guarantees

$$R_{\alpha(\delta, g^\delta)} g^\delta \in \mathcal{S}(\delta, g^\delta)$$

is a convergent regularization method. A typical choice for $\mathcal{S}(\delta, g^\delta)$ is $\{f \in X : \|Tf - g^\delta\|_Y \leq \delta,\ \|Lf\|_X \leq c\}$, for some constant $c > 0$ and L a densely defined

operator with bounded inverse. The choice $L = \mathrm{Id}$ corresponds to Tikhonov regularization. Other popular choices for L are differential operators, in which case the constraint amounts to a smoothness condition on f.

4.2 Regularization for the Truncated Hilbert Transform

We conclude this section by presenting results on the regularized reconstruction from truncated Hilbert transform data. As outlined in Sect. 3, the spectral properties of the operator H_T have been derived in [5, 9]. The main tool for determining the asymptotics of the singular values is to find the global asymptotic behavior of the singular functions. This knowledge has been further used in [6, 10] to derive results on the regularization in the sense of Miller's approach.

More precisely, when studying the asymptotics of the singular functions, one finds two characteristics:

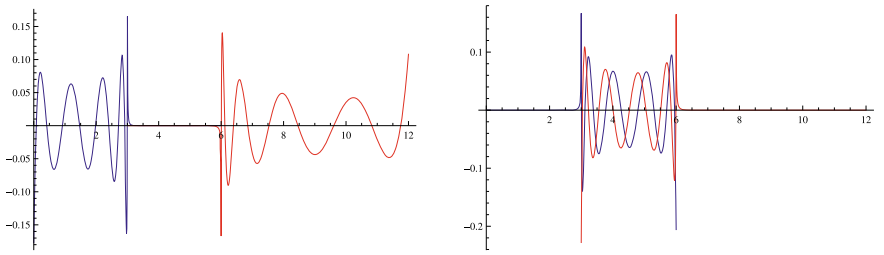
- **Oscillating behavior.** The singular functions u_n corresponding to accumulation in the spectrum at 1 oscillate inside the overlapping region, i.e., on $[a_2, a_3]$, and decay monotonically on $[a_3, a_4]$. On the other hand, the u_n's corresponding to accumulation in the spectrum at 0 oscillate outside of the overlapping region, i.e., on $[a_3, a_4]$ and decay monotonically on $[a_2, a_3]$ (see Fig. 7). This suggests that the part of the spectrum causing instabilities corresponds to signals that are highly oscillating outside of the region-of-interest. Thus, to restore stability, one needs to suppress high oscillations outside of the overlapping region.
- **Logarithmic singularity at the interior singular point.** As we have already seen in Sect. 3, all singular functions have logarithmic singularities at the interior singular point, i.e., at a_3 (for u_n) and at a_2 (for v_n), see Fig. 7. Thus, methods that are SVD based (cf. Examples 43 and 44) might not be ideal: since they use a superposition of the singular functions in the reconstruction, they will most likely create reconstruction artifacts in the form of logarithmic singularities at a_3.

The derivation of stability estimates from the asymptotic expansions of u_n and v_n is very involved and technical and outside of the scope of this chapter. We merely present the results here. The first statement is that if one is only concerned with stability in a slightly restricted region-of-interest of the form $[a_2, a_3 - \mu]$, for some small $\mu > 0$, then the method of Tikhonov already yields a regularization for reconstruction from H_T.

Theorem 46 *Let $g \in \mathcal{R}(H_T)$ and $g^\delta \in L^2([a_1, a_3])$ be noisy data such that $\|g - g^\delta\|_Y \leq \delta$ for some noise level $\delta > 0$. For $E > 0$, define the set of admissible solutions as*

$$\mathcal{S}(\delta, g^\delta) := \{f \in L^2([a_2, a_4]) : \ \|H_T f - g\|_{L^2([a_1,a_3])} \leq \delta, \ \|f\|_{L^2([a_2,a_4])} \leq E\}.$$

Let $\mu > 0$ be constant and consider the reconstruction on $(a_2, a_3 - \mu)$. Then, for sufficiently small δ, any $f_1, f_2 \in \mathcal{S}(\delta, g^\delta)$ satisfy a bound of the form



(a) (b)

Fig. 7 Examples of singular functions u_n and v_n in red and blue, respectively. Here, $a_1 = 0, a_2 = 3$, $a_3 = 6, a_4 = 12$. Singular functions for σ_n close to 0 in (**a**); singular functions for σ_n close to 1 in (**b**). All of the singular functions exhibit singularities at the interior singular points

Fig. 8 Example of a limited data problem: the black circle indicates the field-of-view (FOV)

$$\|f_1 - f_2\|_{L^2([a_2,a_3-\mu])} \leq C_1\delta + C_2 E^{1-\gamma}\delta^{\gamma},$$

where $C_1, C_2, \gamma > 0$ *are constants depending on only* μ *and the relative positions of the points* a_1, a_2, a_3, a_4.

Remark 47 As already stated, the drawback of using Tikhonov regularization here is that it will create artifacts at the boundary of the region-of-interest due to the logarithmic singularities of the singular functions. See Figs. 8 and 9 for an example: Tikhonov regularization clearly exhibits these artifacts on the boundary of the ROI.

As one can see in Fig. 9, the reconstruction using *total variation (TV) regularization* does not show the artifacts on the boundary of the ROI. This is because the method is not SVD based and penalizes singularities. To be more precise, in TV regularization, the set of admissible solutions $\mathcal{S}(\delta, g^{\delta})$ is chosen by restricting the total variation of the admissible functions. For weakly differentiable functions f with derivative f_x, the TV semi-norm is given by

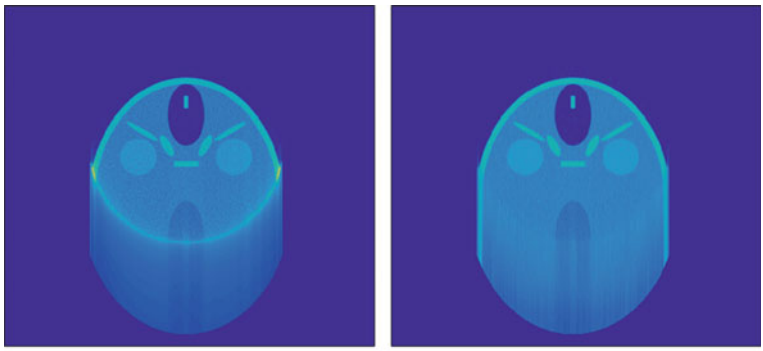


Fig. 9 Regularized reconstruction for the limited data problem in Fig. 8 using Tikhonov regularization in (**a**) and total variation (TV) regularization in (**b**). The artifacts at the boundaries of the ROI are apparent in the Tikhonov reconstruction but not in the TV reconstruction

$$|f|_{TV} := \|f_x\|_{L^1(\mathbb{R})}.$$

The reason we are interested in TV regularization is that a TV penalty is the natural quantity to penalize both the artifacts on the boundary of the ROI, as well as highly oscillating behavior, causing ill-posedness. Again, exploiting the fine properties of the global asymptotic behavior of the singular functions (and now combined with an argument using Helly's selection theorem), one can formulate a stability estimate for TV regularization. In fact, it suffices to penalize the total variation on a subinterval $[a_3 - \mu, a_4]$, for some $\mu > 0$.

Theorem 48 *Let $g \in \mathcal{R}(H_T)$ and $g^\delta \in L^2([a_1, a_3])$ be noisy data such that $\|g - g^\delta\|_Y \leq \delta$ for some noise level $\delta > 0$. For $\mu, \kappa > 0$, define the set of admissible solutions as*

$$\mathcal{S}(\delta, g^\delta) := \{f \in W^{1,1}([a_2, a_4]) : \ \|H_T f - g\|_{L^2([a_1,a_3])} \leq \delta, \ \|f_x\|_{L^1([a_3-\mu,a_4])} \leq \kappa, \int_{a_2}^{a_4} f(x)dx = C\},$$

for some constant C. Then, as $\delta \to 0$, one has that

$$diam\, \mathcal{S}(\delta, g^\delta) = O(|\log \delta|^{-1/2})$$

and the constants in the decay rate depend on only μ and the relative positions of the points a_1, a_2, a_3, a_4.

Remark 49 Note that this decay rate is only logarithmic, while for Tikhonov regularization one has a decay of order δ^γ. However, the imposed prior knowledge in the case of TV regularization is mainly on the region outside of the ROI. Also, the

guarantee that one obtains for the reconstruction is on $\|f_1 - f_2\|_{L^2([a_2,a_4])}$ instead of merely on $\|f_1 - f_2\|_{L^2([a_2,a_3-\mu])}$.

5 Nonlinear Inverse Problems

While the theory of regularization is well understood in the linear case, it is much less straightforward in the case of nonlinear problems. Since for nonlinear operators there is hardly any spectral theory at hand, the analysis of the regularization becomes much more challenging. In this section, we intend to make brief mention of the particularities of treating nonlinear problems and refer to [21, 28] for a detailed discussion of the subject. In what follows, a nonlinear operator is denoted by

$$F(f) = g, \quad F : \mathcal{D}(F) \subset X \to Y$$

and ill-posedness always refers to the lack of continuous dependence of the solution on the data. An important class of nonlinear (typically ill-posed) problems is that of parameter estimation in PDEs.

Example 50 Suppose that for some material with support in $\Omega \subset \mathbb{R}^3$, $u(x)$, $x \in \Omega$ denotes the temperature distribution after sufficiently long time, h denotes internal heat sources and q the heat conductivity of the material. Assuming that u is kept zero at the boundary, the dependencies can be modeled as follows:

$$\begin{aligned} -\nabla \cdot (q(x)\nabla u) &= h(x), \quad x \in \Omega, \\ u(x) &= 0, \quad x \in \partial\Omega. \end{aligned}$$

Further assuming that h is known, the following problem is a typical parameter estimation problem: Given internal measurements of u or boundary measurements of the heat flux $q\frac{\partial u}{\partial n}$, determine the heat conductivity q. Note that the underlying operator $F : q \mapsto u_q$ is not given explicitly but is described through the PDE.

General assumptions typically made when considering nonlinear inverse problems (and also assumed in the remainder of this section) are the following:

- F is continuous,
- F is weakly sequentially closed, i.e.,

$$\left.\begin{array}{rl} f_n & \rightharpoonup f \text{ in } X \\ F(f_n) & \rightharpoonup g \text{ in } Y \end{array}\right\} \Longrightarrow f \in \mathcal{D}(F) \text{ and } F(f) = g.$$

- For simplicity, one assumes $g \in \mathcal{R}(F)$.

For linear problems, the notion of minimum-norm solution has been introduced. For nonlinear problems, one rather considers the f^*-*minimum-norm* solution $f^\dagger$ which

minimizes $\|f - f^*\|_X$. This is because the element 0 no longer plays a special role for nonlinear problems. Typically, one aims at choosing f^* such that it includes some a priori information on the solution. The above assumptions guarantee existence of the f^*-minimum-norm solution. However, since F is nonlinear, it is not necessarily unique.

When analyzing the ill-posedness of linear operators, the closedness of the range is a simple criterion that characterizes the stability of the problem. Therefore, it would be convenient if for nonlinear problems it was possible to consider the linearization of the nonlinear operator. However, in general, there is no guaranteed connection between the ill-posedness of a nonlinear problem and its linearization.

Recall that for linear operators $T : X \to Y$, one has that compactness and injectivity of T implies unboundedness of T^{-1} when X is infinite dimensional. There is a "nonlinear counterpart" to this statement: roughly, when F is compact and locally injective, then $\mathcal{D}(F)$ being "infinite dimensional around $f^\dagger$" implies the non-continuity of the inverse F^{-1}. A precise formulation is the following.

Theorem 51 *Let F be a nonlinear compact and continuous operator with $\mathcal{D}(F)$ weakly closed. Let $F(f^\dagger) = g$ and suppose there exists $\epsilon > 0$ such that $F(f) = \hat{g}$ is uniquely solvable for all $\hat{g} \in \mathcal{R}(F) \cap B_\epsilon(g)$, where $B_\epsilon(g)$ denotes the ball of radius ϵ around g.*

If there exists a sequence $(f_n)_{n\in\mathbb{N}} \subset \mathcal{D}(F)$ with

$$f_n \rightharpoonup f^\dagger, \text{ while } f_n \not\to f^\dagger, \tag{35}$$

then F^{-1} (defined on $\mathcal{R}(F) \cap B_\epsilon(g)$) is not continuous in g.

Note that assumption (35) can roughly be interpreted as infinite dimensionality of $\mathcal{D}(F)$ around $f^\dagger$: If $B_\epsilon(f^\dagger) \subset \mathcal{D}(F)$, one can take $f_n = f^\dagger + \epsilon \cdot e_n$, where e_n form a basis of X (recall that X is assumed to be separable). Then, since $e_n \rightharpoonup 0$, one has $f_n \rightharpoonup f^\dagger$ but $\|f_n - f^\dagger\|_X = \epsilon$.

We conclude this section by mentioning two standard approaches for solving nonlinear inverse problems: Tikhonov regularization and iterative methods.

In the nonlinear setting, Tikhonov regularization amounts to solving the following optimization problem:

$$\arg\min_{f\in\mathcal{D}(F)} \|F(f) - g^\delta\|_Y^2 + \alpha\|f - f^*\|_X^2, \tag{36}$$

where, as before, $g^\delta \in Y$ denotes the noisy data and α the regularization parameter. As already noted, (36) has a solution but it is not necessarily unique due to the nonlinearity of F. Thus, one just searches for *a* solution of (36), which we will denote by f_α^δ. In general, this optimization problem is non-convex and it is possible to get stuck in local minima. The following is a result on Tikhonov regularization for appropriately chosen regularization parameter α.

Theorem 52 *Let $g \in \mathcal{R}(F)$ and $g^\delta \in Y$ with $\|g^\delta - g\|_Y \leq \delta$ for $\delta > 0$. Let $\alpha(\delta)$ be chosen such that*

$$\alpha(\delta) \to 0, \quad \text{as } \delta \to 0,$$
$$\frac{\delta^2}{\alpha(\delta)} \to 0, \quad \text{as } \delta \to 0.$$

Furthermore, suppose that $(\delta_n)_{n\in\mathbb{N}}$ and $(\alpha_n)_{n\in\mathbb{N}}$ are sequences such that $\delta_n \to 0$ as $n \to \infty$ and $\alpha_n := \alpha(\delta_n)$. Then, the sequence $(f_{\alpha_n}^{\delta_n})_{n\in\mathbb{N}}$ of solutions $f_{\alpha_n}^{\delta_n}$ to (36) for $\delta = \delta_n$ and $\alpha = \alpha_n$ has a convergent subsequence. The limit of every convergent subsequence is an f^*-minimum-norm solution. If the f^*-minimum-norm solution $f^\dagger$ is unique, then

$$\lim_{\delta\to 0} f_{\alpha(\delta)}^\delta = f^\dagger.$$

We remark that for nonlinear problems, using Morozov's discrepancy principle straightforwardly for Tikhonov regularization is a bit problematic because in general,

$$\|F(f_\alpha^\delta) - g^\delta\|_Y = \delta$$

is only solvable under restrictive assumptions, and even if it is, this requires solving an additional nonlinear problem simultaneously. On the other hand, the discrepancy principle can be easily implemented for iterative methods. One can take a conventional iterative solver and often restore regularization by early stopping, where the stopping index k_* has to depend on the noise level δ. This is also true (and used) for linear problems. A stopping criterion in accordance to the discrepancy principle takes the form: stop the iteration at k_*, where

$$\|g^\delta - F(f_{k_*}^\delta)\|_Y \leq \tau\delta < \|g^\delta - F(f_k^\delta)\|_Y, \quad k < k_*,$$

for some $\tau > 1$.

6 Phase Retrieval

In this section, we discuss a nonlinear inversion problem that does not very much fit into the theory outlined in the previous section and therefore also highlights the delicacy of studying nonlinear problems. The general setup is as follows.

Let X be a separable Hilbert space and $(\varphi_\lambda)_{\lambda\in\Lambda} \subset X$ some *measurement system* with index set $\Lambda \subseteq \mathbb{C}$. Typically, the requirement on $(\varphi_\lambda)_{\lambda\in\Lambda}$ is that any $f \in X$ can be stably and uniquely recovered from $(\langle f, \varphi_\lambda\rangle_X)_{\lambda\in\Lambda}$. For Λ a discrete index set, this can be conveniently summarized as $(\varphi_\lambda)_{\lambda\in\Lambda}$ being a *discrete frame,* meaning that there exist uniform constants $A, B > 0$ such that

$$A\|f\|_X^2 \leq \sum_{\lambda\in\Lambda} |\langle f, \varphi_\lambda\rangle_X|^2 \leq B\|f\|_X^2, \quad \forall f \in X.$$

The question of *phase retrieval* can then be formulated as follows: When is it possible to uniquely (and stably) recover a function $f \in X$ from magnitude measurements

$$(|\langle f, \varphi_\lambda\rangle_X|)_{\lambda\in\Lambda}?$$

Note that uniqueness of phase retrieval always has to be understood up to a global constant phase factor, i.e., unique recovery of f amounts to finding τf, for any $\tau \in S^1$. Thus, the distance between two elements $f_1, f_2 \in X$ is defined as

$$\mathrm{dist}_X(f_1, f_2) := \inf_{\tau\in S^1} \|f_1 - \tau f_2\|_X.$$

A lot of work has been done on phase retrieval for general frames, see, e.g., [13, 14, 16], as well as for more structured measurement systems, cf. [2, 3, 11, 17–20, 23, 24, 27, 33, 36, 40, 41, 43, 46] for example. However, many questions on uniqueness and stability of phase retrieval remain open. In what follows, we want to highlight a specific phase retrieval problem with a structured measurement system, showing that even if one can (partially) answer the question of uniqueness, the problem is in some sense highly unstable and difficult to regularize.

Gabor Phase Retrieval

Let $X = L^2(\mathbb{R})$ and let the inner product on $L^2(\mathbb{R})$ (or $L^2(\mathbb{R}^2)$) simply be denoted by $\langle\cdot,\cdot\rangle$. We consider *Gabor frames,* i.e., frames that are built from a *window function* that we will choose to be the Gaussian

$$\varphi(t) := e^{-\pi t^2}$$

and its *time-frequency shifts:* for each $\lambda = x + iy \in \mathbb{C}$, which we also identify with the vector $(x, y) \in \mathbb{R}^2$, we define

$$\varphi_\lambda(t) = \varphi_{(x,y)}(t) := M_y T_x \varphi(t),$$

where T_x denotes the translation (or time shift) by $x \in \mathbb{R}$:

$$T_x f(t) := f(t - x),$$

and modulation (or frequency shift) by $y \in \mathbb{R}$ is denoted as

$$M_y f(t) := e^{2\pi i y\cdot t} f(t).$$

While there is a rich theory on Gabor frames in which stable and unique recovery of a signal f is guaranteed from measurements $(\langle f, \varphi_\lambda\rangle)_{\lambda\in\Lambda}$, Λ a *discrete* subset of $\mathbb{C}$, see, e.g., [22], we consider the best case scenario here: we assume that $\Lambda = \mathbb{C}$,

i.e., the measurements are highly *redundant* and the measurements are given by the continuous transform

$$V_\varphi f(x, y) := \int_{\mathbb{R}} f(t)\overline{\varphi(t-x)}e^{-2\pi i t\cdot y}dt = \langle f, \varphi_{(x,y)}\rangle, \quad \forall\, x, y \in \mathbb{R},$$

known as the *Gabor transform* of f. In phase retrieval, the task is to reconstruct f from

$$\left(\left|V_\varphi(x, y) f\right|\right)_{(x,y)\in\mathbb{R}^2}.$$

More precisely, define the forward operator

$$\begin{aligned}\mathcal{A}_\varphi : L^2(\mathbb{R})/S^1 &\to L^2(\mathbb{R}^2, \mathbb{R}_0^+),\\ f &\mapsto \left|V_\varphi f\right|,\end{aligned}$$

then phase retrieval amounts to the inversion of $\mathcal{A}_\varphi$.

Injectivity of Gabor Phase Retrieval

The question of uniqueness can be rather easily settled with the following fundamental formula:

$$\mathcal{F}\left(\left|V_g f\right|^2\right)(x, y) = V_f f(-y, x)\cdot\overline{V_g g(-y, x)}, \tag{37}$$

where $\mathcal{F}$ denotes the two-dimensional Fourier transform on $L^2(\mathbb{R}^2)$. Formula (37) holds, for example, when g is a Schwartz function and f is a tempered distribution [25]. Note that for $g = \varphi$, $V_\varphi\varphi$ is (up to some modulation factor) simply a two-dimensional Gaussian and therefore has no zeroes on all of $\mathbb{C}$. Thus, (37) implies that given $\left|V_\varphi f\right|$, one can recover $V_f f$ uniquely. More precisely, suppose that

$$\begin{aligned}\mathcal{F}\left(\left|V_\varphi f_1\right|^2\right)(x, y) &= V_{f_1} f_1(-y, x)\cdot\overline{V_\varphi\varphi(-y, x)},\\ \mathcal{F}\left(\left|V_\varphi f_2\right|^2\right)(x, y) &= V_{f_2} f_2(-y, x)\cdot\overline{V_\varphi\varphi(-y, x)},\end{aligned}$$

and

$$\left|V_\varphi f_1\right| = \left|V_\varphi f_2\right|.$$

Then, (37) implies

$$\left(V_{f_1} f_1 - V_{f_2} f_2\right)\cdot\overline{V_\varphi\varphi} = 0,$$

and hence $V_{f_1} f_1 = V_{f_2} f_2$. One can further show (by taking one-dimensional Fourier transforms) that

$$V_{f_1} f_1 = V_{f_2} f_2 \Rightarrow f_1 = \tau f_2, \quad \text{for some } \tau \in S^1,$$

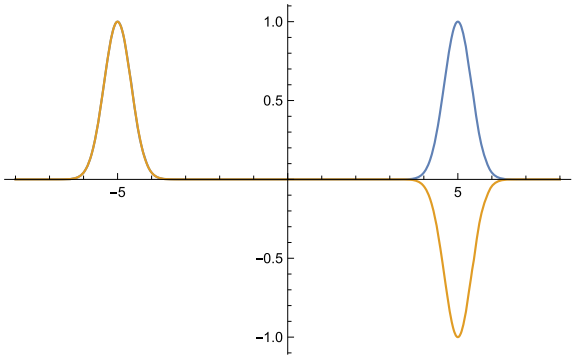


Fig. 10 The functions f_a^+ (blue) and f_a^- (orange) for $a = 5$

which guarantees uniqueness of phase retrieval from measurements $\left|V_\varphi f\right|_{(x,y)\in\mathbb{R}^2}$, cf. [25]. Note, however, that for practical purposes formula (37) is not very useful as the exponential decay in $V_\varphi\varphi$ leads to serious instabilities in the reconstruction.

Stability of Gabor Phase Retrieval

In the strict regularization-theoretic sense, phase retrieval is not an ill-posed problem because of a result in [7], which states that

$$\mathcal{A}_\varphi \text{ injective } \Rightarrow \mathcal{A}_\varphi^{-1} \text{ continuous on } \mathcal{R}(\mathcal{A}_\varphi) \text{ and } \mathcal{R}(\mathcal{A}_\varphi) \text{ closed.}$$

However, in practice, instabilities do occur. As shown in [7], the operator $\mathcal{A}_\varphi^{-1}$ is never uniformly continuous when X is infinite dimensional. More precisely, there is no uniform constant $c_1 > 0$ for which

$$c_1 \operatorname{dist}_X(f_1, f_2) \le \|\mathcal{A}_\varphi(f_1) - \mathcal{A}_\varphi(f_2)\|_{L^2(\mathbb{R}^2)}, \quad \forall\, f_1, f_2 \in X.$$

For Gabor phase retrieval, this lack of stability has been quantified to some extent in [8]. A rather simple example captures the inherent nature of instability. For this, let (f_a^+, f_a^-) be a parameter-dependent pair of functions defined as

$$f_a^+ := T_a\varphi + T_{-a}\varphi, \tag{38}$$
$$f_a^- := T_a\varphi - T_{-a}\varphi, \tag{39}$$

see Fig. 10 for a plot of (f_5^+, f_5^-). As one would expect, the Gabor transforms of such functions are almost the sum of two Gaussian bumps in the complex plane (see Fig. 11). For not too small a, the difference in magnitude of these Gabor transforms $V_\varphi f_a^+$ and $V_\varphi f_a^-$ is very small (see Fig. 12).

This causes the typical instability: while the measurements are very close, this is not true for f_a^+ and f_a^-. More precisely, for this pair of parameter-dependent functions one can show (see [8]):

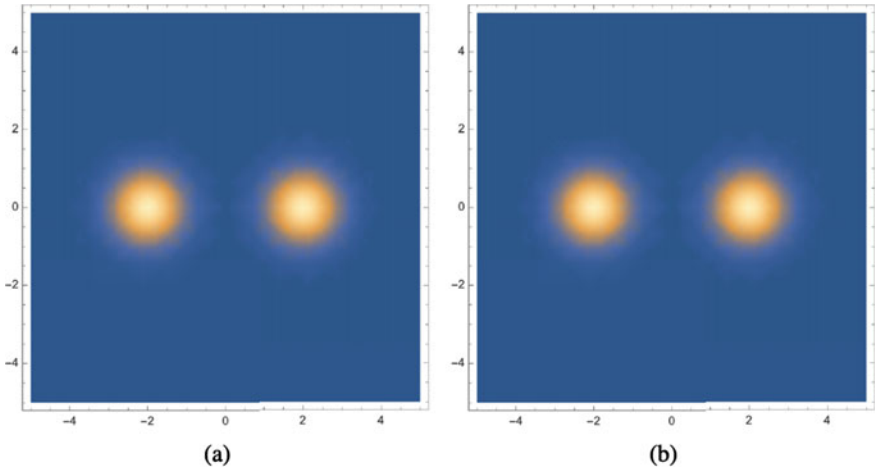
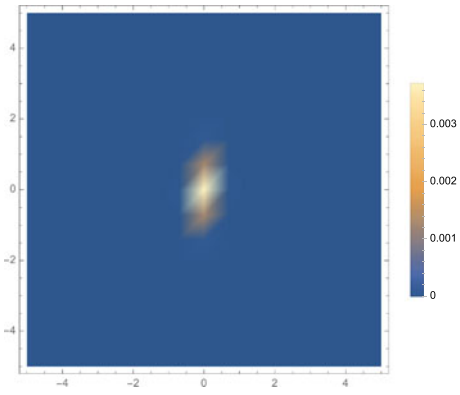


Fig. 11 Magnitudes of the Gabor transforms of f_2^+ in (**a**) and of f_2^- in (**b**)

Fig. 12 $\left|\left|V_\varphi f_a^+\right| - \left|V_\varphi f_a^-\right|\right|$ for $a = 2$



Theorem 53 *Let the functions f_a^+ and f_a^- be defined as in (38) and (39). Then there exists a uniform constant $C > 0$ such that for all $a > 0$ and for all $k \in (0, \pi/2)$:*

$$dist_{L^2(\mathbb{R})}(f_a^+, f_a^-) \geq C e^{k a^2} \left\| \left|V_\varphi f_a^+\right| - \left|V_\varphi f_a^-\right| \right\|_{W^{1,2}(\mathbb{R}^2)}. \tag{40}$$

So already in this explicit example, one has an exponential degradation of stability in the phase retrieval problem, which makes it "severely ill-posed", though not in the classical sense.

Regularization for Gabor Phase Retrieval

We conclude this section by remarking that the above example also reveals that classical regularization will be ineffective in restoring stability of the phase retrieval problem. In essence, one would aim at finding a suitable penalty $|||f|||$ such that the

problem is regularized by minimizing

$$\|\mathcal{A}_\varphi(f) - g\|^2_{L^2(\mathbb{R}^2)} + \alpha \,|||f|||\,,$$

where g denotes the measured data and α a suitably chosen regularization parameter. However, one can verify (see [8] for details) that all classical penalties, such as L^2, Besov space, modulation space norm, etc., do not resolve the occurring instability as they result in

$$|||f_a^+||| \sim |||f_a^-|||$$

for the above example.

The construction of f_a^+ and f_a^-, however, also reveals the source of instability: for these functions, the gap between the two Gaussian bumps centered at $\pm a$, cannot be bridged by the Gabor transform magnitude data in a stable way. The Gabor transform is to some extent well concentrated in both time and frequency and thus the measurements are *disconnected*, which is emphasized more strongly for larger a. This is also reflected in the bound (40) which degrades exponentially in a^2. More generally, whenever the Gabor transform of a signal is mainly concentrated on more than one region and the measurements are small outside of these regions, then phase retrieval will be unstable. We could have, for example, created similar instances to f_a^+ and f_a^- of functions that present a gap in frequency instead of time. One could also think of functions that neither have a gap in frequency nor in time, but a time-frequency gap in their respective Gabor transforms. This observation has led to proposing a novel notion of solution such that stability is restored. In [4], it has been suggested to give up on global phase reconstruction when the Gabor transform is concentrated on more than one *atoll* and is small outside of these regions: then, one would only aim at global phase reconstruction on each individual atoll since this is the best one can do stably for Gabor phase retrieval. Figure 13 illustrates the concept. This *atoll function reconstruction* indeed results in a stability estimate [4]. However, it relies heavily on φ being a Gaussian: in this case, $V_\varphi f$ is related to the Bargmann transform, which is a holomorphic function on $\mathbb{C}$. One can then argue via Cauchy–Riemann equations to obtain stability in regions where the measurements are not small. This has been further formalized in [25], with the concept of the *Cheeger constant* of the measurements describing their connectedness.

We remark that the use of such a "semi-global" phase reconstruction is justified in audio processing applications. There, constant phase factors on almost isolated regions do not audibly change the signal. See Fig. 14 for an example.

7 Instabilities in Image Classification

In the previous sections, we have taken the route of discussing linear inverse problems and the analysis and regularization that can be done for such problems, highlighting limited data reconstruction as an example. For nonlinear problems, the analysis is

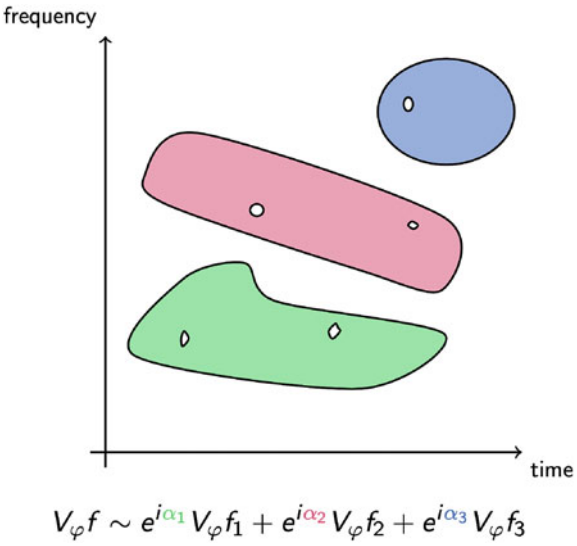


Fig. 13 The Gabor transform of f being concentrated on three regions (highlighted in color). Outside of these regions, $V_\varphi f$ is very small. We aim at reconstructing $V_\varphi f$ up to global phase on each of the three atolls individually. Inside each atoll, one can allow for small *lagoons* (highlighted in white) on which the measurements are small. The number and size of these lagoons both enter the stability estimate

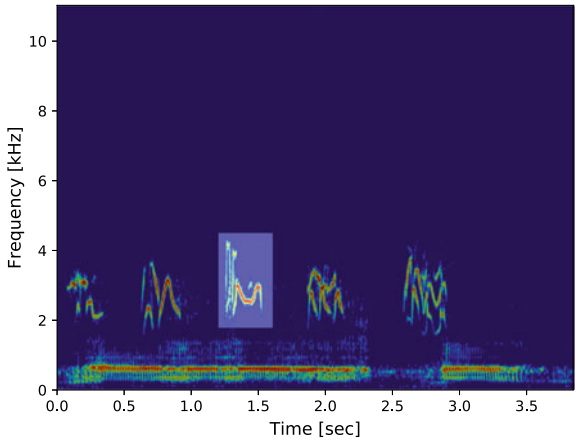


Fig. 14 Gabor transform magnitude of an audio signal containing sounds of a bird and a bison. The region highlighted in light-blue is an example of an atoll: taking the original audio signal, and an audio signal for which the highlighted region has an additional constant phase factor, audibly results in the same signal

already more cumbersome and less straightforward and we showed phase retrieval as an example that does not fit the classical regularization theory but still exhibits instabilities. To conclude this discussion, we provide one more instance of a problem that suffers from instabilities but for which no rigorous theory has been developed so far. Thus, there is no meaningful analysis that can be provided to date but one can still point out the lack of robustness with explicit constructions.

The problem we have in mind is, in a broad sense, using deep neural networks (DNNs) as function approximations. In what follows, we will discuss the problem of object classification in images, but instabilities have, for example, also been reported in using DNNs for medical imaging applications [12].

In its simplest form, image classification aims at recognizing a single object in a given image. Typically, there is a set of *labels* to choose from and the task is to assign to each image the correct label. A function that maps images to labels is called a *classifier.* Suppose that the inputs are all RGB images of fixed square size, with $P = w^2$ pixels, so that they can be described as vectors in $X := \mathbb{R}^{3P}$. Let $L \geq 2$ be the number of labels or categories. Then, a classifier is a map

$$\mathcal{K} : X \to \{1, \ldots, L\}.$$

Typically, one chooses $\mathcal{K}$ to pick the most likely label from a *feature vector,* i.e.,

$$\mathcal{K}(x) := \underset{k=1,\ldots,L}{\arg\max} F_k(x)$$

for some mapping $F : X \to \mathbb{R}^L$. State-of-the-art results are currently achieved by taking F to be a deep neural network. A simplified description of a neural network is the following.

Definition 54 A *feedforward neural network* of depth D is a mapping

$$F = F^D \circ F^{D-1} \circ \ldots \circ F^1,$$

where

$$F^d : \mathbb{R}^{n_{d-1}} \to \mathbb{R}^{n_d}, \quad x \mapsto \rho(W^d x + b^d),$$

for some $W^d \in \mathbb{R}^{n_d \times n_{d-1}}$, $b^d \in \mathbb{R}^{n_d}$ and a nonlinear *activation function* $\rho : \mathbb{R} \to \mathbb{R}$ applied element-wise to $W^d x + b^d$.

In practice, state-of-the-art networks are more sophisticated than in the above definition, but it still gives a good description of their structure: in essence, they are a repeated concatenation of affine transforms and element-wise nonlinearities.

The weight matrices W^d as well as the *bias vectors* b^d are free parameters learned during training: in *supervised learning* a training set of size m, i.e., a set of pairs of images and their correct labels $(x_j, l_j) \in X \times \{1, \ldots, L\}$, $j \in \{1, \ldots, m\}$, is given. One then aims at finding $F : X \to \mathbb{R}^L$ such that it captures the dataset distribution well enough. This is usually done by *empirical risk minimization*, where the objective is to minimize

$$\mathcal{R}\left(F, (x_j, l_j)_{j=1}^m\right) := \frac{1}{m} \sum_{j=1}^{m} J\left(F, x_j, l_j\right),$$

for some *loss function* J. In classification, the "default" is the *cross-entropy loss function* defined as

$$J_{CE}\left(F, x_j, l_j\right) := -\log(F_{l_j}(x_j)),$$

where we assume that F outputs a vector of probabilities. Otherwise, an intermediate step called *softmax layer* needs to be introduced to output probabilities from F.

While the results obtained with supervised learning and deep neural network architectures are really astonishing for large training sets, it has also been pointed out over the last decade that DNNs are vulnerable to so-called *adversarial examples*.

In their seminal work [45], Szegedy et al. demonstrate that DNNs can be rather easily "fooled" by creating small perturbations such that the perturbed image looks (almost) the same upon visual inspection, but the network will no longer be able to correctly classify the object in the image. Since then, adversarial examples have evolved to an entire field of research in machine learning, with a number of contributions that have exploded by now, making it impossible to give a fair account of the existing literature.

To illustrate the phenomenon of adversarial examples, however, we do give an example of an algorithm that searches for small perturbations such that the perturbed image is incorrectly classified. The *DeepFool algorithm* introduced in [38] can be summarized as follows.

Let $\mathcal{K} = \arg\max_{k=1,\dots,L} F$ be a trained classifier and let $x \in X$ be an image with correct label $l = \mathcal{K}(x)$. The DeepFool algorithm searches for $y = x + r$ with $\mathcal{K}(y) \neq l$ as follows:

- Choose a target label $k \neq l$.
- Set $f := F_k - F_l$, where the goal is to achieve $f(y) > 0$.
- Using $f(x + r) \approx f(x) + \nabla f(x) \cdot r$, define the perturbation

$$r := -\frac{f(x)}{\|\nabla f(x)\|_{\ell^2}^2} \nabla f(x)$$

  and set $\hat{x} = x + r$.
- If $\mathcal{K}(\hat{x}) = k$, then we are successful. Otherwise, start at the top with x replaced by $\hat{x}$.

The target label k may be selected at each iteration to minimize $\|r\|$, for some chosen norm $\|\cdot\|$.

In Fig. 15, we give an example of a correctly classified image and a slightly perturbed image, which is classified incorrectly.

We note that adversarial examples can be very effectively constructed: for state-of-the-art networks such as Inception-v3 or ResNet-101, a very high percentage of correctly classified images indeed has an adversarial example in its vicinity. It is therefore very crucial to address this problem of instability. Many *defenses* against adversarial attacks have been suggested but they mainly follow, in one way or the other, the theme of *adversarial training:* adversarial examples are incorporated in the training procedure so that the network learns these instances. While this is a quite effective method, it also has its issues. Typically, networks are adversarially trained with respect to small perturbations. However, as demonstrated in, e.g., [1], small additive perturbations are not the only type of possible attack in image classification. For example, if an image is slightly deformed, the classification output should not be changed either. However, creating *adversarial deformations* is rather straightforward.

Fig. 15 *Left:* The object is correctly classified as a ptarmigan. *Right:* A small perturbation is added, of size $\|r\|_{\ell^\infty} = 0.027$. The object is now incorrectly classified as a partridge. This example has been produced using the DeepFool algorithm on the ImageNet dataset with the Inception-v3 model

The *ADef algorithm* proposed in [1] can be described as follows: First of all, to facilitate the discussion on deformations, we model images as continuous objects, i.e., as elements of the space

$$L^2([0,1]^2, \mathbb{R}^3) = \left\{ \xi : [0,1]^2 \to \mathbb{R}^3 : \int_{[0,1]^2} |\xi(u)|^2 \, du < +\infty \right\}.$$

Given a vector field $\tau : [0,1]^2 \to \mathbb{R}^2$, we define the image after deformation by τ as

$$\xi_\tau(u) := \xi(u + \tau(u)), \quad \forall\, u \in [0,1]^2,$$

where ξ is extended by zero outside of $[0,1]^2$. In this context, the distance between ξ and ξ_τ is not well quantified by a norm of $\xi - \xi_\tau$. Instead, we measure it with a norm on τ which we define to be

$$\|\tau\|_T := \|\tau\|_{L^\infty([0,1]^2)} = \sup_{u \in [0,1]^2} \|\tau(u)\|_{\ell^2(\mathbb{R}^2)}.$$

Suppose that a discrete image $x \in X$ is a discretization of ξ on a regular grid $\{\frac{1}{w+1}, \ldots, \frac{w}{w+1}\}^2$. In return, one can build such a function ξ from x by interpolation. This way, one can make sense of a deformed image x_τ by defining it to be

$$x_\tau(s,t) = \xi\left(\frac{(s,t) + \tau(s,t)}{w+1} \right), \quad \forall\, (s,t) \in \left\{ \frac{1}{w+1}, \ldots, \frac{w}{w+1} \right\}^2.$$

To construct adversarial deformations for a classifier $\mathcal{K} = \arg\max F$ and a correctly classified image $x \in X$ with label $l = \mathcal{K}(x)$, we aim at finding a small vector field

Fig. 16 An image correctly classified as a redfox in (**a**) and a deformed version in (**b**) incorrectly classified as a shopping cart. The deformation is depicted in (**c**) and has size $\|\tau\|_T = 1.178$. This example has been created using the ADef algorithm on the ImageNet dataset with the Inception-v3 model

τ such that $l \neq \mathcal{K}(x_\tau)$. In the spirit of the DeepFool algorithm, one can again take gradient descent steps to achieve this:

- Choose a target label $k \neq l$ and set $f := F_k - F_l$.
- Define the mapping $g : \tau \mapsto f(x_\tau)$ and search for τ such that $g(\tau) > 0$.
- By linear approximation

$$g(\tau) \approx g(0) + (D_0 g)\tau,$$

with (Fréchet) derivative

$$(D_0 g)\tau = \sum_{s,t=1}^{w} A(s,t) \cdot \tau(s,t), \quad \text{with } A(s,t) := \frac{1}{w+1}$$

$$(\nabla f(x))(s,t) \nabla \xi \left(\frac{(s,t)}{w+1} \right).$$

- $(D_0 g)\tau = -g(0)$ does not have a unique solution. We can solve it in the least-squares sense:

$$\tau(s,t) = -\frac{g(0)}{\sum_{s,t=1}^{w} |A(s,t)|^2} A(s,t).$$

- Set $\hat{x}(s,t) = x((s,t) + \tau(s,t))$. If $\mathcal{K}(\hat{x}) = k$, the iteration has been successful. Otherwise repeat with x replaced by $\hat{x}$.

Figure 16 shows an example of a correctly classified image and its adversarially deformed counterpart.

First experiments suggest that networks trained against small perturbations using the celebrated projected gradient descent (PGD) algorithm [35] are less vulnerable to adversarial deformations than standard networks that have not been adversarially

trained. However, the rate of incorrectly classified images when attacked with adversarial deformations is still considerably high, thus suggesting that these adversarially trained networks are far from being robust.

Creating DNNs that are truly robust with respect to *all* types of invariances that we expect (such as rotations, translations, small perturbations, small deformations, etc.) is important, but still an open problem.

References

1. Alaifari, R., Alberti, G.S., Gauksson, T.: ADef: an iterative algorithm to construct adversarial deformations. In: International Conference on Learning Representations (ICLR) (2019)
2. Alaifari, R., Bartolucci, F., Wellershoff, M.: Phase retrieval of bandlimited functions for the wavelet transform (2020). arXiv:2009.05029
3. Alaifari, R., Daubechies, I., Grohs, P., Thakur, G.: Reconstructing real-valued functions from unsigned coefficients with respect to wavelet and other frames. J. Fourier Anal. Appl. **23**(6), 1480–1494 (2017)
4. Alaifari, R., Daubechies, I., Grohs, P., Yin, R.: Stable phase retrieval in infinite dimensions. Found. Comput. Math. **19**(4), 869–900 (2019)
5. Alaifari, R., Defrise, M., Katsevich, A.: Asymptotic analysis of the SVD for the truncated Hilbert transform with overlap. SIAM J. Math. Anal. **47**(1), 797–824 (2015)
6. Alaifari, R., Defrise, M., Katsevich, A.: Stability estimates for the regularized inversion of the truncated Hilbert transform. Inverse Probl. **32**(6), 065,005 (2016)
7. Alaifari, R., Grohs, P.: Phase retrieval in the general setting of continuous frames for Banach spaces. SIAM J. Math. Anal. **49**(3), 1895–1911 (2017)
8. Alaifari, R., Grohs, P.: Gabor phase retrieval is severely ill-posed. Appl. Comput. Harmon. Anal. **50**, 401–419 (2021)
9. Alaifari, R., Katsevich, A.: Spectral analysis of the truncated Hilbert transform with overlap. SIAM J. Math. Anal. **46**(1), 192–213 (2014)
10. Alaifari, R., Pierce, L.B., Steinerberger, S.: Lower bounds for the truncated Hilbert transform. Revista Matemática Iberoamericana **32**(1), 23–56 (2016)
11. Alaifari, R., Wellershoff, M.: Uniqueness of STFT phase retrieval for bandlimited functions. Appl. Comput. Harmon. Anal. **50**, 34–48 (2021)
12. Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A.C.: On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc. Natl. Acad. Sci. **117**(48), 30088–30095 (2020)
13. Balan, R., Casazza, P., Edidin, D.: On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**(3), 345–356 (2006)
14. Bandeira, A.S., Cahill, J., Mixon, D.G., Nelson, A.A.: Saving phase: injectivity and stability for phase retrieval. Appl. Comput. Harmon. Anal. **37**(1), 106–125 (2014)
15. Bertero, M., De Mol, C., Viano, G.A.: The stability of inverse problems. Inverse Scattering Problems in Optics, pp. 161–214. Springer, Berlin (1980)
16. Cahill, J., Casazza, P., Daubechies, I.: Phase retrieval in infinite-dimensional Hilbert spaces. Trans. Am. Math. Soc., Ser. B **3**(3), 63–76 (2016)
17. Candès, E.J., Eldar, Y.C., Strohmer, T., Voroninski, V.: Phase retrieval via matrix completion. SIAM Rev. **57**(2), 225–251 (2015)
18. Candes, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: theory and algorithms. IEEE Trans. Inf. Theory **61**(4), 1985–2007 (2015)
19. Chen, Y., Cheng, C., Sun, Q., Wang, H.: Phase retrieval of real-valued signals in a shift-invariant space. Appl. Comput. Harmon. Anal. **49**(1), 56–73 (2020)

20. Cheng, C., Daubechies, I., Dym, N., Lu, J.: Stable phase retrieval from locally stable and conditionally connected measurements (2020). arXiv:2006.11709
21. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems, vol. 375. Springer Science & Business Media (1996)
22. Gröchenig, K.: Foundations of Time-Frequency Analysis. Birkhäuser, Basel (2001)
23. Gröchenig, K.: Phase-retrieval in shift-invariant spaces with gaussian generator. J. Fourier Anal. Appl. **26**(3), 1–15 (2020)
24. Grohs, P., Liehr, L.: Injectivity of Gabor phase retrieval from lattice measurements (2020). arXiv:2008.07238
25. Grohs, P., Rathmair, M.: Stable Gabor phase retrieval and spectral clustering. Commun. Pure Appl. Math. **72**(5), 981–1043 (2019)
26. Grünbaum, F., Longhi, L., Perlstadt, M.: Differential operators commuting with finite convolution integral operators: some nonabelian examples. SIAM J. Appl. Math. **42**(5), 941–955 (1982)
27. Han, D., Juste, T.: Phase-retrievable operator-valued frames and representations of quantum channels. Linear Algebra Appl. **579**, 148–168 (2019)
28. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods For Nonlinear Ill-Posed Problems, vol. 6. Walter de Gruyter, Berlin (2008)
29. Katsevich, A.: Singular value decomposition for the truncated Hilbert transform. Inverse Probl. **26**(11), 115, 011 (2010)
30. Katsevich, A.: Singular value decomposition for the truncated Hilbert transform: part II. Inverse Probl. **27**(7), 075, 006 (2011)
31. Landau, H.J., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis and uncertainty–II. Bell Syst. Tech. J. **40**(1), 65–84 (1961)
32. Landau, H.J., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis and uncertainty–III: the dimension of the space of essentially time-and band-limited signals. Bell Syst. Tech. J. **41**(4), 1295–1336 (1962)
33. Li, L., Juste, T., Brennan, J., Cheng, C., Han, D.: Phase retrievable projective representation frames for finite abelian groups. J. Fourier Anal. Appl. **25**(1), 86–100 (2019)
34. Maass, P.: The interior Radon transform. SIAM J. Appl. Math. **52**(3), 710–724 (1992)
35. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2017). arXiv:1706.06083
36. Mallat, S., Waldspurger, I.: Phase retrieval for the Cauchy wavelet transform. J. Fourier Anal. Appl. **21**(6), 1251–1309 (2015)
37. Miller, K.: Least squares methods for ill-posed problems with a prescribed bound. SIAM J. Math. Anal. **1**(1), 52–74 (1970)
38. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
39. Natterer, F.: The Mathematics of Computerized Tomography. SIAM (2001)
40. Pfander, G.E., Salanevich, P.: Robust phase retrieval algorithm for time-frequency structured measurements. SIAM J. Imag. Sci. **12**(2), 736–761 (2019)
41. Pohl, V., Yang, F., Boche, H.: Phaseless signal recovery in infinite dimensional spaces using structured modulations. J. Fourier Anal. Appl. **20**(6), 1212–1233 (2014)
42. Reed, M., Simon, B.: Methods of Modern Mathematical Physics: Vol. 1: Functional analysis. Academic, Cambridge (1980)
43. Romero, J.L.: Sign retrieval in shift-invariant spaces with totally positive generator (2020). arXiv:2005.08678
44. Slepian, D., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis and uncertainty–I. Bell Syst. Tech. J. **40**(1), 43–63 (1961)
45. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2013). arXiv:1312.6199
46. Thakur, G.: Reconstruction of bandlimited functions from unsigned samples. J. Fourier Anal. Appl. **17**(4), 720–732 (2011)
47. Zettl, A.: Sturm-Liouville Theory, vol. 121. American Mathematical Society, Providence (2010)

Proximal Gradient Methods for Machine Learning and Imaging

Saverio Salzo and Silvia Villa

1 Introduction

Convex optimization plays a key role in data science and image processing. Indeed, from one hand it provides theoretical frameworks, such as duality theory and the theory of nonexpansive operators, which are indispensable to formally analyze many problems arising in those fields. On the other hand, convex optimization supplies a plethora of algorithmic solutions covering a broad range of applications. In particular, the last decades witnessed an unprecedented development of optimization methods which are now capable of addressing structured and large-scale problems effectively. An important class of such methods, which are at the core of modern nonlinear convex optimization, is that of proximal gradient splitting algorithms. They are first-order methods which are tailored to optimization problems having a composite structure given by the sum of smooth and nonsmooth terms. These methods are splitting algorithms, in the sense that along the iterations they process each term separately by exploiting gradient information when available and the so-called proximity operator for nonsmooth terms.

Even though there is a rich literature on proximal gradient algorithms, in this contribution, we paid particular attention to presenting a self-contained and unifying analysis for the various algorithms, unveiling common theoretical basis. We give state-of-the-art results treating both convergence of the iterates and of objective functions values in infinite-dimensional setting. This work is based on the lecture

S. Salzo (✉)
Istituto Italiano di Tecnologia, Via E. Melen 83, 16152 Genova, Italy
e-mail: saverio.salzo@iit.it

S. Villa
DIMA & MaLGa Center, Università degli Studi di Genova, Via Dodecaneso 35, 16146 Genova, Italy
e-mail: silvia.villa@unige.it

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8_4

notes written for the PhD course "Introduction to Convex Optimization" that was given by the authors at the University of Genoa during the last 5 years.

This chapter is divided into six sections. Section 2 provides an account on convex analysis, recalling the fundamental concepts of subdifferentials, Legendre–Fenchel transform, and duality theory. In Sect. 3, we study the proximal gradient algorithm under different assumptions, addressing also acceleration techniques. Section 4 is about stochastic optimization methods. We study the projected stochastic subgradient method, the proximal stochastic gradient algorithm and the randomized block-coordinate proximal gradient algorithm. Section 5 exploits duality to derive new algorithms. Finally, in Sect. 6, we describe several important applications in which proximal gradient algorithms has been successfully used.

2 Preliminaries on Convex Analysis

2.1 Basic Notations

We set $\mathbb{R}_+ = \{\alpha \in \mathbb{R} \,|\, \alpha \geq 0\}$ and $\mathbb{R}_{++} = \{\alpha \in \mathbb{R} \,|\, \alpha > 0\}$. Throughout the chapter, X is a real Hilbert space and its associated *scalar product* and *norm* is denoted by

$$\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R} \quad \text{and} \quad \|\cdot\| : X \to \mathbb{R}.$$

An *affine* set of X is a set $M \subset X$ such that every straight line joining two distinct points of M is contained in M. In formula this means that, for every $x, y \in M$, and every $\lambda \in \mathbb{R}$, we have $(1-\lambda)x + \lambda y \in M$. If M is affine then $V := M - M$ is a vector subspace of X, which is called the *direction* of M. Moreover, we have $M = V + x$, for every $x \in M$. The intersection of a family of affine sets of X is still affine, so if $C \subset X$ one can define the *affine hull* of C, denoted by $\mathrm{aff}(C)$, which is the intersection of all the affine sets of X containing C. It can be represented as the set of the finite *affine combinations* of elements of C, meaning that $x \in \mathrm{aff}(C)$ if and only if there exists finite number of points $x_1, \ldots, x_n \in C$ and numbers $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ ($n \geq 1$) such that $\sum_{i=1}^n \lambda_i = 1$ and $x = \sum_{i=1}^n \lambda_i x_i$. The *affine dimension* of a set C is the dimension of the affine hull of C. A mapping $T : X \to Y$ between Hilbert spaces is said to be *affine* if $T((1-\lambda)x + \lambda y) = (1-\lambda)Tx + \lambda Ty$, for every $x, y \in X$ and $\lambda \in \mathbb{R}$. An affine mapping T can be uniquely represented as $Tx = Ax + b$ with $A : X \to Y$ be a linear operator and $b \in Y$. The image and the counter image of affine sets through affine mappings are affine sets. An *(affine) hyperplane of* X is a set of the form $\{x \in X \,|\, \varphi(x) = \alpha\}$, where $\varphi : X \to \mathbb{R}$ is a nonzero linear form on X and $\alpha \in \mathbb{R}$.

For every $x \in X$ and every $\delta > 0$ we denote by $B_\delta(x)$ the (closed) ball of center x and radius δ, that is $B_\delta(x) = \{y \in X \,|\, \|y - x\| \leq \delta\}$. Given a subset $C \subset X$, we denote by $\mathrm{int}(C)$, $\mathrm{cl}(C)$, and $\mathrm{bdry}(C)$ its interior, closure and boundary, respectively. An hyperplane $H = \{x \in X \,|\, \varphi(x) = \alpha\}$ is closed if and only if φ is a continuous

linear form on X so that it can be represented as $H = \{x \in X \mid \langle x, u\rangle = \alpha\}$ with $u \in X \setminus \{0\}$. A sequence $(x_k)_{k\in\mathbb{N}}$ in X *converges* to $x \in X$, and we write $x_k \to x$, if $\|x_k - x\| \to 0$, whereas it *weakly converges* to x, and we write $x_k \rightharpoonup x$, if for every $u \in X$, $\langle x_k - x, u\rangle \to 0$. A subset $C \subset X$ is *weakly sequentially closed* if the weak limit of every *weakly convergent* sequence in C belongs to C.

Classically, in optimization, functions and constraints are treated separately. By introducing extended real-valued functions, they can be treated in a unified way. Here with *extended real-valued functions*, we mean functions

$$f\colon X \to \,]-\infty, +\infty],$$

so that the value $-\infty$ will never be allowed. In the rest of the chapter, if not otherwise specified, functions are supposed to be extended real-valued. The *(effective) domain of* f is the set $\operatorname{dom} f = \{x \in X \mid f(x) < +\infty\}$ and the *epigraph of* f is the set

$$\operatorname{epi}(f) = \{(x, t) \in X \times \mathbb{R} \mid f(x) \le t\}. \tag{1}$$

Note that $\operatorname{epi}(f)$ is a subset of $X \times \mathbb{R}$. We also define the *sublevel sets* of f as

$$[f \le t] = \big\{x \in X \mid f(x) \le t\big\}, \qquad t \in \mathbb{R}, \tag{2}$$

and similarly, we define the sets $[f > t]$. An extended real-valued function is called *proper* if $\operatorname{dom} f \neq \varnothing$, meaning that the function admits at least a finite value. The set of minimizers of f is denoted by $\operatorname{argmin} f$.

In optimization problems, extended real-valued functions allow to treat constraints as functions. Indeed let $C \subset X$ and define the *indicator function of* C as

$$\iota_C\colon X \to \,]-\infty, +\infty] : x \mapsto \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C. \end{cases} \tag{3}$$

Then the constrained minimization problem

$$\min_{x\in C} h(x), \qquad h\colon X \to \mathbb{R}$$

can be equivalently written as

$$\min_{x\in X} f(x), \qquad f\colon X \to \,]-\infty, +\infty], \;\; f(x) = h(x) + \iota_C(x).$$

Note that indicator functions and epigraphs allow to establish a one to one correspondence between extended real-valued functions and sets.

2.2 Convex Sets and Functions

A subset $C \subset X$ is said to be *convex* if

$$(\forall\, x, y \in C)(\forall\, \lambda \in [0,1]) \quad (1-\lambda)x + \lambda y \in C, \tag{4}$$

meaning that for every $x, y \in C$, the *segment* $[x, y] = \big\{x + \lambda(y-x) \mid \lambda \in [0,1]\big\}$, joining x and y, is contained in C. A *cone* of X is a subset $C \subset X$ such that

$$(\forall\, x \in C)(\forall\, \lambda \in \mathbb{R}_{++}) \quad \lambda x \in C, \tag{5}$$

meaning that, for every $x \in C$ the *ray* $\mathbb{R}_{++}x = \{\lambda x \mid \lambda \in \mathbb{R}_{++}\}$ is contained in C. The intersection of a family of convex sets of X is still convex, so if $A \subset X$, then one defines the *convex hull of* A, denoted by $\mathrm{co}(A)$, as the intersection of the family of all convex subsets of X containing A. In fact it is the smallest convex subset of X containing A and it can be represented as the set of the finite *convex combinations* of elements of A, meaning that $x \in \mathrm{aff}(A)$ if and only if there exists finite number of points $x_1, \ldots, x_n \in A$ and numbers $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ $(n \geq 1)$ such that $\sum_{i=1}^n \lambda_i = 1$ and $x = \sum_{i=1}^n \lambda_i x_i$.

Let C be a nonempty closed convex subset of X and let $x \in X$. Then the *orthogonal projection* of x onto C is defined as the unique point $p \in C$ such that, for every $y \in C$, $\|p - x\| \leq \|y - x\|$ and is denoted by $P_C(x)$. It is also characterized by the following variational inequality

$$(\forall\, y \in C)\ \langle y - p, x - p\rangle \leq 0.$$

If C is an affine set with direction V, then the above characterization becomes the classical $x - p \in V^\perp$. We recall that for convex sets the property of being closed is equivalent to that of being weakly sequentially closed. We finally recall that the projection operator $P_C \colon X \to X$ is *firmly nonexpansive*, that is, it satisfies

$$(\forall\, x \in X)(\forall\, y \in X) \quad \|P_C(x) - P_C(y)\|^2 \leq \langle P_C(x) - P_C(y), x - y\rangle\,. \tag{6}$$

An extended real-valued function $f \colon X \to\,]-\infty, +\infty\,]$ is *convex* if

$$(\forall\, x, y \in X)(\forall\, \lambda \in [0,1]) \quad f((1-\lambda)x + \lambda y) \leq (1-\lambda) f(x) + \lambda f(y) \tag{7}$$

and is *strictly convex* if in (7) the strict inequality holds when $x, y \in \mathrm{dom}\, f$, $x \neq y$ and $\lambda \in\,]0,1[$. Finally, $g \colon X \to [-\infty, +\infty[$ is *concave* (risp. *strictly concave*) if $-g$ is convex (risp. strictly convex). If f is convex, by induction, definition (7) yields *Jensen's inequality*, that is, for every finite sequence $(x_i)_{1 \leq i \leq m}$ in X and every $(\lambda_i)_{1 \leq i \leq m} \in \mathbb{R}_+^m$ such that $\sum_{i=1}^m \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^{m}\lambda_i x_i\right) \leq \sum_{i=1}^{m}\lambda_i f(x_i). \tag{8}$$

The property of convexity for a function $f\colon X \to\,]-\infty, +\infty]$ is equivalent to the fact that its epigraph $\mathrm{epi}(f)$ is a convex set in $X \times \mathbb{R}$. The function f is *strongly convex* if there exists $\mu > 0$ such that, for every $x, y \in X$ and every $\lambda \in [0, 1]$,

$$f\big((1-\lambda)x + \lambda y\big) \leq (1-\lambda)f(x) + \lambda f(y) - \frac{\mu}{2}(1-\lambda)\lambda\|x-y\|^2. \tag{9}$$

In such case, μ is called the *modulus of strong convexity* of f and the function f is also said to be *μ-strongly convex*. It is easy to see that a function $f\colon X \to\,]-\infty, +\infty]$ is μ-strongly convex if and only if $f - (\mu/2)\|\cdot\|^2$ is convex. Moreover, strongly convex functions admitting a minimizer, say x_*, satisfies the following *quadratic growth* condition

$$(\forall\, x \in X) \qquad f(x) - f(x_*) \geq \frac{\mu}{2}\|x - x_*\|^2. \tag{10}$$

The function $f\colon X \to\,]-\infty, +\infty]$ is *lower semicontinuous* if for every sequence $(x_k)_{k\in\mathbb{N}}$ in X and every $x \in X$, $x_k \to x \;\Rightarrow\; f(x) \leq \liminf_k f(x_k)$. This property is equivalent to the closeness of $\mathrm{epi}(f)$ in $X \times \mathbb{R}$. We denote by $\Gamma_0(X)$ the class of functions $f\colon X \to\,]-\infty, +\infty\,]$ which are proper convex and lower semicontinuous. Such functions are continuous on the interior of their domain. When existence of minimizers is in order, the following definition is needed. The proper function $f\colon X \to\,]-\infty, +\infty]$ is said *coercive* if

$$\lim_{\|x\|\to+\infty} f(x) = +\infty,$$

which is equivalent to say that, for every $\alpha \in \mathbb{R}$, $[f \leq \alpha]$ is bounded. A proper, convex lower semicontinuous and coercive function admits a global minimizer and if the function is strictly convex the minimizer is unique.

