Genuense Athenaeum

Jonathan Eduardo Chirinos Rodríguez

Machine Learning Techniques for Inverse Problems

PhD Thesis

Università
di Genova

Dipartimento di Matematica
May 2024

Università
di Genova

PhD in Mathematics and Applications

Machine Learning Techniques for Inverse Problems

by

Jonathan Eduardo Chirinos Rodríguez

Thesis submitted for the degree of *Doctor of Philosophy* (Cycle XXXVI)

May 2024

Silvia Villa — Supervisor
Curzio Basso — Supervisor

UniGe | DIMA

Dipartimento di Matematica (DIMA)

A Maribel, por iluminar nuestro camino, y a Peter, por acompañarnos.

Sommario

I problemi inversi sono il modello naturale per l'analisi di molte applicazioni del mondo reale. Esempi tipici sono la risonanza magnetica (MRI), la tomografia computerizzata a raggi X (CT) e problemi di recupero delle immagini. Un problema inverso consiste nel ricostruire una sorgente sconosciuta da osservazioni limitate e potenzialmente distorte. Le cosiddette tecniche "data-driven" per risolvere i problemi inversi sono diventate popolari negli ultimi anni grazie alla loro efficacia in molti scenari pratici. Tuttavia, ad oggi sono state fornite poche garanzie teoriche sul loro funzionamento. Questo manoscritto si propone di colmare queste lacune procedendo lungo diverse direzioni chiave.

Gli approcci data-driven sono state oggetto di attenzione poiché richiedono meno conoscenze a priori. Nel primo lavoro, proponiamo e studiamo un approccio di Statistical Learning, basato su Empirical Risk Minimization (ERM), per determinare parametri a partire da esempi. Il nostro principale contributo è un'analisi teorica che dimostra come, se il numero di esempi è abbastanza grande, questo approccio sia ottimale ed adattattivo al livello di rumore e alla regolarità della soluzione. Mostriamo l'applicabilità del nostro framework a una vasta classe di problemi inversi, inclusi i metodi di regolarizzazione spettrale e le norme che promuovono sparsità. Simulazioni numeriche supportano e illustrano ulteriormente i risultati teorici.

Inoltre, introduciamo un approccio data-driven per costruire operatori (fortemente) nonespansivi. Presentiamo l'utilità di tale tecnica nel contesto dei metodi Plug-and-Play, in cui un operatore prossimale in algoritmi classici come Forward-Backward Splitting o l'iterazione primale-duale di Chambolle–Pock viene sostituito da un operatore che mira ad essere fortemente nonespansivo. Stabiliamo un rigoroso quadro teorico per imparare tali operatori utilizzando un approccio ERM. Inoltre, deriviamo una soluzione che è garantita essere fortemente nonespansiva e affine a tratti nell'inviluppo convesso del training set. Dimostriamo che questo operatore converge alla migliore soluzione empirica aumentando il numero di punti all'interno dell'inviluppo. Infine, proponiamo una strategia di implementazione pratica e un'applicazione nel contesto dell'image denoising.

Spesso, i problemi data-driven si scontrano con la sfida di affrontare problemi di dimensione infinita. I teoremi di rappresentazione, introdotti nel contesto dei metodi kernel e recentemente estesi allo studio di problemi variazionali generali, possono essere applicati per affrontare questa questione. Questi teoremi caratterizzano le soluzioni di problemi di dimensione infinita come una combinazione convessa finita di un numero limitato di "atomi". In casi specifici, si può dimostrate che questi atomi sono i punti estremali di una palla unitaria specifica. In questo contesto, contribuiamo caratterizzando l'insieme dei punti estremali della palla unitaria delle funzioni Lipschitziane in spazi metrici finiti. Di conseguenza, verrà fornito un teorema di rappresentazzione in questa impostazione, generalizzando il cosiddetto Teorema di Minkowski-Carathéodory a spazi di dimensione infinita.

Abstract

Inverse problems serve as a general playground for analyzing many real-world applications. Typical examples are MRI, X-Ray CT, and image recovery. An inverse problem involves reconstructing an unknown source from limited and possibly distorted observations. The so-called data-driven techniques for solving inverse problems have become popular in recent years due to their effectiveness in many practical scenarios. Yet, few theoretical guarantees have been provided to date. This manuscript aims to bridge this gap in several key directions.

Data driven approaches have gained attention since they require less prior knowledge. First, we propose and study a statistical machine learning approach, based on Empirical Risk Minimization, to determine the best regularization parameter given a finite set of examples. Our main contribution is a theoretical analysis, showing that, if the number of examples is big enough, this approach is optimal and adaptive to the noise level and the smoothness of the solution. We showcase the applicability of our framework to a broad class of inverse problems, including spectral regularization methods and sparsity-promoting norms. Numerical simulations further support and illustrate the theoretical findings.

Moreover, we introduce a data-driven approach for constructing (firmly) nonexpansive operators. We present the utility of such a technique in the context of Plug-and-Play methods, where one proximal operator in classical algorithms such as Forward-Backward Splitting or the Chambolle–Pock primal-dual iteration is substituted by an operator that aims to be firmly nonexpansive. We establish a rigorous theoretical framework for learning such operators using an ERM approach. Further, we derive a solution that is ensured to be firmly nonexpansive and piecewise affine in the convex envelope of the training data. We prove that such an operator converges to the best empirical solution when increasing the number of points inside the envelope. Finally, we propose a practical implementation strategy and an application in the context of image denoising.

Often, data-driven approaches require to deal with infinite-dimensional problems. Representer theorems, introduced in the context of kernel methods, and recently extended for studying general variational problems, can be applied for tackling this issue. These theorems characterize solutions of infinite-dimensional problems as a finite convex combination of a limited number of "atoms". In specific cases, these atoms can be shown to be the extreme points of a specific unit ball. In this setting, we contribute by characterizing the set of extreme points of the Lipzchitz unit ball in finite metric spaces. Consequently, a representer theorem in this setting will be provided, generalizing the so-called Minkowski-Carathéodory Theorem to infinite-dimensional spaces.

Acknowledgements

First, I would like to thank Dr. Curzio Basso and Prof. Silvia Villa, who gave me the unique opportunity to do this PhD Thesis within the TraDE-OPT ITN project. Notwithstanding this, they also brightly guided and helped me through this beautiful hike which, like most of the ones one may find here, in Liguria, has been full of ups and downs. But always, always, close to the sea. I also want to thank them for the immense patience that they have had with me, almost infinite. I can say, without a doubt, that they always trusted in me and gave me their sincere help and honest opinion in any of the doubts I have had during these three years. In a more personal note, I would particularly like to thank Silvia. For showing me, not only mathematics, but also what it really means to be a researcher. For her patience. For her time. For her never-ending energy. For the advises. For her bright sense of humor, which gave me tons of laughs. For opening me the doors of her house and her family. Thank you, Silvia!

Next, I would like to thank both Lorenzo Rosasco and Ernesto de Vito, both experts and worldwide known researchers in the areas of Machine Learning and Inverse Problems, and with whom I had the pleasure to deepen my knowledge in these subject.

I would also like to thank Gaspare Piemontese and Lara Provenzano, from Camelot, and Giulia Casu and the recent but impressive sign of Nathalie Baxs, from MaLGa. In a world full of anger and stress, they have all helped me in everything I needed throughout these almost four years. Bureaucracy has never been my main strength, and they have always been open to help, always with the best manners and kindness. I am convinced that, without the help of such great, not only professionals, but also human beings, my stay in this country would have been, for sure, more difficult. Thank you all!

Quisiera también dar las gracias a Cesare Molinari, por estar presente desde el principio de mis días en Génova. Cesare fue, y es, una persona con quien siempre pude contar y en quien siempre pude confiar independientemente de lo que necesitara. Además, Cesare fue quien me introdujo en varios de los temas de los que trata este manuscrito: Optimización Convexa y Problemas Inversos y, no en menor importancia, quien me enseñó también, a su particular manera, la ciudad de Génova y muchos (o casi todos) de sus secretos (también llamados 'vicoli' en italiano). En resumen, puedo decir, sin lugar a dudas, que de esta ciudad me llevo un amigo con un corazón gigantesco.

This beautiful path started with other three people: Cristian Vega, Cheik Traoré and Marco Rando. We four started together and lived, always together, this tremendously large and steep hike. We were there when it was needed, helping each other and trying to push the best out of us. Without competitions or childish moves. Always with honesty and friendship. Thank you guys, for everything. Really!

En particular, Cristian no ha sido solo un compañero de trabajo. Cris ha sido un amigo. Mi mano derecha, el hombro en el que siempre me apoyé cuando estuve jodido. Quien

me escuchó en mis peores momentos: esos en los que a nadie le apetece escucharte. Cris con sus puntos de vista particulares y nuevos. Cris con su manera de contar las cosas, de expresarse, de sentir y de ser. Cris. Siempre Cris. Cris fue, y es, también, mi compañero de piso. A quien desperté tantas mañanas con mi molinillo de café y que siempre aguantó mis manías y estrés con el orden extremo. Cris nunca se quejó y eso también se lo agradezco. Mi confidente, mi hermano mayor. Gracias, Cris. Te quiero!

Questi quasi quattro anni in Italia sono stati pieni di bei momenti vissuti con delle persone fantastiche. Persone che ho avuto l'infinito piacere di conoscere sia fuori che dentro MaLGa. Con loro ho condiviso momenti pieni di risate, buon cibo, mare, vicoli, tortillas, arepitas, birre, amari... Per questo e altri motivi, volevo ringraziare Elena, Alessandro, Andrea, Francesco, Matteo Monti, Matteo Levi, Paolo, Giacomo, Noemi, Bastiano, Rosanna, Luca Ratti, Nicola, Marco Letizia, Vassi, Edoardo, Massimiliano, Vito, Pietro, Emilia, Simone Sanna, Matteo Santacesaria, Francesca, Simone di Marino, e anche i miei francesi del cuore Antoine, Mathurin, Romain, Hippolyte e Nicolas. Non meno importanti sono i miei colleghi di calcetto, e alcuni anche di Dipartimento (DIMA) Andrea Poggio, Stefano, Larbi, Giovani Minuto, Francesco Zerman, Luis, Jack, Javier, Laura,... E a Bea. Per arrivare nel momento giusto. Per gli aperitivi e le risate. Grazie davvero a tutti, di cuore.

An important part of these three years was my stay in the beautiful city of Graz, in Austria, where I lived for a total of seven months, in two different periods. Both of the periods where scientific internships under the supervision of Prof. Kristian Bredies, from the University of Graz. I would like to thank him for his time. For all of the days where I knocked at his door with a pletora of questions, which he always answered with patience and calmness. For the ideas. For all the discussions about his point of view in research and science in general. For his passion for mathematics. For the personal discussions about present and future. Thank you, Kristian!

During my first period in Graz, I shared appartment with Emanuele Naldi, Ema. At some point I thought that this was not going to happen, but in the end we managed to organize and coordinate to share the same period together in Graz. With him, and also with Rodolfo and Enis, I learned the Italian I know, but not only. With all of them four I had fun like a kid. We cooked pizza, desserts and handmade pasta. We drunk tons of wine and beer. We discussed about math and politics and economics and love and passion for what we do or we don't do and many other things. We became friends from the beginning and I learned a lot from each of them. From the very first day. I have to say that this has to be seen as luck. I don't see any other way. Four is already a number of people where things could not go well. But they did. They really did. In Graz I also met Lisi. The amazing Lisi. An unbelievable human being with whom I had the pleasure to share hundreds of hours, weirdly cooked cakes that made us laugh without control, parties, dinners and many other things. I will never forget any of them. I have no doubt.

A Carlos. Mandando cartas desde su trinchera particular. Por su paz. Y por transmitirmela. Por Chipre y por Bari. Por aquel penúltimo ron con el Atlántico de fondo. Porque queda al menos uno más. A su familia también, Javi, Manolo y Maria José, por acogerme cuando tocaba. Por hacerme sentir uno más. Por mostrarme su Madrid. A Isabel, por sus Jonathan en mayúsculas, por las infinitas risas hasta la lágrima. Por estar siempre, pero sobre todo en lo malo. A su familia también, Maribel y Antonio. Por su amabilidad eterna. Y a Troy, mi fiel compadre Troy. Por todos y cada uno de esos viajes. Y por los que nos quedan también. Por los vinos y las recetas. Por ese ritmo pausado que tanto nos define.

Doy gracias a la vida por ese regalo que me dio Madrid y no me ha robado aún. Y ya me queda para siempre.

A mis compañeros del LOL, Julio, Sergio, Pablo y, a veces, Pedro. Por las infinitas horas en llamada. Por escucharme, a su manera, cuando lo necesité. Por las discusiones. Por las cervezas en La Concepción. Por las risas también. Sobre todo por las risas. A mis amigos de siempre: Alba, Elisa, Isabel y Coki, que siguen a mi lado, como de costumbre. A Bruno y Ayarith, por los cafés, los juegos de mesa, las discusiones y las risas. Por su calma. Por su espacio y por su tiempo. Por acompañarme también. A Cristo, mi viejo amigo Cristo. Que aunque pasen los años parece que nada cambió. Por mi primer viaje en moto en Tenerife, por no dejar nunca de ser como es. Por creer en él mismo y en mí, a pesar de todo. Gracias a todos, por estar.

Gracias a mi mafia. A Pedro, Héctor y Amine. Los que me llevo a todas partes. Hasta el fin del mundo. Gracias por ser mi norte. Y mi sur también cuando tocaba. Por escucharme. Por decirme lo que tenía que escuchar, no lo que quería. Por venir a verme cuando pudieron. Por las birras. Por los guachinches y las playas. Por los conciertos. Por todos esos momentos en genera. Por las risas también, infinitas. Por ser los hermanos que nunca tuve. Por ser, chicos. Gracias.

A la casa Amine y a las de dentro: Soheila, Alma y Amine, mi negra. Por las charlas eternas. Por los almuerzos y las cenas. Por los cafés. Por ser la familia postiza perfecta. Por los consejos y las risas. Por los libros y los poemas. Por acompañarme y escucharme en este camino tan especial. Por su amor. A pesar de todo, por su amor. Por indicarme el camino cuando hizo falta. Por todos estos años y los que quedan. Siempre. Gracias.

Finalmente, a Maribel. mi madre. Mi luz. El motivo por el que estoy donde estoy. Mi causa y mi consecuencia. Quien me empuja, me acompaña y me apoya. Quien me aguanta cuando no estoy bien. Siempre a mi vera. And to Peter, for the laughs and conversations. For the coffe and the football. For being there when most needed. For being a key element in my life in the past years.

Table of Contents

CHAPTER 1

Introduction

1.1 General context

An inverse problem consists in recovering a solution u^* from a set of linear measurements Au^*, where A is an operator modelling the measurements acquisition process. This task appears in a broad range of practical problems in engineering, signal processing, medical imaging or computer vision. For instance, in signal processing, one often wishes to recover a signal from possibly noisy distorted observations. In medical imaging, Magnetic Resonance Imaging (MRI) or X-ray computed tomography (CT) are typical and give rise to such observations. More generally, image restoration problems, which consist in reconstructing degraded images, also fit into this framework (we refer the reader to [16, 64] for more examples). In mathematical terms, an inverse problem can be written as

$$x = Au^* + \varepsilon, \tag{1.1.1}$$

where x denotes the available measurements and ε is a deterministic quantity modelling the possible presence of noise in the observations. The task of finding u^* from the knowledge of x becomes hard when the problem is ill-posed. In general, a problem is said to be ill-posed whenever (i) solutions do not exist, (ii) solutions are not unique, or (iii) solutions do not depend continuously on the data. For instance, ill-posedness corresponds in our case to the matrix A not being injective (since the inverse would not exist), or having an inverse with large norm. Regularization theory [15, 64] offers a systematic way to address ill-posedness by providing stable approximations of the inverse. A classical approach for restoring well-posedness is based on finding solutions of the following variational problem

$$\min_u \ell(Au, x) + \lambda R(u), \tag{1.1.2}$$

for some $\lambda \in (0, +\infty)$. The term $\ell(A\cdot, x)$ is known as the *data-fitting* term, and constraints the solution to remain close to the available measurements. The function $R\colon \mathcal{U} \to (0, +\infty]$, referred to as the *regularization function*, incorporates prior knowledge about the solution into the problem formulation. For instance, R could be designed in such a way that the values of $R(u)$ are low if u has approximately the same structure as u^* (e.g., is sparse, low-rank, etc.), and high otherwise. Finally, the scalar $\lambda \in (0, +\infty)$ is known as the *regularization parameter*. This parameter allows to choose the relative importance of the data-fitting term and the regularization function, thereby influencing the quality of the recovery results. In addition, for fixed ℓ and R, solutions of (1.1.2) should converge to u^* for a suitable choice of $\lambda = \lambda(\|\varepsilon\|) \to 0$ as $\|\varepsilon\| \to 0$. Consequently, a proper selection of the regularization parameter is essential for achieving optimal reconstruction outcomes. To this day, selecting an appropiate R and a suitable regularization parameter remain challenging problems.

The above strategy for solving inverse problems can be viewed as a *model-based* technique, relying on a mathematical model with well-established properties. For instance, variational methods have for a long time achieved state-of-the-art results [118] in imaging problems. The design of refined regularization terms (e.g. Total Variation [115] or Total Generalized Variation [29]) contributed to achieve remarkable practical performances, while additionally providing robust theoretical guarantees. Notwithstanding this, *data-driven* methodologies have gained significant attention in recent years, since they demonstrate improved performance in various practical scenarios while overcoming some challenges of classical methods (see [5] and references therein). The starting point of data-driven approaches is the assumption that a finite set of pairs of measurements and exact solutions $(\bar{x}_1, \bar{u}_1), \dots, (\bar{x}_n, \bar{u}_n)$, $n \in \mathbb{N}$, is available. This *training set* is then used to define, or refine, a regularization strategy to be applied to any future observation $\bar{x}_{\text{new}}$, for which an exact solution is not known. Here, we will focus on studying data-driven approaches that maintain a fixed underlying variational model, and learn particular elements therein. We list below some relevant examples:

- In the variational model (1.1.2), we fix the regularizer R and we aim to learn the regularization parameter $\lambda \in (0, +\infty)$ from the given training set. This approach is based on the following bilevel optimization problem (see, for instance, [111, 97]). Given a set $\Lambda \subset (0, +\infty)$, we select the regularization parameter as

$$\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} \|u_\lambda^i - \bar{u}_i\|_{\mathcal{U}}^2, \tag{1.1.3}$$

  where $u_\lambda^i := u_\lambda(\bar{x}_i)$ is such that

$$u_\lambda(\bar{x}_i) \in \arg\min_{u \in \mathcal{U}} \ell(Au, \bar{x}_i) + \lambda R(u),$$

  for some discrepancy ℓ (see [53, Chapter 3] and references therein). We then utilize $\widehat{\lambda}$ as regularization parameter for subsequent instances of the same inverse problem: given $\bar{x}_{\text{new}}$, we consider $u_{\widehat{\lambda}}(\bar{x}_{\text{new}})$ as an approximation of $\bar{u}_{\text{new}}$. This approach will be further developed in Chapter 3.

- The previous framework can be generalized by fixing instead regularizers $R \colon \mathcal{U} \times \Theta \to (0, +\infty]$, that are parametrized by a vector $\theta = (\theta_1, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$. The variational model in this cases reads as

$$\min_{u \in \mathcal{U}} \ell(Au, x) + R(u, \theta).$$

  Here, the objective is to learn the vector of parameters $\boldsymbol{\theta}$. Neural Networks that are parametrized by a large set of scalars have been considered, e.g. the Total Deep Variation [96]. A similar approach to the one defined above can be used to find the optimal $\boldsymbol{\theta}$ given a finite training set of input/output pairs.

- Finally, an intriguing recent approach involves learning the entire regularization function R without assuming any prior structure on it. A relevant example can be found in the case where the training set is relative to the denoising problem (A is equal to the identity). If the fidelity term ℓ in (1.1.2) is given by the squared norm, $\ell(x, x') := \frac{1}{2}\|x - x'\|_{\mathcal{X}}^2$ for every x, $x' \in \mathcal{X}$, a possible approach is to learn directly the proximal operator of R, since finding solutions of the variational problem

$$\min_u \frac{1}{2}\|u - x\|_{\mathcal{X}}^2 + R(u),$$

amounts to find elements $u \in \mathcal{U}$ such that $u = \operatorname{prox}_R(x)$. Motivated by the denoising property of proximal operators, the so-called Plug-and-Play (PnP) methods [138] aim to learn instead the best "denoiser" T^* such that $u = T^*(x)$. Further, this operator can then be plugged into any first-order minimization algorithm as a substitute of the proximal map. A possible approach for tackling the problem of learning T^* is to find instead

$$\widehat{T} \in \operatorname*{arg\,min}_{T \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \|T(\bar{x}_i) - \bar{u}_i\|_{\mathcal{X}}^2, \tag{1.1.4}$$

over the space $\mathcal{M}$ of firmly nonexpansive operators [13]. By constraining $\widehat{T}$ to be firmly nonexpansive, we ensure that PnP algorithms converge to a fixed point, see [107]. This problem will be further studied in Chapter 4.

A key observation is that the constrained optimization problem (1.1.4) is defined on an infinite-dimensional space, and hence, it remains somewhat unclear how classical first-order optimization algorithms could be practically implemented. Representer theorems, introduced in the context of kernel methods [119], serve as a natural tool for tackling this issue. In short, a representer theorem enables to write solutions of problems such as (1.1.4) as a convex combination of a finite number of atoms. Consequently, the problem of finding the best operator in an infinite-dimensional space can be reduced to the finite-dimensional problem of finding the right weights for each atom. In Chapter 5, we provide a representation result that can be further applied for solving (1.1.4) in a particular framework.

In this manuscript, we aim to develop the theory for the above mentioned problems in two primary directions: first, to provide theoretical support for data-driven parameter selection methods, and second, to design a data-driven approach for constructing firmly nonexpansive operators with theoretical guarantees. Additionally, we will provide a characterization to the extreme points of the Lipschitz unit ball in finite metric spaces. Further connections between extreme points and representer theorems will be provided. Next, we describe the main motivations and contributions of the thesis.

1.2 Motivation and contributions

We now briefly describe the content of the three main works in this manuscript. We start by introducing the problem of learning the best regularization parameter in inverse problems.

Learning the regularization parameter

Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ and $(\mathcal{U}, \langle\cdot,\cdot\rangle_{\mathcal{U}})$ be real and separable Hilbert spaces, with $x \in \mathcal{X}$ and $u^* \in \mathcal{U}$ representing the measurement/solution pair given by the deterministic inverse problem stated in (1.1.1). As we mentioned above, in order to find stable approximations of u^*, a regularization perspective is essential. Towards this end, we consider a family of regularization operators $u_\lambda : \mathcal{X} \to \mathcal{U}$ parametrized by the positive scalar $\lambda \in (0, +\infty)$. Ideally, for some given discrepancy ℓ, a proper choice of the regularization parameter λ should allow to optimally control the error $\ell(u_\lambda(x), u^*)$.

Strategies such as the Morozov discrepancy principle [104] or the balancing principle [99, 132] were extensively studied in the past and rely on prior knowledge about the noise level $\|\varepsilon\|$. However, this information is often unavailable in many practical scenarios. Motivated by this challenge, subsequent studies have focused on providing parameter choice

rules that solely depend on the measurements x. For instance, we mention generalized cross validation [73, 139] or the quasi-optimality criterion [130, 131]. An exhaustive introduction to parameter choice rules will be given in Section 2.3.1.

In this work, we aim to analyze a particular class of *heuristic* rules that has gained attention in recent years: the so-called *data-driven* parameter selection methods described above. Although these approaches have demonstrated significant success in practice (as shown in [33, 34, 88, 97, 111]), limited convergence guarantees have been established to date. The theoretical analysis we conduct is grounded on the following observations:

- **Transition to a stochastic perspective:** thus far, we have considered the deterministic perspective of inverse problems. In this context, the so-called Bakushinskii veto [8] dictates that no family $(u_\lambda)_{\lambda>0}$ can serve as a convergent regularization method if the regularization parameter λ does not depend on the noise level $\|\varepsilon\|$; i.e., there is no hope to show that $u_\lambda(x) \to u^*$ as $\lambda \to 0$. Consequently, providing convergence guarantees for data-driven approaches within this framework seems infeasible. It therefore seems reasonable to pivot towards a stochastic setting for inverse problems: consider the model
$$\bar{X} = A(\bar{U}) + \varepsilon \tag{1.2.1}$$
where both $\bar{X}$ and $\bar{U}$ are random variables taking values in $\mathcal{X}$ and $\mathcal{U}$ respectively. Here, $\varepsilon \in \mathcal{X}$ is an additive random variable modelling the noise and $A\colon \mathcal{U} \to \mathcal{X}$ defines the forward operator, not necessarily linear. This stochastic perspective has already been explored for yielding provably convergent parameter selection methods, see [11].

- **Empirical Risk Minimization (ERM) paradigm:** the bilevel formulation for selecting the regularization parameter that was stated in (1.1.3) closely resembles a classical regression problem within the context of Supervised Learning [54, 137]. Specifically, it aligns with the well-established theory of ERM. This approach is based on the assumption that the probability distribution of the pair $(\bar{X}, \bar{U})$ is inaccessible, but a finite set $\{(\bar{X}_i, \bar{U}_i)\}_{i=1}^n$, $n \in \mathbb{N}$, of independent copies of $(\bar{X}, \bar{U})$ is available. By fixing a regularization method $\bar{U}_\lambda = \bar{U}_\lambda(\bar{X})$ and a loss function $\ell\colon \mathcal{U} \times \mathcal{U} \to [0, \infty)$, we define $\widehat{\lambda}_\Lambda$ as the minimizer of the empirical risk
$$\widehat{\lambda}_\Lambda \in \operatorname*{arg\,min}_{\lambda \in \Lambda} \widehat{L}(\bar{U}_\lambda), \quad \widehat{L}(\bar{U}_\lambda) := \frac{1}{n} \sum_{i=1}^n \ell(\bar{U}_\lambda(\bar{X}_i), \bar{U}_i)$$
within a set of parameters $\Lambda \subseteq (0, +\infty)$. To perform the theoretical analysis, we borrow ideas from the literature of model selection in statistics and machine learning [60, 75], particularly adapting concepts from [36].

We next outline the primary contributions of this work.

1. We provide a general theoretical analysis that shows that, if $\mathbb{E}[\ell(\bar{U}_\lambda, \bar{U})] \leq \Phi(\lambda)$ for some function Φ and for any $\lambda \in (0, +\infty)$, then the ERM approach for learning the regularization parameter can essentially achieve the same performance as the one by the optimal a-priori choice, up to an error term which decreases with the size of the training set. This result is expressed as follows: given $\eta \in (0, 1)$, with probability greater than $1 - \eta$,
$$\mathbb{E}[\ell(\bar{U}_{\widehat{\lambda}_\Lambda}, \bar{U})] \lesssim \Phi(\lambda_*) + \frac{\log(|\Lambda|/\eta)}{n},$$
where λ_* is a minimizer of Φ and $|\Lambda|$ denotes the cardinality of the set Λ.

2. We show that the scheme presented above can be applied to a broad class of inverse problems. For linear inverse problems in Hilbert spaces, we investigate spectral regularization methods and variational approaches with general convex regularizers. Specifically, we show that the above result is valid for Legendre regularizers [12] and sparsity-promoting regularization functions [7] in the finite-dimensional setting. In addition, we consider non-linear inverse problems in Hilbert spaces and the corresponding Tikhonov regularization [64, Chapter 10].

3. We support the theoretical analysis through numerical experiments, showing the validity of the derived probabilistic bounds for some of the examples mentioned above. We study the cases of Tikhonov and Landweber regularization [64] in a synthetic context, signal denoising and signal deblurring with ℓ^1 regularization [129], and Total Variation regularization [115] for solving an image denoising problem.

Next, we introduce the problem of learning firmly nonexpansive operators.

Learning firmly nonexpansive operators

The focal point of this work is to consider the denoising problem, where we set $\mathcal{X} = \mathcal{U}$ to be a real and separable Hilbert space, and we fix the forward operator A equal to the identity in (1.1.1). In essence, our objective is to construct an operator T^* functioning as a denoiser: given $x \in \mathcal{X}$, T^* should output a denoised version of x, $T^*(x)$, that is close, in some sense, to u^*. As previously discussed, classical approaches involve considering the variational model stated in (1.1.2), with R being a proper, convex and lower semicontinuous function that acts as a denoiser. Several regularization functions for tackling denoising problems have been designed in the literature. Notable examples include Total Variation regularization [38, 115] and its further higher-order generalizations [29, 28, 42]. Prior to the advent of deep learning techniques, these methods represented the state-of-the-art for addressing, for instance, image denoising problems. Next, we develop a different class of methods that can be applied in the same context, and are naturally related to the so-called Plug-and-Play (PnP) approaches.

A novel data-driven approach for solving denoising problems is based on the following observation: finding solutions of (1.1.2) in this case amounts to find $u \in \mathcal{U}$ such that $u = \operatorname{prox}_R(x)$ for a given measurement $x \in \mathcal{X}$. Consequently, it seems reasonable to learn an operator T such that $T = \operatorname{prox}_R$ for some unknown R. The latter approach for learning the regularization operator is related to the so-called Plug-and-Play (PnP) methods. In short, a PnP algorithm substitutes the proximal map in general splitting algorithms such as the Forward-Backward Splitting (FBS) algorithm [6, 50, 72], the Chambolle–Pock primal-dual iteration (CP) [40] or other dual/primal-dual approaches [38, 49, 51], with a general operator T that acts as a *denoiser*. This choice is motivated by the definition of the proximal operator since, for a given input $x \in \mathcal{X}$, $\operatorname{prox}_R(x)$ outputs a denoised version of x. The latter approach would, in some sense, solve two main paradigms that are faced in classical approaches: it avoids choosing an appropriate regularizer R which, as we mentioned, is not an easy task, and it does not require to compute the proximal operator of a convex function which, in general, does not have a closed-form expression.

PnP approaches have demonstrated remarkable empirical performance across a diverse range of imaging applications [32, 61, 123, 127, 138] and, since then, several follow-up works have been published. Notably, the BM3D method [55, 82, 123], although "hand-crafted", has been extensively employed within PnP approaches [93]. Denoising methods based on deep learning techniques have gained attention in recent years [81, 102, 142]. However, theoretical guarantees for this class of methods remain limited. Before diving

into a concise explanation to the term "theoretical guarantees" in this context, it is crucial to highlight that, as we will explore, much of the effort to address the absence of theoretical backing is based on the following rudimentary, yet reasonably accurate, notion:

Constructing proximal operators of convex functions is "equivalent" to constructing nonexpansive operators.

Indeed, the above statement lacks precision, and we will later render it more mathematically rigurous (specifically in Chapter 2). However, constructing nonexpansive operators is, in general, difficult [107], and consequently, many theoretical difficulties arise. Below, we provide two particular consequences that serve as the main motivation of our work.

- **Convergence guarantees.** It has been observed that ensuring convergence guarantees for PnP methods is often challenging due to the dependence on prior properties of the denoiser. Some works have already studied the theoretical properties of denoising functions within the framework of first-order methods. Among them, we mention [41], which proves convergence with a bounded denoiser, or [123, 128], where they prove convergence for PnP Forward-Backward splitting (PnP-FBS) and for PnP-ADMM assuming the denoiser to be nonexpansive. Additionally, authors in [116] provide a convergence analysis for both PnP-FBS and PnP-ADMM under specific contractivity assumptions on the denoiser. However, enforcing a nonexpansivity constraint, or more broadly, a Lipschitz constraint on Neural Networks [107] is challenging in practice. To restore more classical convergence guarantees, and inspired by the fact that the proximal operator of a convex function is a firmly nonexpansive operator, some authors have designed the denoiser as a Neural Network that aims to be an averaged – or firmly nonexpansive – operator [85, 107, 121]. All of the aforementioned studies exhibit promising practical results.

- **Density results.** Another significant motivation arises from the inherent infinite-dimensional nature of the set of nonexpansive operators. As a result, a possible approach to tackle this issue in practice is to discretize this set. For example, methodologies relying on neural networks approximate this space using a vast array of parameters. Consequently, several pertinent questions arise: How closely does the discretized problem resemble its continuous counterpart? Is it possible to provide any density results? Does the solution of the discretized problem converge to that of its continuous counterpart?

The contributions of our work are described below:

1. We study and analyze the properties of the space of nonexpansive operators which, we show, can be characterized as a subset of the dual of a suitable Banach space. Leveraging this, we introduce a natural notion of weak* convergence. With these tools, we provide a rigurous mathematical framework to address the problem of learning an operator through a constrained minimization problem.

2. To construct nonexpansive operators, we adopt the same Statistical Learning approach that we followed for tackling the problem of learning the regularization parameter. We fix $(\bar{X}, \bar{U})$ to be a pair of random variables taking values in $\mathcal{X}$. From a theoretical standpoint, we aim at finding N^*, the minimizer of the expected risk, within the set of nonexpansive operators,

$$N^* \in \operatorname*{arg\,min}_{N \in \mathcal{N}} L(N), \quad L(N) := \mathbb{E}[\|N(\bar{X}) - \bar{U}\|_{\mathcal{X}}^2],$$

where

$$\mathcal{N} := \{N : \mathcal{X} \to \mathcal{X} \,|\, \|N(x) - N(x')\| \leq \|x - x'\|, \text{ for every } x, x' \in \mathcal{X}\}.$$

By simply assuming that the pair $(\bar{X}, \bar{U})$ satisfies $\mathbb{E}[\|\bar{X}\|_{\mathcal{X}}^2 + \|\bar{U}\|_{\mathcal{X}}^2] < \infty$, we show that N^* exists. As before, we do not have access to the exact distribution of the pair $(\bar{X}, \bar{U})$, but to a finite set $\{(\bar{X}_i, \bar{U}_i)\}_{i=1}^n$, $n \in \mathbb{N}$, of independent copies of $(\bar{X}, \bar{U})$. Then, the nonexpansive operator that approximates N^* will be the minimizer of an ERM problem over the space of nonexpansive operators:

$$\widehat{N} \in \operatorname*{arg\,min}_{N \in \mathcal{N}} \widehat{L}(N), \quad \widehat{L}(N) := \frac{1}{n} \sum_{i=1}^{n} \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2. \tag{1.2.2}$$

This operator can also be subsequently utilized for denoising future measurements $\bar{x}_{\text{new}}$.

3. In the aforementioned setting, we prove that the ERM (1.2.2) Γ-converges, almost surely, to the expected risk as the number of points in the training set goes to infinity. This result holds significant importance as it implies, in particular, that if a sequence of minimizers of the ERM, for every $n \in \mathbb{N}$, is considered, then (up to subsequences) it converges almost surely to N^*.

4. As previously mentioned, the class of nonexpansive operators is infinite-dimensional, and a further discretization is required. To address this, we propose a discretization for such a space via piecewise affine functions uniquely determined by simplicial partitions, or triangulations, of the underlying space. This approach yields two main consequences:

   (i) We propose a constructive approach for designing an operator which will be ensured to be piecewise affine and nonexpansive. With this, it is possible to prove that classical minimization algorithms converge to a fixed point.

   (ii) We establish a density result, showing that the class of considered approximating operators is actually large enough: we show that succesively finer triangulations lead to a closer approximation to $\widehat{N}$, minimizer of (1.2.2).

5. We provide a detailed explanation in Section 4.4 on how to design the PnP versions of some algorithms of interest and show their convergence by standard results: PnP Forward-Backward Splitting, PnP ADMM, PnP Douglas–Rachford, and PnP Chambolle–Pock primal-dual iteration.

6. In order to solve the discretized problem in practice, an efficient algorithm must be designed. We propose an algorithm that can work well in practice while having strong theoretical guarantees. Specifically, we design a PnP Chambolle–Pock method by making use of the Moreau's identity. This approach also ensures interpretability of the proposed method.

7. Finally, we evaluate our proposed method in imaging applications, particularly addressing the problem of image denoising. We conclude by comparing our learned denoiser with classic Total Variation regularizers.

We finally describe the third work of this thesis, which tackles the problem of finding the extreme points of the Lipzchitz unit ball and further consequences.

Extreme points and representer theorems for the Lipschitz unit ball

Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ and $(\mathcal{U}, \langle\cdot,\cdot\rangle_{\mathcal{U}})$ be real Hilbert spaces. Denote by Lip the space of Lipschitz functions $f\colon \mathcal{X} \to \mathcal{U}$, and by Lip^M the space of Lipschitz functions with Lipschitz constant $M > 0$. Recall that the space Lip is a Banach space with norm

$$\|f\|_{\mathrm{Lip}} := \|f - f(0)\|_{\mathrm{Lip}_0} + \|f(0)\|_{\mathcal{U}},$$

where $\|\cdot\|_{\mathrm{Lip}_0}$ stands for the smallest Lipschitz constant of f and $\|\cdot\|_{\mathcal{U}}$ is the norm induced by the inner product of $\mathcal{U}$. In addition, given $x_0 \in \mathcal{X}$ we denote by Lip_0 as the space of Lipschitz functions that vanish at x_0 (it is often common to assume $x_0 = 0$). The latter is a Banach space with norm $\|\cdot\|_{\mathrm{Lip}_0}$. Finally, we denote by Lip_0^M its bounded analogue. A detailed work about Lipschitz functions can be found in [140].

In the particular case where $\mathcal{X} = \mathcal{U}$ is a real Hilbert space, the space Lip^1 coincides with the space of nonexpansive operators described in the section above. The main objective of this work to study certain structural properties of the bounded analogue Lip_0^1. In particular, we are interested in characterizing the set of its extreme points. We recall that, given a convex set C, an extreme point of C is a point $x \in C$ such that $C \setminus \{x\}$ remains a convex set. Next, we describe why such a result could be relevant, particularly in the context of variational problems and optimization algorithms.

- **Connection to optimization algorithms.** Recently, it has been observed in [24, 25, 52] that an accurate characterization of extreme points can significantly enhance the efficiency and speed of certain optimization algorithms, particularly the Conditional Gradient Methods (CGM) [71, 91]. These methods, commonly used for constrained optimization problems, are known to move along extreme points during iterations. For instance, the ERM formulated in (1.2.2) is a natural example where CGM could be applied.

- **Representer theorems and dimensionality reduction.** Another motivation for studying such a characterization is drawn by the well-known *representer theorems*. Originally introduced in the context of kernel methods [119] for Machine Learning, these theorems enable expressing some solutions to certain optimization algorithms as a finite convex combination of a few atoms. Recent works [19, 21, 43] explore representation results for more general variational problems than (1.1.2). In our case, we are interested in problems of the form

$$\inf_{u\in\mathrm{Lip}^1} F(\Phi u), \text{ or } \inf_{u\in\mathrm{Lip}} F(\Phi u) + \lambda\|u\|_{\mathrm{Lip}_0} \tag{1.2.3}$$

  where F is an arbitrary function, not necessarily convex nor differentiable, Φ is a linear operator – mapping elements in $\mathcal{U}$ to a finite-dimensional space –, $\|u\|_{\mathrm{Lip}_0}$ is the smallest Lipschitz constant of u, and $\lambda \in (0, +\infty)$ is the regularization parameter. Authors in [19] prove that there are solutions of problem (1.2.3) that can be written as a convex combination of a finite number of atoms.

  Motivated by the so-called Minkowski–Carathéodory Theorem [87], it has been shown that when the dimension of $\mathcal{X}$ is finite, (say $\dim \mathcal{X} = d$, $d \geq 1$), these atoms are the extreme points of the unit ball associated with the regularizer [63]. In our constrained problem, this translates to the set Lip^1, and in the unconstrained case, $\{u \in \mathcal{U} \,|\, \|u\|_{\mathrm{Lip}_0} \leq 1\}$. Consequently, for every $u \in \mathrm{Lip}^1$, there exist $u_0, ..., u_d$ extreme points of Lip^1 and a vector of non-negative scalars $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_d)$, with $\sum_{i=0}^d \alpha_i = 1$, such that

$$u = \sum_{i=0}^{d} \alpha_i u^i.$$

  Hence, finding solutions of (1.2.3) amounts to find the vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, satisfying the conditions above, such that $u \in \mathrm{Lip}$ solves (1.2.3). This establishes a natural

connection between extreme points and representation results. Moreover, recall that the ERM described in (1.2.2) perfectly fits into this setting, where $F \circ \Phi = \widehat{L}$. The study in [19] shows that some solutions to problem (1.2.3) in our scenario can be expressed as a convex combination of finitely many extreme points of the Lipschitz unit ball. However, such a characterization is lacking. Finally, it is worth mentioning that Lipschitz-type constraints are increasingly relevant in the context of Plug-and-Play regularization [116], monotone splitting algorithms [107], and unconstrained variational problems like (1.2.3), see [58].

Providing a characterization of the extreme points of the Lipschitz unit ball has been widely studied during the years. In particular, the case of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ has been analyzed to a large extent, see [48, 65, 110, 113, 114, 122], and more recently in [2, 30], but, no information about the extreme points of the set Lip_0^1 has been provided in the more general case $\mathcal{U} \neq \mathbb{R}$.

While the complete characterization of extreme points in Lip_0^1 presents significant technical challenges, this work aims to partially address this gap by focusing on a specific and intriguing framework. Our key contributions are:

1. We show that the extreme points of Lip_0^1 can be characterized under specific conditions. This characterization applies when $\mathcal{U}$ is a non-trivial, strictly convex real Banach space and $(\mathcal{X}, d)$ a finite metric space with distinct points $x_0, \ldots, x_n$, for $n \geq 1$.

2. We present a representation result for the space Lip_0^1 within the same setting. As we will see, this result generalizes the Minkowski–Carathéodory Theorem for infinite-dimensional spaces.

1.3 Dissemination

Journal articles

This thesis collects the content of the following papers:

- Bredies, K., Chirinos Rodriguez, J. and Naldi, E. “On extreme points and representer theorems for the Lipschitz unit ball on finite metric spaces”, *Archiv der Mathematik*, 2024, https://doi.org/10.1007/s00013-024-01978-y [26].

- Chirinos Rodriguez, J., De Vito, E., Molinari, C., Rosasco, L. and Villa, S. “On learning the optimal regularization parameter in inverse problems”, 2023. *In arxiv [46]*.

- Bredies, K., Chirinos Rodriguez, J. and Naldi, E. “Learning firmly nonexpansive operators”. *In preparation.*

Talks and participation to conferences

- *Regularization in a continuous setting*, TraDE-OPT Winter School, February 2021. Online.

- *A Supervised Learning Approach to Regularization Methods*, TraDE-OPT Summer School, July 2021. Online.

- *Learning Resolvent Operators*, “Mathematical analysis and Applied Mathematics Seminar”, April 2022. Universidad de La Laguna, La Laguna, Spain.

- *A Supervised Learning Approach to Regularization of Inverse Problems*, TraDE-OPT Summer School, July 2022. UCL, Louvain-La-Neuve, Belgium.
- *Learning Firmly Nonexpansive Operators*, "International Conference on Optimization and Decision Science", September 2022. Università degli Studi di Firenze, Florence, Italy.
- *Learning Firmly Nonexpansive Operators*, "Workshop on Mathematical Models for Plug-and-play Image Restoration", December 2022. Centre Culturel Irlandaise, Paris, France.
- *A Supervised Learning Approach to Regularization of Inverse Problems*, "Conference on Deep Learning for Computational Physics", July 2023. UCL, London, England.
- *A Supervised Learning Approach to Regularization of Inverse Problems*, "Seminars MDS", October 2023. University of Twente, Enschede, Netherlands.
- *On Learning the Optimal Regularization Parameter in Inverse Problems*, "Journées SMAI MODE 2024", March 2024. ENS Lyon, Lyon, France.

1.4 Outline

In Chapter 2, we introduce the notation and the main tools and results that we use. In particular, we introduce basic notions of Supervised Learning, Convex Analysis, and Inverse Problems. In Chapter 3 we set the problem of learning the regularization parameter in inverse problems, corresponding to [46]. Next, we dedicate Chapter 4 to the problem of learning firmly nonexpansive operators. Finally, Chapter 5 is devoted to the characterization of extreme points of the Lipschitz unit ball in finite metric spaces, and corresponds to [26]. We conclude this manuscript in Chapter 6, where we present the conclusions and further directions.

CHAPTER 2

Preliminaries

In this chapter, we will introduce all the preliminary tools that will be needed throughout the thesis. In particular, we fix the general notation that we will consider, and we recall basic notions of Supervised Learning, Convex Analysis and Inverse Problems. First, we introduce the general notation.

General notation

In the following, $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ denotes a real separable Hilbert space and $\mathcal{X}^*$ its topological dual space. The space $\mathcal{X}$ its endowed with the norm $\|\cdot\|_{\mathcal{X}}$ induced by the inner product; i.e. $\|x\|_{\mathcal{X}} = \sqrt{\langle x, x\rangle}$. If $\mathcal{X}$ is finite-dimensional, then we assume that the associated norm is the Euclidean one, $\|\cdot\|_2$. Moreover, if $(\mathcal{U}, \langle\cdot,\cdot\rangle_{\mathcal{U}})$ is also a real separable Hilbert space, given a linear and bounded operator $A\colon \mathcal{X} \to \mathcal{U}$, we denote by A^* its adjoint operator and, if A is injective, by A^{-1} its inverse. Moreover, we denote by $\sigma(A)$ the spectrum of A and by σ the elements of $\sigma(A)$. With $\|\cdot\|_{\text{op}}$ we denote the operator norm; i.e.

$$\|A\|_{\text{op}} := \sup_{\|x\|_{\mathcal{X}}=1} \|Ax\|_{\mathcal{U}}.$$

In the finite-dimensional case, we associate $\|A\|_{\text{op}} = \|A\|_2$; i.e. the maximum singular value of the matrix A. We next fix some notation and recall basic results in the context of probability theory.

Probability theory

In the following, $(\Omega, \mathcal{A}, P)$ will denote a probability space; i.e. a triple composed of the set of events Ω, the $\sigma-$algebra $\mathcal{A}$ and the probability measure P defined on Ω. In addition, we assume that, given a Hilbert space $\mathcal{X}$, it is endowed with the Borel $\sigma-$ algebra $\mathcal{B}$; i.e. the smallest $\sigma-$algebra that contains the open sets of $\mathcal{X}$. We say that a function $\bar{X} : \Omega \to \mathcal{X}$ is measurable if, for every $B \in \mathcal{B}$, we have

$$\bar{X}^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{A}.$$

If $\mathcal{X} = \mathbb{R}$, then we say that $\bar{X}$ is a *random variable*. Moreover, we say that a property E holds *almost surely* (abbreviated as a.s.) whenever

$$P\left(\{\omega \in \Omega \mid E \text{ holds }\}\right) = 1.$$

The latter concept will constantly appear along the thesis. In particular, whenever random variables are into play, properties must be understood in the almost sure sense, whether or not it has been explicitly mentioned. We finally recall some basic results in probability theory, which are typically known as concentration inequalities, and will be further used in the context of Statistical Learning theory.

Proposition 2.1. *([54, Proposition 2] and [36, Proposition 11]) Let $\bar{Z}_1, \dots, \bar{Z}_n$, $n \in \mathbb{N}$, be a sequence of i.i.d. random variables with mean $\mathbb{E}[\bar{Z}_i] = \mu$, $\mu > 0$, and variance $\mathbb{E}[|\bar{Z}_i - \mu|^2] \leq \sigma^2$ for every $i = 1, ..., n$. We have that*

- (Bernstein ineq.) *If there exist $B > 0$ such that $|\bar{Z}_i - \mu| \leq B$ a.s., then for all $\varepsilon > 0$ we have*
$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\bar{Z}_i - \mu\right| \geq \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{2(\sigma^2+\frac{1}{3}B\varepsilon)}}. \tag{2.0.1}$$
- (Hoeffding ineq.) *If there exist $B > 0$ such that $|\bar{Z}_i - \mu| \leq B$ a.s., then for all $\varepsilon > 0$ we have*
$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\bar{Z}_i - \mu\right| \geq \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{2B^2}}. \tag{2.0.2}$$
- *If there exists $B > 0$ such that $|\bar{Z}_i| \leq B$ a.s. for every $i = 1, ..., n$, then for all $\alpha, \varepsilon > 0$ we have that*
$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\bar{Z}_i - \mu\right| \geq \varepsilon + \alpha\sigma^2\right) \leq 2e^{-\frac{6n\alpha\varepsilon}{3+4\alpha B}}. \tag{2.0.3}$$

Next, we introduce the necessary concepts of Supervised Learning Theory that will be used along this text.

2.1 Supervised learning

Let $\mathcal{X}$ and $\mathcal{U}$ be real separable Hilbert spaces and let $(\Omega, \mathcal{A}, P)$ be a probability space. The goal of Supervised Learning [54, 137] is to determine a functional relationship between a pair of random variables $(\bar{X}, \bar{U}) \in \mathcal{X}\times\mathcal{U}$, with unknown joint Borel probability distribution μ on $\mathcal{X} \times \mathcal{U}$. Given a measurable loss function $\ell : \mathcal{U} \times \mathcal{U} \to [0, +\infty)$, the so-called *expected risk*,

$$L(f) := \mathbb{E}[\ell(f(\bar{X}), \bar{U})] = \int_{\mathcal{X}\times\mathcal{U}} \ell(f(\bar{x}), \bar{u})\, \mathrm{d}\mu(\bar{x}, \bar{u}), \tag{2.1.1}$$

for some $f\colon \mathcal{X} \to \mathcal{U}$, measures the average error, with respect to the loss ℓ, between $\bar{U}$ and $f(\bar{X})$. Ideally, one would like to find f^* minimizing the expected risk

$$f^* \in \arg\min_{f\in\mathcal{M}} L(f), \tag{2.1.2}$$

where $\mathcal{M}$ stands for the space of measurable functions from $\mathcal{X}$ to $\mathcal{U}$. In this context, two challenges need to be faced: first, the space of measurable functions $\mathcal{M}$ is "too big", making the above problem unfeasible, and second, since the probability distribution μ is unknown, a solution can not be computed. A common way to tackle the latter problem is to assume access to a finite training set $\{(\bar{X}_i, \bar{U}_i)\}_{i=1}^n$ of independent copies of $(\bar{X}, \bar{U})$. In this case, a natural discretization of (2.1.1) is given by the *empirical risk*

$$\widehat{L}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(\bar{X}_i), \bar{U}_i). \tag{2.1.3}$$

The function $\hat{f}$ minimizing the empirical risk within the space of measurable functions is known as *learning algorithm*. We say that a learning algorithm *generalizes* if it is able to correctly predict or classify new data.

Next, we recall some examples of loss functions used for linear regression and binary classification.

- Let $\mathcal{U} = \mathbb{R}$. Then, for regression problems it is common to consider, for every u, $u' \in \mathcal{U}$,
  - the square loss: $\ell(u, u') := (u - u')^2$,
  - the absolute loss: $\ell(u, u') := |u - u'|$, or
  - the ε-sensitive loss: $\ell(u, u') := \max\{|u - u'| - \varepsilon, 0\}$.
- Let $\mathcal{U} = \{\pm 1\}$. Then, for binary classification problems, one can consider, for every u, $u' \in \mathcal{U}$,
  - the missclasification loss: $\ell(u, u') := \begin{cases} 1, & \text{if } uu' \leq 0, \\ 0, & \text{else,} \end{cases}$
  - the hinge loss: $\ell(u, u') := \max\{0, 1 - uu'\}$, or
  - the logistic loss: $\ell(u, u') := \ln(1 + e^{-uu'})$.

A natural approach to measure the generalization properties of a learning algorithm $\hat{f}$ is given by the so-called *excess risk*

$$L(\hat{f}) - L(f^*).$$

Observe that the function above is stochastic since it depends on the set of random variables $\{(\bar{X}_i, \bar{U}_i)\}_{i=1}^n$. Hence, estimations of the excess risk should be stated either in expectation, where one aims to show that

$$\lim_{n\to\infty} \mathbb{E}\left[L(\hat{f}) - L(f^*)\right] = 0,$$

or in high probability, where one aims to prove that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(L(\hat{f}) - L(f^*) > \varepsilon\right) = 0.$$

If any of the above conditions are satisfied, we say that our learning algorithm $\hat{f}$ is *consistent*.

Often, it is common to consider minimizers of the empirical risk within a "smaller" set $\mathcal{H}$ that encodes prior information about the problem. The latter is typically known as *hypothesis space*, and popular examples of $\mathcal{H}$ are the space of linear functions (e.g. for linear regression) or the so-called Reproducing Kernel Hilbert Spaces (RKHS). The approach of finding minimizers of the empirical risk within $\mathcal{H}$ is known as Empirical Risk Minimization (ERM):

$$\hat{f}_{\mathcal{H}} \in \arg\min_{f\in\mathcal{H}} \widehat{L}(f) \tag{2.1.4}$$

Here, the objective is to consider $\mathcal{H}$ in such a way that $\hat{f}_{\mathcal{H}}$ is close to f^* in the sense explained above; i.e. to show that $\hat{f}_{\mathcal{H}}$ is consistent. If $L(f_{\mathcal{H}}) = L(f^*)$, where $f_{\mathcal{H}}$ is the minimizer of the expected risk within $\mathcal{H}$, we say that the hypothesis space $\mathcal{H}$ is *universal*. However, it may happen that the hypothesis space is not large enough and $L(f_{\mathcal{H}}) > L(f^*)$. If this is the case, proving any consistency result becomes difficult. In order to circumvent this issue, it is common to assume that $f^* \in \mathcal{H}$ (see, e.g. [36]), or just compare $L(\hat{f}_{\mathcal{H}})$ with $L(f_{\mathcal{H}})$ instead of $L(f^*)$. The objective of this latter approach is to show that, in expectation or in high probability, the quantity

$$L(\hat{f}_{\mathcal{H}}) - L(f_{\mathcal{H}}) \tag{2.1.5}$$

is small when $n \to \infty$. The term above is typically known as *sample error*, and accounts for the fact that a finite set is being used to find approximate solutions of $f_{\mathcal{H}}$. Next, we aim

at presenting a typical statistical learning result, showing that (2.1.5) goes to 0 as $n \to \infty$. To do so, we first recall that

$$\begin{aligned} L(\hat{f}_{\mathcal{H}}) - L(f_{\mathcal{H}}) &= L(\hat{f}_{\mathcal{H}}) - \widehat{L}(\hat{f}_{\mathcal{H}}) + \widehat{L}(\hat{f}_{\mathcal{H}}) - \widehat{L}(f_{\mathcal{H}}) + \widehat{L}(f_{\mathcal{H}}) - L(f_{\mathcal{H}}) \\ &\leq L(\hat{f}_{\mathcal{H}}) - \widehat{L}(\hat{f}_{\mathcal{H}}) + \widehat{L}(f_{\mathcal{H}}) - L(f_{\mathcal{H}}) \\ &\leq 2 \sup_{f \in \mathcal{H}} |L(f) - \widehat{L}(f)|, \end{aligned}$$

where we used that the quantity $\widehat{L}(\hat{f}_{\mathcal{H}}) - \widehat{L}(f_{\mathcal{H}})$ is negative by the definition of $\hat{f}_{\mathcal{H}}$. In order to state the result, we first recall the definition of covering number. Let $s > 0$ be a positive scalar. We define the *covering number* $\mathcal{N}(\mathcal{H}, s)$ to be the minimal number of balls of radius s covering $\mathcal{H}$. In the following, we state an estimate for (2.1.5), in the particular case where $\mathcal{H}$ is a compact subset of the space of real-valued continuous functions from $\mathcal{X}$, $\mathcal{H} \subseteq \mathcal{C}(\mathcal{X})$, and the loss ℓ is the square loss.

Theorem 2.2. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(\mathcal{X})$. Assume that there exists $M > 0$ such that $|f(x) - u| \leq M$ almost everywhere and that $\sup_{f \in \mathcal{H}} (f(x) - u)^2 \leq \sigma^2$. Then, for all $\varepsilon > 0$,*

$$P\left(\sup_{f \in \mathcal{H}} |L(f) - \widehat{L}(f)| \leq \varepsilon \right) \geq 1 - \mathcal{N}\left(\mathcal{H}, \tfrac{\varepsilon}{8M}\right) 2e^{-\frac{-n\varepsilon^2}{4(2\sigma^2)}}$$

Observe that the compactness assumption on $\mathcal{H}$ ensures, by definition, that the covering number is finite. A proof of this result can be found in [54, Theorem B], and is based on Bernstein inequality (2.0.1).

We next introduce the necessary notions of Convex Analysis.

2.2 Convex analysis

All of the results and definitions that will be presented in this section have been taken from [12, 13] when treating with infinite-dimensional Hilbert spaces, and [3, 87] in finite-dimensional settings. For now, we consider $\mathcal{X}$ to be a real separable Hilbert space. Let f be a function mapping from $\mathcal{X}$ to the extended real line, $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. We recall some basic definitions. The *domain* of f is given by

$$\operatorname{dom} f := \{x \in \mathcal{X} \mid f(x) \in \mathbb{R}\},$$

the *graph* of f is defined as

$$\operatorname{gra} f := \{(x, m) \in \mathcal{X} \times \mathbb{R} \mid f(x) = m\},$$

and the *sublevel set* of f at height $m \in \mathbb{R}$, denoted $\{f \leq m\}$, is

$$\{x \in \mathcal{X} \mid f(x) \leq m\}.$$

In addition, we say that f is (assume it has full domain for simplicity)

- *proper* if $-\infty \notin f(\mathcal{X})$ and $\operatorname{dom} f \neq \emptyset$,
- *convex* if for every x, $x' \in \mathcal{X}$ and for every $\lambda \in (0, 1)$ we have that

$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda) f(x'),$$

- *lower semicontinuous* at a point $x \in \mathcal{X}$ if, for every sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathcal{X}$ such that $x_n \to x$ as $n \to \infty$, then
$$f(x) \leq \liminf_{n\to\infty} f(x_n),$$
and
- lower semicontinuous if it is lower semicontinuous at every point in $\mathcal{X}$.

We say that $f \in \Gamma_0(\mathcal{X})$ if it is proper, convex and lower semicontinuous. We conclude this part by recalling some commonly used functions in the context of Convex Optimization. Given C a nonempty convex subset of $\mathcal{X}$, we denote by ι_C the indicator function onto C, defined as
$$\iota_C(x) := \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{else,} \end{cases}$$
and by N_C the *normal cone* of C, defined for every $x \in \mathcal{X}$ as the set
$$N_C(x) := \begin{cases} \{u \in \mathcal{X} \mid \langle x' - x, u\rangle \leq 0, \text{ for every } x' \in C\}, & \text{if } x \in C, \\ \emptyset, & \text{else.} \end{cases}$$
Next, we introduce proximal operators, basic notions of set-valued mappings, and further relationships between these two concepts.

2.2.1 Proximal operators and set-valued maps

Given a function $f \in \Gamma_0(\mathcal{X})$, the *proximal operator* of R is defined as ([13, Definition 12.23])
$$\text{prox}_f : x \mapsto \arg\min_z \; f(z) + \frac{1}{2}\|z - x\|^2_{\mathcal{X}}. \tag{2.2.1}$$
Note that, since for every $x \in \mathcal{X}$ the function $z \mapsto f(z) + (1/2)\|z - x\|_2^2$ is proper, strongly convex and lower semicontinuous, it has a unique minimizer [13, Corollary 11.17]. Therefore, the proximal mapping of any function $f \in \Gamma_0(\mathcal{X})$ is well-defined. We recall now some relevant examples of proximal operators.

Example 2.3. Let $\mathcal{X} = \mathbb{R}^d$, $1 \leq d < +\infty$. Let $C \subseteq \mathcal{X}$ be a nonempty, closed, convex set and consider $f = \iota_C$ the indicator function onto C, Then, for every $x \in \mathcal{X}$,
$$\text{prox}_f(x) = \text{proj}_C(x),$$
where proj_C denotes the orthogonal projection onto C.

Example 2.4. Let $\mathcal{X} = \mathbb{R}^d$, $1 \leq d < +\infty$, and let $f = \lambda\|\cdot\|_1$ for some scalar $\lambda \in (0, +\infty)$. Then, its proximal operator, denoted as $\mathcal{S}_\lambda$, is known as *soft-thresholding* operator and it is defined element-wise for every $x \in \mathcal{X}$ as
$$(\mathcal{S}_\lambda(x))_i := \begin{cases} 0, & \text{if } |x_i| \leq \lambda, \\ x_i - \lambda\text{sign}(x_i), & \text{if } |x_i| > \lambda, \end{cases} \tag{2.2.2}$$
for every $i \leq d$.

We next turn our attention to set-valued mappings: if $\mathcal{U}$ is a real Hilbert space, a *set-valued operator* $\mathcal{T} : \mathcal{X} \to 2^{\mathcal{U}}$, where $2^{\mathcal{U}}$ denotes the power set of $\mathcal{U}$, maps an element $x \in \mathcal{X}$ to the set $\mathcal{T}(x) \subseteq 2^{\mathcal{U}}$. In this setting, basic notions such as graph, domain or inverse can be naturally extended: the graph of a set-valued map is given by
$$\text{gra}\,\mathcal{T} = \{(x, u) \in \mathcal{X} \times \mathcal{U} \mid u \in \mathcal{T}(x)\},$$

the domain of $\mathcal{T}$ is defined as

$$\operatorname{dom}\mathcal{T} := \{x \in \mathcal{X} \mid \mathcal{T}(x) \neq \emptyset\},$$

and the inverse of $\mathcal{T}$, naturally denoted as $\mathcal{T}^{-1}$, is defined through its graph,

$$\operatorname{gra}\mathcal{T}^{-1} := \{(u, x) \in \mathcal{U} \times \mathcal{X} \mid (x, u) \in \operatorname{gra}\mathcal{T}\}.$$

In particular, we are interested in those set-valued operators $\mathcal{T}$ that are monotone and maximally monotone. A set-valued operator $\mathcal{T}\colon \mathcal{X} \to 2^{\mathcal{U}}$ is *monotone* if

$$\langle u - u', x - x'\rangle \geq 0, \quad \text{for every } (x, u),\ (x', u') \in \operatorname{gra}\mathcal{T}.$$

The operator $\mathcal{T}$ is said to be *maximally monotone* if there exists no monotone operator $\mathcal{T}'$ such that $\operatorname{gra}\mathcal{T} \subsetneq \operatorname{gra}\mathcal{T}'$. A relevant example of a maximally monotone operator is the subdifferential of proper, convex and lower semicontinuous functions (see [13, Theorem 20.25]): given a function $f \in \Gamma_0(\mathcal{X})$, its *subdifferential* is the set-valued operator $\partial f\colon \mathcal{X} \to 2^{\mathcal{U}}$, defined for every $x \in \operatorname{dom} f$ as the set

$$\partial f(x) := \{u \in \mathcal{X} \mid \text{for every } y \in \mathcal{X},\ f(x) + \langle y - x,\ u\rangle_{\mathcal{X}} \leq f(y)\}.$$

If $x \notin \operatorname{dom} f$, then $\partial f(x) = \emptyset$. The latter concept is a generalization of the gradient for convex, non-differentiable functions. In short, an element $u \in \mathcal{X}$ belongs to $\partial f(x)$ for some $x \in \mathcal{X}$ if u is the slope of an affine minorant of f that coincides with f at x. In addition, if f is also differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$, where $\nabla f(x)$ denotes the gradient of f at x. In the following result, we recall when the subdifferential of a function in $\Gamma_0(\mathcal{X})$ is nonempty.

Lemma 2.5. *[13, Theorem 9.23] Let $f \in \Gamma_0(\mathcal{X})$ and let $x \in \operatorname{int}(\operatorname{dom} f)$. Then, $x \in \operatorname{dom}\partial f$.*

We recall some common examples of subdifferential maps.

Example 2.6. [13, Example 16.13] Let $\mathcal{X} = \mathbb{R}^d$, $1 \leq d < +\infty$, and let $C \subseteq \mathcal{X}$ be a nonempty convex set. Then $\partial\iota_C = N_C$ where N_C is the normal cone of C above defined.

Example 2.7. [7, Remark 1.1] Let $\mathcal{X} = \mathbb{R}^d$, $1 \leq d < +\infty$, consider a general norm $\|\cdot\|$ in $\mathbb{R}^d$ (not necessarily the Euclidean one) and denote as $\|\cdot\|_*$ its dual norm. Then,

$$\partial\|\cdot\|(x) = \{\eta \in \mathbb{R}^d \mid \langle \eta, x\rangle = \|x\|,\ \|\eta\|_* \leq 1\}.$$

In Chapter 5, we are interested in studying the so-called *resolvent operator* of set-valued operators $\mathcal{T}$, defined as

$$J_{\mathcal{T}} := (\mathrm{Id} + \mathcal{T})^{-1}.$$

Actually, if $\mathcal{T} = \lambda\partial f$ for some $f \in \Gamma_0(\mathcal{X})$ and some $\lambda \in (0, +\infty)$, it is possible to derive the following property.

Lemma 2.8. *[13, Proposition 16.44] Let $f \in \Gamma_0(\mathcal{X})$. Then,*

$$\operatorname{prox}_f = J_{\partial f}.$$

Therefore, the proximal operator of a proper, convex and lower semicontinuous function coincides with the resolvent operator of its subdifferential, establishing in this way a natural link between proximal operators, and subdifferentials, of functions in $\Gamma_0(\mathcal{X})$. However, note that, in general, the resolvent of a maximally monotone operator is not necessarily a proximal operator.

In order to state the next result, we first recall some relevant definitions: we say that an operator $T : \mathcal{X} \to \mathcal{X}$ is *firmly nonexpansive* if

$$\|T(x) - T(x')\|^2 + \|(x - x') + (T(x') - T(x))\|^2 \leq \|x - x'\|^2, \text{ for every } x, x' \in \mathcal{X}.$$

In addition, we say that T is *nonexpansive* if

$$\|T(x) - T(x')\| \leq \|x - x'\|, \text{ for every } x, x' \in \mathcal{X}.$$

In other words, nonexpansive operators are $1-$ Lipschitz functions mapping from $\mathcal{X}$ to itself. We are now ready to state the desired result.

Lemma 2.9. *[13, Corollary 23.9] The mapping $T : \mathcal{X} \to \mathcal{X}$ is the resolvent of a maximally monotone operator if and only if T is firmly nonexpansive.*

As a direct consequence, by combining both Lemma 2.8 and Lemma 2.9, we can write the following corollary.

Corollary 2.10. *Let $f \in \Gamma_0(\mathcal{X})$. Then, prox_f is a firmly nonexpansive operator.*

In addition, the following lemma holds.

Lemma 2.11. *[13, Proposition 4.4] Let $T\colon \mathcal{X} \to \mathcal{X}$ be an operator. Then, T is firmly nonexpansive if and only if the operator $2T - \mathrm{Id}$ is nonexpansive.*

Hence, for every firmly nonexpansive operator $T\colon \mathcal{X} \to \mathcal{X}$, the operator $N := 2T - \mathrm{Id}$ is nonexpansive and, for every nonexpansive operator $N\colon \mathcal{X} \to \mathcal{X}$, we can construct a firmly nonexpansive operator as $T := \frac{1}{2}(N + \mathrm{Id})$. By combining both Corollary 2.10 and Lemma 2.11, we have been able to show that a larger class where to study proximal operators of convex functions is actually the set of nonexpansive operators. This sequence of results will be useful, in particular, in the context of Chapter 5.

We next introduce the so-called Bregman divergence, we recall the definition of Legendre functions, and provide a link between these two concepts.

2.2.2 Bregman divergence and Legendre functions

Given $f \in \Gamma_0(\mathcal{X})$, not necessarily differentiable, the *Bregman divergence* of f is defined, for every x, $x' \in \mathcal{X}$, as

$$D_f(x, x') := \begin{cases} f(x) - f(x') - \langle s_f(x'), x - x' \rangle_{\mathcal{X}}, & \text{if } x' \in \operatorname{int}(\operatorname{dom} f), \\ +\infty, & \text{elsewhere,} \end{cases} \tag{2.2.3}$$

where $s_f(x')$ is an element of $\partial f(x')$. We recall that, by Lemma 2.5, $\partial f(x')$ is nonempty as long as $x' \in \operatorname{int}(\operatorname{dom} f)$. In addition, if both x and x' belong to $\operatorname{int}(\operatorname{dom} f)$, we can consider also the *symmetric Bregman distance*, that is

$$d_f(x, x') := D_f(x, x') + D_f(x', x) = \langle s_f(x) - s_f(x'), x - x' \rangle_{\mathcal{X}}. \tag{2.2.4}$$

Observe that both the Bregman divergence and the symmetric Bregman distance depend on the choice of the specific subgradient $s_f(x)$ (and $s_f(x')$). If such a choice is not specified, then it is common to choose the one with minimal norm. Obviously, for every $x \in \mathcal{X}$ such that f is differentiable at x, then the subdifferential is single valued in x and so $s_f(x) = \nabla f(x)$.

In Chapter 4, we will need to consider a suitable projection operator with respect to the Bregman divergence. Given a set $C \subseteq \mathcal{X}$, $x \in \mathcal{X}$, and $f \in \Gamma_0(\mathcal{X})$, we define the Bregman projection of f onto C as

$$\pi_C(x) := \operatorname*{arg\,min}_{z \in C} D_f(z, x). \tag{2.2.5}$$

Next, we aim to provide necessary conditions so that, for every $x \in \mathcal{X}$, $\pi_C(x)$ exists and is unique. To do so, we first introduce some preliminary concepts.

Definition 2.12. [12, Definition 5.2] A function $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ such that $f \in \Gamma_0(\mathcal{X})$ is said to be:

- *essentially smooth* if its subdifferential ∂f is locally bounded and single valued on its domain,
- *essentially strictly convex* if $(\partial f)^{-1}$ is locally bounded on its domain and f is strictly convex on every convex subset of $\operatorname{dom} \partial f$, and
- *Legendre* if it is both essentially smooth and essentially strictly convex.

In particular, the essencial smoothness property for functions $f \in \Gamma_0(\mathcal{X})$ can be further characterized.

Lemma 2.13. *[12, Theorem 5.6] Let $f \in \Gamma_0(\mathcal{X})$. Then, f is essentially smooth if and only if $\operatorname{dom} \partial f = \operatorname{int}(\operatorname{dom} f) \neq \emptyset$ and ∂f is single valued on its domain.*

We are now ready to characterize the properties of $f \in \Gamma_0(\mathcal{X})$ that make (2.2.5) well-defined.

Lemma 2.14. *[12, Corollary 7.9] Let f be a Legendre function, C a closed, convex subset of $\mathcal{X}$ such that $C \cap \operatorname{int}(\operatorname{dom} f) \neq \emptyset$ and let $x \in \operatorname{int}(\operatorname{dom} f)$. Then, $\pi_C(x)$ is a singleton and $\pi_C(x) \in \operatorname{int}(\operatorname{dom} f)$.*

By assuming that f is Legendre, the Bregman projection is univocally defined, meaning that it does not depend on the choice of the subgradient: if $x \notin \operatorname{int}(\operatorname{dom} f)$, then $D_f(z, x) = +\infty$. Otherwise, $x \in \operatorname{int}(\operatorname{dom} f) = \operatorname{dom} \partial f$ by Lemma 2.13, where the subdifferential of f is single valued.

The following result extends the classical Pythagorean equality in this setting, and involves the Bregman divergence and its projection onto a closed, convex set $C \subset \mathcal{X}$.

Lemma 2.15. *Let f be a Legendre function, let $C \subset \mathcal{X}$ be a closed, convex set, and let $x \in \operatorname{int}(\operatorname{dom} f)$. Then, for every $z \in C$ we have that*

$$D_f(z, x) \geq D_f(z, \pi_C(x)) + D_f(\pi_C(x), x). \tag{2.2.6}$$

Proof. We write the optimality condition of the set of minimizers of $D_f(\cdot, x)$. By definition,

$$\begin{aligned} y \in \operatorname*{arg\,min}_{z \in \mathcal{X}} D_f(z, x) &\iff y \in \operatorname*{arg\,min}_{z \in C} f(z) - f(x) - \langle s_f(x), z - x \rangle + \iota_C(z), \\ &\iff y \in \operatorname*{arg\,min}_{z \in C} f(z) - \langle s_f(x), z \rangle + \iota_C(z). \end{aligned}$$

Moreover, by Example 2.6, $\partial \iota_C = N_C$. We therefore get that

$$\begin{aligned} y \in \operatorname*{arg\,min}_{z \in C} D_f(z, x) &\iff 0 \in \partial f(y) - s_f(x) + N_C(y) \\ &\iff \langle s_f(x) - \partial f(y), u - y \rangle \leq 0, \quad \text{for every } u \in C, \end{aligned}$$

by the definition of normal cone. Next, by Lemma 2.14, we know that such a minimizer exists, is unique, and is denoted as $y = \pi_C(x)$. Observe that the inequality above holds for every element in $\partial f(y)$. In particular, for $s_f(\pi_C(x)) \in \operatorname{dom} \partial f(\pi_C(x))$. With this choice, the latter inequality reads, for every $u \in C$, as

$$\langle s_f(x) - s_f(\pi_C(x)), u - \pi_C(x) \rangle \leq 0. \tag{2.2.7}$$

Finally, (2.2.6) holds true if and only if, for every $z \in C$,

$$-\langle s_f(x), z - x \rangle \geq -\langle s_f(\pi_C(x)), z - \pi_C(x) \rangle - \langle s_f(x), \pi_C(x) - x \rangle.$$

By developing the above condition, we derive that

$$\langle s_f(x) - s_f(\pi_C(x)), z - \pi_C(x) \rangle \leq 0,$$

Since $z \in C$, this latter condition coincides with the optimality condition derived in (2.2.7), proving the result. □

We conclude this preliminary section of Convex Analysis by recalling the notion of extreme point and its connection with the so-called Minkowski–Carathéodory Theorem.

2.2.3 Extreme Points and Minkowski–Carathéodory Theorem

In this section, we restrict ourselves to the finite-dimensional setting by fixing $\mathcal{X} = \mathbb{R}^d$, $1 \leq d < \infty$. The *convex hull* – or *convex envelope* – of C, denoted as $\operatorname{conv}(C)$, is the smallest convex subset of $\mathcal{X}$ containing C. We recall the following result by Carathéodory [87, Theorem III.1.3.6].

Theorem 2.16. *Let $C \subseteq \mathcal{X}$ be a nonempty set. Then, any element $x \in \operatorname{conv}(C)$ can be written as the convex combination of at most $d + 1$ elements of C.*

Moreover, given a convex set $C \subseteq \mathcal{X}$, an *extreme point* of C is a point x in C such that $C \setminus \{x\}$ is convex. Equivalently, an extreme point is an element $x \in C$ such that, if $x = \frac{1}{2}x^1 + \frac{1}{2}x^2$ with $x^1, x^2 \in C$, then $x^1 = x^2 = x$. We denote by $\operatorname{ext}(C)$ the set of extreme points of C.

However, not every convex set has extreme points. For instance, the upper half plane in $\mathbb{R}^2$,

$$H = \{(x, y) \in \mathbb{R}^2 \mid y \geq 0\},$$

is indeed a convex set and does not have extreme points. We next provide a necessary condition for a convex set C to have extreme points.

Lemma 2.17. *Let $C \subseteq \mathbb{R}^d$, $d \geq 1$ be a convex set. If C is compact, then $\operatorname{ext}(C) \neq \emptyset$.*

Proof. Since C is compact, by the Weierstrass' Theorem, there exists $\bar{x} \in C$ such that $\|x\|_2^2 \leq \|\bar{x}\|_2^2$ for every $x \in C$. We will now show that $x \in \operatorname{ext}(C)$. Let x^1, $x^2 \in C$ such that $\bar{x} = \frac{1}{2}x^1 + \frac{1}{2}x^2$ and suppose that $x^1 \neq x^2$. Observe that, for every, a, $b \in \mathbb{R}$ with $a \neq b$, it holds true that

$$\frac{1}{4}\|a + b\|_2^2 = \frac{1}{2}\|a\|_2^2 + \frac{1}{2}\|b\|_2^2 - \frac{1}{4}\|a - b\|_2^2.$$

Since $a \neq b$, we get that

$$\frac{1}{4}\|a + b\|_2^2 < \frac{1}{2}\|a\|_2^2 + \frac{1}{2}\|b\|_2^2.$$

With this, we derive that

$$\|\bar{x}\|_2^2 = \left\| \frac{x^1}{2} + \frac{x^2}{2} \right\|_2^2 < \frac{1}{2}\|x^1\|_2^2 + \frac{1}{2}\|x^2\|_2^2$$
$$\leq \frac{1}{2}\|\bar{x}\|_2^2 + \frac{1}{2}\|\bar{x}\|_2^2 = \|\bar{x}\|_2^2,$$

which is a contradiction. We hence conclude that $\bar{x} \in \text{ext}(C)$ and so $\text{ext}(C) \neq \emptyset$. □

Since we are in the finite-dimensional setting, [3, Corollary 5.33] tells us that the convex hull of a compact set is again compact. Therefore, if C is a compact set in $\mathcal{X}$, its convex hull has extreme points. The following result is due to Minkowski [87, Theorem III.2.3.4].

Theorem 2.18. *Let $C \subseteq \mathcal{X}$ be a nonempty, convex and compact set. We have that*

$$C = \text{conv}(\text{ext}(C)).$$

By combining both Theorem 2.16 and Theorem 2.18, we derive the main result of this section, known as Minkowski-Carathéodory Theorem.

Corollary 2.19. *Let $C \subseteq \mathcal{X}$ be a nonempty, convex and compact set of dimension $k \leq d$. Then, for any $x \in C$, there exist $x^1, ..., x^{k+1} \in \text{ext}(C)$ and scalars $\lambda_1, ..., \lambda_{k+1} \geq 0$ with $\sum_{i=1}^{k+1} \lambda_i = 1$ such that $x = \sum_{i=1}^{k+1} \lambda_i x^i$.*

In other words, the above result gives conditions for a set C so that any point inside C can be written as a convex combination of a finite number of particular atoms in C. As we mentioned in the introduction, Corollary 2.19 motivates the introduction of the so-called *representer theorems*, first introduced in the context of Reproducing Kernel Hilbert Spaces [119].

2.3 Inverse problems

In this section, we provide a brief introduction to inverse problems, defining ill-posed inverse problems and presenting basic definitions and examples of classic regularization techniques. We start by giving a proper definition of the concept of ill-posedness.

2.3.1 Ill-posed (linear) inverse problems

Let $(\mathcal{U}, \langle\cdot,\cdot\rangle_{\mathcal{U}})$ and $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ be real separable Hilbert spaces and let $A : \mathcal{U} \to \mathcal{X}$ be a linear and bounded operator between the Hilbert spaces $\mathcal{U}$ and $\mathcal{X}$. The goal of an inverse problem is, given a measurement operator A and observations of the form $x = Au^* \in \mathcal{X}$, to retrieve $u^* \in \mathcal{U}$. We recall that this amouts to solving the following equation

$$x = Au^*. \tag{2.3.1}$$

In general, finding solutions of the above problem is hard. Most problems of interest are modelled by a linear operator A which is in general not injective or, even if it is, its inverse has large norm. This phenomena makes the reconstruction task difficult. When this is the case, we say that problem (2.3.1) is ill-posed. The inverse problem (2.3.1) is said to be *ill-posed* if at least one of the following conditions is satisfied [76]:

- a solution does not exist,
- the solution is not unique,

- solutions are unstable with respect to the measurements.

We provide some comments regarding these conditions. First, if a solution does not exist, then it is common to relax the notion of solution. For instance, one could consider solutions u in the least-squares sense: an element $u \in \mathcal{U}$ satisfying

$$\|Au - x\|_{\mathcal{X}} = \inf_{z \in \mathcal{U}} \|Az - x\|_{\mathcal{X}}$$

is said to be a *least-squares* solution of (2.3.1). The second concern is that least-square solutions may not be unique. In this case, a natural approach is to consider least-square solutions of minimal norm: an element $u \in \mathcal{U}$ that is the least-squares solution of (2.3.1) and satisfies

$$\|u\|_{\mathcal{U}} = \inf\{\|z\| \mid z \text{ is a least-squares solution of (2.3.1)}\},$$

is said to be the *best-approximate solution* of (2.3.1). The latter definition motivates us to introduce the so-called *Moore–Penrose pseudo-inverse* of A, denoted by $A^\dagger$, and defined as the operator that maps each measurement x to the least-squares solution of minimal norm of (2.3.1). A precise definition is given in the following.

Definition 2.20. [64, Definition 2.2] Let $A : \mathcal{U} \to \mathcal{X}$ be a linear and bounded operator. The Moore–Penrose pseudo-inverse of A is defined as the unique linear extension of $\tilde{A}^{-1}$, where

$$\tilde{A} := A|_{\ker(A)^\perp} : \ker(A)^\perp \to \operatorname{ran}(A),$$

and such that

$$\operatorname{dom}(A^\dagger) := \operatorname{ran}(A) + \operatorname{ran}(A)^\perp, \quad \ker(A^\dagger) := \operatorname{ran}(A)^\perp.$$

Such solution will be denoted as $u^\dagger = u^\dagger(x)$. In fact, we have the following result.

Theorem 2.21. *[64, Theorem 2.5] Let $x \in \operatorname{dom}(A^\dagger)$. Then, (2.3.1) admits a unique best approximate solution $u^\dagger$. The set of all least-squares solutions is $u^\dagger + \ker(A)$.*

So far, we have operated under the assumption of having access to the precise observation x through the forward operator A. However, this is not the case in general. In reality, it is common to observe only noisy versions of the measurement x. This discrepancy may arise, for instance, due to limitations in the measurement process, environmental interferences, or other sources of disturbance affecting the measurements. Consequently, it is more appropiate to reframe problem (2.3.1) as follows:

$$x = Au^* + \varepsilon \tag{2.3.2}$$

where $\varepsilon \in \mathcal{X}$ is an additive term modelling the noise, and can be either deterministic or stochastic. If there exists $\tau > 0$ such that $\|\varepsilon\|_{\mathcal{X}} \leq \tau$ (or $\mathbb{E}[\|\varepsilon\|_{\mathcal{X}}] \leq \tau$ if ε is a random variable) we call τ the noise level of ε. This revised formulation provides a more accurate model. Below, we outline some relevant examples in the context of signal processing, image recovery, and matrix reconstruction.

Example 2.22. Let $\mathcal{U}$ and $\mathcal{X}$ be a real Hilbert spaces. The denoising problem consists in recovering a clean the ground-truth term $u^* \in \mathcal{U}$ from noisy observations $x \in \mathcal{X}$. It writes as

$$x = u^* + \varepsilon.$$

For example, we can think of ε to be sampled from the standard Gaussian distribution $N(0, \tau^2 \operatorname{Id})$, $\tau > 0$, and the forward operator A equal to the identity. Two relevant examples in this framework are

- $\mathcal{X} = \mathcal{U} = \mathbb{R}^d$, $1 \leq d < +\infty$, and $u^* \in \mathbb{R}^d$ an s-sparse signal, $s \ll d$; i.e. a signal with s non-zero entries, which corresponds to the problem of sparse signal denoising.

- $\mathcal{X} = \mathcal{U} = \mathbb{R}^{N\times N}$, $1 \leq N < +\infty$ and $u^* \in \mathbb{R}^{N\times N}$ a clean discrete image, which corresponds to the problem of image denoising.

The latter examples will be relevant in the forthcoming chapters. Additionally, we will deal with the sparse signal deblurring problem in the experimental section of Chapter 3. For this reason, we describe it below.

Example 2.23. Considering the particular framework of $\mathcal{X} = \mathcal{U} = \mathbb{R}^d$, $1 \leq d < +\infty$, the problem of sparse signal deblurring consists in recovering the clean signal u^* from blurry, noisy observations x, and can be written as

$$x = Au^* + \varepsilon,$$

where ε follows the same model as above. In this case, A is a linear convolution operator that adds blur to the image. For example, following [108], we may think of A to be the following map

$$u \in \mathbb{R}^d \mapsto Au = h * u \in \mathbb{R}^d,$$

with h a Gaussian and "$*$" denoting the discrete convolution operator.

Example 2.24. If $\mathcal{U} = \mathbb{R}^{N\times N}$, $1 \leq N < +\infty$, the problem of subsampled MRI consists in recovering a discrete image $u^* \in \mathbb{R}^{N\times N}$ by having access to an incomplete version from a subsampling of its (discrete) Fourier transform in the so-called $k-$space – or frequency domain –. It can be expressed as

$$x = D\mathcal{F}u^* + \varepsilon,$$

where ε models the noise as in the above cases, $\mathcal{F}$ denotes the (discrete) Fourier transform and D denotes a general subsampling operator. In this case, $A = D\mathcal{F}$.

Example 2.25. Let $\mathcal{U} = \mathbb{R}^{m\times d}$, $1 \leq m < d < +\infty$. Consider a matrix $U^* \in \mathcal{U}$, assumed to be low-rank, with real entries $(u^*_{i,j})_{i\leq m,\ j\leq d}$, and let Θ denote a subset of the complete set of entries, $\Theta \subset \{1, ..., m\} \times \{1, ..., d\}$. Then, let P_Θ denote the linear map that encodes the available information about U^*, defined element-wise as

$$(\mathrm{P}_\Theta(U^*))_{i,j} = \begin{cases} u^*_{i,j}, & \text{if } (i,j) \in \Theta, \\ 0, & \text{otherwise.} \end{cases}$$

The inverse problem of matrix completion consists in finding U^* only from the information entailed in X (see [35]), and can be expressed as

$$X = \mathrm{P}_\Theta(U^*).$$

In general, the pseudo-inverse is not a continuous operator (see e.g. [64, Proposition 2.4]). Hence, if we only have access to noisy versions of the measurements x^ε, then $A^\dagger(x^\varepsilon)$ does not necessarily need to be close to $u^\dagger = A^\dagger x$, where $x = Au^*$. This motivates the construction of continuous operators that converge to the best-approximate solution $u^\dagger$. The latter task is precisely the purpose of regularization theory, which we introduce below.

2.3.2 Regularization of inverse problems

The goal of regularization theory is to construct an operator that outputs stable approximations of the solution. In addition, such operator should contain prior knowledge about the problem. For instance, if available, it is common to construct operators that depend on the noise level or on the forward operator itself. The inherent variability in these considerations make the construction of such an operator akin to an art form. As mentioned in the introduction, a classical way to tackle instability is to find solutions of the following variational problem

$$\min_u \ell(Au, x) + \lambda R(u), \tag{2.3.3}$$

for some regularization parameter $\lambda \in (0, +\infty)$, some discrepancy $\ell \colon \mathcal{X} \times \mathcal{X} \to (0, +\infty)$ and fixed regularizer $R \colon \mathcal{U} \to (0, +\infty]$. In this manuscript, and in particular in Chapter 3, we will consider operators $u_\lambda : \mathcal{X} \to \mathcal{U}$ indexed by a positive scalar $\lambda \in (0, +\infty)$, that are solutions of the variational problem above; i.e. for every $\lambda \in (0, +\infty)$, $x \in \mathcal{X}$,

$$u_\lambda(x) = u_\lambda \in \arg\min_{u \in \mathcal{U}} \ell(Au, x) + \lambda R(u).$$

We now recall the definition of a convergent regularization method (see [64, Chapter 3] for more details).

Definition 2.26. Let $u_\lambda : \mathcal{X} \to \mathcal{U}$ be a family of continuous maps indexed by $\lambda \in (0, +\infty)$, and define

$$\tilde{u} := \arg\min_u \{R(u) \,|\, u \in \arg\min_z \ell(Az, x)\}$$

Then, the family $(u_\lambda)_{\lambda>0}$ is said to be a convergent *regularization method* for the inverse problem (2.3.2) if there exists a parameter choice rule $\lambda = \lambda(\tau, x^\varepsilon)$ such that the family $u_{\lambda(\tau,x^\varepsilon)}(x^\varepsilon)$ converges to $\tilde{u}$ as the noise level vanishes. More precisely, we have

$$\limsup_{\tau \to 0} \|u_{\lambda(\tau,x^\varepsilon)}(x^\varepsilon) - \tilde{u}\|_{\mathcal{U}} = 0, \qquad \text{as long as } \limsup_{\tau \to 0} \lambda(\tau, x^\varepsilon) = 0.$$

If ℓ is the square loss and $R = \|\cdot\|^2_{\mathcal{U}}$, then $\tilde{u} = u^\dagger$ and we recover the classical definition of regularization method given in [64], since the family u_λ would converge pointwise to the pseudo-inverse.

The latter definition indicates that the regularization parameter should depend on the noise level τ, in the sense that $\lambda \to 0$ whenever $\tau \to 0$. Actually, the so-called Bakushinskii veto [8] tells us that no family $(u_\lambda)_{\lambda>0}$ can be a convergent regularization method if the regularization parameter λ does not depend on the noise level τ. Therefore, if we have access to such information, it should be used in order to construct a corresponding parameter choice rule. We distinguish the main directions that have been explored among the years:

- *a priori* rules, that depend on the noise level τ and, possibly, on some prior information about the solution u^*,
- *a posteriori* rules, which depend both on τ and on the noisy measurements x^ε, and
- *heuristic* rules, which only depend on the noisy observations x^ε.

The dependence on the solution u^* for a-priori choices is related to some regularity assumption, which is usually characterized by a certain smoothness parameter. Such choices are primarily of theoretical interest. The main reason is that they allow to derive sharp convergence rates, in the sense that they match optimal lower bounds [64, Chapter 3].

A-posteriori rules were widely studied in the past century. We mention, for instance, the Morozov discrepancy principle [104, 135], the balancing principle [99, 132] or the monotone error rule [77, 126]. As we mentioned in the introduction, in most practical problems, neither the noise level τ nor any smoothness assumptions are known, making it impossible to design an automatic parameter selection method. Heuristic rules are the common choice in practice, since they only depend on the measurements x^ε. Among them, we mention generalized cross-validation [73, 139], the quasi-optimality criterion [131, 130], the L-curve method [79], or methods based on an estimation of the Mean Squared Error, e.g. [124] (see [9] for a meticulous comparative study). Another popular example of heuristic rules are the data-driven selection methods, described in the introduction. In Chapter 3 we will focus on providing convergence guarantees for this approach in a stochastic setting.

Next, we introduce a popular family of linear regularization approaches, called spectral regularization methods, that will be further considered in Chapter 3. Therein, we will show that data-driven parameter selection methods are provably convergent when considering spectral regularization methods.

Spectral regularization methods

A standard choice for both both ℓ and R in (2.3.3) is given by fixing ℓ to be the square loss, $\ell(x, x') = \frac{1}{2}\|x - x'\|_{\mathcal{X}}^2$ and $R = \|\cdot\|_{\mathcal{U}}^2$. For such a choice, the solution is unique and can be explicitly computed. It writes, for every $\lambda \in (0, +\infty)$, as,

$$u_\lambda(x) = (A^*A + \lambda \,\mathrm{Id})^{-1} A^* x. \tag{2.3.4}$$

The expression above is typically known as *Tikhonov regularizer* (see [64, Chapter 5]). Actually, it can be naturally generalized by considering a family of spectral functions $g_\lambda : (0, \|A\|_{\mathrm{op}}^2] \to \mathbb{R}$, indexed by a parameter $\lambda \in (0, +\infty)$, called *spectral regularization methods* or *filter functions*, as follows

$$u_\lambda(x) := g_\lambda(A^*A)A^*x. \tag{2.3.5}$$

By definition, every g_λ is defined on $(0, \|A\|_{\mathrm{op}}^2]$, a set containing the spectrum $\sigma(A^*A)$; and each g_λ is a continuous function that approximates the inverse $1/\sigma$; i.e., for every $\sigma \in [0, \|A\|_{\mathrm{op}}^2]$, we have

$$\lim_{\lambda \to 0} g_\lambda(\sigma) = \frac{1}{\sigma}.$$

Remark 2.27. In order to give a precise definition to the expression 2.3.5, we need to recall classical spectral calculus results. First, since g_λ, $\lambda \in (0, +\infty)$, is continuous on $\sigma(A^*A) \subseteq [0, \|A\|_{\mathrm{op}}^2]$ by assumption, it is the uniform limit of a sequence of polynomials $(p_{n_t})_{n_t \in \mathbb{N}}$. Moreover, the sequence $(p_{n_t}(A^*A))_{n \in \mathbb{N}}$ is well defined (see [84, Theorem VI.31.1]) and converges in norm by [84, Theorem VI.32.1]. The operator $g_\lambda(A^*A)$ is defined as its limit.

Following [10, 64], we define the family of regularization methods g_λ, $\lambda \in (0, +\infty)$, as follows.