2.3 Differentiability and Convexity

We recall the definition of differentiable functions. Let $f\colon X \to\,]-\infty, +\infty]$ be a proper extended real-valued function and let $x_0 \in \mathrm{int}(\mathrm{dom}\, f)$. Then f *is Gâteaux differentiable at* x_0 if there exists a vector $\nabla f(x_0) \in X$ such that

$$(\forall\, v \in X) \qquad \lim_{t\to 0}\frac{f(x_0 + tv) - f(x_0)}{t} = \langle v, \nabla f(x_0)\rangle\,. \tag{11}$$

In such case $\nabla f(x_0)$ is called the *gradient of* f *at* x_0 and f admits *directional derivatives* at x_0 in every direction v and the directional derivatives depend linearly

and continuously on v. When f is Gâteaux differentiable at every point of a subset $A \subset \operatorname{int}(\operatorname{dom} f)$ we say that f *is Gâteaux differentiable on* A.

When $\operatorname{dom} f$ is open and f is differentiable on $\operatorname{dom} f$, convexity is characterized by the monotonicity of the gradient operator, i.e., that $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq 0$, for every $x, y \in X$. Similarly, the strong convexity of f is equivalent to the strong monotonicity of the gradient operator, that is,

$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f) \quad \langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \mu \|x - y\|^2. \tag{12}$$

A function $f\colon X \to \mathbb{R}$ is *Lipschitz smooth* if it is Gâteaux differentiable on X and its gradient is Lipschitz continuous. The following result provides several characterizations of Lipschitz smoothness that will be useful in analyzing proximal gradient methods. The implication (i) $\Rightarrow$ (ii) is called the *descent lemma*, whereas the implication (i) $\Rightarrow$ (iv) is called the *Baillon–Haddad theorem*.

Fact 1 Let $f\colon X \to \mathbb{R}$ be a convex differentiable function and let $L \in \mathbb{R}_+$. The following statements are equivalent.

(i) $(\forall x \in X)(\forall y \in X) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

(ii) $(\forall x \in X)(\forall y \in X) \quad f(y) - f(x) - \langle y - x, \nabla f(x)\rangle \leq \frac{L}{2}\|x - y\|^2$.

(iii) $(\forall x \in X)(\forall y \in X) \quad \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle y - x, \nabla f(x)\rangle$

(iv) $(\forall x \in X)(\forall y \in X) \quad \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle x - y, \nabla f(x) - \nabla f(y)\rangle$

(v) $(\forall x \in X)(\forall y \in X) \quad \langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|^2$.

(vi) $\frac{L}{2}\|\cdot\|^2 - f$ is convex.

In case f is twice differentiable on X, the previous statements are equivalent to

(vii) $(\forall x \in X)(\forall v \in X) \quad \langle \nabla^2 f(x)v, v\rangle \leq L\|v\|^2$.

(viii) $(\forall x \in X) \quad \|\nabla^2 f(x)\| \leq L$.

Fact 2 Let $f\colon X \to \mathbb{R}$ be a differentiable function. Then the following are equivalent

(i) f is μ-strongly convex and ∇f is Lipschitz continuous with constant L.

(ii) $\forall x, y \in X, \ \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{L+\mu}\|x - y\|^2 \leq \langle x - y, \nabla f(x) - \nabla f(y)\rangle$.

2.4 Calculus for Nonsmooth Convex Functions

In this section, we recall the concept of subdifferentials and calculus for nonsmooth convex functions. Let $f\colon X \to]-\infty, +\infty]$ be a proper convex function and $x \in \operatorname{dom} f$. The directional derivative of f at x along the vector v is $f'(x, v) = \lim_{t \to 0^+} (f(x + tv) - f(x))/t$. The *subdifferential of* f *at* x is defined as

$$\partial f(x) := \big\{u \in X \,\big|\, (\forall\, y \in X)\ f(y) \geq f(x) + \langle y - x, u\rangle\big\}. \tag{13}$$

Each element of $\partial f(x)$ is called a *subgradient of* f *at* x. If $x \notin \operatorname{dom} f$, by definition, $\partial f(x) = \varnothing$. Finally, the domain of ∂f, denoted by $\operatorname{dom}\partial f$, is defined as the set of points at which the subdifferential is nonempty. It is easy to see that the subdifferential ∂f is a *monotone* operator, that is, for every $x, y \in X$ and $u \in \partial f(x)$, $v \in \partial f(y)$ $\langle x - y, u - v\rangle \geq 0$. If f is Gâteaux differentiable at $x \in \operatorname{int}(\operatorname{dom} f)$, then $\partial f(x) = \{\nabla f(x)\}$. Let $C \subset X$ be a nonempty convex set and let $x \in C$. The set $\partial \iota_C(x)$ is called the *normal cone to* C *at* x and it is denoted by $N_C(x)$, that is,

$$N_C(x) = \big\{u \in X \,|\, (\forall\, y \in C)\ \langle y - x, u\rangle \leq 0\big\}. \tag{14}$$

We have the following important facts

Fact 3 (Fermat's rule) Let $f\colon X \to\,]-\infty, +\infty]$ be a proper convex function and $x \in \operatorname{dom} f$. Then the following are equivalent

(i) x is a minimizer of f;
(ii) $0 \in \partial f(x)$;
(iii) $(\forall\, y \in X)\ f'(x, y - x) \geq 0$;
(iv) $(\forall\, y \in \operatorname{dom} f)\ f'(x, y - x) \geq 0$,

Fact 4 Let $f \in \Gamma_0(X)$ be μ-strongly convex and $x, u \in X$. Then

$$u \in \partial f(x) \Longleftrightarrow \forall\, y \in X\ f(y) \geq f(x) + \langle y - x, u\rangle + \frac{\mu}{2}\|x - y\|^2.$$

Fact 5 (Moreau–Rockafellar) Let $f \in \Gamma_0(X)$, $g \in \Gamma_0(Y)$, and $A\colon X \to Y$ be a continuous linear operator and suppose that $0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$. Then,

$$(\forall\, x \in X) \qquad \partial(f + g \circ A)(x) = \partial f(x) + A^* \partial g(Ax). \tag{15}$$

In particular, if g is Gâteaux differentiable at $x \in \operatorname{int}(\operatorname{dom} g)$, then $\partial(f + g)(x) = \partial f(x) + \{\nabla g(x)\}$.

Fact 6 Let $(X_i)_{1 \leq i \leq m}$ be m Hilbert spaces and let $X = \bigoplus_{i=1}^m X_i$ be their direct product, endowed with the scalar product $\langle x, y\rangle = \sum_{i=1}^m \langle x_i, y_i\rangle$. Let $(f_i)_{1 \leq i \leq m}$ be a family of proper convex functions, $f_i\colon X_i \to\,]-\infty, +\infty]$ and define

$$f\colon X \to\,]-\infty, +\infty], \qquad f(x) = f_1(x_1) + f_2(x_2) + \cdots, +f_m(x_m).$$

So the function f is separable. Then, for all $x \in \operatorname{dom} f = \prod_{i=1}^m \operatorname{dom} f_i$, we have

$$\partial f(x) = \partial f_1(x_1) \times \partial f_2(x_2) \times \cdots \times \partial f_m(x_m).$$

Example 7 Let us consider the case of the ℓ^1-norm on $\mathbb{R}^d$, that is, $\|x\|_1 = \sum_{i=1}^d |x_i|$. Since $\|\cdot\|_1$ is clearly separable with components $|\cdot|$, it follows from Fact 6 that

$$\partial \|\cdot\|_1(x) = \partial|\cdot|(x_1) \times \cdots \times \partial|\cdot|(x_d).$$

Fact 8 Let $(f_i)_{i\in I}$ be a finite family of continuous affine functions on X, say $f_i = \langle \cdot, u_i\rangle + \alpha_i$, for some $u_i \in X$ and $\alpha_i \in \mathbb{R}$. Let $f = \max_{i\in I} f_i$, let $x \in X$ and set $I(x) = \{i \in I \mid f_i(x) = f(x)\}$. Then

$$\partial f(x) = \mathrm{co}\{u_i \mid i \in I(x)\}. \tag{16}$$

2.5 The Legendre–Fenchel Transform

Let $f\colon X \to]-\infty, +\infty]$ be proper. The function

$$f^*\colon X \to]-\infty, +\infty], \quad f^*(u) = \sup_{x\in X} \langle x, u\rangle - f(x)$$

is called the *Fenchel* conjugate of f, which is always convex and lower semicontinuous. The *Fenchel–Moreau theorem* ensures that if $f \in \Gamma_0(X)$ then $f^* \in \Gamma_0(X)$ and $f^{**} = f$. Thus, the transformation $\cdot^*\colon \Gamma_0(X) \to \Gamma_0(X)$ is an involution, which is called the *Legendre–Fenchel transform*. Let $C \subset X$. The *support function* of C is the function ι_C^*, which is denoted by σ_C, that is, $\sigma_C(u) = \sup_{x\in C} \langle x, u\rangle$.

Fact 9 (Properties of the conjugate operation) Let $f\colon X \to]-\infty, +\infty]$ be a proper function. Then the following hold.

(i) Let $g\colon X \to]-\infty, +\infty]$ be a proper function. Then $f \le g \;\Rightarrow\; f^* \ge g^*$.
(ii) Let $\gamma > 0$. Then, for every $u \in X$, $(\gamma f)^*(u) = \gamma f^*(u/\gamma)$.
(iii) (The conjugate of a separable function is separable). Under the same assumptions of Fact 6, we have

$$\forall\, u = (u_1, \ldots, u_m) \in X \qquad f^*(u) = f_1^*(u_1) + f_2^*(u_2) + \cdots + f_m^*(u_m).$$

(iv) $[f(\cdot - x_0)]^* = f^* + \langle x_0, \cdot\rangle$ and $[f + \langle \cdot, u_0\rangle]^* = f^*(\cdot - u_0)$, for $x_0, u_0 \in X$.
(v) Let $x_0 \in X$. Then $\iota_{\{x_0\}}^* = \langle x_0, \cdot\rangle$.

Example 10 Let $f\colon X \to]-\infty, +\infty]$ be proper function. Then the following hold.

(i) If $f = (1/2)\|\cdot\|^2$, then $f^* = (1/2)\|\cdot\|^2$.
(ii) Let $\varphi\colon \mathbb{R} \to]-\infty, +\infty]$ be an even function. Then $[\varphi \circ \|\cdot\|]^* = \varphi^* \circ \|\cdot\|$.
(iii) Suppose that f is positively homogeneous. Then, $f^* = \iota_{\partial f(0)}$. Recall that $\partial f(0)$ is a closed convex cone.

Fact 11 Let $f\colon X \to]-\infty, +\infty]$ be proper and convex and let $x, u \in X$. Then, the following holds.

(i) $\langle x, u\rangle \le f(x) + f^*(u)$ *(Young–Fenchel inequality).*

(ii) $\langle x, u\rangle = f(x) + f^*(u) \;\Leftrightarrow\; u \in \partial f(x)$.
(iii) If $f \in \Gamma_0(X)$, then $u \in \partial f(x) \;\Leftrightarrow\; x \in \partial f^*(u)$.

Fact 12 Let $f \in \Gamma_0(X)$ be strongly convex. Then f is supercoercive, i.e., $f(x)/\|x\| \to +\infty$ as $\|x\| \to +\infty$.

Fact 13 Let $f \in \Gamma_0(X)$ and $\mu > 0$. Then, if f is μ-strongly convex, we have

(a) $\operatorname{dom} f^* = X$, f^* is differentiable on X and ∇f^* is $(1/\mu)$-Lipschitz continuous.

Vice versa if (a) holds, then f is μ-strongly convex on the convex subsets of $\operatorname{dom}\partial f$.

2.6 The Fenchel–Rockafellar Duality

Duality plays a key role in convex optimization. Here we recall the Fenchel–Rockafellar duality. We let $A\colon X \to Y$ be a continuous linear operator between Hilbert spaces, $f \in \Gamma_0(X)$ and $g \in \Gamma_0(Y)$. Consider the problem

$$\min_{x\in X} f(x) + g(Ax) =: \Phi(x). \tag{$\mathscr{P}$}$$

Its *dual problem* (in the sense of Fenchel–Rockafellar) is

$$\min_{u\in Y} f^*(-A^*u) + g^*(u) =: \Psi(u). \tag{$\mathscr{D}$}$$

One can prove that

$$(\forall\, x \in X)(\forall u \in Y) \quad \Phi(x) \geq -\Psi(u), \tag{17}$$

hence

$$\inf_{x\in X} \Phi(x) \geq \sup_{u\in Y} -\Psi(u) = -\inf_{u\in Y} \Psi(u). \tag{18}$$

This means that the function Φ is (uniformly) above the function $-\Psi$ (which is concave). The difference between the infimum of Φ and the supremum of $-\Psi$, that is $\inf \Phi + \inf \Psi$, is called *the duality gap* and we say that *strong duality* holds if the duality gap is zero.[1]

Let $S = \operatorname{argmin} \Phi$ and $S^* = \operatorname{argmin} \Psi$. Then the following are equivalent.

(i) $\hat{x} \in S$, $\hat{u} \in S^*$, and $\inf_X \Phi + \inf_Y \Psi = 0$ (duality gap is zero);
(ii) $\hat{x} \in \partial f^*(-A^*\hat{u})$ and $A\hat{x} \in \partial g^*(\hat{u})$
(iii) $-A^*\hat{u} \in \partial f(\hat{x})$ and $\hat{u} \in \partial g(A\hat{x})$.

[1] Note that if $\inf \Phi = -\infty$, it follows from (18) that $\inf \Phi = \sup(-\Psi) = -\inf \Psi = -\infty$. In this case, $\Psi \equiv +\infty$ and $\inf \Phi + \inf \Psi = -\infty + \infty$ does not make sense. Anyway, since there is no gap between Φ and $-\Psi$, by convention, we set $\inf \Phi + \inf \Psi = 0$. The same situation occurs if $\inf \Psi = -\infty$.

The conditions (ii) and (iii) above are called *KKT (Karush–Kuhn–Tucker) conditions*. Once one ensures that strong duality holds (that is, $\inf \Phi + \inf \Psi = 0$) they provide fully characterizations for a couple $(\hat{x}, \hat{u})$ to be a primal and dual solution.

Fact 14 Suppose that one of the following conditions is satisfied.

(a) $S \neq \varnothing$ and $\partial(f + g \circ A) = \partial f + A^* \partial g A$
(b) $0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$.

Then Φ is proper and

$$\inf_X \Phi = -\min_Y \Psi, \tag{19}$$

meaning that $S^* \neq \varnothing$ and $\inf_X \Phi + \inf_Y \Psi = 0$.

Example 15 *(Equality constraints)* We consider the problem

$$\min_{Ax=b} f(x), \tag{20}$$

where $f \in \Gamma_0(X)$ and $A \colon X \to Y$ is a continuous linear operator with closed range and $b \in Y$. We assume that a solution exists and that f is continuous at some x such that $Ax = b$. This problem can be equivalently formulated as

$$\min_{x \in X} f(x) + \iota_{\{b\}}(Ax), \tag{21}$$

which is in the form $(\mathscr{P})$. Then, in view of Fact 9(v), the dual problem of (20) is

$$\min_{u \in Y} f^*(-A^*u) + \langle b, u \rangle .$$

Recalling Fact 14(a), to ensure the existence of dual solutions and a zero duality gap, we need to find conditions ensuring the validity of the calculus rule (15). We first prove that if $x \in X$ is such that $Ax = b$, then

$$\partial(\iota_{\{b\}} \circ A)(x) = R(A^*) = A^* \partial \iota_{\{b\}}(Ax). \tag{22}$$

Indeed, we note that $\iota_{\{b\}} \circ A = \iota_{A^{-1}(b)}$ and $A^{-1}(b) = x + N(A)$. Then,

$$\begin{aligned} u \in \partial(\iota_{\{b\}} \circ A)(x) &\iff (\forall y \in A^{-1}(b)) \quad \langle y - x, u \rangle \leq 0 \\ &\iff (\forall v \in N(A)) \quad \langle v, u \rangle \leq 0 \\ &\iff u \in N(A)^{\perp} = R(A^*). \end{aligned}$$

Therefore, $\partial(\iota_{\{b\}} \circ A)(x) = R(A^*)$. Moreover, $A^* \partial \iota_{\{b\}}(Ax) = A^* \partial \iota_{\{b\}}(b)$ and the subdifferential of $\iota_{\{b\}}$ is

$$\partial \iota_{\{b\}} \colon Y \to Y \colon y \mapsto \begin{cases} Y & \text{if } y = b \\ \varnothing & \text{if } y \neq b, \end{cases} \tag{23}$$

hence $A^*\partial\iota_{\{b\}}(Ax) = R(A^*)$ and (22) holds. Finally, recalling the calculus rule for subdifferentials in Fact 5 and that we assumed that f is continuous at some $x \in \mathrm{dom}(\iota_{\{b\}} \circ A)$, then, we have $\partial(f + \iota_{\{b\}} \circ A)(x) = \partial f(x) + \partial(\iota_{\{b\}} \circ A)(x) = \partial f(x) + A^*\partial\iota_{\{b\}}(Ax)$ and hence (15) holds. We note in passing that Fermat's rule for (21) is

$$\begin{aligned} 0 \in \partial(f + \iota_{\{b\}} \circ A)(\hat{x}) &\Leftrightarrow 0 \in \partial f(\hat{x}) + A^*\partial\iota_{\{b\}}(A\hat{x}) \\ &\Leftrightarrow 0 \in \partial f(\hat{x}) + R(A^*) \\ &\Leftrightarrow \exists\, \hat{u} \in Y \quad A^*\hat{u} \in \partial f(\hat{x}). \end{aligned}$$

In the differentiable case, this condition reduces to the classical Lagrange multiplier rule, that is, $\hat{x}$ is a solution of (20) if and only if there exists a multiplier $\hat{u}$ such that $A^*\hat{u} = \nabla f(\hat{x})$.

2.7 Bibliographical Notes

Though convexity is a very old concept, the first systematic study of convex sets in finite dimension is due to Minkowski [73]; while concerning convex functions, it was Jensen [58] to introduce the concept now known as midpoint convexity. The lecture notes by Fenchel [48] constitute the first modern exposition on convex analysis in the finite-dimensional case. Indeed, the notions of support function, Legendre–Fenchel conjugate as well as the duality theory presented in Sects. 2.5 and 2.6, for the special case that A is the identity operator, were fully studied there. At the beginning of the 1960s, convex analysis became a mathematical field in his own, thanks to the works by Moreau [74–76] and Rockafellar [99], who established the theory in infinite dimension and developed the concepts of subgradients and subdifferential, among others. Starting from those works, the field flourished, and it is nowadays still a very active research area.

In the following, we list the main references. Concerning the finite-dimensional setting, we refer to the fundamental monography [98] and the book [57]. For Hilbert spaces, a comprehensive treatment is given in [11] (where most of the facts presented can be found). A lot of research has been also devoted to the Banach spaces and general topological vector spaces. For the former case, we refer to [10, 19, 88, 89], and to [46, 99, 115] for the latter.

3 The Proximal Gradient Method

In this section, we focus on the main object of this chapter, which is the *proximal gradient algorithm* (also called the *forward–backward algorithm*). In the following, we describe the basic assumptions and the algorithm, whereas in the next sections, we

study the convergence properties under several additional assumptions. Moreover, we will also address techniques for accelerating the convergence.

Let $f\colon X \to \mathbb{R}$ be a convex differentiable function, let $g \in \Gamma_0(X)$ and set $F = f + g$. We aim at the following composite optimization problem:

$$\underset{x\in X}{\text{minimize}} \quad f(x) + g(x) =: F(x). \tag{24}$$

The algorithm is detailed below.

Algorithm 1 (The proximal gradient method) *Let $x_0 \in X$ and $\gamma > 0$. Then,*

$$\begin{array}{l} \textit{for } k = 0, 1, \dots \\ \quad \lfloor\, x_{k+1} = \operatorname{prox}_{\gamma g}\big(x_k - \gamma \nabla f(x_k)\big). \end{array} \tag{25}$$

In the above algorithm, $\operatorname{prox}_{\gamma g}\colon X \to X$ is the so-called *proximity operator* of γg which will be defined in the next section. Also, $\gamma > 0$ is the *stepsize* which has to be determined according to the smoothness property of f. More precisely, we will assume that the gradient ∇f is L-Lipschitz continuous, for some $L > 0$, and that the stepsize is set as

$$\gamma < \frac{2}{L}. \tag{26}$$

Remark 16 We stress that some restriction on the stepsize γ should be required. Indeed if we take $g = 0$ and $f(x) = (L/2)\|x\|^2$, we have

$$x_{k+1} = (1 - \gamma L)x_k.$$

Thus, if we take $\gamma = 2/L$, we have $x_{k+1} = -x_k$ and the sequence does not converge, unless $x_0 = 0$.

Example 17 (Iterative Soft-Thresholding Algorithm (ISTA) [41]) We consider the so called Lasso problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1. \tag{27}$$

Then, Algorithm 1 reduces to the following. Let $\gamma \in \,]0, 2/\|A^*A\|[$ and $x_0 \in X$, then

$$\begin{array}{l} \text{for } k = 0, 1, \dots \\ \quad \lfloor\, x_{k+1} = \operatorname{soft}_{\gamma\lambda}(x_k - \gamma A^*(Ax_k - y)). \end{array} \tag{28}$$

In the above equation, $\operatorname{soft}_{\gamma\lambda}\colon \mathbb{R} \to \mathbb{R}$ is the so-called *soft-thresholding operator*, that is, the proximity operator of $\lambda|\cdot|$, which is supposed to be applied componentwise (see (43)).

3.1 Nonexpansive and Averaged Operators

In this section, we present the convergence theory for the method of the fixed point iteration. We recall the classical theory for contractive operators and then we address the case of averaged operators which is motivated by the Krasnosel'skiĭ–Mann iteration.

Let X be a real Hilbert space and let $T : X \to X$. Then

(i) T is *nonexpansive* if for all $x, y \in X$, $\|Tx - Ty\| \leq \|x - y\|$
(ii) T is a *contraction* if for all $x, y \in X$, $\|Tx - Ty\| \leq q\|x - y\|$, for some $q \in]0, 1[$.

A fixed point of T is a point $x \in X$ such that $Tx = x$ and the set of such points is denoted by $\operatorname{Fix} T$. In order to compute fixed points of T, we will consider the following *fixed point iteration*. Let $x_0 \in X$ and define, for every $k \in \mathbb{N}$,

$$x_{k+1} = Tx_k. \tag{29}$$

An iterative method of type (29) is also called *Picard iteration* or *the method of successive approximations*.

Remark 18

(i) Nonexpansive operators may have no fixed points. For instance, a translation $T = \mathrm{Id} + a$, with $a \neq 0$, does not have any fixed point.
(ii) For nonexpansive operators, even admitting fixed points, the fixed point iteration may fail to converge. Indeed, this occurs if we take $T = -\mathrm{Id}$ and start with $x_0 \neq 0$. More generally, rotations are nonexpansive operators admitting a fixed point, for which the fixed point iteration does not converge.

The first important result concerning existence of fixed points and the convergence of the fixed point iteration is the following.

Theorem 19 (Banach-Caccioppoli) *Let $T : X \to X$ be a q-contractive mapping for some $0 < q < 1$. Then there exists a unique fixed point of T, that is,* $\operatorname{Fix} T = \{x_*\}$. *Moreover, for the fixed point iteration* (29)*, we have*

$$(\forall k \in \mathbb{N}) \quad \|x_k - x_*\| \leq q^k \|x_0 - x_*\| \quad \text{and} \quad \|x_k - x_*\| \leq \frac{q^k}{1-q}\|x_0 - x_1\|. \tag{30}$$

Proof We first note that

$$(\forall x, y \in X) \qquad \|x - y\| \leq \frac{1}{1-q}\big(\|x - Tx\| + \|y - Ty\|\big). \tag{31}$$

Indeed, $\|x - y\| \leq \|x - Tx\| + \|Tx - Ty\| + \|Ty - y\| \leq \|x - Tx\| + q\|x - y\| + \|y - Ty\|$, hence $(1 - q)\|x - y\| \leq \|x - Tx\| + \|Ty - y\|$ and (31) follows. Inequality (31) shows that there may exist at most one fixed point of T. Moreover, for every

$k, h \in \mathbb{N}$,

$$\begin{aligned}\|x_k - x_h\| &\leq \frac{1}{1-q}\big(\|x_k - x_{k+1}\| + \|x_h - x_{h+1}\|\big)\\ &\leq \frac{1}{1-q}\big(\|T^k x_0 - T^k x_1\| + \|T^h x_0 - T^h x_1\|\big)\\ &\leq \frac{1}{1-q}\big(q^k\|x_0 - x_1\| + q^h\|x_0 - x_1\|\big)\\ &\leq \frac{q^k + q^h}{1-q}\|x_0 - x_1\|, \end{aligned} \tag{32}$$

where we used that T^k is q^k-contractive. Since $0 < q < 1$, q^k and q^h converge to zero as k and h go to $+\infty$. Therefore, $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence and hence it converges, say to x_*. Then $Tx_k \to Tx_*$ and $Tx_k = x_{k+1} \to x_*$, so $Tx_* = x_*$, that is, x_* is a fixed point of T. The second inequality in (30) follows from (32) by letting $h \to +\infty$. The first equality in (30) follows from the following chain of inequalities

$$\|x_k - x_*\| = \|Tx_{k-1} - Tx_*\| \leq q\|x_{k-1} - x_*\| \leq \cdots \leq q^k\|x_0 - x_*\|.$$

The statement follows. □

As we noted in Remark 18 for general non expansive operators, the fixed point iteration (29) may not converge. To overcome this situation, it is enough to slightly modify the iteration. This leads to the following definition.

Let $T : X \to X$ be a nonexpansive operator and let $\lambda \in]0, 1[$. The *Krasnosel'skiĭ–Mann iteration* is defined as follows:

$$x_0 \in X, \qquad x_{k+1} = x_k + \lambda(Tx_k - x_k). \tag{33}$$

If we look at the example given in Remark 18(ii) , now we see that the iteration (33) becomes $x_{k+1} = (1 - 2\lambda)x_k$. Since $|1 - 2\lambda| < 1$, we have that $x_k = (1 - 2\lambda)^k x_0 \to 0$. Iteration (33) can be equivalently written as a fixed point iteration of the operator $T_\lambda = (1 - \lambda)\mathrm{Id} + \lambda T$. This motivates the study of operators that are convex combination of the identity operator and nonexpansive operators and justify the definition below.

Definition 20 Let $\alpha \in]0, 1[$. Then $T : X \to X$ is an *α-averaged operator* if $T = (1 - \alpha)\mathrm{Id} + \alpha R$ for some nonexpansive operator R. An operator which is $1/2$-averaged is also called *firmly nonexpansive*.

Remark 21 Since averaged operators are convex combinations of nonexpansive operators, they are indeed nonexpansive operators. This follows by the following chain of inequalities:

$$\|Tx - Ty\| = \|(1-\alpha)(x-y) + \alpha(Rx - Ry)\| \leq (1-\alpha)\|x-y\| + \alpha\|Rx - Ry\|$$
$$\leq (1-\alpha)\|x-y\| + \alpha\|x-y\| = \|x-y\|.$$

In the following, we give several characterizations of the property of being an averaged operators.

Lemma 22 *Let $x, y \in X$ and $\lambda \in \mathbb{R}$. Then*

$$\|(1-\lambda)x + \lambda y\|^2 = (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - (1-\lambda)\lambda\|x-y\|^2. \tag{34}$$

Proof Indeed

$$\begin{aligned}\|(1-\lambda)x + \lambda y\|^2 &= (1-\lambda)^2\|x\|^2 + \lambda^2\|y\|^2 + 2(1-\lambda)\lambda\,\langle x, y\rangle \\ &= (1-\lambda)\|x\|^2 - \lambda(1-\lambda)\|x\|^2 \\ &\quad + \lambda\|y\|^2 - (1-\lambda)\lambda\|y\|^2 + 2(1-\lambda)\lambda\,\langle x, y\rangle \\ &= (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - (1-\lambda)\lambda\big(\|x\|^2 + \|y\|^2 - 2\,\langle x, y\rangle\big)\end{aligned}$$

and the statement follows. □

Proposition 23 *Let $T : X \to X$ and $\alpha \in]0, 1[$. Then the following statements are equivalent*

(i) *T is α-averaged*

(ii) $\left(1 - \frac{1}{\alpha}\right)\mathrm{Id} + \frac{1}{\alpha}T$ *is nonexpansive*

(iii) *For every $(x, y) \in X^2$,*

$$\|Tx - Ty\|^2 \leq \|x-y\|^2 - \left(\frac{1}{\alpha} - 1\right)\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2.$$

(iv) *For every $(x, y) \in X^2$*

$$\|Tx - Ty\|^2 + (1-2\alpha)\|x-y\|^2 \leq 2(1-\alpha)\langle x-y, Tx - Ty\rangle.$$

Proof (i) $\Leftrightarrow$ (ii): It follows from the following equivalence

$$T = (1-\alpha)\mathrm{Id} + \alpha R \;\Leftrightarrow\; R = \left(1 - \frac{1}{\alpha}\right)\mathrm{Id} + \frac{1}{\alpha}T.$$

(ii) $\Leftrightarrow$ (iii) : Set $R = (1-\alpha^{-1})\mathrm{Id} + \alpha^{-1}T$ and let $x, y \in X$. It follows from Lemma 22 that

$$\begin{aligned}\|Rx - Ry\|^2 &= \|(1-\alpha^{-1})(x-y) + \alpha^{-1}(Tx - Ty)\|^2 \\ &= (1-\alpha^{-1})\|x-y\|^2 + \alpha^{-1}\|Tx - Ty\|^2 \\ &\quad - \alpha^{-1}(1-\alpha^{-1})\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2\end{aligned}$$

and hence

$$\begin{aligned}&\|Rx - Ry\|^2 - \|x-y\|^2 \\ &= \frac{1}{\alpha}\left(\|Tx - Ty\|^2 - \|x-y\|^2 + \left(\frac{1}{\alpha} - 1\right)\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2\right).\end{aligned}$$

So inequality $\|Rx - Ry\|^2 - \|x-y\|^2 \leq 0$ is equivalent to that in (iii).

(iii) $\Leftrightarrow$ (iv): It follows from the inequality

$$\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 = \|x-y\|^2 + \|Tx - Ty\|^2 - 2\langle x - y, Tx - Ty\rangle.$$

□

Remark 24 The inequality in Proposition 23(iii) shows that if T is α-averaged, then it is also α'-averaged for every $\alpha' > \alpha$. So it makes sense to consider the best (smallest) constant of averagedness.

Remark 25 Contractions are averaged operators. More precisely, if T is a contraction with constant q, then it is $(q+1)/2$-averaged. By Proposition 23(i) it is enough to show that $(1 - 2/(q+1))\mathrm{Id} + 2/(q+1)T$ is nonexpansive. Indeed

$$\begin{aligned}(\forall x, y \in X) \quad &\left\|\frac{q-1}{q+1}x + \frac{2}{q+1}Tx - \frac{q-1}{q+1}y - \frac{2}{q+1}Ty\right\| \leq \\ &\leq \frac{1-q}{q+1}\|x-y\| + \frac{2q}{q+1}\|x-y\| \leq \|x-y\|.\end{aligned}$$

Remark 26 In view of Definition 20 and Proposition 23(iii), an operator T is firmly nonexpansive if and only if

$$(\forall\, x, y \in X) \quad \|Tx - Ty\|^2 \leq \langle x - y, Tx - Ty\rangle. \tag{35}$$

The properties of being averaged is preserved by compositions, as the following result shows.

Proposition 27 *Let $T_1 \colon X \to X$ and $T_2 \colon X \to X$ be two averaged operators, with constants α_1 and α_2 respectively. Then $T_1 \circ T_2$ is averaged with constant*

$$\alpha = \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2}.$$

Averaged operators are important since, provided that they have fixed points, the *Picard iteration* always weakly converges to some fixed point. In the rest of the section, we will prove this result.

Lemma 28 *(demiclosedness principle) Let $T : X \to X$ be a nonexpansive operator. Then $I - T$ is demiclosed, that is, for all sequence $(x_k)_{k\in\mathbb{N}}$ in X and $x, z \in X$, we have*

$$x_k \rightharpoonup x \ \text{ and } \ x_k - Tx_k \to z \ \Rightarrow \ x - Tx = z. \tag{36}$$

Proof Let $k \in \mathbb{N}$. Then using the nonexpansivity of T, we have

$$\begin{aligned}\|x - Tx - z\|^2 &= \|x_k - Tx - z\|^2 - \|x_k - x\|^2 - 2\langle x_k - x, x - Tx - z\rangle \\ &= \|x_k - Tx_k - z\|^2 + \|Tx_k - Tx\|^2 + 2\langle x_k - Tx_k - z, Tx_k - Tx\rangle \\ &\quad - \|x_k - x\|^2 - 2\langle x_k - x, x - Tx - z\rangle \\ &\leq \|x_k - Tx_k - z\|^2 + 2\langle x_k - Tx_k - z, Tx_k - Tx\rangle - 2\langle x_k - x, x - Tx - z\rangle.\end{aligned}$$

Since $x_k - Tx_k - z \to 0$, $x_k - x \rightharpoonup 0$, and Tx_k is bounded, the right-hand side of the above inequality goes to zero and hence $\|x - Tx - z\|^2 = 0$. □

Lemma 29 *(Opial) Let $F \subset X$ be nonempty. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X and suppose that the weak cluster points of $(x_k)_{k\in\mathbb{N}}$ belongs to F and that for any $y \in F$, $(\|x_k - y\|)_{k\in\mathbb{N}}$ is convergent. Then $(x_k)_{k\in\mathbb{N}}$ weakly converges to a point in F.*

Proof The assumptions ensure that $(x_k)_{k\in\mathbb{N}}$ is bounded. Therefore, the set of weak cluster points of $(x_k)_{k\in\mathbb{N}}$ is nonempty. Let $y_1, y_2 \in X$ and let $(x_k^1)_{k\in\mathbb{N}}$ and $(x_k^2)_{k\in\mathbb{N}}$ be subsequences of $(x_k)_{k\in\mathbb{N}}$ such that $x_k^1 \rightharpoonup y_1$ and $x_k^2 \rightharpoonup y_2$. Then, for every $k \in \mathbb{N}$,

$$\begin{aligned}\|x_k - y_1\|^2 - \|y_1\|^2 &= \|x_k\|^2 - 2\langle x_k, y_1\rangle \\ \|x_k - y_2\|^2 - \|y_2\|^2 &= \|x_k\|^2 - 2\langle x_k, y_2\rangle\end{aligned}$$

and hence (subtracting)

$$2\langle x_k, y_2 - y_1\rangle = \|x_k - y_1\|^2 - \|x_k - y_2\|^2 - \|y_1\|^2 + \|y_2\|^2. \tag{37}$$

Since y_1 and y_2 are weak cluster points of $(x_k)_{k\in\mathbb{N}}$, by assumptions, $y_1, y_2 \in F$ and $(\|x_k - y_1\|)_{k\in\mathbb{N}}$ and $(\|x_k - y_2\|)_{k\in\mathbb{N}}$ are convergent. Therefore, by (37), we obtain that there exists $\beta \in \mathbb{R}$ such that $\langle x_k, y_2 - y_1\rangle \to \beta$. Now, since $x_k^i \rightharpoonup y_i$, $i = 1, 2$, we have $\langle x_k^i, y_2 - y_1\rangle \to \langle y_i, y_2 - y_1\rangle$, which implies

$$\langle y_1, y_2 - y_1\rangle = \beta = \langle y_2, y_2 - y_1\rangle$$

and hence $\|y_2 - y_1\|^2 = 0$. This proves that the set of weal cluster points of the sequence $(x_k)_{k\in\mathbb{N}}$ is a singleton. So, the sequence $(x_k)_{k\in\mathbb{N}}$ is weakly convergent. □

Theorem 30 *Let $\alpha \in]0,1[$ and let $T\colon X \to X$ be an α-averaged operator such that the set of fixed points is nonempty, that is* $\operatorname{Fix} T \neq \varnothing$. *Let $(x_k)_{k\in\mathbb{N}}$ be generated by the fixed point iteration* (29). *Then the following hold.*

(i) *For every $k \in \mathbb{N}$ and every $x_* \in \operatorname{Fix} T$, $\|x_{k+1} - x_*\| \leq \|x_k - x_*\|$*

(ii) $\displaystyle\sum_{k=0}^{+\infty} \|Tx_k - x_k\|^2 < \frac{\alpha}{1-\alpha} \operatorname{dist}(x_0, \operatorname{Fix} T)^2$

(iii) *$(x_k)_{k\in\mathbb{N}}$ weakly converges to some point $x_* \in \operatorname{Fix} T$.*

Proof (i): Since T is nonexpansive and x_* is a fixed point of T, $\|x_{k+1} - x_*\| = \|Tx_k - Tx_*\| \leq \|x_k - x_*\|$.

(ii): Let $x_* \in S$. Then, by Proposition 23(iii) (with $x = x_k$ and $y = x_*$), we have

$$(\forall\, k \in \mathbb{N})\ \|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \left(\frac{1}{\alpha} - 1\right)\|x_k - Tx_k\|^2. \tag{38}$$

Therefore,

$$\frac{1-\alpha}{\alpha} \sum_{k=0}^{+\infty} \|x_k - Tx_k\|^2 \leq \sum_{k=0}^{+\infty} \left(\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right) \leq \|x_0 - x_*\|^2.$$

(iii): It follows from (ii) that $\|Tx_k - x_k\| \to 0$. Let x_* be a weak cluster point of $(x_k)_{k\in\mathbb{N}}$ and let $(x'_k)_{k\in\mathbb{N}}$ be a subsequence of $(x_k)_{k\in\mathbb{N}}$ such that $x'_k \rightharpoonup x_*$. Then $Tx'_k - x'_k \to 0$. Hence, in virtue of Lemma 28, $Tx_* - x_* = 0$, that is $x_* \in \operatorname{Fix} T$. Moreover, by item (i), for every $x_* \in \operatorname{Fix} T$, $\|x_k - x_*\|$ is decreasing and hence convergent. The statement follows from Lemma 29 with $F = \operatorname{Fix} T$. □

Applying the previous theorem to the operator $T_\lambda = (1-\lambda)\mathrm{Id} + \lambda T$ and noting that $\operatorname{Fix} T_\lambda = \operatorname{Fix} T$, we get the following result.

Corollary 31 *Let $T\colon X \to X$ be a nonexpansive operator admitting fixed points and let $(x_k)_{k\in\mathbb{N}}$ be generated by the Krasnosel'skiĭ–Mann iteration* (33). *Then $(x_k)_{k\in\mathbb{N}}$ converges to some fixed point of T.*

3.2 The Proximity Operator

Motivated by the use of nonsmooth regularization techniques in inverse problems, we introduce the proximity operator of a convex function.

Definition 32 Let $g \in \Gamma_0(X)$. Then, the *proximity operator* of g is

$$\operatorname{prox}_g\colon X \to X, \qquad \operatorname{prox}_g(x) = \operatorname{argmin}_{y\in X} \left\{ g(y) + \frac{1}{2}\|y - x\|^2 \right\}.$$

Note that the definition is well-posed since the function $y \mapsto g(y) + (1/2)\|y - x\|^2$ is lower semicontinuous and strongly convex, hence, it has a unique minimizer. Moreover, let us check that $\mathrm{prox}_g = (\mathrm{Id} + \partial g)^{-1}$. Using the sum rule for the subdifferential, which holds since the square norm is differentiable, we derive

$$\begin{aligned} z = \mathrm{prox}_g(x) &\Leftrightarrow 0 \in \partial g(z) + z - x \\ &\Leftrightarrow x \in (\mathrm{Id} + \partial g)(z) \\ &\Leftrightarrow z \in (\mathrm{Id} + \partial g)^{-1}(x). \end{aligned}$$

This shows that $(\mathrm{Id} + \partial g)^{-1}(x)$ is actually a singleton and its unique element is $\mathrm{prox}_{\gamma g}(x)$. Note that for every $x \in X$, $\mathrm{prox}_g(x) \in \mathrm{dom} g$, since the minimizer of $g + (1/2)\|\cdot\|^2$ is clearly in the domain of g.

Example 33 Let C be a closed and convex set. The proximity operator of ι_C is the projection on C. The projection is nonexpansive (and, indeed, firmly nonexpansive), but in general not a contraction, unless C is a singleton.

Proposition 34 *Let $g \in \Gamma_0(X)$. Then*

$$(\forall x, y \in X) \quad \|\mathrm{prox}_g(x) - \mathrm{prox}_g(y)\|^2 \leq \langle x - y, \mathrm{prox}_g(x) - \mathrm{prox}_g(y)\rangle. \tag{39}$$

In other words, recalling (35), the operator prox_g is firmly nonexpansive.

Proof Let $x, y \in X$ and set $p_x = \mathrm{prox}_g(x)$ and $p_y = \mathrm{prox}_g(y)$. Then, by Fermat's rule, we have

$$x - p_x \in \partial g(p_x) \text{ and } y - p_y \in \partial g(p_y).$$

Therefore,

$$\begin{aligned} g(p_y) &\geq g(p_x) + \langle x - p_x p_y - p_x\rangle \\ g(p_x) &\geq g(p_y) + \langle y - p_y p_x - p_y\rangle \end{aligned}$$

and summing $g(p_y) + g(p_x) \geq g(p_x) + g(p_y) + \langle y - p_y - x + p_x, p_x - p_y\rangle$. Then the statement follows. □

Remark 35 The function

$$g_\lambda(u) = \inf_{x \in X} \left\{ g(x) + \frac{1}{2\lambda}\|x - u\|^2 \right\}, \tag{40}$$

is called the Moreau envelope of g with parameter λ. We have that g_λ is differentiable and the gradient of g_λ is given as

$$\nabla g_\lambda(u) = \frac{u - \mathrm{prox}_{\lambda g}(u)}{\lambda} \in \partial g(\mathrm{prox}_{\lambda g}(x)). \tag{41}$$

In the following, we provide important properties of proximity operators.

Proposition 36 (Separable sum) *Let $(X_i)_{1\le i\le m}$ be Hilbert spaces and let $X = \bigoplus_{i=1}^m X_i$ be their direct product. Let, for every $i = 1, \dots, m$, $g_i \in \Gamma_0(X_i)$ and define $g\colon X \to]-\infty, +\infty]$ by $g(x) = \sum_{i=1}^m g_i(x_i)$, for every $x = (x_1, \dots, x_m) \in X$. Then*

$$(\forall\, x = (x_1, \dots, x_m) \in X) \qquad \operatorname{prox}_g(x) = (\operatorname{prox}_{g_1}(x_1), \dots, \operatorname{prox}_{g_m}(x_m)). \tag{42}$$

Example 37

(i) (Proximity operator of the ℓ_1 norm) Let $X = \mathbb{R}^d$. The ℓ_1 norm on X is separable, thus the proximity operator can be computed componentwise, so it is enough to compute the proximity operator of the absolute value in $\mathbb{R}$. Let $\gamma > 0$. By definition, for every $t \in \mathbb{R}$, $\operatorname{prox}_{\gamma|\cdot|}(t) = (\mathrm{Id} + \gamma\partial|\cdot|)^{-1}(t)$. Thus, if we make the plot of the graph of $\mathrm{Id} + \gamma\partial|\cdot|$ and invert it, we discover that

$$\operatorname{soft}_\gamma(t) := \operatorname{prox}_{\gamma|\cdot|}(t) = \begin{cases} t-\gamma & \text{if } t > \gamma \\ 0 & \text{if } |t| \le \gamma \\ t+\gamma & \text{if } t < -\gamma. \end{cases} \tag{43}$$

Thus, it follows from Proposition 36 that, for every $x \in \mathbb{R}^d$ and every $i = 1, \dots, d$, $(\operatorname{prox}_{\gamma\|\cdot\|_1}(x))_i = \operatorname{prox}_{\gamma|\cdot|}(x_i)$.

(ii) (Proximity operator of the $\ell_1 + \ell_2$ norm)

$$g(x) = \|x\|_1 + \frac{\lambda}{2}\|x\|_2^2$$

$$\operatorname{prox}_{\gamma g}(x) = \operatorname{prox}_{(\gamma/(\gamma\lambda+1))\|\cdot\|_1}(x/(\gamma\lambda+1))$$

$$(\operatorname{prox}_{\gamma g}(x))_i = \begin{cases} (x_i-\gamma)/(\gamma\lambda+1) & \text{if } x_i > \gamma \\ 0 & \text{if } |x_i| \le \gamma \\ (x_i+\gamma)/(\gamma\lambda+1) & \text{if } x_i < -\gamma \end{cases}$$

Proposition 38 (Properties of the proximity operator) *Let $h \in \Gamma_0(X)$ and let $\gamma > 0$. Then the following holds*

(i) *(linear perturbation) Let $g = h + \langle\cdot, u\rangle + a$, with $u \in X$ and $a \in \mathbb{R}$. Then*

$$\operatorname{prox}_{\gamma g}(x) = \operatorname{prox}_{\gamma h}(x - \gamma u).$$

(ii) *Let $g(x) = h(ax + b)$, with $a \in \mathbb{R}$, $a \neq 0$ and $b \in X$. Then*

$$\operatorname{prox}_{\gamma g}(x) = (\operatorname{prox}_{a^2\gamma h}(ax + b) - b)/a.$$

(iii) *(composition with an orthogonal matrix) Let $g = h \circ L$, with $L\colon X \to X$ a bijective linear map such that $L^* = L^{-1}$. Then*

$$(\forall x \in X) \quad \operatorname{prox}_{\gamma g}(x) = L^* \operatorname{prox}_{\gamma h}(Lx).$$

Proof In the following, we let $x \in X$ and set $p = \operatorname{prox}_{\gamma g}(x)$.

(i): Since $p = \operatorname{argmin}_{y\in X}\{\gamma h(y) + \gamma\langle u, y\rangle + a + \frac{1}{2}\|y - x\|^2\}$, Fermat's rule yields

$$\begin{aligned} 0 \in \gamma \partial h(p) + \gamma u + p - x &\Leftrightarrow x - \gamma u \in (\mathrm{Id} + \gamma \partial h)(p) \\ &\Leftrightarrow p = \operatorname{prox}_{\gamma h}(x - \gamma u). \end{aligned}$$

(ii): We have:

$$\begin{aligned} p = \operatorname{prox}_{\gamma g}(x) &\Leftrightarrow p = \operatorname{argmin}_{y\in X} \left\{ \gamma h(ay + b) + \frac{1}{2}\|y - x\|^2 \right\} \\ &\Leftrightarrow p = \operatorname{argmin}_{y\in X} \left\{ \gamma h(ay + b) + \frac{1}{2a^2}\|ay + b - (ax + b)\|^2 \right\} \\ &\Leftrightarrow p = \operatorname{argmin}_{y\in X} \left\{ \gamma a^2 h(ay + b) + \frac{1}{2}\|ay + b - (ax + b)\|^2 \right\} \\ &\Leftrightarrow ap + b = \operatorname{prox}_{a^2\gamma h}(ax + b) \\ &\Leftrightarrow p = (\operatorname{prox}_{a^2\gamma h}(ax + b) - b)/a. \end{aligned}$$

(iii): We have

$$\begin{aligned} p = \operatorname{prox}_{\gamma g}(x) &\Leftrightarrow p = \operatorname{argmin}_{y\in X} \left\{ \gamma h(Ly) + \frac{1}{2}\|y - x\|^2 \right\} \\ &\Leftrightarrow 0 \in \gamma L^* \partial h(Lp) + p - x \\ &\Leftrightarrow x - p \in L^{-1} \partial h(Lp) \\ &\Leftrightarrow Lx \in \gamma \partial h(Lp) + Lp \\ &\Leftrightarrow p = L^* \operatorname{prox}_{\gamma h}(Lx) \end{aligned}$$

The statement follows. □

Remark 39 Regarding Proposition 38(iii), in general, if L is not orthogonal, we can apply a gradient descent on the dual of the minimization problem defining the prox to compute it approximately. See Sect. 5.

We now introduce an important identity, that is, the *Moreau's decomposition formula*. Let V be a closed linear subspace of X. Then we know that x can be uniquely decomposed in two orthogonal components, $P_V x$ and $P_{V^\perp} x$ such that:

$$x = x_V + x_{V^\perp} = P_V x + P_{V^\perp}(x). \tag{44}$$

If we set $f = \iota_V$, we first note that $(\iota_V)^*(u) = \sup_{x\in X} \langle x, u\rangle - \iota_V(x) = \iota_{V^\perp}(u)$. Thus, we can rewrite (44) as

$$x = \mathrm{prox}_{\iota_V}(x) + \mathrm{prox}_{(\iota_V)^*}(x).$$

This last formula can be generalized to every convex function.

Theorem 40 (Moreau's decomposition) *Let $g \in \Gamma_0(X)$ and let $x \in X$. Then*

$$x = \mathrm{prox}_g(x) + \mathrm{prox}_{g^*}(x).$$

More generally, for all $\gamma > 0$, $x = \mathrm{prox}_{\gamma g}(x) + \gamma \mathrm{prox}_{g^*/\gamma}(x/\gamma)$.

Proof It follows from the list of equivalences below.

$$\begin{aligned} p = \mathrm{prox}_g(x) &\Leftrightarrow x - p \in \partial g(p) \\ &\Leftrightarrow p \in \partial g^*(x - p) \\ &\Leftrightarrow x - (x - p) \in \partial g^*(x - p) \\ &\Leftrightarrow x - p = \mathrm{prox}_{g^*}(x). \end{aligned}$$

□

Example 41 *(The proximity operator of the Euclidean norm)* We want to compute the prox of the norm of X (which is a Hilbert space). First note that

$$\|x\| = \sup_{\|u\|\le 1} \langle x, u\rangle = \sigma_{B_1(0)}(x).$$

Hence,

$$\|\cdot\| = \sigma_{B_1(0)} = (\iota_{B_1(0)})^*.$$

Therefore, it follows from Theorem 40 that

$$\mathrm{prox}_{\|\cdot\|}(x) = x - \mathrm{prox}_{\iota_{B_1(0)}}(x) = x - P_{B_1(0)}(x).$$

More explicitly:

$$\mathrm{prox}_{\|\cdot\|}(x) = \begin{cases} x - \dfrac{x}{\|x\|} & \text{if } \|x\| > 1 \\ 0 & \text{if } \|x\| \le 1. \end{cases}$$

Note that this operation corresponds to a vector soft thresholding, which reduces to (43) for $\dim X = 1$ and $\gamma = 1$.

Example 42 *(The proximity operator of the group lasso norm)* Let $\mathcal{J} = \{J_1, \dots, J_l\}$ be a partition of $\{1, \dots, d\}$. We define a norm on $\mathbb{R}^d$ by considering

$$\|x\|_{\mathcal{J}} = \sum_{i=1}^{l} \left(\sum_{j\in J_i} |x_j|^2 \right)^{1/2}.$$

For every $x \in \mathbb{R}^d$, let us call $x_{J_i} = (x_j)_{j \in J_i} \in \mathbb{R}^{J_i}$ and denote by $\|\cdot\|_{J_i}$ the Euclidean norm on $\mathbb{R}^{J_i}$. Then

$$\|x\|_{\mathcal{J}} = \sum_{i=1}^{l} \|x_{J_i}\|_{J_i}.$$

We next compute the proximity operator of $\|\cdot\|_{\mathcal{J}}$. First note that $\|\cdot\|_{\mathcal{J}}$ is the sum of functions depending on groups of variables x_{J_i}. Therefore the prox can be computed group-wise thanks to the decomposability property (42). Thus

$$(\mathrm{prox}_{\|\cdot\|_{\mathcal{J}}}(x))_{J_i} = \mathrm{prox}_{\|\cdot\|_{J_i}}(x_{J_i}),$$

and recalling Example 41, we have

$$(\mathrm{prox}_{\|\cdot\|_{\mathcal{J}}}(x))_{J_i} = \begin{cases} x_{J_i} - \dfrac{x_{J_i}}{\|x_{J_i}\|} & \text{if } \|x_{J_i}\|_{J_i} > 1 \\ 0 & \text{otherwise} \end{cases}$$

The resulting prox operator is called *block soft-thresholding operator*.

3.3 Worst Case Convergence Analysis

Algorithm 1 can be seen as a fixed-point iteration of the following operator

$$T = \mathrm{prox}_{\gamma g} \circ (\mathrm{Id} - \gamma \nabla f), \tag{45}$$

which is the composition of the proximity operator of γg and the operator $\mathrm{Id} - \gamma \nabla f$. We also note that the fixed points of T are the minimizers of $f + g$. Indeed

$$x = Tx \Leftrightarrow x = \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x)) \Leftrightarrow x - \gamma \nabla f(x) - x \in \partial \gamma g(x) \Leftrightarrow 0 \in \partial(f + g)(x).$$

So we need to study the operator T. We already know that $\mathrm{prox}_{\gamma g}$ is firmly non-expansive and hence $(1/2)$-averaged. The following result concerns the operator $\mathrm{Id} - \gamma \nabla f$.

Proposition 43 *Let $f : X \to \mathbb{R}$ be differentiable and let $L > 0$. Let $\gamma > 0$ and set $T_\gamma = \mathrm{Id} - \gamma \nabla f$. Then, the L-Lipschitz continuity of ∇f is equivalent to the property*

$$(\forall\, x, y \in X)\ \|T_\gamma x - T_\gamma y\|^2 \leq \|x - y\|^2 - \left(\frac{2}{\gamma L} - 1\right)\|(\mathrm{Id} - T_\gamma)x - (\mathrm{Id} - T_\gamma)y\|^2. \tag{46}$$

In particular, if $\gamma < 2/L$, T_γ is a α-averaged operator, with $\alpha = \gamma L/2 < 1$.

Proof Multiplying by $\gamma^2 L$ the inequality in Fact 1(iv) and replacing $\gamma \nabla f$ with $\mathrm{Id} - T_\gamma$, we obtain

$$\|(\mathrm{Id} - T_\gamma)x - (\mathrm{Id} - T_\gamma)y\|^2 \leq \gamma L \langle x - y, (\mathrm{Id} - T_\gamma)x - (\mathrm{Id} - T_\gamma)y \rangle.$$

Then, using the identity

$$\begin{aligned} &2\langle x - y, (\mathrm{Id} - T_\gamma)x - (\mathrm{Id} - T_\gamma)y \rangle \\ &\quad = \|(\mathrm{Id} - T_\gamma)x - (\mathrm{Id} - T_\gamma)y\|^2 + \|x - y\|^2 - \|T_\gamma x - T_\gamma y\|^2, \end{aligned}$$

the statement follows. □

Proposition 44 *Let $f : X \to \mathbb{R}$ a differentiable convex function with a Lipschitz continuous gradient with constant L, let $g \in \Gamma_0(X)$ and set T as in (45). Suppose that $\gamma < 2/L$. Then T is α-averaged with $\alpha = 2/(4 - \gamma L)$.*

Proof It follows from Proposition 43 that $I - \gamma \nabla f$ is α_2-averaged with $\alpha_2 = \gamma L/2$. Moreover, Proposition 34 yields that $\mathrm{prox}_{\gamma g}$ is firmly nonexpansive, that is, α_1-averaged with $\alpha_1 = 1/2$. Therefore, by Proposition 27, $T = \mathrm{prox}_{\gamma g} \circ (I - \gamma \nabla f)$ is α-averaged with

$$\alpha = \frac{1/2 + \gamma L/2 - \gamma L/2}{1 - (1/2)(\gamma L/2)} = \frac{2}{4 - \gamma L}.$$ □

Lemma 45 *For any $x, z \in X$, $y \in \mathrm{dom} g$ and for any $u \in \partial g(x)$. We have*

$$F(z) \geq F(x) + \langle z - x, \nabla f(y) + u \rangle - \frac{L}{2}\|x - y\|^2.$$

Proof Let $x, z \in X$ and let $y \in \mathrm{dom} g$. Then, it follows from Fact 1 that

$$f(y) \geq f(x) - \langle x - y, \nabla f(y) \rangle - \frac{L}{2}\|x - y\|^2.$$

Hence, since f is convex,

$$f(z) \geq f(y) + \langle z - y, \nabla f(y) \rangle \geq f(x) + \langle z - x, \nabla f(y) \rangle - \frac{L}{2}\|x - y\|^2. \quad (47)$$

Now, since $u \in \partial g(x)$, $g(z) \geq g(x) + \langle z - x, u \rangle$, which summed with inequality (47) give the statement. □

Lemma 46 *Let $(a_k)_{k \in \mathbb{N}}$ be a decreasing sequence in $\mathbb{R}_+$. If $\sum_{k=0}^{+\infty} a_k < +\infty$, then*

$$(\forall\, k \in \mathbb{N}) \quad a_k \leq \frac{1}{k+1} \sum_{k=0}^{+\infty} a_k, \text{ and } a_k = o\Big(\frac{1}{k+1}\Big).$$

Proof Let $k \in \mathbb{N}$. Since $a_k \leq a_i$, for $i = 0, 1, \ldots, k$, we have $\sum_{i=0}^{k} a_i \geq (k+1)a_k$, hence the first part of the statement. As regard the second part, we note that, for every

integer $k \geq 2$, we have $\sum_{i=\lceil k/2 \rceil}^{+\infty} a_i \geq \sum_{i=\lceil k/2 \rceil}^{k} a_i \geq (k+1-\lceil k/2 \rceil)a_k \geq \frac{k+1}{2} a_k$. Therefore, $(k+1)a_k \leq 2 \sum_{i=\lceil k/2 \rceil}^{+\infty} a_i \to 0$ as $k \to +\infty$. □

The following theorem provides full convergence results concerning the proximal gradient algorithm.

Theorem 47 *Let $f : X \to \mathbb{R}$ a differentiable convex function with a Lipschitz continuous gradient with constant L and $g \in \Gamma_0(X)$. Let S_* be the set of minimizers of $F := f + g$ and suppose that $S_* \neq \varnothing$. Let $\gamma < 2/L$ and $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 1. Then the following statements hold*

(i) $\displaystyle\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\|^2 \leq \frac{2}{2-\gamma L} \mathrm{dist}(x_0, S_*)^2.$

(ii) *For every $k \in \mathbb{N}$ and for every $x \in X$,*

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + 2\gamma\big(F(x) - F(x_{k+1})\big) + (\gamma L - 1)\|x_{k+1} - x_k\|^2.$$

(iii) *For all $k \in \mathbb{N}$,*

$$\Big(\frac{1}{\gamma} - \frac{L}{2}\Big)\|x_{k+1} - x_k\|^2 \leq F(x_k) - F(x_{k+1}),$$

so that the algorithm is descending.

(iv) *Let $F_* = \inf_{x\in X}(f+g)(x)$. Then $F(x_{k+1}) - F_* = o(1/(k+1))$ and, for all $k \in \mathbb{N}$,*

$$F(x_{k+1}) - F_* \leq \frac{\mathrm{dist}(x_0, S_*)^2}{k+1} \times \begin{cases} \dfrac{1}{2\gamma} & \text{if } \gamma \leq 1/L \\ \dfrac{L}{2}\dfrac{1}{2-\gamma L} & \text{if } 1/L < \gamma < 2/L. \end{cases} \tag{48}$$

(v) *The sequence $(x_k)_{k\in\mathbb{N}}$ weakly converges to some $x_* \in S_*$.*

Proof (i): It follows from (25), Theorem 44, and Theorem 30(ii).

(ii): Let $x \in X$ and $k \in \mathbb{N}$. It follows from (25) that $u := (x_k - x_{k+1})/\gamma - \nabla f(x_k) \in \partial g(x_{k+1})$, hence

$$\frac{x_k - x_{k+1}}{\gamma} = \nabla f(x_k) + u, \qquad u \in \partial g(x_{k+1}).$$

Thus, by Lemma 45, we have that

$$\begin{aligned} F(x) &\geq F(x_{k+1}) + \langle x - x_{k+1}, \nabla f(x_k) + u\rangle - \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= F(x_{k+1}) + \frac{1}{\gamma}\langle x - x_{k+1}, x_k - x_{k+1}\rangle - \frac{L}{2}\|x_{k+1} - x_k\|^2; \end{aligned}$$

and identity $\|x_k - x\|^2 = \|x_k - x_{k+1}\|^2 + \|x_{k+1} - x\|^2 + 2\langle x_{k+1} - x_k, x - x_{k+1}\rangle$, yields

$$\begin{aligned} F(x) - F(x_{k+1}) &\geq \frac{1}{2\gamma}\Big[\|x_k - x_{k+1}\|^2 + \|x_{k+1} - x\|^2 - \|x_k - x\|^2\Big] - \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= \frac{1}{2\gamma}\big[(1 - \gamma L)\|x_k - x_{k+1}\|^2 + \|x_{k+1} - x\|^2 - \|x_k - x\|^2\big]. \end{aligned}$$

Therefore,

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + 2\gamma\big(F(x) - F(x_{k+1})\big) - (1 - \gamma L)\|x_k - x_{k+1}\|^2$$

and the statement follows.

(iv): Let $x_* \in S_*$. Then, it follows from (ii) that, for every $k \in \mathbb{N}$,

$$0 \leq 2\gamma\big(F(x_{k+1}) - F(x_*)\big) \leq \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2 + (\gamma L - 1)_+\|x_k - x_{k+1}\|^2.$$

Thus, summing and using (i), we have

$$\begin{aligned} 2\gamma \sum_{k=0}^{+\infty}\big(F(x_{k+1}) - F(x_*)\big) &\leq \|x_0 - x_*\|^2 + \frac{2\big(\gamma L - 1\big)_+}{2 - \gamma L}\|x_0 - x_*\|^2 \\ &= \|x_0 - x_*\|^2 \times \begin{cases} 1 & \text{if } \gamma \leq 1/L \\ \dfrac{\gamma L}{2 - \gamma L} & \text{if } 1/L < \gamma < 2/L. \end{cases} \end{aligned}$$

Then, since $\big(F(x_{k+1}) - F(x_*)\big)_{k\in\mathbb{N}}$ is decreasing and positive, the statement follows from Lemma 46.