Definition 2.28. The family $g_\lambda : (0, \|A\|_{\mathrm{op}}^2] \to \mathbb{R}$, $\lambda \in (0, +\infty)$, is said to be a spectral regularization method if

(i) there exists a constant $B > 0$ such that

$$\sup_{\sigma \in (0, \|A\|_{\mathrm{op}}^2]} |\sigma g_\lambda(\sigma)| \leq B, \quad \text{for every } \lambda \in (0, +\infty),$$

(ii) there exists a constant $C > 0$ such that

$$\sup_{\sigma \in (0, \|A\|_{\mathrm{op}}^2]} |g_\lambda(\sigma)| \leq \frac{C}{\lambda}, \quad \text{for every } \lambda \in (0, +\infty),$$

(iii) there is a constant $\gamma > 0$ such that

$$\sup_{\sigma \in (0, \|A\|_{\mathrm{op}}^2]} |1 - g_\lambda(\sigma)\sigma| \leq \gamma, \quad \text{for every } \lambda \in (0, +\infty),$$

(iv) there is a constant $\bar{\nu} > 0$, called *qualification parameter*, such that, for all $\nu \leq \bar{\nu}$, we have

$$\sup_{\sigma \in (0, \|A\|_{\mathrm{op}}^2]} \sigma^\nu |1 - \sigma g_\lambda(\sigma)| \leq D_\nu \lambda^\nu, \quad \text{for every } \lambda \in (0, +\infty),$$

where D_ν depends on ν but not on λ.

The above conditions are satisfied by a large class of examples. We recall here some of them (see [10, 64]).

Example 2.29 (Tikhonov regularization)**.** By applying Remark 2.27, the expression $(A^*A + \lambda \mathrm{Id})^{-1}$ in (2.3.4) can be rewritten in terms of the singular values of A^*A in order to obtain the Tikhonov filter, which corresponds, for every $\lambda \in (0, +\infty)$, to the function

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

In this example, it is clear that $g_\lambda(\sigma) \to \frac{1}{\sigma}$ as $\lambda \to 0$. Moreover, $B = C = \gamma = D_\nu = 1$ and the qualification is $\bar{\nu} = 1$.

Example 2.30 (Landweber iteration)**.** If we consider the least-squares problem

$$\min_u \|Au - x\|_{\mathcal{X}}^2,$$

performing the Gradient Descent algorithm in the problem above receives the name of *Landweber iteration*. Given $\lambda \in (0, +\infty)$, with the reparametrization $k = \lfloor 1/\lambda \rfloor$, the Landweber filter is defined as $g_\lambda = (\mathrm{Id} - \omega A^*A)^k$ with ω denoting a constant stepsize. For this example, $B = C = \gamma = 1$ and (iv) holds for any qualification parameter $\bar{\nu} \geq 0$. Finally, $D_\nu = 1$ if $v \in (0, 1]$ and $D_\nu = \nu^\nu$ otherwhise.

Example 2.31 (Spectral cut-off)**.** The spectral cut-off filter, or Truncated Singular Value Decomposition (TSVD) corresponds, for every $\lambda \in (0, +\infty)$, to the function

$$g_\lambda(\sigma) = \begin{cases} \frac{1}{\sigma}, & \text{if } \sigma \geq \lambda, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding constants are $B = C = \gamma = D_\nu = 1$ and the qualification parameter is arbitrary.

The class of filter functions that fit in this setting is way larger. We mention for instance heavy-ball methods, the ν-method [64], or Nesterov accelerations [106].

Up to this point, the regularization methods that we have considered are linear with respect to the given measurement. Next, we aim to go beyond linear regularization methods by considering general convex regularizers in (2.3.3), which leads to minimization problems whose minimizer may not have a closed-form expression or may not be unique.

General variational regularization

The objective of this section is to recall some famous choices of convex regularizers R to be used in (2.3.3). In particular, we provide corresponding regularization schemes for solving some of the inverse problems presented at the beginning of this section. We start by considering R to be the ℓ^1 norm.

Soft-thresholding operator. Given the sparse signal denoising problem described in Example 2.22 with $\mathcal{U} = \mathbb{R}^d$, $1 \leq d < +\infty$, we consider the following variational approach

$$\min_u \frac{1}{2}\|u - x\|_2^2 + \lambda\|u\|_1, \tag{2.3.6}$$

where in this case $R = \|\cdot\|_1$ stands for the ℓ^1 norm, included in order to induce sparsity in the solution [44, 62]. In this case, the set of minimizers of the problem above coincides with the proximal operator of $\lambda\|\cdot\|_1$:

$$\arg\min_u \frac{1}{2}\|u - x\|_2^2 + \lambda\|u\|_1 =: \operatorname{prox}_{\lambda\|\cdot\|_1}(x) = \mathcal{S}_\lambda(x), \tag{2.3.7}$$

where the right hand side is the soft-thresholding operator (2.2.2).

Thus far, we have studied examples of regularization methods whose associated reconstructions have explicit expressions. However, this might not be the case in general. Next, we recall three examples where the solution of (2.3.3) does not have a closed-form, and present numerical tools for solving it.

The Lasso Problem. Given the sparse signal deblurring problem described in Example 2.23 with $\mathcal{U} = \mathbb{R}^d$, $1 \leq d < +\infty$, a classical way to tackle it is to consider solutions, for some $\lambda \in (0, +\infty)$, of

$$\min_u \frac{1}{2}\|Au - x\|_2^2 + \lambda\|u\|_1. \tag{2.3.8}$$

The problem above typically receives the name of the Lasso problem (or basis pursuit denoising in the signal processing literature) [7, 129], and does not have a closed-form solution in general. Nevertheless, proximal-gradient-type algorithms such as the *Iterative Shrinkage-Thresholding Algorithm* (ISTA) [56] – or FISTA [14], its accelerated version – can be applied to this problem in order to recover approximate solutions of (2.3.8).

Total Variation Regularization. Consider the image denoising problem introduced in Example 2.22 for $\mathcal{U} = \mathbb{R}^{N\times N}$, $1 \leq N < +\infty$. A popular approach, introduced in [115] (see also [39]), is to find solutions on the following variational problem:

$$\min_u \frac{1}{2}\|u - x\|_2^2 + \lambda \mathrm{TV}(u), \tag{2.3.9}$$

for some $\lambda \in (0, +\infty)$, and where TV denotes the (discrete) Total Variation regularizer. In order to give a proper definition, we first recall the definition of the discrete gradient given in [38]: for $u \in \mathcal{U}$, the discrete gradient Du is a vector in $\mathcal{U} \times \mathcal{U}$, defined pixel-wise for every $1 \leq i, j \leq N$, by $(Du)_{i,j} = ((Du)^1_{i,j}, (Du)^2_{i,j}) \in \mathbb{R}^2$, where

$$(Du)^1_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j}, & \text{if } i < N, \\ 0, & \text{if } i = N, \end{cases}$$

$$(Du)^2_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j}, & \text{if } j < N, \\ 0, & \text{if } j = N, \end{cases}$$

denote the discrete horizontal and vertical derivatives for $i, j = 1, ..., N$. Then, the isotropic total variation of u is defined as

$$\mathrm{TV}_{\text{iso}}(u) = \sum_{1 \leq i,j \leq N} \sqrt{|(Du)^1_{i,j}|^2 + |(Du)^2_{i,j}|^2}. \tag{2.3.10}$$

Moreover, the anisotropic total variation is defined as

$$\mathrm{TV}_{\text{aniso}}(u) = \sum_{1 \leq i,j \leq N} \left(|(Du)^1_{i,j}| + |(Du)^2_{i,j}| \right). \tag{2.3.11}$$

We refer the reader to [28, 29, 42] for higher order versions. Similarly to the case above, problem (2.3.9) does not have a closed form solution. In addition, proximal gradient algorithms cannot be directly applied in this case since a closed-form expression of the proximity operator of the Total Variation is not available. A possible approach to circumvent this issue can be found in [38], which proposes to apply a proximal gradient algorithm on the dual problem. In this case, the proximal operator coincides with the projection onto the ball of radius λ with respect to the $2-$norm.

In the experimental section of Chapter 4, we consider a generalized version of (2.3.9) for dealing with image denoising problems, and writes as

$$\min_u \frac{1}{2}\|u - x\|_2^2 + \lambda R(Du),$$

where R is a general convex regularizer and D denotes the discrete derivative defined above. With this formulation, we cover a wider range of regularization methods such as higher order versions of the TV functional or the so-called H^1 regularization (see [28]), while also including the examples mentioned above.

Nuclear norm regularization. Given the problem of matrix completion presented in Example 2.25, one of the main approaches for solving it has been to consider the so-called Rank Minimization Problem:

$$\min_U \operatorname{rank}(U), \qquad \text{s.t. } \mathrm{P}_\Theta(U) = \mathrm{P}_\Theta(X).$$

This approach is motivated by the fact that the solution is assumed to be low-rank. Nevertheless, it is well known that such problem is computationally hard to solve [136]. Consequently, further relaxations have been proposed. One of the most popular ones consists in solving, for some $\lambda \in (0, +\infty)$, the following variational problem [35, 66, 67]:

$$\min_U \|\mathrm{P}_\Theta(U - X)\|_2^2 + \lambda \|U\|_*, \tag{2.3.12}$$

where $\|\cdot\|_*$ stands for the nuclear norm of the matrix; i.e. the sum of its singular values. With this reformulation, solving (2.3.12) amounts to use the fact that the proximal operator of the ℓ^1 norm is the soft thresholding operator, and therefore ISTA-type algorithms can be applied.

CHAPTER 3

On Learning the Optimal Regularization Parameter in Inverse Problems

3.1 Introduction

The primary motivation of this chapter is to explore convergence guarantees for data-driven parameter choice rules. As mentioned earlier, regularization theory provides a structured approach for providing stable and accurate solutions of ill-posed inverse problems. In this context, an appropriate choice of the regularization parameter is crucial for providing both good reconstructions and convergent guarantees for the chosen regularization method. To do so, we adapt techniques from Statistical Learning Theory. In particular, the learned regularization parameter will be the minimizer of an Empirical Risk Minimization (ERM) problem [54] within a predefined grid of parameters. Drawing inspiration from the Bakushinskii veto [8], we demonstrate that theoretical guarantees can be provided within a stochastic inverse problems framework, where both the measurement and solution are random variables. We will explore the applicability of this approach across various examples of inverse problems.

From a theoretical standpoint, similar approaches have been previously explored for learning the regularization parameter. For example, in [1] a general approach is analyzed to learn a regularizer in Tikhonov-like regularization schemes for linear inverse problems. The results presented therein can be adapted to determine the optimal regularization parameter in certain scenarios. Another learning approach is analyzed in [57] and [94], where an unsupervised approach is studied. Additionally, a bilevel optimization perspective is taken in [70], where some theoretical results are also given, under the assumption that the set of candidate parameters is compact.

We now describe the organization of this Chapter. In Section 3.2, we establish a general framework for learning one parameter functions. We introduce the main assumptions and present the main result, showing the error behavior of the best empirical parameter with respect to the number of points in the training set. A precise formulation of this result will be given in Theorem 3.1. Next, we delve into specific examples of regularization methods to which the result above mentioned can be applied: Section 3.3 explores spectral regularization methods [64] for linear inverse problems, Section 3.4 focuses on Tikhonov regularization for nonlinear inverse problems [64] and, finally, in Section 3.5 we study general convex regularizers for solving linear inverse problems. In the latter, we provide convergence guarantees for the regularization parameter in two different settings: Legendre regularizers [12] and sparsity-inducing norms [7] in the finite-dimensional setting. The structure of Sections 3.3, 3.4 and 3.5 will be similar. We describe it below.

1. We set the inverse problems framework and state the main assumptions on both the model and the regularization method,

2. we verify that the required assumptions for Theorem 3.1 are satisfied,

3. we fix the discrepancy ℓ and, consequently, both the expected and empirical risks, and

4. we present the consequence of Theorem 3.1, showing the desired error bound for the learned regularization parameter.

We conclude this Chapter in Section 3.6, where we validate, experimentally, the theoretical findings obtained in the above sections. In particular, we will focus on Tikhonov and Landweber regularization. In this case, we also perform a comparison between the studied approach and the so-called quasi-optimality criterion, for which convergence guarantees have also been provided in a stochastic setting [11]. Finally, we explore ℓ^1 regularization applied to the problems of sparse signal denoising and deblurring, and Total Variation regularization for solving an image denoising problem.

3.2 Learning one parameter functions

In this section, we derive Statistical Learning results to learn functions parameterized by one parameter. In particular, in the context of learning in inverse problems, this will be the regularization parameter. For the time being, we consider an abstract learning framework.

Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ and $(\mathcal{U}, \langle\cdot,\cdot\rangle_{\mathcal{U}})$ be real and separable Hilbert spaces and let $(\Omega, \mathcal{A}, P)$ be a probability space. Let $(\bar{X}, \bar{U})$ be a pair of random variables with values in $\mathcal{X} \times \mathcal{U}$ and, for $\lambda \in (0, +\infty)$, we fix $f_\lambda : \mathcal{X} \to \mathcal{U}$ to be a family of measurable functions that are parametrized by λ. Moreover, for some $N \in \mathbb{N}$, define Λ, the finite grid of regularization parameters, as

$$\Lambda = \{\lambda_1, \dots, \lambda_N\} \tag{3.2.1}$$

with $0 < \lambda_1 \leq \lambda_2 \cdots \leq \lambda_N < \infty$. Given a measurable loss $\ell : \mathcal{U} \times \mathcal{U} \to [0, +\infty)$, we aim to find λ_Λ minimizing the expected risk (2.1.1) within Λ:

$$\lambda_\Lambda \in \arg\min_{\lambda \in \Lambda} L(f_\lambda) \tag{3.2.2}$$

Such parameter is the ideal regularization parameter when restricting ourselves to the set Λ. However, we next assume that we do not have access to the exact distribution of the pair $(\bar{X}, \bar{U})$, but to the set $\{(\bar{X}_i, \bar{U}_i)\}_{i=1}^n$, $n \in \mathbb{N}$, of independent copies of $(\bar{X}, \bar{U})$. The ERM (2.1.4) in this case writes as

$$\widehat{\lambda}_\Lambda \in \arg\min_{\lambda \in \Lambda} \widehat{L}(f_\lambda). \tag{3.2.3}$$

We aim at characterizing $L(f_{\widehat{\lambda}_\Lambda})$, namely the expected risk evaluated at the regularization parameter chosen accordingly to the rule in (3.2.3). As we did in Section 2.1, a first idea would be to prove an analogous version of Theorem 2.2 and compare it directly to $L(f_{\lambda_\Lambda})$. Instead, as discussed next, we assume that a suitable error bound $L(f_{\lambda_\Lambda}) \leq \Phi(\lambda_*)$ is available, being λ_* the minimizer of Φ, and then we compare $L(f_{\widehat{\lambda}_\Lambda})$ to $\Phi(\lambda_*)$. Next, we list and comment the main assumptions of the chapter.

Assumption 1. *The loss function ℓ is bounded by a constant $M > 0$.*

In the following, we will consider loss functions defined by classic discrepancy errors in inverse problems. In particular, we focus on Hilbertian norms, see Sections 3.3 and 3.4, and Bregman divergences associated with convex functionals, see Section 3.5. While none one of these examples are bounded in general, since we will assume $\bar{U}$ to be almost surely bounded, a bounded loss will be obtained by composing ℓ with suitable truncation operators. We next state the assumption related to an available bound on the expected risk $L(f_\lambda)$.

Assumption 2. *There exists a positive function $\Phi : (0, +\infty) \to (0, +\infty)$ such that, for every $\lambda \in (0, +\infty)$,*

$$L(f_\lambda) \leq \Phi(\lambda). \tag{3.2.4}$$

Moreover, there exists $\lambda_* > 0$ such that

$$\lambda_* \in \operatorname*{arg\,min}_{\lambda \in (0,+\infty)} \Phi(\lambda). \tag{3.2.5}$$

Finally, there exists a non-decreasing function $C : [1, +\infty) \to [0, +\infty)$ such that, for all $q \geq 1$,

$$\Phi(q\lambda_*) \leq C(q)\Phi(\lambda_*). \tag{3.2.6}$$

The main reason for the above assumption is to avoid smoothness conditions on the dependence of f_λ on λ which are required in classic studies of ERM, see e.g. [1, 54]. This assumption might seem unusual for a learning setting but, as shown in Sections 3.3, 3.4 and 3.5, it is naturally satisfied for a large class of inverse problems. Moreover, this is the usual strategy to design a priori choices of the regularization parameter. In this latter setting, it is often possible to derive tight bounds in the sense that the two quantities, $L(f_\lambda)$ and $\Phi(\lambda)$, have the same behavior with respect to λ and the noise level. Consequently, $L(f_{\lambda_\Lambda})$ is comparable to $\Phi(\lambda_*)$ (see e.g. [64, Chapter 4]). We make one last assumption on how *large* is the set of candidate values Λ.

Assumption 3. *Let Λ be defined as in (3.2.1). Assume that*

$$\lambda_* \in [\lambda_1, \lambda_N] \tag{3.2.7}$$

and, for every $j = 1, \dots, N$, $\lambda_j = \lambda_1 Q^{j-1}$, where

$$Q := \left(\frac{\lambda_N}{\lambda_1}\right)^{\frac{1}{N-1}}. \tag{3.2.8}$$

The above assumption states that we can choose a sufficiently large interval for our discretization so that λ_* in (3.2.5) always falls within the interval. This is an approximation assumption which is satisfied in practice by taking λ_1 sufficiently small (and λ_N sufficiently big).

Given the above assumptions, we next show that the choice $\widehat{\lambda}_\Lambda$ achieves an error close to that of λ_*.

Theorem 3.1. *Let Assumptions 1, 2 and 3 be satisfied and let $\eta \in (0, 1)$. Then, with probability at least $1 - \eta$,*

$$L(f_{\widehat{\lambda}_\Lambda}) \leq 2C(Q)\Phi(\lambda_*) + \frac{13M}{2n} \log \frac{2N}{\eta}.$$

The above result shows that $\widehat{\lambda}_\Lambda$ achieves an error of the same order of λ_*, up to a multiplicative factor depending on $C(Q)$, and a corrective term which decreases as $1/n$. Moreover, from the expression (3.2.8), once the minimal and maximal elements of the discretization are fixed, we can see that $Q \approx 1$ if N is large enough. At the same time, taking N large has a minor effect on the bound, since the corrective term depends logarithmically on N. We first provide the proof of Theorem 3.1.

Proof of Theorem 1

We begin providing a sketch of the main steps in the proof. The idea is to first compare the behavior of $\widehat{\lambda}_\Lambda$ to that of λ_Λ. Indeed, we prove in Lemma 3.2 that with high probability

$$L(f_{\widehat{\lambda}_\Lambda}) \leq 2L(f_{\lambda_\Lambda}) + c\frac{\log(2N)}{n},$$

for some constant $c > 0$. After, we show in Lemma 3.3 that there exists $1 \leq q \leq Q$ such that $q\lambda_* \in \Lambda$. Therefore, by definition of λ_Λ, we get that

$$L(f_{\lambda_\Lambda}) \leq L(f_{q\lambda_*}).$$

Combining the above results and using condition (3.2.6), we get with high probability that

$$L(f_{\widehat{\lambda}_\Lambda}) \lesssim 2L(f_{q\lambda_*}) + \frac{\log(2N)}{n} \lesssim 2C(Q)\Phi(\lambda_*) + \frac{\log(2N)}{n},$$

which is the desired result. We next provide the detailed proof. First, we introduce the following probabilistic lemma.

Lemma 3.2. *Let Assumption 1 be satisfied and let $\eta \in (0,1)$. Then, with probability at least $1-\eta$, we have that*

$$L(f_{\widehat{\lambda}_\Lambda}) \leq 2L(f_{\lambda_\Lambda}) + \frac{13M}{2n}\log\frac{2N}{\eta}.$$

The proof adapts ideas from [36], and is based on a classic union bound argument and the concentration inequality stated in (2.0.3).

Proof. (of Lemma 3.2). For $\lambda \in \Lambda$, define for every $i = 1, ..., n$, the random variable $\bar{Z}_i(\lambda) = \ell(f_\lambda(\bar{X}_i), \bar{U}_i)$. Then,

$$\frac{1}{n}\sum_{i=1}^{n} \bar{Z}_i(\lambda) = \widehat{L}(f_\lambda),$$

and

$$\mathbb{E}[\bar{Z}_i(\lambda)] = L(f_\lambda).$$

Moreover, since ℓ is bounded by Assumption 1, then $\bar{Z}_i(\lambda) \leq M$ and this implies

$$\mathbb{E}[|\bar{Z}_i(\lambda)|^2] = \mathbb{E}[\ell(f_\lambda(\bar{X}_i), \bar{U}_i)\ell(f_\lambda(\bar{X}_i), \bar{U}_i)] \leq ML(f_\lambda).$$

Now, we apply (2.0.3) with $B = M$ and, by recalling that

$$\mathbb{E}[|\bar{Z}_i(\lambda) - \mathbb{E}[\bar{Z}_i(\lambda)]|^2] \leq \mathbb{E}[|\bar{Z}_i(\lambda)|^2],$$

we fix $\sigma^2 := ML(f_\lambda)$. Now, by defining for each $\lambda \in \Lambda$ and for all $\alpha, \varepsilon > 0$, the event

$$E_\lambda := \left\{ |\widehat{L}(f_\lambda) - L(f_\lambda)| \geq \varepsilon + \alpha ML(f_\lambda) \right\},$$

we get that, for any $\lambda \in \Lambda$,

$$P\left(E_\lambda\right) \leq 2e^{-\frac{6n\alpha\varepsilon}{3+4\alpha M}}.$$

By definition of probability measure, we derive that, for all $\alpha, \varepsilon > 0$,

$$P\left(\bigcup_{\lambda\in\Lambda} E_\lambda\right) \leq \sum_{\lambda\in\Lambda} P(E_\lambda) = 2|\Lambda|e^{-\frac{6n\alpha\varepsilon}{3+4\alpha M}},$$

where in this case $|\Lambda|$ denotes the cardinality of the set Λ. Now, let $\eta \in (0,1)$. Since the above is valid for any $\alpha > 0$, fix $\alpha = 1/(3M)$. With this choice, let $\varepsilon = \frac{13M}{6n} \log \frac{2|\Lambda|}{\eta}$. Then, with probability at least $1-\eta$, for all $\lambda \in \Lambda$ we have that

$$\widehat{L}(f_\lambda) \leq \frac{4}{3} L(f_\lambda) + \varepsilon$$

and

$$L(f_\lambda) \leq \frac{3}{2} \left(\widehat{L}(f_\lambda) + \varepsilon \right).$$

Using the above inequalities and the definition of $\widehat{\lambda}_\Lambda$ we have that,

$$\begin{aligned} L(f_{\widehat{\lambda}_\Lambda}) &\leq \frac{3}{2} \left(\widehat{L}(f_{\widehat{\lambda}_\Lambda}) + \varepsilon \right) \\ &\leq \frac{3}{2} \left(\widehat{L}(f_{\lambda_\Lambda}) + \varepsilon \right) \\ &\leq 2L(f_{\lambda_\Lambda}) + 3\varepsilon. \end{aligned}$$

By recalling that $|\Lambda| = N$, we conclude that

$$L(f_{\widehat{\lambda}_\Lambda}) \leq 2L(f_{\lambda_\Lambda}) + \frac{13M}{2n} \log \frac{2N}{\eta},$$

proving the result. □

Note that the above result holds by only assuming that the loss ℓ is bounded. Indeed, the structural assumptions that we have introduced are used to prove the following lemma.

Lemma 3.3. *Let Assumptions 2 and 3 be satisfied and consider λ_* as in Assumption 2. Then, there exists $1 \leq q \leq Q$ such that $q\lambda_* \in \Lambda$ and so*

$$L(f_{\lambda_\Lambda}) \leq L(f_{q\lambda_*}).$$

Proof. From Assumption 3, since $\lambda_* \in [\lambda_1, \lambda_N]$, there exists $j_0 \in \{2, \dots, N\}$ such that

$$\lambda_{j_0 - 1} \leq \lambda_* \leq \lambda_{j_0}.$$

If we let $q = \lambda_{j_0}/\lambda_*$, then $q\lambda_* = \lambda_{j_0} \in \Lambda$. It is only left to prove that $1 \leq q \leq Q$. Given the definition of Q and the construction of Λ, if we divide the above inequalities by λ_{j_0}, then

$$\frac{1}{Q} \leq \frac{1}{q} \leq 1,$$

so that

$$1 \leq q \leq Q.$$

Finally, by the definition of λ_Λ, we get

$$L(f_{\lambda_\Lambda}) \leq L(f_{q\lambda_*}),$$

concluding the proof. □

We add one final remark.

Remark 3.4 (Comparison with Theorem 2.2)**.** A slightly different estimate can be obtained following the same techniques for proving Theorem 2.2. In particular, considering Hoeffding's inequality (2.0.2). If we let $\eta \in (0,1)$, the following bound holds with probability at least $1-\eta$:

$$L(f_{\widehat{\lambda}_\Lambda}) \leq L(f_{\lambda_\Lambda}) + 2\sqrt{\frac{2M}{n}\log\frac{2N}{\eta}}. \tag{3.2.9}$$

Compared to the estimate obtained in Lemma 3.2, the above inequality avoids the factor 2 in front of $L(f_{\lambda_\Lambda})$. However, the dependence on the data cardinality n is considerably worse. In addition, note that in this case we use the fact that set of candidate parameters Λ is finite. If the set Λ were a compact set, one could consider a similar covering argument than the one applied in Theorem 2.2. For completeness, we report the proof of inequality (3.2.9). As we did in Section 2.1, we can see that

$$L(f_{\widehat{\lambda}_\Lambda}) - L(f_{\lambda_\Lambda}) \leq 2\sup_{\lambda\in\Lambda}|L(f_\lambda) - \widehat{L}(f_\lambda)|.$$

Next, define for every $i = 1, ..., n$, the random variable $\bar{Z}_i(\lambda) = \ell(f_\lambda(\bar{X}_i), \bar{U}_i)$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\bar{Z}_i(\lambda) = \widehat{L}(f_\lambda),$$

and

$$\mathbb{E}[\bar{Z}_i(\lambda)] = L(f_\lambda),$$

and recall that $L(f_\lambda) \leq M$ by assumption. Then, $|\bar{Z}_i - \mathbb{E}[\bar{Z}_i(\lambda)]| \leq 2M =: B$. By Hoeffding's inequality (2.0.2), we get that, for every $\lambda > 0$ and for all $\varepsilon > 0$,

$$P\left(|L(f_\lambda) - \widehat{L}(f_\lambda)| \geq \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{8M^2}}.$$

By deifnition, the probability of a union of events is less or equal than the sum of their probabilities. Then,

$$\begin{aligned} P\left(\sup_{\lambda\in\Lambda}|L(f_\lambda) - \widehat{L}(f_\lambda)| \geq \varepsilon\right) &= P\left(\bigcup_{\lambda\in\Lambda}|L(f_\lambda) - \widehat{L}(f_\lambda)| \geq \varepsilon\right) \\ &\leq \sum_{\lambda\in\Lambda} P\left(|L(f_\lambda) - \widehat{L}(f_\lambda)| \geq \varepsilon\right) \\ &\leq 2Ne^{-\frac{n\varepsilon^2}{8M^2}}. \end{aligned}$$

Inequality (3.2.9) follows by setting $\eta := 2Ne^{-\frac{n\varepsilon^2}{8M^2}}$ and deriving the expression for ε.

In the following, we provide concrete examples in the context of inverse problems that illustrate and instantiate the above results.

3.3 Spectral regularization for linear inverse problems

In this section, we illustrate Theorem 3.1 considering spectral regularization methods, already introduced in Section 2.3.2, for a class of stochastic linear inverse problems, extending the classical deterministic framework. The key point is to derive a suitable error bound and a corresponding a priori parameter choice, which will correspond to λ_*, so that

Assumption 2 holds. We next fix the framework. Let $(\mathcal{U}, \langle\cdot,\cdot\rangle_{\mathcal{U}})$ and $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ be real and separable Hilbert spaces, let $A\colon \mathcal{U} \to \mathcal{X}$ be a linear and bounded operator and assume, for simplicity, that $\|A\|_{\mathrm{op}} \leq 1$. Then, let $\bar{U}$ and ε be a pair of random variables with values in $\mathcal{U}$ and $\mathcal{X}$ respectively, and consider the model

$$\bar{X} = A\bar{U} + \varepsilon, \quad \text{a.s..} \tag{3.3.1}$$

We make several assumptions. The first is on the noise ε.

Assumption 4. *We assume that*

$$\mathbb{E}[\varepsilon|\bar{U}] = 0$$

and, moreover, that there exists $\tau > 0$ such that

$$\mathbb{E}[\|\varepsilon\|_{\mathcal{X}}^2 |\bar{U}] \leq \tau^2.$$

The above condition is a simple and natural stochastic extension of the classical bounded noise variance assumption. We also assume that $\bar{U}$ satisfies the following stochastic extension of the classical Hölder source condition [64].

Assumption 5. *The random variable $\bar{U}$ is such that $\|\bar{U}\|_{\mathcal{U}} \leq 1$ a.s.. Moreover, there exist a random variable $\bar{Z}$ with values in $\mathcal{U}$ and scalars β, $s > 0$ such that,*

$$\bar{U} = (A^*A)^s \bar{Z},$$

and

$$\mathbb{E}[\|\bar{Z}\|_{\mathcal{U}}^2] \leq \beta^2.$$

In this setting, we recall that the class of spectral regularization methods that we aim to analyze is given by

$$\bar{U}_\lambda := g_\lambda(A^*A)A^*\bar{X}. \tag{3.3.2}$$

defined by the spectral functions $g_\lambda : (0,1] \to \mathbb{R}$ as mentioned in Remark 2.27. Clearly, $\bar{U}_\lambda = \bar{U}_\lambda(\bar{X})$, but we omit the dependence for conciseness. Note that the above expression ensures that $\bar{U}_\lambda$ is measurable, since it is the image of a linear operator applied to $\bar{X}$. Already in Section 2.3.2 we gave the necessary conditions for the class of spectral functions g_λ to be a regularization method. In this Chapter, instead, we will just assume the following conditions.

Assumption 6. *There exists a general constant $C_1 > 0$ such that, for all $\lambda \in (0, +\infty)$,*

$$\sup_{\sigma\in(0,1]} |g_\lambda(\sigma)\sqrt{\sigma}| \leq \frac{C_1}{\sqrt{\lambda}}.$$

Moreover, there is a constant $C_2 > 0$ and $\alpha > 0$ such that, for $s > 0$ as in Assumption 5,

$$\sup_{\sigma\in(0,1]} |(1 - g_\lambda(\sigma)\sigma)\sigma^s| \leq C_2\lambda^\alpha. \tag{3.3.3}$$

It can be easily shown that assumption 6 is also satisfied by all of the examples listed in Section 2.3.2. We add some remarks regarding this assumption. The first assumption implies that the norm of the regularization method $g_\lambda(A^*A)A^*$ is always bounded and controlled by $1/\sqrt{\lambda}$. Moreover, expression (3.3.3) is an approximation condition, which characterizes the extent to which the considered spectral regularization method can take advantage of the regularity of the problem, expressed by the source condition. Finally, it is satisfied for $\alpha = \min(\bar{\nu}, s)$, where $\bar{\nu}$ is the qualification parameter given in Definition 2.28. Both of the above assumptions allow us to derive suitable error bounds and a corresponding a priori parameter choice, extending classical results in the deterministic setting.

Theorem 3.5. *Under Assumptions 4, 5 and 6, for every $\lambda \in (0, +\infty)$ it holds that*

$$\mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{X}}^2] \leq \frac{C_1^2 \tau^2}{\lambda} + C_2^2 \beta^2 \lambda^{2\alpha}. \tag{3.3.4}$$

In particular, taking

$$\lambda_* = \left(\frac{C_1^2}{2\alpha C_2^2}\right)^{1/(2\alpha+1)} \left(\frac{\tau}{\beta}\right)^{2/(2\alpha+1)},$$

the following bound holds

$$\mathbb{E}[\|\bar{U}_{\lambda_*} - \bar{U}\|_{\mathcal{U}}^2] \leq (2\alpha+1) \left[\left(\frac{C_1^2}{2\alpha}\right)^{2\alpha} C_2^2\right]^{1/(2\alpha+1)} \left(\frac{\tau^{2\alpha}}{\beta}\right)^{2/(2\alpha+1)}. \tag{3.3.5}$$

Proof. To relate $\bar{U}_\lambda$ and $\bar{U}$, we observe that

$$\mathbb{E}[\bar{U}_\lambda|\bar{U}] = \mathbb{E}[g_\lambda(A^*A)A^*\bar{X}|\bar{U}] = \mathbb{E}[g_\lambda(A^*A)A^*A\bar{U}|\bar{U}] = g_\lambda(A^*A)A^*A\bar{U},$$

where we used the definition of $\bar{X}$ and Assumption 4. With this, we can decompose the deviation of $\bar{U}_\lambda$ to $\bar{U}$ as

$$\begin{aligned} \bar{U}_\lambda - \bar{U} &= \bar{U}_\lambda - \mathbb{E}[\bar{U}_\lambda|\bar{U}] + \mathbb{E}[\bar{U}_\lambda|\bar{U}] - \bar{U} \\ &= g_\lambda(A^*A)A^*(\bar{X} - A\bar{U}) + (g_\lambda(A^*A)A^*A - \mathrm{Id})\bar{U} \\ &= g_\lambda(A^*A)A^*\varepsilon + (g_\lambda(A^*A)A^*A - \mathrm{Id})(A^*A)^s\bar{Z}. \end{aligned} \tag{3.3.6}$$

Next, recall that, under Assumption 6, the following operator estimates hold

$$\|g_\lambda(A^*A)A^*\|_{\mathrm{op}} \leq \frac{C_1}{\sqrt{\lambda}}, \quad \|(I - g_\lambda(A^*A)A^*A)(A^*A)^s\|_{\mathrm{op}} \leq C_2\lambda^\alpha, \tag{3.3.7}$$

see e.g. [64]. If we take the expectation of the squared norm in (3.3.6) and develop the square, we get

$$\mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2] = \mathbb{E}[\|g_\lambda(A^*A)A^*\varepsilon\|_{\mathcal{X}}^2] + \mathbb{E}[\|(g_\lambda(A^*A)A^*A - \mathrm{Id})\bar{U}\|_{\mathcal{U}}^2],$$

since, by Assumption 4, we have

$$\begin{aligned} &\mathbb{E}[\langle g_\lambda(A^*A)A^*\varepsilon, (g_\lambda(A^*A)A^*A - \mathrm{Id})\bar{U}\rangle_{\mathcal{U}}] \\ &\quad = \mathbb{E}[\langle g_\lambda(A^*A)A^*\mathbb{E}[\varepsilon|\bar{U}], (g_\lambda(A^*A)A^*A - \mathrm{Id})\bar{U}\rangle_{\mathcal{U}}] = 0. \end{aligned}$$

Then, using again Assumptions 4, 5, and 6 as well as the estimates (3.3.7), we derive

$$\begin{aligned} \mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2] &\leq \|g_\lambda(A^*A)A^*\|_{\mathrm{op}}^2 \mathbb{E}[\|\varepsilon\|_{\mathcal{X}}^2] \\ &\quad + \|(\mathrm{Id} - g_\lambda(A^*A)A^*A)(A^*A)^s\|_{\mathrm{op}}^2 \mathbb{E}[\|\bar{Z}\|_{\mathcal{U}}^2] \\ &\leq \frac{C_1^2\tau^2}{\lambda} + C_2^2\beta^2\lambda^{2\alpha}, \end{aligned}$$

since, by the tower property [112],

$$\mathbb{E}[\|\varepsilon\|_{\mathcal{X}}^2] = \mathbb{E}[\mathbb{E}[\|\varepsilon\|^2|\bar{U}]] \leq \tau^2.$$

Finally, the value of λ minimizing the above bound is

$$\lambda_* = \left(\frac{C_1^2\tau^2}{2\alpha C_2^2\beta^2}\right)^{1/(2\alpha+1)},$$

and the corresponding error bound is

$$\mathbb{E}[\|\bar{U}_{\lambda_*} - \bar{U}\|_{\mathcal{U}}^2] \leq (2\alpha+1) \left[\left(\frac{C_1^2}{2\alpha}\right)^{2\alpha} C_2^2\right]^{1/(2\alpha+1)} \left(\frac{\tau^{2\alpha}}{\beta}\right)^{2/(2\alpha+1)},$$

which is the inequality that we were aiming for. □

The expression given in (3.3.4) provides a bound, for any value of the regularization parameter, of the distance between the regularized and the exact solutions. This bound is composed of two terms. The first one is related to τ, the noise level, and decreases with the regularization parameter as $1/\lambda$. The second one is related to β in the source condition, and increases with the regularization parameter as $\lambda^{2\alpha}$. The choice of the parameter λ_* is then obtained by minimizing this upper bound in λ. Once we plug λ_* in (3.3.4), we obtain the bound in (3.3.5). These results are analogous to the ones usually obtained in the deterministic setting (see for instance Corollary 4.4 in [64]), and are known to be optimal in the sense of Definition 3.17 in [64].

Next, we show that, with the aid of the previous result, and in combination with Theorem 3.1, the regularization parameter on the grid learned from data, namely $\widehat{\lambda}_\Lambda$ defined in (3.2.3), achieves a similar performance to the one of λ_*. Toward this end, we consider the truncation operator $T : \mathcal{U} \to \mathcal{U}$, defined for all $u \in \mathcal{U}$ as

$$Tu = \begin{cases} u, & \text{if } \|u\|_{\mathcal{U}} \leq 1, \\ \dfrac{u}{\|u\|_{\mathcal{U}}}, & \text{otherwise.} \end{cases} \tag{3.3.8}$$

The above operator can also be seen as the projection map that sends any element $u \in \mathcal{U}$ into the unit ball of $\mathcal{U}$ centered at 0. To apply the result in Section 3.2, we consider the loss defined, for every $(u, u') \in \mathcal{U} \times \mathcal{U}$, as

$$\ell(u, u') = \left\|Tu - Tu'\right\|_{\mathcal{U}}^2. \tag{3.3.9}$$

Next, we aim to define the expected risk (3.2.2) in this setting. To do so, for every $\lambda \in (0, +\infty)$ let $f_\lambda(\bar{X}) := \bar{U}_\lambda$, where $\bar{U}_\lambda$ is defined as in (3.3.2). Then, the corresponding expected risk in this case writes,

$$L(\bar{U}_\lambda) := \mathbb{E}[\|T\bar{U}_\lambda - T\bar{U}\|_{\mathcal{U}}^2]. \tag{3.3.10}$$

We study now the error obtained in this context by choosing the regularization parameter λ with ERM. Consider a finite set of independent and identical copies $(\bar{X}_i, \bar{U}_i)$, $i = 1, ..., n$, $n \in \mathbb{N}$, of the pair $(\bar{X}, \bar{U})$ distributed as in (3.3.1) and let $\bar{U}_\lambda^i := \bar{U}_\lambda(\bar{X}_i)$. Then, the corresponding ERM is given by

$$\widehat{\lambda}_\Lambda \in \operatorname*{arg\,min}_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} \left\|T\bar{U}_\lambda^i - \bar{U}_i\right\|_{\mathcal{U}}^2, \tag{3.3.11}$$

where we used that $\bar{U}_i = T\bar{U}_i$ a.s. for every $i = 1, ..., n$ since $\left\|\bar{U}\right\|_{\mathcal{U}} \leq 1$ a.s. by Assumption 5. The following corollary provides the desired error estimates.

Corollary 3.6. *Let Assumption 3 be satisfied with λ_* as in Theorem 3.5. Suppose that Assumptions 4, 5 and 6 hold, and choose the loss as in (3.3.9). Let $\eta \in (0, 1)$. Then, with probability at least $1 - \eta$,*

$$L(\bar{U}_{\widehat{\lambda}_\Lambda}) \leq \frac{2(2\alpha + Q^{2\alpha+1})}{Q} \left[\left(\frac{C_1^2}{2\alpha}\right)^{2\alpha} C_2^2\right]^{1/(2\alpha+1)} \left(\frac{\tau^{2\alpha}}{\beta}\right)^{2/(2\alpha+1)} + \frac{26}{n} \log \frac{2N}{\eta}.$$

In this setting, Assumption 1 is trivially satisfied with $M = 4$. The proof will therefore consist in verifying that also Assumption 2 holds, so that Theorem 3.1 can be applied.

Proof. In this case, we just need to show that Assumption 2 is satisfied for $f_\lambda(\bar{X}) = \bar{U}_\lambda$ and L defined as in (3.3.10). Since T is a projection, it is 1-Lipschitz. Then, for all $\lambda \in (0, +\infty)$, it holds that

$$L(\bar{U}_\lambda) = \mathbb{E}[\left\|T\bar{U}_\lambda - T\bar{U}\right\|_{\mathcal{U}}^2] \leq \mathbb{E}[\left\|\bar{U}_\lambda - \bar{U}\right\|_{\mathcal{U}}^2].$$

Moreover, if we define $\Phi(\lambda)$ as the right hand side of equation (3.3.4), then (3.2.4) holds. In addition, λ_* defined as in Theorem 3.5 is the minimizer of Φ. Now, define the function

$$C : [1, +\infty) \to [0, +\infty); \quad C(q) := \frac{2\alpha + q^{2\alpha+1}}{q(2\alpha+1)},$$

and observe that it is non-decreasing. Then, from the error bound (3.3.5), we derive, for any $q \in [1, +\infty)$, that

$$\Phi(q\lambda_*) = C(q)\Phi(\lambda_*) = \frac{2\alpha + q^{2\alpha+1}}{q} \left[\left(\frac{C_1^2}{2\alpha} \right)^{2\alpha} C_2^2 \right]^{1/(2\alpha+1)} \left(\frac{\tau^{2\alpha}}{\beta} \right)^{2/(2\alpha+1)}.$$

Hence, Assumption 2 is satisfied. The result follows by applying Theorem 3.1. □

Corollary 3.6 shows that, under a natural generalization of the classical assumptions in deterministic inverse problems to the stochastic setting, the error obtained with the optimal parameter on the grid for the empirical risk, namely $\widehat{\lambda}_\Lambda$, is close to that of λ_*, up to a logarithmic factor that increases very slowly with N, and decreases with n. We add one final remark for this section.

Remark 3.7 (Comparison with Theorem 4.1 in [1])**.** The paper [1] aims to learn the optimal Tikhonov regularizer, of the form $\|B(\cdot - h)\|^2$, for a linear operator B and a bias vector $h \in \mathcal{U}$. The main result of [1] is Theorem 4.1, which establishes an excess risk bound for parameters $(\hat{B}, \hat{h})$ learned by minimizing the empirical risk. The setting is quite different since, in [1], the authors learn a general Tikhonov regularizer by demonstrating that the optimal pair (B^*, h^*) consists of the covariance operator and the mean of $\bar{U}$, respectively. In this paper, we only learn the regularization parameter, but our setting allows for a large class of spectral filters. The assumptions of Theorem 4.1, as seen in (20) and (21) of [1], are quite different from Assumption 5 and Assumption 6, making a direct comparisong between our Corollary 1 and Theorem 4.1 not meaningful. We only observe that the proof of Theorem 4.1 in [1] relies on learning techniques that exploit the Lipschitz continuity of the empirical risk with respect to the pair (h, B) and a classic covering argument. In this paper, we use instead a different approach introduced in [36] for the cross-validation method.

Next, we consider the problem of selecting the regularization parameter for Tikhonov regularization in the setting of nonlinear inverse problems [64, Chapter 10].

3.4 Tikhonov regularization for non linear inverse problems

Let $A : \mathrm{dom}(A) \subseteq \mathcal{U} \to \mathcal{X}$ be a (nonlinear) operator whose domain has nonempty interior and let $\bar{U}$ and ε be a pair of random variables with values in $\mathcal{U}$ and $\mathcal{X}$ respectively. Then, let

$$\bar{X} = A(\bar{U}) + \varepsilon, \quad \text{a.s.} \tag{3.4.1}$$

with $\bar{U} \in \mathrm{int}(\mathrm{dom}(A))$ a.s.. We make several assumptions. The first one is on the noise ε.

Assumption 7. *There exists a constant $\tau > 0$ such that*

$$\mathbb{E}[\|\varepsilon\|_{\mathcal{X}}^2 \,|\bar{U}] \leq \tau^2 \quad \text{a.s.}$$

Using Jensen's inequality for the conditional expectation [141, 9.7 (h)], we derive from the previous assumption that

$$\mathbb{E}[\|\varepsilon\|_{\mathcal{X}} |\bar{U}] \leq \tau \text{ a.s.} \tag{3.4.2}$$

Next we impose fairly standard conditions on the operator A.

Assumption 8. *The operator $A : \mathrm{dom}(A) \to \mathcal{X}$ is a continuous and weakly closed operator with $\mathrm{int}(\mathrm{dom}(A))$ non-empty, and with $\mathrm{dom}(A)$ convex. Moreover, A is Fréchet differentiable in $\mathrm{int}(\mathrm{dom}(A))$ with derivative denoted by A', and there exists a constant $C_0 > 0$ such that, for all u, $u' \in \mathrm{int}(\mathrm{dom}(A))$,*

$$\left\|A'(u) - A'(u')\right\|_{\mathrm{op}} \leq C_0 \left\|u - u'\right\|_{\mathcal{U}}. \tag{3.4.3}$$

The previous assumption implies that, for all $u \in \mathrm{int}(\mathrm{dom}(A))$ and $u' \in \mathrm{dom}(A)$,

$$\left\|A(u') - A(u) - A'(u)(u' - u)\right\|_{\mathcal{X}} \leq \frac{C_0}{2} \left\|u' - u\right\|_{\mathcal{U}}^2,$$

so that, by the triangle inequality,

$$\left\|A'(u)(u' - u)\right\|_{\mathcal{X}} \leq \left\|A(u') - A(u)\right\|_{\mathcal{X}} + \frac{C_0}{2} \left\|u' - u\right\|_{\mathcal{U}}^2. \tag{3.4.4}$$

Here, we assume global Lipschitz continuity of the derivative to avoid technicalities, but the argument could be extended under a local smoothness assumption as in [47].

For nonlinear inverse problems, the Tikhonov estimator is defined with respect to a suitable initialization. Here, we assume the initialization to be described by a random variable $\bar{U}_0$ with values in $\mathcal{U}$, and that the set of minimizers

$$\operatorname*{arg\,min}_{u \in \mathrm{dom}(A)} \left\|A(u) - \bar{X}(\omega)\right\|_{\mathcal{X}}^2 + \lambda \left\|u - \bar{U}_0(\omega)\right\|_{\mathcal{U}}^2$$

is nonempty for every $\omega \in \Omega$ thanks to Assumption 8, see [47, Theorem 10.1]. A corresponding Tikhonov regularized estimator is a random variable $\bar{U}_\lambda$ defined by setting, for almost every $\omega \in \Omega$,

$$\bar{U}_\lambda(\omega) \in \operatorname*{arg\,min}_{u \in \mathrm{dom}(A)} \left\|A(u) - \bar{X}(\omega)\right\|_{\mathcal{X}}^2 + \lambda \left\|u - \bar{U}_0(\omega)\right\|_{\mathcal{U}}^2. \tag{3.4.5}$$

Note that $\bar{U}_\lambda = \bar{U}_\lambda(\bar{X}, \bar{U}_0)$, but we omit this dependence for simplicity. The existence of a random variable $\bar{U}_\lambda$ taking values in the set of minimizers is ensured under some additional assumptions, see e.g. Filippov's Implicit function Theorem [86, Theorem 7.1]. For that reason, we directly assume that such measurable selection exists. The following assumption will be needed to derive the error bounds and extends analogous conditions in the deterministic case.

Assumption 9. *The random variable $\bar{U}$ is such that $\|\bar{U} - \bar{U}_0\|_{\mathcal{U}} \leq 1$ a.s. and, under Assumption 8, there exists a random variable $\bar{Z}$ with values in $\mathcal{X}$, and a scalar $\beta > 0$ such that a.s.*

$$\bar{U} - \bar{U}_0 = A'(\bar{U})^* \bar{Z},$$

and

$$\left\|\bar{Z}\right\|_{\mathcal{X}} \leq \beta \text{ a.s.}, \quad \textit{with} \quad \beta C_0 < 1,$$

where C_0 is the constant introduced in Assumption 8.

The latter assumption can be seen as a nonlinear version of the source condition considered in Assumption 5 for $s = 1$.

In the next result, which is analogous to Theorem 3.5, we derive a bound on the error of the Tikhonov regularized solution, leading to a priori parameter choices, and its proof is a modification of the one in the deterministic setting, see e.g. [47, 64].

Theorem 3.8. *Let Assumptions 7, 8 and 9 be satisfied. Then, for every $\lambda \in (0, +\infty)$, it holds that*

$$\mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2] \leq \frac{(\tau + \beta\lambda)^2}{\lambda(1 - \beta C_0)}. \tag{3.4.6}$$

In particular, setting $\lambda_* = \frac{\tau}{\beta}$,

$$\mathbb{E}[\|\bar{U}_{\lambda_*} - \bar{U}\|_{\mathcal{U}}^2] \leq \frac{4}{1 - \beta C_0}\tau\beta.$$

Proof. The expressions below are all intended to hold a.s.. By definition of $\bar{U}_\lambda$, $\bar{U}$ and ε, it follows that

$$\begin{aligned}\|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}}^2 + \lambda\|\bar{U}_\lambda - \bar{U}_0\|_{\mathcal{U}}^2 &\leq \|A(\bar{U}) - \bar{X}\|_{\mathcal{X}}^2 + \lambda\|\bar{U} - \bar{U}_0\|_{\mathcal{U}}^2 \\ &= \|\varepsilon\|_{\mathcal{X}}^2 + \lambda\|\bar{U} - \bar{U}_0\|_{\mathcal{U}}^2. \end{aligned} \tag{3.4.7}$$

Since

$$\|\bar{U}_\lambda - \bar{U}_0\|_{\mathcal{U}}^2 = \|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2 + \|\bar{U} - \bar{U}_0\|_{\mathcal{U}}^2 + 2\langle \bar{U}_\lambda - \bar{U}, \bar{U} - \bar{U}_0\rangle_{\mathcal{U}}, \tag{3.4.8}$$

inequality (3.4.7) implies

$$\|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}}^2 + \lambda\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2 \leq \|\varepsilon\|_{\mathcal{X}}^2 - 2\lambda\langle \bar{U}_\lambda - \bar{U}, \bar{U} - \bar{U}_0\rangle_{\mathcal{U}}.$$

Then, Assumption 9 and Cauchy-Schwartz inequality yield

$$\|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}}^2 + \lambda\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2 \leq \|\varepsilon\|_{\mathcal{X}}^2 + 2\lambda\|A'(\bar{U})(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}\|\bar{Z}\|_{\mathcal{X}}. \tag{3.4.9}$$

Since $\bar{X} \in \mathrm{int}(\mathrm{dom}(A))$ and $\bar{U}_\lambda \in \mathrm{dom}(A)$, and $\mathrm{dom}(A)$ is convex by assumption, inequality (3.4.4) with $u = \bar{U}$ and $u' = \bar{U}_\lambda$ yields

$$\|A'(\bar{U})(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}} \leq \|A(\bar{U}_\lambda) - A(\bar{X})\|_{\mathcal{X}} + \frac{C_0}{2}\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2,$$

so that, by adding and subtracting $\bar{X}$ in $\|A(\bar{U}_\lambda) - A(\bar{X})\|_{\mathcal{X}}$, we obtain

$$\|A'(\bar{U})(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}} \leq \|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}} + \|\varepsilon\|_{\mathcal{X}} + \frac{C_0}{2}\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2.$$

Plugging the above inequality into (3.4.9), we get

$$\begin{aligned}\|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}}^2 + \lambda\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2 \leq \|\varepsilon\|_{\mathcal{X}}^2 + 2\lambda\|\bar{Z}\|_{\mathcal{X}}(&\|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}} \\ &+ \|\varepsilon\|_{\mathcal{X}} + \frac{C_0}{2}\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2).\end{aligned}$$

By adding $\lambda^2\|\bar{Z}\|_{\mathcal{X}}^2$ to both sides and rearranging the terms, we get

$$\begin{aligned}\left(\|A(\bar{U}_\lambda) - \bar{X}\|_{\mathcal{X}} - \lambda\|\bar{Z}\|_{\mathcal{X}}\right)^2 + \lambda\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2 \leq \|\varepsilon\|_{\mathcal{X}}^2 + 2\lambda\|\bar{Z}\|_{\mathcal{X}}(&\|\varepsilon\|_{\mathcal{X}} \\ &+ \frac{C_0}{2}\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2) + \lambda^2\|\bar{Z}\|_{\mathcal{X}}^2.\end{aligned}$$

Next, we take expectations on both sides. First, recall that Assumption 7 implies (3.4.2), i.e. $\mathbb{E}[\|\varepsilon\|_{\mathcal{X}}] \leq \tau$ and therefore, with Assumption 9,

$$\mathbb{E}[\|\bar{Z}\|_{\mathcal{X}}\|\varepsilon\|_{\mathcal{X}}] \leq \beta\tau.$$

Assumption 9 implies also that

$$\mathbb{E}[\|\bar{Z}\|_{\mathcal{X}}\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2] \leq \beta\mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2].$$

We then get that

$$\mathbb{E}[\left(\left\|A(\bar{U}_\lambda) - \bar{X}\right\|_{\mathcal{X}} - \lambda\|\bar{Z}\|_{\mathcal{X}}\right)^2] + \lambda\mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2] \leq \tau^2 + 2\lambda\beta\tau + \lambda^2\beta^2 + \lambda C_0\beta\mathbb{E}[\|\bar{U}_\lambda - \bar{U}\|_{\mathcal{U}}^2].$$

In particular,

$$\mathbb{E}[\left\|\bar{U}_\lambda - \bar{U}\right\|_{\mathcal{U}}^2] \leq \frac{(\tau + \beta\lambda)^2}{\lambda(1 - \beta C_0)},$$

where we used that $\beta C_0 < 1$ by assumption. Finally, the value of λ that minimizes the above bound is

$$\lambda_* = \frac{\tau}{\beta},$$

and the corresponding error bound is

$$\mathbb{E}[\left\|\bar{U}_{\lambda_*} - \bar{U}\right\|_{\mathcal{U}}^2] \leq \frac{4}{1 - \beta C_0}\tau\beta,$$

which proves the result. □

To apply Theorem 3.1, we consider the truncated square loss:

$$\ell(u, u') = \|T(u - \bar{U}_0) - T(u' - \bar{U}_0)\|_{\mathcal{U}}^2, \tag{3.4.10}$$

where T is the truncation operator defined in (3.3.8). Next, for every $\lambda \in (0, +\infty)$, consider $f_\lambda(\bar{X}) := \bar{U}_\lambda$, where $\bar{U}_\lambda$ is given by (3.4.5). The corresponding expected risk is given in this case by

$$L(\bar{U}_\lambda) = \mathbb{E}[\|T(U_\lambda - \bar{U}_0) - T(\bar{U} - \bar{U}_0)\|_{\mathcal{U}}^2],$$

We now analyze the error corresponding to the choice of the regularization parameter with ERM. Consider independent and identical copies $(\bar{X}_i, \bar{U}_i)$, $i = 1, ..., n$, $n \in \mathbb{N}$, of the pair of random variables $(\bar{X}, \bar{U})$ as in (3.4.1). The ERM problem is given by

$$\widehat{\lambda}_\Lambda \in \arg\min_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} \left\|T(\bar{U}_\lambda^i - \bar{U}_0) - (\bar{U}_i - \bar{U}_0)\right\|_{\mathcal{U}}^2. \tag{3.4.11}$$

with $\bar{U}_\lambda^i := \bar{U}_\lambda(\bar{X}_i)$ and where we used that $T(\bar{U}_i - \bar{U}_0) = \bar{U} - \bar{U}_0$ for every $i = 1, ..., n$ since $\|\bar{U} - \bar{U}_0\|_{\mathcal{U}} \leq 1$ by Assumption 9. In the following result we derive an upper bound corresponding to the expected risk.

Corollary 3.9. *Suppose that Assumptions 7, 8 and 9 hold. Let Assumption 3 be satisfied with $\lambda_* = \frac{\tau}{\beta}$ and choose the loss as in (3.4.10). Let $\eta \in (0, 1)$. Then, with probability at least $1 - \eta$,*

$$L(\bar{U}_{\widehat{\lambda}_\Lambda}) \leq \frac{(1 + Q)^2}{2Q(1 - \beta C_0)}\tau\beta + \frac{26}{n}\log\frac{2N}{\eta}.$$

Proof. To prove the result, it is enough to show that Assumptions 1 and 2 are satisfied. First, note that Assumption 1 is satisfied since the truncated square loss in (3.4.10) is bounded by 4. We recall that the operator T defined in (3.3.8) is $1-$Lipschitz. Therefore, Theorem 3.8 implies

$$L(\bar{U}_\lambda) \leq \mathbb{E}[\left\|\bar{U}_\lambda - \bar{U}\right\|_{\mathcal{U}}^2] \leq \Phi(\lambda),$$

with $\Phi(\lambda) = \frac{(\tau+\beta\lambda)^2}{\lambda(1-\beta C_0)}$. The minimizer of Φ is $\lambda_* = \frac{\tau}{\beta}$ with $\Phi(\lambda^*) = \frac{4\tau\beta}{1-\beta C_0}$ and, for every $q \geq 1$ we have that

$$\Phi(q\lambda_*) = \frac{(1+q)^2}{q}(1-\beta C_0)^{-1}\tau\beta = \frac{(1+q)^2}{4q}\Phi(\lambda^*).$$

Since the function

$$C : [1, +\infty) \to [0, +\infty); \quad C(q) := \frac{(1+q)^2}{4q}$$

is non-decreasing, Assumption 2 is satisfied. The result then follows from Theorem 3.1. □

Corollary 3.9 establishes an upper bound on the expected risk of $\bar{U}_{\widehat{\lambda}_\Lambda}$, corresponding to the choice of the optimal regularization parameter based on ERM in the grid Λ. Actually, it ensures that the error obtained when considering $\widehat{\lambda}_\Lambda$ is close to that of λ_*, except for an additive error term that decreases with n. Notably, the dependence on the cardinality of the grid N is only logarithmic.

3.5 General variational approaches for linear inverse problems

In this section, we consider the linear inverse problem setting in Section 3.3, with Assumption 4 on the noise. We study Tikhonov regularization with a general function $R : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ instead of the squared norm,

$$\bar{U}_\lambda(\omega) \in \operatorname*{arg\,min}_{u\in\mathcal{U}} \frac{1}{2}\left\|Au - \bar{X}(\omega)\right\|_{\mathcal{X}}^2 + \lambda R(u). \tag{3.5.1}$$

In this section, we assume that the set of minimizers of the function

$$u \mapsto \frac{1}{2}\left\|Au - \bar{X}(\omega)\right\|_{\mathcal{X}}^2 + \lambda R(u)$$

is nonempty for almost every $\omega \in \Omega$, and that $\omega \mapsto \bar{U}_\lambda(\omega)$ is a measurable selection of the set of minimizers. This setting includes various examples of sparsity-inducing regularizers beyond Hilbertian norms, see e.g. [15] for references. We discuss specific examples in Sections 3.5.1 and 3.5.2. For this class of regularization schemes, natural error metrics are given by the Bregman divergence D_R, defined by (2.2.3), and the symmetric Bregman distance d_R, given (2.2.4), and which is well defined as long as both arguments belong to $\operatorname{int}(\operatorname{dom} R)$, as we already mentioned in Section 2.2. To derive an error bound we consider the following assumptions.

Assumption 10. *We assume that* $R \in \Gamma_0(\mathcal{U})$ *and that* $\operatorname{dom}(\partial R) = \operatorname{int}(\operatorname{dom} R)$.

The previous assumption is satisfied in two main settings, which are discussed in the following: in the finite-dimensional setting; i.e., where $\operatorname{dom} R = \mathbb{R}^d$, $1 \leq d < +\infty$, and in the infinite-dimensional setting, where, by Lemma 2.13, R must be essentially smooth. The following assumption extends classic smoothness assumptions for this setting.

Assumption 11. *The random variable $\bar{U}$ takes values in $\operatorname{int}(\operatorname{dom} R)$ a.s.. Moreover, we assume that there exists a random variable $\bar{Z} \in \mathcal{X}$, measurable with respect to the σ-algebra generated by $\bar{U}$, such that $A^*\bar{Z} \in \partial R(\bar{X})$ a.s.. Finally, we assume that there exists $\beta > 0$ such that*

$$\mathbb{E}[\|\bar{Z}\|_{\mathcal{X}}^2] \leq \beta^2.$$

Assumption 11 can be seen as a generalization of the source condition for the squared norm regularization in Assumption 5 for $s = 1$. In the following, we analyze the behavior of $d_R(\bar{U}_\lambda, \bar{U})$ with respect to the regularization parameter λ. We first show that this quantity is well-defined. From the optimality condition for the Tikhonov problem (3.5.1) we derive that, a.s.,

$$\frac{1}{\lambda} A^*(\bar{X} - A\bar{U}_\lambda) \in \partial R(\bar{U}_\lambda). \tag{3.5.2}$$

In particular, we know that $\bar{U}_\lambda \in \operatorname{dom} \partial R$ and so, by Assumption 10, that $\bar{U}_\lambda \in \operatorname{int}(\operatorname{dom} R)$. Moreover, from Assumption 11 we have that $\bar{U} \in \operatorname{int}(\operatorname{dom} R)$ a.s., and

$$A^*\bar{Z} \in \partial R(\bar{U}).$$

Then, the symmetric Bregman distance is well defined, and can be written as

$$d_R(\bar{U}_\lambda, \bar{U}) = \left\langle \frac{1}{\lambda} A^*(\bar{X} - A\bar{U}_\lambda) - A^*\bar{Z}, \bar{U}_\lambda - \bar{U} \right\rangle_{\mathcal{U}}. \tag{3.5.3}$$

The Bregman distances we consider (both the symmetric and the standard one) are based on the specific subdifferentials considered in the latter formula. In the setting above, we have the following upper bound.

Theorem 3.10. *Suppose that Assumptions 4, 10 and 11 hold. Then, for all $\lambda \in (0, +\infty)$, we have*

$$\mathbb{E}[d_R(\bar{U}_\lambda, \bar{U})] \leq \frac{\tau^2}{2\lambda} + \frac{\beta^2 \lambda}{2}. \tag{3.5.4}$$

In particular, taking $\lambda_* = \frac{\tau}{\beta}$, we have

$$\mathbb{E}[d_R(\bar{U}_{\lambda_*}, \bar{U})] \leq \beta\tau. \tag{3.5.5}$$

Proof. The identities and inequalities below are intended to hold a.s.. By Assumption 11,

$$\begin{aligned}
\lambda d_R(\bar{U}_\lambda, \bar{U}) + \|A(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}^2 &= \langle A^*(\bar{X} - A\bar{U}_\lambda) - \lambda A^*\bar{Z}, \bar{U}_\lambda - \bar{U} \rangle_{\mathcal{U}} \\
&\quad + \|A(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}^2 \\
&= \langle \bar{X} - A\bar{U}_\lambda - \lambda\bar{Z} + A\bar{U}_\lambda - A\bar{U}, A(\bar{U}_\lambda - \bar{U}) \rangle_{\mathcal{X}} \\
&= \langle \bar{X} - A\bar{U} - \lambda\bar{Z}, A(\bar{U}_\lambda - \bar{U}) \rangle_{\mathcal{X}} \\
&\leq \frac{1}{2} \|\bar{X} - A\bar{U} - \lambda\bar{Z}\|_{\mathcal{X}}^2 + \frac{1}{2} \|A(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}^2.
\end{aligned}$$

Rearranging the terms, we obtain

$$\lambda d_R(\bar{U}_\lambda, \bar{U}) + \frac{1}{2} \|A(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}^2 \leq \frac{1}{2} \|\bar{X} - A\bar{U} - \lambda\bar{Z}\|_{\mathcal{X}}^2.$$

Taking the conditional expectation with respect to $\bar{U}$, we get

$$\begin{aligned}
\lambda \mathbb{E}[d_R(\bar{U}_\lambda, \bar{U})|\bar{U}] + \frac{1}{2}\mathbb{E}[\|A(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}^2 |\bar{U}] &\leq \frac{1}{2}\mathbb{E}[\|\bar{X} - A\bar{U}\|_{\mathcal{X}}^2 |\bar{U}] + \frac{\lambda^2}{2}\mathbb{E}[\|\bar{Z}\|_{\mathcal{X}}^2 |\bar{U}] \\
&\quad - \lambda \mathbb{E}[\langle \bar{X} - A\bar{U}, \bar{Z} \rangle_{\mathcal{X}} |\bar{U}].
\end{aligned}$$

By Assumption 11, $\bar{Z}$ is a measurable function with respect to $\bar{U}$, and therefore the last term is zero since $\bar{X} = A\bar{U} + \varepsilon$ and by Assumption 4. Thus, if we take the full expectation, the previous inequality implies

$$\begin{aligned}\lambda\mathbb{E}[d_R(\bar{U}_\lambda, \bar{U})] + \frac{1}{2}\mathbb{E}[\|A(\bar{U}_\lambda - \bar{U})\|_{\mathcal{X}}^2] &\leq \frac{1}{2}\mathbb{E}[\|\bar{X} - A\bar{U}\|_{\mathcal{X}}^2] + \frac{\lambda^2}{2}\mathbb{E}[\|\bar{Z}\|_{\mathcal{X}}^2] \\ &\leq \frac{\tau^2}{2} + \frac{\beta^2\lambda^2}{2},\end{aligned}$$

by Assumptions 4 and 11. Therefore,

$$\mathbb{E}[d_R(\bar{U}_\lambda, \bar{U})] \leq \frac{\tau^2}{2\lambda} + \frac{\beta^2\lambda}{2}. \tag{3.5.6}$$

The value of λ minimizing the above upper bound is

$$\lambda_* = \frac{\tau}{\beta}.$$

and the theorem follows. □

We add a small remark regarding the above result.

Remark 3.11. Following [31], the above analysis can be extended considering $\mathcal{U}$ to be a Banach space embedded in a Hilbert space. In this case, the inner product in $\mathcal{U}$ needs to be replaced by the corresponding duality pairing.

In the rest of the section, we will apply Theorem 3.1 to different loss functions, all based on the Bregman divergence. To perform the analysis, additional assumptions are needed on R to ensure that the hypotheses of Theorem 3.1 are satisfied, e.g. the boundedness of the loss. We focus on two different settings: the case of sparsity-inducing regularizers, of the form $R(u) = |Gu|$, where G is a general linear and bounded operator and $|\cdot|$ a general norm (for instance, the ℓ^1-norm), and the case of regularizers R of Legendre type.

3.5.1 Sparsity inducing regularizers

In this section, we focus on the finite-dimensional setting, where $\mathcal{U} = \mathbb{R}^d$, $1 \leq d < +\infty$, is endowed with the Euclidean norm $\|\cdot\|_2$. We study sparsity-inducing regularizers such as the ℓ^1 norm [7] or the Total Variation regularization [115]. Towards this end, we first introduce a generic norm on $\mathbb{R}^m$ (not necessarily the Euclidean one), which we denote by $\|\cdot\|$, and the corresponding dual norm $\|\cdot\|_*$. We then fix a linear and bounded operator $G\colon (\mathcal{U}, \|\cdot\|_2) \to (\mathbb{R}^m, \|\cdot\|)$. We will consider the following structural assumption.

Assumption 12. *The regularizer $R\colon \mathbb{R}^d \to \mathbb{R}$ is defined by setting, for every $x \in \mathbb{R}^d$,*

$$R(u) = \|Gu\|, \tag{3.5.7}$$

and $\|G\|_{\mathrm{op}} \leq C_1$, for some constant $C_1 > 0$ (here the operator norm is meant with respect to the spaces $\mathcal{U} = \mathbb{R}^d$ and $\mathbb{R}^m$ with their norms $\|\cdot\|_2$ and $\|\cdot\|$, respectively).

The above condition describes the class of sparsity inducing regularizers we consider. This class includes, for instance, all of the examples of regularization functions mentioned in Section 2.3.2, but also Graph-Lasso [103], penalties for multitask learning [105], group lasso [117], or ℓ^q penalties [74], among others (see [80] and references therein).

For these regularization functions R, the subdifferential can be written as

$$\partial R(\cdot) = G^* \partial \|\cdot\|(G\cdot),$$

which is nonempty at every point $u \in \mathcal{U}$. Moreover, we recall that $\partial\|\cdot\|$ is given by Example 2.7. In this section, we consider the loss function defined by the Bregman divergence for every u, $u' \in \mathbb{R}^d$:

$$\ell(u, u') = D_R(u, u') \tag{3.5.8}$$

where D_R is defined as in (2.2.3), for some subgradient $s_R(u') \in \partial R(u')$. As before, if we let $f_\lambda(\bar{X}) = \bar{U}_\lambda$ for every $\lambda \in (0, +\infty)$, then the corresponding expected risk is given by

$$L(\bar{U}_\lambda) = \mathbb{E}[D_R(\bar{U}, \bar{U}_\lambda)]. \tag{3.5.9}$$

In this case, and as in Section 3.3, we also assume that the random variable $\bar{U}$ is such that $\|\bar{U}\|_2 \leq 1$ a.s.. In order to write the ERM in this case, we consider independent and identical copies $(\bar{X}_i, \bar{U}_i)$, $i = 1, ..., n$, $n \in \mathbb{N}$, of the pair of random variables $(\bar{X}, \bar{U})$, distributed as in (3.3.1). Finally, if we denote $\bar{U}^i_\lambda := \bar{U}_\lambda(\bar{X}_i)$, then the ERM is given by

$$\widehat{\lambda}_\Lambda \in \arg\min_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} D_R(\bar{U}_i, \bar{U}^i_\lambda). \tag{3.5.10}$$

A similar approach has been applied in order to learn the optimal regularization parameter in the case of R being the Total Variation [45]. We can now state the probabilistic error estimates for this setting.

Corollary 3.12. *In the setting of this subsection, let Assumptions 4, 11 and 12 be satisfied, let Assumption 3 be satisfied with $\lambda_* = \frac{\tau}{\beta}$ as in Theorem 3.10 and choose the loss as in (3.5.8). Let $\eta \in (0, 1)$. Then, with probability at least $1 - \eta$,*

$$L(U_{\widehat{\lambda}_\Lambda}) \leq \frac{1 + Q^2}{Q} \beta\tau + \frac{13 C_1}{n} \log \frac{2N}{\eta}. \tag{3.5.11}$$

Proof. To apply Theorem 3.1, we need to check that Assumptions 1 and 2 are satisfied. For every $u \in \mathbb{R}^d$ with $\|u\| \leq 1$ and $u' \in \mathbb{R}^d$, we have

$$\begin{aligned}
D_R(u, u') &= \|Gu\| - \|Gu'\| - \langle G^* s_{\|\cdot\|}(Gu'), u - u' \rangle_{\mathbb{R}^m} \\
&= \|Gu\| - \|Gu'\| - \langle s_{\|\cdot\|}(Gu'), Gu - Gu' \rangle_{\mathbb{R}^m} \\
&= \|Gu\| - \langle s_{\|\cdot\|}(Gu'), Gu \rangle_{\mathbb{R}^m} \\
&\leq (1 + \|s_{\|\cdot\|}(Gu')\|_*) \|Gu\| \\
&\leq 2 \|G\|_{\mathrm{op}} \|u\|_2 \\
&\leq 2 C_1.
\end{aligned}$$

Hence, the loss function is bounded on the cylinder $\{(u, u') \in \mathbb{R}^{d \times d} : \|u\|_2 \leq 1\}$, and Assumption 1 is therein satisfied with $M = 2C_1$. We are left to show that Assumption 2 is satisfied for $f_\lambda(\bar{X}) = \bar{U}_\lambda$ and L defined as in (3.5.11). From the inequality

$$D_R(\bar{U}, \bar{U}_\lambda) \leq d_R(\bar{U}, \bar{U}_\lambda)$$

and Theorem 3.10, we derive that

$$L(\bar{U}_\lambda) \leq \Phi(\lambda),$$

where $\Phi(\lambda) = \frac{\tau^2}{2\lambda} + \frac{\beta^2 \lambda}{2}$. The latter is minimized by $\lambda_* = \frac{\tau}{\beta}$ and satisfies

$$\Phi(q\lambda_*) \leq \frac{1 + q^2}{2q} \beta\tau,$$

where the multiplicative factor depending on q is a non-decreasing function for $q \geq 1$. The statement then follows from Theorem 3.1. □

3.5.2 Legendre Regularizers

In this section, we consider Legendre regularizers 2.12. We will rely on the following assumption.

Assumption 13. *The function $R\colon \mathcal{U} \to [0,+\infty]$ is Legendre.*

In particular, Assumption 13 implies Assumption 10, since $\mathrm{dom}(\partial R) = \mathrm{int}(\mathrm{dom}\, R)$ by Lemma 2.13.

Now, we need to consider a projection operator in order to prove that Assumption 1 is satisfied. To do so, let $u_0 \in \mathrm{int}(\mathrm{dom}\, R)$ and $r > 0$ such that $B_r := \{u \in \mathcal{U} \,:\, \|u - u_0\| \leq r\}$ is a subset of $\mathrm{int}(\mathrm{dom}\, R)$. With this, we consider π_{B_r} be the Bregman projection onto B_r, defined in (2.2.5). Observe that, by definition, $\pi_{B_r}(u) \in B_r \subseteq \mathrm{int}(\mathrm{dom}\, R)$ for every $u \in \mathcal{U}$. Next, recalling that it always holds $\mathrm{int}(\mathrm{dom}\, R) \subseteq \mathrm{dom}(\partial R)$, we know that the subdifferential of R is non empty at each point of B_r, which is contained in $\mathrm{dom}\, \partial R$. In particular, under Assumption 13, R is essentially smooth and hence, again by Lemma 2.13, the subdifferential of R is single valued on B_r. Then, for every $u \in B_r$, we denote by $\nabla R(u)$ the subdifferential of R at $u \in B_r$.

We need an additional assumption on the function R in the set B, namely a uniform upper-bound for the norm of ∇R.

Assumption 14. *There exists $C_2 > 0$ such that*

$$\sup_{u \in B_r} \|\nabla R(u)\|_{\mathcal{U}} \leq C_2.$$

Note that, since R is Legendre and essentially smooth, then ∇R is locally bounded on $\mathrm{int}(\mathrm{dom}\, R)$. This means that for every $u \in \mathrm{int}(\mathrm{dom}\, R)$ there exists $\varepsilon > 0$ such that $\sup_{z \in B_\varepsilon(u)} \|\nabla R(z)\| < +\infty$, but this does not imply the validity of Assumption 14. In this context, we consider the loss function defined for all u, $u' \in \mathcal{U}$ as the Bregman divergence between the projections onto B, namely

$$\ell(u, u') = D_R(\pi_{B_r}(u), \pi_{B_r}(u')), \tag{3.5.12}$$

which is univocally defined since $\pi_{B_r}(u') \in B$, and the subdifferential of R is non empty and single valued on B. Next, for every $\lambda \in (0, +\infty)$, consider $f_\lambda(\bar{X}) := \bar{U}_\lambda$, where $\bar{U}_\lambda$ is defined as in (3.5.1). The corresponding expected risk for this problem is given by

$$L(f) = \mathbb{E}[D_R(\pi_{B_r}(\bar{U}), \pi_{B_r}(\bar{U}_\lambda))]. \tag{3.5.13}$$

In this case, and in opposition with the other sections where we assumed that $\|\bar{U}\| \leq 1$, we assume that $\bar{U}$ is such that $\bar{U} \in B$ a.s.. As in the previous sections, we want to bound the expected risk of the regularization method $\bar{U}_\lambda$, when the regularization parameter λ is selected by ERM. To do so, we consider independent and identical copies $(\bar{X}_i, \bar{U}_i)$, $i = 1, ..., n$, $n \in \mathbb{N}$, of the pair of random variables $(\bar{X}, \bar{U})$, distributed as in (3.3.1). Let $\bar{U}_\lambda^i := \bar{U}_\lambda(\bar{X}_i)$ for every $i = 1, ..., n$. Then, the ERM writes as

$$\widehat{\lambda}_\Lambda \in \arg\min_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} D_R(\bar{U}_i, \pi_{B_r}(\bar{U}_\lambda^i))), \tag{3.5.14}$$

where we used that $\pi_{B_r}(\bar{U}_i) = \bar{U}_i$, $i = 1, ..., n$, by assumption. The corresponding error bound is given in the following corollary.

Corollary 3.13. *Let Assumptions 4, 11, 13 and 14 be satisfied, let Assumption 3 be satisfied with $\lambda_* = \frac{\tau}{\beta}$ as in Theorem 3.10 and choose the loss as in (3.5.12). Let $\eta \in (0,1)$. Then, with probability at least $1-\eta$,*

$$L(\bar{U}_{\widehat{\lambda}_\Lambda}) \leq \frac{1+Q^2}{Q}\beta\tau + \frac{26C_2 r}{n}\log\frac{2N}{\eta}.$$

Proof. To prove the statement, we will rely again on Theorem 3.1. Therefore we just need to show that Assumptions 1 and 2 hold. We first show that Assumption 1 is satisfied. From $\pi_{B_r}(u), \pi_{B_r}(u') \in B$ and Assumption 14, recalling that ∂R is single valued on B, it follows that

$$\begin{aligned} 0 \;\leq\; \ell(u,u') \;&=\; D_R(\pi_{B_r}(u), \pi_{B_r}(u')) \;\leq\; D_R(\pi_{B_r}(u), \pi_{B_r}(u')) + D_R(\pi_{B_r}(u'), \pi_{B_r}(u)) \\ &=\; \langle \nabla R(\pi_{B_r}(u)) - \nabla R(\pi_{B_r}(u')), \pi_{B_r}(u) - \pi_{B_r}(u')\rangle \;\leq\; 4C_2 r. \end{aligned}$$

Then, the considered loss function (3.5.12) is bounded and Assumption 1 is satisfied with $M = 4C_2 r$. Next, we check Assumption 2 in this setting. First, observe that both $\bar{U}$ and $\bar{U}_\lambda$ belong to $\mathrm{dom}(\partial R)$ a.s. since $\bar{U} \in B \subseteq \mathrm{int}(\mathrm{dom}\, R)$ by assumption and by the optimality condition stated in (3.5.2). Then, the subdifferential of R is non-empty (and so single valued) at $\bar{U}$ and $\bar{U}_\lambda$ and, by Lemma 2.15,

$$d_J(\bar{U}, \bar{U}_\lambda) \geq D_R(\bar{U}, \bar{U}_\lambda) \geq D_R(\bar{U}, \pi_{B_r}(\bar{U}_\lambda)) + D_R(\pi_{B_r}(\bar{U}_\lambda), \bar{U}_\lambda).$$

Again, since $\bar{U} \in B$ a.s., we have that $\pi_{B_r}(\bar{U}) = \bar{U}$ a.s.. Then, the previous inequality implies that

$$L(\bar{U}_\lambda) = \mathbb{E}[D_R(\pi_{B_r}(\bar{U}), \pi_{B_r}(\bar{U}_\lambda))] = \mathbb{E}[D_R(\bar{U}, \pi_{B_r}(\bar{U}_\lambda))] \leq \mathbb{E}[d_J(\bar{U}, \bar{U}_\lambda)]. \tag{3.5.15}$$

Theorem 3.10 gives the bound $\mathbb{E}[d_J(\bar{U}, \bar{U}_\lambda)] \leq \Phi(\lambda)$, where $\Phi(\lambda) = \frac{\tau^2}{2\lambda} + \frac{\beta^2\lambda}{2}$. So, together with (3.5.15), this implies that

$$L(\bar{U}_\lambda) \leq \Phi(\lambda).$$

The minimizer of $\Phi(\lambda)$ is given by $\lambda_* = \frac{\tau}{\beta}$ with $\Phi(\lambda_*) = \beta\tau$. We derive directly from the definition that

$$\Phi(q\lambda_*) = \frac{1+q^2}{2q}\beta\tau = \frac{1+q^2}{2q}\Phi(\lambda_*)$$

for any $q \geq 1$, where the multiplicative term $\frac{1+q^2}{2q}$ is a non-decreasing function for $q \geq 1$. Hence, Assumption 2 is satisfied and we can apply Theorem 3.1 to obtain the desired result. □

3.6 Numerical results

In this section, we provide empirical validation of the theoretical results discussed in the previous sections. We consider different experimental settings and, for each of them, we illustrate the expected risk decay, evaluated at the learned regularization parameter $\widehat{\lambda}_\Lambda$, showing that it goes to a certain constant as n goes to infinity. First, we consider the setting of linear inverse problems with squared norm regularization. In this case, we focus on Tikhonov regularization, defined in Example 2.29, and the Landweber method, given in Example 2.30. For both of them we compare the proposed data-driven procedure with the so-called quasi-optimality criterion [11]. Then, we turn to more general regularization penalties. More precisely, we consider the problem of denoising and deblurring sparse signals with the ℓ^1-norm, and TV denoising for images.

In all the subsequent experiments, the expected risk L is always approximated with the empirical one, computed with either $N = 500$ or $N = 1000$ points, depending on the complexity of the experiment. Similarly, the optimal parameter λ_* is selected on a sufficiently fine grid to approximate the interval $(0, +\infty)$.

Code statement: All of the simulations have been implemented in Python on a laptop with 32GB of RAM and 2.2 GHz Intel Core I7 CPU. In Section 3.6.2 we also use the library Numerical Tours by G. Peyré [108]. The code is available at https://github.com/TraDE-OPT/Learning-the-Regularization-Parameter.

3.6.1 Spectral regularization methods

In this section, we empirically analyze the proposed data-driven parameter selection strategy for Tikhonov regularization and the Landweber method to solve an instance of a linear inverse problem as in Section 3.3. We consider a problem of the form

$$\bar{X} = A\bar{U} + \varepsilon,$$

which we describe next. The operator A is a 70×70 square matrix, with Gaussian entries $a_{i,j} \sim N(0, 1)$, $1 \leq i, j \leq 70$, that will be then normalized by its operator norm, which in this case coincides with the 2-norm. To ensure that Assumption 5 is satisfied with a known exponent, we define the random variable $\bar{U} \in \mathbb{R}^{70}$ as

$$\bar{U} = (A^*A)^s \bar{Z},$$

with $s > 0$ to be fixed later and $\bar{Z}$ sampled uniformly in the unit ball. This, jointly with $\|A\|_2 \leq 1$, ensures that $\|\bar{U}\| \leq 1$ a.s.. Note that, in this setting, Assumption 5 is satisfied with $\beta = 1$. Finally, $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$, which satisfies Assumption 4. The training set is obtained by sampling n independent pairs $(\bar{x}_i, \bar{u}_i)$, $i = 1, ..., n$, from the previous model. The section is divided into two parts:

- empirical validation of the theoretical results,
- comparison of the studied method with the quasi-optimality criterion [130].

In both cases, every experiment is run 30 times, and we report both the mean (in solid lines) and the values between the 5^{th}-percentile and 95^{th}-percentile of the data (in shaded regions).

Illustration of the data-driven parameter choice

We start considering the problem described in Section 3.6.1 with noise level $\tau = 10^{-2}$ and source condition $s = 0.5$. Starting from the training set $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^{50}$, for every $\lambda \in \Lambda$, we define the empirical risk for the Tikhonov regularized solution as

$$\widehat{L}(\bar{U}_\lambda) = \frac{1}{50} \sum_{i=1}^{50} \|T\bar{U}_\lambda^i - \bar{u}_i\|^2, \tag{3.6.1}$$

where $\bar{U}_\lambda^i := (A^*A + \lambda I)^{-1} A^* \bar{x}_i$ and T stands for the truncation operator onto the unit ball. The empirical risk for the Landweber method is defined analogously, where in this case $\bar{U}_\lambda^i = (I - \gamma A^*A)^{\lfloor 1/\lambda \rfloor} A^* \bar{x}_i$ with constant stepsize $\gamma = 0.2$. For both Tikhonov regularization and Landweber iteration, we build a grid of regularization parameters $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ as in Assumption 3, namely with $\lambda_j = \lambda_1 Q^{j-1}$ for $j = 1, \ldots, N$ and $Q = (\lambda_N/\lambda_1)^{1/(N-1)}$. In the case of Tikhonov, we choose $\Lambda \subseteq [10^{-4}, 100]$ with $N = 500$

and so $Q \approx 1.0281$. For Landweber, we choose $\Lambda \subseteq [10^{-3}, 1]$, while N remains the same and $Q \approx 1.0139$. According to Section 3.2, the parameter proposed by our approach is $\widehat{\lambda}$, a minimizer of (3.6.1) within the grid Λ. In Figure 3.6.1, the function $\lambda \in \Lambda \mapsto \widehat{L}(\bar{U}_\lambda)$ is plotted for Tikhonov regularization. For Landweber, we plot the function in terms of number of iterations k.

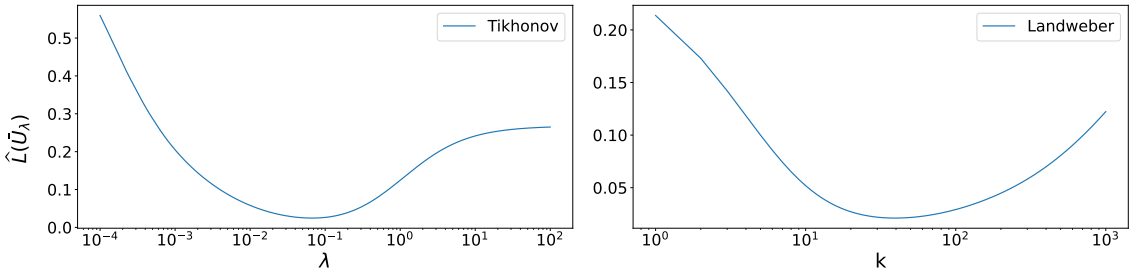

Figure 3.1: Empirical risk trajectories, of the Tikhonov and Landweber regularization methods, with respect to the regularization parameter λ.

Illustration of Theorem 3.5

In this section we investigate the dependence on the noise level τ of the error $L(\bar{U}_{\lambda_*})$, see equation (3.3.5) in Theorem 3.5. For every fixed noise level $\tau > 0$ of ε, let $\lambda_*(\tau)$, or $k_*(\tau)$ in the case of Landweber, be a minimizer of the expected risk,

$$\lambda_*(\tau) \in \underset{\lambda \in (0,+\infty)}{\arg\min}\ L(\bar{U}_\lambda). \tag{3.6.2}$$

As stated in Theorem 3.5, $L(\bar{U}_{\lambda_*(\tau)})$ goes to zero when τ vanishes. The parameter α in Assumption 6 plays an important role in the bound, since $L(\bar{U}_{\lambda_*(\tau)}) \lesssim \tau^{4\alpha/(2\alpha+1)}$. In particular, we expect $L(\bar{U}_{\lambda_*(\tau)})$ to go to 0 faster when α increases. For Tikhonov, $\alpha = \min\{1, s\}$ (since 1 is the qualification parameter for Tikhonov regularization). For Landweber, instead, $\alpha = s$. The influence of s on the decay rate of the reconstruction error is shown in Figure 3.6.1 for the values $s = 0.5$ and $s = 1$. To determine $\lambda_*(\tau)$, we first consider 30 different values of the noise level τ within the interval $[10^{-4}, 10^{-1}]$. The selected smoothness parameters allow us to gain a deeper insight into the behavior of the expected risk with respect to the deterministic rate obtained in Theorem 3.5. In Figure 3.6.1, we illustrate the quantity $L(\bar{U}_{\lambda^*(\tau)})/\tau^{(4s)/(2s+1)}$, where it can be seen that all the curves are bounded when τ goes to zero. We can also observe that the quantity of interest is not going to zero, therefore suggesting that the derived bounds are tight.

In the following experiment, we study the behavior of the best empirical regularization parameters, $\widehat{\lambda}(\tau)$ and $\widehat{k}(\tau)$, with respect to the noise level τ and the smoothness parameter s for both Tikhonov and Landweber methods. Here, the empirical risk is computed with 10 training points for smoothness parameters $s = 0.5$ and 1. We fix 30 different values of the noise level τ in the interval $[10^{-4}, 10^{-1}]$, and we consider the following grids: $\Lambda \subseteq [10^{-5}, 1]$ with $N = 500$ in the case of Tikhonov regularization, and $\Lambda \subseteq [10^{-4}, 1]$ with $N = 5000$ for Landweber. It can be seen that the empirical parameters $\widehat{\lambda}(\tau)$ and $\widehat{k}(\tau)$ exhibit a similar behavior to the a priori optimal ones ([64] and Theorem 3.5): in the case of Tikhonov regularization, $\hat{\lambda}(\tau)$ increases with the noise, and in the case of Landweber, the number of iterations decreases with respect to the noise. The smoothness parameter has also an effect on the optimal regularization parameter: $\hat{\lambda}$ is increasing with respect to s, while the required number of iterations in Landweber is decreasing. This behavior can be observed in Figure 3.6.1.

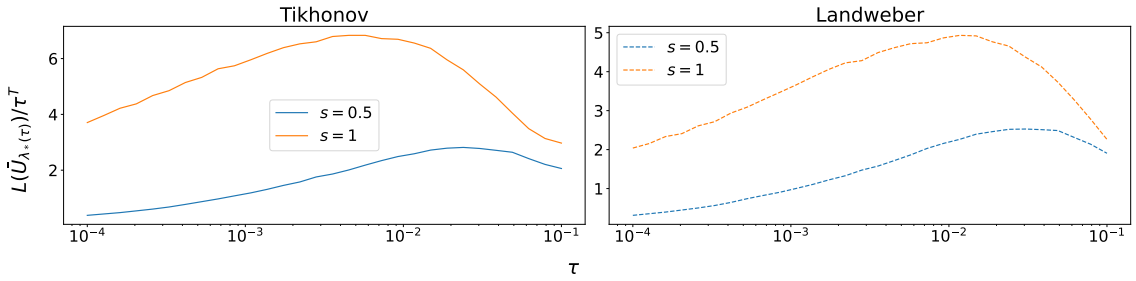

Figure 3.2: Behavior of $L(\bar{U}_{\lambda_*})$ with respect to the rate τ^T, $T = (4s)/(2s+1)$, obtained in Theorem 3.5, for different smoothness parameters s for Tikhonov and Landweber. It can be seen that each trajectory is upper bounded, as suggested by the rate in Theorem 3.5. The horizontal axes are shown in logarithmic scale.

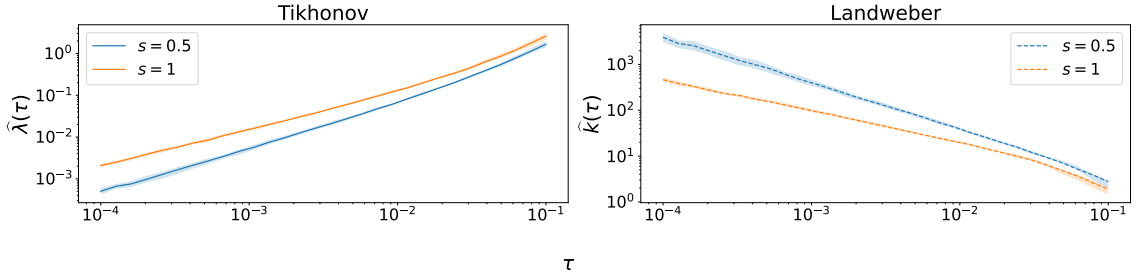

Figure 3.3: Value of $\widehat{\lambda}$, $\widehat{k}$ when varying the noise level for both Tikhonov and Landweber methods. Both parameters have been selected over a training set of 10 points, constructed with different smoothness parameters as shown in the plot. Solid lines represent the mean value, while the shaded regions represent the 5^{th}-percentiles and 95^{th}-percentiles over 30 trials. Both axis are shown in logarithmic scale.

Illustration of error bounds

In this section, we discuss some numerical experiments supporting the error bound stated in Corollary 3.6, both for Tikhonov and Landweber regularization methods. By Corollary 3.6, with high probability, there exist constants c_1, $c_2 > 0$ such that

$$L(\bar{U}_{\widehat{\lambda}_\Lambda}) \leq c_1 \tau^{4\alpha/(2\alpha+1)} + \frac{c_2}{n}.$$

Therefore, the quantity $L(\bar{U}_{\widehat{\lambda}_\Lambda})$ with fixed noise level $\tau > 0$ behaves as $L(\bar{U}_{\lambda_*})$ up to an additive constant. The same holds for fixed n, and $\tau \to 0$. We consider the same setting as for Figure 3.6.1 with noise level $\tau = 0.01$ and smoothness parameter $s = 0.5$. We define the empirical risk, $\widehat{L}(\bar{U}_\lambda)$, for every $n \in \{5, 10, ..., 100\}$, where we sample fresh training points for every different value of n, and we denote by $\widehat{\lambda}(n)$ and $\widehat{k}(n)$ the parameters corresponding to the minimizers of the empirical risk with n points. In Figure 3.6.1 we show that $L(\bar{U}_{\widehat{\lambda}(n)})$ goes to a certain constant, that depends on the noise level, when n increases, see Figure 3.6.1.