(v): It follows from (25), Theorems 44, 30(iii), and the fact that $S_* = \mathrm{Fix}(T)$. □

Remark 48 It follows from (48) that the best bound is achieved when $\gamma = 1/L$.

Remark 49 Suppose that in problem (24) f is the Moreau envelope of a function $h \in \Gamma_0(X)$ with parameter 1, that is $f = h_1$. Then $\nabla f(x) = x - \mathrm{prox}_h(x)$, which is 1-Lipschitz continuous, and the proximal gradient Algorithm 1 with stepsize $\gamma = 1$, becomes

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\lfloor\, x_{k+1} = \mathrm{prox}_{\gamma g}\big(\mathrm{prox}_h(x_k)\big), \end{aligned} \tag{49}$$

which is called the *backward-backward* algorithm. If one takes $g = \iota_{C_1}$ and $h = \iota_{C_2}$, for two closed convex sets $C_1, C_2 \subset X$, we have the alternating projection algorithm

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\lfloor\, x_{k+1} = P_{C_1}\big(P_{C_2}(x_k)\big). \end{aligned} \tag{50}$$

Note that Theorem 47 ensures that the sequence $(x_k)_{k\in\mathbb{N}}$ weakly converges to a point in $\mathrm{argmin}_{x\in C_1} d^2_{C_2}(x)$.

3.4 Convergence Analysis Under Strong Convexity Assumptions

In this section, following the same notation of the previous section, we set

$$T_\gamma = \mathrm{Id} - \gamma \nabla f \quad \text{and} \quad T = \mathrm{prox}_{\gamma g} \circ T_\gamma. \tag{51}$$

We will consider the situation where f and/or g are strongly convex. This will make the corresponding operators T_γ and/or $\mathrm{prox}_{\gamma g}$ contractions.

Proposition 50 *Let $f : X \to \mathbb{R}$ be a differentiable convex function. Suppose that for some $\gamma > 0$, the operator $T_\gamma = \mathrm{Id} - \gamma \nabla f$ is a contraction. Then f is strongly convex and its gradient is Lipschitz continuous.*

Proof Let $x, y \in X$. Then

$$\begin{aligned}
&\|T_\gamma x - T_\gamma y\|^2 \leq q^2 \|x - y\|^2 \\
&\Leftrightarrow \ \|x - y - \gamma(\nabla f(x) - \nabla f(y))\|^2 \leq q^2 \|x - y\|^2 \\
&\Leftrightarrow \ \|x - y\|^2 + \gamma^2 \|\nabla f(x) - \nabla f(y)\|^2 - 2\gamma \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq q^2 \|x - y\|^2 \\
&\Leftrightarrow \ (1 - q^2)\|x - y\|^2 + \gamma^2 \|\nabla f(x) - \nabla f(y)\|^2 \leq 2\gamma \langle \nabla f(x) - \nabla f(y), x - y \rangle \\
&\Rightarrow \begin{cases} \dfrac{1 - q^2}{2\gamma} \|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \\ \dfrac{\gamma}{2} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \end{cases}
\end{aligned}$$

So in virtue of Fact 1(iv) and (12), f is strongly convex and ∇f is Lipschitz continuous. □

Now we assume that f is strongly convex and with Lipschitz continuous gradient. Then we will prove that there exists an interval of values of γ for which T_γ is a contraction.

Theorem 51 *$f : X \to \mathbb{R}$ is Lipschitz smooth with constant $L > 0$ and strongly convex with modulus $\mu > 0$. Then, for every $\gamma \in]0, 2/(L + \mu)]$, $T_\gamma = \mathrm{Id} - \gamma \nabla f$ is a contraction with constant*

$$q_1(\gamma) := \left(1 - \frac{2\gamma\mu L}{L + \mu}\right)^{1/2}. \tag{52}$$

Proof It follows from Fact 2(ii) (multiplied by 2γ) that

$$\frac{2}{\gamma(L + \mu)} \|\gamma \nabla f(x) - \gamma \nabla f(y)\|^2 + \frac{2\gamma\mu L}{L + \mu} \|x - y\|^2 \leq 2 \langle \gamma \nabla f(x) - \gamma \nabla f(y), x - y \rangle$$

Moreover,

$$\|(x-y)-\gamma(\nabla f(x)-\nabla f(y))\|^2 = \|x-y\|^2 + \|\gamma\nabla f(x)-\gamma\nabla f(y)\|^2 \\ -2\langle \gamma\nabla f(x)-\gamma\nabla f(y), x-y\rangle .$$

Hence

$$\|(x-y)-\gamma(\nabla f(x)-\nabla f(y))\|^2 \le \left(1-\frac{2\gamma\mu L}{L+\mu}\right)\|x-y\|^2 \\ -\left(\frac{2}{\gamma(L+\mu)}-1\right)\|\gamma\nabla f(x)-\gamma\nabla f(y)\|^2.$$

Now since $T_\gamma = \mathrm{Id} - \gamma\nabla f$, the inequality above becomes

$$\|T_\gamma x - T_\gamma y\|^2 \le \left(1-\frac{2\gamma\mu L}{\mu+L}\right)\|x-y\|^2 - \left(\frac{2}{\gamma(\mu+L)}-1\right)\|(\mathrm{Id}-T_\gamma)x-(\mathrm{Id}-T_\gamma)y\|^2.$$

Note that if $\gamma(L+\mu)/2 \le 1$, then the above inequality yields

$$\|T_\gamma x - T_\gamma y\| \le \left(1-\frac{2\gamma\mu L}{\mu+L}\right)^{1/2}\|x-y\|, \tag{53}$$

where

$$0 < \frac{2\gamma\mu L}{L+\mu} \le \frac{4\mu L}{(L+\mu)^2} - 1 + 1 = 1 - \left(\frac{L-\mu}{L+\mu}\right)^2 < 1.$$

Therefore, for every $\gamma \in \,]0, 2/(L+\mu)]$, T_γ is a contraction with the constant given in (53). □

If we additionally assume that the function f is twice differentiable the results can be further improved.

Theorem 52 *Let $f\colon X \to \mathbb{R}$ be twice differentiable and suppose that f is μ-strongly convex and that ∇f is L-Lipschitz continuous. Then, for every $\gamma > 0$, $T_\gamma = \mathrm{Id} - \gamma\nabla f$ is Lipschitz continuous with constant*

$$\tilde{q}_1(\gamma) = \max\{|1-\gamma\mu|, |1-\gamma L|\} = \begin{cases} 1-\gamma\mu & \text{if } \gamma \le \dfrac{2}{L+\mu} \\ \gamma L - 1 & \text{if } \gamma \ge \dfrac{2}{L+\mu}. \end{cases} \tag{54}$$

So, if $\gamma \in \,]0, 2/L[$, then T_γ is a contraction.

Proof The mapping T_γ is differentiable and $T_\gamma'(x) = \mathrm{Id} - \gamma\nabla^2 f(x)$. By the mean value theorem, for every $q > 1$,

$$\forall\, x, y \in X \quad \|T_\gamma x - T_\gamma y\| \le q\|x-y\| \;\Leftrightarrow\; \forall\, x \in X \quad \|T_\gamma'(x)\| \le q.$$

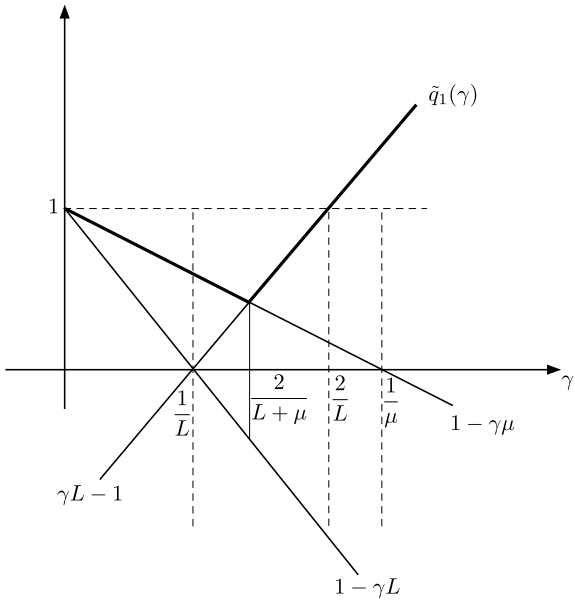


Fig. 1 Explanation of the fact that: $\tilde{q}_1(\gamma) < 1 \iff \gamma < 2/L$

Moreover, $\|T'_\gamma(x)\| = \sup_{\lambda\in\sigma(\nabla^2 f(x))}|1-\gamma\lambda|$. Since f is μ strongly convex and ∇f is L-Lipschitz continuous,

$$(\forall\, x \in X)(\forall\, u \in X) \quad \mu\|u\|^2 \leq \langle \nabla^2 f(x)u, u\rangle \leq L\|u\|^2.$$

So $\sigma(\nabla^2 f(x)) \subset [\mu, L]$ and hence $\|T'_\gamma(x)\| \leq \max_{\lambda\in[\mu,L]}|1-\gamma\lambda| = \tilde{q}_1(\gamma)$. This last equality follows by noting that $\lambda \mapsto |1-\gamma\lambda|$ is a piecewise convex function and hence it achieves its maximum at the end points of the interval $[\mu, L]$. It follows from (54) that $\tilde{q}(\gamma) < 1 \;\Leftrightarrow\; \gamma \in]0, 2/L[$ (see Fig. 1). □

Remark 53 The constant $\tilde{q}_1(\gamma)$ given in Theorem 52 is always better than the constant $q_1(\gamma)$ given in Theorem 51. However, on the minimum value they agree.

Theorem 54 *Let $g \in \Gamma_0(X)$ and suppose that g is σ-strongly convex. Then, for every $\gamma > 1$ the operator* $\text{prox}_{\gamma g}$ *is a contraction with constant* $1/(1+\gamma\sigma)$.

Proof Let $x, y \in X$ and set $p_x = \text{prox}_{\gamma g}(x)$ and $p_y = \text{prox}_{\gamma g}(y)$. Then, by Fermat's rule, we have $(x - p_x)/\gamma \in \partial g(p_x)$ and $(y - p_y)/\gamma \in \partial g(p_y)$. Therefore, recalling Fact 4, we have

$$\begin{aligned} g(p_y) - g(p_x) &\geq \gamma^{-1}\langle p_y - p_x x - p_x\rangle + (\sigma/2)\|p_y - p_x\|^2 \\ g(p_x) - g(p_y) &\geq \gamma^{-1}\langle p_x - p_y y - p_y\rangle + (\sigma/2)\|p_x - p_y\|^2. \end{aligned}$$

and summing, we have $0 \geq \gamma^{-1}\langle p_x - p_y y - x + p_x - p_y\rangle + \sigma\|p_x - p_y\|^2$ and hence

$$\langle p_x - p_y x - y\rangle \geq (1+\gamma\sigma)\|p_x - p_y\|^2. \tag{55}$$

Then, Cauchy-Schwarz inequality yields $\|p_x - p_y\|^2 \leq (1+\gamma\sigma)^{-1}\|p_x - p_y\|\|x - y\|$ and the statement follows. □

Now we are ready to provide the theorem of convergence for the proximal gradient algorithm.

Theorem 55 *Let $f\colon X \to \mathbb{R}$ be Lipschitz smooth with constant $L > 0$ and with modulus of strong convexity $\mu > 0$ and $g \in \Gamma_0(X)$ with modulus of strong convexity $\sigma \geq 0$. Suppose that $\gamma < 2/L$. Let x_* be the minimizer of $F := f + g$ and let $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 1. Then*

$$(\forall k \in \mathbb{N}) \quad \|x_k - x_*\| \leq q^k\|x_0 - x_*\|, \qquad q := \frac{1}{1+\gamma\sigma}\left(1 - \frac{2\gamma\mu L}{L+\mu}\right)^{1/2} \tag{56}$$

Moreover, if f is twice differentiable, then

$$(\forall k \in \mathbb{N}) \quad \|x_k - x_*\| \leq q^k\|x_0 - x_*\|, \qquad \tilde{q} := \begin{cases} \dfrac{1-\gamma\mu}{1+\gamma\sigma} & \text{if } \gamma \leq \dfrac{2}{L+\mu} \\ \dfrac{\gamma L - 1}{1+\gamma\sigma} & \text{if } \gamma \geq \dfrac{2}{L+\mu}. \end{cases} \tag{57}$$

Proof The statement follows from Theorems 51, 52, to 54 and the Banach-Caccioppoli theorem. □

Remark 56

(i) The best value of γ in (57) is achieved for $\gamma = 2/(\mu + L)$.
(ii) When $g = 0$ one can derive an explicit linear rate also in the function values. Indeed, in this case, since $\nabla f(x_*) = 0$, it follows from Fact 1(ii) that $f(x) - f(x_*) \leq (L/2)\|x - x_*\|^2$.

3.5 Convergence Analysis Under Geometric Assumptions

It is possible to show that strongly convex functions satisfy the following condition

$$f(x) - \inf f \leq \frac{1}{2\mu}\|\partial f(x)\|_-^2, \tag{58}$$

where $\|\partial f(x)\|_- = \inf\{\|u\| | u \in \partial f(x)\}$.

This condition is called Łojasiewicz inequality and can hold even for non-strongly convex functions and very recently has been the objective of intense research which has unveiled its connection with the *quadratic growth condition*

$$(\forall\, x \in X) \quad f(x) - \inf_X f \geq \frac{\mu}{2}\mathrm{dist}(x, \operatorname{argmin} f)^2 \tag{59}$$

and ultimately its critical role in achieving linear convergence in optimization algorithms. In this section, we study the convergence of the proximal gradient algorithm under Łojasiewicz-type inequalities.

We start with a major (although simple) example showing a function which is not strongly convex but satisfies the Łojasiewicz inequality and the quadratic growth condition above.

Example 57 Let $A\colon X \to Y$ be a bounded linear operator with closed range between two Hilbert spaces, $b \in Y$, and set

$$f\colon X \to \mathbb{R} \quad f(x) = \frac{1}{2}\|Ax - b\|^2. \tag{60}$$

Note that here we do not assume A^*A to be positive definite. Let b_* be the projection of b onto the range $R(A)$ of A. Then Pytagoras' theorem yields

$$(\forall\, x \in X) \quad f(x) = \frac{1}{2}\big(\|Ax - b_*\|^2 + \|b_* - b\|^2\big).$$

Thus, $f_* := \inf_X f = (1/2)\|b_* - b\|^2$. Now, let $x_* \in S := \operatorname{argmin} f = \big\{x \in X \mid Ax = b_*\big\}$, let $x \in X$ and set $x_p = P_S x$. We have $b_* = Ax_* = Ax_p$, and hence

$$f(x) - f_* = \frac{1}{2}\|Ax - b_*\|^2 = \frac{1}{2}\|A(x - x_*)\|^2 = \frac{1}{2}\|A(x - x_p)\|^2. \tag{61}$$

Moreover, since S is an affine set with direction $N(A)$, we have $x - x_p \in N(A)^\perp$. Now we introduce the pseudo inverse of A, which is a the bounded linear operator $A^\dagger\colon Y \to X$ satisfying, for every $u \in N(A)^\perp$, the equality $A^\dagger Au = u$, hence, $\|u\| \leq \|A^\dagger\|\|Au\|$. Therefore, using (61) we have

$$f(x) - f_* \geq \frac{1}{2}\|A^\dagger\|^{-2}\|x - x_p\|^2 = \frac{1}{2}\|A^\dagger\|^{-2}\mathrm{dist}(x, \operatorname{argmin} f)^2, \tag{62}$$

so that (59) holds with $\mu = \|A^\dagger\|^{-2}$. Moreover, $\nabla f(x) = A^*(Ax - b_*) = A^*A(x - x_*)$, and hence

$$\|\nabla f(x)\|^2 = \|A^*A(x - x_*)\|^2.$$

Thus, inequality (58) in this case reduces to

$$(\forall\, x \in X) \quad \mu\|A(x - x_*)\|^2 \leq \|A^*A(x - x_*)\|^2,$$

which is equivalent to

$$(\forall\, y \in R(A)) \quad \mu\|y\|^2 \leq \|A^* y\|^2. \tag{63}$$

Again, since (as before) for every $y \in R(A) = N(A^*)^\perp$, $\|y\| \leq \|(A^*)^\dagger\|\|A^* y\|$ and $(A^*)^\dagger = (A^\dagger)^*$, we have that (63) and hence (58) holds with $\mu = \|(A^\dagger)^*\|^{-2} = \|A^\dagger\|^{-2}$.

In the following we generalize condition (58).

Definition 58 Let $p \in [1, +\infty[$, let $F \in \Gamma_0(X)$ with $\operatorname{argmin} F \neq \emptyset$. We say that F is p-Łojasiewicz on sublevel sets if for every $t > \inf F$ there exists a constant $c_t > 0$ such that:

$$\forall x \in [\inf F < F \leq t], \quad (F(x) - \inf F)^{1-\frac{1}{p}} \leq c_t \|\partial F(x)\|_-,$$

where for a given set D, $\|D\|_- = \inf_{u \in D} \|u\|$. We will refer to this notion as *global* if $\sup_{t > \inf F} c_t < +\infty$.

Example 59 *(Convex piecewise polynomials)* A function $f : \mathbb{R}^d \to \mathbb{R}$ is a *convex piecewise polynomial* if it is convex, continuous, and $\mathbb{R}^d$ can be partitioned in a finite number of polyhedra $P_1, ..., P_s$ such that for all $i \in \{1, ..., s\}$, the restriction of f to P_i is a convex polynomial of degree $d_i \in \mathbb{N}$. The degree of f is defined as $\deg(f) := \max\{d_i \mid i \in \{1, ..., s\}\}$. Assume $\deg(f) > 0$. Convex piecewise polynomial functions are p-Łojasiewicz on sublevel sets with $p = 1 + (\deg(f) - 1)^d$. This result implies that piecewise linear functions ($\deg(f) = 1$) are 1-Łojasiewicz on sublevel sets and that convex piecewise quadratic functions ($\deg(f) = 2$) are 2-Łojasiewicz.

Example 60 *(L1 regularized least squares)* Let $f(x) = \alpha\|x\|_1 + (1/2)\|Ax - y\|^2$, for some linear operator $A : \mathbb{R}^d \to \mathbb{R}^n$, $y \in \mathbb{R}^n$ and $\alpha > 0$. Then f is convex piecewise polynomial of degree 2, thus it is 2-Łojasiewicz on sublevel sets.

Lemma 61 *Let $(r_k)_{k\in\mathbb{N}}$ be a real sequence being strictly positive and satisfying, for some $\kappa > 0$, $\alpha > 1$ and all $k \in \mathbb{N}$: $r_k - r_{k+1} \geq \kappa r_{k+1}^\alpha$. Define $\tilde{\kappa} := \min\{(\alpha - 1)\kappa, (\alpha - 1)\kappa^{\frac{\alpha-1}{\alpha}}, r_0^{1-\alpha}, \kappa^{1/\alpha} r_0^{1-\alpha}\}$. Then, for all $k \in \mathbb{N}$, $r_k \leq (\tilde{\kappa} k)^{-1/(\alpha-1)}$.*

The proof can be found in [50, Theorem 3.4].

Theorem 62 *Let $f : X \to \mathbb{R}$ be convex and differentiable with L-Lipschitz continuous gradient and let $g \in \Gamma_0(X)$. Set $F = f + g$ and suppose that F has a minimizer and that is p-Łojasiewicz on sublevel sets, for some $p \geq 1$. Let $\gamma < 2/L$ and $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 1 with $x_0 \in \operatorname{dom} F$. Then the sequence $(x_k)_{k\in\mathbb{N}}$ has finite length in X, meaning that $\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\| < +\infty$, and converges strongly to some $x_* \in \operatorname{argmin} F$. Moreover, there exists a constant b_p with explicit expression (see equation (71)), such that the following convergence rates hold, depending on the value of p, and of $\kappa := \gamma(2 - \gamma L)[2c_{F(x_0)}^2]^{-1}$:*

(i) *If* $p = 1$, *then* $x_k = x_*$ *for every* $k \geq (F(x_0) - \inf F)/\kappa$.
(ii) *If* $p \in]1, 2[$, *for all* $k \in \mathbb{N}$,

$$F(x_{k+1}) - \inf F \leq \left(\frac{F(x_k) - \inf F}{\kappa}\right)^{\frac{p}{2(p-1)}} \text{ and } \|x_{k+1} - x_*\| \leq b_p (F(x_k) - \inf F)^{1/2},$$

(iii) *If* $p = 2$, *for all* $k \in \mathbb{N}$,

$$F(x_{k+1}) - \inf F \leq \frac{1}{1+\kappa}(F(x_k) - \inf F) \text{ and } \|x_{k+1} - x_*\| \leq b_2 \frac{(F(x_0) - \inf F)^{1/2}}{(1+\kappa)^{k/2}}.$$

(iv) *If* $p \in]2, +\infty[$, *for all* $k \in \mathbb{N}$,

$$F(x_k) - \inf F \leq c_p k^{-\frac{p}{p-2}} \quad \text{and} \quad \|x_{k+1} - x_*\| \leq b_p c_p^{1/(p-2)} k^{-\frac{1}{p-2}}.$$

Proof We first show that $(x_k)_{k\in\mathbb{N}}$ has finite length. Since $\inf F > -\infty$ then $r_k := F(x_k) - \inf F \in [0, +\infty[$, and Theorem 47(iii) yields

$$a\|x_{k+1} - x_k\|^2 \leq r_k - r_{k+1}, \qquad \text{with } a = \frac{1}{\gamma} - \frac{L}{2}. \tag{64}$$

By definition of Algorithm 1, we have $x_k - \gamma\nabla f(x_{k+1}) - x_{k+1} \in \gamma\partial g(x_{k+1})$ and hence

$$x_k - \gamma\nabla f(x_k) - x_{k+1} + \gamma\nabla f(x_{k+1}) \in \partial\gamma F(x_{k+1}). \tag{65}$$

This implies, together with the nonexpansiveness of $\mathrm{Id} - \gamma\nabla f$ (see Proposition 43), that

$$\gamma \inf_{u\in\partial F(x_{k+1})} \|u\| \leq \|x_k - \gamma\nabla f(x_k) - (x_{k+1} - \gamma\nabla f(x_{k+1}))\| \leq \|x_k - x_{k+1}\|. \tag{66}$$

If there exists $k \in \mathbb{N}$ such that $r_k = 0$ then the algorithm would stop after a finite number of iterations (see (64)), therefore it is not restrictive to assume that $r_k > 0$ for all $k \in \mathbb{N}$. Since $(F(x_k))_{k\in\mathbb{N}}$ is decreasing by Theorem 47(iii), and $x_0 \in \mathrm{dom}\, F$, $x_k \in [\inf F < F \leq F(x_0)]$ for every $k \geq 1$. We set $\varphi(t) := pt^{1/p}$ and $F_0 = F(x_0)$, so that the Łojasiewicz inequality at $x_k \in [\inf F < F \leq F_0]$ can be rewritten as

$$(\forall k \in \mathbb{N}) \quad 1 \leq c_{F_0}\varphi'(r_k)\|\partial F(x_k)\|_-. \tag{67}$$

Combining (64), (66), and (67), and using the concavity of φ, we obtain for all $k \in \mathbb{N}$:

$$\|x_{k+1} - x_k\|^2 \leq \frac{c_{F_0}}{\gamma a}\varphi'(r_k)(r_k - r_{k+1})\|x_k - x_{k-1}\| \leq \frac{c_{F_0}}{\gamma a}(\varphi(r_k) - \varphi(r_{k+1}))\|x_k - x_{k-1}\|.$$

By taking the square root on both sides, and using Young's inequality, we obtain

$$(\forall k \in \mathbb{N}) \quad 2\|x_{k+1} - x_k\| \leq \frac{c_{F_0}}{\gamma a}(\varphi(r_k) - \varphi(r_{k+1})) + \|x_k - x_{k-1}\|. \tag{68}$$

Sum this inequality, and reorder the terms to finally obtain

$$(\forall k \geq 1) \quad \sum_{k=1}^{K} \|x_{k+1} - x_k\| \leq \frac{c_{F_0}}{\gamma a}\varphi(r_1) + \|x_1 - x_0\|.$$

We deduce that $(x_k)_{k\in\mathbb{N}}$ has finite length and therefore converges strongly to some x_*. Moreover, from (66) and the strong closedness of $\partial f : X \rightrightarrows X$, we conclude that $0 \in \partial f(x_*)$. We next show a preliminary inequality which will be useful to prove the rates for $(\|x_k - x_*\|)_{k\in\mathbb{N}}$. Let $K \in \mathbb{N}$ and $1 \leq k \leq K$, recall that $\varphi(t) = pt^{1/p}$, and sum the inequality in (68) between k and K to obtain

$$\|x_K - x_k\| \leq \sum_{n=k}^{K} \|x_{n+1} - x_n\| \leq \frac{pc_{F_0}}{a\gamma} r_k^{1/p} + \|x_k - x_{k-1}\|.$$

Passing to the limit for $K \to \infty$, using (64), and the fact that r_k is decreasing, we derive

$$(\forall k \geq 1) \quad \|x_* - x_k\| \leq \frac{pc_{F_0}}{a\gamma} r_{k-1}^{1/p} + \frac{1}{\sqrt{a}} r_{k-1}^{1/2}. \tag{69}$$

Next we prove the convergence rates. We first derive rates for the sequence of values r_k, from which we will derive the rates for the iterates thanks to (69). Equations (64) and (66) and the Łojasiewicz inequality at $x_{k+1} \in [\inf F < F \leq F_0]$ yield

$$r_k - r_{k+1} \geq a\|x_{k+1} - x_k\|^2 \geq a\gamma^2 \|\partial F(x_{k+1})\|_-^2 \geq a\gamma^2 c_{F_0}^{-2} r_{k+1}^{2-2/p},$$

which we write more compactly as

$$(\forall k \in \mathbb{N}) \quad r_k - r_{k+1} \geq \kappa r_{k+1}^{\alpha}, \quad \text{with } \alpha = 2(p-1)p^{-1} \text{ and } \kappa := a\gamma^2 c_{F_0}^{-2}. \tag{70}$$

The rates for the values are derived from the analysis of the sequences satisfying the inequality in (70), which is recalled in Lemma 61. Depending on the value of p, we obtain different rates.

(i): Since $p = 1$, we deduce from (70) that for all $k \in \mathbb{N}$ $r_{k+1} \leq r_k - \kappa$. Since the sequence $(r_k)_{k\in\mathbb{N}}$ is decreasing and positive, this implies $k \leq r_0\kappa^{-1}$.

(ii): Since $p \in]1, 2[$ we have $\alpha \in]0, 1[$. Thus, the positivity of r_{k+1} and (70) imply that for all $k \in \mathbb{N}$, $r_k \geq \kappa r_{k+1}^{\alpha}$ and hence $r_{k+1} \leq \kappa^{-1/\alpha} r_k^{1/\alpha}$, meaning that r_k converges Q-superlinearly to zero. In addition, we have $r_{k-1}^{1/p} = r_{k-1}^{1/p-1/2} r_{k-1}^{1/2} \leq r_0^{1/p-1/2} r_{k-1}^{1/2}$ and (69) implies $\|x_k - x_*\| \leq b_p r_{k-1}^{1/2}$, with $b_p = pc_{F_0} r_0^{1/p-1/2}/(a\gamma) + (1/\sqrt{a})$.

(iii): If $p = 2$, then $\alpha = 1$ and (70) yields that for all $k \in \mathbb{N}$, $r_{k+1} \leq (1+\kappa)^{-1} r_k$, so that $r_k \leq (1+\kappa)^{-k} r_0$. Moreover, from (69) we derive that,

$$(\forall k \geq 1) \quad \|x_* - x_k\| \leq b_2 r_{k-1}^{1/2}.$$

where $b_2 = 2c_{F_0}/a\gamma + 1/\sqrt{a}$.

(iv): If $p \in]2, +\infty[$, then $\alpha \in]1, 2[$ and (70) and Lemma 61 imply that $r_{k+1} \leq c_p(k+1)^{-p/(p-2)}$, where

$$c_p = \min\left\{\left[\frac{\kappa(p-2)}{p}\right]^{-\frac{p}{p-2}}, \left[\frac{p-2}{p}\right]^{-\frac{p}{p-2}} \kappa^{-\frac{p}{2p-2}}, r_0, \kappa^{-\frac{p^2}{2(p-1)(p-2)}} r_0\right\}. \tag{71}$$

Note that $r_{k-1}^{1/2} \leq r_0^{\frac{1}{2}-\frac{1}{p}} r_{k-1}^{1/p}$, and therefore, defining $b_p = pc_{F_0}/\gamma + (r_0)^{1/2}a^{-1/2} r_0^{-1/p}$, we derive from (69) that $\|x_k - x_*\| \leq b_p r_{k-1}^{1/p}$ for every $k \geq 1$. □

Remark 63 Note that the rates range from the finite termination, for $p = 1$, to the worst-case rates presented in Theorem 47, when p tends to $+\infty$. The bigger is p, the more the rates for the objective function values become closer to $o(k^{-1})$, and the rates of its iterates become arbitrarily slow.

3.6 Accelerations

Proximal gradient methods are very simple and have a very low cost per iteration, but often they converge slowly, both in practice and in theory (see Theorem 47). In this section, we consider the class of accelerated proximal gradient algorithms, which are only slightly more complicated than the basic proximal gradient methods, but have an improved convergence rate. While in the proximal gradient method, only the information obtained in the previous step is used to build the next iterate, accelerated methods are multistep methods, namely they take into account previous iterates to improve the convergence. The most popular accelerated multistep method is due to Nesterov and is also known as *Fast Iterative Soft Thresholding Algorithm* (FISTA). We consider the same setting of the previous sections.

Algorithm 2 *(*Accelerated proximal gradient method*) Let* $0 < \gamma \leq 1/L$ *and let* $(t_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ *be such that* $t_0 = 1$, $t_k \geq 1$, *and for every integer* $k \geq 1$, $t_k^2 - t_k \leq t_{k-1}^2$. *Let* $x_0 = y_0 \in X$ *and define*

$$\begin{aligned} &\textit{for } k = 0, 1, \ldots \\ &\quad x_{k+1} = \operatorname{prox}_{\gamma g}(y_k - \gamma \nabla f(y_k)) \\ &\quad \beta_{k+1} = \frac{t_k - 1}{t_{k+1}} \\ &\quad y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k). \end{aligned} \tag{72}$$

3.6.1 Dynamical Systems Interpretation

One of the crucial observations that lead to a whole stream of literature and allowed to give a physical interpretation of this kind of algorithms is the link of accelerated algorithms with the trajectories of a second-order continuous dynamical system.

Let us consider a heavy ball of mass m in the potential field $\nabla f + \partial g$ under the force of friction, or "viscosity" controlled by a function $p(t) > 0$. The motion $x(t)$ of the heavy ball is described by the following second-order differential inclusion:

$$m\ddot{x} \in -\nabla f(x(t)) - \partial g(x(t)) - p(t)\dot{x}(t) \tag{73}$$

Intuitively, ignoring existence issues, the heavy ball reaches the minimizer of $f + g$ for $t \to +\infty$, due to the loss of energy caused by the friction. In addition, the friction avoids the zig-zagging effect, which is one of the causes that slows down gradient type methods. We consider a scenario where the viscosity coefficient is of the form $p(t) = \alpha/t$ which turned out to be crucial in the achievement of accelerated rates:

$$0 \in \ddot{x} + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) + \partial g(x(t)). \tag{74}$$

We next show that Algorithm 2 can be seen as a discretization of (74). To this aim, we discretize implicitly with respect to the nonsmooth function g and explicitly with respect to the smooth one f. Let $h > 0$ be a fixed time step, and set $t_k = (\tau_0 + k)h$, $x_k = x(t_k)$. The suggested implicit/explicit discretization strategy reads as

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{(\tau_0 + k)h^2}(x_k - x_{k-1}) + \partial g(x_{k+1}) + \nabla f(y_k) \ni 0,$$

where y_k will be suitably chosen as a linear combination of x_k and x_{k-1}. Rearranging the terms in 3.6.1 we derive

$$x_{k+1} + h^2\partial g(x_{k+1}) \ni x_k + \left(1 - \frac{\alpha}{\tau_0 + k}\right)(x_k - x_{k-1}) - h^2\nabla f(y_k).$$

A choice of y_k classically made in the literature is

$$y_k = x_k + \left(1 - \frac{\alpha}{\tau_0 + k}\right)(x_k - x_{k-1}).$$

Recalling the definition of proximal operator, and setting $\gamma = h^2$ we can rewrite 3.6.1 as

$$\begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{\tau_0 + k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{\gamma g}(y_k - \gamma\nabla f(y_k)), \end{cases}$$

which is an instance of Algorithm 2 for a specific choice of parameters t_k's (see next section).

3.6.2 Convergence Analysis

We start with few results concerning the sequence of the parameters t_k's.

Proposition 64 *Suppose that* $t_0 = 1$ *and for every integer* $k \geq 1$

$$t_k \geq 0 \ \text{ and } \ t_k^2 - t_k - t_{k-1}^2 = -b - ct_k \tag{75}$$

for some $c \in [0, 1[$ *and* $b \in [0, 1 - c]$. *Then condition* (75) *is equivalent to*

$$t_k = \frac{1-c}{2} + \sqrt{\left(\frac{1-c}{2}\right)^2 + t_{k-1}^2 - b}. \tag{76}$$

Moreover, the following hold.

(i) *For every integer* $k \geq 1$, $1 \leq t_{k-1} \leq t_k \leq 1 - c + t_{k-1}$.
(ii) *Suppose that* $2\sqrt{b} \leq 1 - c$. *Then, for every integer* $k \geq 1$, $(1-c)/2 + t_{k-1} \leq t_k$. *Hence* $k(1-c)/2 \leq t_k - 1 \leq k(1-c)$.

Proof The discriminant of the quadratic equation in (75) (in the unknown t_k) is $\Delta_k = (1-c)^2 + 4(t_{k-1}^2 - b)$. Then it is clear that if $t_{k-1} \geq 1$, then $\Delta_k > 0$, the positive solution of (75) is (76) and $t_k \geq (1-c)/2 + \sqrt{(1-c)^2/4 + 1 - b} \geq 1$, since $b \leq 1 - c$. Vice versa, if $t_{k-1} \geq 1$, then (76) $\Rightarrow$ (75). In the end, if $t_{k-1} \geq 1$, then (76) and (75) are equivalent and in such case $t_k \geq 1$. So, the first part of the statement follows by an induction argument since $t_0 = 1$.

(i): We derive from (75) that $t_k^2 - t_{k-1}^2 = -b + (1-c)t_k \geq -b + 1 - c \geq 0$, hence $t_{k-1} \leq t_k$. Moreover, it follows from (76) that

$$\left(t_k - \frac{1-c}{2}\right)^2 \leq \left(\frac{1-c}{2}\right)^2 + t_{k-1}^2 \leq \left(\frac{1-c}{2} + t_{k-1}\right)^2. \tag{77}$$

Thus, $t_k - (1-c)/2 \leq (1-c)/2 + t_{k-1}$ and hence $t_k \leq 1 - c + t_{k-1}$. The statement follows.

(ii): Suppose that $1 - c \geq 2\sqrt{b}$. Then $(1-c)^2/4 - b \geq 0$ and hence, we have $t_k = (1-c)/2 + \sqrt{(1-c)^2/4 - b + t_{k-1}^2} \geq (1-c)/2 + t_{k-1}$ and the first part of the statement follows. Next, summing the inequalities $(1-c)/2 \leq t_i - t_{i-1} \leq (1-c)$ from $i = 1$ to $i = k$, we have $k(1-c)/2 \leq t_k - 1 \leq k(1-c)$. □

Remark 65 The following are two special cases of (75).

$$t_k = \frac{1}{2} + \sqrt{\frac{1}{4} + t_{k-1}^2} \quad \text{and} \quad t_k = \frac{k+a}{a} \ (\text{with}) \ a \geq 2, \tag{78}$$

which are obtained from (75) with $(b, c) = (0, 0)$ and $(b, c) = (1/a^2, (a-2)/a)$ respectively. Note that in both cases $1 - c \geq 2\sqrt{b}$ (and in the first case, in virtue of Proposition 64(ii), we have $t_k \geq (k+2)/2$).

Remark 66 Suppose that the t_k's satisfy (75) with $2\sqrt{b} \leq 1-c$. Then, since $t_k^2 - (1-c)t_k \leq t_{k-1}^2$, we have, for $k \geq 2$,

$$\frac{t_k^2}{t_{k-1}^2} \leq \frac{t_k^2}{t_k(t_k-(1-c))} = 1 + \frac{1-c}{t_k-(1-c)} \leq 1 + \frac{1-c}{t_{k-2}} \leq 2-c, \tag{79}$$

where in the second last inequality we used that $t_k \geq (1-c)/2 + t_{k-1} \geq 1-c+t_{k-2}$. Note that, in view of Proposition 64(i), $t_1 \leq 2-c$. Therefore, since $2-c > 1$, we have

$$(\forall\, k \in \mathbb{N},\ k \geq 1) \qquad t_k \leq (2-c)t_{k-1}. \tag{80}$$

Lemma 67 *Let $y \in X$ and set $x = \operatorname{prox}_{\gamma g}(y - \gamma \nabla f(y))$, with $\gamma \leq 1/L$. Then*

$$(\forall\, z \in X) \quad F(x) + \frac{\|x-z\|^2}{2\gamma} \leq F(z) + \frac{\|z-y\|^2}{2\gamma}.$$

Proof It follows from the definition of the proximity operator that

$$\begin{aligned} x &= \operatorname{argmin}_{z\in X} \left\{ \gamma g(z) + \frac{1}{2}\|y-z-\gamma\nabla f(y)\|^2 \right\} \\ &= \operatorname{argmin}_{z\in X} \left\{ g(z) + \frac{1}{2\gamma}\|y-z\|^2 + \langle z-y, \nabla f(y)\rangle \right\}. \end{aligned}$$

Therefore, since $z \mapsto g(z) + \frac{1}{2\gamma}\|y-z\|^2 + \langle z-y, \nabla f(y)\rangle$ is γ^{-1}-strongly convex and x is its minimizer, it follows from (10) that

$$\begin{aligned} \frac{1}{2\gamma}\|z-x\|^2 \leq\ & g(z) + \frac{1}{2\gamma}\|y-z\|^2 + \langle z-y, \nabla f(y)\rangle \\ & - \Big(g(x) + \frac{1}{2\gamma}\|y-x\|^2 + \langle x-y, \nabla f(y)\rangle \Big) \end{aligned}$$

hence

$$\begin{aligned} g(x) + \underbrace{\frac{1}{2\gamma}\|y-x\|^2 + \langle x-y, \nabla f(y)\rangle}_{(a)} + \frac{1}{2\gamma}\|z-x\|^2 \\ \leq g(z) + \frac{1}{2\gamma}\|y-z\|^2 + \langle z-y, \nabla f(y)\rangle\,. \end{aligned}$$

Now, since f is L-Lipschitz continuous and $\gamma \leq 1/L$, it follows from Theorem 1 that

$$f(x) - f(y) \leq \langle x-y, \nabla f(y)\rangle + \frac{L}{2}\|x-y\|^2 \leq \underbrace{\langle x-y, \nabla f(y)\rangle + \frac{1}{2\gamma}\|x-y\|^2}_{(a)}.$$

Therefore,

$$f(x) + g(x) + \frac{1}{2\gamma}\|z - x\|^2 \leq f(y) + g(z) + \frac{1}{2\gamma}\|y - z\|^2 + \langle z - y, \nabla f(y)\rangle$$
$$\leq f(z) + g(z) + \frac{1}{2\gamma}\|y - z\|^2,$$

where in the last inequality we used that $f(y) + \langle z - y, \nabla f(y)\rangle \leq f(z)$, due to the convexity of f. □

We now present the first of the two results of the section, which concerns the convergence in value for Algorithm 2. Next, we will address the convergence of the iterates under slightly stronger assumptions on the sequence of parameters t_k's.

Theorem 68 *Let $f : X \to \mathbb{R}$ be convex and differentiable with L-Lipschitz continuous gradient and let $g \in \Gamma_0(X)$. Set $F = f + g$ and suppose that F has a minimizer. Define $(x_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$ according to Algorithm 2. Then*

$$(\forall k \in \mathbb{N}, k \geq 1) \quad F(x_k) - \min F \leq \frac{\operatorname{dist}(x_0, \operatorname{argmin} F)^2}{2\gamma t_{k-1}^2}.$$

Moreover, if the parameters t_k's are defined according to Proposition 64 with $1 - c \geq 2\sqrt{b}$, then $F(x_k) - \min F = O(1/k^2)$.

Proof It follows from the definition of y_{k+1} in Algorithm 2 that, for every $k \in \mathbb{N}$,

$$y_{k+1} = \left(1 - \frac{1}{t_{k+1}}\right)x_{k+1} + \frac{1}{t_{k+1}}\big(\underbrace{x_k + t_k(x_{k+1} - x_k)}_{v_{k+1}}\big)$$

Therefore, for every $k \in \mathbb{N}$,

$$y_k = \left(1 - \frac{1}{t_k}\right)x_k + \frac{1}{t_k}v_k \qquad (v_0 := y_0) \tag{81}$$

Moreover, it follows from the definition of v_{k+1} that $v_{k+1} - x_k = t_k(x_{k+1} - x_k)$ and hence

$$x_{k+1} = x_k + \frac{1}{t_k}(v_{k+1} - x_k) = \left(1 - \frac{1}{t_k}\right)x_k + \frac{1}{t_k}v_{k+1}. \tag{82}$$

Also, by Lemma 67, with $y = y_k$ and $x = x_{k+1}$, we have

$$(\forall z \in X) \quad F(x_{k+1}) + \frac{\|x_{k+1} - z\|^2}{2\gamma} \leq F(z) + \frac{\|z - y_k\|^2}{2\gamma}. \tag{83}$$

Now, let $x_* \in \operatorname{argmin} F$ and set

$$z = \left(1 - \frac{1}{t_k}\right) x_k + \frac{1}{t_k} x_*.$$

Then, we derive from (81) and (151) that

$$x_{k+1} - z = \frac{1}{t_k}(v_{k+1} - x_*) \quad \text{and} \quad y_k - z = \frac{1}{t_k}(v_k - x_*).$$

Therefore, it follows from (83) and the convexity of F (considering that z is a convex combination of x_k and x_*) that

$$\begin{aligned} F(x_{k+1}) + \frac{\|v_{k+1} - x_*\|^2}{2\gamma t_k^2} &\le F(z) + \frac{\|v_k - x_*\|^2}{2\gamma t_k^2} \\ &\le \left(1 - \frac{1}{t_k}\right) F(x_k) + \frac{1}{t_k} F(x_*) + \frac{\|v_k - x_*\|^2}{2\gamma t_k^2}. \end{aligned}$$

Summing $-F(x_*)$ to both terms of the above inequality and setting $r_k = F(x_k) - F(x_*)$, we get

$$r_{k+1} + \frac{\|v_{k+1} - x_*\|^2}{2\gamma t_k^2} \le \left(1 - \frac{1}{t_k}\right) r_k + \frac{\|v_k - x_*\|^2}{2\gamma t_k^2}$$

and hence, multiplying by t_k^2

$$t_k^2 r_{k+1} + \frac{\|v_{k+1} - x_*\|^2}{2\gamma} \le t_k(t_k - 1) r_k + \frac{\|v_k - x_*\|^2}{2\gamma}. \tag{84}$$

Now we set, for every integer $k \ge 1$, $\mathcal{E}_k = t_{k-1}^2 r_k + \|v_k - x_*\|^2/(2\gamma)$. Then, by using $t_k^2 - t_k - t_{k-1}^2 \le -c t_k$, we have

$$(\forall\, k \in \mathbb{N},\ k \ge 1) \quad \mathcal{E}_{k+1} \le t_k(t_k - 1) r_k + \frac{\|v_k - x_*\|^2}{2\gamma} \le -c t_k r_k + \mathcal{E}_k. \tag{85}$$

Therefore, $\mathcal{E}_k$ is decreasing and hence, using (84) with $k = 0$, we have, for all $k \ge 1$

$$t_{k-1}^2 r_k \le \mathcal{E}_k \le \mathcal{E}_1 = r_1 + \frac{\|v_1 - x_*\|^2}{2\gamma} \le \frac{\|v_0 - x_*\|^2}{2\gamma} = \frac{\|x_0 - x_*\|^2}{2\gamma}.$$

Since x_* is an arbitrary element of $\operatorname{argmin} F$, the first part of the statement follows. The second part of the statement follows from Proposition 64(ii) and the fact that, for every integer $k \ge 1$, $2t_{k-1} \ge 2 + (k-1)(1-c) = k(1-c) + 1 + c \ge k(1-c)$. □

Remark 69 The quantity $\mathcal{E}_k$ introduced in the proof of Theorem 68 can be seen as a discretization of a Lyapunov function of the continuous dynamical system (74).

We now start the analysis of the convergence of the iterates.

Proposition 70 *Under the assumptions of Theorem 68 suppose additionally that for every integer $k \in \mathbb{N}$,*

$$t_k \geq 1 \ \text{ and } \ t_k^2 - t_k - t_{k-1}^2 \leq -ct_k \tag{86}$$

for some $c \in]0, 1[$. Then the following hold.

(i) $\sum_{k=0}^{\infty} t_k(F(x_k) - \inf F) < +\infty$.
(ii) $\sum_{k=1}^{\infty} t_k \|x_{k+1} - x_k\|^2 < +\infty$.

Proof Let r_k and $\mathcal{E}_k$ be defined as in the proof of Theorem 68. It follows from (85) that, for every integer $k \geq 1$,

$$ct_k r_k \leq \mathcal{E}_k - \mathcal{E}_{k+1}. \tag{87}$$

Hence $c\sum_{k=1}^{\infty} t_k r_k \leq \mathcal{E}_1 \leq \|x_0 - x_*\|^2/(2\gamma)$. Concerning the second statement, it follows from (83) with $z = x_k$, that

$$F(x_{k+1}) + \frac{\|x_{k+1} - x_k\|^2}{2\gamma} \leq F(x_k) + \frac{\|x_k - y_k\|^2}{2\gamma}. \tag{88}$$

Subtracting $-\inf F$ and recalling the definition of y_k in Algorithm 2, we get

$$r_{k+1} + \frac{\|x_{k+1} - x_k\|^2}{2\gamma} \leq r_k + \frac{(t_{k-1} - 1)^2}{t_k^2} \frac{\|x_k - x_{k-1}\|^2}{2\gamma}, \tag{89}$$

which, multiplied by t_k^2 yields

$$\frac{1}{2\gamma}\Big(t_k^2\|x_{k+1} - x_k\|^2 - (t_{k-1} - 1)^2\|x_k - x_{k-1}\|^2\Big) \leq t_k^2(r_k - r_{k-1}).$$

Since $(t_{k-1} - 1)^2 = t_{k-1}^2 + 1 - 2t_{k-1}$, we have

$$\begin{aligned}\frac{1}{2\gamma}\Big(t_k^2\|x_{k+1} - x_k\|^2 - t_{k-1}^2\|x_k - x_{k-1}\|^2 + (2t_{k-1} - 1)\|x_k - x_{k-1}\|^2\Big)\\ \leq t_{k-1}^2 r_k - t_k^2 r_{k+1} + (t_k^2 - t_{k-1}^2)r_k.\end{aligned} \tag{90}$$

Summing the above inequality from $k = 1$ to $k = K$, and recalling that $t_0 = 1$, we have

$$\frac{1}{2\gamma}\Big(t_K^2\|x_{K+1}-x_K\|^2+\sum_{k=2}^{K}(2t_{k-1}-1)\|x_k-x_{k-1}\|^2\Big)$$
$$\leq r_1-t_K^2 r_{K+1}+\sum_{k=1}^{K}(t_k^2-t_{k-1}^2)r_k$$
$$\leq r_1+(1-c)\sum_{k=1}^{K}t_k r_k,$$

where we used the fact that, by 70, we have $t_k^2-t_{k-1}^2\leq(1-c)t_k$. Therefore, since $t_{k-1}\leq 2t_{k-1}-1$ (being $t_{k-1}\geq 1$),

$$\sum_{k=2}^{K}t_{k-1}\|x_k-x_{k-1}\|^2\leq\sum_{k=2}^{K}(2t_{k-1}-1)\|x_k-x_{k-1}\|^2\leq 2\gamma\Big(r_1+(1-c)\sum_{k=1}^{K}t_k r_k\Big) \tag{91}$$

and the statement follows from (i). □

We need two additional results concerning the convergence of numerical sequences.

Lemma 71 *Let $(a_k)_{k\in\mathbb{N}}$, $(\varepsilon_k)_{k\in\mathbb{N}}$ be sequences in $\mathbb{R}_+$ such that $\sum_{k\in\mathbb{N}}\varepsilon_k<+\infty$ and*

$$(\forall\, k\in\mathbb{N})\quad a_{k+1}\leq a_k+\varepsilon_k. \tag{92}$$

Then $(a_k)_{k\in\mathbb{N}}$ is convergent.

Proof Define $u_k=a_k+\sum_{i=k}^{+\infty}\varepsilon_i$. Then it follows from (92) that $u_{k+1}=a_{k+1}+\sum_{i=k+1}^{+\infty}\varepsilon_i\leq a_k+\sum_{i=k}^{+\infty}\varepsilon_i=u_k$, so that $(u_k)_{k\in\mathbb{N}}$ is decreasing and hence convergent. Then, by definition of u_k, $a_k=u_k-\sum_{i=k}^{+\infty}\varepsilon_i$ and hence $(a_k)_{k\in\mathbb{N}}$ is convergent too. □

Lemma 72 *Suppose that the sequence of parameters t_k's satisfy equation (75) in Proposition 64 with $2\sqrt{b}\leq 1-c$. Let $(a_k)_{k\geq 1}$ and $(b_k)_{k\geq 1}$ be two positive sequences such that*

$$(\forall\, k\in\mathbb{N},\ k\geq 1)\qquad a_{k+1}\leq\frac{t_{k-1}-1}{t_k}a_k+b_k, \tag{93}$$

If $(t_k b_k)_{k\in\geq 1}$ is summable, then $(a_k)_{k\in\geq 1}$ is summable.

Proof Let $k\in\mathbb{N}$ with $k\geq 1$. Multiplying equation (93) by t_k^2 and using the relation $t_k^2-t_k\leq t_{k-1}^2$ and the fact that $t_{k-1}\leq t_k$, we have

$$t_k^2 a_{k+1}\leq t_k(t_{k-1}-1)a_k+t_k^2 b_k\leq t_k(t_k-1)a_k+t_k^2 b_k\leq t_{k-1}^2 a_k+t_k^2 b_k. \tag{94}$$

Hence

$$t_{k-1}^2 a_k - a_1 = \sum_{i=1}^{k-1} (t_i^2 a_{i+1} - t_{i-1}^2 a_i) \leq \sum_{i=1}^{k-1} t_i^2 b_i. \tag{95}$$

Then, dividing by t_{k-1}^2, we obtain

$$a_k \leq \frac{a_1}{t_{k-1}^2} + \frac{1}{t_{k-1}^2} \sum_{i=1}^{k-1} t_i^2 b_i \tag{96}$$

and hence

$$\begin{aligned} \sum_{j=1}^{k} a_j &\leq \sum_{j=1}^{k} \frac{a_1}{t_{j-1}^2} + \sum_{j=1}^{k} \sum_{i=1}^{j-1} \frac{1}{t_{j-1}^2} t_i^2 b_i \\ &= \sum_{j=1}^{k} \frac{a_1}{t_{j-1}^2} + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{1}{t_{j-1}^2} t_i^2 b_i. \end{aligned} \tag{97}$$

Now we analyze the term $\sum_{j=i+1}^{k} 1/t_{j-1}^2$. Let $j \in \mathbb{N}$ with $j \geq 2$. Since, by assumption, $t_j(t_j - (1-c)) \leq t_{j-1}^2$ and $t_j \geq (1-c)/2 + t_{j-1} \geq (1-c) + t_{j-2}$, we have

$$\begin{aligned} \frac{1}{t_{j-1}^2} &\leq \frac{1}{t_j(t_j - (1-c))} \\ &= \frac{1}{1-c} \left(\frac{1}{t_j - (1-c)} - \frac{1}{t_j} \right) \\ &\leq \frac{1}{1-c} \left(\frac{1}{t_{j-2}} - \frac{1}{t_{j-1}} + \frac{1}{t_{j-1}} - \frac{1}{t_j} \right). \end{aligned}$$

Hence, for $i \geq 1$ and $k \geq 2$,

$$\begin{aligned} \sum_{j=i+1}^{k} \frac{1}{t_{j-1}^2} &\leq \frac{1}{1-c} \left[\sum_{j=i+1}^{k} \left(\frac{1}{t_{j-2}} - \frac{1}{t_{j-1}} \right) + \sum_{j=i+1}^{k} \left(\frac{1}{t_{j-1}} - \frac{1}{t_j} \right) \right] \\ &= \frac{1}{1-c} \left(\frac{1}{t_{i-1}} - \frac{1}{t_{k-1}} + \frac{1}{t_i} - \frac{1}{t_k} \right) \\ &\leq \frac{3-c}{1-c} \frac{1}{t_i}, \end{aligned}$$

where in the last inequality we used that $t_j \leq (2-c) t_{j-1}$ (see Remark 66). In the end, it follows from (97) that

$$\sum_{j=1}^{k} a_j \leq \sum_{j=1}^{k} \frac{a_1}{t_{j-1}^2} + \frac{3-c}{1-c} \sum_{i=1}^{k-1} t_i b_i$$
$$\leq a_1 + \frac{a_1(3-c)}{1-c} \frac{1}{t_1} + \frac{3-c}{1-c} \sum_{i=1}^{k-1} t_i b_i.$$

The statement follows. □

We are finally ready for the second main result of this section which addresses the convergence of the iterates of Algorithm 2.

Theorem 73 *Under the assumptions of Theorem 68, suppose additionally that the parameters t_k's satisfy the equation (75) with $c > 0$. Then $x_k \rightharpoonup x_*$ for some $x_* \in$ argmin F.*

Proof We invoke Opial's Lemma 29. We first prove that weak cluster points of $(x_k)_{k\in\mathbb{N}}$ belong to argmin F. We note that Theorem 68 yields that $F(x_k) \to \inf F$. Let $(x_{k_n})_{n\in\mathbb{N}}$ be a weakly convergent subsequence with $x_{k_n} \rightharpoonup x_*$. Since F is weakly lower semicontinuous, we have $F(x_*) \leq \liminf_n F(x_{k_n}) = \lim_k F(x_k) = \inf F$ and hence $x_* \in$ argmin F. We now prove that for every $x_* \in$ argmin F, the sequence $(\|x_k - x_*\|)_{k\in\mathbb{N}}$ is convergent. Let $x_* \in$ argmin F and set $h_k = \|x_k - x_*\|^2/2$ and $\delta_k = (1/2)\|x_k - x_{k+1}\|^2$. Then, since $\|x_k - x_*\|^2 = \|x_k - x_{k+1}\|^2 + \|x_{k+1} - x_*\|^2 + 2\langle x_k - x_{k+1}, x_{k+1} - x_*\rangle$ and $y_k - x_k = \beta_k(x_k - x_{k-1})$, we have

$$h_k - h_{k+1} = \delta_k + \langle x_k - x_{k+1}, x_{k+1} - x_*\rangle \tag{98}$$
$$= \delta_k - \beta_k \langle x_k - x_{k-1}, x_{k+1} - x_*\rangle + \langle y_k - x_{k+1}, x_{k+1} - x_*\rangle . \tag{99}$$

Now we note that, by definition of x_{k+1} and the fact that $x_* \in$ argmin F, we have

$$y_k - x_{k+1} - \gamma\nabla f(y_k) \in \partial\gamma g(x_{k+1}) \quad \text{and} \quad -\gamma\nabla f(x_*) \in \partial\gamma g(x_*).$$

Hence, using the monotonicity of ∂g (see Sect. 2.4), we have

$$\langle x_{k+1} - x_*, y_k - x_{k+1} - \gamma\nabla f(y_k) + \gamma\nabla f(x_*)\rangle \geq 0$$

which yields, in virtue of Fact 1(iv), that

$$\begin{aligned}\langle x_{k+1} - x_*, y_k - x_{k+1}\rangle &\geq \gamma\langle x_{k+1} - x_*, \nabla f(y_k) - \nabla f(x_*)\rangle \\ &= \gamma\langle y_k - x_*, \nabla f(y_k) - \nabla f(x_*)\rangle + \gamma\langle x_{k+1} - y_k, \nabla f(y_k) - \nabla f(x_*)\rangle \\ &\geq \frac{\gamma}{L}\|\nabla f(y_k) - \nabla f(x_*)\|^2 - \gamma\|x_{k+1} - y_k\|\|\nabla f(y_k) - \nabla f(x_*)\| \\ &\geq -\frac{\gamma L}{4}\|x_{k+1} - y_k\|^2,\end{aligned}$$

where in the last inequality we minorized the function $\alpha \mapsto (1/L)\alpha^2 - \|x_{k+1} - y_k\|\alpha$ with $-\|x_{k+1} - y_k\|^2 L/4$. Hence it follows from (99) that

$$h_k - h_{k+1} \geq \delta_k - \beta_k \langle x_k - x_{k-1}, x_{k+1} - x_* \rangle - \frac{\gamma L}{4} \|x_{k+1} - y_k\|^2. \tag{100}$$

Now, (98), written for $k-1$, yields $h_{k-1} - h_k = \delta_{k-1} + \langle x_{k-1} - x_k, x_k - x_* \rangle$ and hence, we have

$$\begin{aligned} h_{k+1} - h_k - \beta_k(h_k - h_{k-1}) &\leq -\delta_k + \beta_k \langle x_k - x_{k-1}, x_{k+1} - x_* \rangle + \frac{\gamma L}{4} \|x_{k+1} - y_k\|^2 \\ &\quad + \beta_k \delta_{k-1} - \beta_k \langle x_k - x_{k-1}, x_k - x_* \rangle \\ &= -\delta_k + \frac{\gamma L}{4} \|x_{k+1} - y_k\|^2 \\ &\quad + \beta_k \delta_{k-1} + \beta_k \langle x_k - x_{k-1}, x_{k+1} - x_k \rangle). \end{aligned}$$

Now, using the definition of y_k, we have

$$\begin{aligned} \frac{1}{2} \|x_{k+1} - y_k\|^2 &= \frac{1}{2} \|x_{k+1} - x_k - \beta_k(x_k - x_{k-1})\|^2 \\ &= \frac{1}{2} \|x_{k+1} - x_k\|^2 + \frac{\beta_k^2}{2} \|x_k - x_{k-1}\|^2 - \beta_k \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle \\ &= \delta_k + \beta_k^2 \delta_{k-1} - \beta_k \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle . \end{aligned}$$

Therefore,

$$h_{k+1} - h_k - \beta_k(h_k - h_{k-1}) \leq -\frac{1}{2} \left(1 - \frac{\gamma L}{2} \right) \|x_{k+1} - y_k\|^2 + (\beta_k + \beta_k^2) \delta_{k-1}. \tag{101}$$

Since $\gamma L < 2$ and $\beta_k + \beta_k^2 \leq 2$ we finally have

$$h_{k+1} - h_k \leq \beta_k(h_k - h_{k-1}) + 2\delta_{k-1}, \tag{102}$$

which yields

$$(h_{k+1} - h_k)_+ \leq \beta_k(h_k - h_{k-1})_+ + 2\delta_{k-1}. \tag{103}$$

Since $t_k \delta_{k-1} \leq (2-c) t_{k-1} \delta_{k-1}$ and $t_{k-1} \delta_{k-1}$ is summable in virtue of Proposition 70(ii), Lemma 72 yields that $((h_{k+1} - h_k)_+)_{k \in \mathbb{N}}$ is summable. Finally, since

$$h_{k+1} \leq h_k + (h_{k+1} - h_k)_+ \tag{104}$$

and h_k is positive, the statement follows from Lemma 71. □

Remark 74 In order to have convergence of the iterates in Algorithm 2, possible choices of the parameters t_k's are (76) with $c > 0$ and $b = 0$ (which looks as a perturbed version of the classical choice given in the first of (78)) and, recalling Remark 18, the second in (78) with $a > 2$.

3.7 Bibliographical Notes

Section 3.1. Fixed-point iterations, also known as the method of successive approximations, was developed by Picard, starting from ideas by Cauchy and Liouville. For the case of Banach spaces, Theorem 19 was first formulated and proved by Banach in his famous dissertation from 1922. Later and independently it was rediscovered by Caccioppoli in 1931. Since then, numerous generalizations or extensions have been obtained which deal with more general classes of operators and iterations. Krasnosel'skiĭ–Mann iteration, as presented in (33), were first studied in [63] with $\lambda = 1/2$. For general $\lambda \in]0, 1[$, these mappings have been studied by Schaefer [104], Browder and Petryshyn [25, 26], and Opial [85]. Mann in [70] considered the more general case of this iteration where λ may vary. Later this case was also studied in [43, 54]. The concept of averaged operator was introduced in [9]. Later, the properties of compositions and convex combinations of averaged nonexpansive operators (Proposition 27) have been applied to the design of new fixed-point algorithms in [38].

Section 3.2. The proximity operator was introduced by Moreau in 1962 [74] and further investigated in [75, 76] as a generalization of the notion of a convex projection operator. Later was considered within the proximal point algorithm in [97]. Since then, it appears in most of the splitting algorithms used in practice [34].

Sections 3.3–3.4. The proximal gradient algorithm finds its roots in the projected gradient method [53, 64] and was originally devised in [72] in the more general context of monotone operators. Weak convergence of the iterates were proved in [51, 72]. An error tolerant version, with variable stepsize is presented in [39], whereas worst-case rate of convergence in values was studied in [12, 24]. The proximal gradient algorithm is also a generalization of the iterative soft thresholding algorithm, first proposed in [41].