Next, we illustrate the behavior of the expected risk L with respect to the noise level τ. First, we fix as smoothness parameter $s = 0.5$ and consider 30 different values of the noise level τ within the interval $[10^{-4}, 10^{-1}]$. Next, for every τ, we find $\lambda_*(\tau)$, $k_*(\tau)$ as the minimizers of the expected risk L. Then, we fix the grid $\Lambda \subseteq [10^{-5}, 1]$ with $N = 500$ and $Q \approx 1.0233$ in the case of Tikhonov, and $\Lambda \subseteq [10^{-4}, 1]$ with $N = 3000$ and $Q \approx 1.0031$ in the case of Landweber. With this, we find $\widehat{\lambda}_\Lambda(\tau)$, $\widehat{k}_\Lambda(\tau)$ as the minimizers of the empirical risk $\widehat{L}(\bar{U}_\lambda)$, constructed with $n = 5$ freshly generated training points. In Figure 3.6.1 we

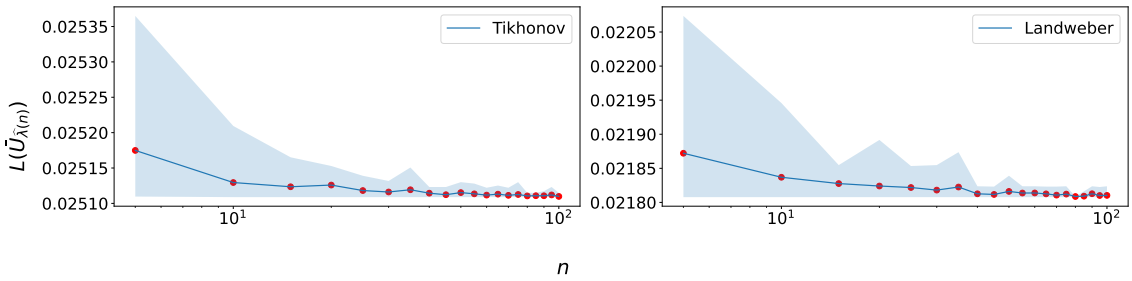

Figure 3.4: Behavior of $L(\bar{U}_{\widehat{\lambda}(n)})$, both for Tikhonov and Landweber regularization, as a function of n. The solid lines represent the mean value, while the shaded regions represent the 5^{th}-percentiles and 95^{th}-percentiles over 30 trials. The x-axis is shown in logarithmic scale.

plot, for every noise level τ, the values $L(\bar{U}_{\lambda_*(\tau)})$ and $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$ in the cases of Tikhonov and Landweber, showing that their behavior with respect to τ is comparable.

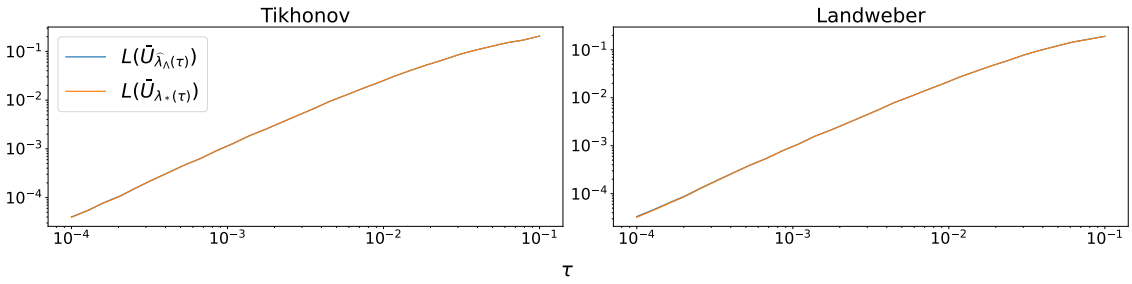

Figure 3.5: Comparison between $L(\bar{U}_{\lambda_*(\tau)})$, in orange, and $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$, in blue, when varying the noise level τ both for Tikhonov and Landweber regularization. As it can be observed, in such a scale, the lines almost coincide.

Comparison with the quasi-optimality criterion

In this section we compare our data-driven approach to the quasi-optimality criterion [130]. The latter is one of the most common and simple-to-implement heuristic parameter selection methods and does not require the noise level to be computed. Theoretical guarantees on its performance are available in the stochastic inverse problems setting [11]. First, note that the computational cost of the two methods can be very different. The quasi-optimality criterion performs instance-wise as all the usual parameter selection methods; i.e. given a set of test data $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^{n_{\text{test}}}$, $n_{\text{test}} \in \mathbb{N}$, and a regularization method $\bar{U}_\lambda$, it outputs the best regularization parameter $\widehat{\lambda}_i$ for each $\bar{x}_i$, $i = 1, ..., n_{\text{test}}$. This could lead to high computational costs when the number of test points is big. Indeed, the method needs to be run as many times as the number of points, and for each test point the computation of the whole regularization path is required (see below). On the contrary, our algorithm requires to have access to a training set, but then, on test problems, the learned parameter $\widehat{\lambda}$ will be the same for every $i = 1, ..., n_{\text{test}}$, and only one regularized problem needs to be solved. In the following we compare the two approaches in terms of average performance on the test problems for Tikhonov and Landweber methods.

For Tikhonov regularization, we fix a grid of $N = 1000$ regularization parameters $\Lambda \subseteq [10^{-5}, 10]$, with $Q \approx 1.0139$ and we denote $\bar{U}^i_{\lambda_j}$ the solution of the regularized problem for the parameter λ_j and datum $\bar{x}_i$, $i \in \{1, \dots, n_{\text{test}}\}$. We fix $n_{\text{test}} = 50$. For each pair

$L^{\text{learn}} - L^{\text{qo}}$, Tikhonov				
noise lev.	$\tau = 10^{-3}$	$\tau = 10^{-2}$	$\tau = 10^{-1}$	$\tau = 0.5$
mean	-0.0025	-0.0665	-0.6071	-0.9935
std	4.07×10^{-7}	4.27×10^{-6}	4.22×10^{-5}	0.0

Table 3.1: Mean value and standard deviation of the error difference between our method and the quasi-optimality criterion. Above, we compare methods in the case of Tikhonov regularization for different values of the noise level.

$L_{\text{learn}} - L^{\text{qo}}$, Landweber				
noise lev.	$\tau = 10^{-3}$	$\tau = 10^{-2}$	$\tau = 10^{-1}$	$\tau = 0.5$
mean	-0.9987	-0.9348	-0.5042	0.5775
std	4.60×10^{-7}	1.56×10^{-6}	1.11×10^{-16}	0.0

Table 3.2: Mean value and standard deviation of the error difference between our method and the quasi-optimality criterion with different values of the noise level.

$(\bar{x}_i, \bar{u}_i)$ in the test set, we select the parameter with the quasi-optimality criterion, namely we set $\lambda_i^{\text{qo}} = \lambda_{j_*(i)}$, where $j_*(i)$ is defined as

$$j_*(i) \in \underset{j \in \{1,\dots,1000\}}{\arg\min} \|\bar{U}^i_{\lambda_j} - \bar{U}^i_{\lambda_{j+1}}\|.$$

Our method instead provides a unique $\widehat{\lambda}_\Lambda$, depending on the training set. For this experiment, we fix a training set of 1000 points. We then compare the average test error corresponding to the two methods, where, for the quasi-optimality criterion we consider

$$L^{\text{qo}} = \frac{1}{50} \sum_{i=1}^{50} \|\bar{U}^i_{\lambda_i^{\text{qo}}} - \bar{u}_i\|^2.$$

For the Landweber iteration, we fix a grid of $N = 800$ regularization parameters $\Lambda \subseteq [1/10^3, 1]$, with $Q \approx 1.0087$ we follow the implementation of the quasi-optimality criterion proposed in [9], and we define $\lambda_i^{\text{qo}} = \lambda_{j_*}$, where $j_*(i)$ is defined as

$$j_*(i) \in \underset{j \in \{1,\dots,800\}}{\arg\min} \|\bar{U}^i_{2\lfloor 1/\lambda_{j+1} \rfloor} - \bar{U}^i_{\lfloor 1/\lambda_{j+1} \rfloor}\|,$$

and we compare the average test error as for the Tikhonov method.

We denote the test error corresponding to our method L^{learn} (for both Tikhonov and Landweber) and we compute the quantity $L^{\text{learn}} - L^{\text{qo}}$ for 30 different realizations of the training set. We show in tables 3.6.1 and 3.6.1 the mean value and standard deviation of the proposed experiment for both Tikhonov and Landweber with source condition $s = 0.5$. As the tables suggest, the data-driven selection method performs differently than the quasi-optimality criterion for both Tikhonov and Landweber methods. On the one hand, in the case of Tikhonov regularization, the difference between the two studied methods is small when the noise level is small. Instead, when such noise level increases, the learned regularization parameter performs considerably better. In the case of Landweber, it can be seen in 3.6.1 that the learned regularization parameter performs better for lower values of the noise level.

3.6.2 Sparsity inducing regularizers

In this section, we explore the theoretical results in Section 3.5.1 for three different examples: denoising and deblurring of a sparse signal, and Total Variation regularization

for image denoising, which were already discussed in Section 2.3.2. In particular, we will focus on illustrating, experimentally, Corollary 3.12. To do so, we will perform the same experiments that we did for the spectral case: first, we show that the expected risk, evaluated at the best empirical parameter $\widehat{\lambda}_\Lambda$ with fixed noise level τ, tends to a certain constant when the number of training points goes to infinity. Second, we show that the expected risk, when evaluated at the best empirical parameter $\widehat{\lambda}_\Lambda$ for a fixed number of training points, has a comparable behavior, with respect to the noise level τ, to the expected risk evaluated at its minimum. We start with the simplest case: denoising of a sparse signal.

Denoising of a sparse signal

Let $\mathcal{X} = \mathcal{U} = \mathbb{R}^d$, $1 \leq d < \infty$ and consider the sparse signal denoising problem stated in Example 2.22:

$$x = u^* + \varepsilon. \tag{3.6.3}$$

Consider the white noise model $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$, with noise level $\tau > 0$, and assume the solution u^* to be such that $\|u^*\|_2 \leq 1$ as required by assumption. The sparsity paramter We consider the same regularization approach as in Section 2.3.2; i.e. $\bar{U}_\lambda = \mathcal{S}_\lambda(x)$, for every $\lambda \in (0, +\infty)$, to be the soft-thresholding operator as regularization method.

As an illustrative example, we show in Figure 3.6 the behavior of the empirical risk in this setting, where the training set $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^n$ is generated according to (3.6.3) with $d = 1024$, $s = 64$ and noise level $\tau = 0.25$,

$$\widehat{L}(\mathcal{S}_\lambda) = \frac{1}{n} \sum_{i=1}^{n} D_{\|\cdot\|_1}(\bar{u}_i, \mathcal{S}_\lambda(\bar{x}_i)), \tag{3.6.4}$$

for $n = 10$ and λ in a grid $\Lambda \subseteq [10^{-4}, 10]$ with $N = 1000$ and so $Q \approx 1.0116$. As it can be seen, this empirical behavior matches the theoretical one: we first recall the Bregman divergence for this case: for every u, $x \in \mathbb{R}^d$, we have

$$D_{\|\cdot\|_1}(u, \mathcal{S}_\lambda(x)) = \|u\|_1 - \langle s_{|\cdot|}(x), u\rangle = \|u\|_1 - \langle \mathrm{sign}(x), u\rangle.$$

On the one hand, observe that, for every $i = 1, ..., n$, $\mathrm{sign}(\mathcal{S}_\lambda(\bar{u}_i)) \to \mathrm{sign}(\bar{x}_i)$ as $\lambda \to 0$. This leads to

$$\widehat{L}(\mathcal{S}_\lambda) \to \frac{1}{n} \sum_{i=1}^{n} D_{\|\cdot\|_1}(\bar{u}_i, \bar{x}_i), \quad \text{as } \lambda \to 0,$$

where the right hand side is constant. On the other hand, the for every $\lambda \in (0, +\infty)$ with $\lambda > \sup_{i=1,...,n} \|\bar{x}_i\|_\infty$, we have that $\mathcal{S}_\lambda(\bar{x}_i) = 0$ for every $i = 1, ..., n$. Therefore, $D_{\|\cdot\|_1}(\bar{u}_i, \mathcal{S}_\lambda(\bar{x}_i)) = \|\bar{u}_i\|$ for every $i = 1, ..., n$ and so

$$\widehat{L}(\mathcal{S}_\lambda) \to \frac{1}{n} \sum_{i=1}^{n} \|\bar{u}_i\|_1, \quad \text{as } \lambda \to +\infty,$$

where the right hand side is again constant in this case.

Next, we illustrate, from a numerical point of view, Corollary 3.12 for this setting. First, we show that the expected risk for this problem, when evaluated at the learned regularization parameter $\widehat{\lambda}_\Lambda$, goes to a certain constant as n goes to infinity. First, we fix as noise level $\tau = 0.25$ and a grid of regularization parameters of $N = 1000$ points $\Lambda \subseteq [10^{-5}, 1]$, with $Q \approx 1.0116$ and, for every $n \in \{1, 2, ..., 20\}$ we define $\widehat{\lambda}(n)$ as a minimizer of the empirical risk (3.6.4) where, for every n, we consider an independent set of

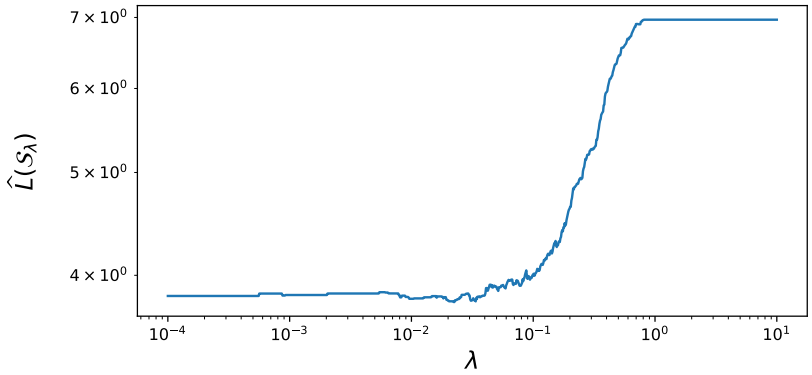

Figure 3.6: Behavior of $\widehat{L}(\mathcal{S}_\lambda)$ with respect to the regularizaton parameter λ for the signal denoising problem.

training points $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^n$, generated according to (3.6.3). In Figure 3.6.2, we plot the quantity $L(\mathcal{S}_{\widehat{\lambda}(n)})$ for different values of the dimension, $d \in \{512, 1024, 2048\}$, and fixed sparsity $s = 16$, showing that, empirically, it is converging to a certain constant when the number of training points goes to infinity. As expected, this convergence does not depend on the dimension of the problem.

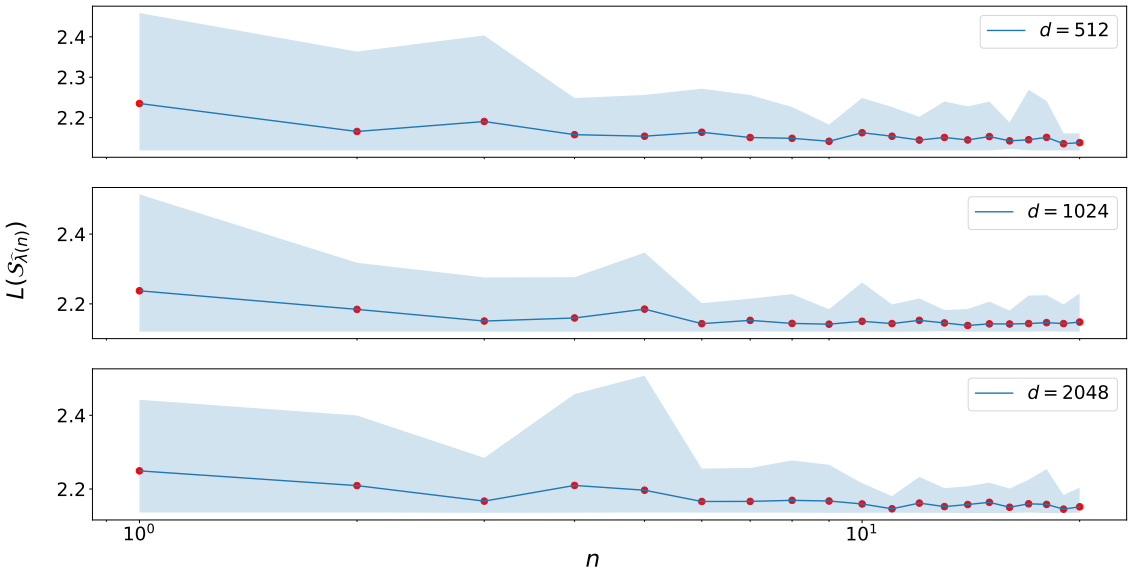

Figure 3.7: Behavior of $L(\mathcal{S}_{\widehat{\lambda}(n)})$ as a function of n for different values of the dimension. The solid lines represent the mean value, while the shaded regions represent the 5^{th}-percentiles and 95^{th}-percentiles over 30 trials. The x-axis is shown in logarithmic scale.

Finally, we show the behavior of the expected risk L with respect to the noise level τ. First, we fix $d = 1024$ and sparsity $s = 16$. Next, we fix 30 different values of the noise level $\tau \in [0.1, 1]$. Then, for every value of the noise level τ, we find $\lambda_*(\tau)$ as the minimizer of the expected risk L. After, we consider the grid $\Lambda \subseteq [10^{-5}, 1]$ with $N = 500$ and $Q \approx 1.0233$. With this, we find $\widehat{\lambda}_\Lambda$ as the minimizer of the empirical risk $\widehat{L}$, constructed with $n = 5$ fresh training points. n Figure 3.6.2, we confirm that the behavior of both $L(\mathcal{S}_{\lambda_*(\tau)})$ and $L(\mathcal{S}_{\widehat{\lambda}_\Lambda(\tau)})$ with respect to τ is comparable.

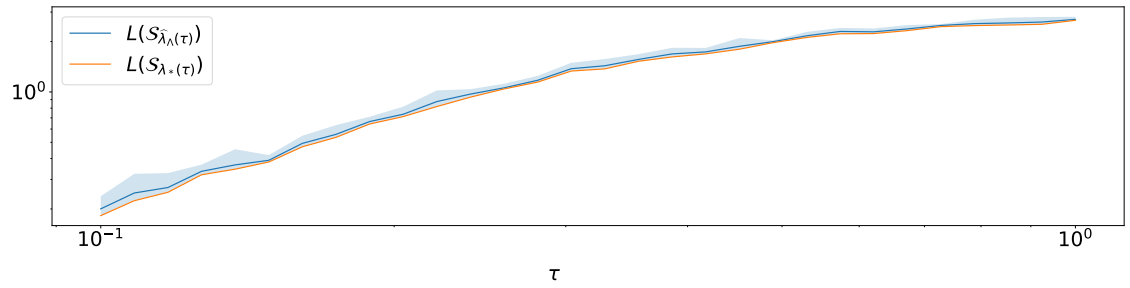

Figure 3.8: Behavior of the expected risk L with respect to the noise level τ. Recall that $L(\mathcal{S}_{\widehat{\lambda}_\Lambda(\tau)})$ has been computed 30 times. We therefore report the mean value, in a solid line, and the values between the 5^{th}-percentile and 95^{th}-percentile, which corresponds to the shaded region. Both axis are shown in logarithmic scale.

Deblurring of a sparse signal

In this section, we consider the problem of deblurring a sparse signal explained in Example 2.23. Let $\mathcal{X} = \mathcal{U} = \mathbb{R}^d$, $1 \leq d < \infty$ and consider

$$x = Au^* + \varepsilon, \tag{3.6.5}$$

where A is the blur operator defined therein and $\varepsilon \sim N(0, \tau^2 \operatorname{Id})$ with noise level $\tau > 0$. In order to define a regularization method, we consider the Lasso problem, developed in Section 2.3.2. We let $\bar{U}_\lambda$ to be the output of FISTA [14] until convergence; i.e. until the difference between iterates is smaller than 10^{-6}. We now aim at illustrating Corollary 3.12; i.e., showing the error behavior of the learned regularization parameter when n goes to infinity. For this example, we fix $\tau = 0.1$ and the grid of admissible regularization parameters to be $\Lambda \subseteq [10^{-2}, 1]$ with $N = 50$ and $Q \approx 1.0985$. The ERM writes as

$$\widehat{\lambda}(n) \in \arg\min_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} D_{\|\cdot\|_1}(\bar{u}_i, \bar{U}_\lambda^i).$$

where $\bar{U}_\lambda^i := \bar{U}_\lambda(\bar{x}_i)$ and, for every $n \in \{5, 10, ..., 50\}$, we consider independent sets of training points $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^n$, that have been generated according to (3.6.5). According to Corollary 3.12, the expected risk L evaluated at $\widehat{\lambda}_\Lambda(n)$, $L(\bar{U}_{\widehat{\lambda}_\Lambda(n)})$, should converge to certain constant when $n \to \infty$. We plot this behavior in Figure 3.9.

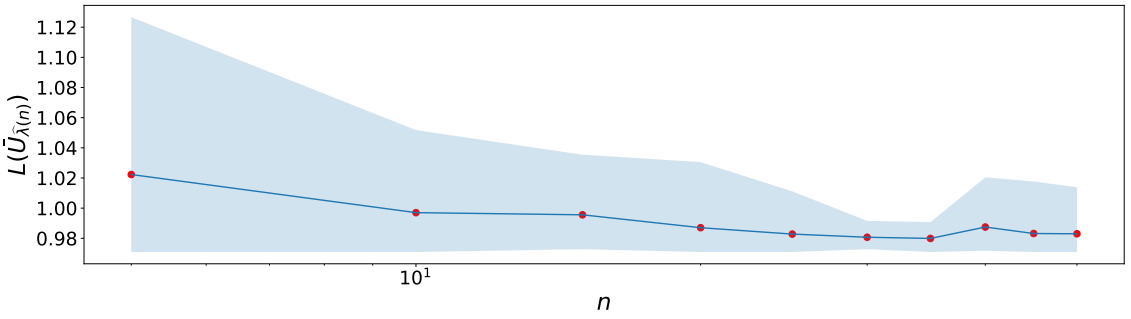

Figure 3.9: Behavior of $L(\bar{U}_{\widehat{\lambda}(n)})$ for the signal deblurring problem, showing that it goes to a certain constant as n increases. The solid line represents the mean value, while the shaded region represents the values between 5^{th}-percentile and 95^{th}-percentile over 30 trials. The x-axis is shown in logarithmic scale.

Next, we aim to show, empirically, that the behavior of the learned regularization parameter and the optimal one is comparable with respect to the noise level τ. We therefore fix 30 different values of the noise level within the interval $[0.1, 1]$ and define, for every τ, $\lambda_*(\tau)$ as the minimizer of the expected risk L. After, we fix a grid of regularization parameters $\Lambda \subseteq [10^{-2}, 1]$ with $N = 10$ and $Q \sim 2.1544$. Hence, $\widehat{\lambda}_\Lambda$ will be the minimizer of the empirical risk $\widehat{L}$, constructed with $n = 5$ freshly generated training points for every value of the noise level τ. In Figure 3.6.2, we plot the quantities $L(\bar{U}_{\lambda_*(\tau)})$ and $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$, showing that their behavior is comparable with respect to the noise level τ.

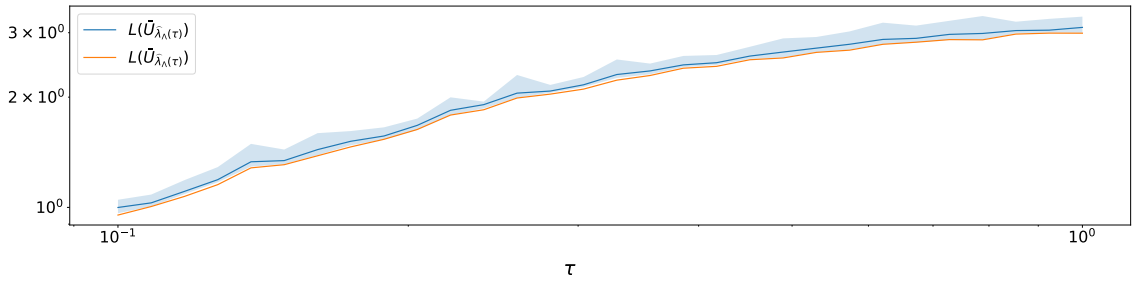

Figure 3.10: Behavior of the expected risk L, with respect to the noise level τ, for the signal deblurring problem. Recall that $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$, in blue, has been computed 30 times. We therefore report the mean value, in a solid line, and the values between the 5^{th}-percentile and 95^{th}-percentile, which corresponds to the shaded region. Both axis are shown in logarithmic scale.

Finally, we show one example of a reconstructed signal using our regularization parameter choice. In order to learn the parameter $\widehat{\lambda}$, we first construct a training set of $n_{\text{train}} = 100$ clean/corrupted signals with the same distribution as the test element that we want to reconstruct, with noise level $\tau = 0.1$. Then, the regularization parameter will be the minimizer of the empirical risk (3.5.10) with respect to the fixed training set. We show in the third row of Figure 3.6.2, the resulting regularized solution with the learned regularization parameter.

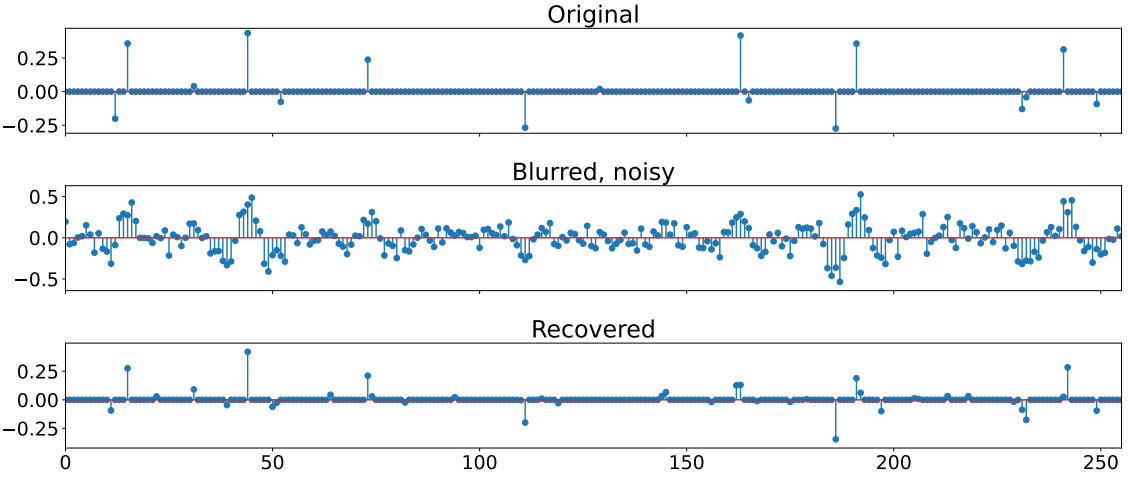

Figure 3.11: Deblurring of a sparse noisy, blurred signal with learned regularization parameter. In the first row, we show the original signal; in the second, its blurred and noisy version; and in the third row, the regularized solution with learned regularization parameter.

Total Variation for image denoising

In this section, we use our data-driven algorithm for choosing the regularization parameter of the anisotropic Total Variation regularizer (2.3.11). To do so, we focus on the image denoising problem as in Example 2.22: we let $\mathcal{X} = \mathcal{U} = \mathbb{R}^{d\times d}$, $1 \leq d < \infty$ and

$$x = u^* + \varepsilon, \tag{3.6.6}$$

where $\varepsilon \sim N(0, \tau^2 \,\mathrm{Id})$ with noise level $\tau > 0$. Moreover, we consider $\bar{U}_\lambda$ as a solution of the variational problem (2.3.9), via FISTA on the dual problem [38], as already indicated in Section 2.3.2, until convergence; i.e. until the difference between iterates is smaller than 10^{-8}. In order to illustrate Corollary 3.12, we first show the behavior of the expected risk, evaluated at the learned regularization parameter $\widehat{\lambda}_\Lambda$ for this example.

We consider the MNIST dataset [59] of 28×28 images of digits from 0 to 9, and corrupt them as in (3.6.6). In order to give empirical evidence of Corollary 3.12, we fix the noise level $\tau = 0.25$. Then, we fix a grid of $N = 50$ points $\Lambda \subseteq [10^{-3}, 1]$, with $Q \approx 1.1514$. For every $n \in \{5, 10, ..., 50\}$, we let $\widehat{\lambda}(n)$ be a minimizer of the empirical risk,

$$\widehat{\lambda}(n) \in \arg\min_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{TV}}(\bar{u}_i, \bar{U}^i_\lambda).$$

where, for every n, we consider an independent training set of points $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^n$ randomly selected from a set of 3000 images. We therefore plot in Figure 3.6.2 the behavior of the expected risk L, evaluated a $\widehat{\lambda}(n)$. As it can be seen, it converges to a certain constant as $n \to \infty$.

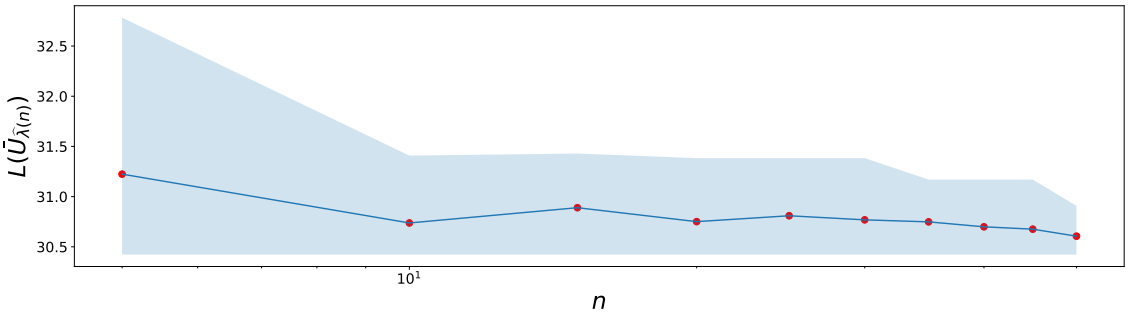

Figure 3.12: Behavior of $L(\bar{U}_{\widehat{\lambda}(n)})$ as a function of n for the image denoising problem, showing that it goes to a certain constant as n increases. The solid line represents the mean value, while the shaded regions represent the 5^{th}-percentiles and 95^{th}-percentiles over 30 trials. The x-axis is shown in logarithmic scale.

We now want to illustrate the behavior of the expected risk, with respect to the noise level τ, for the TV regularization problem. To do so, we consider the exact same experimental setting as we did for Figure 3.6.2 for the signal deblurring problem, and we show in Figure 3.6.2 that the behavior of both $L(\bar{U}_{\lambda_*(\tau)})$ and $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$ is comparable with respect to τ.

Finally, as an illustrative example, we explore the performance of the studied parameter selection method on test images from the MNIST dataset. We compute four different data-driven regularization parameters for four different training sets, each of 100 training points, and check the reconstruction results of the TV regularized solution for two

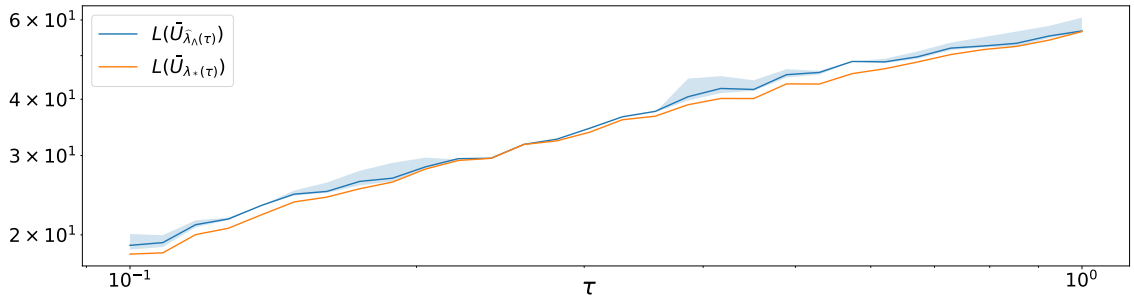

Figure 3.13: Comparison between $L(\bar{U}_{\lambda_*(\tau)})$, in orange, and $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$, in blue, when varying the noise level τ for the Total Variation regularization. In the case of $L(\bar{U}_{\widehat{\lambda}_\Lambda(\tau)})$, the solid line represents the mean value, while the shaded region represents the values between the 5^{th}-percentile and 95^{th}-percentile over 30 trials. Bot axes are shown in logarithmic scale.

different digits in the test set. The results are shown in Figure 3.6.2. We observe that the recovery results on single test images may vary depending on the set of points that was used for training. This is expected, since our parameter selection method has been designed in order to perform effectively on average.

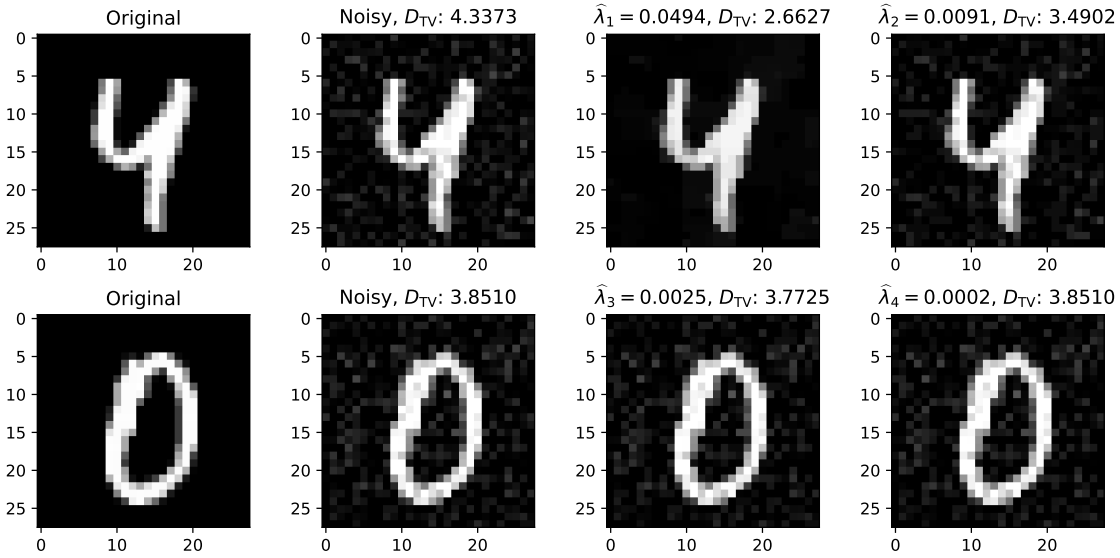

Figure 3.14: Total Variation denoising algorithm for two digits in the test set. From left to right, in every row, we plot the original image, its noisy version, and the recovery obtained with different regularization parameters. We also include, accordingly, the Bregman divergence with respect to the original image and the value of the regularization parameter that has been used for such recovery.

CHAPTER 4

Learning Firmly Nonexpansive Operators

4.1 Introduction

In this work, we aim to devise a data-driven approach for constructing firmly nonexpansive operators. This class of functions has gained significant popularity with the emergence of the so-called Plug-and-Play (PnP) methods [138]. In short, a PnP method substitutes the proximal operator in first-order optimization algorithms with a general "denoising" function. In particular, if such a denoiser is firmly nonexpansive, then the resulting algorithm still converges to a fixed point. This strategy stems from the common challenge of lacking closed-form expressions for proximal operators of convex functions. PnP approaches have proven to show promising results in many applications [93, 142], but few theoretical guarantees have been provided.

This work aims to address this gap by constructing an operator as the minimizer of an ERM problem within the space of nonexpansive operators. By borrowing classical tools from the theory of Lipschitz functions, we show the existence of a minimizer. Further, since such a space is infinite-dimensional, a proper discretization will be required in order to construct it in practice. Our proposal is to define a space of certain piecewise affine functions on simplices. With such a selection, we prove that the problem has a nice and easy-to-implement finite-dimensional characterization.

Some works have already tried to tackle the theoretical difficulties in this context. For instance, in order to provide convergence guarantees to a fixed point, authors in [107] employ a stochastic penalization of the Lipschitz norm of the gradient of the proposed network. Additionally, the studies conducted in [85, 121] are, to the best of our knowledge, pioneering efforts in constructing α-averaged, nonexpansive, neural networks. Furthermore, in [107], the authors demonstrated that their proposed set of denoisers is dense within a subclass of maximal monotone operators. This work remains the only known instance of a density analysis in this context, highlighting the gap in understanding and the potential for further research.

We now describe the organization of this chapter. In Section 4.2 we start by recalling a crucial observation of this work: in order to analyze the problem of learning firmly nonexpansive operators, it is enough to consider nonexpansive operators. We observe that the latter can be see as a subset of the space of Lipschitz functions and, in particular, we notice that this space is isometrically isomorphic with the dual of the so-called Arens–Eels space. Therefore, considering the weak* topology induced by this characterization, we also present some preliminary results which will be useful for the forthcoming analysis. In Section 4.3 we present the abstract constrained optimization setting. By combining

Theorem 4.8 and Proposition 4.9, we show that the our learning problem has a minimizer. Next, in Theorem 4.10 we show that the ERM $\Gamma-$converges to its continuous version. In Sections 4.3.3 and 4.3.4 we construct the class of operators that will be then used in practice. As we mentioned above, such a construction will be based on defining piecewise affine operators on simplicial partitions. We conclude the theoretical part in Section 4.4, where we explaining in detail how to design the PnP versions of some algorithms of interest and show their convergence by standard results: PnP Forward-Backward Splitting, PnP ADMM, PnP Douglas–Rachford, and PnP Chambolle–Pock primal-dual iteration. Finally, in Section 4.5 we develop the experimental setting that we consider and show its applicability to the problem of image denoising.

4.2 Preliminaries

Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ be a real separable Hilbert space. In this section, we introduce the preliminary mathematical tools in order to study the desired learning problem. To do so, we recall that the main motivation of this work was to construct proximal operators of functions in $\Gamma_0(\mathcal{X})$. For such class of mappings, we know that they can be written, by Lemma 2.8, as the resolvent of their subdifferential operator. Hence, a first step would be analyze the properties of the set

$$\mathcal{M} := \{T : \mathcal{X} \to \mathcal{X} \,:\, T = J_A, \text{ where } A \text{ is maximally monotone}\}.$$

We recall that, indeed, every $T \in \mathcal{M}$ is not necessarily a proximal operator, but every proximal operator of a function in $\Gamma_0(\mathcal{X})$ is contained in $\mathcal{M}$. Moreover, by Lemma 2.9, $T \in \mathcal{M}$ if and only if T is firmly nonexpansive. Finally, by Lemma 2.11, for every firmly nonexpansive operator T, there exists a nonexpansive operator N such that $T = (1/2)(N + \mathrm{Id})$, and viceversa. Therefore, from now on, we will restrict our study to the space of nonexpansive operators; i.e. the space

$$\mathcal{N} := \{T : \mathcal{X} \to \mathcal{X} \,:\, T \text{ is nonexpansive}\}.$$

In the context of PnP methods, this observation has been already pointed out, for instance, in [107]. In the next section, and motivated by the fact that nonexpansive operators are 1-Lipschitz operators from $\mathcal{X}$ to $\mathcal{X}$, we study the space of Lipschitz operators and its topological structure. Therein, we describe the required mathematical tools for studying the problem of learning nonexpansive operators.

4.2.1 The spaces $\mathrm{Lip}_0(\mathcal{X})$ and $\mathrm{Lip}(\mathcal{X})$

Given an operator $T : \mathcal{X} \to \mathcal{X}$, the Lipschitz space $\mathrm{Lip}_0(\mathcal{X})$ is defined as the space of Lipschitz operators which vanish at 0. If we endow the space $\mathrm{Lip}_0(\mathcal{X})$ with the norm

$$\|T\|_{\mathrm{Lip}_0} := \sup_{x \neq y} \frac{\|T(x) - T(y)\|_{\mathcal{X}}}{\|x - y\|_{\mathcal{X}}},$$

it becomes a Banach space (see [140]). This norm corresponds to the smallest Lipschitz constant of T. Moreover, the space of Lipschitz operators mapping from $\mathcal{X}$ to $\mathcal{X}$, denoted as $\mathrm{Lip}(\mathcal{X})$ is also a Banach space by endowing it with the norm

$$\|T\|_{\mathrm{Lip}} := \|T(0)\|_{\mathcal{X}} + \|T - T(0)\|_{\mathrm{Lip}_0}$$

Such property can also be derived from the following result.

Proposition 4.1. *Given the space $\mathcal{Y} = \mathcal{X} \times \mathrm{Lip}_0(\mathcal{X})$, we have that*

$$\mathrm{Lip}(\mathcal{X}) \cong \mathcal{Y},$$

where the right-hand side is a Banach space with norm

$$\|(x, T_0)\|_{\mathcal{Y}} := \|x\|_{\mathcal{X}} + \|T_0\|_{\mathrm{Lip}_0},$$

for every $x \in \mathcal{X}$, $T_0 \in \mathrm{Lip}_0(\mathcal{X})$.

Proof. Define the following mapping

$$\begin{aligned} \varphi : \mathrm{Lip}(\mathcal{X}) &\to \mathcal{X} \times \mathrm{Lip}_0(\mathcal{X}) \\ T &\mapsto (T(0), T - T(0)), \end{aligned}$$

and observe that φ is an isometry:

$$\|\varphi(T)\|_{\mathcal{Y}} = \|T(0)\|_{\mathcal{X}} + \|T - T(0)\|_{\mathrm{Lip}_0} = \|T\|_{\mathrm{Lip}}.$$

Moreover, φ is bijective. If we take T_1, $T_2 \in \mathrm{Lip}(\mathcal{X})$ such that

$$(T_1(0), T_1 - T_1(0)) = (T_2(0), T_2 - T_2(0)),$$

we get first that $T_1(0) = T_2(0)$, and by using this fact in the second component we obtain that $T_1 = T_2$. Thus, φ is injective. Finally, let $(x, T_0) \in \mathcal{X} \times \mathrm{Lip}_0(\mathcal{X})$. We aim at proving that there exists $T \in \mathrm{Lip}(\mathcal{X})$ such that $\varphi(T) = (T(0), T - T(0)) = (x, T_0)$. To do so, we choose $T = x + T_0$ (note that $T(0) = x + T_0(0) = x$). □

Due to this result, we can identify every $T \in \mathrm{Lip}(X)$ as $T \equiv (T(0), T - T(0))$. Moreover, by [140, Theorem 3.3], we know that

$$\mathrm{Lip}_0(\mathcal{X}) \cong (\text{Æ}(\mathcal{X}))^* ,$$

where $\text{Æ}(\mathcal{X})$ is the so-called Arens–Eels space of $\mathcal{X}$. Combining this with Proposition 4.1, we get that

$$(\mathcal{X} \times \text{Æ}(\mathcal{X}))^* = \mathcal{X}^* \times \mathrm{Lip}_0(\mathcal{X}) = X \times \mathrm{Lip}_0(\mathcal{X}) \cong \mathrm{Lip}(\mathcal{X}).$$

The above result allows us to gain further insights about the topological structure of the set $\mathrm{Lip}(\mathcal{X})$. First, observe that, being $\mathrm{Lip}(\mathcal{X})$ a dual space, it has an associated weak* topology which, by Proposition 4.1, can be decomposed as $\tau_{\mathcal{X}} \times \tau_{\mathrm{Lip}_0(\mathcal{X})}$, where $\tau_{\mathcal{X}}$ is the weak topology on $\mathcal{X}$ and $\tau_{\mathrm{Lip}_0(\mathcal{X})}$ is the weak* topology on $\mathrm{Lip}_0(\mathcal{X})$. First, observe that, since $\mathcal{X}$ is separable, then so is $\text{Æ}(\mathcal{X})$ (this can be seen as a consequence of [140, Theorem 3.14]) and, hence, every bounded set of $\mathrm{Lip}_0(\mathcal{X})$ is first countable. This implies that we can restrict ourselves to sequences when studying convergence. Next, by referring again to [140, Theorem 3.3], $\tau_{\mathrm{Lip}_0(\mathcal{X})}$ corresponds to the topology of pointwise convergence in bounded sets; i.e. a bounded sequence $(T_k)_{k\in\mathbb{N}}$ in $\mathrm{Lip}_0(\mathcal{X})$ (i.e., there exists $C > 0$ such that $\mathcal{L}(T_k) \leq C$ for every $k \in \mathbb{N}$) is converging to $T \in \mathrm{Lip}_0(\mathcal{X})$ if $T_k(x) \to T(x)$ for every $x \in \mathcal{X}$. Combining all of the above comments, we characterize the weak* topology in $\mathrm{Lip}(X)$ in the following Corollary.

Corollary 4.2. *Let $(T_k)_{k\in\mathbb{N}}$ be a sequence in $\mathrm{Lip}(\mathcal{X})$ and $T \in \mathrm{Lip}(\mathcal{X})$. Then, $T_k \overset{*}{\rightharpoonup} T$ in $\mathrm{Lip}(\mathcal{X})$ if and only if $T_k(0) \rightharpoonup T(0)$ and $T_k(0) - T_k(x) \to T(0) - T(x)$ for every $x \in \mathcal{X}$.*

We now note, as a further consequence of the corollary above, that the weak* convergence becomes uniform convergence when considering compact sets.

Corollary 4.3. *Let $(T_k)_{k\in\mathbb{N}}$ be a sequence in $\mathrm{Lip}(\mathcal{X})$ and let $T \in \mathrm{Lip}(\mathcal{X})$ such that $T_k \overset{*}{\rightharpoonup} T$ as $k \to \infty$. Then,*

(i) Let $K \subset \mathcal{X}$ be a compact set. Then, the sequence $(T_k(x) - T_k(0))_{k\in\mathbb{N}}$ converges to $T(x) - T(0)$ as $k \to \infty$ uniformly for $x \in K$.

(ii) Let $f : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ be a continuous function and let $K \subset \mathcal{X} \times \mathcal{X}$ be a compact set. Then, the sequence $(\langle f(x,y),\ T_k(0) - T(0)\rangle_{\mathcal{X}})_{k\in\mathbb{N}}$ converges uniformly to zero as $k \to \infty$ for every $(x, y) \in K$.