Section 3.5. The idea of imposing geometric conditions on the function to be optimized to derive improved convergence rates of first-order methods is old, and was already used in [27, 91, 97]. A systematic study of the class of functions satisfying favorable geometric conditions is more recent and is the result of a series of papers, among which we mention [14, 16, 17]. The fact that convex piecewise polynomial functions are p-Łojasiewicz on sublevel sets is due to [66, Corollary 3.6], in agreement with [27, Corollary 3.6], for the special case of piecewise linear convex functions and with [65, Theorem 2.7] for convex piecewise quadratic functions. The fact that the lasso problem is 2-Łojasiewicz has been observed in [17, Sect. 3.2.1]. Kurdyka–Łojiasiewicz inequality is a powerful tool to analyze convergence of first-order splitting algorithms as shown in a whole line of work [3–5, 17, 18, 50, 69] ranging from the analysis of the proximal point algorithm to a whole class of descent gradient based techniques. These results had an impressive impact on the machine learning community, see e.g., [60]. Theorem 62 is a special case of [52, Theorem 4.1].

Section 3.6. The idea of adding an inertial term in 74 to mitigate zig-zagging was due to Polyak, and gave raise to the heavy ball method [93] (see also [1]), which

is optimal in the sense of Nemirovski and Yudin [81] for the class of convex twice continuously differentiable functions. A simple, but not very intuitive, modification of Polyak's method was due to Nesterov [83], and is the famous accelerated gradient method for convex smooth objective functions [82, 83]. The acceleration technique has been first extended to the proximal point algorithm by Güler [55] and finally extended to the composite optimization problems in [12]. Various modifications of these accelerated algorithms are nowadays the methods of choice to optimize objective functions in a large scale scenario, even in a nonconvex setting: despite convergence issues, the ADAM algorithm is probably the most used in the deep learning context [62]. The first papers studying accelerated algorithms were focused on convergence of the objective function values. Convergence of the iterates has been established much more recently, starting from the paper by Chambolle and Dossal [29] and further devoloped later. Only many years later its introduction, Nesterov accelerated method has been shown to be a specific discretization of the heavy ball system introduced by Polyak with a vanishing inertial coefficient [111], and this key observation started a very active research activity on the subject (see [7], [6] and references therein).

4 Stochastic Minimization Algorithms

In this section, we analyze stochastic versions of the algorithms previously presented. We will consider problems of type

$$\underset{x\in X}{\text{minimize}} \quad f(x) + g(x), \tag{105}$$

where $f\colon X \to \mathbb{R}$ is a convex function and $g\colon X \to \,]-\infty, +\infty]$ is a proper convex and lower semicontinuous function, and depending on the hypotheses only a stochastic subgradient/gradient of f will be available. One of the main examples for such situation is when f is given in the form of an expectation, that is,

$$f(x) = \mathsf{E}[\varphi(x, \zeta)], \tag{106}$$

which corresponds to the setting of *stochastic optimization*. In this case, a stochastic subgradient/gradient of f is obtained through a subgradient/gradient of $\varphi(x, \zeta)$. Finally, in general we will assume that the proximity operator of g is given explicitly. However, in the last section we will consider a situation in which the proximity operator of g is actually given through a stochastic oracle.

We start by recalling few facts on conditional expectation.

Fact 75 The following hold.

(i) Let ζ be a random variable with value in the measurable space $\mathcal{Z}$. Then the operator $\mathsf{E}[\cdot \mid \zeta]\colon L^1 \to L^1$ is linear and monotone increasing.

(ii) Let ξ be a real-valued summable random variable and ζ be a random variable with value in a measurable space $\mathcal{Z}$. Then, $\mathsf{E}[\mathsf{E}[\xi, \mid \zeta]] = \mathsf{E}[\xi]$.

(iii) Let ζ be a random variable with value in the measurable space $\mathcal{Z}$ and let $\varphi\colon \mathcal{Z} \to \mathbb{R}$ be a measurable real function such that $\mathsf{E}[|\varphi(\zeta)|] < +\infty$. Then $\mathsf{E}[\varphi(\zeta) \mid \zeta] = \varphi(\zeta)$.

(iv) Let X be a separable Hilbert space and let ζ_1 and ζ_2 be two X-valued random vectors such that $\mathsf{E}[|\langle \zeta_1, \zeta_2\rangle|] < +\infty$ and $\mathsf{E}[\|\zeta_2\|] < +\infty$. Then $\mathsf{E}[\langle \zeta_1, \zeta_2\rangle \mid \zeta_1] = \langle \zeta_1, \mathsf{E}[\zeta_2 \mid \zeta_1]\rangle$.

(v) Let ζ_1 and ζ_2 be two independent random variables with values in the measurable spaces $\mathcal{Z}_1$ and $\mathcal{Z}_2$ respectively. Let $\varphi\colon \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R}$ be measurable and such that $\mathsf{E}[|\varphi(\zeta_1, \zeta_2)|] < +\infty$. Then $\mathsf{E}[\varphi(\zeta_1, \zeta_2) \mid \zeta_1] = \psi(\zeta_1)$, where, for every $z_1 \in \mathcal{Z}_1$, $\psi(z_1) = \mathsf{E}[\varphi(z_1, \zeta_2)]$.

4.1 The Stochastic Subgradient Method

Here we take g in (105) as an indicator function of a closed convex set. Thus, we assume that $C \subset X$ is a nonempty closed and convex set and $f\colon X \to \mathbb{R}$ is a convex function and we want to solve the following problem

$$\underset{x \in C}{\text{minimize}} \quad f(x), \tag{107}$$

where the projection onto C can be computed explicitly but, only a stochastic subgradient of f is available. The algorithm is detailed below.

Algorithm 3 *(The stochastic subgradient projection method) Let $x_0 \in X$ and $(\gamma_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}_{++}$. Then,*

$$\begin{aligned} &\textit{for } k = 0, 1, \dots \\ &\quad \left\lfloor \begin{array}{l} \hat{u}_k \textit{ is a summable } X\textit{-valued random vector s.t. } \mathsf{E}[\hat{u}_k \mid x_k] \in \partial f(x_k), \\ x_{k+1} = P_C(x_k - \gamma_k \hat{u}_k). \end{array} \right. \end{aligned} \tag{108}$$

Moreover, define, for every $k \in \mathbb{N}$,

$$f_k = \min_{0 \le i \le k} \mathsf{E}[f(x_i)], \qquad \bar{x}_k = \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i x_i.$$

Remark 76 In addition to the sequence x_k, Algorithm 3 requires keeping track of the sequences $\Gamma_k := \sum_{i=0}^{k} \gamma_i$ and $\bar{x}_k$, which can updated recursively, as $\Gamma_{k+1} = \Gamma_k + \gamma_k$ and $\bar{x}_{k+1} = \Gamma_{k+1}^{-1}(\Gamma_k \bar{x}_k + \gamma_{k+1} x_{k+1})$.

The following theorem gives the main convergence results about the algorithm.

Theorem 77 *Let $C \subset X$ be a nonempty closed convex set and let $f : X \to \mathbb{R}$ be convex. Let $(x_k)_{k\in\mathbb{N}}$, $(f_k)_{k\in\mathbb{N}}$, and $(\bar{x}_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 3. We make the following additional assumption*

A1 *There exists $B \geq 0$, such that, for every $k \in \mathbb{N}$, $\mathsf{E}[\|\hat{u}_k\|^2] \leq B^2$.*

Then, for every $k \in \mathbb{N}$, x_k is square summable in norm and $f(x_k)$ is summable and the following statements hold.

(i) *Suppose that $\gamma_k \to 0$ and $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$. Then $\liminf_k \mathsf{E}[f(x_k)] = \lim_k f_k = \inf_C f$.*

(ii) *Let $x \in C$ and let $m, k \in \mathbb{N}$ with $m \leq k$. Then*

$$\sum_{j=m}^{k} \frac{\gamma_j}{\sum_{i=m}^{k} \gamma_i} \mathsf{E}[f(x_j)] - f(x) \leq \frac{\mathsf{E}[\|x_m - x\|^2]}{2} \frac{1}{\sum_{i=m}^{k} \gamma_i} + \frac{B^2}{2} \frac{\sum_{j=m}^{k} \gamma_i^2}{\sum_{i=m}^{k} \gamma_i}. \tag{109}$$

(iii) *Suppose that $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$ and $\sum_{i=0}^{k} \gamma_i^2 / \sum_{i=0}^{k} \gamma_i \to 0$. Then $f_k \to \inf_C f$ and $\mathsf{E}[f(\bar{x}_k)] \to \inf_C f$.*

Moreover, if $\operatorname{argmin}_C f \neq \varnothing$, the right hand side of (109), with $m = 0$ and $x \in \operatorname{argmin}_C f$, yields a rate of convergence for both $f_k - \min_C f$ and $\mathsf{E}[f(\bar{x}_k)] - \min_C f$.

Proof Let $k \in \mathbb{N}$ and $x \in C$ and set $u_k = \mathsf{E}[\hat{u}_k \mid x_k]$. First of all, note that assumption A1 actually implies that $\|\hat{u}_k\|$ is square summable and hence summable. Then we prove the following inequality

$$2\gamma_k \langle x_k - x, \hat{u}_k \rangle \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \gamma_k^2 \|\hat{u}_k\|^2. \tag{110}$$

Indeed setting $y_k = x_k - \gamma_k \hat{u}_k$ and using the relation $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, we have

$$\begin{aligned} 2\gamma_k \langle x_k - x, \hat{u}_k \rangle &= 2\langle x_k - x, x_k - y_k \rangle \\ &= \|x_k - x\|^2 + \|x_k - y_k\|^2 - \|y_k - x\|^2. \end{aligned} \tag{111}$$

Now, since P_C is nonexpansive, we have $\|x_{k+1} - x\| = \|P_C(y_k) - P_C(x)\| \leq \|y_k - x\|$ and hence (110) follows.

We prove by induction that $\|x_k - x\|$ is square summable for every $k \in \mathbb{N}$. The statement is true for $k = 0$. Suppose that $\|x_k - x\|$ is square summable for some $k \in \mathbb{N}$. Then it follows from (110) that

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + 2\gamma_k \|x_k - x\| \|\hat{u}_k\| + \gamma_k^2 \|\hat{u}_k\|^2.$$

The right-hand side is summable, and hence $\|x_{k+1} - x\|$ is square summable. So, all the terms in (110) are summable. Therefore, taking the conditional expectation given x_k of both terms of inequality (110) and using the fact that $u_k = \mathsf{E}[\hat{u}_k \mid x_k] \in \partial f(x_k)$ and the properties in Fact 75, we have almost surely

$$2\gamma_k(f(x_k) - f(x)) \le 2\gamma_k\langle x_k - x, \mathsf{E}[\hat{u}_k \mid x_k]\rangle \\ \le \|x_k - x\|^2 - \mathsf{E}[\|x_{k+1} - x\|^2 \mid x_k] + \gamma_k^2 \mathsf{E}[\|\hat{u}_k\|^2 \mid x_k]. \quad (112)$$

Now, being f subdifferentiable, there exists $(a, \beta) \in H \times \mathbb{R}$, $a \neq 0$, such that $\langle \cdot, a\rangle + \beta \le f$, hence $\langle x_k, a\rangle + \beta \le f(x_k)$. Therefore, we have $(f(x_k))_- \le \|x_k\|\,\|a\| + |\beta|$, which together with (112) yields the summability of $f(x_k)$. Taking the expectation in (112) and recalling that $\mathsf{E}[\|\hat{u}_k\|^2] \le B^2$, we get

$$2\gamma_k(\mathsf{E}[f(x_k)] - f(x)) \le \mathsf{E}[\|x_k - x\|^2] - \mathsf{E}[\|x_{k+1} - x\|^2] + \gamma_k^2 B^2. \quad (113)$$

(i): Since $(f_k)_{k\in\mathbb{N}}$ is decreasing, we have $\inf_C f \le \lim_k f_k = \inf_k f_k = \inf_k \mathsf{E}[f(x_k)] \le \liminf_k \mathsf{E}[f(x_k)]$. Therefore it is sufficient to prove that $\liminf_k \mathsf{E}[f(x_k)] \le \inf_C f$. Suppose that $x \in C$ is such that $f(x) < \liminf_k \mathsf{E}[f(x_k)] = \sup_n \inf_{k\ge n} \mathsf{E}[f(x_k)]$. Then there exists $n \in \mathbb{N}$ such that $f(x) < \inf_{k\ge n} \mathsf{E}[f(x_k)]$. Set $\rho = \inf_{k\ge n} \mathsf{E}[f(x_k)] - f(x) > 0$. Then, (113) yields

$$(\forall\, k \ge n) \quad \gamma_k \rho \le \mathsf{E}[\|x_k - x\|^2] - \mathsf{E}[\|x_{k+1} - x\|^2] - \gamma_k\big(\rho - \gamma_k B^2\big).$$

Now, since $\gamma_k \to 0$, there exists $m \in \mathbb{N}$ such that for every integer $k \ge m$, we have $\rho - \gamma_k B^2 \ge 0$ and hence, setting $\nu := \max\{n, m\}$, we have

$$\rho \sum_{k\ge\nu} \gamma_k \le \mathsf{E}[\|x_\nu - x\|^2] < +\infty.$$

This contradicts the assumption $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$. Therefore, we showed that there is no $x \in C$ such that $f(x) < \liminf_k \mathsf{E}[f(x_k)]$, that is, $\liminf_k \mathsf{E}[f(x_k)] \le \inf_C f$.

(ii): It follows from (113) that

$$(\forall\, i \in \mathbb{N}) \quad \gamma_i(\mathsf{E}[f(x_i)] - f(x)) \le \frac{1}{2}\big(\mathsf{E}[\|x_i - x\|^2] - \mathsf{E}[\|x_{i+1} - x\|^2\big) + \frac{B^2}{2}\gamma_i^2. \quad (114)$$

So, summing from m to k, we have

$$\sum_{i=m}^{k} \gamma_i(\mathsf{E}[f(x_i)] - f(x)) \le \frac{1}{2}\mathsf{E}[\|x_m - x\|^2] + \frac{B^2}{2}\sum_{i=m}^{k} \gamma_i^2.$$

Dividing the above inequality by $\sum_{i=m}^{k} \gamma_i$ yields (109).

(iii): We first note that, since f is convex and $\bar{x}_k$ is a convex combination of the x_i's, with coefficients $\eta_i = \gamma_i / \sum_{j=0}^{k} \gamma_j$, with $0 \le i \le k$, we have $\mathsf{E}[f(\bar{x}_k)] \le \sum_{i=0}^{k} \eta_i \mathsf{E}[f(x_i)]$. Moreover, $f_k = \sum_{i=0}^{k} \eta_i f_k \le \sum_{i=0}^{k} \eta_i \mathsf{E}[f(x_i)]$. Therefore,

$$(\forall\, k \in \mathbb{N}) \quad h_k := \max\{f_k, \mathsf{E}[f(\bar{x}_k)]\} \le \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i \mathsf{E}[f(x_i)]. \quad (115)$$

Let $x \in C$. Then it follows from (109) and (115) that $\limsup_k h_k \leq f(x)$. Since x is arbitrary in C, we have $\limsup_k h_k \leq \inf_C f$. Moreover, clearly we have $\inf_C f \leq \liminf_k h_k$. Therefore, $h_k \to \inf_C f$. Since $\inf_C f \leq f_k \leq h_k$ and $\inf_C f \leq \mathsf{E}[f(\bar{x}_k)] \leq h_k$, the statement follows. □

Lemma 78 *Let $m, k \in \mathbb{N}$ with $2 \leq m < k$. Then, the following inequalities hold.*

(i) $\log\left(\frac{k}{m}\right) + \frac{1}{2}\left(\frac{1}{m} + \frac{1}{k}\right) \leq \sum_{i=m}^{k} \frac{1}{i} \leq \log\left(\frac{k}{m-1}\right)$

(ii) $2(\sqrt{k} - \sqrt{m}) + \frac{1}{2}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{k}}\right) \leq \sum_{i=m}^{k} \frac{1}{\sqrt{i}}.$

(iii) $\sum_{i=0}^{+\infty} \frac{1}{i^2} = \frac{\pi}{6}.$

Lemma 79 *Let $a \in \mathbb{R}^n_{++}$ and $\alpha, \beta \in \mathbb{R}_{++}$. Then*

$$\min_{\gamma \in \mathbb{R}^n_{++}} \frac{\alpha}{2a^\top \gamma} + \frac{\beta}{2} \frac{\|\gamma\|^2}{a^\top \gamma} = \sqrt{\frac{\alpha\beta}{\|a\|^2}}$$

and the minimum is achieved at $\gamma = \left(\sqrt{\alpha/\beta\|a\|^2}\right)a$.

Proof Define $\varphi\colon \mathbb{R} \times \mathbb{R}^n \to \,]-\infty, +\infty\,]$ such that

$$\varphi(t, \gamma) = \begin{cases} \dfrac{\alpha + \beta\|\gamma\|^2}{2t} & \text{if } t > 0 \text{ and } \gamma \in \mathbb{R}^n_+ \\ +\infty & \text{otherwise.} \end{cases}$$

Clearly φ is closed, convex, and differentiable in $\mathbb{R}_{++} \times \mathbb{R}^n_{++}$, and, for all $(t, \gamma) \in \mathbb{R}_{++} \times \mathbb{R}^n_{++}$,

$$\nabla\varphi(t, \gamma) = \left(-\frac{\alpha + \beta\|\gamma\|^2}{2t^2}, \frac{\beta}{t}\gamma \right). \tag{116}$$

Then,

$$\inf_{\gamma \in \mathbb{R}^n_{++}} \frac{\alpha}{2a^\top \gamma} + \frac{\beta}{2} \frac{\|\gamma\|^2}{a^\top \gamma} = \inf_{t>0} \inf_{\substack{\gamma \in \mathbb{R}^n_{++} \\ a^\top \gamma = t}} \frac{\alpha + \beta\|\gamma\|^2}{2t} = \inf_{\substack{(t,\gamma) \in \mathbb{R} \times \mathbb{R}^n \\ a^\top \gamma = t}} \varphi(t, \gamma),$$

and the right hand side can be written as

$$\inf_{(t,\gamma) \in \mathbb{R} \times \mathbb{R}^n} \varphi(t, \gamma) + \iota_{\{0\}}\big((-1, a)^\top (t, \gamma)\big).$$

So, Fermat's rule yields

$$0 \in \nabla\varphi(t,\gamma) + A^*\partial\iota_{\{0\}}(A(t,\gamma)),$$

where $A\colon \mathbb{R}^{n+1} \to \mathbb{R}$ is the linear form $A = (-1,a)^\top\cdot$ and A^* is the map $s \mapsto s(-1,a)$. Therefore, we have

$$(-1,a)^\top(t,\gamma) = 0 \quad \text{and} \quad -\nabla\varphi(t,\gamma) \in \mathbb{R}(-1,a),$$

which, in view of (116), implies that there exists $s \in \mathbb{R}$ such that

$$\begin{cases} \dfrac{\alpha+\beta\|\gamma\|^2}{2t^2} = -s \\ -\dfrac{\beta}{t}\gamma = sa \\ a^\top\gamma = t \end{cases}$$

Now, it follows from the last two equations above that $-\beta = -\beta a^\top\gamma/t = s\|a\|^2$ and hence

$$\begin{cases} \dfrac{\alpha+\beta\|\gamma\|^2}{2t^2} = \dfrac{\beta}{\|a\|^2} \\ \gamma = \dfrac{t}{\|a\|^2}a \\ a^\top\gamma = t. \end{cases}$$

It follows from the second equation above that $\|\gamma\|^2 = t^2/\|a\|^2$ which, substituted into the first equation, gives $t = \sqrt{\alpha/\beta}\|a\|$. Therefore, finally, we have

$$\gamma = \left(\sqrt{\frac{\alpha}{\beta\|a\|^2}}\right)a \quad \text{and} \quad \varphi(t,\gamma) = \sqrt{\frac{\alpha\beta}{\|a\|^2}}. \qquad \square$$

Corollary 80 *Under the same assumptions of Theorem 77, the following hold.*

(i) *Suppose that* $\operatorname{argmin}_C f \neq \varnothing$ *and let* $D \geq \operatorname{dist}(x_0, \operatorname{argmin}_C f)$ *and* $k \in \mathbb{N}$. *Then,*

$$\max\{f_k, \mathsf{E}[f(\bar{x}_k)]\} - \min_C f \leq \frac{D^2}{2}\frac{1}{\sum_{i=0}^k \gamma_i} + \frac{B^2}{2}\frac{\sum_{j=0}^k \gamma_i^2}{\sum_{i=0}^k \gamma_i}. \tag{117}$$

Moreover, the right hand side of (117) *is minimized when, for every* $i = 0,\dots,k$, $\gamma_i = D/(B\sqrt{k+1})$ *and in that case we have*

$$\max\left\{f_k, \mathsf{E}[f(\bar{x}_k)]\right\} - \min_C f \leq \frac{BD}{\sqrt{k+1}}.$$

(ii) *Let, for every* $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/(k+1)$. *Then,* $f_k \to \inf_C f$ *and* $\mathsf{E}[f(\bar{x}_k)] \to \inf_C f$. *Moreover, if* $\operatorname{argmin}_C f \neq \varnothing$, *we have, for every* $k \in \mathbb{N}$,

$$\max\left\{f_k, \mathsf{E}[f(\bar{x}_k)]\right\} - \min_C f \leq \left(\frac{\operatorname{dist}(x_0, \operatorname{argmin}_C f)^2}{2\bar{\gamma}} + \frac{\pi \bar{\gamma} B^2}{12}\right)\frac{1}{\log(k+1)}. \tag{118}$$

(iii) *Let, for every* $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$. *Then,* $f_k \to \inf_C f$ *and* $\mathsf{E}[f(\bar{x}_k)] \to \inf_C f$. *Moreover, if* $\operatorname{argmin}_C f \neq \varnothing$, *for every integer* $k \geq 2$, *we have*

$$\max\left\{f_k, \mathsf{E}[f(\bar{x}_k)]\right\} - \min_C f \leq \frac{\operatorname{dist}(x_0, \operatorname{argmin}_C f)^2}{2\bar{\gamma}}\frac{1}{\sqrt{k+1}} + \bar{\gamma} B^2 \frac{\log(k+1)}{\sqrt{k+1}}. \tag{119}$$

(iv) *Let, for every* $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$ *and suppose that* C *is bounded with diameter* $\bar{D} > 0$ *and that* $\operatorname{argmin}_C f \neq \varnothing$. *Set, for every* $k \in \mathbb{N}$, $\tilde{f}_k = \min_{\lfloor k/2 \rfloor \leq i \leq k} f(x_i)$ *and* $\tilde{x}_k = \left(\sum_{i=\lfloor k/2 \rfloor}^k \gamma_i\right)^{-1} \sum_{i=\lfloor k/2 \rfloor}^k \gamma_i x_i$. *Then, for every integer* $k \geq 2$,

$$\max\left\{\tilde{f}_k, \mathsf{E}[f(\tilde{x}_k)]\right\} - \min_C f \leq \left(\frac{3\bar{D}^2}{2\bar{\gamma}} + \frac{5\bar{\gamma}B^2}{2}\right)\frac{1}{\sqrt{k+1}}. \tag{120}$$

Proof (i): Equation (117) follows from (115) and by minimizing the right hand side of (109), with $m = 0$, w.r.t. $x \in \operatorname{argmin}_C f$. Now, it follows from Lemma 79 that the minimum of the right-hand side of (117) is $BD/\sqrt{k+1}$ and is achieved at $(\gamma_i)_{0 \leq i \leq k} \equiv D/(B\sqrt{k+1})$. Note that is this case $\bar{x}_k = (k+1)^{-1}\sum_{i=0}^k x_i$.

(ii): We derive from Lemma 78(i), with $m = 1$, that $\sum_{i=0}^k \gamma_i = \bar{\gamma}\sum_{i=1}^{k+1}(1/i) \geq \bar{\gamma}\log(k+1)$. Moreover, we have $\sum_{i=0}^k \gamma_i^2 = \bar{\gamma}^2 \sum_{i=1}^{k+1} 1/i^2 \leq \bar{\gamma}^2\pi/6$. So, the first part follows from Theorem 77 (iii), while the inequality in (118) follows from (117) with $D = \operatorname{dist}(x_0, \operatorname{argmin}_C f)$.

(iii) Lemma 78(ii), with $m = 1$, yields $\sum_{i=1}^k 1/\sqrt{i} \geq 2(\sqrt{k} - 1) + (1/2)(1 + 1/\sqrt{k}) \geq 2\sqrt{k} - 3/2$. Moreover, $2\sqrt{k} - 3/2 \geq \sqrt{k}$ for $k \geq 3$ and clearly for $k \leq 2$, $\sum_{i=1}^k 1/\sqrt{i} \geq \sqrt{k}$. Therefore, for every $k \in \mathbb{N}$, $\sum_{i=0}^k \gamma_i = \bar{\gamma}\sum_{i=1}^{k+1} 1/\sqrt{i} \geq \bar{\gamma}\sqrt{k+1}$. Moreover, by Lemma 78(i), we have $\sum_{i=1}^k 1/i = 1 + \sum_{i=2}^k 1/i \leq 1 + \log k \leq 2\log k$, for $k \geq 3$. Therefore, for every $k \in \mathbb{N}$, $k \geq 2$, we have $\sum_{i=0}^k \gamma_i^2 = \bar{\gamma}^2 \sum_{i=1}^{k+1} 1/i \leq 2\bar{\gamma}^2 \log(k+1)$. Again, the first part follows from Theorem 77(iii), while (119) follows from (117) with $D = \operatorname{dist}(x_0, \operatorname{argmin}_C f)$.

(iv): Let $k \in \mathbb{N}$, $k \geq 2$. It follows from Lemma 78(i) that

$$\sum_{i=\lfloor k/2 \rfloor}^k \gamma_i^2 = \bar{\gamma}^2 \sum_{i=\lfloor k/2 \rfloor + 1}^{k+1} \frac{1}{i} \leq \bar{\gamma}^2 \log\left(\frac{k+1}{\lfloor k/2 \rfloor}\right) \leq \bar{\gamma}^2 \log 4 \leq \bar{\gamma}^2 \frac{5}{3}.$$

Moreover, Lemma 78(ii) yields

$$\sum_{i=\lfloor k/2\rfloor}^{k} \gamma_i = \bar{\gamma} \sum_{i=\lfloor k/2\rfloor+1}^{k+1} \frac{1}{\sqrt{i}}$$
$$\geq 2\bar{\gamma}(\sqrt{k+1} - \sqrt{\lfloor k/2\rfloor + 1}) \geq 2\bar{\gamma}\sqrt{k+1}\Big(1 - \sqrt{\frac{\lfloor k/2\rfloor + 1}{k+1}}\Big).$$

Now, since $(\lfloor k/2\rfloor + 1)/(k+1) \leq 2/3$, we have

$$\sum_{i=\lfloor k/2\rfloor}^{k} \gamma_i \geq 2\bar{\gamma}\left(1 - \sqrt{\frac{2}{3}}\right)\sqrt{k+1} \geq \frac{\bar{\gamma}}{3}\sqrt{k+1}$$

The statement follows from Theorem 77(ii), with $m = \lfloor k/2\rfloor$ and $x \in \operatorname{argmin}_C f$, taking into account that, as in (115), $\max\{\tilde{f}_k, f(\tilde{x}_k)\} \leq \big(\sum_{i=\lfloor k/2\rfloor}^{k} \gamma_i\big)^{-1} \sum_{i=\lfloor k/2\rfloor}^{k} \gamma_i f(x_i)$. □

Example 81 A case in which the above stochastic algorithm arises is in the *incremental subgradient method*. We aim at solving

$$\min_{x\in C} f(x) := \frac{1}{m}\sum_{j=1}^{m} f_j(x),$$

where every $f_j\colon X \to \mathbb{R}$ is convex and Lipschitz continuous with constant L_j. The projected incremental subgradient method is as follows. Let, for every j, $\tilde{\nabla} f_j\colon X \to X$ be a selection of ∂f_j. Let $x_0 \in X$. Then,

$$\left.\begin{array}{l}\text{for } k = 0, 1, \dots\\ \quad \text{chose an index } j_k \in \{1, \dots, m\} \text{ at random}\\ \quad x_{k+1} = P_C(x_k - \gamma_k \underbrace{\tilde{\nabla} f_{j_k}(x_k)}_{\hat{u}_k}).\end{array}\right. \tag{121}$$

Since $\partial f = (1/m)\sum_{j=1}^{m} \partial f_j$, we have that $(1/m)\sum_{j=1}^{m} \tilde{\nabla} f_j(x) \in \partial f(x)$. Let $k \in \mathbb{N}$. Then, x_k is a random variable, depending on $j_0, \dots, j_{k-1}$. Hence, $\hat{u}_k := \tilde{\nabla} f_{j_k}(x_k)$ is a random variable, where x_k and j_k are independent random variables, and Fact 75 yields

$$u_k := \mathsf{E}\big[\tilde{\nabla} f_{j_k}(x_k) \,|\, x_k\big] = \frac{1}{m}\sum_{j=1}^{m} \tilde{\nabla} f_j(x_k) \in \partial f(x_k)$$

and

$$\mathsf{E}[\|\tilde{\nabla} f_{j_k}(x_k)\|^2 \,|\, x_k] = \frac{1}{m}\sum_{j=1}^{m} \|\tilde{\nabla} f_j(x_k)\|^2 \leq \frac{1}{m}\sum_{j=1}^{m} L_j^2,$$

and hence $\mathsf{E}[\|\tilde{\nabla} f_{j_k}(x_k)\|^2] \leq (1/m)\sum_{j=1}^m L_j^2$. In the end assumptions of Theorem 77 are satisfied with $B^2 = (1/m)\sum_{j=1}^m L_j^2$.

Example 82 *(Stochastic optimization)* We generalize the previous example. We consider the following optimization problem

$$\underset{x\in C}{\text{minimize}} \quad f(x), \qquad f(x) = \mathsf{E}[\varphi(x,\zeta)], \tag{122}$$

where $f\colon X \to \mathbb{R}$, ζ is a random variable with values in a measurable space $\mathcal{Z}$ with distribution μ and $\varphi\colon X \times \mathcal{Z} \to \mathbb{R}$ is such that

- (SO_1) $\forall\, z \in \mathcal{Z}$, $\varphi(\cdot, z)$ is convex and $L(z)$-Lipschitz continuous and $\int_{\mathcal{Z}} L(z)^2 d\mu < +\infty$.
- (SO_2) $\varphi(0,\cdot) \in L^1(\mathcal{Z},\mu)$.

The above assumptions ensure that, for every $x \in X$, $\varphi(x,\cdot) \in L^1(\mathcal{Z},\mu)$. Indeed, for every $z \in \mathcal{Z}$, $|\varphi(x,z)| \leq |\varphi(x,z) - \varphi(0,z)| + |\varphi(0,z)| \leq L(z)\|x\| + |\varphi(0,z)|$. Hence $\varphi(x,z) \in L^1(\mathcal{Z},\mu)$, since $L(z)$ and $\varphi(0,z)$ are so. We let $\partial\varphi\colon X \times \mathcal{Z} \to 2^X$ be such that $\partial\varphi(x,z) = \partial\varphi(\cdot,z)(x)$ and we make the following additional assumptions

- (SO_3) there exists a measurable $\tilde{\nabla}\varphi\colon X \times \mathcal{Z} \to X$, such that, for every $x \in X$ and for μ-a.e. $z \in \mathcal{Z}$, $\tilde{\nabla}\varphi(x,z) \in \partial\varphi(x,z)$.
- (SO_4) $(\zeta_k)_{k\in\mathbb{N}}$ is a sequence of independent copies of ζ.

Then we consider the following algorithm. Let $x_0 \in X$. Then,

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\quad x_{k+1} = P_C\big(x_k - \gamma_k \underbrace{\tilde{\nabla}\varphi(x_k,\zeta_k)}_{\hat{u}_k}\big). \end{aligned} \tag{123}$$

We have, for every $x_1, x_2 \in X$,

$$|f(x_1) - f(x_2)| \leq \int_{\mathcal{Z}} |\varphi(x_1,z) - \varphi(x_2,z)| d\mu(z) \leq \|x_1 - x_2\| \int_{\mathcal{Z}} L(z) d\mu(z).$$

Therefore, f is Lipschitz continuous with constant $\int_{\mathcal{Z}} L(z)d\mu(z) \leq \big(\int_{\mathcal{Z}} L(z)^2 d\mu(z)\big)^{1/2}$. Moreover, assumption ($\mathrm{SO}_3$) implies that

$$\text{for all } x, y \in X \text{ and for } \mu\text{-a.e. } z \in \mathcal{Z} \quad \varphi(y,z) \geq \varphi(x,z) + \langle y - x, \tilde{\nabla}\varphi(x,z)\rangle. \tag{124}$$

Note that all terms of the above inequality are μ-summable, in particular, since $\|\tilde{\nabla}\varphi(x,z)\| \leq L(z)$ and $L(z)$ is μ-summable, $\tilde{\nabla}\varphi(x,\cdot)$ is μ-summable. Hence, integrating (124) w.r.t. μ we get

$$(\forall\, x, y \in X) \qquad f(y) \geq f(x) + \langle y - x, \int_{\mathcal{Z}} \tilde{\nabla}\varphi(x,z) d\mu(z)\rangle.$$

Therefore, for every $x \in X$, $\mathsf{E}[\tilde{\nabla}\varphi(x,\zeta)] \in \partial f(x)$. Now, let $k \in \mathbb{N}$, $k \geq 1$. Then, it follows from (123) that

$$x_k = x_k(\zeta_0, \dots, \zeta_{k-1}),$$

hence x_k and ζ_k are independent random variables. Therefore, Fact 75(v) yields that $u_k := \mathsf{E}[\tilde{\nabla}\varphi(x_k,\zeta_k) \,|\, x_k] = \int_{\mathcal{Z}} \tilde{\nabla}\varphi(x_k,z)d\mu(z) \in \partial f(x_k)$ and

$$\mathsf{E}\big[\|\tilde{\nabla}\varphi(x_k,\zeta_k)\|^2 \,|\, x_k\big] = \int_{\mathcal{Z}} \|\tilde{\nabla}\varphi(x_k,z)\|^2 d\mu(z) \leq \int_{\mathcal{Z}} L(z)^2 d\mu(z) < +\infty,$$

and hence $\mathsf{E}[\|\tilde{\nabla}\varphi(x_k,\zeta_k)\|^2] \leq \int_{\mathcal{Z}} L(z)^2 d\mu(z)$. In the end Theorem 77 applies with $B^2 = \int_{\mathcal{Z}} L(z)^2 d\mu(z)$, so that the stochastic algorithm (123) provides a solution to problem (122).

4.2 Stochastic Proximal Gradient Method

We address again problem (105) where now f is Lipschitz smooth, and we consider a stochastic version of Algorithm 1. In the following we set $F = f + g$.

Algorithm 4 *(The stochastic proximal gradient method) Let $x_0 \in X$ and $(\gamma_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}_{++}$. Then,*

$$\left\lfloor \begin{array}{l} \text{for } k = 0, 1, \dots \\ \quad \hat{u}_k \text{ is a square summable } X\text{-valued random vector s.t. } \mathsf{E}[\hat{u}_k \,|\, x_k] = \nabla f(x_k), \\ \quad x_{k+1} = \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \hat{u}_k). \end{array} \right. \tag{125}$$

Moreover, define, for every $k \in \mathbb{N}$,

$$F_k = \min_{0\leq i\leq k} \mathsf{E}[F(x_{i+1})], \qquad \bar{x}_k = \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i x_{i+1}.$$

The following theorem gives the main convergence results about the algorithm.

Theorem 83 *Let $f : X \to \mathbb{R}$ be convex and differentiable with a L-Lipschitz continuous gradient, let $g \in \Gamma_0(X)$, and define $F = f + g$. Let $(x_k)_{k\in\mathbb{N}}$, $(F_k)_{k\in\mathbb{N}}$, and $(\bar{x}_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 4. We make the following additional assumption*

A1 *There exists $\sigma \geq 0$, such that, for every $k \in \mathbb{N}$, the random variable $\|\hat{u}_k - \nabla f(x_k)\|$ is square summable and $\mathsf{E}[\|\hat{u}_k - \nabla f(x_k)\|^2 | x_k] \leq \sigma^2$.*

A2 *For every $k \in \mathbb{N}$, $\gamma_k \leq 1/L$.*

Then, for every $k \in \mathbb{N}$, x_k is square summable in norm and $F(x_k)$ is summable and the conclusions (i), (ii), and (iii) of Theorem 77 and those of Corollary 80(i)(iii)(iv)

remain valid in expectation, with the constant B^2 replaced by σ^2 and f_k, $\mathsf{E}[f(x_k)]$, and $\inf_C f$ *replaced by F_k, $\mathsf{E}[F(x_k)]$, and* $\inf F$ *respectively. In particular, the following hold.*

(i) *Suppose that $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$ and that $\sum_{i=0}^k \gamma_i^2 / \sum_{i=0}^k \gamma_i \to 0$. Then $F_k \to \inf F$, $\liminf_k \mathsf{E}[F(x_k)] = \inf F$ and $\mathsf{E}[F(\bar{x}_k)] \to \inf F$.*

(ii) *Suppose that $S_* := \operatorname{argmin} F \neq \varnothing$ and let, for every $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$, with $\bar{\gamma} \leq 1/L$. Then, for every integer $k \geq 2$,*

$$\max\{F_{k+1}, \mathsf{E}[F(\bar{x}_{k+1})]\} - \min F \leq \frac{\operatorname{dist}(x_0, S_*)^2}{2\bar{\gamma}} \frac{1}{\sqrt{k+1}} + \bar{\gamma}\sigma^2 \frac{\log(k+1)}{\sqrt{k+1}}.$$

Proof Since $\gamma_k \leq 1/L$ for every $k \in \mathbb{N}$, it follows from Lemma 45, that, for every $(x, y) \in X^2$, $z \in \operatorname{dom} \partial g$, and every $\eta \in \partial g(z)$ we have

$$F(x) \geq F(z) + \langle x - z, \nabla f(y) + \eta \rangle - \frac{1}{2\gamma_k} \|z - y\|^2 . \tag{126}$$

Let $x \in X$. Applying the previous inequality with $z = x_{k+1}$, $\eta = \gamma_k^{-1}(x_k - x_{k+1}) - \hat{u}_k$, and $y = x_k$ we obtain

$$F(x) \geq F(x_{k+1}) + \langle x - x_{k+1}, \nabla f(x_k) - \hat{u}_k + \frac{x_k - x_{k+1}}{\gamma_k} \rangle - \frac{1}{2\gamma_k} \|x_{k+1} - x_k\|^2 \tag{127}$$

and thus, setting $(\forall k \in \mathbb{N})\ \tilde{x}_{k+1} = \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$,

$$\begin{aligned} F(x_{k+1}) - F(x) &\leq \langle x - x_{k+1}, \hat{u}_k - \nabla f(x_k) - \frac{x_k - x_{k+1}}{\gamma_k} \rangle + \frac{1}{2\gamma_k} \|x_{k+1} - x_k\|^2 \\ &= \langle x - x_{k+1}, \hat{u}_k - \nabla f(x_k) \rangle + \frac{1}{2\gamma_k} \Big(-2\langle x - x_{k+1}, x_k - x_{k+1} \rangle + \|x_{k+1} - x_k\|^2 \Big) \\ &= \langle x - x_{k+1}, \hat{u}_k - \nabla f(x_k) \rangle + \frac{1}{2\gamma_k} \big(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 \big) \\ &= \langle x - \tilde{x}_{k+1}, \hat{u}_k - \nabla f(x_k) \rangle + \big\langle \tilde{x}_{k+1} - x_{k+1}, \hat{u}_k - \nabla f(x_k) \big\rangle \\ &\qquad + \frac{1}{2\gamma_k} \big(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 \big). \end{aligned} \tag{128}$$

We next want to take the conditional expectation of this inequality. To this aim we first prove by induction that $\|x_k\|$ and $\|\nabla f(x_k)\|$ are square summable and $F(x_k)$ is summable. The statement is clearly true for $k = 0$. Suppose that it holds for $k \geq 0$. Then it follows from (128) and the nonexpansivity of $\operatorname{prox}_{\gamma_k g}$ that

$$\begin{aligned}
&\|x_{k+1}-x\|^2 + 2\gamma_k(F(x_{k+1})-F(x)) \\
&\quad\le 2\gamma_k(\|x-\tilde{x}_{k+1}\|+\|x_{k+1}-\tilde{x}_{k+1}\|)\|\hat{u}_k-\nabla f(x_k)\|+\|x_k-x\|^2 \\
&\quad\le 2\gamma_k(\|x-\operatorname{prox}_{\gamma_k g}(x)\|+\|x-x_k+\gamma_k\nabla f(x_k)\|+\gamma_k\|\hat{u}_k-\nabla f(x_k)\|) \\
&\qquad\times\|\hat{u}_k-\nabla f(x_k)\|+\|x_k-x\|^2 \qquad (129)
\end{aligned}$$

and hence we derive that $\|x_{k+1}\|$ is square summable and $F(x_{k+1})$ is summable. Moreover, since ∇f is Lipschitz continuous, we have $\|\nabla f(x_{k+1})\| \le L\|x_{k+1}-x\|+\|\nabla f(x)\|$, which implies that $\|\nabla f(x_{k+1})\|$ is square summable too. given x_k in (128) and recalling that $\mathsf{E}[\hat{u}_k \mid x_k] = \nabla f(x_k)$, we get

$$\begin{aligned}
&\mathsf{E}[\|x_{k+1}-x\|^2|x_k] + 2\gamma_k\mathsf{E}[F(x_{k+1})-F(x)|x_k] \\
&\le \|x_k-x\|^2 + 2\gamma_k\mathsf{E}[\langle\tilde{x}_{k+1}-x_{k+1},\hat{u}_k-\nabla f(x_k)\rangle|x_k]. \qquad (130)
\end{aligned}$$

Since $\operatorname{prox}_{\gamma g}$ is nonexpansive by Proposition 34, we derive

$$\mathsf{E}[\|x_{k+1}-x\|^2|x_k] + 2\gamma_k\mathsf{E}[F(x_{k+1})-F(x)|x_k] \le \|x_k-x\|^2 + 2\gamma_k^2\sigma^2, \qquad (131)$$

and this yields

$$2\gamma_k(\mathsf{E}[F(x_{k+1})]-F(x)) \le \mathsf{E}[\|x_k-x\|^2] - \mathsf{E}[\|x_{k+1}-x\|^2] + 2\gamma_k^2\sigma^2. \qquad (132)$$

The above equation is the same as (113) except for the fact that $F(x_k)$ and B^2 are replaced by $F(x_{k+1})$ and σ^2 respectively. The proof thus essentially continues as the one of Theorem 77.

4.3 Randomized Block-Coordinate Descent

In this section, we address the following problem

$$\operatorname*{minimize}_{\boldsymbol{x}\in X} F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}), \qquad g(\boldsymbol{x}) = \sum_{i=1}^m g_i(x_i), \qquad (133)$$

where X is the direct sum of m separable real Hilbert spaces $(X_i)_{1\le i\le m}$, i.e.,

$$X = \bigoplus_{i=1}^m X_i \quad \text{and} \quad (\forall\, \boldsymbol{x}=(x_1,\cdots,x_m), \boldsymbol{y}=(y_1,\cdots,y_m)\in X) \quad \langle \boldsymbol{x},\boldsymbol{y}\rangle = \sum_{i=1}^m \langle x_i, y_i\rangle$$

and the following assumptions hold

A1 $f\colon X \mapsto \mathbb{R}$ is convex and differentiable with Lipschitz continuous gradient.
A2 $(\forall i \in [m] := \{1,\dots,m\})$, $g_i \in \Gamma_0(X_i)$.

We study the following algorithm.

Algorithm 5 *(The randomized block-coordinate proximal gradient method) Let* $\boldsymbol{x}^0 = (x_1^0, \dots, x_m^0) \in X$ *and* $(\gamma_i)_{1\le i\le m} \in \mathbb{R}^m_{++}$. *Then,*

$$\begin{array}{l} \textit{for } k = 0, 1, \dots \\ \quad \textit{for } i = 0, 1, \dots, m \\ \quad\quad x_i^{k+1} = \begin{cases} \operatorname{prox}_{\gamma_k g_{i_k}}\left(x_{i_k}^k - \gamma_{i_k} \nabla_{i_k} f(\boldsymbol{x}^k)\right) & \textit{if } i = i_k \\ x_i^k & \textit{if } i \neq i_k \end{cases} \end{array} \tag{134}$$

where $(i_k)_{k\in\mathbb{N}}$ *are independent random variables taking values in* $\{1, \dots, m\}$ *with* $\mathsf{p}_i := P(i_k = i) > 0$ *for all* $i \in \{1, \dots, m\}$.

In the following we denote by $J_i \colon X_i \to X$ the canonical embedding of X_i into X, that is, $J_i(x_i) = (0, \dots, x_i, \dots, 0)$, where x_i occurs in the i-th position. Thus, the algorithm can be equivalently written as

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + J_{i_k}\left(\operatorname{prox}_{\gamma_{i_k} g_{i_k}}(x_{i_k}^k - \gamma_{i_k} \nabla_{i_k} f(\boldsymbol{x}^k)) - x_{i_k}^k\right). \tag{135}$$

Moreover, we set

$$\Gamma^{-1} = \bigoplus_{i=1}^m \frac{1}{\gamma_i} \mathrm{Id}_i, \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\Gamma^{-1}} = \sum_{i=1}^m \frac{1}{\gamma_i} \langle x_i, y_i \rangle \tag{136}$$

and

$$W = \bigoplus_{i=1}^m \frac{1}{\gamma_i \mathsf{p}_i} \mathrm{Id}_i, \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle_W = \sum_{i=1}^m \frac{1}{\gamma_i \mathsf{p}_i} \langle x_i, y_i \rangle . \tag{137}$$

Remark 84 Algorithm 5 can be interpreted as a stochastic optimization algorithm which uses special stochastic gradients and proximity operators oracles. Indeed, let ξ be a random variable with values in $\{1, \dots, m\}$ distributed as i_k and let

$$\hat{g}(\boldsymbol{x}, \xi) = \frac{1}{\mathsf{p}_\xi} g_\xi(x_\xi). \tag{138}$$

Then, clearly $\mathsf{E}[\hat{g}(\boldsymbol{x}, \xi)] = \sum_{i=1}^m g_i(x_i) = g(\boldsymbol{x})$. Moreover,

$$\begin{aligned} \operatorname{prox}^W_{\hat{g}(\cdot,\xi)}(\boldsymbol{x}) &= \operatorname{argmin}_{\boldsymbol{y}\in X} \left\{ \frac{1}{\mathsf{p}_\xi} g_\xi(x_\xi) + \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_W^2 \right\} \\ &= \operatorname{argmin}_{y\in X} \left\{ \frac{1}{\mathsf{p}_\xi} g_\xi(x_\xi) + \frac{1}{2\gamma_\xi, \mathsf{p}_\xi} (y_\xi - x_\xi)^2 + \sum_{i\neq\xi} \frac{1}{2\gamma_i \mathsf{p}_i} (y_i - x_i)^2 \right\} \end{aligned}$$

and hence

$$(\forall\, i \in \{1,\dots,m\}) \quad [\mathrm{prox}^W_{\hat{g}(\cdot,\xi)}(\boldsymbol{x})]_i = \begin{cases} x_i & \text{if } i \neq \xi \\ \mathrm{prox}_{\gamma_\xi g_\xi}(x_\xi) & \text{if } i = \xi. \end{cases} \tag{139}$$

Also, if we set $\hat{\nabla}^W_\xi f(\boldsymbol{x}) = \gamma_\xi J_\xi(\nabla_\xi f(\boldsymbol{x}))$, we have

$$\mathsf{E}[\hat{\nabla}^W_\xi f(\boldsymbol{x})] = (\gamma_i \mathsf{p}_i \nabla_i f(\boldsymbol{x}))_{1\le i\le m} = W^{-1}\nabla f(\boldsymbol{x}) = \nabla^W f(\boldsymbol{x}). \tag{140}$$

Therefore, it is clear that Algorithm 5 can be rewritten as a stochastic proximal gradient algorithm in the metric W as follows

$$\boldsymbol{x}^{k+1} = \mathrm{prox}^W_{\hat{g}(\cdot,i_k)}(\boldsymbol{x}^k - \hat{\nabla}^W_{i_k} f(\boldsymbol{x}^k)). \tag{141}$$

Proposition 85 *Let $f\colon X \to \mathbb{R}$ be a convex differentiable function. Then the following statements are equivalent.*

(i) *∇f is Lipschitz continuous.*
(ii) *There exists $(L_i)_{1\le i\le m} \in \mathbb{R}^m_+$ such that for all $i \in \{1,\dots,m\}$ and $\boldsymbol{x} = (x_1,\dots,x_m) \in X$, the mapping $\nabla_i f(x_1,\dots,x_{i-1},\cdot,x_{i+1},\cdots,x_m)\colon X_i \to X_i$ is Lipschitz continuous with constant L_i.*

Proof (i) $\Rightarrow$ (ii): Let L be a Lipschitz constant of ∇f. Then (ii) holds with $(L_i)_{1\le i\le m} \equiv L$.

(ii) $\Rightarrow$ (i): Let, for every $i \in [m]$, $q_i = L_i / \sum_{j=1}^m L_j$. Then $(q_i)_{1\le i\le m} \in \mathbb{R}^m_+$ and $\sum_{i=1}^m q_i = 1$. Let $x, v \in X$. Then

$$\begin{aligned}
f(x+v) &= f\Big(x + \sum_{i=1}^m J_i(v_i)\Big) \\
&= f\Big(\sum_{i=1}^m q_i (x + q_i^{-1} J_i(v_i))\Big) \\
&\le \sum_{i=1}^m q_i f(x + q_i^{-1} J_i(v_i)) \\
&\le \sum_{i=1}^m q_i \big(f(x) + \langle q_i^{-1} v_i, \nabla_i f(x)\rangle + \frac{L_i}{2}\|q_i^{-1} v_i\|^2\big) \\
&= f(x) + \langle v, \nabla f(x)\rangle + \sum_{i=1}^m \frac{L_i}{2q_i}\|v_i\|^2 \\
&= f(x) + \langle v, \nabla f(x)\rangle + \frac{\sum_{i=1}^m L_i}{2}\|v\|^2.
\end{aligned}$$

Therefore, Fact 1(ii) yields that ∇f is Lipschitz continuous. □

Remark 86 Let $f\colon X \to \mathbb{R}$ be a convex differentiable function with Lipschitz continuous gradient. The constants $(L_i)_{1\le i\le m}$ defined in Proposition 85 are called the *block-Lipschitz constants of the partial gradients* $\nabla_i f$. Then the following block-coordinate descent lemma holds

$$(\forall\, v_i \in X_i) \quad f(\boldsymbol{x} + J_i(v_i)) \le f(\boldsymbol{x}) + \langle v_i, \nabla f_i(\boldsymbol{x})\rangle + \frac{L_i}{2}\|v_i\|^2. \tag{142}$$

Lemma 87 *Let X be a real Hilbert space. Let $\varphi\colon X \to \mathbb{R}$ be differentiable and convex and $\psi \in \Gamma_0(X)$. Let $x \in X$ and set $x^+ = \operatorname{prox}_{\psi}(x - \nabla\varphi(x))$. Then, for all $z \in X$,*

$$\begin{aligned}\langle z - x, x - x^+\rangle \le \big((\varphi+\psi)(z) - (\varphi+\psi)(x) - \|z-x\|^2\big)\\ + \big(\psi(x) - \psi(x^+) + \langle x - x^+, \nabla\varphi(x)\rangle\big) - \|x - x^+\|^2.\end{aligned}$$

Proof Let $z \in X$. By definition of x^+ we have $x - x^+ - \nabla\varphi(x) \in \partial\psi\left(x^+\right)$. Therefore, $\psi(z) \ge \psi\left(x^+\right) + \langle z - x^+, x - x^+ - \nabla\varphi(x)\rangle$, and hence

$$\langle z - x^+, x - x^+\rangle \le \psi(z) - \psi\left(x^+\right) + \langle z - x^+, \nabla\varphi(x)\rangle. \tag{143}$$

Now, we note that $\|x^+ - z\|^2 = \|x^+ - x\|^2 + \|x - z\|^2 + 2\langle x^+ - x, x - z\rangle$. Then,

$$\begin{aligned}\langle z - x, x - x^+\rangle + \langle x - x^+, x - x^+\rangle\\ \le \psi(z) - \psi\left(x^+\right) + \langle z - x, \nabla\varphi(x)\rangle + \langle x - x^+, \nabla\varphi(x)\rangle\end{aligned}$$

and hence

$$\begin{aligned}\langle z - x, x - x^+\rangle \le \psi(z) - \psi(x) + \langle z - x, \nabla\varphi(x)\rangle + \psi(x) - \psi(x^+)\\ + \langle x - x^+, \nabla\varphi(x)\rangle - \|x - x^+\|^2.\end{aligned}$$

Since $\langle z - x, \nabla\varphi(x)\rangle \le \varphi(z) - \varphi(x) - (\mu_\varphi/2)\|z - x\|^2$, the statement follows. □

Now we set

$$\begin{aligned}\bar{\boldsymbol{x}}^{k+1} &= \left(\operatorname{prox}_{\gamma_i g_i}(x_i^k - \gamma_i \nabla_i f(\boldsymbol{x}^k))\right)_{1\le i\le m}\\ \boldsymbol{\Delta}^k &= \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}.\end{aligned} \tag{144}$$

Then, recalling (135), we have

$$\bar{x}_{i_k}^{k+1} = \operatorname{prox}_{\gamma_{i_k} g_{i_k}}\left(x_{i_k}^k - \gamma_{i_k}\nabla_{i_k} f(\boldsymbol{x}^k)\right) = x_{i_k}^{k+1} \qquad \Delta_{i_k}^k = x_{i_k}^k - x_{i_k}^{k+1}. \tag{145}$$

Also note that

$$\boldsymbol{x}^k = \boldsymbol{x}^k(i_0, \ldots, i_{k-1}) \quad \text{and} \quad \bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^{k+1}(i_0, \ldots, i_{k-1}).$$

We derive from (145) that

$$\frac{x_{i_k}^k - x_{i_k}^{k+1}}{\gamma_{i_k}} - \nabla_{i_k} f(\boldsymbol{x}^k) \in \partial g_{i_k}(x_{i_k}^{k+1}) \tag{146}$$

Proposition 88 *Let f and g satisfy Assumptions 4.3 and 4.3. Let $(L_i)_{1\le i\le m}$ be the block-Lipschitz constants of the partial gradients $\nabla_i f$ as defined in Proposition 85. Let $(\gamma_i)_{1\le i\le m} \in \mathbb{R}^m_{++}$ be such that $\gamma_i < 2/L_i$. Set $\delta = \max_{1\le i\le m} \gamma_i L_i$ and $\mathsf{p}_{\min} = \min_{1\le i\le m} \mathsf{p}_i$. Let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ be generated by Algorithm 5. Then, for all $\boldsymbol{x} \in X$,*

$$\begin{aligned}\langle \boldsymbol{x} - \boldsymbol{x}^k, \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\rangle_{\Gamma^{-1}} \le{}& \frac{1}{\mathsf{p}_{\min}} \mathsf{E}\big[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \,|\, i_0, \dots, i_{k-1}\big] \\ &+ \big(F(\boldsymbol{x}) - F(\boldsymbol{x}^k)\big) + \frac{\delta - 2}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}}. \end{aligned}\tag{147}$$

Proof First note that $\bar{\boldsymbol{x}}^{k+1} = \operatorname{prox}_g^{\Gamma^{-1}}(\boldsymbol{x} - \nabla^{\Gamma^{-1}} f(\boldsymbol{x}^k))$, where the prox and the gradient are computed in the weighted norm $\|\cdot\|_{\Gamma^{-1}}$. Then we derive from Lemma 87 written in the norm $\|\cdot\|_{\Gamma^{-1}}$ that

$$\begin{aligned}\langle \boldsymbol{x} - \boldsymbol{x}^k, \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\rangle_{\Gamma^{-1}} \le{}& \big(F(\boldsymbol{x}) - F(\boldsymbol{x}^k)\big) \\ &+ g(\boldsymbol{x}^k) - g(\bar{\boldsymbol{x}}^{k+1}) + \langle \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}, \nabla f(\boldsymbol{x}^k)\rangle \\ &- \|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\|^2_{\Gamma^{-1}}. \end{aligned}\tag{148}$$

Next, we have

$$\begin{aligned}&g(\boldsymbol{x}^k) - g(\bar{\boldsymbol{x}}^{k+1}) + \langle \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}, \nabla f(\boldsymbol{x}^k)\rangle \\ &= \mathsf{E}\left[\frac{1}{\mathsf{p}_{i_k}}\big(g_{i_k}(x_{i_k}^k) - g_{i_k}(\bar{x}_{i_k}^{k+1}) + \langle x_{i_k}^k - \bar{x}_{i_k}^{k+1}, \nabla_{i_k} f(\boldsymbol{x}^k)\rangle\big) \,|\, i_0, \dots, i_{k-1}\right]\end{aligned}$$

Moreover, since $x_{i_k}^{k+1} = \bar{x}_{i_k}^{k+1}$ and $\boldsymbol{x}^k$ and $\boldsymbol{x}^{k+1}$ differ only for the i_k-th component

$$\begin{aligned}&\frac{1}{\mathsf{p}_{i_k}}\big(g_{i_k}(x_{i_k}^k) - g_{i_k}(\bar{x}_{i_k}^{k+1}) + \langle x_{i_k}^k - \bar{x}_{i_k}^{k+1}, \nabla_{i_k} f(\boldsymbol{x}^k)\rangle\big) \\ &\quad= \frac{1}{\mathsf{p}_{i_k}}\big(g(\boldsymbol{x}^k) - g(\boldsymbol{x}^{k+1}) + \langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \nabla f(\boldsymbol{x}^k)\rangle\big) \\ &\quad= \frac{1}{\mathsf{p}_{\min}}\big(g(\boldsymbol{x}^k) - g(\boldsymbol{x}^{k+1}) + \langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \nabla f(\boldsymbol{x}^k)\rangle\big) \\ &\qquad- \Big(\underbrace{\frac{1}{\mathsf{p}_{\min}} - \frac{1}{\mathsf{p}_{i_k}}}_{\ge 0}\Big)\big(g_{i_k}(x_{i_k}^k) - g_{i_k}(x_{i_k}^{k+1}) + \langle x_{i_k}^k - x_{i_k}^{k+1}, \nabla_{i_k} f(\boldsymbol{x}^k)\rangle\big) \\ &\quad\le \frac{1}{\mathsf{p}_{\min}}\big(g(\boldsymbol{x}^k) - g(\boldsymbol{x}^{k+1}) + \langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \nabla f(\boldsymbol{x}^k)\rangle\big)\end{aligned}$$

$$- \left(\frac{1}{\mathsf{p}_{\min}} - \frac{1}{\mathsf{p}_{i_k}} \right) \frac{1}{\gamma_{i_k}} \|\Delta_{i_k}^k\|^2,$$

where in the last inequality we used that

$$- \left(g_{i_k}(x_{i_k}^k) - g_{i_k}(x_{i_k}^{k+1}) + \langle x_{i_k}^k - x_{i_k}^{k+1}, \nabla_{i_k} f(\boldsymbol{x}^k) \rangle \right) \leq -\frac{1}{\gamma_i} \|\Delta_i^k\|^2 \tag{149}$$

which was obtained by the fact that $v_i = (x_{i_k}^k - x_{i_k}^{k+1})/\gamma_{i_k} - \nabla_{i_k} f(\boldsymbol{x}^k) \in \partial g_{i_k}(x_{i_k}^{k+1})$. So

$$\begin{aligned} & g(\boldsymbol{x}^k) - g(\bar{\boldsymbol{x}}^{k+1}) + \langle \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}, \nabla f(\boldsymbol{x}^k) \rangle \\ & \quad \leq \frac{1}{\mathsf{p}_{\min}} \mathsf{E}[g(\boldsymbol{x}^k) - g(\boldsymbol{x}^{k+1}) + \langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \nabla f(\boldsymbol{x}^k) \rangle \mid i_0, \ldots, i_{k-1}] \\ & \qquad - \frac{1}{\mathsf{p}_{\min}} \sum_{i=1}^m \frac{\mathsf{p}_i}{\gamma_i} \|\Delta_i^k\|^2 + \|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\|_{\Gamma^{-1}}^2. \end{aligned} \tag{150}$$

Now, we derive from the block-coordinate descent lemma (142) and the fact that $\boldsymbol{x}^k$ and $\boldsymbol{x}^{k+1}$ differ only in the i_k-th component, that

$$\begin{aligned} & \mathsf{E}[\langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \nabla f(\boldsymbol{x}^k) \rangle \mid i_0, \ldots, i_{k-1}] \\ & \qquad \leq \mathsf{E}[f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1}) + \frac{L_{i_k}}{2} \|\Delta_{i_k}^k\|^2 \mid i_0, \ldots, i_{k-1}] \\ & \qquad \leq \mathsf{E}[f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1}) \mid i_0, \ldots, i_{k-1}] + \frac{1}{2} \sum_{i=1}^m \mathsf{p}_i L_i \|\Delta_i^k\|^2. \end{aligned}$$

Therefore it follows from the above inequality and (150) that

$$\begin{aligned} & g(\boldsymbol{x}^k) - g(\bar{\boldsymbol{x}}^{k+1}) + \langle \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}, \nabla f(\boldsymbol{x}^k) \rangle \\ & \quad \leq \frac{1}{\mathsf{p}_{\min}} \mathsf{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \mid i_0, \ldots, i_{k-1}] \\ & \qquad + \frac{1}{2\mathsf{p}_{\min}} \sum_{i=1}^m \frac{\mathsf{p}_i}{\gamma_i} (\gamma_i L_i - 2 - \sigma_{\Gamma^{-1}}) \|\Delta_i^k\|^2 + \|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\|_{\Gamma^{-1}}^2 \\ & \quad \leq \frac{1}{\mathsf{p}_{\min}} \mathsf{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \mid i_0, \ldots, i_{k-1}] \\ & \qquad + \frac{\delta - 2}{2} \|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\|_{\Gamma^{-1}}^2 + \|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\|_{\Gamma^{-1}}^2, \end{aligned}$$

where in the last inequality we used that $\gamma_i L_i - 2 \leq \delta - 2 \leq 0$ and that $\mathsf{p}_i \geq \mathsf{p}_{\min}$. The statement follows from (148). □

Proposition 89 *Under the assumptions of Proposition 88 suppose additionally that* $\boldsymbol{x}$ *is an* X*-valued random variable which is measurable w.r.t. to the* σ*-algebra generated by* $i_0, \dots, i_{k-1}$*. Then*

$$\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\|_W^2 \,|\, i_0, \dots, i_{k-1}] - \|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2 = \|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}\|_{\Gamma^{-1}}^2 - \|\boldsymbol{x}^k - \boldsymbol{x}\|_{\Gamma^{-1}}^2 \tag{151}$$

and $\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_W^2 \,|\, i_0, \dots, i_{k-1}] = \|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}^k\|_{\Gamma^{-1}}^2$.