Proof. We start with the proof of (i). Let $\varepsilon > 0$. Since K is compact, there exists a finite collection of points $x_1, ..., x_m \in K$, $m \in \mathbb{N}$, and balls $B_{\frac{\varepsilon}{2}}(x_i)$, centered at x_i with radius $\frac{\varepsilon}{2}$, such that

$$K \subset \bigcup_{i=1}^{m} B_{\frac{\varepsilon}{2}}(x_i)$$

Now, recall that, by Corollary 4.2, $T_k(x_i) - T_k(0) \to T(x_i) - T(0)$ as $k \to \infty$ for every $i = 1, ..., m$. On the other hand, for every $x \in K$, we have that

$$\begin{aligned}\|T_k(x) - T_k(0) - T(x) + T(0)\|_{\mathcal{X}} \leq \|T_k(x) - T_k(x_i)\|_{\mathcal{X}} + \|T(x_i) - T(x)\|_{\mathcal{X}} \\ + \|T_k(x_i) - T_k(0) - T(x_i) + T(0)\|_{\mathcal{X}}.\end{aligned}$$

Observe that, as $T \in \mathrm{Lip}(\mathcal{X})$, there exists $c_1 > 0$ such that $\|T(x) - T(x_i)\|_{\mathcal{X}} \leq c_1\|x - x_i\|_{\mathcal{X}}$. In addition, since the sequence $(T_k)_{k\in\mathbb{N}}$ is weak* convergent, it is bounded; i.e. there exists a constant $c_2 > 0$ such that $\|T_k\|_{\mathrm{Lip}} = \|T_k - T_k(0)\|_{\mathrm{Lip}_0} + \|T_k(0)\|_{\mathcal{X}} \leq c_2$, which immediately implies that $\|T_k - T(0)\|_{\mathrm{Lip}_0} = \|T_k\|_{\mathrm{Lip}_0} \leq c_2$ for every $k \in \mathbb{N}$. Hence, $\|T_k(x) - T_k(x_i)\|_{\mathcal{X}} \leq c_2\|x - x_i\|_{\mathcal{X}}$ for every $k \in \mathbb{N}$. Therefore, if the added point x_i is chosen in such a way that $\|x - x_i\|_{\mathcal{X}} \leq \frac{\varepsilon}{2}$, we get that

$$\|T_k(x) - T_k(0) - T(x) + T(0)\|_{\mathcal{X}} \leq (c_1 + c_2)\frac{\varepsilon}{2} + \|T_k(x_i) - T_k(0) - T(x_i) + T(0)\|_{\mathcal{X}}$$

Finally, for $k \geq k_0$, $k_0 \in \mathbb{N}$ appropriate,

$$\|T_k(x_i) - T_k(0) - T(x_i) + T(0)\|_{\mathcal{X}} < \frac{\varepsilon}{2},$$

concluding the proof of (i). Now, we want to prove (ii). Since the sequence $(T_k(0))_{k\in\mathbb{N}}$ converges weakly to $T(0)$ by Corollary 4.2, there exists a constant $C > 0$ such that $\|T_k(0)\|_{\mathcal{X}} \leq C$ for every $k \in \mathbb{N}$. Let $\varepsilon > 0$. Since f is continuous, and hence uniformly continuous in compact sets, for each $(x, y) \in K$, we can choose $\delta = \delta(x, y) > 0$ such that

$$\|f(x', y') - f(x, y)\|_{\mathcal{X}} \leq \frac{\varepsilon}{4C} \tag{4.2.1}$$

for $\|(x', y') - (x, y)\|_{\mathcal{X}\times\mathcal{X}} < \delta(x, y)$. Next, since K is compact, there exists a finite collection of points $\{(x_i, y_i)\}_{i=1}^{m}$, $m \in \mathbb{N}$, in K and balls $B_{\delta_i}(x_i, y_i)$ centered at (x_i, y_i), with radius $\delta_i := \delta(x_i, y_i)$, such that

$$K \subset \bigcup_{i=1}^{m} B_{\delta_i}(x_i, y_i).$$

By the weak convergence of the sequence $(T_k(0))_{k\in\mathbb{N}}$, we have, for every $i = 1, ..., m$,

$$|\langle f(x_i, y_i), T_k(0) - T(0)\rangle| \leq \frac{\varepsilon}{2}.$$

Finally, for $(x, y) \in K$, choose $i_0 \in \{1, ..., m\}$ such that $\|(x_{i_0}, y_{i_0}) - (x, y)\|_{\mathcal{X}\times\mathcal{X}} < \delta_{i_0}$. We conclude that

$$\begin{aligned}|\langle f(x,y), T_k(0) - T(0)\rangle| &\leq |\langle f(x_{i_0}, y_{i_0}), T_k(0) - T(0)\rangle| \\ &\quad + |\langle f(x,y) - f(x_{i_0}, y_{i_0}), T_k(0) - T(0)\rangle| \\ &\leq \frac{\varepsilon}{2} + \|f(x,y) - f(x_{i_0}, y_{i_0})\|_{\mathcal{X}}\|T_k(0) - T(0)\|_{\mathcal{X}} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4C}2C = \varepsilon.\end{aligned}$$

Since ε was arbitrary, we conclude the proof. □

As we will see in the forthcoming sections, the above results turn out to be key in our analysis, since we have now the correct tools for providing good structural results about the set $\mathcal{N}$.

4.2.2 Properties of $\mathcal{N}$

The next step is to study the topological properties of the subset $\mathcal{N}$. To do so, we recall that nonexpansive operators are 1-Lipschitz operators; i.e., a subset of $\mathrm{Lip}(\mathcal{X})$. As we did in Proposition 4.1, it is easy to prove that

$$\mathcal{N} \cong \mathcal{X} \times \mathcal{N}_0,$$

where $\mathcal{N}_0$ is the space of nonexpansive operators that vanish at 0. Note that, as a subset of $\mathrm{Lip}_0(\mathcal{X})$, it is bounded. Therefore, its topology is the one of pointwise convergence and Corollary 4.2 is also valid for $\mathcal{N}$. Now, for further convenience, let us define, for every $x, y \in \mathcal{X}$, the function

$$\begin{aligned}\varphi_{x,y} : \mathrm{Lip}(\mathcal{X}) &\to \mathbb{R}, \\ T &\mapsto \|T(x) - T(y)\|_{\mathcal{X}} - \|x - y\|_{\mathcal{X}}.\end{aligned}$$

With this, we will prove two results that will be useful in the forthcoming analysis. First, we prove the following set equality.

Lemma 4.4. *We have that*

$$\mathcal{N} = \bigcap_{x,y\in\mathcal{X}} \varphi_{x,y}^{-1}\left((-\infty, 0]\right). \tag{4.2.2}$$

Proof. First, let us observe that, given x, $y \in \mathcal{X}$,

$$\varphi_{x,y}^{-1}\left((-\infty, 0]\right) = \left\{T \in \mathrm{Lip}(\mathcal{X}) \,:\, \|T(x) - T(y)\|_{\mathcal{X}} - \|x - y\|_{\mathcal{X}} \leq 0\right\},$$

i.e., the pre-image of the function φ in the interval $(-\infty, 0]$ coincides with the set of functions that are nonexpansive for the given pair $(x, y) \in \mathcal{X} \times \mathcal{X}$. By taking the intersection over every pair $(x, y) \in \mathcal{X} \in \mathcal{X}$, we recover the definition of the set of nonexpansive operators. Conversely, if $N \in \mathcal{N}$, then N is nonexpansive for every x, $y \in \mathcal{X}$ and hence $N \in \varphi_{x,y}^{-1}\left((-\infty, 0]\right)$ for every x, $y \in \mathcal{X}$. □

Now, we prove several properties of the set $\varphi_{x,y}^{-1}\left((-\infty, 0]\right)$.

Lemma 4.5. *For every x, $y \in \mathcal{X}$, the set $\varphi_{x,y}^{-1}\left((-\infty, 0]\right)$ is non-empty, convex and weak* closed.*

Proof. Let $x, y \in \mathcal{X}$. First, observe that the set is non-empty since the identity mapping is nonexpansive. Let T, $S \in \varphi_{x,y}^{-1}((-\infty, 0])$ and $\alpha \in (0,1)$. Then,

$$\begin{aligned}\|[(1-\alpha)T + \alpha S](x) - [(1-\alpha)T + \alpha S](y)\|_{\mathcal{X}} \\ &\leq (1-\alpha)\|T(x) - T(y)\|_{\mathcal{X}} + \alpha\|S(x) - S(y)\|_{\mathcal{X}} \\ &\leq \|x - y\|_{\mathcal{X}},\end{aligned}$$

hence $(1-\alpha)T + \alpha S \in \varphi_{x,y}^{-1}((-\infty, 0])$, showing convexity. It is only left to prove that the set $\varphi_{x,y}^{-1}((-\infty, 0])$ is weak* closed. To do so, let $T \in \mathrm{Lip}(\mathcal{X})$ and choose a sequence $(T_k)_{k\in\mathbb{N}}$ in $\varphi_{x,y}^{-1}((-\infty, 0])$ such that $T_k \overset{*}{\rightharpoonup} T$ in $\mathrm{Lip}(\mathcal{X})$. We have to show that $T \in \varphi_{x,y}^{-1}((-\infty, 0])$. Using Corollary 4.2, we see that $T_k(x) \rightharpoonup T(x)$ and $T_k(y) \rightharpoonup T(y)$ as $k \to \infty$. Consequently

$$\|T(x) - T(y)\|_{\mathcal{X}} \leq \liminf_{k\to\infty} \|T_k(x) - T_k(y)\|_{\mathcal{X}} \leq \|x - y\|_{\mathcal{X}}.$$

Thus, $T \in \varphi_{x,y}^{-1}((-\infty, 0])$. □

We are ready to state the main result of the class $\mathcal{N}$.

Corollary 4.6. *The set $\mathcal{N}$ is non-empty, weak* closed and convex. Moreover, $\mathcal{N}_0$ is weak* compact.*

Proof. First, note that both convexity and closedness are preserved by arbitrary intersections. Therefore, combining both Lemma 4.4 and Lemma 4.5, we get that $\mathcal{N}$ is convex and weak* closed. Also, the identity mapping belongs to $\mathcal{N}$, so $\mathcal{N}$ is non-empty. Finally, recall that functions in $\mathcal{N}_0$ are nonexpansive, hence bounded in $\mathrm{Lip}_0(\mathcal{X})$. By [101, Corollary 2.6.19], we get that $\mathcal{N}_0$ is weak* compact. □

Finally, in Section 4.3.5, we make use of the following density result, taken from [92], and which we report here for completeness.

Theorem 4.7. *Let $\mathcal{X}$ be a separable Banach space and $\mathcal{Y}$ a general Banach space. Then, for every Lipschitz function $f\colon \mathcal{X} \to \mathcal{Y}$ with Lipschitz constant $L > 0$ and every $\varepsilon > 0$, there exists a function $g\colon \mathcal{X} \to \mathcal{Y}$ which is uniformly Gâteaux differentiable, Lipschitz with the same constant $L > 0$, and such that $\|f - g\|_\infty < \varepsilon$.*

We have all the necessary tools to analyze the problem of learning firmly nonexpansive operators.

4.3 Learning firmly nonexpansive operators

We turn now our attention to the main problem, which is the one of constructing, in a feasible way, nonexpansive operators approximating some given data. First, we state the problem in the set $\mathcal{N}$ from a very general optimization point of view and prove the existence of minimizers in the class $\mathcal{N}$.

4.3.1 A general problem

We are interested in finding solutions of the following problem

$$\inf_{N\in\mathcal{N}} F(N). \tag{P}$$

By mimicking classical results [13], we observe that existence is satisfied if we assume that $F : \mathcal{N} \to (-\infty, \infty]$ is proper, convex, weak* lower semicontinuous; i.e.,

$$\text{if } N_k \overset{*}{\rightharpoonup} N, \text{ then } F(N) \leq \liminf_{k\to\infty} F(N_k), \tag{4.3.1}$$

and coercive in the following sense:

$$\text{if } \|N_k(0)\|_{\mathcal{X}} \to \infty, \text{ then } F(N_k) \to \infty. \tag{4.3.2}$$

We state such result in the following.

Theorem 4.8. *Let $F : \mathcal{N} \to (-\infty, +\infty]$ be a proper and convex operator that satisfies assumptions (4.3.1) and (4.3.2). Then, there exists a minimizer of (P).*

Proof. Define

$$m_0 := \inf_{N \in \mathcal{N}} F(N) \geq -\infty,$$

and observe that, for every $m > m_0$, the sublevel set of F, $\{F \leq m\}$, is bounded by the coercivity of F. Indeed, for every $N \in \mathcal{N}$ we have that $\|N\|_{\mathrm{Lip}_0} \leq 1$, so $\|N_k\|_{\mathrm{Lip}} \to \infty$ as $k \to \infty$ for some sequence $(N_k)_{k\in\mathbb{N}}$ implies $\|N_k(0)\|_{\mathcal{X}} \to \infty$ as $k \to \infty$, and hence, $F(N_k) \to \infty$ as $k \to \infty$, a contradiction. Also, by weak* lower semicontinuity of F, $\{F \leq m\}$ is weak* closed. Therefore, the sublevel sets $\{F \leq m\}$ are weak* compact. Moreover, for every $m > m_0$, each sublevel set is non-empty (since m_0 is the infimum of F). Now, observe that

$$\bigcap_{m>m_0} \{F \leq m\} \neq \emptyset,$$

since the weak* closed family $(\{F \leq m\})_{m>m_0}$ has the finite intersection property (it is nested) and $\{F \leq m\}$ is weak* compact for each $m > m_0$. This also implies that m_0 is finite, as otherwise, F would admit $-\infty$. Finally, as

$$\bigcap_{m>m_0} \{F \leq m\} = \{F \leq m_0\} = F^{-1}(m_0),$$

we have the thesis. □

4.3.2 The statistical model

Let $(\Omega, \mathcal{A}, P)$ be a probability space. The statistical model that we want to study can be expressed in terms of a Supervised Learning problem [54]: we consider the pair of random variables $(\bar{X}, \bar{Z})$ with joint Borel probability measure μ' on $\mathcal{X} \times \mathcal{X}$ which is unknown. Theoretically, we want to find a firmly nonexpansive operator $T^* : \mathcal{X} \to \mathcal{X}$ such that $T^*(\bar{X})$ is close to $\bar{Z}$ in the following sense: : using a quadratic loss, T^* will be the minimizer of the *expected risk*,

$$T^* \in \operatorname*{arg\,min}_{T \in \mathcal{M}} L(T); \quad L(T) := \int_{X\times X} \|T(\bar{x}) - \bar{z}\|^2 \mathrm{d}\mu'(\bar{x}, \bar{z}), \tag{4.3.3}$$

within the space of firmly nonexpansive operators $\mathcal{M}$. As we have already mentioned at the beginning of Section 4.2, this problem can be equivalently formulated in the language of nonexpansive operators by applying the formula $T = \frac{1}{2}(\mathrm{Id} + N)$: indeed, if we define $\bar{U} := 2\bar{Z} - \bar{X}$, then the probability measure μ of the pair $(\bar{X}, \bar{U})$ is given by the push-forward of μ' via the following transformation:

$$\varphi : \mathcal{X} \times \mathcal{X} \to \mathcal{X} \times \mathcal{X}; \quad (x, z) \mapsto (x, 2z - x).$$

We now aim at finding a nonexpansive operator N^* minimizing the expected risk (2.1.1) with respect to the square loss, which in this case writes as

$$N^* \in \operatorname*{arg\,min}_{N \in \mathcal{N}} L(N); \quad L(N) := \int_{\mathcal{X}\times\mathcal{X}} \|N(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \, \mathrm{d}\mu(\bar{x}, \bar{u}), \tag{CP}$$

To prove that there exists a minimizer for (CP), we need to assume that

$$\int_{\mathcal{X}\times\mathcal{X}} \left(\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2\right) \mathrm{d}\mu(\bar{x},\bar{u}) < \infty. \tag{4.3.4}$$

We obtain the desired existence result by combining Assumption (4.3.4) with both Theorem 4.8 and the following result.

Proposition 4.9. *The expected risk* (CP) *is proper, convex, weak* lower semicontinuous, and coercive.*

Proof. First, let us prove that L is proper. If we take $N = \mathrm{Id}$, we get

$$L(\mathrm{Id}) = \int_{\mathcal{X}\times\mathcal{X}} \|\bar{x}-\bar{u}\|_{\mathcal{X}}^2 \,\mathrm{d}\mu(\bar{x},\bar{u}) \leq 2\int_{\mathcal{X}\times\mathcal{X}} \left(\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2\right) \,\mathrm{d}\mu(\bar{x},\bar{u}),$$

where the right-hand side is bounded by Assumption 4.3.4. We want to see now that L is convex. Let N, $S \in \mathcal{N}$. Take $\alpha \in (0,1)$ and observe that

$$\begin{aligned} L(\alpha N + (1-\alpha)S) &= \int_{\mathcal{X}\times\mathcal{X}} \|(\alpha N + (1-\alpha)S)(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \,\mathrm{d}\mu(\bar{x},\bar{u}) \\ &= \int_{\mathcal{X}\times\mathcal{X}} \|\alpha(N(\bar{x})-\bar{u}) + (1-\alpha)(S(\bar{x})-\bar{u})\|_{\mathcal{X}}^2 \,\mathrm{d}\mu(\bar{x},\bar{u}) \\ &\leq \int_{\mathcal{X}\times\mathcal{X}} \left(\alpha\|N(\bar{x})-\bar{u}\|_{\mathcal{X}}^2 + (1-\alpha)\|S(\bar{x})-\bar{u}\|_{\mathcal{X}}^2\right) \,\mathrm{d}\mu(\bar{x},\bar{u}), \end{aligned}$$

where we use that the squared norm is convex. By the linearity of the integral, we obtain the desired result. Let us prove now that L is coercive in the sense of (4.3.2). Let $(N_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{N}$ such that $\|N_k(0)\|_{\mathcal{X}} \to \infty$ when $k \to \infty$. Observe that, for every $\bar{x}$, $\bar{u} \in \mathcal{X}$,

$$\|N_k(\bar{x}) - \bar{u}\|_{\mathcal{X}} \geq \|N_k(0)\|_{\mathcal{X}} - \|N_k(\bar{x}) - N_k(0)\|_{\mathcal{X}} - \|\bar{u}\|_{\mathcal{X}} \geq \|N_k(0)\|_{\mathcal{X}} - (\|\bar{x}\|_{\mathcal{X}} + \|\bar{u}\|_{\mathcal{X}}),$$

since $\|N_k\|_{\mathrm{Lip}_0} \leq 1$ and, if $\|\bar{x}\|_{\mathcal{X}} + \|\bar{u}\|_{\mathcal{X}} \leq \frac{1}{2}\|N_k(0)\|_{\mathcal{X}}$, then

$$\|N_k(\bar{x}) - \bar{u}\|_{\mathcal{X}} \geq \frac{1}{2}\|N_k(0)\|_{\mathcal{X}}. \tag{4.3.5}$$

Now, consider $K > 0$ such that

$$\int_{\{\|\bar{x}\|_{\mathcal{X}}+\|\bar{u}\|_{\mathcal{X}}\leq K\}} \mathrm{d}\mu(\bar{x},\bar{u}) > 0$$

and $K_0 \in \mathbb{N}$ such that, for every $k \geq K_0$, $\frac{1}{2}\|N_k(0)\|_{\mathcal{X}} \geq K$. We obtain that, for every $k \geq K_0$,

$$\begin{aligned} L(N_k) &= \int_{\mathcal{X}\times\mathcal{X}} \|N_k(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \,\mathrm{d}\mu(\bar{x},\bar{u}) \\ &\geq \int_{\{\|\bar{x}\|_{\mathcal{X}}+\|\bar{u}\|_{\mathcal{X}}\leq K\}} \|N_k(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \,\mathrm{d}\mu(\bar{x},\bar{u}) \\ &\geq \frac{1}{4}\|N_k(0)\|_{\mathcal{X}}^2 \int_{\{\|\bar{x}\|_{\mathcal{X}}+\|\bar{u}\|_{\mathcal{X}}\leq K\}} \mathrm{d}\mu(\bar{x},\bar{u}), \end{aligned}$$

where in the last inequality we used (4.3.5). Since the last integral is strictly positive by hypothesis, we obtain the desired result. It is left to prove that L is weak* lower

semicontinuous in the sense of (4.3.1). Let $(N_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{N}$ and $N \in \mathcal{N}$ such that $N_k \overset{*}{\rightharpoonup} N$ as $k \to \infty$. Observe that, by Fatou's Lemma,

$$\begin{aligned}\liminf_{k\to\infty} L(N_k) &= \liminf_{k\to\infty} \int_{\mathcal{X}\times\mathcal{X}} \|N_k(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \, \mathrm{d}\mu(\bar{x},\bar{u}) \\ &\geq \int_{\mathcal{X}\times\mathcal{X}} \liminf_{k\to\infty} \|N_k(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \, \mathrm{d}\mu(\bar{x},\bar{u}) \\ &\geq \int_{\mathcal{X}\times\mathcal{X}} \|N(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \, \mathrm{d}\mu(\bar{x},\bar{u}) = L(N),\end{aligned}$$

where we have used the fact that the squared norm is lower semicontinuous w.r.t. the weak topology on $\mathcal{X}$ and, by Corollary 4.2, that $N_k(\bar{x}) - \bar{u} \rightharpoonup N(\bar{x}) - \bar{u}$ as $k \to \infty$ for every $\bar{x}, \bar{u} \in \mathcal{X}$. □

In practice, the minimizer of (CP) cannot be computed since, as we mentioned, the probability measure μ is unknown. Instead, we suppose to have access to a finite set $\{(\bar{X}_i, \bar{U}_i)\}_{i=1}^n$ of identical and independent copies of the pair $(\bar{X}, \bar{U})$, and we aim at finding minimizers of the empirical risk (2.1.3) with respect to the square loss, which for this problem writes as

$$N_n^* \in \operatorname*{arg\,min}_{N\in\mathcal{N}} L_n(N), \quad L_n(N) := \frac{1}{n}\sum_{i=1}^n \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2 \tag{EP}$$

in the space of nonexpansive operators. Note the slight change of notation with respect to (2.1.3). This will be later justified during the statement of Theorem 4.10. Now, we aim to show that problem (EP) is a good approximation of (CP) for n large enough. As we have explored in Section 2.1, this is a standard question Supervised Learning theory [54]. A corresponding version of Theorem 2.2 in this setting would show that

$$L(N_n^*) - L(N^*) \sim \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

i.e. that the the *excess risk* goes to zero as $1/\sqrt{n}$. Note that, on top of showing that the error when considering N_n^* as an approximant of N^* converges to 0, it also gives a precise rate of convergence. As it can be seen in Theorem 2.2, it is common to assume that the underlying space is compact. However, since the space $\mathcal{N}$ is not compact in general, such an assumption can not be satisfied in our setting. We therefore avoid this approach. Instead, we present in the following theorem a rather more qualitative result, by showing that problem (EP) Γ-converges (see [20]) to (CP) when n goes to infinity.

Theorem 4.10. *We have that L_n Γ-converges a.s. to the expected risk L as $n \to \infty$; i.e., both of the following conditions are satisfied:*

(i) (“$\liminf$” inequality) *For every $N \in \mathcal{N}$ and for every sequence $(N_n)_{n\in\mathbb{N}}$ in $\mathcal{N}$ with $N_n \overset{*}{\rightharpoonup} N$ in $\mathrm{Lip}(X)$, we have that*

$$L(N) \leq \liminf_{n\to\infty} L_n(N_n);$$

(ii) (existence of a recovery sequence) *for every $N \in \mathcal{N}$, there exists a sequence $(N_n)_{n\in\mathbb{N}}$ in $\mathcal{N}$ with $N_n \overset{*}{\rightharpoonup} N$ in $\mathrm{Lip}(X)$ such that*

$$\limsup_{n\to\infty} L_n(N_n) \leq L(N).$$

Proof. We start first with the existence of a recovery sequence. Let N be a nonexpansive operator and define the random variable

$$\bar{G} := \|N(\bar{X}) - \bar{U}\|_{\mathcal{X}}^2$$

and observe that, since

$$\begin{aligned}\mathbb{E}[|\bar{G}|] &= \int_{\mathcal{X}\times\mathcal{X}} \|N(\bar{x}) - \bar{u}\|_{\mathcal{X}}^2 \, \mathrm{d}\mu(\bar{x},\bar{u}) \\ &\leq 2\int_{\mathcal{X}\times\mathcal{X}} (\|N(\bar{x}) - N(0)\|_{\mathcal{X}}^2 + \|N(0) - \bar{u}\|_{\mathcal{X}}^2) \, \mathrm{d}\mu(\bar{x},\bar{u}) \\ &\leq 2\int_{\mathcal{X}\times\mathcal{X}} (\|\bar{x}\|_{\mathcal{X}}^2 + \|N(0) - \bar{u}\|_{\mathcal{X}}^2) \, \mathrm{d}\mu(\bar{x},\bar{u}),\end{aligned}$$

there exists a constant $C > 0$ such that

$$\mathbb{E}[|\bar{G}|] \leq C\int_{\mathcal{X}\times\mathcal{X}} \left(\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2 + 1\right) \, \mathrm{d}\mu(\bar{x},\bar{u}) < \infty.$$

Now, define for every $i = 1, ..., n$,

$$\bar{G}_i := \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2$$

which are also random variables that are independent and identically distributed as G. By the law of large numbers [68, Theorem 10.13],

$$L_n(N) = \frac{1}{n}\left(\bar{G}_1 + ... + \bar{G}_n\right) \to \mathbb{E}[\bar{G}] = L(N) \text{ as } n \to \infty \text{ a.s..} \tag{4.3.6}$$

Hence,

$$\limsup_{n\to\infty} L_n(N_n) \leq L(N) \quad \text{a.s.},$$

for $N_n = N$ for every $n \in \mathbb{N}$.

We want now to prove the "$\liminf$" inequality. Let $N_n : \mathcal{X} \to \mathcal{X}$, $n \in \mathbb{N}$, be a sequence of nonexpansive operators such that $N_n \xrightarrow{*} N$ for some $N \in \mathcal{N}$. Let $\varepsilon > 0$ and observe that, by Ulam's tightness Theorem [17], the measure μ, being a probability measure, is tight. Then, since it is finite, the measure $\mu' = (\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2)\mu$ is tight too by hypothesis. Therefore, we can choose a compact set $K \subset \mathcal{X} \times \mathcal{X}$ such that

$$\int_{\mathcal{X}\times\mathcal{X}\setminus K} (\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2) \, \mathrm{d}\mu(\bar{x},\bar{u}) < \varepsilon.$$

Moreover, since $L_n(N_n) = L_n(N_n) - L_n(N) + L_n(N)$, we obtain

$$L_n(N_n) - L_n(N) = \frac{1}{n}\sum_{i=1}^{n} \left(\|N_n(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2 - \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2\right),$$

where

$$\begin{aligned}&\|N(\bar{X}_i) - \bar{U}_i + N_n(\bar{X}_i) - N(\bar{X}_i)\|_{\mathcal{X}}^2 - \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2 \\ &\qquad = 2\langle N(\bar{X}_i) - \bar{U}_i, N_n(\bar{X}_i) - N(\bar{X}_i)\rangle_{\mathcal{X}} + \|N_n(\bar{X}_i) - N(\bar{X}_i)\|_{\mathcal{X}}^2 \\ &\qquad \geq 2\langle N(\bar{X}_i) - \bar{U}_i, N_n(\bar{X}_i) - N(\bar{X}_i)\rangle_{\mathcal{X}}.\end{aligned}$$

Observe that there exists a constant $C > 0$ such that $\|N_n(0)\|_{\mathcal{X}} \leq C$. Now, we define, for every $i = 1, ..., n$,

$$\bar{G}_i' := \varepsilon\chi_K(\bar{X}_i, \bar{U}_i), \quad \bar{H}_i := \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}} 2(C + \|\bar{X}_i\|_{\mathcal{X}})\chi_{\mathcal{X}\setminus K}(\bar{X}_i, \bar{U}_i),$$

where χ_K denotes the *characteristic function* of K, defined as

$$\chi_K(x) := \begin{cases} 1, & \text{if } x \in K, \\ 0, & \text{otherwise.} \end{cases}$$

Then, we claim that, for n large enough,

$$\langle N(\bar{X}_i) - \bar{U}_i, N_n(\bar{X}_i) - N(\bar{X}_i)\rangle_{\mathcal{X}} \geq -\bar{G}'_i - \bar{H}_i.$$

On one hand, observe that, for every $i = 1, ..., n$,

$$\begin{aligned} -\langle N(\bar{X}_i) - \bar{U}_i, &N_n(\bar{X}_i) - N(\bar{X}_i)\rangle_{\mathcal{X}} \chi_K(\bar{X}_i, \bar{U}_i) \\ &\leq |\langle N(\bar{X}_i) - \bar{U}_i, N_n(\bar{X}_i) - N_n(0) - N(\bar{X}_i) + N(0)\rangle_{\mathcal{X}}| \chi_K(\bar{X}_i, \bar{U}_i) \\ &\quad + |\langle N(\bar{X}_i) - \bar{U}_i, N_n(0) - N(0)\rangle_{\mathcal{X}}| \chi_K(\bar{X}_i, \bar{U}_i) \\ &\leq \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}} \|N_n(\bar{X}_i) - N_n(0) - N(\bar{X}_i) + N(0)\|_{\mathcal{X}} \chi_K(\bar{X}_i, \bar{U}_i) \\ &\quad + |\langle N(\bar{X}_i) - \bar{U}_i, N_n(0) - N(0)\rangle_{\mathcal{X}}| \chi_K(\bar{X}_i, \bar{U}_i). \end{aligned}$$

First, since K is compact, it is bounded. Then, there exists a ball $B_r(0)$ centered at 0 with radius $r > 0$ such that $K \subset B_r(0)$. This implies that $\|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}} \leq 2r$ as long as $(\bar{X}_i, \bar{U}_i) \in K$. In addition, by Corollary 4.3 (i), for $(\bar{X}_i, \bar{U}_i) \in K$ and for n large enough we get that

$$\|N_n(\bar{X}_i) - N_n(0) - N(\bar{X}_i) + N(0)\|_{\mathcal{X}} \leq \frac{\varepsilon}{4r}.$$

Finally, by Corollary 4.3 (ii), and for n large enough, we obtain

$$|\langle N(\bar{X}_i) - \bar{U}_i, N_n(0) - N(0)\rangle_{\mathcal{X}}| < \frac{\varepsilon}{2},$$

as long as $(\bar{X}_i, \bar{U}_i) \in K$. Combining everything, we get

$$\langle N(\bar{X}_i) - \bar{U}_i, N_n(\bar{X}_i) - N(\bar{X}_i)\rangle_{\mathcal{X}} \chi_K(\bar{X}_i, \bar{U}_i) \geq -\bar{G}'_i.$$

On the other hand,

$$\begin{aligned} -\langle N(\bar{X}_i) - \bar{U}_i, &N_n(\bar{X}_i) - N(\bar{X}_i)\rangle_{\mathcal{X}} \chi_{\mathcal{X}\setminus K}(\bar{X}_i, \bar{U}_i) \\ &\leq \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}} \|N_n(\bar{X}_i) - N(\bar{X}_i)\|_{\mathcal{X}} \chi_{\mathcal{X}\setminus K}(\bar{X}_i, \bar{U}_i) \\ &\leq \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}} (\|N_n(\bar{X}_i) - N_n(0)\|_{\mathcal{X}} + \|N(0) - N(\bar{X}_i)\|_{\mathcal{X}} \\ &\quad + \|N_n(0) - N(0)\|_{\mathcal{X}}) \chi_{\mathcal{X}\setminus K}(\bar{X}_i, \bar{U}_i) \\ &\leq \|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}} 2(C + \|\bar{X}_i\|_{\mathcal{X}}) \chi_{\mathcal{X}\setminus K}(\bar{X}_i, \bar{U}_i) = \bar{H}_i, \end{aligned}$$

and the claim has been proven. Now, note that $\bar{G}' = \varepsilon \chi_K(\bar{X}, \bar{u})$ is a random variable with

$$\mathbb{E}[|\bar{G}'|] = \varepsilon \int_{\mathcal{X}\times\mathcal{X}} |\chi_K(\bar{x}, \bar{u})| \, \mathrm{d}\mu(\bar{x}, \bar{u}) = \varepsilon \mu(K) \leq \varepsilon,$$

and $\bar{H} = \|N(\bar{X}) - \bar{U}\|_{\mathcal{X}} 2(C + \|\bar{X}\|_{\mathcal{X}})(1 - \chi_K(\bar{X}, \bar{U}))$ is a random variable with

$$\begin{aligned} \mathbb{E}[|\bar{H}|] &= 2 \int_{\mathcal{X}\times\mathcal{X}\setminus K} \|N(\bar{x}) - \bar{u}\|_{\mathcal{X}} (C + \|\bar{x}\|_{\mathcal{X}}) \, \mathrm{d}\mu(\bar{x}, \bar{u}) \\ &\leq 2 \int_{\mathcal{X}\times\mathcal{X}\setminus K} (\|\bar{x}\|_{\mathcal{X}} + C + \|\bar{u}\|_{\mathcal{X}})(C + \|\bar{x}\|_{\mathcal{X}}) \, \mathrm{d}\mu(\bar{x}, \bar{u}) \\ &\leq C' \int_{\mathcal{X}\times\mathcal{X}\setminus K} (\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2 + 1) \, \mathrm{d}\mu(\bar{x}, \bar{u}) < C'\varepsilon \end{aligned}$$

for a suitable constant $C' > 0$. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{G}'_i + \bar{H}_i) \to \mathbb{E}[\bar{G}] + \mathbb{E}[\bar{H}] \text{ as } n \to \infty \text{ a.s.},$$

which implies that, for n large enough, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{G}'_i + \bar{H}_i) \leq \varepsilon + \mathbb{E}[\bar{G}] + \mathbb{E}[\bar{H}] \leq (2 + C')\varepsilon, \quad \text{a.s..}$$

Now, for n large enough,

$$L_n(N_n) - L_n(N) \geq -\frac{2}{n}\sum_{i=1}^{n}(\bar{G}'_i + \bar{H}_i) \geq -2(2 + C')\varepsilon, \quad \text{a.s..}$$

On the other hand, recall that, by (4.3.6), $L_n(N) \to F(N)$ a.s. as $n \to \infty$. Combining this with the inequality above, we derive that

$$\liminf_{n\to\infty} L_n(N_n) \geq L(N) - 2(2 + C')\varepsilon.$$

Since ε was arbitrary, we conclude. □

The Fundamental Theorem of $\Gamma-$convergence [20, Theorem 2.1] states that, if $(L_n)_{n\in\mathbb{N}}$ is an equicoercive sequence of functions that is $\Gamma-$converging to L, then, up to subsequences, the sequence of minimizers $(N_n^*)_{n\in\mathbb{N}}$ of (EP) converges a.s. to a minimizer of the continuous problem (CP). In our case, it is only left to prove that the sequence $(L_n)_{n\in\mathbb{N}}$ is equicoercive, as we do in the following.

Proposition 4.11. *The sequence $(L_n)_{n\in\mathbb{N}}$, defined as in (EP), is equicoercive a.s. on $\mathcal{N}$ with respect to the weak* convergence.*

Proof. First, observe that

$$\begin{aligned}
\|N(\bar{X}_i) - \bar{U}_i\|_{\mathcal{X}}^2 &= \|N(\bar{X}_i) - N(0) - \bar{U}_i\|_{\mathcal{X}}^2 + \|N(0)\|_{\mathcal{X}}^2 + 2\langle N(\bar{X}_i) - N(0) - \bar{U}_i, N(0)\rangle_{\mathcal{X}} \\
&\geq \|N(\bar{X}_i) - N(0) - \bar{U}_i\|_{\mathcal{X}}^2 + \|N(0)\|_{\mathcal{X}}^2 - 2\|N(\bar{X}_i) - N(0) - \bar{U}_i\|_{\mathcal{X}}\|N(0)\|_{\mathcal{X}} \\
&\geq \frac{1}{2}\|N(0)\|_{\mathcal{X}}^2 - \|N(\bar{X}_i) - N(0) - \bar{U}_i\|_{\mathcal{X}}^2 \\
&\geq \frac{1}{2}\|N(0)\|_{\mathcal{X}}^2 - 2\left(\|\bar{X}_i\|_{\mathcal{X}}^2 + \|\bar{U}_i\|_{\mathcal{X}}^2\right)
\end{aligned}$$

since, for every a, $b \in \mathbb{R}$, it holds that $ab \leq a^2/4 + b^2$, and this immediately implies that $-2ab \geq -a^2/2 - 2b^2$. Then,

$$L_n(N) \geq \frac{1}{2}\|N(0)\|_{\mathcal{X}}^2 - \frac{2}{n}\sum_{i=1}^{n}\left(\|\bar{X}_i\|_{\mathcal{X}}^2 + \|\bar{U}_i\|_{\mathcal{X}}^2\right)$$

where, by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}(\|\bar{X}_i\|_{\mathcal{X}}^2 + \|\bar{U}_i\|_{\mathcal{X}}^2) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}\times\mathcal{X}}\left(\|\bar{x}\|_{\mathcal{X}}^2 + \|\bar{u}\|_{\mathcal{X}}^2\right)\mathrm{d}\mu(\bar{x},\bar{u}), \quad \text{a.s..}$$

Since it is a converging sequence, it is also bounded by some constant $C > 0$, which leads to

$$L_n(N) \geq c\|N(0)\|_{\mathcal{X}}^2 - C, \text{ a.s.},$$

for some $c > 0$. Hence, for every $C' > 0$ we have that

$$\{L_n(N) \leq C'\} \subseteq \{N \in \mathcal{N} : \|N(0)\|_{\mathcal{X}} \leq C''\}$$

for some $C'' > 0$ (in fact, $C'' = ((C' + C)/c)^{1/2}$) and where the set on the right-hand side is weak* compact in $\mathrm{Lip}(\mathcal{X})$. Indeed, by [101, Corollary 2.6.19], we only need to show that the set above is bounded in $\mathrm{Lip}(X)$:

$$\|N\|_{\mathrm{Lip}} = \|N(0)\|_{\mathcal{X}} + \|N - N(0)\|_{\mathrm{Lip}_0} \leq C'' + 1.$$

This concludes the proof. □

As we already mentioned above, the following corollary is a direct consequence of both Theorem 4.10 and Theorem 4.11. Its proof can be found in [20] in a more general setting.

Corollary 4.12. *Let $(N_n^*)_{n\in\mathbb{N}}$, be the sequence of minimizers of (EP) for every $n \in \mathbb{N}$. Then, there exists a minimizer N^* of (CP) such that, up to subsequences, $N_n^* \overset{*}{\rightharpoonup} N^*$, a.s., as $n \to \infty$.*

The above results provide a theoretical analysis that highlights the fact that the ERM (EP) is indeed approximating its continuous version (CP) for large values of n.

In the next section, we will introduce and analyze a practical approach to learn nonexpansive operators. To do so, we shift to the deterministic setting and fix a training set of noisy measurements/solutions $\{(\bar{x}_i, \bar{u}_i)\}_{i=1}^n$, $n \in \mathbb{N}$, instead of a set of random variables. We consider, from now on, the following problem

$$\widehat{N} \in \operatorname*{arg\,min}_{N\in\mathcal{N}} \widehat{L}(N), \quad \widehat{L}(N) := \frac{1}{n}\sum_{i=1}^{n} \|N(\bar{x}_i) - \bar{u}_i\|_{\mathcal{X}}^2 \tag{DP}$$

as a natural substitute of (EP) in the deterministic case. In order to solve this, classical optimization algorithms can be considered, but the class of nonexpansive operators is infinite-dimensional, and so, Problem (DP) remains computationally unfeasible. Therefore, a further approximation needs to be considered. To do so, a first idea could be to consider Problem (DP) restricted to the set $\mathcal{N}(\bar{D})$ of nonexpansive operators on $\bar{D} := \{\bar{x}_i\}_{i=1}^n$, i.e., operators $N : \bar{D} \to \mathcal{X}$ such that $\|N(\bar{x}_i) - N(\bar{x}_j)\|_{\mathcal{X}} \leq \|\bar{x}_i - \bar{x}_j\|_{\mathcal{X}}$ for $i, j = 1, \ldots, n$. Actually, every N in $\mathcal{N}(\bar{D})$ is uniquely characterized by n values $u_i := N(\bar{x}_i)$ that are required to satisfy the nonexpansivity condition. The corresponding reformulation of this discrete version of Problem (DP) would be

$$\begin{aligned} \min_{u_1,\ldots,u_n\in\mathcal{X}} \ & \frac{1}{n}\sum_{i=1}^{n} \|u_i - \bar{u}_i\|_{\mathcal{X}}^2 \\ \text{s.t. } & \|u_i - u_j\|_{\mathcal{X}} \leq \|\bar{x}_i - \bar{x}_j\|_{\mathcal{X}}, \text{ for every } i, j \in \{1, \ldots, n\}. \end{aligned} \tag{4.3.7}$$

This leaves us with the following challenge: given a solution of (4.3.7), how can we extend it to the whole space? In fact, in practical scenarios, we need to know the value of such an operator at any point in $\mathcal{X}$. For instance, we know that a nonexpansive extension to the whole space always exists [95], but it is difficult to construct it in practice. For this reason, our proposed approach will be to consider, on the one hand, finite-dimensional spaces $\mathcal{X}$; i.e. $\mathcal{X} = \mathbb{R}^d$, $d \geq 1$ and, on the other hand, operators N that are, in addition, piecewise affine in d-simplices (or just simplices). In this way, we will show that problem (DP) can be reduced to finding N in the finitely many points $\bar{D}$. This idea will be explained in the following.

In the following, we develop the main tool that we consider in order to discretize the set $\mathcal{N}$: simplicial partitions.

4.3.3 Simplicial partitions

Let $\mathcal{X} = \mathbb{R}^d$, $2 \leq d < \infty$, endowed with the norm $\|\cdot\|_{\mathcal{X}} := \|\cdot\|_2$, and let $D := \{x_1, \ldots, x_m\}$, with $x_i \in \mathbb{R}^d$, $d+1 \leq m$ be a general finite set of points, independent of the training inputs above considered, that do not lie on a $(d-1)-$dimensional hyperplane. Consider its convex envelope $\mathrm{conv}(D)$, which has non-empty interior. We want to consider a simplicial partition of $\mathrm{conv}(D)$. Let $\ell \in \mathbb{N}$ and

$$\mathfrak{T} = \{S_1, ..., S_\ell\},$$

where for every $t = 1, ..., \ell$, $S_t = \mathrm{conv}\{x_{t,0}, \ldots, x_{t,d}\}$, for some subcollection of $d+1$ distinct elements of D. We assume $\mathfrak{T}$ to have the following properties:

(P1) $\mathfrak{T}$ forms a partition of $\mathrm{conv}(D)$;

(P2) the interior of every simplex S_t is non-empty;

(P3) the intersection of every two simplices has to be, either empty or coincide with the convex envelope of its common vertices.

A partition defined by these conditions is also known as *face-to-face simplicial partition* for a polytope [78], where *face* stands for the convex hull of any collection of "d" vertices. For example, in dimension $d = 2$, *face* denotes the edge of a triangle, while in dimension $d = 3$, the *faces* of a tetrahedron are triangles. An example of a classical partition satisfying all of the above conditions is the so-called *Delaunay Triangulation* [69]. Nevertheless, we are not interested in fixing a particular method, but only one satisfying the above conditions.

4.3.4 Piecewise affine nonexpansive operators

In this section, we want to construct a finite-dimensional set of operators which will be characterized only by their value on each node x_i, $i = 1, \ldots, m$. To do so, we first introduce an arbitrary set of points $D' = \{u_1, \ldots, u_m\}$, independent of D. Such elements will determine the value of the constructed nonexpansive operator at every point in D, i.e. the nonexpansive operator N will be uniquely defined by $N(x_i) := u_i$. In addition, we want to construct N in such a way that it is also nonexpansive. To do so, we first construct an operator N that will be affine on every simplex S_t and, second, find a suitable condition for the vertices so that the resulting operator is nonexpansive. Let us remark that, once the nonexpansive operator N is well defined, in order to then construct a firmly nonexpansive operator T, one just needs to recall the formula $T = \frac{1}{2}(\mathrm{Id} + N)$, since the samples $(\bar{x}_1, \bar{u}_1), \ldots, (\bar{x}_n, \bar{u}_n)$ were drawn from $(\bar{X}, \bar{U})$, and $\bar{U} = 2\bar{Z} - \bar{X}$, see Section 4.3.2.

Consider the convex envelope of the set D, $\mathrm{conv}(D)$, and let $\mathfrak{T} = \{S_1, ..., S_\ell\}$, $\ell \in \mathbb{N}$, denote a simplicial partition of $\mathrm{conv}(D)$ with properties (P1), (P2) and (P3) given in the section above. Consider

$$\lambda_1, \ldots, \lambda_m \colon \mathrm{conv}(D) \to [0, 1]$$

the Lagrange elements of order 1 associated with the simplicial partition $\mathfrak{T}$ (see [109, Chapter 4] for more details); i.e., such that

(i) $\lambda_i(x_j) = \delta_{ij}$, for $i,\ j = 1, \ldots, m$ (here, δ_{ij} denotes the Kronecker delta),

(ii) $\lambda_i|_{S_t}$ is a polynomial of degree at most 1 for each $i = 1, \dots, m$ and $t = 1, \dots, \ell$.

Note that each λ_i is continuous in $\mathrm{conv}(D)$. In addition, if $i_0, \dots, i_d$ denote the indices of the vertices $x_{i_0}, \dots, x_{i_d}$ of the simplex S_t, then $\lambda_{i_0}|_{S_t}, \dots, \lambda_{i_d}|_{S_t}$ correspond to the barycentric coordinates on S_t; i.e. the unique non-negative functions that satisfy

$$\sum_{j=0}^{d} \lambda_{i_j}(x) = 1, \quad \sum_{j=0}^{d} \lambda_{i_j}(x) x_{i_j} = x, \tag{4.3.8}$$

for each $x \in S_t$. Further, as $\lambda_i|_{S_t} = 0$ for each $i \in \{1, \dots, m\} \setminus \{i_0, \dots, i_d\}$, we have that $\sum_{i=1}^{m} \lambda_i(x) = 1$ for any $x \in \mathrm{conv}(D)$, which immediately implies that

$$\sum_{i=1}^{m} \lambda_i = \chi_{\mathrm{conv}(D)}, \quad \text{and } \sum_{i=1}^{m} \lambda_i x_i = \mathrm{Id}\,|_{\mathrm{conv}(D)}.$$

Next, given $D' = \{u_1, ..., u_m\}$, we define the operator

$$\widetilde{N} \colon \mathrm{conv}(D) \to \mathbb{R}^d; \quad \widetilde{N}(x) := \sum_{i=1}^{m} \lambda_i(x) u_i. \tag{4.3.9}$$

Finally, for every simplex S_t, $t = 1, \dots, \ell$, with vertices $i_0, \dots, i_d$, we have that

$$\widetilde{N}|_{S_t} = \sum_{j=0}^{d} \lambda_{i_j}|_{S_t} u_{i_j},$$

which, since each $\lambda_{i_j}|_{S_t}$ is affine linear, implies that $\widetilde{N}|_{S_t}$ is also affine linear.