Proof If follows from Fact 75(v) that

$$\begin{aligned}
&\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\|_W^2 \,|\, i_0, \dots, i_{k-1}] \\
&= \mathsf{E}\Big[\sum_{i=1}^m \frac{1}{\gamma_i \mathsf{p}_i} \|x_i^{k+1} - x_i\|^2 \,\big|\, i_0, \dots, i_{k-1}\Big] \\
&= \mathsf{E}\Big[\|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2 - \frac{1}{\gamma_{i_k} \mathsf{p}_{i_k}} \|x_i^k - x_i\|^2 + \frac{1}{\gamma_{i_k} \mathsf{p}_{i_k}} \|\bar{x}_i^{k+1} - x_i\|^2 \,\big|\, i_0, \dots, i_{k-1}\Big] \\
&= \|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2 - \|\boldsymbol{x}^k - \boldsymbol{x}\|_{\Gamma^{-1}}^2 + \|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}\|_{\Gamma^{-1}}^2
\end{aligned}$$

The second equation follows from (151), by choosing $\boldsymbol{x} = \boldsymbol{x}^k$. □

Proposition 90 *Under the assumptions of Proposition 88 set* $F = f + g$*. Then, the following hold.*

(i) $(\mathsf{E}[F(\boldsymbol{x}^k)])_{k\in\mathbb{N}}$ *is decreasing.*

(ii) *Suppose that* $\inf_{k\in\mathbb{N}} \mathsf{E}[F(\boldsymbol{x}^k)] > \infty$*. Then,*

$$\sum_{k\in\mathbb{N}} \|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}^k\|_{\Gamma^{-1}}^2 = \sum_{k\in\mathbb{N}} \mathsf{E}\big[\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_W^2 \big| i_0, \dots, i_{k-1}\big] < +\infty \quad P \text{ a.s.}$$

(iii) *For every* $k \in \mathbb{N}$ *and every* $\boldsymbol{x} \in \operatorname{dom} F$

$$\begin{aligned}
&\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\|_W^2 \,|\, i_0, \dots, i_{k-1}] \\
&\qquad \le \|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2 - 2\big(F(\boldsymbol{x}^k) - F(\boldsymbol{x})\big) \\
&\qquad\quad + \frac{2}{\mathsf{p}_{\min}} \Big(\frac{(\delta-1)_+}{2-\delta} + 1\Big) \mathsf{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \,|\, i_0, \dots, i_{k-1}].
\end{aligned} \tag{152}$$

Proof Let $k \in \mathbb{N}$ and $\boldsymbol{x} \in \operatorname{dom} F$. Since

$$\|\boldsymbol{x}^k - \boldsymbol{x}\|_{\Gamma^{-1}}^2 - \|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}\|_{\Gamma^{-1}}^2 = -\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}\|_{\Gamma^{-1}}^2 + 2\langle \boldsymbol{x}^k - \bar{\boldsymbol{x}}^{k+1}, \boldsymbol{x}^k - \boldsymbol{x}\rangle_{\Gamma^{-1}},$$

we derive from (147), multiplied by 2, that

$$\|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}\|_{\Gamma^{-1}}^2 \leq \|\boldsymbol{x}^k - \boldsymbol{x}\|_{\Gamma^{-1}}^2 + (\delta - 1)\|\bar{\boldsymbol{x}}^{k+1} - \boldsymbol{x}^k\|_{\Gamma^{-1}}^2 + \frac{2}{\mathsf{p}_{\min}}\mathsf{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \,|\, i_0, \ldots, i_{k-1}] - 2\big(F(\boldsymbol{x}^k) - F(\boldsymbol{x})\big). \tag{153}$$

Then for an X-valued random variable $\boldsymbol{x}'$ measurable with respect to $i_0, \ldots, i_{k-1}$, Proposition 89 yields

$$\begin{aligned}&\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}'\|_W^2 \,|\, i_0, \ldots, i_{k-1}] \\ &\quad\leq \|\boldsymbol{x}^k - \boldsymbol{x}'\|_W^2 + (\delta - 1)\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_W^2 \,|\, i_0, \ldots, i_{k-1}] \\ &\qquad + \frac{2}{\mathsf{p}_{\min}}\mathsf{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \,|\, i_0, \ldots, i_{k-1}] \\ &\qquad - 2\big(F(\boldsymbol{x}^k) - F(\boldsymbol{x}')\big).\end{aligned} \tag{154}$$

Taking $\boldsymbol{x}' = \boldsymbol{x}^k$ in (154), we have

$$\frac{\mathsf{p}_{\min}}{2}(2 - \delta)\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_W^2 \,|\, i_0, \ldots, i_{k-1}] \leq \mathsf{E}\big[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \,|\, i_0, \ldots, i_{k-1}\big], \tag{155}$$

which plugged into (154), with $\boldsymbol{x}' \equiv \boldsymbol{x} \in \operatorname{dom} F$, gives (iii). Moreover, taking the expectation in (155), we obtain

$$\frac{\mathsf{p}_{\min}}{2}(2 - \delta)\mathsf{E}\big[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_W^2\big] \leq \mathsf{E}[F(\boldsymbol{x}^k)] - \mathsf{E}[F(\boldsymbol{x}^{k+1})], \tag{156}$$

which gives (i). Finally, set for all $k \in \mathbb{N}$, $\xi_k = \mathsf{E}\big[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1})\big|i_0, \ldots, i_{k-1}\big] \geq 0$. Then

$$\mathsf{E}\Big[\sum_{k=0}^{+\infty} \xi_k\Big] = \sum_{k=0}^{+\infty} \mathsf{E}[\xi_k] = \sum_{k=0}^{+\infty} \mathsf{E}[F(\boldsymbol{x}^k)] - \mathsf{E}[F(\boldsymbol{x}^{k+1})] \leq \mathsf{E}[F(\boldsymbol{x}^0)] - \inf_{k\in\mathbb{N}} \mathsf{E}[F(\boldsymbol{x}^k)].$$

This shows that if $\inf_{k\in\mathbb{N}} \mathsf{E}[F(\boldsymbol{x}^k)] > -\infty$, then $\sum_{k=0}^{+\infty} \xi_k$ is P-integrable and hence it is P-a.s. finite. Then (ii) follows from (155) and Proposition 89. □

Proposition 91 *Under the assumptions of Proposition 90, suppose in addition that F is bounded from below. Then, there exist $(\boldsymbol{y}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{v}^k)_{k\in\mathbb{N}}$, sequences of X-valued random variables, such that the following hold.*

(i) $\boldsymbol{v}^k \in \partial F(\boldsymbol{y}^k)$ *P-a.s.*
(ii) $\boldsymbol{y}^k - \boldsymbol{x}^k \to 0$ *and* $\boldsymbol{v}^k \to 0$ *P-a.s.*

Proof It follows from (144) that, $(x_i^k(\omega) - \bar{x}_i^{k+1}(\omega))/\gamma_i - \nabla_i f(\boldsymbol{x}^k(\omega)) \in \partial g_i(\bar{x}_i^{k+1}(\omega))$, for all $i \in [m]$ and $\omega \in \Omega$. Hence

$$\Big(\frac{x_i^k(\omega) - \bar{x}_i^{k+1}(\omega)}{\gamma_i}\Big)_{1\le i\le m} - \nabla f(\boldsymbol{x}^k(\omega)) \in \partial g(\bar{\boldsymbol{x}}^{k+1}(\omega)).$$

Set $\boldsymbol{y}^k = \bar{\boldsymbol{x}}^{k+1}$ and let $\boldsymbol{v}^k \colon \Omega \to X$ be such that, for every $\omega \in \Omega$,

$$\begin{aligned}\boldsymbol{v}^k(\omega) = \Big(\frac{x_i^k(\omega) - y_i^k(\omega)}{\gamma_i}\Big)_{1\le i\le m} &+ \nabla f(\boldsymbol{y}^k(\omega)) - \nabla f(\boldsymbol{x}^k(\omega)) \\ &\in \partial g(\boldsymbol{y}^k(\omega)) + \nabla f(\boldsymbol{y}^k(\omega)) = \partial F(\boldsymbol{y}^k(\omega)).\end{aligned}$$

Clearly $\boldsymbol{v}^k$ is measurable and hence it is a random variable. Moreover, for every $\omega \in \Omega$,

$$\|\boldsymbol{v}^k(\omega)\| \le \frac{1}{\gamma_{\min}}\|\boldsymbol{x}^k(\omega) - \boldsymbol{y}^k(\omega)\| + \|\nabla f(\boldsymbol{y}^k(\omega)) - \nabla f(\boldsymbol{x}^k(\omega))\|.$$

Now, since F is bounded from below, Proposition 90(ii) yields that $(\|\boldsymbol{y}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}})_{k\in\mathbb{N}}$ is summable P-a.s. and hence $\boldsymbol{y}^k - \boldsymbol{x}^k \to 0$ P-a.s. The statement follows from the fact that ∇f is Lipschitz continuous (see Proposition 85). □

Lemma 92 *(Stochastic Opial) Let X be a Hilbert space, let S be a nonempty subset of X be a subset and let $(x^k)_{k\in\mathbb{N}}$ be a random sequence on $(\Omega, \mathcal{A}, P)$ with values in X. Assume that*

(a) S is separable;
(b) for every $z \in S$, there exists Ω_z with $P(\Omega_z) = 1$ such that, for every $\omega \in \Omega_z$,

$$\exists \lim_k \|x^k(\omega) - z\|;$$

(c) there exists $\hat{\Omega}$ with $P(\hat{\Omega}) = 1$ such that, for every $\omega \in \hat{\Omega}$, every weak cluster point of $(x^k(\omega))$ belongs to S.

Then there exists a S-valued random variable $\bar{x}$ such that $x^k \rightharpoonup \bar{x}$ a.s.

Proof We first show that there exists $\tilde{\Omega}$ such that, for every $\omega \in \tilde{\Omega}$ and for every $z \in S$, there exists

$$\lim_n \|x^k(\omega) - z\|.$$

Let $W \subseteq S$ countable dense in S and let $\tilde{\Omega} = \bigcap_{w\in W} \Omega_w$. Then $P(\tilde{\Omega}) = 1$ and, for every $\omega \in \tilde{\Omega}$ and for every $w \in W$, there exists

$$\lim_n \|x^k(\omega) - w\|.$$

Fix $\omega \in \tilde{\Omega}$ and $z \in S$. Since W is dense in S, there exists a sequence (w_j) in W such that $w_j \to z$. Since $w_j \in W$ for every $j \geq 0$, we know that there exists

$$\lim_k \|x^k(\omega) - w_j\| = \tau_j(\omega). \tag{157}$$

Note that

$$-\|w_j - z\| \leq \|x^k(\omega) - z\| - \|x^k(\omega) - w_j\| \leq \|w_j - z\|. \tag{158}$$

Then, (157) and (158) yield

$$\begin{aligned} -\|w_j - z\| &\leq \liminf_k \left[\|x^k(\omega) - z\| - \|x^k(\omega) - w_j\|\right] \\ &= \liminf_k \|x^k(\omega) - z\| - \tau_j(\omega) \leq \limsup_k \|x^k(\omega) - z\| - \tau_j(\omega) \\ &= \limsup_k \left[\|x^k(\omega) - z\| - \|x^k(\omega) - w_j\|\right] \leq \|w_j - z\|. \end{aligned}$$

Taking the limit for $k \to +\infty$ and recalling that $w_k \to z$, we get that there exists $\lim_k \|x^k(\omega) - z\|$. So we proved that for every $\omega \in \tilde{\Omega}$ the limit of $\lim_k \|x^k(\omega) - z\|$ exists. Now suppose that $\bar{\Omega} := \tilde{\Omega} \cap \hat{\Omega}$. Then, for every $\omega \in \bar{\Omega}$, we have both that: for every $z \in \mathcal{Z}$, $\exists \lim_k \|x^k(\omega) - z\|$; every weak cluster point of $x^k(\omega)$ belongs to $\mathcal{Z}$. We conclude by Lemma 29 that, for every $\omega \in \bar{\Omega}$, there exists $\bar{x}(\omega) \in \mathcal{Z}$ such that $x^k(\omega) \rightharpoonup \bar{x}(\omega)$. □

Now we give the main convergence results, which extends to the stochastic setting the convergence rate of the (deterministic) proximal gradient algorithm given in Theorem 47.

Theorem 93 *Under the assumptions of Proposition 88 set $F = f + g$, $F_* = \inf F$, and $S_* = \operatorname{argmin} F \subset X$. Then, the following hold.*

(i) $\mathsf{E}[F(\boldsymbol{x}^k)] \to F_*$.
(ii) *Suppose that $S_* \neq \varnothing$. Then $\mathsf{E}[F(\boldsymbol{x}^k)] - F_* = o(1/k)$ and, for all integer $k \geq 1$,*

$$\mathsf{E}[F(\boldsymbol{x}^k)] - F_* \leq \left[\frac{\operatorname{dist}_W^2(\boldsymbol{x}^0, S_*)}{2} + \left(\frac{\max\{1, (2-\delta)^{-1}\}}{\mathsf{p}_{\min}} - 1\right)(F(\boldsymbol{x}^0) - F_*)\right]\frac{1}{k}.$$

Moreover, there exists a random variable $\boldsymbol{x}_*$ taking values in S_* such that $\boldsymbol{x}^k \rightharpoonup \boldsymbol{x}_*$ P-a.s.

Proof Proposition 90(iii) gives, for all $\boldsymbol{x} \in \operatorname{dom} F$ and $k \in \mathbb{N}$,

$$\begin{aligned} &\mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\|_W^2 \mid i_0, \ldots, i_{k-1}] \\ &\quad \leq \|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2 + 2\mathsf{E}[F(\boldsymbol{x}) - F(\boldsymbol{x}^{k+1}) \mid i_0, \ldots, i_{k-1}] + \xi_k, \end{aligned} \tag{159}$$

where

$$\xi_k = b_1 \mathsf{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) \,|\, i_0, \dots, i_{k-1}], \qquad b_1 = 2\left(\frac{\max\{1, (2-\delta)^{-1}\}}{\mathsf{p}_{\min}} - 1\right).$$

Note that the random variables $\boldsymbol{x}^k$'s are discrete with finite range and $(\mathsf{E}[F(\boldsymbol{x}^k)])_{k\in\mathbb{N}}$ is decreasing. Moreover, $\sum_{k\in\mathbb{N}} \mathsf{E}[\xi_k] \leq b_1(F(\boldsymbol{x}^0) - F_*)$. Therefore, taking the expectation in (159) we have

$$2\mathsf{E}[F(\boldsymbol{x}^{k+1})] - F(\boldsymbol{x}) \leq \mathsf{E}[\|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2] - \mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\|_W^2] + \mathsf{E}[\xi_k] \tag{160}$$

Since $(\mathsf{E}[F(\boldsymbol{x}^k)])_{k\in\mathbb{N}}$ is decreasing, $\mathsf{E}[F(\boldsymbol{x}^k)] \to \inf_{k\in\mathbb{N}} \mathsf{E}[F(\boldsymbol{x}^k)] \geq F_*$. Thus, the statement (i) is true if $\inf_{k\in\mathbb{N}} \mathsf{E}[F(\boldsymbol{x}^k)] = -\infty$. Suppose that $\inf_{k\in\mathbb{N}} \mathsf{E}[F(\boldsymbol{x}^k)] > -\infty$ and let $\boldsymbol{x} \in \operatorname{dom} F$. Then, the right hand side of (160), being summable, converges to zero. Therefore, $F_* \leq \lim_{k\to+\infty} \mathsf{E}[F(\boldsymbol{x}^{k+1})] \leq F(\boldsymbol{x})$. Since $\boldsymbol{x}$ is arbitrary in $\operatorname{dom} F$, (i) follows. Let $\boldsymbol{x} \in S_*$. Then, $F(\boldsymbol{x}) = F_*$ and (160) yields

$$2\sum_{k\in\mathbb{N}} \left(\mathsf{E}[F(\boldsymbol{x}^{k+1})] - F_*\right) \leq \mathsf{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}\|^2\right] + \sum_{k\in\mathbb{N}} \mathsf{E}[\xi_k] \leq \|\boldsymbol{x}^0 - \boldsymbol{x}\|^2 + b_1(F(\boldsymbol{x}^0) - F_*).$$

Therefore, we have $\sum_{k\in\mathbb{N}}(\mathsf{E}[F(\boldsymbol{x}^{k+1})] - F_*) \leq (\|\boldsymbol{x}^0 - \boldsymbol{x}\|^2 + b_1(F(\boldsymbol{x}^0) - F_*))/2$. Since $(\mathsf{E}[F(\boldsymbol{x}^{k+1})] - F_*)_{k\in\mathbb{N}}$ is decreasing, the first part of statement (ii) follows from Fact 46. Concerning the convergence of the iterates, we will use the stochastic Opial's Lemma 92. Let $\boldsymbol{x} \in \operatorname{argmin} F$. Then it follows from (159) that

$$(\forall\, k \in \mathbb{N}) \quad \mathsf{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\|_W^2 \,|\, i_0, \dots, i_{k-1}] \leq \|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2 + \xi_k.$$

Since $\mathsf{E}[\sum_{k\in\mathbb{N}} \xi_k] = \sum_{k\in\mathbb{N}} \mathsf{E}[\xi_k] < +\infty$, we have $\sum_{k\in\mathbb{N}} \xi_k < +\infty$ P-a.s. and hence $(\|\boldsymbol{x}^k - \boldsymbol{x}\|_W^2)_{k\in\mathbb{N}}$ is an almost supermartingale in the sense of Robbins and Siegmund [96]. Thus, there exists $\Omega_1 \subset \Omega$ such that $P(\Omega_1) = 1$ and for every $\omega \in \Omega_1$, $(\|\boldsymbol{x}^k(\omega) - \boldsymbol{x}\|_W^2)_{k\in\mathbb{N}}$ is convergent. Now, it follows from Proposition 91 that there exists $\Omega_2 \subset \Omega$ with $P(\Omega_2) = 1$, such that, for every $\omega \in \Omega_2$, $\boldsymbol{v}^k(\omega) \in \partial F(\boldsymbol{y}^k(\omega))$, $\boldsymbol{y}^k(\omega) - \boldsymbol{x}^k(\omega) \to 0$, and $\boldsymbol{v}^k(\omega) \to 0$. Therefore, let $\omega \in \Omega_2$ and let $(\boldsymbol{x}^{n_k}(\omega))_{k\in\mathbb{N}}$ be a subsequence of $(\boldsymbol{x}_k(\omega))_{k\in\mathbb{N}}$ such that $\boldsymbol{x}^{n_k}(\omega) \rightharpoonup \bar{\boldsymbol{x}}$. Then,

$$\boldsymbol{v}^{n_k}(\omega) \in \partial F(\boldsymbol{y}^{n_k}(\omega)) \quad \boldsymbol{y}^{n_k}(\omega) - \boldsymbol{x}^{n_k}(\omega) \to 0 \quad \boldsymbol{v}^{n_k}(\omega) \to 0. \tag{161}$$

Then, it follows from (161) that $\boldsymbol{y}^{n_k}(\omega) \rightharpoonup \boldsymbol{x}$ and, since ∂F is weak-strong closed, that $0 \in \partial F(\boldsymbol{x})$. Therefore the two conditions in the stochastic Opial's Lemma 92 are satisfied with $S = \operatorname{argmin} F$ and hence the statement follows.

4.4 Bibliographical Notes

Stochastic methods in optimization were initiated by Robbins and Monro [95], Kiefer and Wolfowitz [61], and Ermoliev [47]. These methods are nowadays very popular due to applications in deep machine learning [22]. The projected stochastic subgradient method was studied in [44, 80]. In the last years rate of convergence in the last iterates were also derived [105]. The proximal stochastic gradient which explicitly assumes the Lipschitz continuity of the gradient was studied in [2, 100]. The worst case convergence rate in expectation of proximal stochastic gradient method is much worse with respect to the one of proximal gradient method. Recently, variance reduction techniques have been studied to improve the convergence behavior of stochastic methods [59], at the cost of keeping previously computed gradients in memory. These techniques are particularly useful for empirical risk minimization problems, see [42, 56] and references therein. Randomized strategies in block coordinate descent methods were popularized by Nesterov in [84]. Since then a number of works appeared extending and improving the analysis under several aspects. We cite among others [35, 79, 94, 103, 114].

5 Dual Algorithms

In this section, we show how proximal gradient algorithms can be used on the dual problem, to derive new algorithmic solutions for the primal.

5.1 A Framework for Dual Algorithms

We consider the same setting of Sect. 2.6. Here we additionally assume that f is strongly convex with modulus of convexity $\mu > 0$. In this situation, it follows from Fact 13 that f^* is differentiable on X and ∇f^* is $1/\mu$-Lipschitz continuous. Moreover, since f is strongly convex, the primal problem $(\mathscr{P})$ admits a (unique) solution, say $\hat{x}$. We also assume that the calculus rule for subdifferentials (15) holds. Thus, in view of Fact 14, we have that a dual solution $\hat{u}$ also exists, the duality gap is zero, and the following KKT conditions hold

$$\hat{x} = \nabla f^*(-A^*\hat{u}) \quad \text{and} \quad A\hat{x} \in \partial g^*(\hat{u}). \tag{162}$$

So, in this case, *a dual solution uniquely determines the primal solution*. Actually, the map $u \mapsto \nabla f^*(-A^*u)$ provides a way to go from the dual space Y into the primal space X. See Fig. 2. The following proposition tells us even more.

Proposition 94 *Under the notation of Sect. 2.6, let $u \in Y$ and set $x = \nabla f^*(-A^*u)$. Then*

$$\frac{\mu}{2}\|x-\hat{x}\|^2 \leq \Psi(u) - \Psi(\hat{u}).$$

Proof It follows from the KKT conditions (162), Fact 11, and the definition of u that

$$f(\hat{x}) + f^*(-A^*\hat{u}) = \langle \hat{x}, -A^*\hat{u}\rangle \quad \text{and} \quad f(x) + f^*(-A^*u) = \langle x, -A^*u\rangle.$$

Thus, since $-A^*u \in \partial f(x)$ and f is μ-strongly convex,

$$\begin{aligned} f^*(-A^*u) - f^*(-A^*\hat{u}) &= f(\hat{x}) - f(x) + \langle \hat{x}, A^*\hat{u}\rangle - \langle x, A^*u\rangle \\ &\geq \langle \hat{x} - x, -A^*u\rangle + \frac{\mu}{2}\|\hat{x}-x\|^2 + \langle \hat{x}, A^*\hat{u}\rangle - \langle A^*u, x\rangle \\ &= \langle A\hat{x}, \hat{u} - u\rangle + \frac{\mu}{2}\|\hat{x}-x\|^2. \end{aligned}$$

Now, since $A\hat{x} \in \partial g^*(\hat{u})$, we have

$$g^*(u) - g^*(\hat{u}) \geq \langle A\hat{x}, u - \hat{u}\rangle.$$

Summing the two inequalities above, we have

$$\big(f^*(-A^*u) + g^*(u)\big) - \big(f^*(-A^*\hat{u}) + g^*(\hat{u})\big) \geq \frac{\mu}{2}\|x-\hat{x}\|^2$$

and the statement follows. □

We define the *duality gap function*

$$G\colon X \times Y \to \,]-\infty, +\infty\,], \qquad G(x,u) = \Phi(x) + \Psi(u).$$

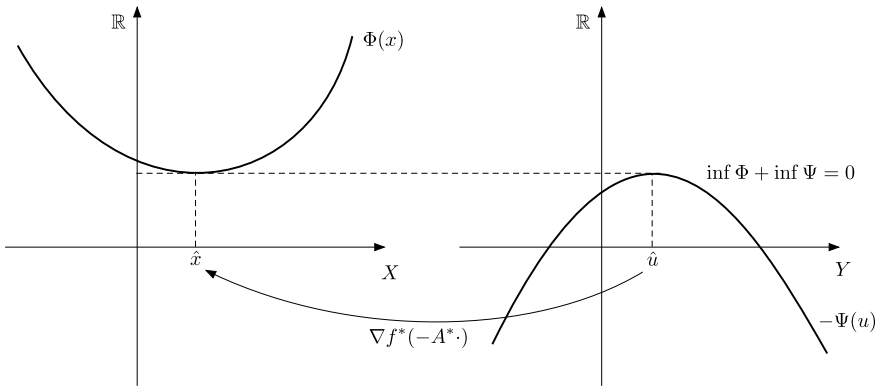


Fig. 2 Duality in strongly convex problems

Recall that if strong duality holds $\inf \Phi = -\inf \Psi$, and hence

$$(\Phi(x) - \inf \Phi) + (\Psi(u) - \inf \Psi) = G(x, u),$$

so the duality gap function bounds the primal and dual objectives. We have the following theorem

Theorem 95 *Under the notation of Sect. 2.6, suppose that $R(A) \subset \mathrm{dom}\partial g$. Then the following holds:*

(i) *Suppose that g^* is α-strongly convex. Let $u \in \mathrm{dom} g^*$ and set $x = \nabla f^*(-A^*u)$. Then,*

$$G(x, u) \leq \left(1 + \frac{\|A\|^2}{\alpha\mu}\right)(\Psi(u) - \inf \Psi). \tag{163}$$

(ii) *Suppose that g is L-Lipschitz continuous. Let $u \in \mathrm{dom} g^*$ be such that $\Psi(u) - \inf \Psi < \|A\|^2 L^2/\mu$ and set $x = \nabla f^*(-A^*u)$. Then, we have*

$$G(x, u) \leq 2\frac{\|A\|L}{\mu^{1/2}}(\Psi(u) - \inf \Psi)^{1/2}. \tag{164}$$

Proof Let $u \in \mathrm{dom} g^*$ and let $x = \nabla f^*(-A^*u)$. Since $R(A) \subset \mathrm{dom}\partial g$, we have $\partial g(Ax) \neq \varnothing$. Let $v \in \partial g(Ax)$. Then we first prove that for every $s \in [0, 1]$,

$$\Psi(u) - \inf \Psi \geq sG(x, u) + \frac{s}{2}\left(\alpha(1 - s) - \frac{s}{\mu}\|A\|^2\right)\|u - v\|^2. \tag{165}$$

Indeed, let $s \in [0, 1]$. Then

$$\begin{aligned}\Psi(u) - \inf \Psi &\geq \Psi(u) - \Psi(u + s(v - u)) \\ &= g^*(u) - g^*(u + s(v - u)) \\ &\quad + f^*(-A^*u) - f^*\big(-A^*u - sA^*(v - u)\big).\end{aligned} \tag{166}$$

Now, since f^* is $(1/\mu)$-Lipschitz smooth, we have

$$\begin{aligned}&f^*\big(-A^*u - sA^*(v - u)\big) - f^*(-A^*u) \\ &\leq \big\langle -sA^*(v - u), \nabla f^*(-A^*u)\big\rangle + \frac{1}{2\mu}s^2\|A\|^2\|v - u\|^2.\end{aligned}$$

Moreover, since g^* is α-strongly convex ($\alpha \geq 0$),

$$g^*(u + s(v - u)) - g^*(u) \leq s(g^*(v) - g^*(u)) - \alpha\frac{s(1 - s)}{2}\|u - v\|^2.$$

Therefore, it follows from (166) that

$$\begin{aligned}\Psi(u) - \inf \Psi &\geq \Psi(u) - \Psi(u + s(v-u)) \\ &\geq s\big(g^*(u) - g^*(v) - \langle x, A^*(u-v)\rangle\big) \\ &\quad + \frac{s}{2}\left(\alpha(1-s) - \frac{s}{\mu}\|A\|^2\right)\|v-u\|^2. \end{aligned} \tag{167}$$

Now, we note that

$$\begin{aligned}G(x,u) &= \Phi(x) + \Psi(u) \\ &= \big(f(x) + f^*(-A^*u) - \langle -A^*u, x\rangle\big) + \big(g(Ax) + g^*(u) - \langle Ax, u\rangle\big).\end{aligned}$$

Moreover, since $x = \nabla f^*(-A^*u)$ and $v \in \partial g(Ax)$, Young equality yields

$$f(x) + f^*(-A^*u) - \langle -A^*u, x\rangle = 0 \quad \text{and} \quad g(Ax) + g^*(v) - \langle Ax, v\rangle = 0.$$

Therefore,

$$G(x,u) = g^*(u) - g^*(v) - \langle Ax, u - v\rangle. \tag{168}$$

In conclusion, (165) follows from (167) and (168).

(i): If in (165) we chose $s = \alpha/(\alpha + \|A\|^2/\mu)$ we have $\alpha(1-s) - s\|A\|^2/\mu = 0$ and hence

$$\frac{\alpha}{\alpha + \|A\|^2/\mu} G(x,u) \leq \Psi(u) - \inf \Psi.$$

Then (163) follows.

(ii): It follows from (165) with $\alpha = 0$ that, for every $s \in [0,1]$,

$$sG(x,u) \leq \Psi(u) - \inf \Psi + \frac{s^2}{2\mu}\|A\|^2\|u-v\|^2.$$

Since g is L-Lipschitz continuous, we have $\mathrm{dom} g^* \subset B_L(0)$. Moreover, $u \in \mathrm{dom} g^*$ and $v \in \partial g(Ax) \Rightarrow Ax \in \partial g^*(v) \Rightarrow v \in \mathrm{dom} g^*$. Therefore, $\|u-v\|^2 \leq 2(\|u\|^2 + \|v\|^2) \leq 2L^2$. Then,

$$G(x,u) \leq \inf_{s\in[0,1]} \frac{1}{s}(\Psi(u) - \inf \Psi) + \frac{s}{\mu}\|A\|^2 L^2.$$

Since, if $0 < a < b$, $\min_{s\in[0,1]}(a/s + bs) = 2\sqrt{ab}$, the statement follows. □

5.2 *Dual Proximal Gradient Algorithms*

It follows from Proposition 94 and Theorem 95 that if an algorithm, applied to the dual problem $(\mathscr{D})$, provides a minimizing sequence, that is, a sequence $(u_k)_{k\in\mathbb{N}}$ such

that $\Psi(u_k) \to \inf \Psi$, then, the sequence $(x_k)_{k\in\mathbb{N}}$, defined as $x_k = \nabla f^*(-A^*u_k)$ is converging (possibly also in function values) to the solution of the primal problem. In particular, we have

$$\|x_k - \hat{x}\|^2 \leq \frac{2}{\mu}\big(\Psi(u_k) - \inf \Psi\big) \to 0,$$

and, depending on the assumptions in Theorem 95,

$$\Phi(x_k) - \inf \Phi \leq O\big(\Psi(u_k) - \inf \Psi\big) \to 0$$
$$\text{or}$$
$$\Phi(x_k) - \inf \Phi \leq O\big(\sqrt{\Psi(u_k) - \inf \Psi}\big) \to 0.$$

Since the gradient of the term $f^*(-A^*\cdot)$ in $(\mathscr{D})$ is Lipschitz continuous with constant $\|A\|^2/\mu$, the proximal gradient algorithm applied to $(\mathscr{D})$ leads to the following

Algorithm 6 (Dual proximal gradient algorithm) *Let $u^0 \in Y$ and $0 < \gamma < \frac{2\mu}{\|A\|^2}$. Then,*

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\quad x_k = \nabla f^*(-A^*u_k) \\ &\quad u_{k+1} = \operatorname{prox}_{\gamma g^*}(u_k + \gamma A x_k). \end{aligned} \tag{169}$$

Then, since Theorem 47(iv) ensures that $\Psi(u_k) - \Psi(\hat{u}) = o(1/(k+1))$, we have

$$\|x_k - \hat{x}\| \leq o(1/\sqrt{k+1})$$

and, again, in the settings of Theorem 95,

$$\Phi(x_k) - \inf \Phi \leq o(1/(k+1)) \quad \text{or} \quad \Phi(x_k) - \inf \Phi \leq o(1/\sqrt{k+1}).$$

Similarly, we can apply Algorithm 2 to the dual problem $(\mathscr{D})$ and this yields the following dual algorithm.

Algorithm 7 (Dual accelerated proximal gradient algorithm) *Let $0 < \gamma \leq \mu/\|A\|^2$ and let $(t_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be defined as Proposition 64 with $1 - c \geq 2\sqrt{b}$. Let $u_0 = v_0 \in Y$ and define*

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\quad y_k = \nabla f^*(-A^*v_k) \\ &\quad u_{k+1} = \operatorname{prox}_{\gamma g^*}(v_k + \gamma A y_k) \\ &\quad \beta_{k+1} = \frac{t_k - 1}{t_{k+1}} \\ &\quad v_{k+1} = u_{k+1} + \beta_{k+1}(u_{k+1} - u_k). \end{aligned} \tag{170}$$

Then, defining $x_k = \nabla f^*(-A^*u_k)$, Theorem 68 yield

$$\|x_k - \hat{x}\| \leq O(1/k) \tag{171}$$

and, under the assumptions of Theorem 95, that

$$\Phi(x_k) - \inf \Phi \leq O(1/k^2) \quad \text{or} \quad \Phi(x_k) - \inf \Phi \leq O(1/k).$$

Finally, suppose that g is separable, meaning that

$$g\colon Y := \bigoplus_{i=1}^m Y_i \to \left]-\infty, +\infty\right], \quad g(y_1, \ldots, y_m) = \sum_{i=1}^m g_i(y_i), \tag{172}$$

and $A\colon X \to Y$ with $Ax = (A_1 x, \ldots, A_m x)$, where $A_i\colon X \to Y_i$ are bounded linear operators. Then g^* is separable as well and $A^*\colon Y \to X$ is such that $A^* u = \sum_{i=1}^m A_i^* u_i$. Hence, one can apply Algorithm 5 to the dual problem $(\mathscr{D})$, yielding the following stochastic dual algorithm.

Algorithm 8 (stochastic dual block coordinate gradient ascent method) *Let* $\boldsymbol{u}^0 = (u_1^0, \ldots, u_m^0) \in Y$ *and let* $(\gamma_i)_{1\leq i\leq m} \in \mathbb{R}^m_{++}$ *be such that* $0 < \gamma_i < 2\mu/\|A_i\|^2$. *Then,*

$$\begin{array}{l} \textit{for } k = 0, 1, \ldots \\ \left|\begin{array}{l} \boldsymbol{x}^k = \nabla f^*(-A^*\boldsymbol{u}^k) \\ \textit{for } i = 0, 1, \ldots, m \\ \left|\begin{array}{l} u_i^{k+1} = \begin{cases} \operatorname{prox}_{\gamma_k g_{i_k}^*}\left(u_{i_k}^k + \gamma_{i_k} A_{i_k} \boldsymbol{x}^k\right) & \textit{if } i = i_k \\ u_i^k & \textit{if } i \neq i_k, \end{cases} \end{array}\right. \end{array}\right. \end{array} \tag{173}$$

where $(i_k)_{k\in\mathbb{N}}$ *are independent random variables taking values in* $\{1, \ldots, m\}$ *with* $\mathsf{p}_i := P(i_k = i) > 0$ *for all* $i \in \{1, \ldots, m\}$.

Remark 1 Note that in the setting of Algorithm 8, the primal problem can be written as

$$\min_{\boldsymbol{x}\in X} \sum_{i=1}^m g_i(A_i \boldsymbol{x}) + f(\boldsymbol{x}). \tag{174}$$

Now, suppose that f^* is quadratic, so that $\nabla f^* = H$ is a linear operator. Then, since $\boldsymbol{u}^{k+1}$ and $\boldsymbol{u}^k$ differ on the i_k component only, denoting by J_{i_k} the canonical injection of Y_{i_k} into Y, we have

$$\begin{aligned} \boldsymbol{x}^{k+1} &= -H A^* \boldsymbol{u}^{k+1} \\ &= -H A^* J_{i_k}(u_{i_k}^{k+1} - u_k^k) + \boldsymbol{x}^k. \end{aligned}$$

Thus, Algorithm 8 can be written as follows. Set $\boldsymbol{u}^0 = 0$, $\boldsymbol{x}^0 = 0$. Then

$$\begin{array}{l} \text{for } k = 0, 1 \dots \\ \quad \text{for } i = 0, 1, \dots, m \\ \quad\quad u_i^{k+1} = \begin{cases} \operatorname{prox}_{\gamma_k g_{i_k}^*}\left(u_{i_k}^k + \gamma_{i_k} A_{i_k} \boldsymbol{x}^k\right) & \text{if } i = i_k \\ u_i^k & \text{if } i \neq i_k, \end{cases} \\ \quad \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - H A_{i_k}^* (u_{i_k}^{k+1} - u_{i_k}^k). \end{array} \tag{175}$$

This shows that Algorithm 8 can be used as an incremental stochastic method for the minimization of (174), in which at each iteration one selects at random a single component in the sum (say i_k) and uses only the knowledge related to that component $(A_{i_k}, A_{i_k}^*, g_{i_k}^*, \gamma_{i_k})$ to make an update of the algorithm.

Example 96 (Linearly constrained problems) We consider the minimization problem

$$\min_{Ax=b} f(x),$$

where $f\colon X \to \,]-\infty, +\infty]$ is closed and strongly convex with constant $\mu > 0$. Then the dual problem is

$$\min_{u \in Y} f^*(-A^*u) + \langle u, b \rangle,$$

which is an *uncostrained and smooth* optimization problem. Thus, since $g^* = \langle \cdot, b \rangle$ and $\operatorname{prox}_{\gamma g^*}(u) = u - \gamma b$, Algorithm 6 becomes

$$\begin{array}{l} \text{for } k = 0, 1, \dots \\ \quad x_k = \nabla f^*(-A^* u_k) \\ \quad u_{k+1} = u_k + \gamma(A x_k - b), \end{array} \tag{176}$$

where $\gamma < 2\mu / \|A\|^2$.

5.3 Bibliographical Notes

Proposition 94 is standard, while Theorem 95 was essentially given (in a less explicit form) in [45]. Dual algorithms have been proposed several times in the literature. We mention among others the works [28, 37] for deterministic algorithms, while [107] for stochastic algorithms in the context of machine learning. The dual accelerated proximal gradient Algorithm 7 was presented in [15] with the standard choice of the parameters t_k's given by the first of (78). The gradient descent on the dual of the linearly constrained optimization problem described in Example 96 coincides, up to a change of variables, with the linearized Bregman method studied in a series of papers, see [86, 116] and references therein.

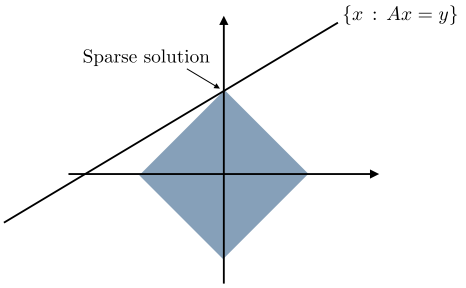


Fig. 3 Solution of problem (P_1) for $A\colon \mathbb{R}^2 \to \mathbb{R}$. Here A satisfies the NSP relative to $S = \{2\}$

6 Applications

In this section, we present three main applications where convex optimization plays a key role, providing fundamental tools and computational solutions.

6.1 Sparse Recovery

In many applications throughout science and engineering, one often needs to solve ill-posed inverse problems, where the number of available measurements is smaller than the dimension of the vector (signal) to be estimated. More formally, the setting is the following: given an observation $y \in \mathbb{R}^n$, and a linear measurement process $A\colon \mathbb{R}^d \to \mathbb{R}^n$ the goal is to

$$\text{find } x_* \in \mathbb{R}^d \text{ such that } \quad Ax_* = y, \tag{177}$$

under the assumption that $d >> n$. In general, more than one solution of the above problem exists, but reconstruction of x_* is often possible since in many practical situations of interest, the vectors of interest are *sparse*, namely they only have a few nonzero entries or few degrees of freedom compared to their dimension. In compress sensing it is shown that reconstruction of sparse vectors is not only feasible in theory, but efficient algorithms also exist to perform the reconstruction in practice. One of the most popular strategies is *basis pursuit* and consists in solving the following convex optimization problem

$$\min_{Ax=y} \|x\|_1. \tag{P_1}$$

In realistic situations, the measurements y will be always affected by noise, i.e.:

$$\|Ax_* - y\| \leq \delta$$

thus it makes more sense to consider the problem

$$\min_{\|Ax-y\|\leq\delta} \|x\|_1. \tag{$P_{1,\delta}$}$$

Then, the constrained problem ($P_{1,\delta}$) is usually transformed into a penalized problem, i.e (Fig. 3).

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-y\|^2 + \lambda\|x\|_1, \tag{178}$$

which is advantageous from the algorithmic point of view. It is possible to show that the problems ($P_{1,\delta}$) and (178) are equivalent, for suitable choices of the regularization parameter.

Proposition 97 *Let $A \in \mathbb{R}^{n\times d}$ and let $y \in \mathbb{R}^n$. Then the following hold:*

(i) *If x is a minimizer of (178) with $\lambda > 0$, then there exists $\delta = \delta(x) \geq 0$ such that x is a minimizer of ($P_{1,\delta}$).*
(ii) *If x is a minimizer of ($P_{1,\delta}$) with $\delta \geq 0$, then there exists $\lambda = \lambda(x) \geq 0$ such that x is a minimizer of (178).*

Proof Fermat's rule for problem (178) yields

$$0 \in A^*(Ax-y) + \lambda\partial\|\cdot\|_1(x),$$

that is,

$$(\forall\, i \in \{1,\dots,d\}) \quad (A^*(y-Ax))_i \in \lambda\partial|\cdot|(x_i) = \begin{cases} \lambda\,\mathrm{sign}(x_i) & \text{if } x_i \neq 0 \\ [-\lambda,\lambda] & \text{if } x_i = 0. \end{cases}$$

This shows that 0 is a minimizer of (178) if and only if $\|A^*y\|_\infty \leq \lambda$. Moreover, if $\|A^*y\|_\infty > \lambda$ and x is a minimizer of (178), then $x \neq 0$ and $\lambda = \|A^*(Ax-y)\|_\infty$ (so λ is uniquely determined by any minimizer).

Now, problem ($P_{1,\delta}$) can be equivalently written as

$$\min_{x\in X} \|x\|_1 + \iota_{B_\delta(y)}(Ax),$$

where $B_\delta(y)$ is the ball of radius δ centered at y. Moreover, 0 is a minimizer of ($P_{1,\delta}$) if and only if $\|y\| \leq \delta$. We therefore suppose that $\|y\| > \delta$, so that 0 is not a minimizer of ($P_{1,\delta}$). Then, the minimizers of ($P_{1,\delta}$) are different from zero and characterized by the following equation

$$0 \in \partial\|\cdot\|_1(x) + A^*\partial\iota_{B_\delta(y)}(Ax).$$

which is equivalent to

$$\exists u \in \partial\iota_{B_\delta(y)}(Ax) \quad \text{such that} \quad -A^*u \in \partial\|\cdot\|_1(x). \tag{179}$$

Recall that

$$\partial \iota_{B_\delta(y)}(Ax) = N_{B_\delta(y)}(Ax) = \begin{cases} \{0\} & \text{if } \|Ax - y\| < \delta \\ \mathbb{R}_+(Ax - y) & \text{if } \|Ax - y\| = \delta. \end{cases}$$

If $\|Ax - y\| < \delta$, then $u = 0$ and hence $0 \in \partial \|\cdot\|_1(x)$ which yields $x = 0$. Therefore since 0 is not a minimizer of $(P_{1,\delta})$, then necessarily $\|Ax - y\| = \delta$ and equation (179) is equivalent to

$$\|Ax - y\| = \delta \quad \text{and} \quad \exists \alpha > 0 \quad \text{such that} \quad \alpha A^*(y - Ax) \in \partial \|\cdot\|_1(x),$$

which yields

$$\exists \alpha > 0 \text{ s. t. } \forall\, i \in \{1, \dots, d\}\ (A^*(y - Ax))_i \in \frac{1}{\alpha} \partial |\cdot|(x_i) = \begin{cases} \alpha^{-1} \operatorname{sign}(x_i) & \text{if } x_i \neq 0 \\ [-\alpha^{-1}, \alpha^{-1}] & \text{if } x_i = 0. \end{cases}$$

Taking into account the above equations one can see that, if x is a minimizer of (178), then x is a minimizer of $(P_{1,\delta})$ with $\delta = \|Ax - y\|$ and, vice versa, if x is a minimizer of $(P_{1,\delta})$, then x is a minimizer of (178) with $\lambda = \|A^*(Ax - y)\|_\infty$. □

Remark 98 Analogous equivalence results relate $(P_{1,\delta})$ and (178) to another constrained problem:

$$\min_{\|x\|_1 \leq \tau} \|Ax - y\|_2, \quad \tau > 0.$$

6.1.1 Proximal Gradient Algorithms for Lasso

In this section, we specialized several proximal gradients algorithms we studied in the previous sections to the case of the lasso problem (178). As already anticipated in Example 17, (the proximal gradient) Algorithm 1 become the so called *Iterative Soft-Thresholding Algorithm* (ISTA), which is described below. Let $\gamma \in]0, 2/\|A^*A\|[$ and $x_0 = y_0 \in X$. Then,

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\quad \lfloor\, x_{k+1} = \operatorname{soft}_{\gamma\lambda}(x_k - \gamma A^*(Ax_k - y)), \end{aligned} \tag{180}$$

where $\operatorname{soft}_{\gamma\lambda} : \mathbb{R} \to \mathbb{R}$ is the so called *soft-thresholding operator*, which is the proximity operator of $\lambda|\cdot|$ (see (43)) and which is supposed to be applied component-wise. We stress that according to Example 60 and Theorem 62(iii), algorithm (180) provides a sequence that converges linearly to a solution of problem (178).

Now, according to Algorithm 2, its accelerated version is as follows. Let $x_0 = y_0 \in X$ and $\gamma \in \,]0, 1/\|A^*A\|\,]$. Then,

$$\begin{aligned} &\text{for } k = 0, 1, \dots \\ &\left\lfloor \begin{aligned} u_k &= x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}) \\ x_{k+1} &= \mathrm{soft}_{\gamma\lambda}(u_k - \gamma A^*(Au_k - y)). \end{aligned} \right. \end{aligned} \tag{181}$$

This algorithm is known as *Fast Iterative Soft-Thresholding Algorithm* (FISTA) and when the parameters t_k's are defined according to Proposition 64 with $1 - c \geq 2\sqrt{b}$, Theorem 68 yields that it converges in values with rate $O(1/k^2)$. Finally, we specialize the randomized proximal gradient Algorithm 5. We denote by a^i and a_k the i-th column and k-th row of A respectively. Since $\nabla_i[(1/2)\|Ax - b\|^2] = \langle a^i, Ax - b\rangle$, condition (ii) in Proposition 85 is satisfied with $L_i = \|a^i\|^2$. Then, Algorithm 5 (assuming that each block is made of one coordinate only) writes as

$$x^{k+1} = x^k + \left[\mathrm{soft}_{\gamma_i\lambda}\left(x_i^k - \gamma_i {a^i}^\top (Ax^k - b)\right) - x_i^k\right] e_{i_k}, \tag{182}$$

where $\gamma_i < 2/\|a^i\|^2$. Then, Theorem 93 ensures that $\mathsf{E}[F(x^k)] - \inf F = o(1/k)$ and that $(x_k)_{k\in\mathbb{N}}$ there exists a random vector x_* taking values in the solution set of problem (178) such that $x_k \to x_*$ almost surely.

6.2 Image Denoising

One of the most popular denoising models for imaging, is based on the *total variation regularizer*, and is known under the name "ROF" (Rudin, Osher and Fatemi). We consider a scalar-valued digital image $x \in \mathbb{R}^{m\times n}$ of size $m \times n$ pixels. A standard approach for defining the discrete total variation is to use a finite difference scheme acting on the pixels. The discrete gradient operator $D\colon \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \approx (\mathbb{R}^2)^{m\times n}$ is defined by

$$(Dx)_{i,j} = ((D_1x)_{i,j}, (D_2x)_{i,j}) \in \mathbb{R}^2,$$

where

$$(D_1x)_{i,j} = \begin{cases} x_{i+1,j} - x_{i,j} & \text{if } 1 \leq i \leq m-1 \\ 0 & i = m \end{cases}$$

$$(D_2x)_{i,j} = \begin{cases} x_{i,j+1} - x_{i,j} & \text{if } 1 \leq j \leq n-1 \\ 0 & j = n \end{cases}$$

The discrete ROF model is then defined by

$$\min_{x\in\mathbb{R}^{m\times n}} \lambda\|Dx\|_{2,1} + \|x - y\|_2^2, \tag{183}$$

where $y \in \mathbb{R}^{m\times n}$ is the given noisy image, and the discrete total variation is defined by

$$\|Dx\|_{2,1} = \sum_{i,j} \|(Dx)_{i,j}\|_2 = \sum_{i,j} \left((D_1 x)_{i,j}^2 + (D_2 x)_{i,j}^2\right)^{1/2},$$

that is, the ℓ_1-norm of the 2-norm of the pixelwise image gradients. We can interpret the total variation regularization from a sparsity point of view, establishing analogies with lasso approach in (178). Indeed, the ℓ_1-norm induces sparsity in the gradients of the image. More precisely, this regularizer can be interpreted as a group lasso one (see Example 42), where each group include the two directional derivatives at each pixel. Hence, this norm favors vectors with sparse gradients, namely piecewise constant images. This favorable property, a.k.a. staircasing effect has also some drawbacks in the applications, and other regularizations have been proposed. In the next section we describe an algorithm to solve (183).

6.2.1 Algorithms for Total Variation Denoising

Solving the discrete ROF (Rudin–Osher–Fatemi) model

$$\min_{x\in\mathbb{R}^{m\times n}} \lambda\|Dx\|_{2,1} + \frac{1}{2}\|x-y\|_2^2, \tag{184}$$

is equivalent to compute the proximity operator of the total variation, which is not available in closed form. Here we show how to solve the above problem by a dual algorithm. Indeed the problem is of the form ($\mathscr{P}$) considered by the Fenchel–Rockafellar duality theory with $f(x) = (1/2)\|x-y\|^2$,

$$g(\boldsymbol{v}) = \lambda\|\boldsymbol{v}\|_{2,1} = \sum_{i,j} \lambda\|\boldsymbol{v}_{i,j}\|_2, \quad \boldsymbol{v} = (\boldsymbol{v}_{i,j})_{\substack{1\le i\le m,\\ 1\le j\le n}},\ \boldsymbol{v}_{i,j} \in \mathbb{R}^2,$$

and $A = D$. We first compute $\|D\|$ since it will be useful later to set the steplength. For every $x \in \mathbb{R}^{m\times n}$

$$\begin{aligned}
\|Dx\|^2 &= \sum_{\substack{1\le i<m\\ 1\le j\le n}} (x_{i+1,j} - x_{i,j})^2 + \sum_{\substack{1\le i\le m\\ 1\le j<n}} (x_{i,j+1} - x_{i,j})^2 \\
&\le 2 \sum_{\substack{1\le i<m\\ 1\le j\le n}} ((x_{i+1,j})^2 + (x_{i,j})^2) + 2 \sum_{\substack{1\le i\le m\\ 1\le j<n}} ((x_{i,j+1})^2 + (x_{i,j})^2) \\
&\le 8\|x\|^2,
\end{aligned}$$

therefore $\|D\|^2 \le 8$. We next prove that the dual problem is

$$\min_{\boldsymbol{u}\in(\mathbb{R}^2)^{m\times n}} \frac{1}{2}\big(\|y - D^*\boldsymbol{u}\|^2 - \|y\|^2\big) + \iota_{B_\lambda(0)^{m\times n}}(\boldsymbol{u}), \quad B_\lambda(0) \subset \mathbb{R}^2, \tag{185}$$

where $B_\lambda(0)$ is the ball of $\mathbb{R}^2$ of radius λ centered at zero. Indeed, it is easy to check that $D^* = -\mathrm{div}\colon (\mathbb{R}^2)^{m\times n} \cong (\mathbb{R}^{m\times n})^2 \to \mathbb{R}^{m\times n}$ where, for every $(\boldsymbol{u}^1, \boldsymbol{u}^2) \in (\mathbb{R}^{m\times n})^2$

$$(\mathrm{div}(\boldsymbol{u}^1, \boldsymbol{u}^2))_{i,j} = \begin{cases} \boldsymbol{u}^1_{i,j} - \boldsymbol{u}^1_{i-1,j} & \text{if } 1 < i < m, \\ \boldsymbol{u}^1_{1,j} & \text{if } i = 1, \\ -\boldsymbol{u}^1_{m-1,j} & \text{if } i = m, \end{cases} + \begin{cases} \boldsymbol{u}^2_{i,j} - \boldsymbol{u}^2_{i,j-1} & \text{if } 1 < j < n, \\ \boldsymbol{u}^1_{i,1} & \text{if } j = 1, \\ -\boldsymbol{u}^1_{i,n-1} & \text{if } j = n, \end{cases}$$

and

$$f^*(z) = \frac{1}{2}(\|z + y\|_2^2 - \|y\|_2^2).$$

Moreover, $g(\boldsymbol{v}) = \sum_{i,j} \lambda\|\boldsymbol{v}_{i,j}\|_2 = \sum_{i,j} \sigma_{B_\lambda(0)}(\boldsymbol{v}_{i,j})$, which shows that g is separable. Then it follows from Fact 9(iii) that g^* is separable as well, so

$$g^*(\boldsymbol{v}) = \sum_{i,j} \iota_{B_\lambda(0)}(\boldsymbol{v}_{i,j}) = \iota_{B_\lambda(0)^{m\times n}}(\boldsymbol{v}).$$

Finally, since $\nabla f^*(z) = z + y$, the way one goes from the dual variable $\boldsymbol{u} \in (\mathbb{R}^2)^{m\times n}$ to the primal variable $x \in \mathbb{R}^{m\times n}$ is through the formula

$$x = \nabla f^*(-D^*\boldsymbol{u}) = y - D^*\boldsymbol{u}.$$

The dual proximal gradient algorithm (176) writes down as follows

$$\begin{array}{l} \text{for } k = 0, 1, \ldots \\ \left\lfloor \begin{array}{l} x^{(k)} = y - D^*\boldsymbol{u}^{(k)} \\ \boldsymbol{u}^{(k+1)} = P_{B_\lambda(0)^{n\times m}}(\boldsymbol{u}^{(k)} + \gamma D x^{(k)}), \end{array} \right. \end{array} \tag{186}$$

where $\gamma < 2/\|D\|^2 = 1/4$. Note also that the projection onto $B_\lambda(0)^{m\times n}$ is separable too and can be computed as

$$P_{B_\lambda(0)^{m\times n}}(\boldsymbol{u}) = \big(P_{B_\lambda(0)}(\boldsymbol{u}_{i,j})\big)_{\substack{1\le i\le m,\\ 1\le j\le n}}, \qquad P_{B_\lambda(0)}(\boldsymbol{u}_{i,j}) = \begin{cases} \boldsymbol{u}_{i,j} & \text{if } \|\boldsymbol{u}_{i,j}\|_2 \le \lambda \\ \dfrac{\boldsymbol{u}_{i,j}}{\|\boldsymbol{u}_{i,j}\|_2} & \text{if } \|\boldsymbol{u}_{i,j}\|_2 > \lambda. \end{cases}$$

Then it follows from the theory given in Sect. 5 that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to the minimizer of (184) as an $O(1/\sqrt{k})$.

We next specialize Algorithm 2 to problem (185). Let $\boldsymbol{u}_0 = \boldsymbol{v}_0 \in X$, $z_0 = y - D^*\boldsymbol{u}^{(0)}$, and $\gamma \in]0, 1/8[$. Define

$$\begin{aligned}&\text{for } k = 0, 1, \dots \\ &\left\lfloor \begin{aligned} x^{(k)} &= y - D^* \boldsymbol{u}^{(k)} \\ \boldsymbol{u}^{(k+1)} &= P_{B_\lambda(0)^{n\times m}}(\boldsymbol{v}^{(k)} + \gamma D z^{(k)}), \\ \boldsymbol{v}^{(k+1)} &= \boldsymbol{u}^{(k+1)} + \beta_{k+1}(\boldsymbol{u}^{(k+1)} - \boldsymbol{u}^{(k)}) \\ z^{(k+1)} &= x^{(k+1)} + \beta_{k+1}(x^{(k+1)} - x^{(k)}) \end{aligned} \right.\end{aligned} \tag{187}$$

With the choice of parameters as in Theorem 68, from the results in Sect. 5, we derive that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to the minimizer of (184) as an $O(1/k)$.

Finally, we specialize the randomized proximal gradient Algorithm 5. Note that, condition (ii) in Proposition 85 is satisfied with $L_{i,j} = \sqrt{17}$. Then, Algorithm 5 (assuming that each block is made of one $\mathbb{R}^2$ block only and (i_k, j_k) is uniformly distributed on $\{1, \dots, n\} \times \{1, \dots, m\}$) writes as

$$\begin{aligned}&\text{for } k = 0, 1, \dots \\ &\left\lfloor \begin{aligned} x^{(k)} &= x^{k-1} + D^*(\boldsymbol{u}^{k-1} - \boldsymbol{u}^k) \\ \boldsymbol{u}^{(k+1)} &= \boldsymbol{u}^{(k)} + J_{(i_k,j_k)}\big[P_{B_\lambda(0)}(\boldsymbol{u}^{(k)}_{i_k,j_k} + \gamma_{i_k,j_k}(Dx^{(k)})_{i_k,j_k}) - \boldsymbol{u}^{(k)}_{i_k,j_k}\big], \end{aligned} \right.\end{aligned} \tag{188}$$

where $\gamma_{i,j} < 2/\sqrt{17}$ and $J_{(i_k,j_k)} \colon \mathbb{R}^2 \to (\mathbb{R}^2)^{m\times n}$ is the canonical injection. Then, denoting by x_* the unique solution of (184), Theorem 93 and the results in Sect. 5 ensure that $\mathsf{E}[\|x^k - x_*\|^2] \le o(1/\sqrt{k})$.

6.3 Machine Learning

In statistical machine learning we are given two random variables ξ and η, with values in $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ respectively, with joint distribution μ. We let $\ell \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be a convex loss function and the goal is to find a function $h \colon \mathcal{X} \to \mathcal{Y}$ in a given hypothesis function space which minimizes the averaged risk $R(h) = \mathsf{E}[\ell(\xi, \eta, h(\eta))]$ without knowing the distribution μ but based on some sequence $(\xi_k, \eta_k)_{k\in\mathbb{N}}$ of independent copies of (ξ, η).

In this problem, concerning the hypothesis function space one option is that of considering reproducing kernel Hilbert spaces (RKHS). They indeed are defined through kernel functions and are flexible enough to model even infinite-dimensional function spaces. They are defined as follows. We let $\Lambda \colon \mathcal{X} \to H$ be a general map from the input space $\mathcal{X}$ to a separable Hilbert space H, endowed with a scalar product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Then the corresponding RKHS is defined as

$$\mathcal{H} = \{h \in \mathbb{R}^{\mathcal{X}} \mid \exists\, w \in H \text{ s.t. } h = \langle w, \Lambda(\cdot)\rangle\} \quad \|h\| = \inf\{\|w\| \mid h = \langle w, \Lambda(\cdot)\rangle\}. \tag{189}$$

In this context, the map Λ is called the *feature map* and the corresponding *kernel function* is defined as

$$K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \qquad K(x, x') = \big\langle \Lambda(x), \Lambda(x') \big\rangle. \tag{190}$$

In this way, the above statistical learning problem becomes

$$\min_{w\in H} R(w) = \mathsf{E}[\ell(\xi, \eta, \langle w, \Lambda(\xi)\rangle)] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, \langle w, \Lambda(x)\rangle) d\mu(x, y), \quad (191)$$

which is supposed to be solved via some sequence $(\xi_k, \eta_k)_{k\in\mathbb{N}}$ of independent copies of (ξ, η).

In order to approach problem (191) we consider two strategies. The first one consists in considering the problem as an instance of a stochastic optimization problem as described in Example 82. The second one is to consider a regularized empirical version of (191) based on the available sample. In the following, we describe these two approaches.

6.3.1 Statistical Learning as Stochastic Optimization

We make the following assumptions.

- SL_1 For every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(x, y, \cdot)\colon \mathbb{R} \to \mathbb{R}$ is positive, convex and Lipschitz continuous with constant $\alpha > 0$ and $\mathsf{E}[\ell(\xi, \eta, 0)] < +\infty$.
- SL_2 The feature map Λ is measurable and $\mathsf{E}[\|\Lambda(\xi)\|^2] < +\infty$.

We show that problem (191) is an instance of Example 82. Indeed, we let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and, for every $w \in H$ and $z = (x, y) \in \mathcal{Z}$, $\varphi(w, z) = \ell(z, \langle w, \Lambda(x)\rangle)$. Then,

$$(\forall\, w_1, w_2 \in H)(\forall\, z = (x, y) \in \mathcal{X} \times \mathcal{Y})$$
$$|\varphi(w_1, z) - \varphi(w_2, z)| \le \alpha |\langle w_1 - w_2, \Lambda(x)\rangle| \le \alpha \|\Lambda(x)\| \|w_1 - w_2\|.$$

Hence, conditions $(\mathrm{SO}_1) - (\mathrm{SO}_2)$ in Example 82 hold with $L(z) = \alpha\|\Lambda(x)\|$. Moreover,

$$(\forall\, z \in \mathcal{Z})(\forall\, w \in H) \quad \partial\varphi(w, z) = \partial\ell(z, \langle w, \Lambda(x)\rangle)\Lambda(x), \quad (192)$$

where $\partial\varphi(w, z) = \partial\varphi(\cdot, z)(w)$. Now, let, for every $(z, t) \in \mathcal{Z} \times \mathbb{R}$, $\tilde{\ell}'(z, t)$ be a subgradient of $\ell(z, \cdot)$ at t and define

$$\tilde{\nabla}\varphi\colon H \times \mathcal{Z} \to H\colon (w, z) \mapsto \tilde{\ell}'(z, \langle w, \Lambda(x)\rangle)\Lambda(x) \in \partial\varphi(w, z).$$

Therefore, assumptions $(\mathrm{SO}_3) - (\mathrm{SO}_4)$ in Example 82 are satisfied and

$$\mathsf{E}[\tilde{\nabla}\varphi(w, \zeta)] = \int_{\mathcal{Z}} \tilde{\ell}'(x, y, \langle w, \Lambda(x)\rangle)\Lambda(x) d\mu(x, y) \in \partial R(w).$$

Then algorithm (175) becomes

$$w_{k+1} = w_k - \gamma_k \tilde{\ell}'(\xi_k, \eta_k, \langle w_k, \Lambda(\xi_k)\rangle)\Lambda(\xi_k). \quad (193)$$

If we define $h_k(x) = \langle w_k, \Lambda(x)\rangle$ and the kernel $K(x, x') = \langle \Lambda(x), \Lambda(x')\rangle$, then it follows from (193) that

$$h_{k+1}(x) = h_k(x) - \gamma_k \tilde{\ell}'(\xi_k, \eta_k, h_k(\xi_k))K(x, \xi_k). \tag{194}$$

Moreover, set

$$\bar{w}_k = \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i w_i, \qquad h_k(x) = \langle \bar{w}_k, \Lambda(x)\rangle = \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i g_i(x).$$

Then, the risk of $\bar{h}_k$ is $R(\bar{w}_k)$ and according to Theorem 77 we have that $R(\bar{w}_k) \to \inf_H R$, and if $S_* := \operatorname{argmin}_H R \neq \varnothing$, $D \geq \operatorname{dist}(x_0, S_*)$, and $\gamma_k = \bar{\gamma}/\sqrt{k+1}$, we have

$$(\forall\, k \in \mathbb{N}) \quad \mathsf{E}[R(\bar{w}_k)] - \min_H R \leq \frac{D^2}{2\bar{\gamma}} \frac{1}{\sqrt{k+1}} + \bar{\gamma} B^2 \frac{\log(k+1)}{\sqrt{k+1}},$$

where $B^2 = \alpha^2 \mathsf{E}[\|\Lambda(\xi)\|^2]$. Moreover, for all $k \in \mathbb{N}$, if $(\gamma_i)_{0\leq i\leq k} \equiv D/(B\sqrt{k+1})$, then

$$\mathsf{E}[R(\bar{w}_k)] - \min_H R \leq \frac{BD}{\sqrt{k+1}}. \tag{195}$$

Note that algorithm (194) is fully practicable, since it depends only on the kernel function K and on the data (ξ_k, η_k). In the following, we provide a list of 1-Lipschitz continuous losses:

- the *hinge loss*: $\mathcal{Y} = \{-1, 1\}$ and $\ell(x, y, t) = \max\{0, 1 - yt\}$;
- the *logistic loss for classification*: $\mathcal{Y} = \{-1, 1\}$ and $\ell(x, y, t) = \log(1 + e^{-yt})$;
- *L^1-loss*: $\mathcal{Y} = \mathbb{R}$ and $\ell(x, y, t) = |y - t|$;
- *logistic loss for regression*: $\mathcal{Y} = \mathbb{R}$ and $\ell(x, y, t) = -\log \dfrac{4e^{y-t}}{(1 + e^{y-t})^2}$.
- *ε-insensitive loss*: $\mathcal{Y} = \mathbb{R}$ and $\ell(x, y, t) = \max\{0, |y - t| - \varepsilon\}$.

6.3.2 Regularized Empirical Risk Minimization

Regularized empirical risk estimation solves the following optimization problem

$$\min_{w\in H} \frac{\lambda}{n} \sum_{i=1}^{n} \ell(y_i, \langle w, \Lambda(x_i)\rangle) + \frac{1}{2}\|w\|^2 =: \Phi(w), \tag{196}$$

where $(x_i, y_i)_{1\leq i\leq n}$ are realizations of the random variables $(\xi_i, \eta_i)_{1\leq i\leq n}$ and we assume for simplicity that the loss function is $\ell\colon \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ (convex in the second variable), and $\lambda > 0$ is a regularization parameter. Essentially the goal here is to find a function $h = \langle w, \Lambda(\cdot)\rangle$ that best fits the data $(x_i, y_i)_{1\leq i\leq n}$ according the to given

loss ℓ. Depending on the choice of the loss function the techniques take different names. If ℓ is the square loss, that is, $\mathcal{Y} = \mathbb{R}$ and $\ell(s,t) = (s-t)^2$, one talks about *ridge regression*. If ℓ is the *Vapnik ε-insensitive loss*

$$\ell(s,t) = \max\{0, |s-t| - \varepsilon\},$$

then we have *support vector regression*. Finally, if ℓ is the *hinge loss*, that is $\mathcal{Y} = \{-1, 1\}$ and $\ell(s,t) = (1-st)_+$, then we get *support vector machines*. Another important loss for classification is the *logistic loss*, which is defined as $\ell(s,t) = \log(1+e^{-st})$.

We are going to compute the dual problem of (196) in the sense of Fenchel–Rockafellar (see Sect. 2.6). Define the operator

$$\Lambda(\mathsf{X})\colon H \to \mathbb{R}^n, \quad \Lambda(\mathsf{X})w = \begin{bmatrix} \langle w, \Lambda(x_1)\rangle \\ \cdots \\ \langle w, \Lambda(x_n)\rangle \end{bmatrix} \in \mathbb{R}^n$$

and the functions

$$g\colon \mathbb{R}^n \to \mathbb{R}, \quad g(z) = \frac{\lambda}{n}\sum_{i=1}^n \ell(y_i, -z_i), \quad \text{and} \quad f\colon H \to \mathbb{R}, \quad f(w) = \frac{1}{2}\|w\|^2. \tag{197}$$

Then problem (196) can be written as

$$\min_{w\in H} f(w) + g(-\Lambda(\mathsf{X})w), \tag{198}$$

which is in the form ($\mathscr{P}$) considered by the Fenchel–Rochafellar duality. We recall that the dual problem is

$$\min_{\alpha\in\mathbb{R}^n} f^*(\Lambda(\mathsf{X})^*\alpha) + g^*(\alpha) \tag{199}$$

and the corresponding KKT optimality conditions are (see Sect. 2.6)

$$\bar{w} \in \partial f^*(\Lambda(\mathsf{X})^*\bar{\alpha}) \qquad \text{and} \qquad \bar{\alpha} \in \partial g(-\Lambda(\mathsf{X})\bar{w}). \tag{200}$$

So, since $f^* = (1/2)\|\cdot\|^2$ and

$$(\forall\, \alpha \in \mathbb{R}^n) \qquad \Lambda(\mathsf{X})^*\alpha = \sum_{i=1}^n \alpha_i \Lambda(x_i),$$

the first term in the dual objective function (199) is

$$
\begin{aligned}
f^*(\Lambda(\mathsf{X})^*\alpha) &= \frac{1}{2}\|\Lambda(\mathsf{X})^*\alpha\|^2 \\
&= \frac{1}{2}\Big\|\sum_{i=1}^n \alpha_i \Lambda(x_i)\Big\|^2 \\
&= \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j \big\langle \Lambda(x_i), \Lambda(x_j)\big\rangle \\
&= \frac{1}{2}\alpha^\top \mathsf{K}\alpha,
\end{aligned}
$$

where $\mathsf{K} \in \mathbb{R}^{n\times n}$ is the Gram matrix, defined as $\mathsf{K} = (K(x_i, x_j))_{i,j=1}^n$ and K is the kernel function associated to the feature map Λ as defined in (190). Now we compute the form of g^*. According to (197), the function g is separable, that is, it can be written as $g(z) = \sum_{i=1}^n g_i(z_i)$, where $g_i = (\lambda/n)\ell(y_i, -\cdot)$. Therefore

$$
g^*(\alpha) = \sum_{i=1}^n g_i^*(\alpha_i).
$$

Moreover, recalling the properties of the Fenchel conjugation, we have

$$
g_i^*(s) = \frac{\lambda}{n}\ell^*\Big(y_i, -s\frac{n}{\lambda}\Big).
$$

Therefore we are lead to the following theorem

Theorem 99 *The dual problem of* (196) *is*

$$
\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\alpha^\top \mathsf{K}\alpha + \frac{\lambda}{n}\sum_{i=1}^n \ell^*\Big(y_i, -\alpha_i\frac{n}{\lambda}\Big) =: \Psi(\alpha). \tag{201}
$$

where $\mathsf{K} = (K(x_i, x_j))_{i,j=1}^n$ *and* K *is the kernel function associated to the feature map (see* (190)*),* $\ell^*(y_i, \cdot)$ *is the Fenchel conjugate of* $\ell(y_i, \cdot)$*. Moreover,* (i) *the primal problem* (196) *has a unique solution, the dual problem has solutions and* $\min\Phi = -\min\Psi$ *(strong duality holds); and* (ii) *the solutions* $(\bar w, \bar\alpha)$ *of the primal and dual problems are characterized by the following KKT conditions*

$$
\begin{cases}
\bar w = \Lambda(\mathsf{X})^*\bar\alpha = \displaystyle\sum_{i=1}^n \bar\alpha_i \Lambda(x_i), \\
\forall\, i \in \{1, \dots, n\} \quad -\dfrac{\bar\alpha_i n}{\lambda} \in \partial L(y_i, \langle \Lambda(\boldsymbol{x}_i), \bar w\rangle),
\end{cases} \tag{202}
$$

where $\partial\ell(y_i, \cdot)$ *is the subdifferential of* $\ell(y_i, \cdot)$*. Finally for the estimated function it holds*

$$\langle \bar{w}, \Lambda(\cdot) \rangle = \sum_{i=1}^{n} \bar{\alpha}_i K(x_i, \cdot).$$

Remark 100 The first equation in (202) says that the primal solution can be written as a finite linear combination of feature map evaluations on the training points. This is known as the *representer theorem* in the related literature. Moreover, the coefficients of this representation can be obtained through the solution of the dual problem (201).

We now specialize Theorem 99 to distance-based and margin-based losses.

Corollary 101 *Suppose that ℓ is a convex* distance-based loss*, that is, of the form $\ell(s,t) = \chi(s-t)$ with $\mathcal{Y} = \mathbb{R}$, for some convex function $\chi : \mathbb{R} \to \mathbb{R}_+$. Then the dual problem* (201) *becomes*

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^\top \mathsf{K} \alpha - y^\top \alpha + \frac{\lambda}{n} \sum_{i=1}^{n} \chi^* \Big(\frac{\alpha_i n}{\lambda} \Big). \tag{203}$$

Suppose that ℓ is a convex margin-based loss*, that is, of the form $\ell(s,t) = \chi(st)$ with $\mathcal{Y} = \{-1, 1\}$, for some convex function $\chi : \mathbb{R} \to \mathbb{R}_+$. Then the dual problem* (201) *becomes*

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^\top \mathsf{K} \alpha + \frac{\lambda}{n} \sum_{i=1}^{n} \chi^* \Big(- \frac{y_i \alpha_i n}{\lambda} \Big). \tag{204}$$

The following example shows that all the losses commonly used in machine learning admit explicit Fenchel conjugates.

Example 102 (i) The *least squares loss* is $\ell(s,t) = \chi(s-t)$ with $\chi = (1/2)|\cdot|^2$. In that case (203) reduces to

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^\top \mathsf{K} \alpha - y^\top \alpha + \frac{n}{2\lambda} \|\alpha\|^2.$$

which is strongly convex with modulus n/λ and has the explicit solution $\bar{\alpha} = (\mathsf{K} + (n/\lambda)\mathsf{Id})^{-1} y$.

(ii) The *Vapnik-ε-insensitive loss* for regression is $\ell(s,t) = \chi(s-t)$ with $\chi = |\cdot|_\varepsilon$. Then, $\chi^* = \varepsilon|\cdot| + \iota_{[-1,1]}$ and the dual problem (203) turns out to be

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^\top \mathsf{K} \alpha - y^\top \alpha + \varepsilon \|\alpha\|_1 + \iota_{\lambda/n[-1,1]^n}(\alpha)$$

(iii) The *Huber loss* is the distance-based loss defined by

$$\chi(r) = \begin{cases} r^2/2 & \text{if } |r| \le \rho \\ \rho|r| - \rho^2/2 & \text{otherwise.} \end{cases}$$

Then $\chi^* = \iota_{[-\rho,\rho]} + (1/2)|\cdot|^2$ and (203) becomes

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\alpha^\top \mathsf{K}\alpha - y^\top\alpha + \frac{n}{2\lambda}\|\alpha\|_2^2 + \iota_{\rho\lambda/n[-1,1]^n}(\alpha)$$

(iv) The *logistic loss* for classification is the margin-based loss with $\chi(r) = \log(1 + e^{-r})$. Thus

$$\chi^*(s) = \begin{cases} (1+s)\log(1+s) - s\log(-s) & \text{if } s \in]-1, 0[\\ 0 & \text{if } s = -1 \text{ or } s = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

It is easy to see that χ has Lipschitz continuous derivative with constant $1/4$ and hence χ^* is strongly convex with modulus 4. Thus, referring to (203) and (199), we see that in this case $\mathrm{dom} g^* = \prod_{i=1}^n (y_i[0, \lambda/n])$ and g^* is differentiable on $\mathrm{int}(\mathrm{dom} g^*)$ with locally Lipschitz continuous gradient. Moreover, since $\lim_{s\to 1}|(\chi^*)'(s)| = \lim_{s\to 0}|(\chi^*)'(s)| = +\infty$, we have that $\|\nabla g^*(\alpha)\| = +\infty$ on the boundary of $\mathrm{dom} g^*$. Finally, it follows from (202) that $0 < y_i\bar{\alpha}_i < \lambda/n$, for $i = 1, \ldots, n$.

(v) The *hinge loss* is the margin-based loss with $\chi(r) = (1 - r)_+$. We have $\chi^*(s) = s + \iota_{[-1,0]}(s)$. So the dual problem (204) is

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\alpha^\top \mathsf{K}\alpha - y^\top\alpha + \iota_{\lambda/n[0,1]^n}(y \odot \alpha).$$

where $y \odot \alpha = (y_i\alpha_i)_{1\le i\le n}$ is the Hadamard product of y and α.

The connection between the primal and dual problem is clarified by the following result, which follows from Proposition 94, Theorem 95 and (197).

Corollary 103 *Let $\bar{\alpha} \in \mathbb{R}^n$ be a solution of the dual problem (201) and let $\bar{w} = \Lambda(\mathsf{X})^\top\bar{\alpha}$ be the solution of the primal problem (196). Let $\alpha \in \mathbb{R}^n$ and set $w = \Lambda(\mathsf{X})^\top\alpha$. Then the following hold.*

(i) $\Psi(\alpha) - \min\Psi \ge \frac{1}{2}\|w - \bar{w}\|_2^2$.

(ii) *If $\ell(y_i, \cdot)$ Lipschitz smooth with constant a_1, then*

$$\Phi(w) - \inf\Phi \le \left(1 + \frac{\lambda a_1\|\Lambda(\mathsf{X})\|^2}{n}\right)(\Psi(\alpha) - \inf\Psi).$$

(iii) *if $\ell(y_i, \cdot)$ is Lipschitz continuous with constant a_2, then*

$$\Phi(w) - \inf\Phi \le 2\|\Lambda(\mathsf{X})\|\frac{a_2\lambda}{n}(\Psi(\alpha) - \inf\Psi)^{1/2}.$$

Remark 104 The above proposition ensures that if an algorithm generates a sequence $(\alpha^k)_{k\in\mathbb{N}}$ that is minimizing for the dual problem (201), i.e., $\Psi(\alpha^k) \to \min\Psi$, then the sequence defined by $w^k = \Lambda(\mathsf{X})^*\alpha^k$, $k \in \mathbb{N}$, converges to the solution of the primal problem. More precisely, for the function $\langle w_k, \Lambda(\cdot)\rangle$ we have

$$|\langle w_k, \Lambda(x)\rangle - \langle \bar{w}, \Lambda(x)\rangle| \leq \|w_k - \bar{w}\|\|\Lambda(x)\| \to 0 \tag{205}$$

and the function $\langle w_k, \Lambda(\cdot)\rangle$ can be expressed in terms of the kernel only, indeed $\langle w_k, \Lambda(x)\rangle = \sum_{i=1}^n \alpha_i^k \langle \Lambda(x_i), \Lambda(x)\rangle = \sum_{i=1}^n \alpha_i^k K(x_i, x)$.

Proximal gradient algorithms for SVM. For all the cases treated in Example 102, the dual problem (201) has the following form

$$\min_{\alpha\in\mathbb{R}^n} q(\alpha) + \sum_{i=1}^n h_i(\alpha_i) = \Psi(\alpha), \tag{206}$$

where $q\colon \mathbb{R}^n \to \mathbb{R}$ is convex and smooth with Lipschitz continuous gradient (locally Lipschitz for the logistic loss) and includes the quadratic term $(1/2)\alpha^\top \mathsf{K}\alpha$, and $h_i\colon \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ is proper, lower semicontinuous, convex, and admitting a closed-form proximity operator. So, the form (201) is amenable to proximal gradient type algorithms studied in the previous sections. We note that due to Corollary 103 if $\Psi(\alpha^k)_{k\in\mathbb{N}}$ converges linearly (resp. sublinearly) to $\inf\Psi$, then $(w^k)_{k\in\mathbb{N}}$ as well as $\Phi(w^k) - \inf\Phi$ converges linearly (resp. sublinearly) too. In particular recalling Example 59, we have that the dual proximal gradient Algorithm 6 converges linearly on the dual problem (206) for all the losses presented in Example 102 (except for the logistic one) and yields a linearly convergent sequence for the primal problem too. Similarly to the lasso problem, additional algorithmic solutions are obtained by applying on the dual problem the accelerated proximal gradient Algorithm 2 and the randomized block-coordinate proximal gradient Algorithm 5. In the case of the logistic loss considered in Example 102 (iv), proximal gradient algorithm with linesearch should be considered. See [102].

6.3.3 Structured Sparsity in Machine Learning

Sparse estimation methods are very popular in machine learning. The most natural one is the minimization of the empirical risk regularized with the ℓ^1 norm, in the very same way that we described in Sect. 6.1. In several applications of interest, it is beneficial to impose more structure in the regularization process and several extensions of the ℓ^1 regularization, such as group lasso or multitask learning, are common. It turns out that proximal gradient algorithms play a key role in the solution of the related variational problems, which we write using the notation introduced in the previous subsection

$$\min_{w\in\mathbb{R}^d} \frac{\lambda}{n}\sum_{i=1}^n \ell(y_i, \langle w, \Lambda(x_i)\rangle) + \Omega(w), \tag{207}$$

where the loss function is supposed to be differentiable with a Lipschitz continuous gradient (e.g., the square loss) and $\Omega\colon \mathbb{R}^d \to \mathbb{R}$ is a structured sparsity inducing

penalty. In this section, we briefly summarize some examples and the related proximal gradient algorithms.

When the input variables are supposed to be grouped together according to predefined groups forming a partition of the variables, the group lasso penalty discussed in Example 42 promotes solutions w_* depending only on few groups. The algorithms and the considerations made for the lasso problem in Sect. 6.1.1 can be generalized to the group LASSO, replacing the soft-thresholding operator with the proximal operator of the group lasso computed in Example 42. If the support of the solution is a union of potentially overlapping groups defined a priori then a different penalty should be used.

Let $\mathcal{J} = \{J_1, \dots, J_m\}$ be a family of subsets of $\{1, \dots, d\}$ whose union is $\{1, \dots, d\}$ itself. Let us call $v_{J_\ell} = (v_j)_{j \in J_\ell} \in \mathbb{R}^{J_\ell}$. Denote by $\|\cdot\|_{J_\ell}$ the Euclidean norm on $\mathbb{R}^{J_\ell}$ and by $J_{J_\ell} \colon \mathbb{R}^{J_\ell} \to \mathbb{R}^d$ the canonical embedding. We define a penalty on $\mathbb{R}^d$ by considering

$$\Omega(w) = \inf \left\{ \sum_{\ell=1}^m \|v_{J_\ell}\| \,\Big|\, \sum_{\ell=1}^m J_{J_\ell}(v_{J_\ell}) = w \right\}. \tag{208}$$

When the groups do not overlap, the above penalty coincides with the group lasso norm. If some groups overlap, then this penalty induces the selection of w_* sparsely supported on a union of groups. The regularized empirical risk in this case can be written in terms of the vectors v_{J_ℓ}:

$$\min_{(v_1, \dots, v_l) \in \mathbb{R}^{J_1} \times \dots \times \mathbb{R}^{J_m}} \frac{\lambda}{n} \sum_{i=1}^n \ell \left(y_i, \sum_{\ell=1}^m \langle v_{j_\ell}, J_{j_\ell}^* \Lambda(x_i) \rangle \right) + \sum_{\ell=1}^m \|v_{J_\ell}\|,$$

and the problem in these new variables coincide with a regularized group lasso without overlap.

Learning multiple tasks simultaneously has been shown to improve performance relative to learning each task independently, when the tasks are related in the sense that they all share a small set of features. For example, given T tasks modeled as $x \mapsto \langle w_t, \Lambda(x) \rangle$, for $t = 1, \dots, T$, multi-task learning amounts to the minimization of

$$\min_{(w_1, \dots, w_T) \in \mathbb{R}^{d \times T}} \sum_{t=1}^T \frac{\lambda}{n_t} \sum_{i=1}^n \ell(y_i, \langle w_t, \Lambda(x_i) \rangle) + \sum_{j=1}^d \Big(\sum_{t=1}^T w_{j,t}^2 \Big)^{1/2},$$

where n_t is the number of samples for each task. Note that the regularization is an instance of a group lasso norm of the vector $(w_1, \dots, w_T) \in \mathbb{R}^{d \times T}$, and the multitask problem can therefore be solved as described above.

6.4 Bibliographical Notes

Section 6.1 The connections between the lasso minimization problem and the problem of determining the sparsest solutions of linear systems is the topic of interest for the compressive sensing community. We refer to [49] for a mathematical introduction on this subject. The solution of the lasso problem motivated a huge amount of research at the interface between convex optimization, signal processing, inverse problems, and machine learning. The Iterative Soft thresholding algorithm has been proposed in [41] and around the same time the application of the proximal gradient algorithm to the lasso problem, but also to other signal processing problems was discussed in [39]. Strong convergence of the sequence of iterates generated by the proximal gradient algorithm for the objective function in (178) was proved in [41] and generalized in [36]. The FISTA algorithm was proposed by Beck and Teboulle in the seminal paper [12]. Block coordinate versions of the ISTA algorithm are considered e.g., in [78, 103].

Section 6.2 The ROF model has been introduced by Rudin, Osher and Fatemi in [101], and studied theoretically in [31]. The approach based on duality has been considered in [28, 30, 33]. The application of FISTA and a monotone modification to the dual problem has been considered in [13].

Section 6.3 Stochastic optimization approaches for machine learning are very popular, and in particular stochastic gradient descent [21], see the related discussion in Sect. 4.4. One of the most well known stochastic methods to solve SVM in the primal variables is PEGASOS [106].

Proximal methods have been immediately the methods of choice to deal with structured sparsity in machine learning. The literature on the topic is vast, see the surveys [8, 77] and references therein.

Support vector machines are due to Vapnik and have been introduced in [20, 40]. There, the case of the hinge loss for classification with a general kernel function (so to cover nonlinear classifiers) was treated. The dual problem was derived via the Lagrange theory. The analysis for general losses as well as the connection with reproducing kernel Hilbert spaces and the formulation via general feature maps is given, e.g., in [110].

Acknowledgements The work of S. Villa has been supported by the ITN-ETN project TraDE-OPT funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie grant agreement No 861137 and by the project "Processi evolutivi con memoria descrivibili tramite equazioni integro-differenziali" funded by Gruppo Nazionale per l' Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. Alvarez, F., Attouch, H.: An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. Set-Valued Anal. **9**, 3–11 (2001)
2. Atchadé, Y.F., Fort, G., Moulines, E.: On perturbed proximal gradient algorithms. J. Mach. Learn. Res. **18**, 1–33 (2017)
3. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Progr. **116**, 5–16 (2009)
4. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Ł ojasiewicz inequality, Math. Oper. Res. **35**, 438–457 (2010)
5. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Progr. **137**, 91–129 (2013)
6. H. Attouch, Z. Chbani, J. Peypouquet, P. Redont, Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Math. Prog. Ser. B **168**, 123–175 (2018)
7. Aujol, J.-F., Dossal, C., Rondepierre, A.: Optimal convergence rates for Nesterov Acceleration. SIAM J. Optim. **29**, 3131–3153 (2019)
8. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with Sparsity-Inducing Penalties. Optim. Mach. Learn. **5**, 19–53 (2011)
9. Baillon, J.B., Bruck, R.E., Reich, S.: On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces. Houston J. Math. **4**, 1–9 (1978)
10. Barbu, V., Precupanu, T.: Convexity and Optimization in Banach Spaces. Springer, Dordrecht (2012)
11. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, New York (2017)
12. Beck, A., Teboulle, M.: A fast iterative Shrinkage-Thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
13. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**, 2419–2434 (2009)
14. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**, 1205–1223 (2006)
15. Beck, A., Teboulle, M.: A fast dual proximal gradient algorithm for convex minimization and applications. Oper. Res. Lett. **42**, 1–6 (2014)
16. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**, 556–572 (2007)
17. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. **165**, 471–507 (2017)
18. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Prog. **146**, 459–494 (2013)
19. Borwein, J.M., Vanderwerff, J.D.: Convex Functions: Constructions, Characterizations and Counterexamples. Cambridge University Press, Cambridge (2010)
20. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory—COLT '92, p. 144 (1992)
21. Bottou, L., Bousquet, O.: The tradeoffs of large-scale learning. In: Optimization for Machine Learning, pp. 351–368, The MIT Press, Cambridge (2012)
22. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**, 223–311 (2018)
23. Bourbaki, N.: General Topology, 2nd edn. Springer, New York (1989)
24. Bredies, K.: A forward-backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space. Inv. Prob. **25**, Art. 015005 (2009)

25. Browder, F.E., Petryshyn, W.V.: The solution by iteration of nonlinear functional equations in Banach spaces. Bull. Am. Math. Soc. **72**, 571–575 (1966)
26. Browder, F.E., Petryshyn, W.V.: Construction of fixed points of nonlinear mappings in Hilbert space. J. Math. Anal. Appl. **20**, 197–228 (1967)
27. Burke, J.V., Ferris, M.C.: Weak sharp minima in mathematical programming. SIAM J. Control Optim. **31**, 1340–1359 (1993)
28. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**, 89–97 (2004)
29. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". J. Optim. Theory Appl. **166**, 968–982 (2015)
30. Chambolle, A., Lions, P.-L.: Image restoration by constrained total variation minimization and variants. In: Investigative and Trial Image Processing, San Diego, CA (SPIE), vol. 2567, pp. 50–59 (1995)
31. Chambolle, A., Lions, P.-L.: Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
32. Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numerica **25**, 161–319 (2016)
33. Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal-dual method for total variation-based image restoration. SIAM J. Sci. Comput. **20**, 1964–1977 (1999)
34. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing, In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York, NY (2011)
35. Combettes, P.L., Pesquet, J.-C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**, 1121–1248 (2015)
36. Combettes, P.L., Pesquet, J.-C.: Proximal thresholding algorithms for minimization over orthonormal bases. SIAM J. Optim. **18**, 1351–1376 (2007)
37. Combettes, P.L., V u, B.C.: Dualization of signal recovery problems. Set-Valued Anal. **18**, 373–404 (2010)
38. Combettes, P.L., Yamada, I.: Compositions and convex combinations of averaged nonexpansive operators. J. Math. Anal. Appl. **425**, 55–70 (2015)
39. Combettes, P.L., Wajs, V.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**, 1168–1200 (2005)
40. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. **20**, 273–297 (1995)
41. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. **57**, 1413–1457 (2004)
42. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: *Advances in Neural Information Processing Systems*, vol. 27 (2014)
43. Dotson, W.G.: On the Mann iterative process. Trans. Am. Math. Soc. **149**, 65–73 (1970)
44. Duchi, J., Singer, Y.: Efficient online and batch learning using forward backward splitting. J. Mach. Learn. Res. **10**, 2899–2934 (2009)
45. Dünner, C., Forte, S., Takac, M., Jaggi, M.: Primal-dual rates and certificates. In: *Proceedings of The 33rd International Conference on Machine Learning, PMLR*, vol. 48, pp. 783–792 (2016)
46. Ekeland, I., Témam, R.: Roger. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, Convex analysis and variational problems (1999)
47. Ermoliev, Yu.M.: On the method of generalized stochastic gradients and quasi-Fejér sequences. Cybernetics **5**, 208–220 (1969)
48. Fenchel, W.: Convex Cones, Sets, and Functions. Princeton University (1953)
49. Foucart, S., Rauhut, H.: A mathematical introduction to compressive sensing. Birkäuser. Springer, New York (2010)
50. Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. **165**, 874–900 (2015)

51. Gabay, D.: Applications of the method of multipliers to variational inequalities. In: Fortin, M., Glowinski, R. (eds.) Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, North-Holland, Amsterdam, vol. 15, pp. 299–331 (1983)
52. Garrigos, G., Rosasco, L., Villa, S.: Convergence of the Forward-Backward Algorithm: Beyond the Worst Case with the Help of Geometry (2017). https://arxiv.org/abs/1703.09477
53. Goldstein, A.A.: Convex programming in Hilbert space. Bull. Am. Math. Soc. **70**, 709–710 (1964)
54. Groetsch, C.W.: A note on segmenting Mann iterates. J. Math. Anal. Appl. **40**, 369–372 (1972)
55. Guler, O.: New proximal point algorithms for convex minimization. SIAM J. Optim. **2**, 649–664 (1992)
56. Blatt, D., Hero, A., Gauchman, H.: A convergent incremental gradient method with a constant step size. SIAM J. Optim. **18**, 29–51 (2007)
57. Hiriart-Urruty, J.-B., Lemaréchal, C.: Fundamentals of Convex Analysis. Springer, Berlin (2001)
58. Jensen, J.L.W.V.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math. **30**, 175–193 (1906)
59. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst. **26**, 315–323 (2013)
60. Karimi, H., Nutini, J., Schmidt, M.: Linear Convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016. Lecture Notes in Computer Science, vol. 9851. Springer, Cham
61. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. **23**, 462–466 (1952)
62. Kingma, D.P., Ba, L.J.: Adam: a method for stochastic optimization. In: Proceedings of Conference on Learning Representations (ICLR), San Diego (2015)
63. Krasnoselski, M.A.: Two remarks on the method of successive approximations. Uspekhi Mat. Nauk. **10**, 123–127 (1955)
64. Levitin, E.S., Polyak, B.T.: Constrained minimization methods. U.S.S.R. Comput. Math. Math. Phys. **6**, 1–50 (1966)
65. Li, W.: Error bounds for piecewise convex quadratic programs and applications. SIAM J. Control Optim **33**, 1510–1529 (1995)
66. Li, G.: Global error bounds for piecewise convex polynomials. Math. Prog. Ser. A **137**, 37–64 (2013)
67. Lions, P.L., Mercier, I.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
68. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**, 157–178 (1993)
69. Luque, F.: Asymptotic convergence analysis of the proximal point algorithm. SIAM J. Control Optim. **22**, 277–293 (1984)
70. Mann, W.R.: Mean value methods in iteration. Proc. Am. Math. Soc. **4**, 506–510 (1953)
71. Martinet, B.: Régularisation d'in Opér. 4, Sér. R-3, pp. 154–158 (1970)
72. Mercier, B.: Inéquations Variationnelles de la Mécanique. No. 80.01 in Publications Mathématiques d'Orsay. Université de Paris-XI, Orsay, France (1980)
73. Minkowski, H.: Theorie der konvexen Körper, insbesondere Begründung ihres Oberflächenbegriffs. In: Hilbert, D. (ed.) Gesammelte abhandlungen von Hermann Minkowski [Collected Papers of Hermann Minkowski], vol. 2, pp. 131–229. B.G. Teubner, Leipzig (1911)
74. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien, C. R. Acad. Sci. Paris Ser. A Math. **255**, 2897–2899 (1962)
75. Moreau, J.J.: Propriétés des applications "prox", C. R. Acad. Sci. Paris Ser. A Math. **256**, 1069–1071 (1963)
76. Moreau, J.J.: Proximité et dualité dans un espace Hilbertien. Bull. de la Société Mathématique de France **93**, 273–299 (1965)

77. Mosci, S., Rosasco, L., Santoro, M., Verri, A., Villa, S.: Solving structured sparsity regularization with proximal methods. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 418–433. Springer, Berlin, Heidelberg (2010)
78. Necoara, I., Clipici, D.: Parallel random coordinate descent method for composite minimization: convergence analysis and error bounds. SIAM J. Optim. **26**, 197–226 (2016)
79. Necoara, I., Nesterov, Y., Glineur, F.: Random block coordinate descent methods for linearly constrained optimization over networks. J. Optim. Theory Appl. **173**, 227–254 (2017)
80. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**, 1574–1609 (2009)
81. Nemirovsij, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience, New York (1983)
82. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, London (2004)
83. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
84. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**, 341–362 (2012)
85. Opial, Z.: Weak convergence of the sequence of successive approximations for nonexpansive mappings. Bull. Am. Math. Soc. **73**, 591–597 (1967)
86. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation- based image restoration. Multiscale Model. Sim. **4**, 460–489 (2005)
87. Passty, G.B.: Ergodic convergence of a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–390 (1979)
88. Peypouquet, J.: Convex Optimization in Normed Spaces. Springer, Cham (2015)
89. Phelps, R.R.: Convex Functions, Monotone Operators and Differentiability. Springer, Berlin (1993)
90. Polyak, B.T.: Dokl. Akad. Nauk SSSR **174**
91. Polyak, B.T.: Gradient methods for minimizing functionals. Zh. Vychisl. Mat. Mat. Fiz. **3**, 643–653 (1963)
92. Polyak, B.T.: Subgradient methods: a survey of Soviet research. In: Lemaréchal, C.L., Mifflin, R. (eds.) Proceedings of a IIASA Workshop, Nonsmooth Optimization, pp. 5–28. Pergamon Press, New York (1977)
93. Polyak, B.T.: Introduction to Optimization. Optimization Software, Inc. (1987)
94. Richtàrik, P., Takàč, M.: Parallel coordinate descent methods for big data optimization. Math. Program. Ser. A **156**, 56–484 (2016)
95. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
96. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Optimizing Methods in Statistics, pp. 233–257. Academic Press (1971)
97. Rockafellar, T.: Monotone operators and the proximal point algorithm. SIAM J. Optim. **14**, 877–898 (1976)
98. Rockafellar, T.: Convex Analysis. Princeton University Press, Princeton (1970)
99. Rockafellar, T.: Conjugate duality and optimization. Society for Industrial and Applied Mathematics, Philadelphia (1974)
100. Rosasco, L., Villa, S., Vũ, B.C.: Convergence of stochastic proximal gradient method. Appl. Math. Optim. **82**, 891–917 (2020)
101. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
102. Salzo, S.: The variable metric forward-backward splitting algorithm under mild differentiability assumptions. SIAM J. Optim. **27**(4), 2153–2181 (2017)
103. Salzo, S. Villa, S.: Parallel random block-coordinate forward-backward algorithm: a unified convergence analysis. Math. Program. Ser. A. https://doi.org/10.1007s10107-020-01602-1

104. Schaefer, H.: Über die Methode sukzessiver Approximationen. Jber. Deutsch. Math.-Verein. **59**, 131–140 (1957)
105. Shamir, O., Zhang, T.: Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. In: Proceedings of the 30th International Conference on Machine Learning, pp. 71–79 (2013)
106. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: primal estimated sub-gradient solver for SVM. Math. Program. **127**, 3–30 (2011)
107. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. J. Mach. Learn. Res. **14**, 567–599 (2013)
108. Shor, N.: Minimization Methods for Non-differentiable Functions. Springer, New York (1985)
109. Sibony, M.: Méthodes itéraratives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone. Calcolo **7**, 65–183 (1970)
110. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
111. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**, 1–43 (2016)
112. Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim. **29**, 119–138 (1991)
113. Wolfe, P.: A method of conjugate subgradients for minimizing nondifferentiable functions. Nondifferentiable optimization. Math. Program. Stud. **3**, 145–173 (1975)
114. Wright, S.: Coordinate descent algorithms. Math. Program. **151**, 3–34 (2015)
115. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific Publishing Co. Inc, River Edge, NJ (2002)
116. Zhang, X., Burger, M., Bresson, X., Osher, S.: Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. SIAM J. Imaging Sci. **3**, 253–276 (2010)

Regularization: From Inverse Problems to Large-Scale Machine Learning

Ernesto De Vito, Lorenzo Rosasco, and Alessandro Rudi

1 Introduction

Inverse problem theory provides a general and elegant framework to understand and model a variety of estimation/recovery problems. Machine learning is one such problem. Indeed, connections between inverse problems and learning have been known for a while and made mathematically precise. An inverse problem perspective to learning brings to light the importance of the notion of stability both from a statistical and a numerical point of view. This latter perspective turns out to be of particular importance when dealing with large-scale problems.

This chapter provides a brief introduction to machine learning from an inverse problem perspective with an emphasis on large-scale problems. After recalling an inverse problem perspective on supervised learning in Hilbert spaces, we discuss regularization methods for large-scale machine learning. In particular, we derive and contrast different regularization schemes. Starting from classic Tikhonov regularization, we then introduce iterative regularization, the idea of early stopping, and discuss different variants including accelerated and stochastic versions. Finally, we discuss projection with regularization and introduce stochastic extensions. Our discussion shows how the different methods are grounded in common estimation principles, but their computational properties are different. Iterative regularization allows to

E. De Vito (✉) · L. Rosasco
DIMA & MaLGA, Universita di Genova, Via Dodecaneso 35, Genova, Italy
e-mail: ernesto.devito@unige.it

L. Rosasco
e-mail: lorenzo.rosasco@unige.it

A. Rudi
Inria and Ecole Normale Supérieure, PSL Research University, Paris, France
e-mail: alessandro.rudi@inria.fr

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8_5

combine statistical and time complexities, while regularization with stochastic projections allows to simultaneously control statistical, time, and space complexity.

We refer to the appendix for the notation.

2 Learning as an Inverse Problem

In this section, we revisit supervised learning from an inverse problems perspective. This allows to later draw results and algorithms from classical regularization theory in inverse problems and to adapt them to supervised learning. Our presentation is based on [19].

2.1 Inverse Problems

Here we provide a short review on linear inverse problems, see, for example, [22] for a full exposition. We also refer to [29] for technical facts on functional analysis and operator theory.

Consider a linear continuous operator $A : \mathcal{H} \to \mathcal{G}$ between two Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$. Given a datum $g \in \mathcal{G}$, the problem of finding the solution $f \in \mathcal{H}$ of the equation

$$Af = g \tag{1}$$

is called the inverse problem associated with Eq. (1). The goal is to recover an unknown $f \in \mathcal{H}$ from the knowledge of A and g. The inverse problem is called ill-posed if at least one of the following conditions occurs:

- the solution does not exist, i.e., $g \notin \operatorname{ran} A$;
- the solution, if it exists, is not unique, i.e., $\ker A \neq \emptyset$;
- if the solution is unique, it does not depend continuously on the datum g.

Here $\operatorname{ran} A$ and $\ker A$ denote the range and the kernel of A. The last property is also referred to as the stability property, see later.

The question is how to find well-posed approximate solutions to the above problem. The first step is to replace Problem (1) with the least-squares problem

$$\inf_{f \in \mathcal{H}} \| Af - g \|_{\mathcal{G}}^2,$$

which admits a solution provided that $P_{\overline{\operatorname{ran} A}}\, g \in \operatorname{ran} A$, where $P_{\overline{\operatorname{ran} A}}$ denotes the projection onto the closure in $\mathcal{G}$ of the subspace $\operatorname{ran} A$. Under this condition there exists a canonical solution defined by

$$f^\dagger = \operatorname*{argmin}_{f \in \mathcal{H}_0} \| f \|_{\mathcal{H}},$$

where

$$\mathcal{H}_0 = \operatorname*{argmin}_{f \in \mathcal{H}} \|Af - g\|_{\mathcal{G}}^2$$

is a closed convex subset of $\mathcal{H}$, so that $f^\dagger$ always exists and is unique.

The vector $f^\dagger$ is called Moore–Penrose solution (or pseudo-solution) solution. The densely defined Moore–Penrose inverse operator from $\mathcal{G}$ to $\mathcal{H}$ is defined as

$$A^\dagger g = f^\dagger$$

whose dense domain is the set of $g \in \mathcal{G}$ such that $P_{\overline{\text{ran } A}} g \in \text{ran } A$, *i.e.,*

$$\text{dom}(A^\dagger) = \text{ran } A \oplus \ker A^*,$$

where A^* denotes the adjoint.

Typically, the subspace ran A is not closed so that closed graph theorem implies that $A^\dagger$ is not continuous. In this case we say that $f^\dagger$ is not stable with respect to the datum g. This question is particularly important since in practice the data might be affected by noise. A common way to formalize this idea is to replace the true datum g in Problem (1) with a noisy version g_δ such that

$$\|g - g_\delta\|_{\mathcal{G}} \leq \delta,$$

where $\delta > 0$ is seen as a noise level. Note that, though it is very reasonable to assume that $g \in \text{ran } A \subseteq \text{dom}(A^\dagger)$, in general, $g_\delta \notin \text{dom}(A^\dagger)$ and, even if it happens, $\|A^\dagger g - A^\dagger g_\delta\|$ can be very large since $A^\dagger$ is unbounded.

Regularization theory provides a general framework to derive stable solutions. Broadly speaking, regularization refers to a procedure to derive a sequence of solutions that converge to $f^\dagger$ and is stable to noise. A classic example is Tikhonov regularization given by

$$f_\delta^\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \left(\|Af - g_\delta\|_{\mathcal{G}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right),$$

where $\lambda > 0$ is a regularization parameter. Classical results in regularization theory [22] show that if $\lambda = \lambda_\delta$ is chosen as function of the noise level δ in such a way that

$$\lim_{\delta \to 0} \lambda_\delta = 0, \qquad \lim_{\delta \to 0} \frac{\delta}{\lambda_\delta} = 0,$$

then

$$\lim_{\delta \to 0} \|f_\delta^{\lambda_\delta} - f^\dagger\|_{\mathcal{H}} = 0.$$

Rates of convergence and error bounds can also be derived under suitable conditions.

2.2 Statistical Learning Theory

In this section, we briefly introduce the basic concepts in statistical supervised learning with least squares. Among a variety of references we mention [15, 16, 20, 24, 45, 48]. We refer to [21] for technical facts on probability and measure theory.

Supervised learning is concerned with the problem of learning a function from random samples of input–output pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We assume that the input space $\mathcal{X}$ is a Polish space, for example, $\mathbb{R}^d$, and the output space is $\mathcal{Y} = \mathbb{R}$. The product space $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ is endowed with a probability measure ρ defined on the Borel σ-algebra of $\mathcal{X} \times \mathbb{R}$.

The probability distribution ρ is known only through a training set

$$\mathbf{z}_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{Z}^n$$

of pairs, sampled independently and identically according to ρ. Given $\mathbf{z}_n$, the goal of supervised learning is to find an estimate $f_n : \mathcal{X} \to \mathbb{R}$ such that, given a new unlabeled input $x \in \mathcal{X}$, the value $f_n(x)$ is a good approximation of the true label y.

To make this precise, the error of any measurable function $f : \mathcal{X} \to \mathbb{R}$ is measured introducing the expected (square) loss

$$\mathcal{L}(f) = \int_{\mathcal{X}\times\mathbb{R}} (f(x) - y)^2 \, d\rho(x, y), \tag{2}$$

so that pairs (x, y) that are more likely to be sampled have more influence on the error. We will see that other error measures are also possible.

The expected loss can be written in a different way under the assumption that

$$\int_{\mathcal{X}\times\mathbb{R}} y^2 \, d\rho(x, y) < +\infty. \tag{3}$$

We recall the following integral decomposition. Given a measurable function $h : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, then

$$\int_{\mathcal{X}\times\mathbb{R}} h(x, y) d\rho(x, y) = \int_{\mathcal{X}} \left(\int_{\mathbb{R}} h(x, y) d\rho_{\mathcal{X}}(x) \right) d\rho(y|x), \tag{4}$$

where $\rho_{\mathcal{X}}$ is called the marginal measure on $\mathcal{X}$ and $\rho(\cdot\,|x)$ is the conditional probability measure on $\mathbb{R}$ given $x \in \mathcal{X}$.

Remark 1 As for the classical Fubini's theorem, if h is a positive measurable function such that the right-hand side of (4) is finite, then h is integrable and (4) holds true. Indeed, for absolutely continuous distributions, the above property follows from Fubini's theorem. For the general case, an ad hoc analysis is needed.

It is then easy to see that Eq. (4) implies

$$\mathcal{L}(f) = \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x) + L(f_\rho) = \|f - f_\rho\|_\rho^2 + \mathcal{L}(f_\rho), \tag{5}$$

where f_ρ is the *regression function*, defined for $\rho_{\mathcal{X}}$-almost all $x \in \mathcal{X}$ as

$$f_\rho(x) = \int_{\mathbb{R}} y \, d\rho(y|x),$$

and $\|\cdot\|_\rho$ is the norm of the Hilbert space

$$L^2(\mathcal{X}, \rho_{\mathcal{X}}) = \left\{ f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\rho^2 = \int_{\mathcal{X}} |f(x)|^2 d\rho_{\mathcal{X}}(x) < \infty \right\}.$$

Equation (5) makes clear that the function of interest is f_ρ and the goal is to find for any training set $\mathbf{z}_n$ an estimate f_n such that its excess expected risk

$$\|f_n - f_\rho\|_\rho^2 = \mathcal{L}(f) - \mathcal{L}(f_\rho)$$

is small with high probability. Indeed, the above quantity is stochastic through its dependence to the dataset $\mathbf{z}_n$. More precisely, in statistical learning theory the focus is on studying the convergence as well as explicit bounds on the probability

$$\rho^n \left\{ \mathbf{z}_n \in Z^n \mid \|f_n - f_\rho\|_\rho^2 \geq \epsilon \right\},$$

for all $\epsilon > 0$.

Remark 2 The quantities depending on the training set $\mathbf{z}_n$ are called empirical quantities and are denoted by the subscript $_n$, instead of $\mathbf{z}_n$ for the easy of notation.

Remark 3 We note that in statistical learning the regression function f_ρ is viewed as the solution of an optimization problem

$$\min_{f \in L^2(\mathcal{X}, \rho_{\mathcal{X}})} \mathcal{L}(f),$$

where the functional $\mathcal{L}$ is well defined, convex, and continuous.

We next discuss how the above problem can be reformulated as a linear inverse problem. We first discuss two basic examples of the above framework.

Example 4 (*Regression*) For all $i = 1, \ldots, n, n \in \mathbb{N}$, let x_i be a sequence of random points in $\mathcal{X}$ sampled according to a fixed probability distribution $\rho_{\mathcal{X}}$, and ϵ_i a sequence of random numbers with zero mean, bounded variance, and possibly dependent on x_i. Given a bounded function $f_* : \mathcal{X} \to \mathbb{R}$, assume

$$y_i = f_*(x_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{6}$$

In other words, data are samples of a function corrupted with noise and evaluated at random locations, while the learning problem is to recover f_* from the training set. The above is the classical model for regression. It is a special case of the general framework in this section where $f_\rho = f_*$ and the conditional distribution is defined by the noise distribution. We observe that a natural approach to the above regression problem is to solve the interpolation problem of finding f such that

$$y_i = f(x_i) \quad i = 1, \dots, n. \tag{7}$$

Example 5 (*Binary Classification*) Consider the case where the conditional distribution $\rho(y|x)$ is supported on $\{-1, 1\}$, that is, it corresponds to the pair of point masses $\rho(1|x)$, $\rho(-1|x)$ for almost all $x \in \mathcal{X}$. In this case, the natural error measure is the misclassification risk

$$R(f) = \rho\{(x, y) \in \mathcal{Z} \mid f(x)y < 0\},$$

that is, the expected number of misclassifications. It is a classic result that the misclassification risk is minimized by the so-called Bayes decision rule $b_\rho = \operatorname{sign}(f_\rho)$ and moreover

$$R(f) - R(b_\rho) \leq \left\| f - f_\rho \right\|_\rho.$$

This latter observation can be seen as justification for using least squares for classification problems.

2.3 Learning as an Inverse Problem

In this section, we show that if we restrict the search for learning solutions to a suitable Hilbert space (see below), then supervised learning can be reformulated as a linear inverse problem under a data model different from the classic ones. Connections between different estimation/statistical problems and inverse problems are classical. The idea that machine learning algorithms can often be seen as regularization and learning interpreted as an ill-posed problem has been discussed in [37], but also in [49]. A mathematical treatment close to the one presented in this chapter is in [5], albeit restricted to function approximation from a fixed, finite set of input points.

The treatment of learning from an inverse problem described here is introduced in [19] and further elaborated in [18]. These latter papers stem from a line of work moving its steps from the analytical perspective on supervised learning pushed forward in [15] and developed in [17]. It is relevant the study in [43], which considers the connection between learning and classical Shannon sampling theorem. We refer to [8, 31] for recent developments in this line of works.

We begin our discussion considering a preliminary step where only the training set is considered.

2.4 Linear Inverse Problem Associated to Finite Data

As a starter, we note that it is well known that the interpolation problem defined by Eq. (7) can be formulated as a discrete linear inverse problem [5].

We first observe that, in order to make Problem (7) meaningful, we have to fix a suitable space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ where we are looking for the solution of Eq. (7). The space $\mathcal{H}$ is usually called *hypotheses space* and can be seen as an a priori assumption on the function f^* generating the data. The key assumption we make is the following.

Assumption 6 The hypotheses space $\mathcal{H}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ and for all $x \in \mathcal{X}$, the evaluation functionals

$$\mathcal{H} \ni f \mapsto f(x) \in \mathbb{R},$$

are continuous.

The above requirement ensures that (7) is well defined and, as we show later, also stable under small perturbation of f. More importantly, the above assumption allows to view supervised learning as a linear inverse problem as we discuss next. Indeed, a direct consequence of this assumption is that the Riesz representation theorem ensures that for all $x \in \mathcal{X}$ there exists a function $K_x \in \mathcal{H}$ such that the following reproducing formula holds true:

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}} \qquad f \in \mathcal{H}, \tag{8}$$

function evaluation is given by a linear functional. This is the key property that effectively allows to formulate (7) as a *linear* inverse problem. As we discuss later, this condition corresponds to considering hypothesis spaces that are so-called *reproducing kernel Hilbert spaces*. To make the exposition simple, in the rest of the section, we also assume that for some constant $\kappa > 0$,

$$\|K_x\|_{\mathcal{H}} \leq \kappa \quad \rho_{\mathcal{X}} - \text{almost surely.} \tag{9}$$

Given the above observation we have that, any training set $\mathbf{z}_n$ defines a *sampling operator*

$$S_n : \mathcal{H} \to \mathbb{R}^n \qquad (S_n f)^i = \langle f, K_{x_i} \rangle_{\mathcal{H}}, \quad i = 1, \ldots, n,$$

and Problem (7) can be formulated as the linear inverse problem corresponding to finding $f \in \mathcal{H}$ such that

$$S_n f = \mathbf{y}, \tag{10}$$

where $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$.

The above problem is a promising start, but essentially corresponds to a "*n*oisy" inverse problem, in the sense that we have only an empirical problem based on the

data. While this is the basis for practical algorithms, it is not clear how it relates to the problem of estimating the regression function f_ρ, which is the target of learning. The question is then, if the problem of estimating the regression function can itself be formulated as an ideal "*noiseless*" linear inverse problem, of which Problem (10) is a "*noisy*" empirical instantiation. Indeed, this is the case as we discuss next.

Linear Inverse Problem Associated to Infinite Data

Roughly speaking, the answer follows identifying the ideal/noiseless problem with the *infinite data* limit of Problem (10), sometimes called population setting. Indeed, this corresponds to considering the operator

$$S_\rho : \mathcal{H} \to L^2(\mathcal{X}, \rho_\mathcal{X}), \qquad (S_\rho f)(x) = \langle f, K_x \rangle_\mathcal{H}, \quad \rho_\mathcal{X} - \text{almost surely},$$

that by (9) is well defined and bounded, and considering the associated linear inverse problem defined by

$$S_\rho f = f_\rho. \tag{11}$$

The above inverse problem can be seen as the one corresponding to estimating the regression function. We add three remarks to illustrate the above idea.

Remark 7 (*Risk and Moore–Penrose Solution*) The inverse problem associated to (11) corresponds to looking for a function in $\mathcal{H}$ providing a good approximation of the regression function. This problem is typically ill-posed, in particular note that generally the regression function does not belong to $\mathcal{H}$. The associated least squares problem is

$$\inf_{f \in \mathcal{H}} \left\| S_\rho f - f_\rho \right\|_\rho^2, \tag{12}$$

which, in light of Remark 3, corresponds to considering,

$$\inf_{f \in \mathcal{H}} \mathcal{L}(f).$$

The solutions of the above problem are the set $\mathcal{H}_0$ of generalized solutions of Problem (11). If $\mathcal{H}_0$ is not empty, then we denote by $f_\mathcal{H}^\dagger$ the Moore–Penrose solution, that is, the generalized solution with minimal norm in $\mathcal{H}$. Such a solution might in general not exist, but if it does it often replaces the regression function as the target of learning. Note that $f_\mathcal{H}^\dagger$ can be written as $f_\mathcal{H}^\dagger = S_\rho^\dagger f_\rho$. Also, note that the empirical Problem (10) always has a Moore–Penrose solution given by $f_n^\dagger = S_n^\dagger \mathbf{y}$, but in general the latter does not play any special role.

Remark 8 (*Empirical and population problems*) Let $\rho_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ be the empirical measure on the data. For sake of simplicity, assume that $x_i \neq x_j$ if $i \neq j$, then we can identify $L^2(\mathcal{X}, \rho_n)$ with $\mathbb{R}^n$ endowed with the scalar product

$$\langle w, w' \rangle_n = \frac{1}{n} w^\top w', \tag{13}$$

and S_ρ reduces to S_n if we replace ρ by ρ_n. Developing this latter observation we can view Problem (11) as the ideal inverse problem we would wish to solve, and to Problem (10) as a corresponding empirical problem. It is important to note that unlike classical inverse problems, here the operators defining the two problems have same domains but different ranges. We will see next how the distance (noise) between the two problems can be quantified. Note that, if $x_i = x_j$ for some pair $i \neq j$, the space $L^2(\mathcal{X}, \rho_n)$ can be identified with the subspace of vectors $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ such that $c_i = c_j$ and S_n takes values in this subspace.

Finally, the following remark proposes a data model that is tuned to regard learning framework as an inverse problem. This model differs from the usual setting in inverse problems.

Remark 9 (*Noise and sampling*) Following the above setting, Problem (10) can be seen as a noisy randomly discretized version of Problem (11). Note, however, that it is not immediately clear how this idea can be formalized since the operators defining the two problems have different range ($\mathbf{y}$ is a vector and f_ρ a function!). One idea is to consider the normal equations associated to the two problems that is

$$S_n^* S_n f = S_n^* \mathbf{y}, \qquad S_\rho^* S_\rho f = S_\rho^* f_\rho.$$

This suggests to consider the quantities

$$\|S_\rho^* f_\rho - S_n^* \mathbf{y}\|_{\mathcal{H}}, \qquad \left\|S_\rho^* S_\rho - S_n^* S_n\right\|_\infty$$

as a measure of the perturbation due to random noise and random sampling (here $\|\cdot\|_\infty$ is the operator norm).

As seen in the following, they will play a role similar to the noise level in classical inverse problems, but will also require probabilistic tools, such as concentration inequalities.

The above discussion raises at least two lines of questions. The first concerns the nature of the inverse problem describing supervised learning. We investigate this formulation by discussing the nature of the space $\mathcal{H}$ considered and analyzing the operators defining the problem. Further, we comment on the connection with related problems. The second question regards the extension of algorithmic ideas from regularization theory to learning. This will be the topic of all the rest of the chapter. First, we provide a self-contained introduction on the theory of reproducing kernel Hilbert spaces.

3 Reproducing Kernel Hilbert Spaces and Related Operators

As seen above, the key condition to cast supervised learning as a linear inverse problem is assuming the solution space to be a Hilbert space with continuous evaluation

functionals. This is a vast class of function spaces called Reproducing Kernel Hilbert Spaces (RKHS) and next we recall a few important properties. Two observations are important for our discussion. First, as we show below, different choices of RKHS are possible, effectively introducing different parameterizations of the solution space. Second, as we discuss, the operators defined in the previous section can be seen as restriction/extension operators and are closely related to integral operators and corresponding integral equations. A classic reference on RKHS is [1] and a self-contained introduction can also be found in [45].

3.1 Reproducing Kernels

Each RKHS has an associated *reproducing kernel*

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad K(x, x') = \langle K_{x'}, K_x \rangle_{\mathcal{H}}, \tag{14}$$

which is a symmetric positive definite function, that is, such that the matrix with entries $K(x_i, x_j)$ is symmetric and positive semi-definite for all $x_1, \ldots, x_N \in \mathcal{X}$ and $n \in \mathbb{N}$. In particular, the vector $K_x \in \mathcal{H}$ corresponds to the function $K(x, \cdot)$.

Examples of kernels and RKHS abound, here we provide three basic ones.

Example 10 (*Linear kernel*) Let $\mathcal{X} = \mathbb{R}^d$ and consider the kernel $K(x, x') = x^\top x'$, for all $x, x' \in \mathcal{X}$. The corresponding RKHS is the space of linear functions on $\mathbb{R}^d$

$$\mathcal{H} = \{f_w : \mathbb{R}^d \to \mathbb{R} : f_w(x) = w^\top x \text{ where } w \in \mathbb{R}^d\} \qquad \|f_w\|_{\mathcal{H}}^2 = w^\top w.$$

Example 11 (*Finite dictionaries*) Consider a finite family $\{\phi_i : \mathcal{X} \to \mathbb{R} \mid i = 1, \ldots, p\}$ of p functions and define the kernel

$$K(x, x') = \sum_{j=1}^{p} \phi_j(x)\phi_j(x') = \Phi_p^\top(x)\Phi_p(x) \qquad x, x' \in \mathcal{X}, \tag{15}$$

where $\Phi_p : \mathcal{X} \to \mathbb{R}^p$, $\Phi_p(x)^i = \phi_i(x)$, is called the feature map. The corresponding RKHS is

$$\mathcal{H} = \{f_w : \mathcal{X} \to \mathbb{R} \ : \ f_w(x) = \sum_{j=1}^{p} w^j \phi_j(x) \text{ where } w \in \mathbb{R}^p\},$$

with the norm

$$\|f\|_{\mathcal{H}}^2 = \inf\{w^\top w \ : \ w \in \mathbb{R}^p \text{ such that } f_w = f\}.$$

It is easy to check that the operator

$$U : \mathbb{R}^p \to \mathcal{H} \qquad Uw = f_w$$

is a partial isometry from $(\ker U)^\perp$ onto $\mathcal{H}$.

Example 12 (*Gaussian kernel*) Let $\mathcal{X} = \mathbb{R}^d$. Given $\gamma > 0$, consider the kernel

$$K(x, x') = e^{-\|x-x'\|^2\gamma} \qquad x, x' \in \mathcal{X}.$$

The corresponding RKHS $\mathcal{H}$ can be seen as the subspace of

$$\mathcal{H} = \{f \in L^2(\mathbb{R}^d, dx) \ : \|f\|_{\mathcal{H}}^2 := C_\gamma \int_{\mathbb{R}^d} |\tilde{f}(\omega)|^2 e^{\frac{\|\omega\|^2}{2}} d\omega < \infty\},$$

where $L^2(\mathbb{R}^d, dx)$ is the Hilbert space of functions $f : \mathbb{R}^d \to \mathbb{R}$, which are square-integrable with respect to the Lebesgue measure dx of $\mathbb{R}^d$, $\tilde{f}$ denotes the Fourier transform of f, and C_γ is a suitable constant.