In order to extend such an operator to the whole space, we consider the following

$$N := \tilde{N} \circ \pi_{\mathrm{conv}(D)}$$

where $\pi_{\mathrm{conv}(D)}$ stands for the projection onto $\mathrm{conv}(D)$. Given this construction, we can define the space of piecewise affine operators on $\mathrm{conv}(D)$, and therefore on its simplicial partition $\mathfrak{T}$, that are extended to the whole space via projections:

$$\mathrm{PA}(\mathfrak{T}) := \left\{ N : \mathbb{R}^d \to \mathbb{R}^d \,:\, N := \tilde{N} \circ \pi_{\mathrm{conv}(D)} \right\},$$

where $\tilde{N}$ has been defined above in every simplex of $\mathrm{conv}(D)$. We are now ready to introduce the finite-dimensional space of nonexpansive operators $\mathcal{N} \cap \mathrm{PA}(\mathfrak{T})$. We focus on solving the deterministic problem (DP) in this class:

$$\min_{N \in \mathcal{N} \cap \mathrm{PA}(\mathfrak{T})} \widehat{L}(N). \tag{PAP}$$

Observe that the function $\widehat{L}$ is only defined on the set $\bar{D}$, but the constraint does not necessarily need to be imposed only on $\bar{D}$. We therefore assume that $\mathrm{PA}(\mathfrak{T})$ is defined on a bigger set D and that $\bar{D} \subseteq D$. Moreover, notice that the problem above is finite-dimensional since, as we already mentioned, an element N in $\mathrm{PA}(\mathfrak{T})$ is uniquely characterized by the set D. Our objective now is to study whether we can propose an equivalent formulation of the above problem by imposing conditions on the finitely many points $(x_1, N(x_1)), \dots (x_m, N(x_m))$, as we mentioned at the beginning of the section. A first attempt could be to impose the nonexpansivity condition on these points i.e.,

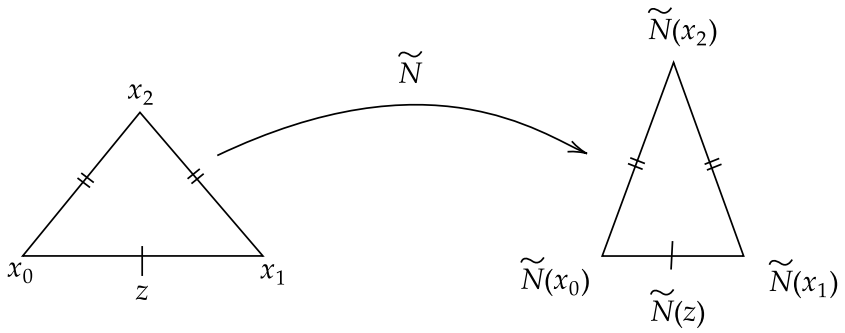

Figure 4.1: In dimension $d = 2$, we define the nonexpansive operator as described in the picture and such that $\|\tilde{N}(x_0) - \tilde{N}(x_2)\|_2 = \|x_0 - x_2\|_2$, $\|\tilde{N}(x_1) - \tilde{N}(x_2)\|_2 = \|x_1 - x_2\|_2$ and $\|\tilde{N}(x_0) - \tilde{N}(x_1)\|_2 < \|x_0 - x_1\|_2$. If we consider the point $z = \frac{1}{2}x_0 + \frac{1}{2}x_1$, by construction, $\|\tilde{N}(z) - \tilde{N}(x_2)\|_2 > \|z - x_2\|_2$.

$\|N(x_i) - N(x_j)\|_2 \leq \|x_i - x_j\|_2$, for every $i, j = 1, ..., m$. In Figure 4.3.4, we show a counterexample in dimension $d = 2$. The defined operator satisfies the nonexpansivity condition $\|N(x_i) - N(x_j)\|_2 \leq \|x_i - x_j\|_2$, for every $i, j = 1, ..., m$, over all the vertices, but fails to be nonexpansive in general.

In order to find the right condition, we first observe that the operator $\tilde{N}$ can be rewritten in a more convenient way. First, let us introduce some useful tools. For every simplex $S_t \in \mathfrak{T}$, denote by $i_0, \ldots, i_d$ the indices associated to the vertices $x_{i_0}, \ldots, x_{i_d}$ in S_t. Define the matrices

$$A_t := [x_{i_1} - x_{i_0} \mid \cdots \mid x_{i_d} - x_{i_0}], \quad B_t = [u_{i_1} - u_{i_0} \mid \ldots \mid u_{i_d} - u_{i_0}]. \tag{4.3.10}$$

Moreover, for every $x \in \mathrm{conv}(D)$, there exists $t = 1, \ldots, \ell$ such that $x \in S_t$. Then, since the barycentric coordinates sum up to 1, we can write

$$x = \sum_{j=0}^{d} \lambda_{i_j}(x) x_{i_j} = \sum_{j=1}^{d} \lambda_{i_j}(x)(x_{i_j} - x_{i_0}) + x_{i_0}.$$

Then,

$$x - x_{i_0} = \sum_{j=1}^{d} \lambda_{i_j}(x)(x_{i_j} - x_{i_0}) = A_t[\lambda_{i_1}(x), \ldots, \lambda_{i_d}(x)]^T.$$

By defining $\Lambda_t(x) := [\lambda_{i_1}(x), \ldots, \lambda_{i_d}(x)]^T$, we derive that

$$\Lambda_t(x) = A_t^{-1}(x - x_{i_0}).$$

Hence, we obtain the following affine expression for the operator $\tilde{N}$ in every simplex:

$$\tilde{N}|_{S_t}(x) = \sum_{j=1}^{d} \lambda_{i_j}(u_{i_j} - u_{i_0}) + u_{i_0} = B_t \Lambda_t(x) + u_{i_0} = B_t A_t^{-1}(x - x_{i_0}) + u_{i_0}. \tag{4.3.11}$$

We are now ready to present a sufficient condition for $\tilde{N}$ (hence for N) to be nonexpansive. Actually, we are able to characterize the elements in $\mathcal{N} \cap \mathrm{PA}(\mathfrak{T})$. To do so, we first recall a preliminary result about the differentiability of Lipschitz functions, see [83] for a proof.

Proposition 4.13. *Let N be a nonexpansive operator. Then, it is differentiable almost everywhere, and, for every differentiability point x, it holds that $\|\nabla N(x)\|_2 \leq 1$.*

Recall that $\nabla N(x)$ denotes the Jacobian matrix of N at the point x, and that the operator norm of a matrix is defined as its 2- norm; i.e., its maximum singular value. The nonexpansivity characterization reads as follows.

Theorem 4.14. *The operator $N = \tilde{N} \circ \pi_{\mathrm{conv}(D)}$ is nonexpansive if and only if, for every $t = 1, ..., \ell$ we have $\|B_t A_t^{-1}\|_2 \leq 1$.*

Proof. Let $N = \tilde{N} \circ \pi_{\mathrm{conv}(D)}$ be a nonexpansive operator. By Proposition 4.13 $\|\nabla N(x)\|_2 \leq 1$. In particular, for every $t = 1, \ldots, \ell$ we can take a point x in the interior of S_t and observe that the derivative of N at x coincides with the derivative of $\tilde{N}$ at x, which is $\nabla \tilde{N}(x) = B_t A_t^{-1}$ by the characterization (4.3.11). Conversely, observe that projections are nonexpansive operators. Therefore, if we prove that $\tilde{N}$ is nonexpansive, then the operator N, as we have defined it, will also be nonexpansive, since the composition of two nonexpansive operators is, again, nonexpansive. Let x, $x' \in \mathrm{conv}(D)$ and suppose first that both points belong to the same simplex S_t. We consider the path $\gamma : [0,1] \to \mathbb{R}^d$; $\gamma(s) = (1-s)x + sx'$ and observe that the function $\tilde{N} \circ \gamma$ is differentiable since it is the composition of a segment and $\tilde{N}$, which is an affine function in every simplex. Therefore, by the fundamental theorem of calculus, we get that

$$\begin{aligned}
\|\tilde{N}(x') - \tilde{N}(x)\|_2 = \|(\tilde{N} \circ \gamma)(1) - (\tilde{N} \circ \gamma)(0)\|_2 &= \left\| \int_0^1 (\tilde{N} \circ \gamma)'(s)\, \mathrm{d}s \right\|_2 \\
&\leq \int_0^1 \|(\tilde{N} \circ \gamma)'(s)\|_2 \, \mathrm{d}s = \int_0^1 \|\nabla \tilde{N}(\gamma(s))\gamma'(s)\|_2 \, \mathrm{d}s \\
&\leq \int_0^1 \|B_t A_t^{-1}\|_2 \|x' - x\|_2 \, \mathrm{d}s \leq \|x' - x\|_2,
\end{aligned}$$

where we used that $\nabla \tilde{N}(x) = B_t A_t^{-1}$ for every $x \in \mathbb{R}^d$ since (4.3.11) holds. Now, if $x \in S_t$, $x' \in S_{t'}$ for $t \neq t'$, we consider again the segment connecting both points

$$\gamma : [0,1] \to \mathbb{R}^d; \quad \gamma(s) = (1-s)x + sx',$$

and observe that the whole line is contained in $\mathrm{conv}(D)$ since, by definition, $\mathfrak{T}$ partitions the convex hull $\mathrm{conv}(D)$. Moreover, the segment intersects a finite number of distinct simplices. We consider the collection of the entry and exit points of the segment to these simplices. In this way, we obtain a set of $k+1$ points as follows: $p_0 = x$; p_i, $i = 1, ..., k-1$, are the entry or exit points of the segment intersecting a simplex (as we explained above), and finally, $p_k = x'$. To each of these points, we associate $k+1$ numbers $(s_i)_{i=0}^k$ with $\gamma(s_i) = x_h$, $i = 0, \ldots, k$. Now, for every $i = 1, ..., k$, we define the subsegment $\gamma_h : [s_{i-1}, s_i] \to \mathbb{R}^d$ from p_{i-1} to p_i. By construction, we have that each one of these segments lies entirely in a simplex that we denote by S_{t_i} and that $\mathrm{length}(\gamma) = \sum_{i=1}^k \mathrm{length}(\gamma_i)$. Noticing that $\tilde{N} \circ \gamma_i$ is differentiable for every $i = 0, \ldots, k$, we obtain

$$\begin{aligned}
\|\tilde{N}(x') - \tilde{N}(x)\|_2 \leq \sum_{i=1}^k \left\|\tilde{N}(p_i) - \tilde{N}(p_{i-1})\right\|_2 &= \sum_{h=1}^k \left\| \int_{s_{i-1}}^{s_i} (\tilde{N} \circ \gamma_i)'(s) \mathrm{d}s \right\|_2 \\
&\leq \sum_{i=1}^k \int_{s_{i-1}}^{s_i} \|(\tilde{N} \circ \gamma_i)'(s)\|_2 \, \mathrm{d}s = \sum_{i=1}^k \int_{s_{i-1}}^{s_i} \|\nabla \tilde{N}(\gamma_i(s))\gamma_i'(s)\|_2 \, \mathrm{d}s \\
&\leq \sum_{i=1}^k \int_{s_{i-1}}^{s_i} \|B_{t_i} A_{t_i}^{-1}\|_2 \|p_i - p_{i-1}\|_2 \, \mathrm{d}s \leq \sum_{h=1}^k \int_{s_{i-1}}^{s_i} \|p_i - p_{i-1}\|_2 \, \mathrm{d}s \\
&= \sum_{i=1}^k \mathrm{length}(\gamma_i) = \mathrm{length}(\gamma) = \|x - x'\|_2.
\end{aligned}$$

□

Now that we have found the right condition. We now aim at tackling (PAP) from a computational point of view. First, for simplicity we assume that $x_i = \bar{x}_i$ for $i = 1, \dots, n$ and then consider the following problem:

$$\begin{aligned} \min_{u_1,\dots,u_m \in \mathbb{R}^d} & \frac{1}{n} \sum_{i=1}^{n} \|u_i - \bar{u}_i\|_2^2 \\ \text{s.t. } & \|B_t(u_1, ..., u_m) A_t^{-1}\|_2 \leq 1, \text{ for every } t \in \{1, \dots, \ell\}, \end{aligned} \tag{FP}$$

where the matrices A_t, $B_t(u_1, ..., u_m)$, $t = 1, ..., \ell$, are defined as in (4.3.10). Note that here we introduce the notation $B_t(u_1, ..., u_m)$ in order to recall the fact that B_t depends on the variables u_i, while A_t is fixed. Observe that only the first n vectors u_i are taken into consideration in the minimization, while all the m points are required to satisfy the constraint. A direct consequence of Theorem 4.14 is that problems (FP) and (PAP) are equivalent.

4.3.5 A density result

Next, we present a convergence result of (PAP) to (DP). The idea will be that when increasing the number of points in the triangulation, piecewise affine functions approximate better the solution of (DP). To do so, we adapt the definition given in [78, Definition 2.2] and recall some preliminary results.

Definition 4.15. Let $\Omega \subset \mathbb{R}^d$ be a bounded polytope. We say that $\mathfrak{F} = (\mathfrak{T}_k)_{k\in\mathbb{N}}$ is a *vanishing* family of simplicial partitions for Ω if, for every $k \in \mathbb{N}$, the longest edge of $\mathfrak{T}_k$ is of length at most $1/k$. In addition, we say that $\mathfrak{F}$ is strongly regular if there exists a constant $c > 0$ such that for all partitions $\mathfrak{T}_k \in \mathfrak{F}$ and all simplices $S \in \mathfrak{T}_k$ we have

$$\operatorname{meas}_d S \geq ck^{-d}.$$

In the sequel, the bounded polytope Ω that we will consider is $\operatorname{conv}(D)$. We are ready to present the first preliminary lemma.

Lemma 4.16. *Let $\mathfrak{F} = (\mathfrak{T}_k)_{k\in\mathbb{N}}$ be a strongly regular vanishing family of simplicial partitions for $\operatorname{conv}(D)$. Then, there exists a constant $c' > 0$ such that, for every partition $\mathfrak{T}_k \in \mathfrak{F}$ and all simplices $S_t \in \mathfrak{T}_k$, $t = 1, \dots, \ell(k)$, $\ell(k) \in \mathbb{N}$, if we denote by A_t the matrix relative to the simplex S_t defined in (4.3.10), we have*

$$\|A_t^{-1}\|_2 \leq \frac{k}{c'}.$$

Proof. Fix t and consider the singular values of the matrix A_t, $\sigma(A_t) = (\sigma_i)_{i=1}^d$, $\sigma_1 \geq \ldots \geq \sigma_d > 0$, and observe that for every $i = 1, \dots, d$,

$$\sigma_i \leq \sigma_1 := \|A_t\|_2 \leq \|A_t\|_{\mathrm{F}},$$

where

$$\|A_t\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{d} a_{ij}^2 = \sum_{i=1}^{d} \|a_i\|^2 \leq \frac{d}{k^2},$$

with $a_i = (x_{ti} - x_{t0})$, $a_{ij} = (x_{ti} - x_{t0})_j$, $i, j = 1, \dots, d$, where the sequence $(x_{ti})_{i=0}^d$ denotes the vertices of S_t. Then, for every $i = 1, \dots, d$,

$$\sigma_i \leq \frac{\sqrt{d}}{k}. \tag{4.3.12}$$

Now, let us recall that $(1/d!)|\det A_t| = \operatorname{meas}_d S_t$. Thus, by strong regularity we obtain

$$\prod_{i=1}^{d} \sigma_i = |\det A_t| \geq \operatorname{meas}_d S_t \geq ck^{-d},$$

for some $c > 0$. Thus, for every $j = 1, \dots, d$,

$$\sigma_j = \frac{\prod_{i=1}^{d} \sigma_i}{\prod_{i \neq j}^{d} \sigma_i} \geq \frac{ck^{-d}}{(\sqrt{d}/k)^{d-1}} \geq \frac{c'}{k},$$

for some $c' > 0$. Since $\|A_t^{-1}\|_2 = \sigma_d^{-1}$, we conclude. □

From now on, we will assume that, for every $k \in \mathbb{N}$, the partition $\mathfrak{T}_k$, formed with the points $(x_i^k)_{i=1}^{n_k}$, $n_k \in \mathbb{N}$, contains the set $\bar{D}$; i.e., for every $k \in \mathbb{N}$, $x_i^k = \bar{x}_i$ for every $i = 1, \dots, n$.

The following lemma is key to prove our convergence result. As it will be seen, we prove that the minimizer of $\widehat{L}$ on the set $\mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}$ converges, in some sense, to the minimizer of (DP) as k goes to infinity. We recall that the function $\widehat{L}$ remains the same for every $k \in \mathbb{N}$. In other words, although we augment the number of points at every iteration, the penalization is done only for the initial training points. Therefore, once a training set is fixed, we are able to prove that our constructed operators converge to a minimizer of (DP).

Lemma 4.17. *Let $\mathfrak{F} = (\mathfrak{T}_k)_{k\in\mathbb{N}}$ be a vanishing family of simplicial partitions for $\operatorname{conv}(D)$ and define, for every $k \in \mathbb{N}$,*

$$\widehat{N}_k \in \operatorname*{arg\,min}_{N \in \mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}} \widehat{L}(N), \tag{4.3.13}$$

where $\widehat{L}$ is defined as in (DP). Then, the sequence $(\widehat{N}_k)_{k\in\mathbb{N}}$ is bounded in $\mathrm{Lip}(\mathcal{X})$ and there exists a subsequence $\{k_j\}_{j\in\mathbb{N}}$ and an operator $\widehat{N} \in \mathcal{N}$ such that $\widehat{N}_{k_j} \overset{*}{\rightharpoonup} \widehat{N}$.

Proof. Let $k \in \mathbb{N}$ and let $\widehat{N}_k$ be a minimizer of $\widehat{L}$ in $\mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}$. Define, for every $k \in \mathbb{N}$, the operator $N_k^0 := \widehat{N}_k - \widehat{N}_k(0)$ and observe that $N_k^0 \in \mathcal{N}_0$. Since $\mathcal{N}_0$ is weak* compact by Corollary 4.6, there exists a subsequence $\{k_j\}_{j\in\mathbb{N}}$ and an operator $N_\infty^0 \in \mathcal{N}_0$ such that $N_{k_j}^0 \overset{*}{\rightharpoonup} N_\infty^0$ as $j \to \infty$. By Corollary 4.2, this implies in particular that, for every $x \in X$,

$$\widehat{N}_{k_j}(x) - \widehat{N}_{k_j}(0) \to N_\infty^0(x) - N_\infty^0(0) = N_\infty^0(x), \quad \text{as } j \to \infty. \tag{4.3.14}$$

Next, we claim that the sequence $(\widehat{N}_k(0))_{k\in\mathbb{N}}$ is bounded in $\mathcal{X}$. Indeed, observe first that, for every $k \in \mathbb{N}$, $\widehat{L}(\widehat{N}_k) \leq \widehat{L}(\pi_{\operatorname{conv}(D)})$ as, for every $k \in \mathbb{N}$, $\pi_{\operatorname{conv}(D)} \in \mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}$ and $\widehat{N}_k$ minimizes $\widehat{L}$ in $\mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}$. Since $\widehat{L}$ is coercive in the sense of (4.3.2) by Proposition 4.9 (applied to the empirical measure $\widehat{\mu} := \frac{1}{n}\sum_{i=1}^{n} \delta_{(\bar{x}_i, \bar{u}_i)}$), we conclude the proof of the claim. Therefore, up to subsequences, there exists $u \in \mathcal{X}$ such that $\widehat{N}_{k_j}(0) \rightharpoonup u$. Now, let us define the operator $N_\infty := N_\infty^0 + u$, observe that $N_\infty \in \mathcal{N}$ and note that, up to subsequences

$$\widehat{N}_{k_j}(0) \rightharpoonup u = N_\infty(0). \tag{4.3.15}$$

Finally, by (4.3.14), we obtain that, for every $x \in X$,

$$\widehat{N}_{k_j}(x) - \widehat{N}_{k_j}(0) \to N_\infty^0(x) = N_\infty(x) - N_\infty(0), \quad \text{as } j \to \infty.$$

By combining this with (4.3.15) and the characterization provided in Corollary 4.2, we have shown that, up to subsequences,

$$\widehat{N}_{k_j} \overset{*}{\rightharpoonup} N_\infty =: \widehat{N}, \quad \text{as } j \to \infty,$$

which is the result that we were aiming for. □

We are now ready to present the main result of the section.

Theorem 4.18. *Let $\mathfrak{F} = (\mathfrak{T}_k)_k$ be a strongly regular vanishing family of simplicial partitions for $\mathrm{conv}(D)$. Then, for every sequence of operators defined as in (4.3.13) there exists a subsequence $\{k_j\}_{j\in\mathbb{N}}$ such that*

$$\widehat{N}_{k_j} \overset{*}{\rightharpoonup} \widehat{N} \quad \textit{with} \quad \widehat{N} \in \operatorname*{arg\,min}_{N\in\mathcal{N}} \widehat{L}(N).$$

Proof. For every $k \in \mathbb{N}$, let $\{x_i^k\}_{i=1}^{n_k}$ be the vertices of all the simplices of the partition $\mathfrak{T}_k$ (notice that in particular $\{\bar{x}_i\}_{i=1}^n \subset \{x_i^k\}_{i=1}^{n_k}$ for every k, since $\mathfrak{T}_k$ are subsequent refinements). Let $\widehat{M}$ be a minimizer of (DP) and let $\delta > 0$. From Theorem 4.7, there exists a differentiable operator $N^\delta \in \mathcal{N}$ such that $\|N^\delta - \widehat{M}\|_\infty \leq \delta$. We define for all k the operators $N_k^\delta(x_i^k) := N^\delta(x_i^k)$ in the points x_i^k and extend them in a piecewise affine manner following $\mathfrak{T}_k$ (in particular for all k it holds $N_k^\delta(x_i) = N^\delta(x_i)$, $i = 1, \dots, n$). For every simplex $S_t \in \mathfrak{T}_k$ we indicate the vertices as $\{x_{i(t,j)}^k\}_{j=0}^d$, we introduce the numbers $\{u_{i(t,j)}^k = N^\delta\left(x_{i(t,j)}^k\right)\}_{j=0}^d$ and we define the matrices A_t, B_t as in (4.3.10). We recall the fact that $\|x_{i(t,j)}^k - x_{i(t,0)}^k\|_2 \leq \frac{1}{k}$ for every $S_t \in \mathfrak{T}_k$ and $j \in \{1, \dots, d\}$. Since N^δ is differentiable, we there exists a function $q : \mathbb{N} \to \mathbb{R}$, independent on j and t, such that $kq(k) \to 0$ and

$$N^\delta(x_{i(t,j)}^k) = N^\delta(x_{i(t,0)}^k) + \nabla N^\delta\left(x_{i(t,0)}^k\right)(x_{i(t,j)}^k - x_{i(t,0)}^k) + q(k),$$

for every $S_t \in \mathfrak{T}_k$ and $j \in \{1, \dots, d\}$. We can write $B_t = \nabla N^\delta(x_{i(t,0)}^k)A_t + q(k)$. It follows that

$$\|B_t A_t^{-1}\|_2 \leq \|\nabla N^\delta\|_\infty + q(k)\|A_t^{-1}\|_2 \leq 1 + q(k)\|A_t^{-1}\|_2,$$

where we used that $\|\nabla N^\delta\|_\infty \leq 1$ again by Theorem 4.7. By Lemma 4.16 we have $\|B_t A_t^{-1}\| \leq 1 + kq(k)/c'$. Since this estimate is independent of t, we have

$$\|N_k^\delta\|_{\mathrm{Lip}_0} = \max_{t=1,\dots,\ell(k)} \|B_t A_t^{-1}\| \leq 1 + kq(k)/c'.$$

In particular, we obtain $\max\{\|N_k^\delta\|_{\mathrm{Lip}_0}, 1\} \to 1$ as $k \to \infty$. We define now for every $x \in \mathcal{X}$,

$$\widetilde{N}_k^\delta(x) := \frac{N_k^\delta(x)}{\max\{\|N_k^\delta\|_{\mathrm{Lip}_0}, 1\}}.$$

We notice that $\widetilde{N}_k^\delta \in \mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}$ and, for every $i = 1, \dots, n$,

$$\widetilde{N}_k^\delta(\bar{x}_i) = \frac{N_k^\delta(\bar{x}_i)}{\max\{\|N_k^\delta\|_{\mathrm{Lip}_0}, 1\}} = \frac{N^\delta(\bar{x}_i)}{\max\{\|N_k^\delta\|_{\mathrm{Lip}_0}, 1\}} \to N^\delta(\bar{x}_i) \quad \text{as } k \to \infty. \tag{4.3.16}$$

By Lemma 4.17, there exists a subsequence $\{\widehat{N}_{k_j}\}_j$ of $\{\widehat{N}_k\}_k$ weakly converging to an operator $\widehat{N} \in \mathcal{N}$. By weak* lower semicontinuity of the functional $\widehat{L}$ and the fact that $\widehat{L}(\widehat{N}_{k_j}) \leq \widehat{L}(\widetilde{N}_{k_j}^\delta)$ (since $\widetilde{N}_{k_j}^\delta \in \mathrm{PA}(\mathfrak{T}_k) \cap \mathcal{N}$) we obtain

$$\widehat{L}(\widehat{N}) \leq \liminf_{j\to\infty} \widehat{L}(\widehat{N}_{k_j}) \leq \liminf_{j\to\infty} \widehat{L}(\widetilde{N}_{k_j}^\delta).$$

On the other hand, from 4.3.16 we get

$$\widehat{L}(\widetilde{N}_{k_j}^\delta) = \frac{1}{n}\sum_{i=1}^n \|\widetilde{N}_{k_j}^\delta(\bar{x}_i) - \bar{u}_i\|^2 \to \widehat{L}(N^\delta), \quad \text{as } k \to \infty.$$

Hence, $\widehat{L}(\widehat{N}) \leq \widehat{L}(N^\delta)$. Using Young's inequality and the definition of $\widehat{L}$ we obtain

$$\begin{aligned}\widehat{L}(\widehat{N}) \leq \widehat{L}(N^\delta) &= \frac{1}{n}\sum_{i=1}^{n} \|N^\delta(\bar{x}_i) - \bar{u}_i\|^2 \\ &\leq \frac{1}{n}\left[\sum_{i=1}^{n}\left(1+\frac{1}{\delta}\right)\|N^\delta(\bar{x}_i) - \widehat{M}(\bar{x}_i)\|^2 + (1+\delta)\,\|\widehat{M}(\bar{x}_i) - \bar{u}_i\|^2\right] \\ &\leq \delta^2\left(1+\frac{1}{\delta}\right) + (1+\delta)\,\widehat{L}(\widehat{M}).\end{aligned}$$

As δ was arbitrary we obtain $\widehat{L}(\widehat{N}) \leq \widehat{L}(\widehat{M})$ and the thesis. □

Remark 4.19. The problem of constructing strongly regular refinements is not a trivial task. In the case of $d = 2$, there are examples of strongly regular sequences of refinements. In particular, one can use the face-to-face longest-edge bisection algorithm [98], which was proven to provide a strongly regular sequence of refinements. Whether this technique gives strongly regular sequences in higher dimensions is still an open problem. However, it seems to perform well in practice and numerical experiments have been already provided for such issue, see for example [78].

We finish this section by providing the convergent PnP versions for some of the most well-known optimization algorithms: the Forward-Backward splitting algorithm, the Douglas–Rachford algorithm, and the Chambolle–Pock primal-dual iteration.

4.4 Convergent PnP methods

As it has been already mentioned in the introduction, the purpose of constructing firmly nonexpansive operators is that classical splitting methods require this property in order to achieve good convergence guarantees. In the following, we show how to rephrase classical methods in the PnP framework and recall some standard convergence results.

Finding a solution to a minimization problem of the form $\min_x f(x) + R(x)$, where both f and R are in $\Gamma_0(\mathcal{X})$, is equivalent (under some mild assumptions) to finding a point $x \in \mathcal{X}$ such that $0 \in \partial f(x) + \partial R(x)$. We have already mentioned in Chapter 2 that subdifferentials are a particular case of maximal monotone operators. This is the reason why, from a more general point of view, splitting methods and their relative convergence theory are often developed in the context of finding zeros in the sum of maximal monotone operators. In particular, these methods aim to find solutions of

$$0 \in A_1 x + A_2 x, \tag{4.4.1}$$

where A_i, $i = 1, 2$, are maximal monotone operators. A first example of such methods is the Forward-Backward Splitting algorithm (FBS), which iterates

$$x^{k+1} = J_{\tau A_2}(x^k - \tau A_1 x^k),$$

where $J_{\tau A_2}$ denotes the resolvent operator of A_2 with step-size τ. The weak convergence of the sequence $(x_k)_k$ generated by the algorithm to a solution of Problem (4.4.1) is guaranteed under the assumption that A_1 is β-cocoercive, i.e.,

$$\langle A_1 x - A_1 x', x - x'\rangle \geq \beta\|A_1 x - A_1 x'\| \quad \text{for all } x,\ x' \in \mathcal{X},$$

and $\tau \in (0, 2/\beta)$. We recall that, by Lemma 2.9, the set of resolvents of maximal monotone operators coincides with the set of firmly nonexpansive operators. For this reason,

substituting $J_{\tau A_2}$ with any firmly nonexpansive operator does not alter the convergence guarantees of the method. If we consider a firmly nonexpansive operator T, then the PnP-FBS algorithm iterates can be written as

$$x^{k+1} = T(x^k - \tau A_1 x^k). \tag{PnP-FBS}$$

If τ satisfies the same hypothesis stated above, then we can still guarantee convergence of the generated sequence. In particular, the sequence will converge (weakly) to a solution of the maximal monotone inclusion problem $0 \in A_1 x + A_T x$, where A_T is the (unique) maximal monotone operator that satisfies $T = (I + \tau A_T)^{-1}$.

Many other splitting methods have similar properties. We mention the so-called ADMM algorithm or, equivalently, the Douglas-Rachford algorithm (see [100] for the equivalence), which iterates

$$\begin{aligned} x_1^{k+1} &= J_{\tau A_1}(w^k), \\ x_2^{k+1} &= J_{\tau A_2}(2x_1^{k+1} - w^k), \\ w^{k+1} &= w^k + (x_2^{k+1} - x_1^{k+1}). \end{aligned} \tag{4.4.2}$$

It is known from [125] that, without further assumptions, both the sequences $(x_1^k)_k$ and $(x_2^k)_k$ generated by the algorithm, converge weakly to the same solution of Problem (4.4.1). Substituting again $J_{\tau A_2}$ with a firmly nonexpansive operator T, we obtain the PnP-DR algorithm

$$\begin{aligned} x_1^{k+1} &= J_{\tau A_1}(w^k), \\ x_2^{k+1} &= T(2x_1^{k+1} - w^k), \\ w^{k+1} &= w^k + (x_2^{k+1} - x_1^{k+1}), \end{aligned} \tag{PnP-DR}$$

generating two sequences $(x_1^k)_k$ and $(x_2^k)_k$ weakly converging to a solution of $0 \in A_1 x + A_T x$, with the same operator A_T introduced above.

When the problem is more structured, further splitting methods can be applied. Let us consider the problem $\min_x f(x) + R(Gx)$, where $G : \mathcal{X} \to \mathcal{X}'$ is a linear operator, $\mathcal{X}'$ a Hilbert space, $f \in \Gamma_0(\mathcal{X})$, and $R \in \Gamma_0(\mathcal{X}')$. The corresponding monotone inclusion problem becomes finding $x \in \mathcal{X}$ such that $0 \in \partial f(x) + G^* \partial R(Gx)$, and in general, finding $x \in \mathcal{X}$ such that

$$0 \in A_1 x + G^* A_2 G x, \tag{4.4.3}$$

with A_1, A_2 maximal monotone operators on $\mathcal{X}$ and $\mathcal{X}'$ respectively. In order to solve such a problem, one can employ the so-called Chambolle–Pock primal-dual iteration. The iterations in the context of maximal monotone inclusions read as

$$\begin{aligned} x^{k+1} &= J_{\tau A_1}(x^k - \tau G^* y^k), \\ y^{k+1} &= J_{\sigma A_2^{-1}}\left(y^k + \sigma G(2x^{k+1} - x^k)\right). \end{aligned}$$

Without assuming any further properties on the operators A_1 or A_2, the convergence of the method is guaranteed when $\tau\sigma\|G\|^2 \leq 1$ (see [40], or [27] for the limiting case). In particular, the sequence $(x^k)_k$ generated by the algorithm weakly converges to a solution of (4.4.3). This time, in order to find the corresponding PnP version of the algorithm, we first make use of Moreau's identity, which gives, for every $x \in \mathcal{X}$,

$$J_{\sigma A_2^{-1}}(x) = \sigma(\mathrm{Id} - J_{\sigma^{-1} A_2})(\sigma^{-1} x).$$

We can now write the iterations of the algorithm as

$$\begin{aligned} x^{k+1} &= J_{\tau A_1}(x^k - \tau G^* y^k), \\ y^{k+1} &= \sigma(\mathrm{Id} - J_{\sigma^{-1} A_2})\left(\sigma^{-1} y^k + G(2x^{k+1} - x^k)\right). \end{aligned}$$

Now again, we substitute the resolvent operator of A_2 with a firmly nonexpansive operator T and obtain the PnP-CP method

$$\begin{aligned} x^{k+1} &= J_{\tau A_1}(x^k - \tau G^* y^k), \\ y^{k+1} &= \sigma(\mathrm{Id} - T)\left(\sigma^{-1} y^k + G(2x^{k+1} - x^k)\right). \end{aligned} \tag{PnP-CP}$$

Assuming $\tau\sigma\|G\|_2^2 \leq 1$ we guarantee convergence of the method to a solution of the monotone inclusion problem $0 \in A_1 x + G^* A_T G x$, where A_T is defined by the equality $T = (\mathrm{Id} + \sigma^{-1} A_T)^{-1}$. Therefore, we have

$$0 \in A_1 x + G^* A_T G x = A_1 x + \sigma G^*(T^{-1} - \mathrm{Id}) G x. \tag{4.4.4}$$

In applications, we consider T coming from a learning process and thus, we consider the operator $(T^{-1} - \mathrm{Id})$ as fixed. This means that changing σ does change the underlying problem and thanks to the explicit expression in (4.4.4), we can think of σ as a regularization parameter (see also Section 4.5.1). This was only possible by making use of Moreau's identity which provides us with more interpretability of the solution than the one we could have achieved by only substituting $J_{\sigma A_2^{-1}}$ directly.

4.5 Experiments

In applications, given a set of data points $\{(\bar{x}_i, \bar{z}_i)\}_{i=1}^n$ we are interested in finding the best firmly nonexpansive operator that approximates these points in terms of the least squares distance. In order to do so, as explained at the beginning of Section 4.2 we can focus on learning a nonexpansive operator. We first define the points $\bar{u}_i = 2\bar{z}_i - \bar{x}_i$, we then fix a triangulation $\mathfrak{T}$ for $D = \bar{D} = \{\bar{x}_i\}_{i=1}^n$ (as explained in Section 4.3.3), and try to solve the following problem

$$\begin{aligned} \min_{u_1,\dots u_n \in \mathbb{R}^d} & \frac{1}{n}\sum_{i=1}^n \|u_i - \bar{u}_i\|_2^2 \\ & \text{s.t. } \|B_t(u_1,\dots,u_n)A_t^{-1}\|_2 \leq 1, \text{ for every } S_t \in \mathfrak{T}, \end{aligned} \tag{4.5.1}$$

where A_t and B_t are defined as in Section 4.3.4 considering $m = n$. This problem can be written in matricial form as follows

$$\min_{U \in \mathbb{R}^{d\times n}} \frac{1}{n}\|U - \bar{U}\|_F^2 + \sum_{S_t \in \mathfrak{T}} \iota_{\{\|B_t(\cdot)A_t^{-1}\|_2 \leq 1\}}(U).$$

Let us introduce the functions $f, g : \mathbb{R}^{d\times d} \to \mathbb{R} \cup \{+\infty\}$, defined by

$$f(U) = \frac{1}{n}\|U - \bar{U}\|_F^2, \quad R(M) = \iota_{\{\|\cdot\|_2 \leq 1\}}(M).$$

Moreover, define $G_t : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times d}$ linear operators that map U into $B_t(U)A_t^{-1}$ for each $t = 1, \dots, \ell$. The problem can be written in the form

$$\min_{\substack{U \in \mathbb{R}^{d\times n} \\ V_t \in \mathbb{R}^{d\times d}}} f(U) + \sum_{S_t \in \mathfrak{T}} R(V_t) \quad \text{s.t. } V_t = G_t U \quad t = 1, \dots, \ell \tag{4.5.2}$$

In order to solve it, we make use of the ADMM algorithm [18], considering a sequence of non-decreasing penalty parameters $\rho_k > 0$, that becomes stationary after a finite number of iterations (see [18, Section 3.4.1] for further details). The algorithm we use to solve Problem (4.5.2) is not relevant and one can make use of other suited iterative methods.

4.5.1 Image denoising

In the following, we focus on the image denoising problem described in Example 2.22. Let $\mathcal{X} = \mathbb{R}^{N\times N}$. We aim to reconstruct the clean image u^* by having only access to the measurement x, corrupted via a random additive noise $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$. We set a parameter $\lambda > 0$ and consider a rather more general problem than the one stated in Section 2.3.2, which writes as

$$\min_u \frac{1}{2}\|u - x\|_F^2 + \lambda R(Du) \tag{4.5.3}$$

where D denotes the discrete gradient defined in Section 2.3.2. Moreover, the function $R : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, assumed be convex, determines the penalization one wants to impose on the (discrete) gradient of the image. In practice, such a regularization function is difficult to choose this function since it must entail some prior knowledge of the image one wants to reconstruct. Some examples of R include $R(v) = \frac{\lambda}{2}\|v\|_F^2$, known as H^1 regularization, $R(v) = \lambda\|v\|_{1,1}$, where $\lambda > 0$ denotes the regularization parameter or, if we define $v_{ij} := (Du)_{i,j} \in \mathbb{R}^2$ as the gradient of the image u in the position (i, j), and define the functions $\|.\|_{p,1}$, for $p = 1, 2$,

$$\|v\|_{p,1} = \sum_{i=1}^{n}\sum_{j=1}^{m} \|(Du)_{i,j}\|_p,$$

then the choice $\lambda\|v\|_{1,1}$ correspond to the anisotropic TV (2.3.11), and the choice $\lambda\|v\|_{2,1}$ to the isotropic TV (2.3.10).

Our goal is to learn the operator R from scratch using a data-driven approach, and we do this by learning its proximal operator. In particular, we learn a firmly nonexpansive operator that we think of as a denoiser of gradients of images. In particular, observe that all of the possible choices for the function R that we presented above are separable with respect to the pixels and, moreover, the function is the same in every pixel (in the examples above, respectively, $\frac{\lambda}{2}\|\cdot\|_2^2$, $\lambda\|\cdot\|_1$ and $\lambda\|\cdot\|_2$). Such observation is key in our study since we will assume that the gradient can be locally and equally treated, pixel per pixel, by the learned resolvent. This corresponds to assume that R is of the form

$$R(v) = \sum_{i=1}^{n}\sum_{j=1}^{m} r(v_{ij}),$$

with $r : \mathbb{R}^2 \to \mathbb{R}$. Then, we learn $\partial r : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ through its resolvent $(\mathrm{Id} + \partial r)^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$. In practice, the resulting learned operator should preserve the properties of prox_R, but it will not necessarily be a proximal operator: we only guarantee that it is firmly nonexpansive. With the assumptions introduced above, instead of learning the firmly nonexpansive operator associated with prox_R, we just need to do such thing for prox_r instead, which is defined on $\mathbb{R}^2$ and can be well approximated by our method.

In practice, for the learning process, we consider gradients pixel by pixel of noisy images as inputs (called $\bar{x}_i \in \mathbb{R}^2$ in Section 4.3) and their corresponding clean gradients on the same pixels as outputs (called $\bar{z}_i \in \mathbb{R}^2$ at the beginning of Section 4.5). Once this firmly nonexpansive operator is learned, we denote it by T_r and we can plug it inside a splitting method to solve (4.5.3). In our experiments, we make use of the PnP-CP method

introduced in Section 4.4. The resulting algorithm reads as follows

$$\begin{aligned} u^{k+1} &= (u^k - \tau D^* v^k + \tau \hat{u}^\delta)/(1+\tau) \\ z^{k+1} &= v^k + \sigma D(2x^{k+1} - x^k) \\ &\text{for } i = 1, \dots, n, j = 1, \dots, m, \text{ do} \\ v_{ij}^{k+1} &= \sigma(I - T_r)(z_{ij}/\sigma) \end{aligned} \tag{4.5.4}$$

where the step-sizes $\tau > 0$ and $\sigma > 0$, are such that $\tau\sigma \leq \|D\|_2^2$, and T_r is the learned firmly nonexpansive operator (corresponding to the proximal operator of r). The variable u lives in $\mathbb{R}^{n\times m}$ and is the one that converges to the clean image. The variables v and z both live in $\mathbb{R}^{n\times m\times 2}$, with components $v_{ij}, z_{ij} \in \mathbb{R}^2$. Finally, the loop plays the role of the proximal operator of the dual of R. Notice that we used Moreau's identity to use directly T_r, as we explained in Section 4.4.

The learning process of our proposed denoiser was done over samples from an image of a butterfly and images from the MNIST dataset, shown in Figure 4.2. This choice was made to learn an operator that would allow the reconstruction of both edges and smoother parts. Notice, that it is possible to choose any image sample. To achieve better performance, a good choice for training images could be using images similar to the one that has to be reconstructed, if available. If the interest is focused more on the reconstruction of edges, a possibility is to choose manually pixels near edges or images with many edges. We show this in our experiments. After some analysis, we noticed that we arrived at good results also with a low number of data points (this made the learning process easy and relatively fast), for this reason, we chose $n = 1000$ pairs of data points $\{(\bar{x}_i, \bar{z}_i)\}_{i=1}^n$ and perform the learning process as described in the first part of Section 4.5. The noisy data were generated using Gaussian distributions with variance equal to 10. As discussed at the end of Section 4.4, we use the dual stepsize as a regularization parameter. Tuning such parameters allows us to deal with different noise levels (in the experiments we will set the variance of the Gaussian noise to 10, 20, and 30).

Code statement: All of the simulations have been implemented in Matlab on a laptop with Intel Core i7 1165G7 CPU @ 2.80GHz and 8 Gb of RAM. The code is available at https://github.com/TraDE-OPT/Learning-firmly-nonexpansive-operators.

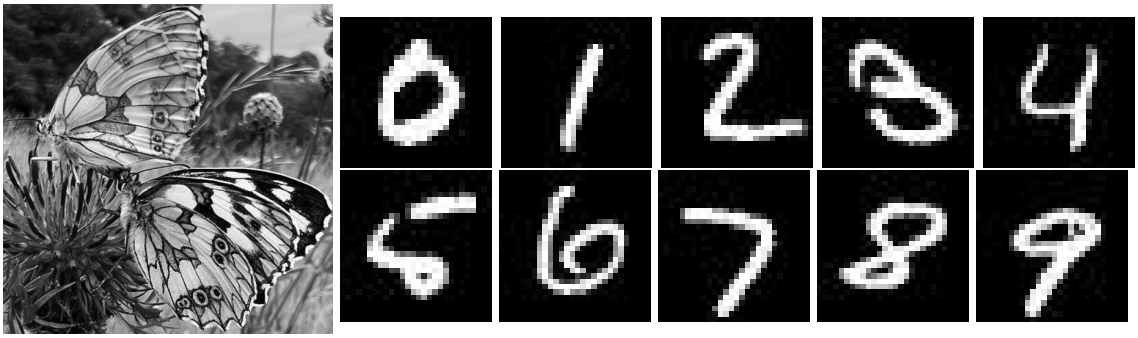

Figure 4.2: Clean data images. In order to construct the data set, we make use of Gaussian noise with noise variance $\tau^2 = 10$.

First experiments: butterfly In the first experiment, we test the learned operator using images similar to the ones used during the learning process. We use as training set clean and noisy data from the picture on the left in Figure 4.2 and as test image we consider a noisy version of the picture on the left in Figure 4.3. We perform experiments using

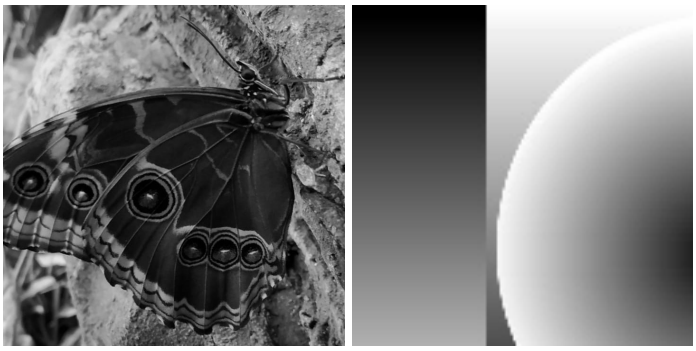

Figure 4.3: Clean version of the test images for experiments 1 and 2, respectively.

three different levels of noise applied to the test image, considering Gaussian noise with variance $\tau^2 = 10,\ 20$, and 30, respectively. In Figure 4.4, we report some results given by solving (4.5.3) with $R = \frac{\alpha}{2}\|\cdot\|_F^2$ (indicated as "H1" in the experiments), isotropic TV (indicated as "TV"), and the learned denoiser (indicated as "Learned"). Parameter and step-size choices were done manually, looking for best performance. In Table 4.5.1, we report the results in terms of PSNR and SSIM for reconstructions given by solving (4.5.3). As it can be seen in Table 4.5.1, our method can perform well in practice. We do not claim to have a method that improves state-of-the-art approaches. As we have already mentioned, the main contribution of our work is to propose a data-driven approach in the context of learning the proximal operator of convex regularizers, while still providing good theoretical guarantees.