Note that Assumption (9) corresponds to assume that

$$K(x, x) \leq \kappa^2 \qquad \rho_X - \text{a.e. } x \in \mathcal{X}, \tag{16}$$

where it is understood that K is measurable.

3.2 The Operators Defined by the Kernel

As mentioned before, it is also useful to analyze the operators defined by the kernel, since they define the inverse problem associated to supervised learning.

We begin noting that functions in the reproducing kernel Hilbert space $\mathcal{H}$ are defined over the whole space $\mathcal{X}$, whereas functions in $L^2(\mathcal{X}, \rho_X)$ are defined ρ_X-almost everywhere. Roughly speaking, elements in $L^2(\mathcal{X}, \rho_X)$ are uniquely defined[1] only on the support $\mathcal{X}_\rho$ of the marginal distribution ρ_X (we recall that $\mathcal{X}_\rho$ is the smallest closed subset of $\mathcal{X}$ having ρ_X-measure 1).

If $\mathcal{X} \subseteq \mathbb{R}^d$ and ρ_X has a strictly positive density with respect to the Lebesgue measure, then $\mathcal{X} = \mathcal{X}_\rho$, however in machine learning $\mathcal{X}_\rho$ can be strictly contained in $\mathcal{X}$. Indeed, it is often interesting to think of $\mathcal{X}$ as a high-dimensional Euclidean space and $\mathcal{X}_\rho$ as smaller set, for example, a low-dimensional sub-manifold.

In this view, the operator S_ρ can be seen as a *restriction operator* that given a function defined over the whole space $\mathcal{X}$ provides a restriction to $\mathcal{X}_\rho$. The corresponding adjoint operator $S_\rho^* : L^2(\mathcal{X}, \rho_X) \to \mathcal{H}$ can be shown to have the form

$$S_\rho^* g = \int_{\mathcal{X}} g(x) K_x \, d\rho_X(x), \qquad \forall g \in L^2(\mathcal{X}, \rho_X),$$

[1] More precisely, two continuous functions that are equal almost everywhere, they coincide only on $\mathcal{X}_\rho$.

where the integral converges as a Bochner $\mathcal{H}$-valued integral since $\|K_x\|_{\mathcal{H}}$ is bounded. Note that $S_\rho^* g$ only depends on g which is itself defined only on $\mathcal{X}_\rho$. Since $S_\rho^* g$ is an element of $\mathcal{H}$, it is defined on the whole space $\mathcal{X}$ and S_ρ^* can be seen as an *extension* operator. Given a function g defined on $\mathcal{X}_\rho$, $S_\rho^* g$ provides a harmonic extension on the whole space $\mathcal{X}$ defined by the kernel K. The interpretation of kernel operators as restriction/extension operators is discussed in [14] and connected to manifold learning [4].

We stress that the composition of the restriction and extension operators is not identity. Indeed, it is easy to check that the operator

$$L_\rho = S_\rho S_\rho^* : L^2(\mathcal{X}, \rho_\mathcal{X}) \to L^2(\mathcal{X}, \rho_\mathcal{X})$$

is the integral operator defined by the kernel K

$$L_\rho g(x) = \int_\mathcal{X} K(x, x') g(x') d\rho_\mathcal{X}(x'), \qquad \rho_\mathcal{X} - \text{a.e.} x \in \mathcal{X}, \quad g \in L^2(\mathcal{X}, \rho_\mathcal{X}), \quad (17)$$

and the operator $T_\rho = S_\rho^* S_\rho : \mathcal{H} \to \mathcal{H}$ can be written as

$$T_\rho = \int_\mathcal{X} K_x \otimes K_x \, d\rho_\mathcal{X}(x),$$

where $K_x \otimes K_x : \mathcal{H} \to \mathcal{H}$ is the rank-one operator

$$(K_x \otimes K_x)(f) = \langle K_x, f \rangle_\mathcal{H} K_x = f(x) K_x,$$

and the integral converges in the Banach space $\mathcal{S}_1(\mathcal{H})$ of trace class operators, see Appendix. Furthermore, Eq. (8) shows that

$$\langle T_\rho f, f' \rangle_\mathcal{H} = \int_\mathcal{X} f(x) f'(x) \, d\rho_\mathcal{X}(x), \qquad \forall f, f' \in \mathcal{H}.$$

As discussed below T_ρ can be seen as a suitable covariance operator.

Remark 13 (*Properties of the kernel operators*) The operators L_ρ, T_ρ are trace class positive operators and S_ρ, S_ρ^* are Hilbert–Schmidt operators, see Appendix.

3.2.1 Empirical Kernel Operators

Following the above discussion, the sampling operator S_n can be seen as the restriction operator associated to the input points $x_1, \ldots, x_n$. Given a function in $\mathcal{H}$ it evaluates the function at the training set inputs. Note that, if we endow $\mathbb{R}^n$ with the normalized scalar product (13), then

$$\left\| S_n f - \mathbf{y} \right\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2, \tag{18}$$

where the right-hand side is called the empirical error of f.

The adjoint operator $S_n^* : \mathbb{R}^n \to \mathcal{H}$ can be shown to have the following form:

$$S_n^* c = \frac{1}{n} \sum_{i=1}^{n} K_{x_i} c^i, \qquad \forall c \in \mathbb{R}^n. \tag{19}$$

As discussed above $\mathbb{R}^n$ can be identified with $L^2(\mathcal{X}, \rho_n)$, whereas the latter can be seen as space of functions defined on the training set inputs. In this view, by identifying c with $(f(x_1), \ldots, f(x_n))$, the action of S_n^* can be seen as an extension operator providing the value of the functions outside of the training set inputs. Such an operator is called an *out-of-sample extension*.

The operator $L_n = S_n S_n^* : L^2(\mathcal{X}, \rho_n) \to L^2(\mathcal{X}, \rho_n)$ can be written as

$$(L_n c)^i = \frac{1}{n} \sum_{j=1}^{n} K(x_i, x_j) c^j = K_n / n, \tag{20}$$

where, in the last equality, we identify $L^2(\mathcal{X}, \rho_n)$ with $\mathbb{R}^n$ and K_n is the $n \times n$ matrix

$$(K_n)_{ij} = K(x_i, x_j) \qquad i, j = 1, \ldots, n.$$

The operator L_n can be seen as discretization of the integral operator in (17) [27]. The matrix K_n is called the kernel matrix. The operator $T_n = S_n^* S_n : \mathcal{H} \to \mathcal{H}$ can be written as

$$T_n = \frac{1}{n} \sum_{j=1}^{n} K_{x_i} \otimes K_{x_i},$$

so that

$$\langle T_n f, f' \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{j=1}^{n} f(x_i) f'(x_i), \qquad \forall f, f' \in \mathcal{H}.$$

As discussed below T_n can be seen as a suitable empirical covariance operator.

3.3 The Linear Kernel Case and Compressed Sensing

The above operators take a simple form if we consider the linear kernel in $\mathbb{R}^d$. In this case, the RKHS can be identified with $\mathbb{R}^d$ itself and the sampling operator S_n with the $n \times d$ data matrix X_n whose rows are the input points. The adjoint S_n^* is the

transpose of X_n (multiplied by $1/n$) and $S_n^* S_n$ is the empirical covariance matrix[2]

$$\Sigma_n = \frac{1}{n} X_n^\top X_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top .$$

In the population case, the only operator that have a familiar form is $S_\rho^* S_\rho$ that can be seen as the population covariance

$$\Sigma = \mathbb{E}[\frac{1}{n} X_n^\top X_n] = \int_{\mathbb{R}^d} x x^\top \, d\rho_X(x).$$

Remark 14 (*Connection to compressed sensing and linear regression*) Note that the sampling operator can be seen as a collection of measurements defined by random linear functionals. This suggests a connection to classical linear regression but also to compressed sensing [23]. Indeed, if we consider the linear kernel, then Problem (10) can be written as

$$X_n w = \mathbf{y},$$

where X_n is the n by d data matrix, $y_i = x_i^\top w_* + \epsilon_i$, and w_* is a parameter to be estimated. Unlike in compressed sensing, the source of randomness in the sampling operator lies in the nature of the data and it is not a design choice.

4 Tikhonov Regularization

In this section, we introduce Tikhonov regularization, discuss its numerical realization, and finally develop a suitable learning error analysis. The idea of considering Tikhonov regularization for statistical estimation problem dates back to the work on ridge regression [25]. The idea of combining Tikhonov regularization with kernels has been emphasized in [49]. The analysis of Tikhonov regularization as discussed in this chapter was first proposed in [17], where it was shown how to avoid empirical process and covering number estimates in the learning error analysis, bringing in the idea of using covariance estimates to approximate integral operators. These ideas were then developed in several papers. In particular, [44] where bounds are sharpened, and [12] where they were further improved under refined assumptions.

Following, the connection discussed before, consider the family of variational problems,

$$\min_{f \in \mathcal{H}} \left(\|S_n f - \mathbf{y}\|_n^2 + \lambda \|f\|_{\mathcal{H}}^2 \right), \tag{21}$$

parametrized by $\lambda > 0$, called the regularization parameter. By (19), the above problem can be written as

[2] Or rather the second moment matrix.

$$\min_{f\in\mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

The first term is the empirical error of f and promotes functions that fit well the data $(x_1, y_1), \ldots, (x_n, y_n)$, whereas the second term encourages functions having small norm in $\mathcal{H}$, *i.e.,* that are more regular.

We next comment on its numerical realization.

4.1 Numerical Aspects

A direct computation shows that the minimizer of Problem (21) is given by

$$f_n^\lambda = (S_n^* S_n + \lambda I)^{-1} S_n^* \mathbf{y}. \tag{22}$$

Note that, while the sampling operator has finite rank, in general, the above expression is not directly applicable since $(S_n^* S_n + \lambda I)$ is an operator from $\mathcal{H}$ to $\mathcal{H}$.

If the space $\mathcal{H}$ is finite dimensional, it might be possible to identify f_n^λ with a finite-dimensional vector and use (22) directly. For example, if we consider the linear kernel, f_n^λ can be identified with some $w \in \mathbb{R}^d$, see Remark 14, and S_n with X_n, the n by d the data matrix. Similarly, if we consider the kernel defined by a dictionary $\phi_1, \ldots, \phi_p$, f_n^λ can be identified with some $w \in \mathbb{R}^p$ and S_n with the n by p matrix with rows $(\phi_1(x_i), \ldots, \phi_p(x_i))$ for $i = 1, \ldots, n$.

The following lemma provides a finite-dimensional representation formula, usually referred to as the representer theorem [49].

Lemma 15 *For all $\lambda > 0$, let f_n^λ be defined as in (22), then*

$$f_n^\lambda(x) = \sum_{i=1}^{n} K(x, x_i) c_i, \qquad \boldsymbol{c} = (K_n + \lambda n I)^{-1} \mathbf{y}, \tag{23}$$

where $\boldsymbol{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$.

Proof Since $S_n^S (S_n S_n^* + \lambda I) = (S_n^* S_n + \lambda I) S_n^*$, then

$$(S_n^* S_n + \lambda I)^{-1} S_n^* = S_n^* (S_n S_n^* + \lambda I)^{-1},$$

so that

$$f_n^\lambda = S_n^* (S_n S_n^* + \lambda I)^{-1} \mathbf{y}. \tag{24}$$

Further, (20) gives

$$(S_n S_n^* + \lambda I)^{-1} \mathbf{y} = (L_n + \lambda I)^{-1} \mathbf{y} = n (K_n + \lambda n I)^{-1} \mathbf{y} = n \mathbf{c},$$

where $\mathbf{c} = (K_n + \lambda n I)^{-1}\mathbf{y}$. Equation (24) with (19) reads

$$f_n^\lambda = nS_n^*\mathbf{c} = \sum_{i=1}^n K_{x_i} c^i,$$

so that the reproducing Formulas (8) and (14) give

$$f_n^\lambda(x) = \left\langle K_x, f_n^\lambda \right\rangle_{\mathcal{H}} = \sum_{i=1}^n \left\langle K_x, K_{x_i} \right\rangle_{\mathcal{H}} c_i = \sum_{i=1}^n K(x, x_i) c_i.$$

□

We add two remarks related to computational requirements, before discussing the statistical properties of this method.

Remark 16 (*Complexity of Tikhonov regularization*) Note that, for finite-dimensional spaces with dimension p, as in Example 11, the time and memory complexity for Tikhonov regularization are, respectively, the minimum between $O(\min np^2 + p^3,\ pn^2 + n^3)$ and $O(np)$, respectively. For infinite-dimensional spaces, they are $O(n^3)$ and $O(n^2)$, or more precisely $O(c_K n^2 + n^3)$ and $O(n)$, where c_K is the cost of computing $K(x, x')$ for $x, x' \in \mathcal{X}$ and usually $c_K = O(d)$, so that the time complexity is $O(dn^2 + n^3)$.

Remark 17 (*Model selection and regularization path*) In practice, the regularization parameter λ needs to be chosen and this often requires computing the family solutions corresponding to different regularization levels, often called regularization path. In this case, the above time complexities need to be multiplied by the number of regularization parameter values to be tried.

4.2 Error Analysis for Tikhonov Regularization

We next provide an error analysis for Tikhonov regularization. We make a few simplifying assumptions. We assume there exists $f_* \in \mathcal{H}$ such that

$$y_i = f_*(x_i) + \epsilon_i,$$

where x_i are random points and ϵ_i zero mean bounded random numbers. In particular, the couples $(x_1, \epsilon_1), \ldots, (x_n, \epsilon_n)$ are i.i.d. Note that the above conditions are equivalent to assuming that

$$S_\rho f_* = f_\rho \qquad |y| \le M \text{ almost surely.} \tag{25}$$

As discussed before, in general, the support X_ρ of the marginal distribution $\rho_\mathcal{X}$ is not the whole space $\mathcal{X}$ so that $\ker S_\rho$ is not empty. For example, it is easy to show that if

K is continuous, then

$$\ker S_\rho = \{K_x \ : \ x \in X_\rho\}^\perp.$$

If $\ker S_\rho \neq \{0\}$, $f_{\mathcal{H}}$ is not uniquely defined. In order to restore uniqueness, according to the discussion in Remark 7, we consider $f_{\mathcal{H}}^\dagger = S_\rho^\dagger f_\rho$, which always exists and satisfies

$$S_\rho f_{\mathcal{H}}^\dagger = f_\rho. \qquad f_{\mathcal{H}}^\dagger \in \ker S_\rho{}^\perp, \tag{26}$$

as well as

$$S_\rho^* S_\rho f^\dagger = S_\rho^* f_\rho.$$

The main result of this section is the following theorem, whose proof is provided in Sect. 4.6. While the natural norm to consider is the $L^2(\mathcal{X}, \rho)$ norm, here we consider estimates in $\mathcal{H}$ since they are easier to prove. The proof of estimates in $L^2(\mathcal{X}, \rho)$ follows similar, albeit more complex, ideas.

Theorem 18 *Fix $\tau > \ln 4$, the following results hold true:*

- *for all $\lambda > 0$ with probability at least $1 - 4e^{-\tau}$*

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c \frac{\tau}{\lambda \sqrt{n}} + \|f^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}},$$

  where c is a suitable constant independent of n, λ, and τ;
- *if $\lambda = \lambda_n$ is chosen as a function of the number of points so that*

$$\lim_{n\to\infty} \lambda_n = 0, \qquad \lim_{n\to\infty} \frac{1}{\lambda_n \sqrt{n}} = 0,$$

  then, with probability at least $1 - 4e^{-\tau}$,

$$\lim_{n\to\infty} \|f_n^{\lambda_n} - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} = 0. \tag{27}$$

Furthermore, assume that

$$f_{\mathcal{H}}^\dagger \in \operatorname{ran} (S_\rho^* S_\rho)^r, \tag{28}$$

for some $0 < r \leq 1$, then

- *for all $\lambda > 0$, with probability at least $1 - 4e^{-\tau}$,*

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq C \left(\frac{\tau}{\lambda \sqrt{n}} + \lambda^r \right),$$

  where C is a suitable constant independent of n, λ, and τ;
- *if λ is chosen as a $\lambda_n = n^{-\frac{1}{2(r+1)}}$, then with probability at least $1 - 4e^{-\tau}$*

$$\|f_n^{\lambda_n} - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq C' \tau n^{-\frac{r}{2(r+1)}}, \tag{29}$$

where C' is a suitable constant independent of n and τ.

The assumption that $\tau > \ln 4$ ensures that $1 - 4e^{-\tau} > 0$. Equation (27) states that Tikhonov estimator $f_n^{\lambda_n}$ with λ_n going to zero slower than $1/\sqrt{n}$ converges in probability to $f_{\mathcal{H}}^{\dagger}$, which agrees almost everywhere with the regression function f_ρ, see (26). Bound (29) provides a convergence rate under the a priori Assumption (28), we now comment. Recall that $S_\rho^* S_\rho$ is a positive trace class operator such that

$$\overline{\operatorname{ran} S_\rho^* S_\rho} = \ker S_\rho^* S_\rho{}^\perp = \ker S_\rho{}^\perp,$$

so that for any $r \in \mathbb{R}$ the operator $(S_\rho^* S_\rho)^r$ is defined by functional calculus, see Appendix. Moreover, Hilbert–Schmidt theorem implies that there exists an orthonormal base $(v_\ell)_{\ell\in\Lambda}$ of $\ker S_\rho{}^\perp$ and a sequence of strictly positive numbers $(\sigma_\ell)_{\ell\in\Lambda}$ such that

$$S_\rho^* S_\rho v_\ell = \sigma_\ell v_\ell \qquad \ell \in \Lambda \qquad \text{and} \qquad \sum_{\ell\in\Lambda} \sigma_\ell < +\infty,$$

so that $(S_\rho^* S_\rho)^r_\ell = \sigma_\ell^r v_\ell$ for all $\ell \in \Lambda$. Hence, the source condition (28) is equivalent to

$$\sum_{\ell\in\Lambda} \frac{\left\langle f_{\mathcal{H}}^{\dagger}, v_\ell \right\rangle_{\mathcal{H}}^2}{\sigma_\ell^{2r}} < +\infty. \tag{30}$$

If Λ is finite, i.e., S_ρ has finite rank, (30) holds true for any $r \geq 0$. If Λ is infinite, without loss of generality we can assume that $\Lambda = \mathbb{N}$. Since $f_{\mathcal{H}}^{\dagger} \in \mathcal{H}$ Condition (30) is always satisfied with $r = 0$. However, since the sequence $(\sigma_\ell)_{\ell\in\Lambda}$ goes to zero for ℓ going to infinity, if $r > 0$, (30) requires that the Fourier coefficients $\left\langle f_{\mathcal{H}}^{\dagger}, v_\ell \right\rangle_{\mathcal{H}}$ of $f_{\mathcal{H}}^{\dagger}$ on the base $(v_\ell)_\ell$ go to zero faster than an ℓ_2-sequence. Hence, the source Condition (30) can be seen as a regularity requirement on $f_{\mathcal{H}}^{\dagger}$ with respect to the spectral properties of $S_\rho^* S_\rho$, where r plays the role of a smoothness parameter.

The bound in (29) states that the learning rate of the Tikhonov estimator $f_n^{\lambda_n}$ is $1/n^{\frac{r}{2r+1}}$ where the exponent is an increasing function of r. This means that if $f_{\mathcal{H}}^{\dagger}$ is more regular, we need less training points to achieve the same error. However, there is a saturation effect stated by the assumption that $0 < r \leq 1$. This means that even if $f_{\mathcal{H}}^{\dagger}$ satisfies (28) with $r > 1$, the best rate achieved by Tikhonov estimator $f_n^{\lambda_n}$ is $1/n^{1/4}$, corresponding to the choice $r = 1$.

The above result can be extended to cover the case when f_ρ does not belong to the hypotheses space $\mathcal{H}$, as well as to derive estimates in $L^2(\mathcal{X}, \rho_n)$. To conclude a matching lower bound can be derived, showing that the obtained bounds are optimal in a precise sense. We refer the interested reader to [8, 12, 26, 31, 46] and references therein.

4.3 Error Decomposition

To derive the above bound, we first consider a suitable error decomposition and then study the various error terms. The idea is to first study the difference $\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}$ for any $\lambda > 0$, and then derive a suitable choice for λ. To this aim, we decompose such an error into several terms. We begin by considering

$$f^\lambda = (S_\rho^* S_\rho + \lambda I)^{-1} S_\rho^* f_\rho, \tag{31}$$

which is the unique solution of the problem

$$\min_{f \in \mathcal{H}} \left(\|S_\rho f - f_\rho\|_\rho^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

Then, we have the following equality:

$$f_n^\lambda - f_{\mathcal{H}}^\dagger = f_n^\lambda - f^\lambda + f^\lambda - f_{\mathcal{H}}^\dagger.$$

In the above expression:

- The term $f^\lambda - f_{\mathcal{H}}^\dagger$ does not depend on the data, but only on the distribution and is called approximation error or bias.
- The term $f_n^\lambda - f^\lambda$ depends on the data; is stochastic; and is called variance, estimation, or sample error.

We study these two terms next.

4.4 Approximation Error

A first question is whether the approximation error converges to zero and a second question we can ask is if it is possible to derive the rate of convergence for (32) under the source Condition (28). The theory of inverse problems provides a positive answer to both questions [22], that we state in the following theorem.

Theorem 19 *Under the assumption that $f_{\mathcal{H}}^\dagger$ exists, see (26), then*

$$\lim_{\lambda \to 0} \|f^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} = 0. \tag{32}$$

Furthermore, if $f_{\mathcal{H}}^\dagger$ satisfies the source Condition (28) with $0 < r \leq 1$, then

$$\|f^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq \lambda^r \|(S_\rho^* S_\rho)^{-r} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}. \tag{33}$$

Proof By construction $f^\lambda, f_{\mathcal{H}}^\dagger \in \ker S_\rho^* S_\rho{}^\perp$ and $(v_\ell)_{\ell \in \Lambda}$ is a base $\ker S_\rho^* S_\rho{}^\perp$ of eigenvectors of $S_\rho^* S_\rho$ with strictly positive eigenvalues σ_ℓ. Using (31) we have

$$f^\lambda - f_{\mathcal{H}}^\dagger = ((S_\rho^* S_\rho + \lambda I)^{-1} S_\rho^* S_\rho - I) f_{\mathcal{H}}^\dagger = -\lambda (S_\rho^* S_\rho + \lambda I)^{-1} f_{\mathcal{H}}^\dagger,$$

so that

$$\| f^\lambda - f_{\mathcal{H}}^\dagger \|_{\mathcal{H}}^2 = \sum_{\ell \in \Lambda} \left(\frac{\lambda}{\sigma_\ell + \lambda} \right)^2 \left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2 .$$

Fix ℓ, $\frac{\lambda}{\sigma_\ell + \lambda}$ goes to zero if λ goes to zero and $0 < \frac{\lambda}{\sigma_\ell + \lambda} \leq 1$. Moreover, the series $\sum_\ell \left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2$ converges, so that dominated convergence theorem implies that $\lim_{\lambda \to 0} \| f^\lambda - f_{\mathcal{H}}^\dagger \|_{\mathcal{H}}^2$ goes to zero, which is (32).

Assume (28) with $0 < r \leq 1$. Then

$$\begin{aligned} \left(\frac{\lambda}{\sigma_\ell + \lambda} \right)^2 \left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2 &= \left(\frac{\lambda \sigma_\ell^r}{\sigma_\ell + \lambda} \right)^2 \frac{\left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2}{\sigma_\ell^{2r}} = \lambda^{2r} \left(\frac{(\sigma_\ell \lambda^{-1})^r}{(\sigma_\ell \lambda^{-1}) + 1} \right)^2 \frac{\left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2}{\sigma_\ell^{2r}} \\ &\leq \lambda^{2r} \frac{\left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2}{\sigma_\ell^{2r}}, \end{aligned}$$

where the last inequality holds true since the function x^r is convex with tangent line at $x = 1$ given by $y = 1 + rx$, so that

$$x^r \leq 1 + r(x - 1) \leq 1 + x \qquad x > 0.$$

Then

$$\| f^\lambda - f_{\mathcal{H}}^\dagger \|_{\mathcal{H}}^2 \leq \lambda^{2r} \sum_{\ell \in \Lambda} \frac{\left\langle f_{\mathcal{H}}^\dagger, v_\ell \right\rangle_{\mathcal{H}}^2}{\sigma_\ell^{2r}} = \lambda^{2r} \| (S_\rho^* S_\rho)^{-r} f_{\mathcal{H}}^\dagger \|_{\mathcal{H}}^2,$$

which is (33). □

4.5 Sample Error

We now consider the sample error. The idea is to further decompose the above expression to isolate the perturbations due to noise and random sampling as shown by the following result.

Lemma 20 *For all* $\lambda > 0$

$$\| f_n^\lambda - f_\lambda \|_{\mathcal{H}} \leq \lambda^{-1} \left(\| (S_n^* \boldsymbol{y} - S_n^* S_n f_{\mathcal{H}}^\dagger) \|_{\mathcal{H}} + \left\| S_n^* S_n - S_\rho^* S_\rho \right\|_\infty \| f_{\mathcal{H}}^\dagger \|_{\mathcal{H}} \right). \quad (34)$$

Proof By the explicit form of f_n^λ and f^λ we get

$$f_n^\lambda - f_\lambda = (S_n^* S_n + \lambda I)^{-1} S_n^* \mathbf{y} - (S_\rho^* S_\rho + \lambda I)^{-1} S_\rho^* f_\rho. \tag{35}$$

We add and subtract $(S_n^* S_n + \lambda I)^{-1} S_n^* S_n f_{\mathcal{H}}^\dagger$ in (35) so that

$$f_n^\lambda - f_\lambda = (S_n^* S_n + \lambda I)^{-1}(S_n^* \mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger) + [(S_n^* S_n + \lambda I)^{-1} S_n^* S_n - (S_\rho^* S_\rho + \lambda I)^{-1} S_\rho^* S_\rho) f_{\mathcal{H}}^\dagger],$$

where we used the fact that $S_\rho^* f_\rho = S_\rho^* S_\rho f_{\mathcal{H}}^\dagger$ by (25). Furthermore

$$\begin{aligned}(S_n^* S_n + \lambda I)^{-1} S_n^* S_n - (S_\rho^* S_\rho + \lambda I)^{-1} S_\rho^* S_\rho &= (S_n^* S_n + \lambda I)^{-1} S_n^* S_n - S_\rho^* S_\rho (S_\rho^* S_\rho + \lambda I)^{-1} = \\ &= (S_n^* S_n + \lambda I)^{-1} \left(S_n^* S_n (S_\rho^* S_\rho + \lambda I) - (S_n^* S_n + \lambda I) S_\rho^* S_\rho \right) (S_\rho^* S_\rho + \lambda I)^{-1} = \\ &= \lambda (S_n^* S_n + \lambda I)^{-1} \left(S_n^* S_n - S_\rho^* S_\rho \right) (S_\rho^* S_\rho + \lambda I)^{-1}.\end{aligned}$$

Hence

$$\begin{aligned}[(S_n^* S_n + \lambda I)^{-1} S_n^* S_n - (S_\rho^* S_\rho + \lambda I)^{-1} S_\rho^* S_\rho)] f_{\mathcal{H}}^\dagger = \\ = \lambda (S_n^* S_n + \lambda I)^{-1} \left(S_n^* S_n - S_\rho^* S_\rho \right) (S_\rho^* S_\rho + \lambda I)^{-1} f_{\mathcal{H}}^\dagger.\end{aligned}$$

Since $S_n^* S_n$ and $S_\rho^* S_\rho$ are positive operators,

$$\left\| (S_n^* S_n + \lambda I)^{-1} \right\|_\infty \leq \frac{1}{\lambda}, \qquad \left\| (S_\rho^* S_\rho + \lambda I)^{-1} \right\|_\infty \leq \frac{1}{\lambda},$$

so that triangular inequality gives

$$\| f_n^\lambda - f_\lambda \|_{\mathcal{H}} \leq \lambda^{-1} \left(\| (S_n^* \mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger) \|_{\mathcal{H}} + \left\| S_n^* S_n - S_\rho^* S_\rho \right\|_\infty \| f_{\mathcal{H}}^\dagger \|_{\mathcal{H}} \right).$$

□

We isolated the analytic part of the error analysis, and now the above error is expressed in terms of random quantities taking into account noise and random sampling. We have the following concentration inequality.

Proposition 21 *Fix $\tau > \ln 4$, with probability at least $1 - 4e^{-\tau}$*

$$\max\{ \| (S_n^* \mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger) \|_{\mathcal{H}}, \left\| S_n^* S_n - S_\rho^* S_\rho \right\|_\infty \} \leq c \frac{\tau}{\sqrt{n}}, \tag{36}$$

where c is a suitable constant independent of n and τ.

Proof We first estimate $\| (S_n^* \mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger) \|_{\mathcal{H}}$. We introduce the family $(\xi_i)_i$ of i.i.d random variables taking value in $\mathcal{H}$ defined as

$$\xi_i = y_i K_{x_i}.$$

Assumptions (9) and (25) give that

$$\|\xi_i\|_{\mathcal{H}} \leq M\kappa,$$

and a direct computation shows that

$$\mathbb{E}[\xi_i] = S_\rho^* f_\rho = S_\rho^* S_\rho f_{\mathcal{H}}^\dagger \qquad \frac{1}{n}\sum_{i=1}^n \xi_i = S_n^* \mathbf{y}.$$

Hence, Höeffding inequality (76) implies that, with probability at least $1 - 2e^{-\tau}$,

$$\|(S_n^* \mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger)\|_{\mathcal{H}} \leq 2M\kappa\sqrt{\frac{\tau}{n}}.$$

In order to estimate $\left\|S_n^* S_n - S_\rho^* S_\rho\right\|_\infty$, we first note that

$$\left\|S_n^* S_n - S_\rho^* S_\rho\right\|_\infty \leq \left\|S_n^* S_n - S_\rho^* S_\rho\right\|_{\mathcal{S}_2(\mathcal{H})},$$

where $\|\cdot\|_{\mathcal{S}_2(\mathcal{H})}$ is the Hilbert–Schmidt norm in the Hilbert space $\mathcal{S}_2(\mathcal{H})$ of Hilbert–Schmidt operators. Let $(\zeta_i)_i$ be the family of i.i.d. random variables taking value in $\mathcal{S}_2(\mathcal{H})$

$$\zeta_i = K_{x_i} \otimes K_{x_i}.$$

Assumption (25) gives

$$\|\zeta_i\|_{\mathcal{S}_2(\mathcal{H})} \leq \kappa,$$

and, as above,

$$\mathbb{E}[\zeta_i] = S_\rho^* S_\rho \qquad \frac{1}{n}\sum_{i=1}^n \zeta_i = S_n^* S_n,$$

so that by (76), with probability at least $1 - 2e^{-\tau}$

$$\left\|S_n^* S_n - S_\rho^* S_\rho\right\|_{\mathrm{HS}} \leq 2\kappa\sqrt{\frac{\tau}{n}}.$$

An union bound and the fact that $\tau > 1$ provide the claim.

Remark 22 In the above proof, we bound the operator norm $\left\|S_n^* S_n - S_\rho^* S_\rho\right\|_\infty$ with the Hilbert–Schmidt norm $\left\|S_n^* S_n - S_\rho^* S_\rho\right\|_{\mathcal{S}_2(\mathcal{H})}$ in order to use concentration inequality in Hilbert spaces. However, this is a rough bound and it is possible to use concentration inequalities in the operator norm to have tight constants [31, 32, 47].

4.6 Proof of Theorem 18

Proof Since

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq \|f_n^\lambda - f^\lambda\|_{\mathcal{H}} + \|f^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}},$$

combining (34) and (36), with high probability

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c\frac{\tau}{\lambda\sqrt{n}}(1 + \|f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}) + \|f^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}.$$

Up to redefining the constant c, we prove the first claim, whereas (27) is a consequence of (32) and the choice of $\lambda = \lambda_n$.

Furthermore, if $f_{\mathcal{H}}^\dagger$ satisfies (28), bound (33) gives

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c\frac{\tau}{\lambda\sqrt{n}}(1 + \|f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}) + \lambda^r\|(S_\rho^* S_\rho)^{-r} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}},$$

which is the third claim with a suitable constant C. By balancing the two terms

$$\frac{1}{\lambda\sqrt{n}} = \lambda^r,$$

it follows that $\lambda_n = n^{-\frac{1}{2(r+1)}}$ and, with this choice,

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq n^{-\frac{r}{2(r+1)}}\left(c\tau(1 + \|f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}) + \|(S_\rho^* S_\rho)^{-r} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}\right),$$

which allows to derive (29). □

4.7 Optimization Enters the Game: Statistical and Computational Trade-Offs

So far we analyzed the numerical realization and statistical properties of the estimator induced by Tikhonov regularization. From a statistical point of view, we obtained optimal statistical rates. From a numerical point of view, we derived a solution in closed form that requires solving a linear system with roughly $O(n^3)$, $O(n^2)$ time and space requirements, respectively (for general kernels). An observation is that, so far, we considered numerical and statistical aspects in isolation. Questioning this way to proceed is important, especially in large-scale regimes where computations can be massive. Computational requirements depend only on the size of data. But, by assumption, the data are noisy and scattered, and are only a proxy to approximate an ideal problem. Shouldn't the computational requirements depend on the quality of the data? Further, shouldn't they depend on how easy or hard is the underlying problem (e.g., as expressed by the source condition)? This is the line of reasoning we will pursue in the rest of the chapter. A first step is to consider an optimization

perspective on Tikhonov regularization and then couple it with the estimation point of view in the previous section. The idea is to begin by replacing the direct solver with an iterative one. Note that the approach of exploiting self-regularizing properties of optimization is classical in inverse problems [22]. We refer to [9] for the idea of combining statistics and optimization, and to [10] for basic ideas in optimization.

The basic gradient descent iteration applied to Problem (21) gives

$$f_n^j = f_n^{j-1} - \gamma[S_n^*(S_n f_n^{j-1} - \mathbf{y}) + \lambda f_n^{j-1}], \qquad j = 1, \dots t-1.$$

By induction, it is easy to show that the above iteration also admits the following representation all $t \in \mathbb{N}$:

$$f_n^t(x) = \sum_{i=1}^n K(x, x_i) c_i^t, \qquad \mathbf{c}^t = \mathbf{c}^{t-1} - \gamma[\frac{1}{n}(K_n \mathbf{c}^{t-1} - \mathbf{y}) + \lambda \mathbf{c}^{t-1}],$$

where $\mathbf{c}^t = (c_1^t, \dots, c_n^t)$ and $\mathbf{c}^0 = 0$.

5 Iterative Regularization

In this section, we introduce a class of algorithms based on iterative regularization, also called implicit regularization and closely related to the idea of early stopping. We begin illustrating the general idea by discussing a basic example, namely, Landweber iteration [28]. Iterative regularization for learning, corresponding to Landweber iteration, was dubbed L2 boosting and first considered in [11] for a fixed design setting and in [13] for the statistical learning setting. A number of variants have been recently considered, an incomplete list including conjugate gradient [7], accelerated [34] and stochastic gradient methods [2, 30, 39], as well as different averaging schemes [33]. Here, we provide an overview of these results highlighting the interplay between statistics and optimization.

5.1 Landweber Iteration

Consider the algorithm defined by the following sequence:

$$f_n^j = f_n^{j-1} - \gamma S_n^*(S_n f_n^{j-1} - \mathbf{y}), \qquad j = 1, \dots t-1, \tag{37}$$

where $f_n^0 = 0$, $\gamma > 0$, and $t \in \mathbb{N}$. The above iteration can be seen to be the gradient descent iteration of the empirical error (18). It is called Landweber iteration in the context of inverse problems.

Following the discussion for Tikhonov regularization, one can see that, if the space $\mathcal{H}$ is finite dimensional, it is possible to identify f_n^λ with a finite-dimensional vector and use (37) directly. For example, if we consider the linear kernel or the kernel defined by a dictionary $\phi_1, \ldots, \phi_p$, in this latter case, f_n^t can be identified with a vector $w \in \mathbb{R}^p$ and S_n with the n by p matrix with rows $(\phi_1(x_i), \ldots, \phi_p(x_i))$ for $i = 1, \ldots, n$. For more general kernels, the following representer theorem holds.

Lemma 23 *For all $t \in \mathbb{N}$, let f_n^λ be defined as in (37), then*

$$f_n^t(x) = \sum_{i=1}^{n} K(x, x_i) c_i^t, \qquad \boldsymbol{c}^{t+1} = \boldsymbol{c}^t - \frac{\gamma}{n}(K_n \boldsymbol{c}^t - \boldsymbol{y}), \tag{38}$$

where $\boldsymbol{c}^t = (c_1^t, \ldots, c_n^t)$ and $\boldsymbol{c}^0 = 0$

Remark 24 (*Complexity of Landweber iteration*) Note that for finite-dimensional spaces with dimension p, the time and memory complexity for Landweber iteration are, respectively, the minimum between $O(npt)$ and $O(pn^2 + n^2 t)$, and $O(np)$. For infinite-dimensional spaces, they are $O(n^2 t)$ and $O(n^2)$. More precisely the latter complexities included the cost of evaluating the kernel which is often proportional to the data dimension. Also note that, when the latter is small, memory requirements can be reduced, recomputing the kernel on the fly.

5.2 A Regularization View on Gradient Descent

The following result allows to draw a connection to Tikhonov regularization and sheds light on the regularization properties of Landweber iteration.

Lemma 25 *The iteration in (37) can be written as*

$$f_n^t = \gamma \sum_{j=0}^{t-1} (I - \gamma S_n^* S_n)^j S_n^* \boldsymbol{y}.$$

The proof of the above follows from a basic induction argument. It shows that Landweber iteration can be seen as the linear operator

$$G_t = \sum_{j=0}^{t-1} (I - \gamma S_n^* S_n)^j,$$

applied to $S_n^* \mathbf{y}$. Then, using spectral calculus and properties of the geometric series, if γ is chosen so that

$$\left\| I - \gamma S_n^* S_n \right\| < 1, \tag{39}$$

for example taking $\gamma < \kappa^2$, then

$$\gamma \sum_{j=0}^{\infty} (I - \gamma S_n^* S_n)^j = (S_n^* S_n)^{-1},$$

if $(S_n^* S_n)^{-1}$ is assumed to exist for the sake of simplicity. More generally, it can be shown that

$$\gamma \sum_{j=0}^{\infty} (I - \gamma S_n^* S_n)^j S_n^* = S_n^\dagger.$$

Then, if the step size is chosen to satisfy (39), the operator corresponding to Landweber iteration can be seen as truncated series expansion. The only free parameter is the number of iterations which corresponds to the number of terms in such an expansion. It is easy to see that the condition number of the operator G_t is controlled by t, and the bigger t the larger is the condition number. Indeed, the operators

$$(S_n^* S_n + \lambda I)^{-1}, \qquad \gamma \sum_{j=0}^{\infty} (I - \gamma S_n^* S_n)^j,$$

are similar and one can consider roughly a correspondence $t \sim 1/\lambda$. The number of iteration t acts as the regularization parameter for Landweber iteration. This latter observation has crucial computational implications.

5.3 Landweber Iteration and Iterative Regularization

Landweber iteration is an instance of so-called iterative regularization, sometimes called early stopping regularization. The remarkable property of this class of method is that they couple computational and statistical properties. The number of iterations controls at the same time the stability, and hence the learning properties, of the solution as well the computational requirements. More computations are needed if the data can be exploited, whereas fewer computations must be considered to ensure stability when data are poor or scarce.

The above reasoning is made precise by the following result.

Theorem 26 *Fix* $\tau > \ln 4$. *Assume that for some* $r > 0$

$$f_{\mathcal{H}}^\dagger \in \operatorname{ran} (S_\rho^* S_\rho)^r, \tag{40}$$

then the following results hold:

- *for all* $t \in \mathbb{N} > 0$ *with probability at least* $1 - 4e^{-\tau}$

$$\|f_n^t - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c\left(\frac{\tau t}{\sqrt{n}} + \frac{1}{t^r}\right),$$

where c is a suitable constant independent of n, t, and τ;

- *if t is chosen as a $t_n = n^{\frac{1}{2(r+1)}}$ with probability at least $1 - 4e^{-\tau}$*

$$\|f_n^{\lambda_n} - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq Cn^{-\frac{r}{2(r+1)}},$$

where C is a suitable constant independent of n and τ.

The proof follows the same line of one of Theorem 18.

5.4 Proof Sketch

The starting point is to consider the population version of the algorithm

$$f^t = \gamma \sum_{j=0}^{t-1} (I - \gamma S_\rho^* S_\rho)^j S_\rho^* f_\rho,$$

and the error decomposition

$$f_n^t - f_{\mathcal{H}}^\dagger = f_n^t - f^t + f^t - f_{\mathcal{H}}^\dagger.$$

For bounding the sample error $f_n^t - f^t$ it is useful to rewrite the empirical and population gradient iterations as

$$f_n^{t+1} = (I - \gamma S_n^* S_n) f_n^t + \gamma S_n^* \mathbf{y},$$

and

$$f^{t+1} = (I - \gamma S_\rho^* S_\rho) f^t + \gamma S_\rho^* f_\rho,$$

respectively. Then subtracting the above equations, we obtain

$$\begin{aligned} f_n^{t+1} - f^{t+1} &= (I - \gamma S_n^* S_n) f_n^t - (I - \gamma S^* S) f^t + \gamma (S_n^* \mathbf{y} - S_\rho^* f_\rho) \\ &= (I - \gamma S_n^* S_n) f_n^t - (I - \gamma S_n^* S_n + \gamma S_n^* S_n - \gamma S^* S) f_t + \gamma (S_n^* \mathbf{y} - S_\rho^* f_\rho) \\ &= (I - \gamma S_n^* S_n)(f_n^t - f^t) + \gamma [(S^* S - S_n^* S_n) f_t + (S_n^* \mathbf{y} - S_\rho^* f_\rho)], \end{aligned}$$

where, in the second equality, we added and subtracted $\gamma S^* S f_t$. Then, by induction

$$f_n^t - f^t = \gamma \sum_{j=0}^{t-1} (I - \gamma S_n^* S_n)[(S^* S - S_n^* S_n) f^j + (S_n^* \mathbf{y} - S_\rho^* f_\rho)].$$

The bound on the sample error follows taking the norm of the above expression and using the triangle inequality, which reduces to the problem of controlling the stochastic terms

$$\left\| S_\rho^* S_\rho - S_n^* S_n \right\|, \quad \left\| S_n^* \mathbf{y} - S_\rho^* f_\rho \right\|,$$

already studied in the analysis of Tikhonov regularization but also the norm of $\left\| f_t \right\|_{\mathcal{H}}$ which can be shown to be bounded by $\| f_{\mathcal{H}}^\dagger \|_{\mathcal{H}}$ since using spectral calculus properties of the geometric series,

$$f^t = (I - (I - \gamma S_\rho^* S_\rho)^t) f_{\mathcal{H}}^\dagger.$$

The bound on the approximation error also follows from this last equation since it implies that

$$\left\| f^t - f_{\mathcal{H}}^\dagger \right\| = \left\| (I - \gamma S^* S)^t f_{\mathcal{H}}^\dagger \right\|.$$

This latter expression can be controlled with the same approach used for controlling the approximation error of Tikhonov regularization.

5.5 A Regularization View on Optimization

The result in Theorem 26 shows that the statistical properties of Landweber iteration are essentially the same as Tikhonov regularization. However, the computational complexities are different:

- **From statistical to time complexity**. The number of iteration controls the time complexity. For example, in the infinite-dimensional case, we have $O(n^2 n^{\frac{1}{2(r+1)}})$. Note that unlike Tikhonov regularization, Landweber iteration does not suffer from a saturation effect. For easy problems, i.e., large r, the computational difference is dramatic.
- **Regularization path and warm restart**. In iterative regularization, the computation of the regularization path is embedded in the method—unlike Tikhonov regularization.

Another interesting aspect of the above discussion is that it provides a different perspective on optimization methods in the context of machine learning. The classical optimization viewpoint would be to consider the convergence properties of the gradient iteration (37) to a minimizer of the empirical error (18). The above discussion provides an alternative point of view, by looking at gradient descent from a regularization perspective. The iteration (37) is only an empirical iteration whereas the real objective in machine learning is to solve (12). From this perspective, to obtain a good approximation of $f_{\mathcal{H}}^\dagger$, rather than just letting the iteration run to convergence, it is also possible to stop early. This is what we mean here with early stopping. Following this discussion, it is natural to ask whether other optimization methods can

also be analyzed within a regularization framework. This is indeed the case as we discuss in the next.

5.6 Accelerated Iterative Regularization

A key problem in optimization is to find efficient methods to minimize an objective function of interest. The literature on the topic is vast and here we discuss two ideas in the context of machine learning. In particular, we revisit these ideas with respect to the expected error rather than the training error.

Nesterov acceleration. The first idea is the so-called Nesterov acceleration of the gradient method defining Landweber iteration. In our context, it defines the following iteration:

$$f_n^j = f_n^{j-1} - \gamma S_n^*(S_n h_n^{j-1} - \mathbf{y}), \qquad h_n^{j-1} = f_n^{j-1} + \alpha_j(f_n^{j-1} + f_n^{j-2}),$$

for $f_n^0 = f_n^{-1} = 0$, γ satisfying (39) and

$$\alpha_j = \frac{j-1}{j+\beta}, \quad \beta \geq 1.$$

The ν-method. The second method is also known as Chebyshev method, it is related to the heavy-ball method, and is given for $\nu > 0$ by

$$f_n^j = f_n^{j-1} - \omega_t S_n^*(S_n h_n^{j-1} - \mathbf{y}) + \alpha_j(f_n^{j-1} + f_n^{j-2}),$$

for $f_n^0 = f_n^{-1} = 0$ and $\alpha_1 = 0$, $\omega_1 = \frac{4\nu+2}{\kappa^2(4\nu+1)}$ and

$$\omega_j = 4\frac{(2j+2\nu-1)(j+\nu-1)}{(j+2\nu-1)(2j+4\nu-1)},$$

$$\alpha_j = \frac{(j-1)(2j-3)(2j+2\nu-1)}{(j+2\nu-1)(2j+4\nu-1)(2j+2\nu-3)}.$$

The numerical realization of the above methods can be derived analogously to Tikhonov regularization and Landweber iteration. The computational time/space complexities per iteration are the same as Landweber iteration. The key difference with Landweber iteration is seen considering the corresponding error bounds.

5.7 Error Bounds and the Effect of Acceleration

Error bounds for the above accelerated methods can be proved following similar arguments to Tikhonov regularization and Landweber iteration.

Theorem 27 *Fix $t > \ln 4$ and set $r^* = 1/2$ for Nesterov acceleration, and $r_* = \nu - 1/2$ for the ν-method. Assume that for some $0 < r \leq r_*$*

$$f_{\mathcal{H}}^{\dagger} \in \operatorname{ran} (S_\rho^* S_\rho)^r, \tag{41}$$

then the following results hold:

- *for all $t \in \mathbb{N} > 0$ with probability at least $1 - 4e^{-\tau}$*

$$\| f_n^t - f_{\mathcal{H}}^{\dagger} \|_{\mathcal{H}} \leq c \left(\frac{\tau t^2}{\sqrt{n}} + \frac{1}{t^{2r}} \right),$$

  where c is a suitable constant independent of n, t, and τ;
- *if t is chosen as $t_n = n^{\frac{1}{4(r+1)}}$ with probability at least $1 - 4e^{-\tau}$*

$$\| f_n^{\lambda_n} - f_{\mathcal{H}}^{\dagger} \|_{\mathcal{H}} \leq C \tau n^{-\frac{r}{2(r+1)}},$$

  where C is a suitable constant independent of n and τ.

The above results quantify the effect of acceleration in a statistical learning setting. From the above bound, one can see that the approximation error decreases faster than for Landweber acceleration. Indeed, the approximation error term can also be seen as an optimization error. This faster convergence is the direct effect of acceleration. On the other hand, one can also observe that acceleration affects the sample error negatively. This is a well-known instability property of acceleration methods. When combining these two terms, the effect of acceleration cancels out in the final bound. Indeed, the above methods yield again the same optimal bound obtained for Tikhonov regularization and Landweber iteration.

The remarkable property of the accelerate methods above is that they allow for a much more aggressive stopping rule. Indeed, now the regularization parameter is t^2, so that

$$t_n = \sqrt{n^{\frac{1}{2(r+1)}}}$$

iteration suffices. The effect of acceleration is to effectively reduce the time complexity needed for optimal statistical bounds. In the infinite-dimensional setting, the time complexity is given by $O(n^2 \sqrt{n^{\frac{1}{2(r+1)}}})$, which greatly improves the time complexity of Landweber iterations (note, however, that accelerated methods suffer from saturation).

5.8 Incremental and Stochastic Iterative Regularization

Stochastic gradient techniques are often advocated to deal with large-scale problems. We review this techniques within our context. Consider the following iteration:

$$f_n^j = f_n^{j-1} - \gamma_t K_{x_{p(j)}}(f_n^{j-1}(x_{p(j)}) - y_{p(j)}), \quad j = 0, \ldots, q,$$

for $f_n^0 = 0$. We add a few comments. First, the selection function p has values in $\{1, \ldots, n\}$ and can be deterministic or stochastic. Three common choices are: (1) cyclic, p is deterministic; (2) stochastic, p is a uniformly distributed; and (3) reshuffling, p describes a random permutation chosen every n steps. Second, in machine learning, the above iteration is often broadly referred to as stochastic gradient descent although it does not define a descent method. In optimization, the name incremental gradient method is also used. Third, compared to Landweber iteration only an input–output pair is used to compute a point-wise gradient in each iteration. The term *pass* or *epoch* refers to n iterations (note that for stochastic gradient, it corresponds to one pass over the data only on average). Finally, the numerical realization of the above methods can be derived analogously to Tikhonov regularization and Landweber iteration. Keeping in mind the difference between iterations and epochs, it is useful to compare the complexity of the above method to Landweber iteration.

Remark 28 (*Time and space complexity*) The cost of each iteration is the minimum between $O(p)$ and $O(n)$ in the finite-dimensional case, and $O(n)$ in the infinite-dimensional case, omitting the cost of computing the kernel. The memory cost is also per iteration $O(p)$ and $O(n)$ in the finite-dimensional case and $O(n)$ in the finite-dimensional case. Note that, if we consider an epoch rather than an iteration, then time/space complexity of each epoch is the same as Landweber iteration or the accelerated variants, yet the final result is essentially the same as Landweber iteration.

5.9 Error Bounds

The proof of the error bounds for the incremental gradient methods discussed above is considerably more complex than the one for Landweber iteration. However, the obtained bound is essentially the same as the one for Landweber. The following bounds hold for the cyclic and stochastic incremental gradient for $\gamma_t = c/n$.

Theorem 29 *Fix* $\tau > \ln 4$. *Assume that for some* $r > 0$

$$f_{\mathcal{H}}^{\dagger} \in \operatorname{ran}(S_\rho^* S_\rho)^r, \tag{42}$$

choose $\gamma_t = c/n$, *the following results hold:*

- *for all number of epochs* $t \in \mathbb{N} > 0$ *with probability at least* $1 - 4e^{-\tau}$

$$\|f_n^t - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c\left(\frac{\tau t}{\sqrt{n}} + \frac{1}{t^r}\right),$$

where c is a suitable constant independent of n, t, and τ;

- *if t is chosen as a $t_n = n^{\frac{1}{2(r+1)}}$, with probability at least $1 - 4e^{-\tau}$, $\tau > 1$*

$$\|f_n^{\lambda_n} - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq C\tau n^{-\frac{r}{2(r+1)}}, \quad r > 0,$$

where C is a suitable constant independent of n and τ.

The above result shows that again the obtained bounds are optimal; however, in light of Remark 28, the time space complexity of incremental methods is essentially the same as standard gradient descent and worse than accelerated methods. This suggests that there is no gain in considering incremental techniques. We add two final remarks.

Remark 30 (*One pass SGD*) We note that a different variant of the above result shows that only one pass over the data is sufficient provided that

- averaging of the iterates is considered,
- and the step size is chosen as

$$\gamma_t = n^{r/2(r+1)}.$$

While theoretically interesting in practice this might require running multiple passes to choose the step size adaptively.

Remark 31 (*Mini-batches*) Finally, we note that an hybrid between gradient descent and incremental gradient methods is obtained considering mini-batches, that is,

$$f_n^j = f_n^{j-1} - \frac{\gamma_t}{b} \sum_{i=b(j-2)+1}^{b(j-1)} K_{x_{p(i)}}(f_n^{j-1}(x_{p(i)}) - y_{p(i)}), \quad j = 0, \ldots, q,$$

for $f_n^0 = 0$. Here, at each iteration, b points are used to compute the gradient (rather than n or 1). It can be shown that this approach allows to obtain optimal rates, but again does not yield computational improvements. It shows, however, how the choices of mini-batch cardinality and the step size can be done to preserve optimal rates. In particular, larger mini-batches allow to consider larger step size, for example, with $b = \sqrt{n}$ we can consider $\gamma_t = 1/\sqrt{n}$.

6 Regularization with Stochastic Projections

In this section, we discuss how the combination of Tikhonov regularization with stochastic projections allows to retain good statistical bounds while controlling time as well as memory requirements.

We begin some preliminary consideration. We will be interested in the infinite-dimensional setting. In this context, the solution given by Tikhonov regularization is

$$f_n^{\lambda}(x) = \sum_{i=1}^{n} c_i K(x_i, x), \quad c = (K_n + \lambda n I)^{-1}\mathbf{y}. \tag{43}$$

The above procedure has $O(n^3 d)$ and $O(n^2)$ complexity in time and space, respectively. While the time complexity is improved for iterative regularization the space complexity remains the same. We next discuss two basic ideas, namely, *Nÿstrom approximations* and *random features*. Both methods introduce finite-dimensional approximations of the space of functions to be considered, albeit in different ways. We refer to [40, 41] for large-scale kernel methods using projection methods, and to [38] for random features.

We start by observing that regularization with projections is well known in inverse problems [22] as we are going to recall in this section. We next introduce Nÿstrom approximations and we show that they can be seen as a form of regularization with projections, where the latter are stochastic.

6.1 Projection Regularization

Projection regularization is classical in inverse problems [22] and is based on considering a family of finite-dimensional subspaces and corresponding projection operators. In our context, given a RKHS $\mathcal{H}$ the classic least-squares projection method corresponds to considering a family of finite-dimensional nested subspaces $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}$, with corresponding projection operators $P_1, P_2, \ldots$, and define a family of approximate solutions,

$$f_n^M = S_{n,M}^{\dagger}\mathbf{y} \qquad M \in \mathbb{N},$$

where $S_n : \mathcal{H} \to \mathbb{R}^n$ is the sampling operator and $S_{n,M} = P_M S_n$. The classical example of projection regularization method is truncated singular values decomposition (TSVD), aka principal component regression (PCR) in statistics. In our context, this amounts to consider the spaces $\mathcal{H}_M$ to be the span of the first M eigenfunctions of the operator $S_n^* S_n$. This latter method is essentially known to be optimal among projection regularization methods and indeed can also be analyzed in the context of supervised learning, with an analysis following the one for Tikhonov regularization. Without entering into details, crucial quantities in this analysis are

$$\left\| S_{n,M}^{\dagger} \right\|, \qquad \text{and} \qquad \left\| (I - P_M) S_{\rho}^* S_{\rho} \right\|. \tag{44}$$

The first term appears in the analysis of the sample error and controls stability, whereas the second term appears in the control of the approximation error. Choosing the eigenfunctions of $S_n^* S_n$ to build the spaces $\mathcal{H}_1 \subset \mathcal{H}_2 \dots$ ensures the sharpest control of the above quantities [6].

Furthermore, it is possible to combine projections with other regularization techniques. In our context, this corresponds to considering the minimization problem

$$\min_{f \in \mathcal{H}} \left(\| S_n P_M f - \mathbf{y} \|_n^2 + \lambda \| f \|_{\mathcal{H}}^2 \right) \qquad \lambda > 0, \tag{45}$$

whose minimizer is given by

$$f_n^{\lambda, M} = (P_M S_n^* S_n P_M + \lambda I)^{-1} P_M S_n^* \mathbf{y}.$$

The above scheme is common inverse problems where the interplay between M and λ is known to be crucial to obtain good error bounds [22].

We next introduce a different approximation often referred to as Nÿstrom approximations.

6.2 Nÿstrom Approximations

We begin noting that in light of Lemma 23, Problem (21) can be written as

$$\min_{f \in \mathcal{H}_n} \| S_n f - \mathbf{y} \|_n^2 + \lambda \| f \|_{\mathcal{H}}^2, \tag{46}$$

where

$$\mathcal{H}_n = \operatorname{span}\{K_{x_1}, \dots, K_{x_n}\}.$$

The basic idea of Nÿstrom approximations is to consider a set of *centers* $\tilde{x}_1, \dots, \tilde{x}_M$ sampled from $x_1, \dots, x_n$ according to some distribution, e.g., uniformly at random and then to consider,

$$\min_{f \in \mathcal{H}_M} \left(\| S_n f - \mathbf{y} \|_n^2 + \lambda \| f \|_{\mathcal{H}}^2 \right) \qquad \lambda > 0, \tag{47}$$

where

$$\mathcal{H}_M = \operatorname{span}\{K_{\tilde{x}_1}, \dots, K_{\tilde{x}_M}\}.$$

An obvious, yet important observation, is that we are not subsampling the training set, but only considering a smaller set of inputs to build a function space.

It is easy to check that the solution of Problem (47) is given by

$$f_n^{\lambda,M}(x) = \sum_{i=1}^{M} \alpha_i K(\tilde{x}_i, x), \quad \alpha = (K_{n,M}^\top K_{n,M} + \lambda n K_M)^\dagger K_{n,M}^\top \mathbf{y}, \tag{48}$$

where $K_{n,M}$ and K_M are $n \times M$ and $M \times M$ matrices, respectively, with entries

$$(K_{n,M})_{i,j} = K(x_i, \tilde{x}_j) \qquad (K_M)_{i,j} = K(\tilde{x}_i, \tilde{x}_j).$$

In this case, the complexity becomes $O(nM)$ in space, and $O(nMd + nM^2 + M^3)$ in time, which again can be much lower than standard Tikhonov regularization if $M \ll n$.

As clear from the above derivation, Nÿstrom approximations hold for any kernel. A common way to asses the accuracy of such an approximation is to consider

$$\tilde{K}_M = K_{n,M}^\top K_M^\dagger K_{n,M},$$

and analyzing

$$\left\| K_n - \tilde{K}_M \right\|.$$

Indeed, considering $\tilde{K}_M$ in place of K_n is related to the Nÿstrom approximation used to discretize integral equations and it is the reason for the name of these class of approximations. While interesting the above reasoning does not yield any direct insight on the effect of Nÿstrom approximation in terms of prediction accuracy. Before analyzing this we discuss the connection between Nÿstrom approximations and classical regularization techniques in inverse problems.

6.2.1 Regularization with Stochastic Projections

The interpretation of Tikhonov regularization with Nÿstrom approximation as a form of regularization with stochastic projections rests on the following lemma.

Lemma 32 *Problem* (47) *is equivalent to*

$$\min_{f \in \mathcal{H}} \left\| S_n P_M f - \mathbf{y} \right\|_n^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2 \tag{49}$$

and the solution of both minimization problems is

$$f_n^{\lambda,M} = (P_M S_n^* S_n P_M + \lambda I)^{-1} P_M S_n^* \mathbf{y}, \tag{50}$$

with P_M the projection operator with range $\mathcal{H}_M$.

Proof Note that Problem (47) and Problem (49) are strictly convex and coercive, therefore they admit a unique solution that is denoted by $f_n^{\lambda,M}$ and $g_n^{\lambda,M}$, respectively. To show that $f_n^{\lambda,M} = g_n^{\lambda,M}$, let $g_n^{\lambda,M} = a + b$ with $a \in \mathcal{H}_M$ and $b \in \mathcal{H}_M^\perp$. A necessary

condition for $g_n^{\lambda,M}$ to be optimal is that $b = 0$, indeed, considering that $P_M b = 0$, we have

$$\begin{aligned}\left\|S_n P_M g_n^{\lambda,M} - \mathbf{y}\right\|_n^2 + \lambda \left\|g_n^{\lambda,M}\right\|_{\mathcal{H}}^2 &= \left\|S_n P_M a - \mathbf{y}\right\|_n^2 + \lambda \|a\|_{\mathcal{H}}^2 + \lambda \|b\|_{\mathcal{H}}^2 \\ &\geq \left\|S_n P_M a - \mathbf{y}\right\|_n^2 + \lambda \|a\|_{\mathcal{H}}^2.\end{aligned}$$

This means that $g_n^{\lambda,M} \in \mathcal{H}_M$, but on $\mathcal{H}_M$ the functionals defining Problem (47) and Problem (49) are identical because $P_M f = f$ for any $f \in \mathcal{H}_M$ and so $f_n^{\lambda,M} = g_n^{\lambda,M}$. Therefore, by computing the gradient of the objective function in Problem (49), we see that $f_n^{\lambda,M}$ is given by Eq. (50). □

The above result shows that indeed Tikhonov regularization with Nÿstrom approximation is a special form of regularization with projection, where the projections are stochastic.

6.3 Error Bounds

We next discuss error bounds for Nyström approximations [40]. The analysis in this section is a simplified version of the one in [40].

Theorem 33 *Assume that for some* $0 < r \leq 1/2$

$$f_{\mathcal{H}}^{\dagger} \in \text{ran}\,(S_\rho^* S_\rho)^r. \tag{51}$$

Select M *points* $\tilde{x}_1,\ldots,\tilde{x}_M$ *uniformly without replacement. Let* $f^{\lambda,M}$ *be defined as in* (48) *and, fix* $\tau \geq 1$, *choose*

$$\lambda_n = \left(\frac{\tau^2}{n}\right)^{\frac{1}{2(r+1)}}, \qquad M_n = \frac{C_0}{\lambda_n}(\tau + \log \tfrac{c_2}{\lambda_n}).$$

With probability at least $1 - Ce^{-\tau}$,

$$\|f^{\lambda_n,M_n} - f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \leq c\tau^2 n^{-\frac{r}{2(r+1)}},$$

provided that $n \geq c_0 + c_1\tau^2$, *where the constants* c_0, c_1, c_2, c, C_0, C *are independent of* n, τ.

The obtained bound is again statistically optimal. It is interesting to see the behavior of M_n. Note that M_n is always smaller than $\sqrt{n}$ and decreases for increasing r from $\sqrt{n}$ (if r goes to 0), to $\sqrt[3]{n}$ (if $r = 1/2$). This shows the problem is "easy" the computational complexity can be dramatically reduced without incurring in any loss in accuracy.

We now prove the above result. First we provide an algebraic decomposition of the error as follows.

Theorem 34 *Let $\lambda > 0$. The following bound holds:*

$$\begin{aligned}\|f_n^{\lambda,M} - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \;\leq\; & \frac{1}{\lambda}\|S_n^*\mathbf{y} - S_n^*S_n f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \\ & + \left(2 + \frac{1}{\sqrt{\lambda}}\|S_n(I-P_M)\|_\infty\right)\|(I-P_M)f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \\ & + \sqrt{\lambda}\|(S_n^*S_n + \lambda I)^{-1/2} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}.\end{aligned}$$

Proof To derive bounds for Nyström method, we will consider the following decomposition:

$$\begin{aligned} f_n^{\lambda,M} - f_{\mathcal{H}}^\dagger = {} & (P_M S_n^* S_n P_M + \lambda I)^{-1} P_M (S_n^*\mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger) && (52)\\ & + (P_M S_n^* S_n P_M + \lambda I)^{-1} P_M S_n^* S_n (I - P_M) f_{\mathcal{H}}^\dagger && (53)\\ & + ((P_M S_n^* S_n P_M + \lambda I)^{-1} P_M S_n^* S_n P_M - I) P_M f_{\mathcal{H}}^\dagger && (54)\\ & - (I - P_M) f_{\mathcal{H}}^\dagger && (55)\end{aligned}$$

that holds since $P_M^2 = P_M$. Denote by $T_{n,M}$ the operator $T_{n,M} = P_M S_n^* S_n P_M$, and by $T_{n,M,\lambda}$ the operator $T_{n,M,\lambda} = T_{n,M} + \lambda I$, and note that $\|T_{n,M,\lambda}^{-1}\|_\infty \leq \lambda^{-1}$ and $\|T_{n,M,\lambda}^{-1/2}\|_\infty \leq \lambda^{-1/2}$. The first term (52) controls the variance of the estimator and is bounded by

$$\begin{aligned}\|T_{n,M,\lambda}^{-1} P_M (S_n^*\mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger)\|_{\mathcal{H}} &\leq \|T_{n,M,\lambda}^{-1}\|_\infty \|P_M\|_\infty \|S_n^*\mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \\ &\leq \lambda^{-1}\,\|S_n^*\mathbf{y} - S_n^* S_n f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}.\end{aligned}$$

The second term (53) depends on how well the projected regularization approximates S_n and $f_{\mathcal{H}}^\dagger$. To bound it, observe that

$$\|T_{n,M,\lambda}^{-1} P_M S_n^* S_n (I - P_M) f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq \|T_{n,M,\lambda}^{-1/2}\|_\infty \|T_{n,M,\lambda}^{-1/2} P_M S_n^*\|_\infty \|S_n(I - P_M) f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}.$$

Now by (68) we have

$$\|T_{n,M,\lambda}^{-1/2} P_M S_n^*\|_\infty^2 = \|T_{n,M,\lambda}^{-1/2} T_{n,M} T_{n,M,\lambda}^{-1/2}\|_\infty \leq 1.$$

Moreover, since $I - P_M$ is a projection operator, we have

$$\|S_n(I - P_M) f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} = \|S_n(I - P_M)^2 f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq \|S_n(I - P_M)\|_\infty \|(I - P_M) f_{\mathcal{H}}^\dagger\|_{\mathcal{H}},$$

so that

$$\|T_{n,M,\lambda}^{-1} P_M S_n^* S_n (I - P_M) f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq \lambda^{-\frac{1}{2}} \|S_n(I - P_M)\|_\infty \|(I - P_M) f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}.$$

The third term (54) controls the bias of the estimator. To bound it, by (72) we have

$$\begin{aligned}\|(T_{n,M,\lambda}^{-1}T_{n,M}-I)P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} &= \lambda\|T_{n,M,\lambda}^{-1}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}\\ &\le \lambda\|T_{n,M,\lambda}^{-1/2}\|_{\infty}\|T_{n,M,\lambda}^{-1/2}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \le \lambda^{1/2}\|T_{n,M,\lambda}^{-1/2}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}.\end{aligned}$$

Moreover, by (73)

$$\begin{aligned}\|T_{n,M,\lambda}^{-1/2}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}^2 &= \lambda\left\langle f_{\mathcal{H}}^{\dagger}, P_M(P_M S_n^* S_n P_M+\lambda I)^{-1}P_M f_{\mathcal{H}}^{\dagger}\right\rangle\\ &\le \left\langle f_{\mathcal{H}}^{\dagger}, P_M(S_n^* S_n+\lambda I)^{-1}P_M f_{\mathcal{H}}^{\dagger}\right\rangle\\ &= \|(S_n^* S_n+\lambda I)^{-1/2}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}^2.\end{aligned}$$

Hence, we get

$$\|(T_{n,M,\lambda}^{-1}T_{n,M}-I)P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \le \lambda^{\frac{1}{2}}\|(S_n^* S_n+\lambda I)^{-1/2}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}.$$

Finally, by $P_M f_{\mathcal{H}}^{\dagger} = (f_{\mathcal{H}}^{\dagger}-(I-P_M)f_{\mathcal{H}}^{\dagger})$ we have

$$\begin{aligned}\|(S_n^* S_n+\lambda I)^{-1/2}P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} &\le \|(S_n^* S_n+\lambda I)^{-1/2}f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}\\ &\quad+\|(S_n^* S_n+\lambda I)^{-1/2}\|_{\infty}\|(I-P_M)f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}},\end{aligned}$$

so that

$$\|(T_{n,M,\lambda}^{-1}T_{n,M}-I)P_M f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \le \lambda^{\frac{1}{2}}\|(S_n^* S_n+\lambda I)^{-1/2}f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}+\|(I-P_M)f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}.$$

The theorem is concluded by combining the bounds derived above for the four terms. □

Analyzing the decomposition above, we see that the error associated to the Nyström estimator decomposes in one variance term, one term that accounts for the effect of the stochastic projection on S_n and on $f_{\mathcal{H}}^{\dagger}$, and a third term that resembles a bias term. We need the following lemma.

Lemma 35 *Take $\lambda > 0$ and $\tau > 0$, with probability at least $1-4e^{-\tau}$.*

$$\|(S_n^* S_n+\lambda I)^{-1/2}f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \le \sqrt{2}\|(S_\rho^* S_\rho+\lambda I)^{-1/2}f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}$$

provided that

$$n \ge 4c^2\tau^2\lambda^{-2}, \tag{56}$$

where c is a suitable constant.

Proof Taking into account (56), bound (36) implies that, with probability at least $1-4e^{-\tau}$,

$$\|S_n^* S_n - S_\rho^* S_\rho\|_\infty \le \lambda/2 < \lambda. \tag{57}$$

By (75) with $A = S_n^* S_n + \lambda I$ and $B = S_\rho^* S_\rho + \lambda I$, we get with the same probability

$$\|(S_n^* S_n + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty \le \sqrt{\frac{1}{1 - \lambda^{-1}\|S_n^* S_n - S_\rho^* S_\rho\|_\infty}} \le \sqrt{2}, \tag{58}$$

where (74) is satisfied taking into account that

$$\begin{aligned}\left\|(S_\rho^* S_\rho + \lambda I)^{-1/2}\left(S_n^* S_n - S_\rho^* S_\rho\right)(S_\rho^* S_\rho + \lambda I)^{-1/2}\right\|_\infty &\le \left\|(S_\rho^* S_\rho + \lambda I)^{-1}\right\|_\infty \left\|S_n^* S_n - S_\rho^* S_\rho\right\|_\infty \\ &\le \lambda^{-1}\left\|S_n^* S_n - S_\rho^* S_\rho\right\|_\infty \le \frac{1}{2}\end{aligned}$$

since $\left\|S_\rho^* S_\rho + \lambda I\right\|_\infty^{-1} \le \lambda^{-1}$ and (57). Hence,

$$\begin{aligned}\|(S_n^* S_n + \lambda I)^{-1/2} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} &= \|(S_n^* S_n + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}(S_\rho^* S_\rho + \lambda I)^{-1/2} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \\ &\le \|(S_n^* S_n + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty \|(S_\rho^* S_\rho + \lambda I)^{-1/2} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \\ &\le \sqrt{2}\|(S_\rho^* S_\rho + \lambda I)^{-1/2} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}.\end{aligned}$$

□

More advanced versions of the result above are in [40–42].

We need the following concentration inequality. Set

$$d_\infty(\lambda) = \sup_{x \in X} \left\langle (S_\rho^* S_\rho + \lambda I)^{-1} K_x, K_x \right\rangle.$$

Lemma 36 *Take $\tau > 0$ and $0 < \lambda \le \left\|S_\rho\right\|_\infty^2$, with probability at least $1 - 2e^{-\tau}$*

$$\left\|(S_\rho^* S_\rho + \lambda I)^{-1/2}(S_n^* S_n - S_\rho^* S_\rho)(S_\rho^* S_\rho + \lambda I)^{-1/2}\right\| \le \frac{2\beta(1 + d_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta d_\infty(\lambda)}{n}}, \tag{59}$$

where $\beta = \tau + \ln(8\lambda^{-1}\kappa^2)$.

Proof See Proposition 6 in [41]. □

Let $\tilde{x}_1, \ldots, \tilde{x}_M$ be the selected Nyström points and set

$$Z_M : \mathcal{H} \to \mathbb{R}^M \qquad (Z_M f)^i = \left\langle f, K_{\tilde{x}_i} \right\rangle \qquad i = 1, \ldots, M.$$

The result above is crucial to control the effect of the randomized projection, as shown in the next lemma.

Lemma 37 *Let $M \in \mathbb{N}$, $\lambda > 0$, $\tau \ge 1$. With probability at least $1 - 2e^{-\tau}$*

$$\|(I - P_M) S_\rho^*\|_\infty \le \sqrt{2\lambda},$$

provided that

$$M \geq \frac{13\kappa^2}{\lambda}(\tau + \log \tfrac{8\kappa^2}{\lambda}), \quad \lambda \leq \|S_\rho\|_\infty^2. \tag{60}$$

Proof Equation (71) implies that

$$\begin{aligned}\|(I - P_M)S_\rho^*\|_\infty &\leq \|(I - P_M)(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty \\ &= \|(I - P_M)(Z_M^* Z_M + \lambda I)^{1/2}(Z_M^* Z_M + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty \\ &\leq \|(I - P_M)(Z_M^* Z_M + \lambda I)^{1/2}\|_\infty \|(Z_M^* Z_M + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty.\end{aligned}$$

Now note that, since $\operatorname{ran} P_M = \operatorname{ran} Z_M^*$, then $(I - P_M)Z_M = 0$, so

$$\begin{aligned}\|(I - P_M)(Z_M^* Z_M + \lambda I)^{1/2}\|_\infty^2 &= \|(I - P_M)(Z_M^* Z_M + \lambda I)(I - P_M)\|_\infty \\ &\leq \|(I - P_M)Z_M^* Z_M (I - P_M)\|_\infty + \lambda \|(I - P_M)^2\|_\infty \\ &\leq \lambda.\end{aligned}$$

To conclude the proof we have to show that

$$\|(Z_M^* Z_M + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty \leq \sqrt{2}.$$

Indeed, by (75) with $A = (Z_M^* Z_M + \lambda I)^{-1/2}$, $B = (S_\rho^* S_\rho + \lambda I)^{1/2}$ and $\Delta = (Z_M^* Z_M + \lambda I)^{-1/2}(Z_M^* Z_M - S_\rho^* S_\rho)(S_\rho^* S_\rho + \lambda I)^{1/2}$, we have

$$\|(Z_M^* Z_M + \lambda I)^{-1/2}(S_\rho^* S_\rho + \lambda I)^{1/2}\|_\infty \leq \sqrt{\frac{1}{1 - \|\Delta\|_\infty}} \leq \sqrt{2},$$

provided that

$$\|\Delta\|_\infty = \|(S_\rho^* S_\rho + \lambda I)^{-1/2}(Z_M^* Z_M - S_\rho^* S_\rho)(S_\rho^* S_\rho + \lambda I)^{1/2}\| < \frac{1}{2}.$$

Since the Nyström points are uniformly selected without replacement, $\tilde{x}_1, \ldots, \tilde{x}_M$ are independently and identically distributed according to ρ_X, so that we can apply Lemma 36 by replacing S_n with Z_M. Hence, with probability at least $1 - 2e^{-\tau}$,

$$\|\Delta\|_\infty \leq \frac{2\beta(1 + d_\infty(\lambda))}{3M} + \sqrt{\frac{2\beta d_\infty(\lambda)}{M}} \leq \frac{1}{2},$$

with $\beta = \tau + \ln(8\lambda^{-1}\kappa^2)$, where the last bound is a consequence of (60), taking into account that by (9)

$$d_\infty(\lambda) \leq \kappa^2 \lambda^{-1} \qquad \|S_\rho\|^2 \leq \kappa^2.$$

Indeed, since $\kappa^2\lambda^{-1} \geq 1$ and $M \geq 13\beta\kappa^2/\lambda$,

$$\frac{2\beta(1+d_\infty(\lambda))}{3M}+\sqrt{\frac{\beta d_\infty(\lambda)}{M}}\le\frac{4\beta\kappa^2}{3M\lambda}+\sqrt{\frac{2\beta\kappa^2}{M\lambda}}\le\frac{1}{2}.$$

□

We need the following bound, which depends on the source condition.

Lemma 38 *Assume that* $f_{\mathcal{H}}^{\dagger}\in\operatorname{ran}(S_\rho^*S_\rho)^r$ *for some* $r\in(0,1/2]$, *then*

$$\|(I-P_M)f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}\le C\|(I-P_M)S_\rho^*\|_\infty^{2r}, \tag{61}$$

where C *is a constant depending on* ρ.

Proof By assumption, there exists $g\in\mathcal{H}$ such that $f_{\mathcal{H}}^{\dagger}=(S_\rho^*S_\rho)^r g$, then

$$\|(I-P_M)f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}=\|(I-P_M)(S_\rho^*S_\rho)^r g\|_{\mathcal{H}}\le C\|(I-P_M)(S_\rho^*S_\rho)^r\|_\infty,$$

where $C=\|g\|_{\mathcal{H}}$. By Cordes inequality (69) with $s=2r$, $A=(I-P_M)^s=(I-P_M)$ and $B=(S_\rho^*S_\rho)^{\frac{1}{2}}$

$$\begin{aligned}\|(I-P_M)(S_\rho^*S_\rho)^r\|_\infty&=\|(I-P_M)^s(S_\rho^*S_\rho)^{s/2}\|_\infty\le\|(I-P_M)(S_\rho^*S_\rho)^{1/2}\|_\infty^s\\&=\|(I-P_M)S_\rho^*\|_\infty^{2r}.\end{aligned}$$

□

Proof (*Theorem* 33) We bound the three terms in the analytic decomposition of $\|f_n^{\lambda,M}-f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}$ in Theorem 34. By (36), with probability at least $1-4e^{-\tau}$

$$\frac{1}{\lambda_n}\|S_n^*\mathbf{y}-S_n^*S_nf_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}\lesssim\frac{1}{\lambda_n\sqrt{n}}=n^{-\frac{r}{2(r+1)}},$$

since $\lambda_n=n^{-\frac{1}{2(r+1)}}$. To bound the second term, we need the following two considerations: first, by Lemma 37, selecting $C_0=13\kappa^2$ and $c_2=8\kappa^2$, we have

$$\|(I-P_M)S_\rho^*\|_\infty\le\sqrt{2\lambda},$$

with probability $1-2e^{-\tau}$. Second, we bound $\|(I-P_M)S_n^*\|_\infty$ as follows:

$$\begin{aligned}\|(I-P_M)S_n^*\|_\infty^2&=\|(I-P_M)S_n^*S_n(I-P_M)\|_\infty^2\\&\le\|(I-P_M)(S_n^*S_n-S_\rho^*S_\rho)(I-P_M)\|_\infty+\|(I-P_M)S_\rho^*S_\rho(I-P_M)\|_\infty\\&\le\|I-P_M\|_\infty^2\|S_n^*S_n-S_\rho^*S_\rho\|_\infty+\|(I-P_M)S_\rho^*\|_\infty^2\\&\le\frac{c\tau}{\sqrt{n}}+2\lambda,\end{aligned}$$

with probability $1-4e^{-\tau}$, where for the last step we used (36) and the fact that $\|I-P_M\|_\infty\le 1$ since $I-P_M$ is a projection operator. Now, note that, by (61) and the fact that $\lambda_n\sqrt{n}\ge 1$,

$$\left(2+\frac{1}{\sqrt{\lambda}}\|S_n(I-P_M)\|_\infty\right)\|(I-P_M)f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \le C\left(2+\frac{1}{\sqrt{\lambda}}\|S_n(I-P_M)\|_\infty\right)\|(I-P_M)S_\rho^*\|_\infty^{2r}$$
$$\le C(2+\sqrt{2+c\tau})\lambda_n^r,$$

where the last bound holds true with probability at least $1-4e^{-\tau}$. Finally, by Lemma 35 with probability at least $1-4e^{-\tau}$.

$$\sqrt{\lambda_n}\|(S_n^*S_n+\lambda I)^{-1/2}f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \le \sqrt{2\lambda_n}\|(S_\rho^*S_\rho+\lambda I)^{-1/2}f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}$$
$$= \sqrt{2\lambda_n}\|(S_\rho^*S_\rho+\lambda I)^{-1/2}(S_\rho^*S_\rho)^r g\|_{\mathcal{H}} \le \sqrt{2}\|g\|_{\mathcal{H}}\lambda_n^r,$$

where the last bound has the same proof of (33) and condition (56) is satisfied. □

6.4 Regularization by Subsampling

A useful observation is derived exchanging λ and M, that is, considering

$$M_n = \tilde{O}(n^{\frac{1}{2(r+1)}}), \qquad \lambda_n = \tilde{O}(\frac{1}{M_n}).$$

Clearly, in this case, the same bound holds. However, we now naturally think of M as a regularization parameter rather than a parameter controlling the computational budget. This shows that M can be used to control at the same time the statistical time and space complexity of the obtain solution.

Remark 39 (*Regularization path*) An advantage of parameterizing the algorithm by M is that an easy incremental implementation can be considered. Indeed, it can be shown that the solution computed for some set of centers can be efficiently updated if one center is added. This suggests to build a regularization path by computing solutions corresponding to an increasing number of centers via incremental updates.

6.5 Random Features

The approach of random features is based on the following basic idea [26, 38, 41, 50].

Recall that for finite dictionaries of Example 11, the kernel can be written as

$$K_M(x,x') = \Phi_M(x)^\top \Phi_M(x'), \tag{62}$$

where $\Phi_M : \mathcal{X} \to \mathbb{R}^M$ is the feature map. We can identify $\mathcal{H}_M$ with $\mathbb{R}^M$ since for any $f \in \mathcal{H}_M$ there exists a vector $w \in \mathbb{R}^M$ such that

$$f(x) = \Phi_M(x)^t w \qquad x \in \mathcal{X}.$$

With this identification,

$$f_n^\lambda(x) = \Phi(x)^t w_n^\lambda \qquad w_n^\lambda = (S_{n,M}^t S_{n,M} + \lambda I)^{-1} S_n^t \mathbf{y},$$

where $S_{n,M}$ can be seen as the n by M matrix defined as

$$S_{n,M} = \begin{pmatrix} \Phi_M(x_1)^t \\ \dots \\ \Phi_M(x_n)^t \end{pmatrix}.$$

In this case, the complexity becomes $O(nM)$ in space, and $O(nMd + nM^2 + M^3)$ in time, which can be much lower than standard Tikhonov regularization if $M \ll n$. Despite the simplicity of this approach, considering kernels of the form (62) can be too much of a restriction. Indeed, classic examples of kernels such as the Gaussian kernel $e^{-\left\|x-x'\right\|^2\gamma}$ do not satisfy (62). It is then natural to ask if the above reasoning can still be useful to reduce the computational burden for more complex kernels such as the Gaussian kernel.

Random features provide one possible approach. The basic idea is to relax Eq. (62) assuming it holds only approximately, that is,

$$K(x, x') \approx \Phi_M(x)^\top \Phi_M(x'). \tag{63}$$

Clearly, if one such approximation exists the approach described in the previous section can still be used. The question is then for which kernels an approximation of the form (63) can be derived. A simple manipulation of the Gaussian kernel of Example 12 provides one basic example.

Example 40 (*Random Fourier features*) Using basic properties of the Fourier transform, it is easy to see that

$$e^{-\left\|x-x'\right\|^2\gamma} = \left(\frac{1}{2\sqrt{\pi\gamma}}\right)^d \int_{\mathbb{R}^d} e^{-\frac{\left\|\omega\right\|^2}{4\gamma}} e^{i\omega^\top x} e^{-i\omega^\top x} d\omega,$$

and the above expression can be further simplified considering

$$e^{-\left\|x-x'\right\|^2\gamma} = \int_0^1 \left(\int_{\mathbb{R}^d} \cos(\omega^\top x + b)\cos(\omega^\top x + b) \quad \frac{1}{(\sqrt{2\pi(2\gamma)})^d} e^{-\frac{\left\|\omega\right\|^2}{2(2\gamma)}} d\omega \right) db.$$

Then the idea is to view the above integral as an expectation and consider a Monte Carlo approximation

$$e^{-\left\|x-x'\right\|^2\gamma} \simeq \frac{1}{M}\sum_{j=1}^{M}\cos(\omega_j^\top x + b_j)\cos(\omega_j^\top x + b_j),$$

where $(\omega_j, b_j)_{j=1}^M$ are independent samples of the probability distribution which is the product of a Gaussian distribution with variance 2γ and the uniform distribution over $[0, 1]$. Then we can define the feature map

$$\Phi_M(x) = \frac{1}{\sqrt{M}}(\cos(\omega_1^\top x + b_M), \ldots, \cos(\omega_M^\top x + b_M)).$$

The above example can be abstracted to a general approximation strategy. Assume that the kernel K has an integral representation

$$K(x, x') = \int_\Omega \psi(x, \omega)\psi(x', \omega)d\pi(\omega), \quad \forall x, x' \in \mathcal{X}, \tag{64}$$

where (Ω, π) is probability space and $\psi : \mathcal{X} \times \Omega \to \mathbb{R}$. The random feature approach consists in approximating K via Monte Carlo sampling of its integral representation: given $\omega_1, \ldots, \omega_M$ independently and identically distributed according to the probability distribution π, consider

$$K_M(x, x') := \frac{1}{M}\sum_{j=1}^{M}\psi(x, \omega_j)\psi(x', \omega_j) = \Phi_M(x)^\top \Phi_M(x'), \tag{65}$$

with $\Phi_M(x) := M^{-1/2}\,(\psi(x, \omega_1), \ldots, \psi(x, \omega_M))$.

As discussed before, this leads to the following learning algorithm. Let $\lambda > 0$ and $M \in \mathbb{N}$, for any $x \in \mathcal{X}$, $f_n^{\lambda,M}$ is defined as

$$f_n^{\lambda,M}(x) := \Phi_M(x)^\top w_n^{\lambda,M}, \quad \text{with} \quad w_n^{\lambda,M} := (S_{n,M}^\top S_{n,M} + \lambda I)^{-1} S_{n,M}^\top \mathbf{y}, \tag{66}$$

where $S_{n,M}$ is the n by M matrix with rows $(\Phi_M(x_1), \ldots, \Phi_M(x_n))$.

The above discussion shows that the random features approach, based on deriving a representation (64) and then approximating it via (65), is applicable to a wide range of kernels, e.g., all translation-invariant kernels, and can be used to reduce the computational costs as soon as we choose $M \ll n$. However, since using random features rely on an approximation (63), it is natural to ask whether the approach leads to a loss of accuracy. A possible way to tackle this question is to characterize the error incurred replacing a kernel K with an approximation K_M, and indeed results in this direction abound. A more compelling question for supervised learning is whether using random features leads to computational advantages at the expanses of a decrease in prediction performances.

Note that the random feature algorithm $f_n^{\lambda,M}$ defined by (66) is the solution of the minimization problem

$$\min_{w\in\mathbb{R}^M} \left(\left\| S_{n,M} w - \mathbf{y} \right\|_n^2 + \lambda \left\| w \right\|_{\mathbb{R}^M}^2 \right). \tag{67}$$

The following lemma shows that random features algorithm can be seen as a special form of regularization with projection, where the projections are stochastic. To this aim, we assume that there exists a Hilbert space $\mathcal{H}$ such that $\Phi(\cdot, \omega) \in \mathcal{H}$ for all $\omega \in \Omega$ and, for any sample $(\omega_1, \ldots, \omega_M) \in \Omega^M$, we set $\psi_i = \psi(\cdot, \omega_i)/\sqrt{M}$ for $i = 1, \ldots, M$ and

$$\mathcal{H}_M = \operatorname{span}\{\psi_1, \ldots, \psi_M\}.$$

Note that, random features might not belong to the RKHS defined by the kernel being approximated. It suffices to think of Fourier random features and the Gaussian kernel. Here, we make this assumption to illustrate a difference in the form of the estimator considered while using random features and Nyström approximations. In the rest of the section, we restrict to the case of finite dictionary of Example 11. In this case, $\mathcal{H}_M$ is a subspace of $\mathcal{H}$, and we denote $J_M : \mathcal{H}_M \to \mathcal{H}$ the inclusion, moreover $\mathcal{H}_M$ is the RKHS with reproducing kernel K_M given by (65). We stress that, in general, $\| J_M f \|_{\mathcal{H}} \neq \left\| f \right\|_{\mathcal{H}_M}$.

Lemma 41 *Problem (67) is equivalent to*

$$\min_{f\in\mathcal{H}} \left(\left\| S_n P_M f - \mathbf{y} \right\|_n^2 + \lambda \left\| J_M^\dagger f \right\|_{\mathcal{H}_M}^2 \right),$$

where $S_n : \mathcal{H} \to \mathbb{R}^n$ is the sampling operator and P_M is the projection onto $\mathcal{H}_M$, regarded as closed subspace of $\mathcal{H}$.

Proof Let U_M be the operator

$$U_M : \mathbb{R}^M \to \mathcal{H}_M, \quad U_M w = \sum_{j=1}^M w^j \psi_i, \quad \forall w \in \mathbb{R}^M,$$

then U_M is a partial isometry with kernel $\ker U_M = \ker S_{n,M}$ and range $\operatorname{ran} U_M = \mathcal{H}_M$ and $S_{n,M} = S_n J_M U_M$. Hence, Problem (67) is equivalent to

$$\begin{aligned}
\min_{w\in\mathbb{R}^M} \left(\left\| S_{n,M} w - \mathbf{y} \right\|_n^2 + \lambda \left\| w \right\|_{\mathbb{R}^M}^2 \right) &= \min_{w\in\ker U_M{}^\perp} \left(\left\| S_{n,M} w - \mathbf{y} \right\|_n^2 + \lambda \left\| w \right\|_{\mathbb{R}^M}^2 \right) \\
&= \min_{g\in\mathcal{H}_M} \left(\left\| S_{n,M} U_M^* g - \mathbf{y} \right\|_n^2 + \lambda \left\| g \right\|_{\mathcal{H}_M}^2 \right) \\
&= \min_{g\in\mathcal{H}_M} \left(\left\| S_n J_M g - \mathbf{y} \right\|_n^2 + \lambda \left\| g \right\|_{\mathcal{H}_M}^2 \right) \\
&= \min_{f\in\mathcal{H}} \left(\left\| S_n J_M J_M^\dagger f - \mathbf{y} \right\|_n^2 + \lambda \left\| J_M^\dagger f \right\|_{\mathcal{H}_M}^2 \right) \\
&= \min_{f\in\mathcal{H}} \left(\left\| S_n P_M f - \mathbf{y} \right\|_n^2 + \lambda \left\| J_M^\dagger f \right\|_{\mathcal{H}_M}^2 \right),
\end{aligned}$$

where the last two inequalities are consequence of the fact that ran $J_M^\dagger = \mathcal{H}_M$ and $J_M J_M^\dagger = P_M$. □

Note that Lemma 41 is not strictly convex, so that set of minimizer is not a singleton, however its minimal norm solution is the minimizer of Problem (67).

The above result shows that Tikhonov regularization with random features is indeed a form of projection regularization but where we consider a special regularizer $\|J_M^\dagger\|_{\mathcal{H}_M}$ depending on the random projection. An analysis random features in full generality is in [41]. Essentially the same bounds hold for random features [41], but obtained with different techniques.

7 Conclusions

In this chapter, we reviewed the connection of supervised learning with the square loss and inverse problems. Then, we used the obtained formulation to provide a unifying description of the computational and statistical properties of a different regularization technique. In particular, we contrasted variational regularization (Tikhonov regularization), with iterative regularization and regularization with stochastic regularization.

The purpose of this discussion was to highlight how all these techniques share common estimation principles, and indeed have very similar statistical properties, but different computational properties.

- Variational regularization introduces a dichotomy between statistics and computations/optimization.
- Iterative regularization controls at once time and statistical complexities. Optimization and statistics are seen as aspects of a common underlying problem.
- Finally, regularization with stochastic projections allows to deal simultaneously with time, statistical, and space complexities.

We conclude with a few remarks.

- **Extensions to other loss/regularizers**. It is straightforward for variational regularization. It is a subject of study for iterative regularization where several extensions can be made. It is an open problem for regularization with projections.
- **A regularization view on optimization**. The material presented in this chapter proposes regularization theory to study optimization and numerical algorithms to solve statistical problems. This latter framework allows to consider at the same time the computational and stability properties of the considered procedures.
- **From supervised learning to inverse problems and beyond.** Drawing a connection between machine learning and inverse problems, allowed to exploit ideas, algorithms, and results from inverse problems in the context of machine learning. It would be interesting to investigate the opposite direction, and see if machine learning ideas could be applied in classical inverse problems arising in signal processing, PDE, and integral equations.

Acknowledgements This chapter evolved from a set of notes for the Summer School "Applied Harmonic Analysis and Machine Learning" in Genova, as well as the Summer School "Structured Regularization for High-Dimensional Data Analysis" at IHP Paris, and the Spring School "Structural Inference" Malente, Germany. The results we present are an overview of the work with and of a number of collaborators and colleagues. Among others, we would like to thank Gilles Blanchard, Andrea Caponnetto, Luigi Carratino, Raffaello Camoriano, Junhong Lin, Nicole Mücke, Nicoló Pagliana, Steve Smale, Alessandro Verri, Silvia Villa and Ding-Xuan Zhou.
This material is based on work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPUs and the Tesla k40 GPU used for this research. L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-18-1-7009, FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS-DLV-777826. Part of this work was funded by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d' avenir program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). A. R. acknowledges the financial support of the European Research Council (grant REAL 947908). E.D.V. is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

Appendix

In this section, we collect the notation and some basic mathematical facts, adapted to our setting.

$\mathcal{H}, \mathcal{G}$	Hilbert spaces
$\|f\|_{\mathcal{H}}$	norm of $f \in \mathcal{H}$
$\langle f, g\rangle_{\mathcal{H}}$	scalar product between $f, h \in \mathcal{H}$
$\mathcal{K}^{\perp} = \{f \in \mathcal{H} : \langle f, f'\rangle_{\mathcal{H}} = 0 \, \forall f' \in \mathcal{K}\}$	complement of a closed subspace $\mathcal{K} \subset \mathcal{H}$
$P_{\mathcal{K}}$	projection onto a closed subspace $\mathcal{K} \subset \mathcal{H}$
$\ker A = \{f \in \mathcal{H} : Af = 0\}$	kernel of $A : \mathcal{H} \to \mathcal{G}$
$\operatorname{ran} A = \{Af \in \mathcal{G} : f \in \mathcal{H}\}$	range of $A : \mathcal{H} \to \mathcal{G}$
$A^* : \mathcal{G} \to \mathcal{H}$	adjoint of $A : \mathcal{H} \to \mathcal{G}$
$\lvert A\rvert = \sqrt{A^*A}$	absolute value of $A : \mathcal{H} \to \mathcal{G}$
$\|A\|_{\infty} = \sup_{f \in \mathcal{H}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{H}}}$	operator norm of $A : \mathcal{H} \to \mathcal{G}$
$w \otimes v = \langle \cdot, v\rangle_{\mathcal{H}} w : \mathcal{H} \to \mathcal{G}$	rank-one operator $v \in \mathcal{H}, w \in \mathcal{G}$
$(\Omega, \mathcal{F}, \mathbb{P})$	probability space
$\mathbb{E}[\xi]$	expectation of the random variable $\xi : \Omega \to \mathcal{H}$

We recall the following results.

- Trace class operators: an operator $A : \mathcal{H} \to \mathcal{H}$ is trace class if there exists (any) a base $(v_\ell)_{\ell \in \Lambda}$ of $\mathcal{H}$ such that the series $\sum_{\ell \in \Lambda} \langle |A| v_\ell, v_\ell\rangle_{\mathcal{H}}$ converges and

$$\operatorname{Tr} A = \sum_{\ell \in \Lambda} \langle A v_\ell, v_\ell \rangle_{\mathcal{H}}$$

is called the trace of A. An $A : \mathcal{H} \to \mathcal{G}$ is trace class if $|A|$ is trace class and the space $\mathcal{S}_1(\mathcal{H}, \mathcal{G})$ of trace class operators from $\mathcal{H}$ to $\mathcal{G}$ is a Banach space with respect the norm

$$\|A\|_{\mathcal{S}_1(\mathcal{H},\mathcal{G})} = \operatorname{Tr}(|A|).$$

- Hilbert–Schmidt operators: $A : \mathcal{H} \to \mathcal{G}$ is a Hilbert–Schmidt operator if $A^* A$ is a trace class operator. The space $\mathcal{S}_2(\mathcal{H}, \mathcal{G})$ of Hilbert–Schmidt operators from $\mathcal{H}$ to $\mathcal{G}$ is a Hilbert space with respect to the scalar product

$$\langle A, B \rangle_{\mathcal{S}_2(\mathcal{H},\mathcal{G})} = \operatorname{Tr}(B^* A),$$

and $\mathcal{S}_1(\mathcal{H}, \mathcal{G}) \subset \mathcal{S}_2(\mathcal{H}, \mathcal{G})$.
- Hilbert–Schmidt theorem: if $A : \mathcal{H} \to \mathcal{H}$ is a Hilbert–Schmidt self-adjoint operator, there exists a base $(v_\ell)_{\ell \in \Lambda}$ of $\mathcal{H}$ and a sequence $(v_\ell)_{\ell \in \Lambda}$ of real numbers such that

$$A v_\ell = \sigma_\ell v_\ell.$$

- Functional calculus: if $A : \mathcal{H} \to \mathcal{H}$ is a Hilbert–Schmidt self-adjoint operator and $\varphi : \mathbb{R} \to \mathbb{R}$ is a bounded function, by the functional calculus the operator $\varphi(A) : \mathcal{H} \to \mathcal{H}$ is defined as

$$\varphi(A) f = \sum_{\ell \in \Lambda} \varphi(\sigma_\ell) \, \langle f, v_\ell \rangle_{\mathcal{H}} \, v_\ell.$$

where the series converges in $\mathcal{H}$.
- For any operator $A : \mathcal{H} \to \mathcal{G}$

$$\|A\|_\infty^2 = \|A A^*\|_\infty \tag{68}$$

- Cordes inequality: for any pair of positive operator $A, B : \mathcal{H} \to \mathcal{H}$ and $s \in (0, 1]$

$$\|A^s B^s\|_\infty \le \|AB\|_\infty^s \tag{69}$$

- Take two operators $B, C : \mathcal{H} \to \mathcal{G}$ such that

$$B B^* \le C C^*, \tag{70}$$

then

$$\|AB\|_\infty \le \|AC\|_\infty \tag{71}$$

for any operator $A : \mathcal{G} \to \mathcal{G}'$. Indeed, by (70)

$$ABB^*A^* \leq ACC^*A^* \qquad \Longrightarrow \qquad \left\|ABB^*A^*\right\|_\infty \leq \left\|ACC^*A^*\right\|_\infty.$$

Then,

$$\left\|AB\right\|_\infty^2 = \left\|ABB^*A^*\right\|_\infty \leq \left\|CBB^*C^*\right\|_\infty = \left\|AC\right\|_\infty^2.$$

- If $A : \mathcal{H} \to \mathcal{G}$ and $\lambda > 0$

$$((A^*A + \lambda I)^{-1}A^*A - I) = \lambda(A^*A + \lambda I)^{-1} \tag{72}$$

see [22].

- For any positive operator $C : \mathcal{H} \to \mathcal{H}$ and projection $P : \mathcal{H} \to \mathcal{H}$

$$P(PCP + \lambda I)^{-1}P \leq P(C + \lambda I)^{-1}P \tag{73}$$

see, for example, Theorem V.2.3-(iv) of [3] for the proof of this property of operator convex functions, and Cor. V.2.6 together with Exercise V.1.10-(ii) and Exercise V.2.11, in the same book, for the proof of operator convexity of $(\cdot + \lambda)^{-1}$.

- For any pair $A, B : \mathcal{H} \to \mathcal{H}$ of positive operators with bounded inverse such that

$$\left\|B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}}\right\|_\infty < 1 \tag{74}$$

then

$$\left\|A^{-\frac{1}{2}}B^{\frac{1}{2}}\right\|_\infty \leq \sqrt{\frac{1}{1 - \left\|B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}}\right\|_\infty}}. \tag{75}$$

Indeed, observe that

$$\begin{aligned}(A^{-\frac{1}{2}}B^{\frac{1}{2}})^*A^{-\frac{1}{2}}B^{\frac{1}{2}} &= B^{\frac{1}{2}}A^{-1}B^{\frac{1}{2}} = B^{\frac{1}{2}}(A - B + B)^{-1}B^{\frac{1}{2}} \\ &= B^{\frac{1}{2}}\left(B^{\frac{1}{2}}\left(B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}} + I\right)B^{\frac{1}{2}}\right)^{-1}B^{\frac{1}{2}} \\ &= \left(B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}} + I\right)^{-1} = \sum_{\ell=0}^{+\infty}(-1)^\ell\left(B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}}\right)^\ell,\end{aligned}$$

where the Neumann series converges by (74). Hence, triangular inequality gives

$$\left\|A^{-\frac{1}{2}}B^{\frac{1}{2}}\right\|_\infty^2 = \left\|\left(B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}} + I\right)^{-1}\right\|_\infty \leq \frac{1}{1 - \left\|B^{-\frac{1}{2}}(A - B)B^{-\frac{1}{2}}\right\|_\infty}.$$

- Höeffidng inequality in separable Hilbert spaces [35, 36, 51] : take a family

$$\xi_1, \ldots, \xi_n : \Omega \to \mathcal{H}$$

of independent zero mean random variables such that $\|\xi_i\|_{\mathcal{H}} \leq \kappa$, then for all $\epsilon > 0$

$$\mathbb{P}\left[\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\|_{\mathcal{H}} > \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{4\kappa^2}\right),$$

i.e., for all $\tau > 0$ with probability at least $1 - 2e^{-\tau}$

$$\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\|_{\mathcal{H}} \leq \frac{2\kappa\sqrt{\tau}}{\sqrt{n}}. \tag{76}$$

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**(3), 337–404 (1950). http://dx.doi.org/10.2307/1990404
2. Aymeric Dieuleveut, F.B.: Nonparametric stochastic approximation with large step-sizes. Ann. Stat. **44**(4), 1363–1399 (2016). https://doi.org/10.1214/15-AOS1391
3. Bathia, R., Johnson, C.R.: Matrix analysis. SIAM Rev. **40**(2), 413–413 (1998)
4. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. **7**, 2399–2434 (2006)
5. Bertero, M., De Mol, C., Pike, E.R.: Linear inverse problems with discrete data. i. general formulation and singular system analysis. Inverse Probl. **1**(4), 301 (1985)
6. Blanchard, G., Bousquet, O., Zwald, L.: Statistical properties of kernel principal component analysis. Mach. Learn. **66**(2–3), 259–294 (2007)
7. Blanchard, G., Krämer, N.: Optimal learning rates for kernel conjugate gradient regression. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada, pp. 226–234. Curran Associates, Inc. (2010). https://proceedings.neurips.cc/paper/2010/hash/b2f627fff19fda463cb386442eac2b3d-Abstract.html
8. Blanchard, G., Mücke, N.: Optimal rates for regularization of statistical inverse learning problems. Found. Comput. Math. **18**(4), 971–1013 (2018)
9. Bottou, L., Bousquet, O.: The Tradeoffs of Large Scale Learning. In: NIPS (2007)
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
11. Bühlmann, P., Yu, B.: Boosting with the l 2 loss: regression and classification. J. Am. Stat. Assoc. **98**(462), 324–339 (2003)
12. Caponnetto, A., De Vito, E.: Optimal rates for the regularized least-squares algorithm. Found. Comput. Math. **7**(3), 331–368 (2007)
13. Caponnetto, A., Yao, Y.: Cross-validation based adaptation for regularization operators in learning theory. Anal. Appl. **8**, 161–3183 (2010)
14. Coifman, R.R., Lafon, S.: Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. Appl. Comput. Harmon. Anal. **21**(1), 31–52 (2006)
15. Cucker, F., Smale, S.: On the mathematical foundation of learning. Am. Math. Soc. **39**(1), 1–49 (2001)
16. Cucker, F., Zhou, D.X.: Learning theory: an approximation theory viewpoint, vol. 24. Cambridge University Press (2007)
17. De Vito, E., Caponnetto, A., Rosasco, L.: Model selection for regularized least-squares algorithm in learning theory. Found. Comput. Math. **5**(1) (2005)

18. De Vito, E., Rosasco, L., Caponnetto, A.: Discretization error analysis for Tikhonov regularization. Anal. Appl. (Singap.) **4**(1), 81–99 (2006)
19. De Vito, E., Rosasco, L., Caponnetto, A., Giovannini, U.D., Odone, F.: Learning from examples as an inverse problem. J. Mach. Learn. Res. **6**(May), 883–904 (2005)
20. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition, vol. 31. Springer Science & Business Media (2013)
21. Dudley, R.: Real Analysis and Probability, Cambridge Studies in Advanced Mathematics, vol. 74. Cambridge University Press, Cambridge (2002). https://doi.org/10.1017/CBO9780511755347
22. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems, vol. 375. Springer (1996)
23. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser Basel (2013)
24. Györfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer Science & Business Media (2006)
25. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)
26. Hsu, D., Kakade, S.M., Zhang, T.: Random design analysis of ridge regression. J. Mach. Learn. Res.-Proc. Track **23**, 9–1 (2012)
27. Kress, R.: Linear Integral Equations, vol. 82. Springer Science & Business Media (1989)
28. Landweber, L.: An iteration formula for fredholm integral equations of the first kind. Am. J. Math. **73**, 615–624 (1951)
29. Lang, S.: Real and functional analysis, Graduate Texts in Mathematics, vol. 142, 3rd edn. Springer, New York (1993). https://doi.org/10.1007/978-1-4612-0897-6
30. Lin, J., Rosasco, L.: Optimal rates for multi-pass stochastic gradient methods. J. Mach. Learn. Res. **18**, 97:1–97:47 (2017). http://jmlr.org/papers/v18/17-176.html
31. Lin, J., Rudi, A., Rosasco, L., Cevher, V.: Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. Appl. Comput. Harmon. Anal. **48**(3), 868–890 (2020)
32. Minsker, S.: On some extensions of bernstein's inequality for self-adjoint operators. Stat. Probab. Lett. **127**, 111–119 (2017)
33. Mücke, N., Neu, G., Rosasco, L.: Beating SGD saturation with tail-averaging and minibatching. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 12,568–12,577 (2019). https://proceedings.neurips.cc/paper/2019/hash/4d0b954f0bef437c29dfa73fafdf3fa5-Abstract.html
34. Pagliana, N., Rosasco, L.: Implicit regularization of accelerated methods in hilbert spaces. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 14,454–14,464 (2019). https://proceedings.neurips.cc/paper/2019/hash/c61aed648da48aa3893fb3eaadd88a7f-Abstract.html
35. Pinelis, I.: Optimum bounds for the distributions of martingales in Banach spaces. Ann. Probab. **22**(4), 1679–1706 (1994)
36. Pinelis, I.: Correction: Optimum bounds for the distributions of martingales in Banach spaces. [Ann. Probab. **22**(4), 1679–1706 (1994); MR1331198 (96b:60010)]. Ann. Probab. **27**(4), 2119 (1999)
37. Poggio, T., Girosi, F.: Networks for approximation and learning. Proc. IEEE **78**(9), 1481–1497 (1990)
38. Rahimi, A., Recht, B.: Random Features for Large-Scale Kernel Machines. In: NIPS, pp. 1177–1184. Curran Associates, Inc. (2007)
39. Rosasco, L., Villa, S.: Learning with incremental iterative regularization. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015,

December 7-12, 2015, Montreal, Quebec, Canada, pp. 1630–1638 (2015). https://proceedings.neurips.cc/paper/2015/hash/1587965fb4d4b5afe8428a4a024feb0d-Abstract.html
40. Rudi, A., Camoriano, R., Rosasco, L.: Less is more: Nyström computational regularization. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 1648–1656. Curran Associates, Inc. (2015). http://papers.nips.cc/paper/5936-less-is-more-nystrom-computational-regularization.pdf
41. Rudi, A., Camoriano, R., Rosasco, L.: Generalization Properties of Learning with Random Features. (2016)
42. Rudi, A., Canas, G.D., Rosasco, L.: On the sample complexity of subspace learning. Adv. Neural. Inf. Process. Syst. **26**, 2067–2075 (2013)
43. Smale, S., Zhou, D.X.: Shannon sampling and function reconstruction from point values. Bull. Am. Math. Soc. **41**(3), 279–305 (2004)
44. Smale, S., Zhou, D.X.: Learning theory estimates via integral operators and their approximations. Constr. Approx. **26**(2), 153–172 (2007)
45. Steinwart, I., Christmann, A.: Support Vector Machines. Springer New York (2008). https://books.google.de/books?id=HUnqnrpYt4IC
46. Steinwart, I., Hush, D.R., Scovel, C., et al.: Optimal rates for regularized least squares regression. In: COLT, pp. 79–93 (2009)
47. Tropp, J.A.: User-friendly tools for random matrices: an introduction (2012)
48. Vapnik, V.N., Vapnik, V.: Statistical Learning Theory, vol. 1. Wiley, New York (1998)
49. Wahba, G.: Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia (1990)
50. Yen, E.H., Lin, T.W., Lin, S.D., Ravikumar, P.K., Dhillon, I.S.: Sparse Random Feature Algorithm as Coordinate Descent in Hilbert Space. In: Advances in Neural Information Processing Systems 27 (2014). http://papers.nips.cc/paper/5479-sparse-random-feature-algorithm-as-coordinate-descent-in-hilbert-space.pdf
51. Yurinsky, V.: Sums and Gaussian Vectors, vol. 1617. Springer, Berlin (1995)

Applied and Numerical Harmonic Analysis (104 Volumes)

1. A. I. Saichev and W. A. Woyczyński: *Distributions in the Physical and Engineering Sciences* (ISBN: 978-0-8176-3924-2)
2. C. E. D'Attellis and E. M. Fernandez-Berdaguer: *Wavelet Theory and Harmonic Analysis in Applied Sciences* (ISBN: 978-0-8176-3953-2)
3. H. G. Feichtinger and T. Strohmer: *Gabor Analysis and Algorithms* (ISBN: 978-0-8176-3959-4)
4. R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN: 978-0-8176-3918-1)
5. T. M. Peters and J. C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN: 978-0-8176-3941-9)
6. G. T. Herman: *Geometry of Digital Spaces* (ISBN: 978-0-8176-3897-9)
7. A. Teolis: *Computational Signal Processing with Wavelets* (ISBN: 978-0-8176-3909-9)
8. J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN: 978-0-8176-3963-1)
9. J. M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN: 978-0-8176-3967-9)
10. Procházka, N. G. Kingsbury, P. J. Payner, and J. Uhlir: *Signal Analysis and Prediction* (ISBN: 978-0-8176-4042-2)
11. W. Bray and C. Stanojevic: *Analysis of Divergence* (ISBN: 978-1-4612-7467-4)
12. G. T. Herman and A. Kuba: *Discrete Tomography* (ISBN: 978-0-8176-4101-6)
13. K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN: 978-0-8176-4022-4)
14. L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN: 978-0-8176-4104-7)
15. J. J. Benedetto and P. J. S. G. Ferreira: *Modern Sampling Theory* (ISBN: 978-0-8176-4023-1)
16. D. F. Walnut: *An Introduction to Wavelet Analysis* (ISBN: 978-0-8176-3962-4)

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8

17. A. Abbate, C. DeCusatis, and P. K. Das: *Wavelets and Subbands* (ISBN: 978-0-8176-4136-8)
18. O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN: 978-0-8176-4280-80)
19. H. G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN: 978-0-8176-4239-6)
20. O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN: 978-0-8176-4295-2)
21. L. Debnath: *Wavelets and Signal Processing* (ISBN: 978-0-8176-4235-8)
22. G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN: 978-0-8176-4279-2)
23. J. H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN: 978-0-8176-4331-7)
24. J. J. Benedetto and A. I. Zayed: *Sampling, Wavelets, and Tomography* (ISBN: 978-0-8176-4304-1)
25. E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN: 978-0-8176-4125-2)
26. L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN: 978-0-8176-3263-2)
27. W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN: 978-0-8176-4105-4)
28. O. Christensen and K. L. Christensen: *Approximation Theory* (ISBN: 978-0-8176-3600-5)
29. O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN: 978-0-8176-4354-6)
30. J. A. Hogan: *Time—Frequency and Time—Scale Methods* (ISBN: 978-0-8176-4276-1)
31. C. Heil: *Harmonic Analysis and Applications* (ISBN: 978-0-8176-3778-1)
32. K. Borre, D. M. Akos, N. Bertelsen, P. Rinder, and S. H. Jensen: *A Software-Defined GPS and Galileo Receiver* (ISBN: 978-0-8176-4390-4)
33. T. Qian, M. I. Vai, and Y. Xu: *Wavelet Analysis and Applications* (ISBN: 978-3-7643-7777-9)
34. G. T. Herman and A. Kuba: *Advances in Discrete Tomography and Its Applications* (ISBN: 978-0-8176-3614-2)
35. M. C. Fu, R. A. Jarrow, J.-Y. Yen, and R. J. Elliott: *Advances in Mathematical Finance* (ISBN: 978-0-8176-4544-1)
36. O. Christensen: *Frames and Bases* (ISBN: 978-0-8176-4677-6)
37. P. E. T. Jorgensen, J. D. Merrill, and J. A. Packer: *Representations, Wavelets, and Frames* (ISBN: 978-0-8176-4682-0)
38. M. An, A. K. Brodzik, and R. Tolimieri: *Ideal Sequence Design in Time-Frequency Space* (ISBN: 978-0-8176-4737-7)
39. S. G. Krantz: *Explorations in Harmonic Analysis* (ISBN: 978-0-8176-4668-4)
40. B. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN: 978-0-8176-4915-9)

41. G. S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 1* (ISBN: 978-0-8176-4802-2)
42. C. Cabrelli and J. L. Torrea: *Recent Developments in Real and Harmonic Analysis* (ISBN: 978-0-8176-4531-1)
43. M. V. Wickerhauser: *Mathematics for Multimedia* (ISBN: 978-0-8176-4879-4)
44. B. Forster, P. Massopust, O. Christensen, K. GrÃchenig, D. Labate, P. Vandergheynst, G. Weiss, and Y. Wiaux: *Four Short Courses on Harmonic Analysis* (ISBN: 978-0-8176-4890-9)
45. O. Christensen: *Functions, Spaces, and Expansions* (ISBN: 978-0-8176-4979-1)
46. J. Barral and S. Seuret: *Recent Developments in Fractals and Related Fields* (ISBN: 978-0-8176-4887-9)
47. O. Calin, D.-C. Chang, and K. Furutani, and C. Iwasaki: *Heat Kernels for Elliptic and Sub-elliptic Operators* (ISBN: 978-0-8176-4994-4)
48. C. Heil: *A Basis Theory Primer* (ISBN: 978-0-8176-4686-8)
49. J. R. Klauder: *A Modern Approach to Functional Integration* (ISBN: 978-0-8176-4790-2)
50. J. Cohen and A. I. Zayed: *Wavelets and Multiscale Analysis* (ISBN: 978-0-8176-8094-7)
51. D. Joyner and J.-L. Kim: *Selected Unsolved Problems in Coding Theory* (ISBN: 978-0-8176-8255-2)
52. G. S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN: 978-0-8176-4943-2)
53. J. A. Hogan and J. D. Lakey: *Duration and Bandwidth Limiting* (ISBN: 978-0-8176-8306-1)
54. G. Kutyniok and D. Labate: *Shearlets* (ISBN: 978-0-8176-8315-3)
55. P. G. Casazza and P. Kutyniok: *Finite Frames* (ISBN: 978-0-8176-8372-6)
56. V. Michel: *Lectures on Constructive Approximation* (ISBN : 978-0-8176-8402-0)
57. D. Mitrea, I. Mitrea, M. Mitrea, and S. Monniaux: *Groupoid Metrization Theory* (ISBN: 978-0-8176-8396-2)
58. T. D. Andrews, R. Balan, J. J. Benedetto, W. Czaja, and K. A. Okoudjou: *Excursions in Harmonic Analysis, Volume 1* (ISBN: 978-0-8176-8375-7)
59. T. D. Andrews, R. Balan, J. J. Benedetto, W. Czaja, and K. A. Okoudjou: *Excursions in Harmonic Analysis, Volume 2* (ISBN: 978-0-8176-8378-8)
60. D. V. Cruz-Uribe and A. Fiorenza: *Variable Lebesgue Spaces* (ISBN: 978-3-0348-0547-6)
61. W. Freeden and M. Gutting: *Special Functions of Mathematical (Geo-)Physics* (ISBN: 978-3-0348-0562-9)
62. A. I. Saichev and W. A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 2: Linear and Nonlinear Dynamics of Continuous Media* (ISBN: 978-0-8176-3942-6)
63. S. Foucart and H. Rauhut: *A Mathematical Introduction to Compressive Sensing* (ISBN: 978-0-8176-4947-0)

64. G. T. Herman and J. Frank: *Computational Methods for Three-Dimensional Microscopy Reconstruction* (ISBN: 978-1-4614-9520-8)
65. A. Paprotny and M. Thess: *Realtime Data Mining: Self-Learning Techniques for Recommendation Engines* (ISBN: 978-3-319-01320-6)
66. A. I. Zayed and G. Schmeisser: *New Perspectives on Approximation and Sampling Theory: Festschrift in Honor of Paul Butzer's 85th Birthday* (ISBN: 978-3-319-08800-6)
67. R. Balan, M. Begue, J. Benedetto, W. Czaja, and K. A. Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN: 978-3-319-13229-7)
68. H. Boche, R. Calderbank, G. Kutyniok, and J. Vybiral: *Compressed Sensing and its Applications* (ISBN: 978-3-319-16041-2)
69. S. Dahlke, F. De Mari, P. Grohs, and D. Labate: *Harmonic and Applied Analysis: From Groups to Signals* (ISBN: 978-3-319-18862-1)
70. A. Aldroubi: *New Trends in Applied Harmonic Analysis* (ISBN: 978-3-319-27871-1)
71. M. Ruzhansky: *Methods of Fourier Analysis and Approximation Theory* (ISBN: 978-3-319-27465-2)
72. G. Pfander: *Sampling Theory, a Renaissance* (ISBN: 978-3-319-19748-7)
73. R. Balan, M. Begue, J. Benedetto, W. Czaja, and K. A. Okoudjou: *Excursions in Harmonic Analysis, Volume 4* (ISBN: 978-3-319-20187-0)
74. O. Christensen: *An Introduction to Frames and Riesz Bases, Second Edition* (ISBN: 978-3-319-25611-5)
75. E. Prestini: *The Evolution of Applied Harmonic Analysis: Models of the Real World, Second Edition* (ISBN: 978-1-4899-7987-2)
76. J. H. Davis: *Methods of Applied Mathematics with a Software Overview, Second Edition* (ISBN: 978-3-319-43369-1)
77. M. Gilman, E. M. Smith, and S. M. Tsynkov: *Transionospheric Synthetic Aperture Imaging* (ISBN: 978-3-319-52125-1)
78. S. Chanillo, B. Franchi, G. Lu, C. Perez, and E. T. Sawyer: *Harmonic Analysis, Partial Differential Equations and Applications* (ISBN: 978-3-319-52741-3)
79. R. Balan, J. Benedetto, W. Czaja, M. Dellatorre, and K. A. Okoudjou: *Excursions in Harmonic Analysis, Volume 5* (ISBN: 978-3-319-54710-7)
80. I. Pesenson, Q. T. Le Gia, A. Mayeli, H. Mhaskar, and D. X. Zhou: *Frames and Other Bases in Abstract and Function Spaces: Novel Methods in Harmonic Analysis, Volume 1* (ISBN: 978-3-319-55549-2)
81. I. Pesenson, Q. T. Le Gia, A. Mayeli, H. Mhaskar, and D. X. Zhou: *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science: Novel Methods in Harmonic Analysis, Volume 2* (ISBN: 978-3-319-55555-3)
82. F. Weisz: *Convergence and Summability of Fourier Transforms and Hardy Spaces* (ISBN: 978-3-319-56813-3)
83. C. Heil: *Metrics, Norms, Inner Products, and Operator Theory* (ISBN: 978-3-319-65321-1)
84. S. Waldron: *An Introduction to Finite Tight Frames: Theory and Applications* (ISBN: 978-0-8176-4814-5)

85. D. Joyner and C. G. Melles: *Adventures in Graph Theory: A Bridge to Advanced Mathematics* (ISBN: 978-3-319-68381-2)
86. B. Han: *Framelets and Wavelets: Algorithms, Analysis, and Applications* (ISBN: 978-3-319-68529-8)
87. H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, and R. Mathar: *Compressed Sensing and Its Applications* (ISBN: 978-3-319-69801-4)
88. A. I. Saichev and W. A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 3: Random and Fractal Signals and Fields* (ISBN: 978-3-319-92584-4)
89. G. Plonka, D. Potts, G. Steidl, and M. Tasche: *Numerical Fourier Analysis* (978-3-030-04305-6)
90. K. Bredies and D. Lorenz: *Mathematical Image Processing* (ISBN: 978-3-030-01457-5)
91. H. G. Feichtinger, P. Boggiatto, E. Cordero, M. de Gosson, F. Nicola, A. Oliaro, and A. Tabacco: *Landscapes of Time-Frequency Analysis* (ISBN: 978-3-030-05209-6)
92. E. Liflyand: *Functions of Bounded Variation and Their Fourier Transforms* (ISBN: 978-3-030-04428-2)
93. R. Campos: *The XFT Quadrature in Discrete Fourier Analysis* (ISBN: 978-3-030-13422-8)
94. M. Abell, E. Iacob, A. Stokolos, S. Taylor, S. Tikhonov, J. Zhu: *Topics in Classical and Modern Analysis: In Memory of Yingkang Hu* (ISBN: 978-3-030-12276-8)
95. H. Boche, G. Caire, R. Calderbank, G. Kutyniok, R. Mathar, P. Petersen: *Compressed Sensing and its Applications: Third International MATHEON Conference 2017* (ISBN: 978-3-319-73073-8)
96. A. Aldroubi, C. Cabrelli, S. Jaffard, U. Molter: *New Trends in Applied Harmonic Analysis, Volume II: Harmonic Analysis, Geometric Measure Theory, and Applications* (ISBN: 978-3-030-32352-3)
97. S. Dos Santos, M. Maslouhi, K. Okoudjou: *Recent Advances in Mathematics and Technology: Proceedings of the First International Conference on Technology, Engineering, and Mathematics, Kenitra, Morocco, March 26-27, 2018* (ISBN: 978-3-030-35201-1)
98. Á. Bényi, K. Okoudjou: *Modulation Spaces: With Applications to Pseudodifferential Operators and Nonlinear Schrödinger Equations* (ISBN: 978-1-0716-0330-7)
99. P. Boggiato, M. Cappiello, E. Cordero, S. Coriasco, G. Garello, A. Oliaro, J. Seiler: *Advances in Microlocal and Time-Frequency Analysis* (ISBN: 978-3-030-36137-2)
100. S. Casey, K. Okoudjou, M. Robinson, B. Sadler: *Sampling: Theory and Applications* (ISBN: 978-3-030-36290-4)
101. P. Boggiatto, T. Bruno, E. Cordero, H. G. Feichtinger, F. Nicola, A. Oliaro, A. Tabacco, M. Vallarino: *Landscapes of Time-Frequency Analysis: ATFA 2019* (ISBN: 978-3-030-56004-1)

102. M. Hirn, S. Li, K. Okoudjou, S. Saliana, Ö. Yilmaz: *Excursions in Harmonic Analysis, Volume 6: In Honor of John Benedetto's 80th Birthday* (ISBN: 978-3-030-69636-8)
103. F. De Mari, E. De Vito: *Harmonic and Applied Analysis: From Radon Transforms to Machine Learning* (ISBN: 978-3-030-86664-8)