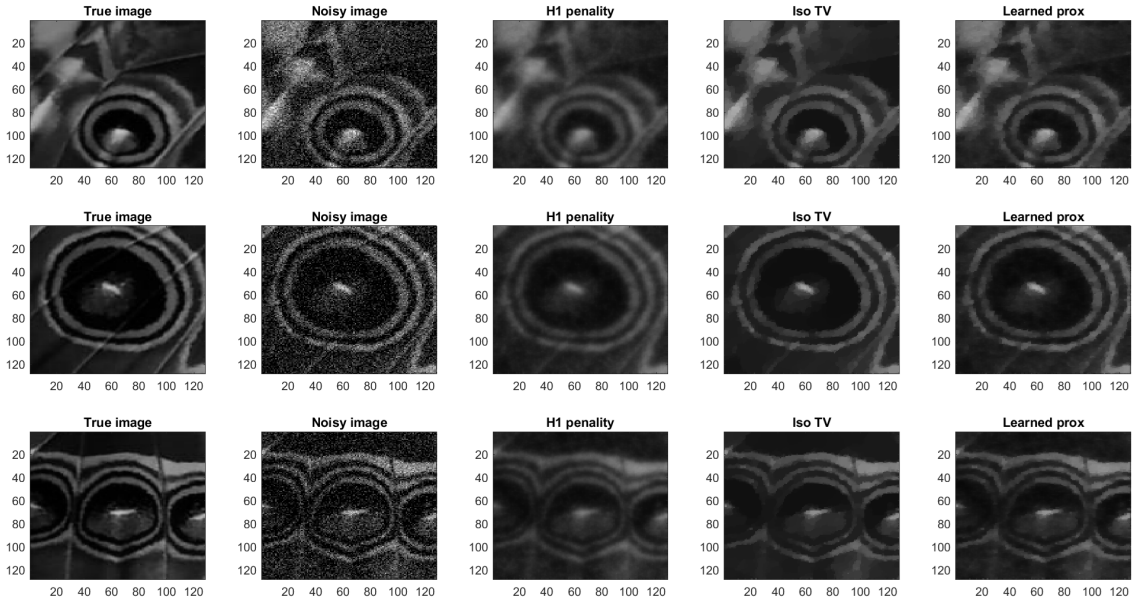

Figure 4.4: Butterfly images. Results of the experiment performed using a noisy image with Gaussian noise with noise variance $\tau^2 = 30$.

Second experiment: circles and edges In the second experiment we performed tests on images of circles and shadows, using two different data sets and thus, two different learned operators. We use data derived from the pictures shown in Figure 4.2. This choice was made to understand how Plug-and-Play methods were able to reconstruct both edges and smooth parts when provided with operators learned using completely different datasets. In Figure 4.5, we report some results given by solving (4.5.3) with $R = \frac{\alpha}{2}\|\cdot\|_F^2$

Butterfly images				
	Noisy	H1	TV	Learned
$\tau^2 = 10$				
PSNR (dB)	28.2806	32.1127	33.3655	**33.7698**
SSIM	0.67053	0.88882	0.82597	**0.90678**
$\tau^2 = 20$				
PSNR (dB)	22.5353	29.5202	29.8667	**29.8926**
SSIM	0.41033	**0.83430**	0.79383	0.82949
$\tau^2 = 30$				
PSNR (dB)	19.3036	27.2511	27.4723	**27.5141**
SSIM	0.27491	0.76491	0.75531	**0.77291**

Table 4.1: Comparison of performance for denoising using various regularization methods: H^1 penalty (indicated by H1), Isotropic Total Variation (indicated by TV), and ours (Learned). The comparison is given in terms of Peak Signal-to-Noise ratio (PSNR), and Structural Similarity Index Measure (SSIM).

(H1), isotropic TV (TV), the denoiser learned using the image of butterflies on the left in Figure 4.2 (Learned 1) and the denoiser learned using the MNIST images on the right in Figure 4.2 (Learned 2). Parameter and step-size choices were done manually, looking for best performance. We show the values in terms of PSNR and SSIM for three different values of the noise variance in Table 4.5.1. It is possible to see that (for both datasets) our data-driven method reconstructs well edges while not introducing too many artifacts in the image. It is interesting to notice that while the operator learned using more natural images (the images of butterflies) reconstructs smoother solutions, the one learned using the MNIST dataset seems to reconstruct better edges while introducing some artifacts (similarly to what happens for TV regularization).

Since we have the full expression of the learned operators, we can analyze them further. In Figure 4.6 (on the left) we plot the Lipschitz constants for the learned operator in every element of the triangulation $S_t \in \mathfrak{T}$. In particular, we plot the norm of the linear operator $B_t A_t^{-1}$ for each triangle. Recall that the Lipschitz constant for the learned operator is the maximum of the Lipschitz constant in every triangle. Even if the Lipschitz constant is not less or equal to 1 (but still close to 1), the PnP algorithm converges. The reason why the Lipschitz constant is not less than one is that we use an algorithm to find a solution of Problem (4.5.2) and we do not compute the minimizer exactly. To find a 1-Lipschitz operator one can try to replace the constraints $\|B_t(\cdot)A_t^{-1}\|_2 \leq 1$ with $\|B_t(\cdot)A_t^{-1}\|_2 \leq 1-\varepsilon$, $\varepsilon \in (0,1)$, searching in this way for $(1-\varepsilon)$-Lipschitz operators. We did some experiments in this direction, where we chose $\varepsilon = 0.01$. We learned in this way actual nonexpansive operators but we got similar results and the performance didn't improve. In the same figure, on the right, it is possible to see the decreasing behavior of the PnP Chambolle-Pock method described in (4.5.4). Figure 4.7 shows how the action of the learned operator from $\mathbb{R}^2$ to $\mathbb{R}^2$ looks like. The images show how the learned operator has both features of the prox of the 2-norm and the 2-norm squared.

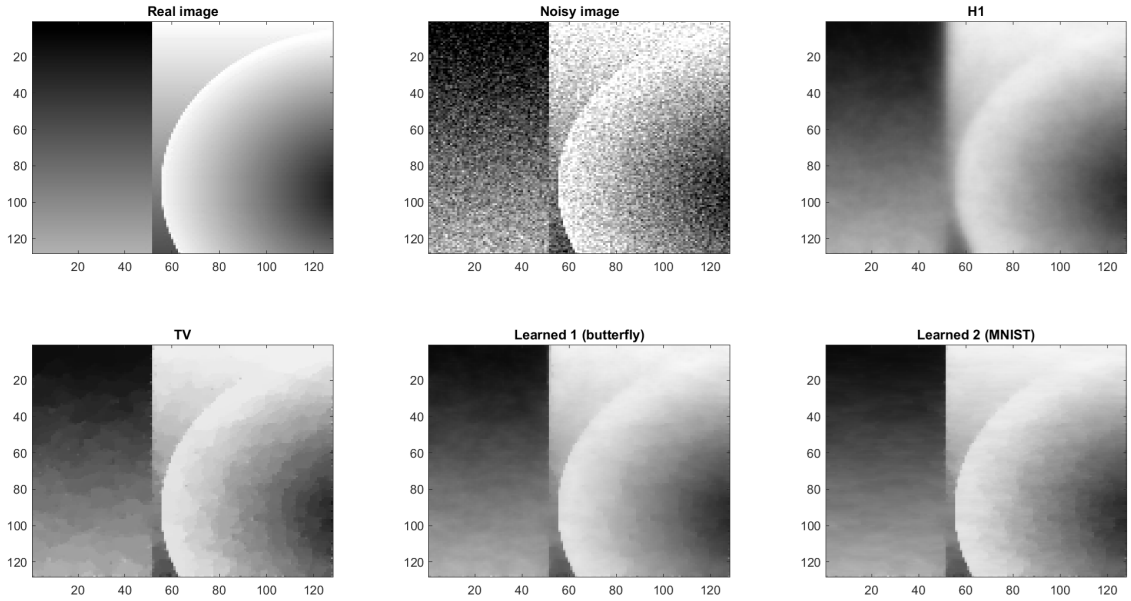

Figure 4.5: Circles. Results of the experiment performed using a noisy image with Gaussian noise, with noise variance $\tau^2 = 30$. Comparison for different training sets.

Circle images					
	Noisy	H1	TV	Learned 1	Learned 2
$\tau^2 = 10$					
PSNR (dB)	28.3301	30.2135	36.1745	36.4605	**37.6887**
SSIM	0.54501	0.94343	0.89702	**0.96562**	0.96444
$\tau^2 = 20$					
PSNR (dB)	22.5004	27.6047	31.1407	33.4358	**33.8865**
SSIM	0.28188	0.92580	0.78223	**0.93806**	0.93652
$\tau^2 = 30$					
PSNR (dB)	19.1533	25.4987	30.9583	30.8230	**31.1913**
SSIM	0.17782	**0.92462**	0.87905	0.92247	0.90443

Table 4.2: Comparison of performance for denoising using various regularization methods: H^1 penalty (H1), Isotropic Total Variation (TV), the learned operator using the butterfly dataset (Learned 1), and the learned operator using the MNIST dataset (Learned 2). The comparison is given in terms of Peak Signal-to-Noise ratio (PSNR), and Structural Similarity Index Measure (SSIM).

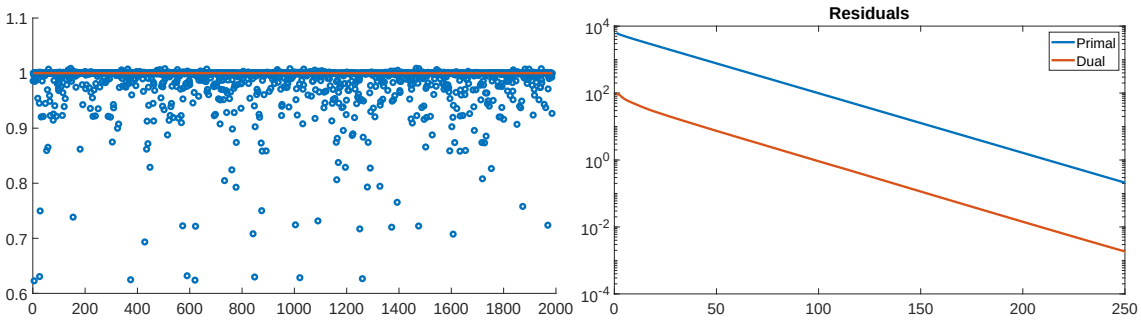

Figure 4.6: On the left: Lipschitz constants for the learned operator (Learned 1) in every element of the triangulation $S_t \in \mathfrak{T}$. On the right: decreasing behavior of the primal and dual residuals for the PnP Chambolle-Pock method described in (4.5.4), using the learned operator using butterfly data images (Learned 1).

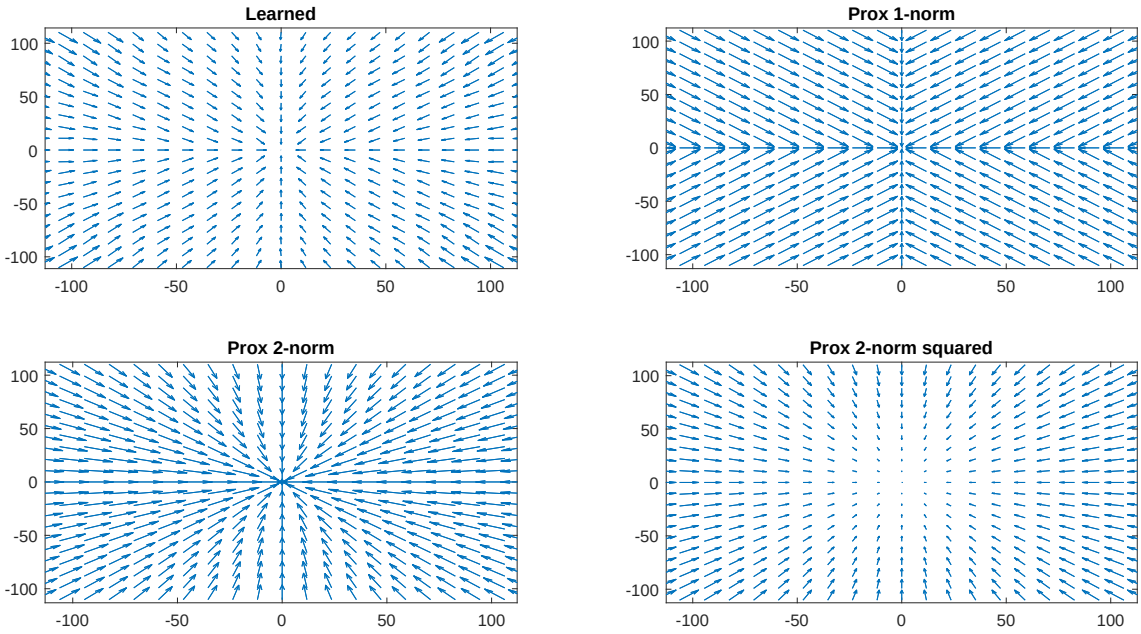

Figure 4.7: Action of proximal operators on a section of $\mathbb{R}^2$. Top-left: our learned operator (Learned 1), top-right: prox of the 1-norm (used for the anisotropic TV), bottom-left: prox of the 2-norm (used in the isotropic TV), bottom-right: prox of the 2-norm squared. It is possible to see that it shares some properties of the prox of the 1-norm and 2-norm. However, contrary to those, the magnitude of shift for the learned operator depends on the distance to zero, as it happens also for the prox of the 2-norm squared.

CHAPTER 5

On extreme points and representer theorems for the Lipschitz unit ball on finite metric spaces

5.1 Introduction

Let $(\mathcal{X}, d)$ be a metric space, $(\mathcal{U}, \|\cdot\|_{\mathcal{U}})$ a non-trivial, strictly convex real Banach space. We denote by Lip_0 to be the Banach space of Lipschitz functions from $\mathcal{X}$ to $\mathcal{U}$ that vanish at the distinct point $x_0 \in \mathcal{X}$ [140]. Let $L \geq 0$. We define the set

$$\mathrm{Lip}_0^L = \{f : \mathcal{X} \to \mathcal{U} \mid f(x_0) = 0 \text{ and } \|f(x) - f(x')\|_{\mathcal{U}} \leq L d(x, x') \text{ for all } x, x \in \mathcal{X}\}.$$

The objective of this chapter is to characterize the set of extreme points of the set Lip_0^1, which remains to be a challenging problem to date. Although the case $\mathcal{U} = \mathbb{R}$ has been studied to a large extent [65], no information about the general setting $\mathcal{U} = \mathbb{R}^d$, $d \geq 2$, has been provided. We partially fill this gap in Theorem 5.2, where we consider the general setting mentioned above with $\mathcal{U}$ being a strictly convex real Banach space, and $\mathcal{X}$ a finite metric space; i.e., $\mathcal{X} = \{x_0, ..., x_n\}$, being $x_0, \ldots, x_n$, $n \geq 1$, distinct points.

The second contribution of this work is to provide a representer theorem in this setting. These results have recently become popular in the context of variational inverse problems [19, 21, 133]. In the finite-dimensional setting, a natural connection can be provided: a representer theorem enables the expression of some solutions of variational problems through a finite number of extreme points: the extreme points associated to the unit ball of the regularizer [63]. For these reasons, there is increasing recent interest in characterizing extreme points associated with various regularizers, see [21, 23] and [25] and [4, 22, 37, 90, 134]. We provide in Theorem 5.3 a representation result for the space Lip_0^1 that improves Theorem 2.19. As we will see, the number of required extreme points is independent of the dimension of the space. In this sense, a generalization to the Minkowski-Carathéodory Theorem in infinite-dimensional spaces is provided.

5.2 Extreme points

We recall that, given C a convex subset of a real vector space, an *extreme point* of C is a point $u \in C$ such that, if $u = \frac{1}{2}u^1 + \frac{1}{2}u^2$ with $u^1,\ u^2 \in C$, then $u^1 = u^2 = u$. We denote with $\mathrm{ext}(C)$ the set of extreme points of C.

For further convenience, we consider the following definition of the set Lip_0^1.

$$\mathrm{Lip}_0^1 := \{u = (u_0, \dots, u_n) \in \mathcal{U}^{n+1} \,:\, u_0 = 0 \text{ and } \|u_i - u_j\|_{\mathcal{U}} \leq d(x_i, x_j), \\ \text{for every } i, j = 1, ..., \, n\}.$$

Note that both of the definitions of the set Lip_0^1 are equivalent, since in this case, we are only considering the images of the finite set $\mathcal{X} = \{x_0, ..., x_n\}$ through functions f mapping to $\mathcal{U}$. We provide now a preliminary lemma.

Lemma 5.1. *Let $u \in \mathrm{Lip}_0^1$. For every $i = 1, ..., n$, there exists $0 = i_0, i_1, .., i_k = i$, $k \geq 1$, such that $\|u_{i_{j+1}} - u_{i_j}\|_{\mathcal{U}} = d(x_{i_{j+1}}, x_{i_j})$, for every $j = 0, ..., k-1$ if and only if there does not exist a nonempty subset $S \subset \{1, ..., n\}$ such that $\|u_i - u_j\|_{\mathcal{U}} < d(x_i, x_j)$, for every $i \in S$, $j \in S^c$.*

Proof. We proceed by contradiction: let $S \subset \{1, ..., n\}$, $S \neq \emptyset$, such that $\|u_i - u_j\|_{\mathcal{U}} < d(x_i, x_j)$ for every $i \in S$, $j \in S^c$, and let $i \in S$. By hypothesis, we can choose $k \geq 1$ with $0 = i_0, ..., i_k = i$ such that $\|u_{i_{j+1}} - u_{i_j}\|_{\mathcal{U}} = d(x_{i_j}, x_{i_{j+1}})$ for every $j = 0, ..., k-1$. As $i_0 = 0 \in S^c$ and $i_k = i \in S$, we derive that there must exist $j = 0, ..., k-1$ such that $i_{j+1} \in S$ and $i_j \in S^c$. It follows that $\|u_{i_{j+1}} - u_{i_j}\|_{\mathcal{U}} < d(x_{i_{j+1}}, x_{i_j})$ but this contradicts the hypothesis and, hence, concludes the first part of the proof.

Conversely, let us consider the set

$$T = \{i \in \{1, ..., n\} \mid \text{there exists } \; 0 = i_0, ..., i_k = i \text{ s.t.} \\ \|u_{i_{\ell+1}} - u_{i_\ell}\|_{\mathcal{U}} = d(x_{i_{\ell+1}}, x_{i_{\ell+1}}), \, \ell = 0, ..., k-1\},$$

and suppose that $T \neq \{1, ..., n\}$. Define the set $S := T^c \subset \{1, ..., n\}$, and observe that $S \neq \emptyset$. It is left to prove that, for every $i \in S$, $j \in S^c$, $\|u_i - u_j\|_{\mathcal{U}} < d(x_i, x_j)$. Let us suppose that there exists $i \in S$ and $j \in S^c$ such that $\|u_i - u_j\|_{\mathcal{U}} = d(x_i, x_j)$. Since $j \in S^c$, there exists $0 = i_0, ..., i_k = j$, $k \geq 1$, such that $\|u_{i_{\ell+1}} - u_{i_\ell}\|_{\mathcal{U}} = d(x_{i_\ell}, x_{i_{\ell+1}})$ for $\ell = 0, ..., k-1$. Since, by hypothesis, we have that $\|u_i - u_j\|_{\mathcal{U}} = d(x_i, x_j)$, and defining $i_{k+1} := i$ we obtain a path from 0 to i satisfying the equalities for $\ell = 1, ..., k$. This implies that $i \in T = S^c$, a contradiction. We therefore have found that there exists $S \subset \{1, ..., n\}$, $S \neq \emptyset$ such that $\|u_i - u_j\|_{\mathcal{U}} < d(x_i, x_j)$, for every $i \in S$, $j \in S^c$. □

Define now the set

$$\mathcal{E} := \{u \in \mathcal{U}^{n+1} \mid u_0 = 0 \text{ and, for every } i = 1, ..., n, \text{ there exists} \\ 0 = i_0, ..., i_k = i, \; k \geq 1 : \|u_{i_{j+1}} - u_{i_j}\|_{\mathcal{U}} = d(x_{i_j}, x_{i_{j+1}}), \; j = 0, ..., \; k-1\}.$$

Observe that the definition of $\mathcal{E}$ is motivated by the previous lemma, since every point $u \in \mathcal{E}$ satisfies the first condition of Lemma 5.1. We are now ready to characterize the extreme points of the set Lip_0^1.

Theorem 5.2. *We have that $\mathrm{ext}(\mathrm{Lip}_0^1) = \mathcal{E}$.*

Proof. First, we will prove that, if $u \notin \mathcal{E}$, then $u \notin \mathrm{ext}(\mathrm{Lip}_0^1)$. Let $u \in \mathrm{Lip}_0^1$ such that $u \notin \mathcal{E}$. By the previous lemma, we get that there exists $S \subset \{1, ..., n\}$, $S \neq \emptyset$, such that $\|u_i - u_j\|_{\mathcal{U}} < d(x_i, x_j)$, for every $i \in S$, $j \in S^c$. Choose now

$$\varepsilon = \min_{i \in S, \, j \in S^c} d(x_i, x_j) - \|u_i - u_j\|_{\mathcal{U}},$$

and observe that $\varepsilon > 0$. Moreover, choose $v \in \mathcal{U}$ such that $\|v\|_{\mathcal{U}} = 1$ (which exists since $\mathcal{U}$ is non-trivial) and set

$$u_i^1 := \begin{cases} u_i + \varepsilon v, & \text{if } i \in S; \\ u_i, & \text{else,} \end{cases} \qquad u_i^2 := \begin{cases} u_i - \varepsilon v, & \text{if } i \in S; \\ u_i, & \text{else.} \end{cases}$$

Indeed, if we define $u^k := (u_0^k, u_1^k, ..., u_n^k)$, $k = 1, 2$, then $u^1 \neq u^2$. Moreover, observe that

$$\|u_i^k - u_j^k\|_{\mathcal{U}} = \|u_i - u_j\|_{\mathcal{U}} \leq d(x_i, x_j), \quad \text{for every } i, j \in S \text{ or } i, j \in S^c,\ k = 1, 2,$$

since $u \in \mathrm{Lip}_0^1$ and

$$\begin{aligned}\|u_i^k - u_j^k\|_{\mathcal{U}} = \|u_i \pm \varepsilon v - u_j\|_{\mathcal{U}} &\leq \|u_i - u_j\|_{\mathcal{U}} + \varepsilon \\ &\leq \|u_i - u_j\|_{\mathcal{U}} + d(x_i, x_j) - \|u_i - u_j\|_{\mathcal{U}} \\ &= d(x_i, x_j), \quad \text{for } i \in S,\ j \in S^c,\ k = 1, 2.\end{aligned}$$

Therefore, $u^k \in \mathrm{Lip}_0^1$, $k = 1, 2$ and $u = \frac{1}{2}u^1 + \frac{1}{2}u^2$, $u^1 \neq u^2$ and so $u \notin \mathrm{ext}(\mathrm{Lip}_0^1)$. Hence, $\mathrm{ext}(\mathrm{Lip}_0^1) \subset \mathcal{E}$.

We would like to prove now that $\mathcal{E} \subset \mathrm{ext}(\mathrm{Lip}_0^1)$. Let $u \in \mathrm{Lip}_0^1 \setminus \mathrm{ext}(\mathrm{Lip}_0^1)$. We will prove that there exists $S \subset \{1, ..., n\}$, $S \neq \emptyset$, such that $\|u_i - u_j\|_{\mathcal{U}} < d(x_i, x_j)$, for every $i \in S$, $j \in S^c$. If so, by the previous lemma, this would mean that $u \notin \mathcal{E}$. Since $u \notin \mathrm{ext}(\mathrm{Lip}_0^1)$, there exist u^1, $u^2 \in \mathrm{Lip}_0^1$, $u^1 \neq u^2$, such that $u = \frac{1}{2}u^1 + \frac{1}{2}u^2$. Now, define the set

$$S = \{i \in \{1, ..., n\} \mid u_i^1 \neq u_i^2\},$$

and observe that it is nonempty since $u^1 \neq u^2$ by hypothesis. Now, let $i \in S$, $j \in S^c$. Then,

$$\|u_i - u_j\|_{\mathcal{U}} = \left\|\frac{1}{2}u_i^1 - \frac{1}{2}u_j^1 + \frac{1}{2}u_i^2 - \frac{1}{2}u_j^2\right\|_{\mathcal{U}} = \left\|\frac{1}{2}u_i^1 - \frac{1}{2}u_j + \frac{1}{2}u_i^2 - \frac{1}{2}u_j\right\|_{\mathcal{U}}.$$

In order to finish the proof, define $a := u_i^1 - u_j$, $b := u_i^2 - u_j$, and observe that $a \neq b$. Now, we distinguish two cases: if a is not proportional to b, we get

$$\|u_i - u_j\|_{\mathcal{U}} < \frac{1}{2}\|a\|_{\mathcal{U}} + \frac{1}{2}\|b\|_{\mathcal{U}} \leq d(x_i, x_j),$$

since we assumed that $\mathcal{U}$ is a strictly convex space. If they are proportional, then, by possibly interchanging a and b, we have $b = \lambda a$ for some $\lambda \neq 1$, we can further assume that $-1 \leq \lambda < 1$, and obtain that

$$\|u_i - u_j\|_{\mathcal{U}} = \left\|\frac{a}{2} + \frac{\lambda a}{2}\right\|_{\mathcal{U}} \leq \frac{|1 + \lambda|}{2}\|a\|_{\mathcal{U}} < d(x_i, x_j).$$

The result immediately follows. □

In the next section, we provide a representation result as a natural consequence of the extreme points characterization shown above.

5.3 Representer theorems

We are now ready to state the representer theorem for the space Lip_0^1. In the case of $\mathcal{U} = \mathbb{R}^d$, the Minkowski–Carathéodory theorem would imply that every function in Lip_0^1 can be represented as a convex combination of at most $nd + 1$ extreme points. We are able to improve this number up to $n + 1$ extreme points, which is independent of d, and covers the infinite-dimensional case as well.

Theorem 5.3. *For every $u \in \mathrm{Lip}_0^1$, there exist $k \leq n + 1$, $u^1, ..., u^k \in \mathrm{ext}(\mathrm{Lip}_0^1)$, and scalars $\lambda_1, ..., \lambda_k \geq 0$ with $\sum_{i=1}^k \lambda_i = 1$ such that $u = \sum_{i=1}^k \lambda_i u^i$.*

Proof. Let $u \in \mathrm{Lip}_0^1$ and recall that it is of the form $u = (u_0, ..., u_n)$, with $u_0 = 0$. Choose $v \in \mathcal{U}$ such that $\|v\|_{\mathcal{U}} = 1$. Define the set

$$D = \left\{ t = (t_0, ..., t_n) \in \mathbb{R}^{n+1} \mid u + tv \in \mathrm{Lip}_0^1 \right\},$$

being $(u + tv)_i := u_i + t_i v$, for every $i = 0, ..., n$. Moreover, observe that $t_0 = 0$ for every $t \in D$ since, if $t_0 \neq 0$ then $(u + tv)_0 \neq 0$. Now, we claim that, if $t \in \mathrm{ext}(D)$, then $u + tv \in \mathrm{ext}(\mathrm{Lip}_0^1)$. Indeed, if $t \in D$ and $u + tv \notin \mathrm{ext}(\mathrm{Lip}_0^1)$, then there exists a subset $S \subset \{1, ..., n\}$, $S \neq \emptyset$, such that $\|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}} < d(x_i, x_j)$, for every $i \in S$, $j \in S^c$. Choose

$$\varepsilon = \min_{i \in S,\, j \in S^c} d(x_i, x_j) - \|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}},$$

and observe that $\varepsilon > 0$. Moreover, define

$$t_i^1 := \begin{cases} t_i + \varepsilon, & \text{if } i \in S; \\ t_i, & \text{else,} \end{cases} \qquad t_i^2 := \begin{cases} t_i - \varepsilon, & \text{if } i \in S; \\ t_i, & \text{else.} \end{cases}$$

With such definitions, observe that $t^1 \neq t^2$. Now, $u + t^k v \in \mathrm{Lip}_0^1$, for $k = 1, 2$, because

$$\begin{aligned} \|u_i - u_j + (t_i^k - t_j^k)v\|_{\mathcal{U}} &= \|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}} \\ &\leq d(x_i, x_j), \quad \text{for every } i, j \in S \text{ or } i, j \in S^c, \end{aligned}$$

since $t \in D$ and

$$\begin{aligned} \|u_i - u_j + (t_i^k - t_j^k)v\|_{\mathcal{U}} &\leq \|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}} + \varepsilon \|v\|_{\mathcal{U}} \\ &\leq \|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}} + d(x_i, x_j) - \|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}} \\ &= d(x_i, x_j), \quad \text{for } i \in S,\, j \in S^c,\, k = 1, 2. \end{aligned}$$

Then, $t^1, t^2 \in D$ and $t = \frac{1}{2}t^1 + \frac{1}{2}t^2$, $t^1 \neq t^2$, which implies that $t \notin \mathrm{ext}(D)$. Consequently, $t \in \mathrm{ext}(D)$ implies $u + tv \in \mathrm{ext}(\mathrm{Lip}_0^1)$. Now, we show that D is a nonempty, convex, compact subset of $\mathbb{R}^{n+1}$. First, note that $0 \in D$ and that convexity follows from fact that D is the preimage of the convex set Lip_0^1 through the affine mapping $t \mapsto u + tv$. Moreover, boundedness follows because, for every $t \in D$, we have

$$d(x_i, x_0) \geq \|(u + tv)_i - (u + tv)_0\|_{\mathcal{U}} = \|u_i - u_0 + t_i v\|_{\mathcal{U}} \geq |t_i| - \|u_i - u_0\|_{\mathcal{U}}$$

and so, for every $i = 1, \ldots, n$ we have that

$$|t_i| \leq d(x_i, x_0) + \|u_i - u_0\|_{\mathcal{U}} \leq 2d(x_i, x_0).$$

It is only left to prove that D is closed. Let $(t^k)_{k \in \mathbb{N}}$ be a sequence in D converging to some $t \in \mathbb{R}^{n+1}$. We have that

$$\begin{aligned} \|u_i - u_j + (t_i - t_j)v\|_{\mathcal{U}} &= \|u_i - u_j + (t_i^k - t_j^k)v - (t_i^k - t_j^k)v + (t_i - t_j)v\|_{\mathcal{U}} \\ &\leq \|u_i - u_j + (t_i^k - t_j^k)v\|_{\mathcal{U}} + \|(t_i - t_j)v - (t_i^k - t_j^k)v\|_{\mathcal{U}} \\ &\leq d(x_i, x_j) + |t_i - t_i^k| + |t_j - t_j^k|, \end{aligned}$$

for every $i, j = 0, \ldots, n$.We obtain the result by taking limits when $k \to \infty$. By the Krein–Milman theorem, we know that $D = \overline{\mathrm{conv}}(\mathrm{ext}(D))$. Moreover, we can apply the Minkowski–Carathéodory theorem, and since $0 \in D$, $\mathrm{span}\ D \subset \{0\} \times \mathbb{R}^n$, we have $\dim \mathrm{span}\ D \leq n$. Consequently, there exist $k \leq n + 1$ and scalars $\lambda_1, ..., \lambda_k \geq 0$ with $\sum_{i=1}^k \lambda_i = 1$ such that $0 = \sum_{i=1}^k \lambda_i t^i$, with $t^i \in \mathrm{ext}(D)$, $i = 1, ..., k$. Finally, by the previous claim, we know that for every $i = 1, ..., k$, if we define $u^i := u + t^i v$, then $u^i \in \mathrm{ext}(\mathrm{Lip}_0^1)$ and, hence

$$\sum_{i=1}^k \lambda_i u^i = \sum_{i=1}^k \lambda_i (u + t^i v) = u + \left(\sum_{i=1}^k \lambda_i t^i \right) v = u,$$

concluding the proof. □

CHAPTER 6

Conclusions

6.1 Summary

In this work, we studied data-driven approaches for solving inverse problems from different points of view. In particular, we focused on providing, or maintaining, the same theoretical guarantees obtained classical approaches. Below, we provide a detailed list of the topics that have been treated.

- We analyzed the problem of learning the regularization parameter for a large class of regularization methods in inverse problems. Such topic has gained atention in the past years due to its promising results in many applications [96, 97, 120], since it does not require to have any prior knowledge neither on the noise level nor on the ground truth. By applying statistical learning techniques [54, 137], we were able to characterize the error performance of this method following an Empirical Risk Minimization approach. Our analysis studies a wide variety of regularization methods, including spectral regularization methods (Tikhonov regularization, Landweber iteration, the ν-method), non-linear Tikhonov regularization [64] and general convex regularizers such as sparsity inducing norms [7]. Various numerical experiments have been included in order to validate and illustrate the theoretical findings. We believe that our results are a step forward towards understanding the theoretical principles of data-driven approaches applied to classical regularization techniques.

- We designed a data-driven approach for constructing firmly nonexpansive operators. This particular class of maps have recently become popular in the context of Plug-and-Play methods [138], which have turned out to be very successful in a wide variety of applications such as image recovery problems. In this work, considering a supervised learning approach, we present a complete theoretical framework for studying the problem of learning firmly nonexpansive operators. In addition, we propose a constructive method to produce firmly nonexpansive operators that adapts well to the characteristics of a given training set of noisy measurements/solutions. Our proposed approach, based on simplicial partitions and their refinements, gives theoretical guarantees for the convergence of PnP algorithms. In practice, the method is well-suited for low-dimensional problems.

- Finally, we explored certain geometric properties of the space of Lipschitz functions. In particular, we provided a characterization to the extreme points of the Lipschitz unit ball. Consequently, we studied representation results for such space and improved the Minkowski-Carathéodory Theorem in this context, showing that the required number of extreme points does not depend on the dimension of the space.

6.2 Future directions

In the context of Chapter 3, we mention an intriguing future direction.

- **Extension to deep-learning methods.** As we mentioned in the introduction, deep-learning techniques have been considered in the literature [96]. In this context, it is common to consider regularizers R that are implicitly parametrized by a large vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_k) \in \mathbb{R}^k$, for some $k \gg 1$. However, they are in general nonconvex, and therefore our approach can not be directly applied. Therefore, it would be natural to extend the provided framework in order to derive error bounds for this setting, which have not yet been studied.

The possible future directions of the work developed in Chapters 4 and 5 are many. We list below some of them

- **Extension to proximal operators of nonconvex regularizers.** Recent works [89, 116] have pointed out that going beyond convexity can be beneficial and help achieving better reconstruction outcomes. In all these cases, it is required to learn an operator with a fixed Lipschitz constant. Since the theory developed in the Chapter 4 easily adapts to the case of learning L-Lipschitz operator for any $L > 0$, we believe that our contribution could be an important complement to the existing analyses.

- **Dimensionality reduction techniques.** The experimental setting that we propose is based on learning firmly nonexpansive operators from $\mathbb{R}^2$ to $\mathbb{R}^2$, since the task of constructing simplicial partitions is costly in higher dimensions. Hence, we would like to explore whether any dimensionality reduction technique could be applied in order to learn firmly nonexpansive operators in higher dimensions.

- **Conditional Gradient Methods.** Recent works [24, 25, 52] have pointed out that a proper extreme points characterization could be useful to improve the performance of Conditional Gradient methods (CGM) in various settings. Both the extreme points characterization and the representation result provided in [26] fit perfectly into the framework of problem (4.3.7). Hence, a possible approach would be to consider our extreme points characterization and check whether classical CGMs can provide an efficient approach for solving (4.3.7).

- **Extension to general metric spaces.** Finally, much work needs to be done in order to provide a complete characterization to the extreme points of the Lipschitz unit ball; i.e. when $\mathcal{X}$ is a general metric space. We indeed aim to extend our result to this latter case.

References

[1] Alberti, G. S., De Vito, E., Lassas, M., Ratti, L., and Santacesaria, M. "Learning the optimal Tikhonov regularizer for inverse problems". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 25205–25216. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/d3e6cd9f66f2c1d3840ade4161cf7406-Paper.pdf.

[2] Alimohammadi, D. and Pazandeh, H. "Extreme points of the unit ball in the dual space of some real subspaces of Banach spaces of Lipschitz functions". In: *ISRN Math. Anal.* (2012), Art. ID 735139, 13. DOI: 10.5402/2012/735139.

[3] Aliprantis, C. D. and Border, K. C. *Infinite dimensional analysis*. Third. A hitchhiker's guide. Springer, Berlin, 2006, pp. xxii+703. ISBN: 978-3-540-32696-0; 3-540-32696-0.

[4] Ambrosio, L., Aziznejad, S., Brena, C., and Unser, M. "Linear inverse problems with Hessian–Schatten total variation". In: *Calc. Var.* 63.1 (2023), p. 9. ISSN: 1432-0835. DOI: 10.1007/s00526-023-02611-6.

[5] Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. "Solving inverse problems using data-driven models". In: *Acta Numerica* 28 (2019), pp. 1–174. DOI: 10.1017/S0962492919000059.

[6] Attouch, H., Bolte, J., and Svaiter, B. F. "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods". In: *Mathematical Programming, Series A* 137.1 (2011), pp. 91–124. DOI: 10.1007/s10107-011-0484-9.

[7] Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. "Optimization with sparsity-inducing penalties". In: *Foundations and Trends® in Machine Learning* 4.1 (2012), pp. 1–106.

[8] Bakushinskii, A. "Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion". In: *USSR Computational Mathematics and Mathematical Physics* 24.4 (1984), pp. 181–182. ISSN: 0041-5553. DOI: https://doi.org/10.1016/0041-5553(84)90253-2.

[9] Bauer, F. and Lukas, M. A. "Comparing parameter choice methods for regularization of ill-posed problems". In: *Mathematics and Computers in Simulation* 81.9

(2011), pp. 1795–1841. ISSN: 0378-4754. DOI: https://doi.org/10.1016/j.matcom.2011.01.016.

[10] Bauer, F., Pereverzev, S., and Rosasco, L. "On regularization algorithms in learning theory". In: *J. Complexity* 23.1 (2007), pp. 52–72.

[11] Bauer, F. and Reiß, M. "Regularization independent of the noise level: an analysis of quasi-optimality". In: *Inverse Problems* 24.5 (2008), p. 055009. DOI: 10.1088/0266-5611/24/5/055009.

[12] Bauschke, H. H., Borwein, J. M., and Combettes, P. L. "Essential Smoothness, Essential Strict Convexity, and Legendre Functions in Banach Spaces". In: *Communications in Contemporary Mathematics* 03.04 (2001), pp. 615–647. DOI: 10.1142/S0219199701000524.

[13] Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. Second. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, 2017, pp. xix+619. ISBN: 978-3-319-48310-8; 978-3-319-48311-5. DOI: 10.1007/978-3-319-48311-5.

[14] Beck, A. and Teboulle, M. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202. DOI: 10.1137/080716542.

[15] Benning, M. and Burger, M. "Modern regularization methods for inverse problems". In: *Acta Numerica* 27 (2018), pp. 1–111. DOI: 10.1017/S0962492918000016.

[16] Bertero, M. and Boccacci, P. *Introduction to inverse problems in imaging*. Institute of Physics Publishing, Bristol, 1998, pp. xii+351. ISBN: 0-7503-0439-1; 0-7503-0435-9. DOI: 10.1887/0750304359.

[17] Billingsley, P. *Convergence of probability measures*. Second. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1999, pp. x+277. ISBN: 0-471-19745-9. DOI: 10.1002/9780470316962.

[18] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundations and Trends in Machine Learning* 3 (Jan. 2011), pp. 1–122. DOI: 10.1561/2200000016.

[19] Boyer, C., Chambolle, A., De Castro, Y., Duval, V., Gournay, F. de, and Weiss, P. "On representer theorems and convex regularization". In: *SIAM J. Optim.* 29.2 (2019), pp. 1260–1281. ISSN: 1052-6234. DOI: 10.1137/18M1200750.

[20] Braides, A. *A handbook of Γ-convergence*. cvgmt preprint. 2006. URL: http://cvgmt.sns.it/paper/57/.

[21] Bredies, K. and Carioni, M. “Sparsity of solutions for variational inverse problems with finite-dimensional data”. In: *Calc. Var. Partial Differential Equations* 59.1 (2020), Paper No. 14, 26. ISSN: 0944-2669. DOI: 10.1007/s00526-019-1658-1.

[22] Bredies, K., Carioni, M., and Fanzon, S. “A superposition principle for the inhomogeneous continuity equation with Hellinger–Kantorovich-regular coefficients”. In: *Communications in Partial Differential Equations* 47 (Sept. 2022), pp. 1–47. DOI: 10.1080/03605302.2022.2109172.

[23] Bredies, K., Carioni, M., Fanzon, S., and Romero, F. “On the extremal points of the ball of the Benamou-Brenier energy”. In: *Bull. Lond. Math. Soc.* 53.5 (2021), pp. 1436–1452. ISSN: 0024-6093. DOI: 10.1112/blms.12509.

[24] Bredies, K., Carioni, M., Fanzon, S., and Romero, F. “A Generalized Conditional Gradient Method for Dynamic Inverse Problems with Optimal Transport Regularization”. In: *Foundations of Computational Mathematics* (2022). ISSN: 1615-3375. DOI: 10.1007/s10208-022-09561-z.

[25] Bredies, K., Carioni, M., Fanzon, S., and Walter, D. “Asymptotic linear convergence of fully-corrective generalized conditional gradient methods”. In: *Mathematical Programming* (2023). DOI: 10.1007/s10107-023-01975-z.

[26] Bredies, K., Chirinos Rodriguez, J., and Naldi, E. “On extreme points and representer theorems for the Lipschitz unit ball on finite metric spaces”. In: *Arch. Math.* (2024). DOI: 10.1007/s00013-024-01978-y.

[27] Bredies, K., Chenchene, E., Lorenz, D. A., and Naldi, E. “Degenerate Preconditioned Proximal Point Algorithms”. In: *SIAM Journal on Optimization* 32.3 (2022), pp. 2376–2401. DOI: 10.1137/21M1448112.

[28] Bredies, K. and Holler, M. “Higher-order total variation approaches and generalisations”. In: *Inverse Problems* 36.12 (2020), p. 123001. ISSN: 1361-6420. DOI: 10.1088/1361-6420/ab8f80.

[29] Bredies, K., Kunisch, K., and Pock, T. “Total Generalized Variation”. In: *SIAM Journal on Imaging Sciences* 3.3 (2010), pp. 492–526. DOI: 10.1137/090769521.

[30] Bungert, L., Korolev, Y., and Burger, M. “Structural analysis of an L-infinity variational problem and relations to distance functions”. In: *Pure and Applied Analysis* 2.3 (2020), pp. 703–738.

[31] Burger, M., Resmerita, E., and He, L. “Error estimation for Bregman iterations and inverse scale space methods in image restoration”. In: *Computing* 81 (Nov. 2007), pp. 109–135. DOI: 10.1007/s00607-007-0245-z.

[32] Buzzard, G. T., Chan, S. H., Sreehari, S., and Bouman, C. A. “Plug-and-Play Unplugged: Optimization-Free Reconstruction Using Consensus Equilibrium”. In: *SIAM Journal on Imaging Sciences* 11.3 (2018), pp. 2001–2020. DOI: 10.1137/17M1122451.

[33] Calatroni, L., Cao, C., Reyes, J. C. De los, Schönlieb, C.-B., and Valkonen, T. "Bilevel approaches for learning of variational imaging models". In: *Variational Methods*. De Gruyter, 2017, pp. 252–290. ISBN: 9783110430394. DOI: doi:10.1515/9783110430394-008.

[34] Calatroni, L. and Papafitsoros, K. "Analysis and automatic parameter selection of a variational model for mixed Gaussian and salt-and-pepper noise removal". In: *Inverse Problems* 35.11 (2019), p. 114001. DOI: 10.1088/1361-6420/ab291a.

[35] Candès, E. J. and Recht, B. "Exact matrix completion via convex optimization". In: *Found. Comput. Math.* 9.6 (2009), pp. 717–772. ISSN: 1615-3375,1615-3383. DOI: 10.1007/s10208-009-9045-5.

[36] Caponnetto, A. and Yao, Y. "Cross-Validation Based Adaptation for Regularization Operators in Learning Theory". In: *Analysis and Applications* 08.02 (2010), pp. 161–183. DOI: 10.1142/S0219530510001564.

[37] Carioni, M., Iglesias, J. A., and Walter, D. "Extremal Points and Sparse Optimization for Generalized Kantorovich–Rubinstein Norms". In: *Foundations of Computational Mathematics* (2023), pp. 1–42.

[38] Chambolle, A. "An Algorithm for Total Variation Minimization and Applications". In: *Journal of Mathematical Imaging and Vision* 20 (2004), pp. 89–97. DOI: 10.1023/B:JMIV.0000011325.36760.1e.

[39] Chambolle, A. and Lions, P.-L. "Image recovery via total variation minimization and related problems". In: *Numerische Mathematik* 76 (1997), pp. 167–188.

[40] Chambolle, A. and Pock, T. "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging". In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145.

[41] Chan, S. H., Wang, X., and Elgendy, O. A. "Plug-and-Play ADMM for Image Restoration: Fixed-Point Convergence and Applications". In: *IEEE Transactions on Computational Imaging* 3.1 (2017), pp. 84–98. DOI: 10.1109/TCI.2016.2629286.

[42] Chan, T., Marquina, A., and Mulet, P. "High-Order Total Variation-Based Image Restoration". In: *SIAM Journal on Scientific Computing* 22.2 (2000), pp. 503–516. DOI: 10.1137/S1064827598344169.

[43] Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. "The Convex Geometry of Linear Inverse Problems". In: *Foundations of Computational Mathematics* 12.6 (2012), 805—849. ISSN: 1615-3383. DOI: 10.1007/s10208-012-9135-7. URL: http://dx.doi.org/10.1007/s10208-012-9135-7.

[44] Chen, S. S., Donoho, D. L., and Saunders, M. A. "Atomic Decomposition by Basis Pursuit". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 33–61. DOI: 10.1137/S1064827596304010.

[45] Chenchene, E., Hosseini, A., and Bredies, K. “A hybrid proximal generalized conditional gradient method and application to total variation parameter learning”. In: *2023 European Control Conference (ECC)*. IEEE. 2023, pp. 1–6.

[46] Chirinos Rodriguez, J., De Vito, E., Molinari, C., Rosasco, L., and Villa, S. *On Learning the Optimal Regularization Parameter in Inverse Problems*. 2023. arXiv: 2311.15845 [math.ST].

[47] Clason, C. *Regularization of Inverse Problems*. 2021. arXiv: 2001.00617 [math.FA].

[48] Cobzaş, S. “Extreme points in Banach spaces of Lipschitz functions”. In: *Mathematica (Cluj)* 31(54).1 (1989), pp. 25–33. ISSN: 0025-5505.

[49] Combettes, P. L. and Pesquet, J.-C. “Proximal splitting methods in signal processing”. In: *Fixed-point algorithms for inverse problems in science and engineering*. Vol. 49. Springer Optim. Appl. Springer, New York, 2011, pp. 185–212. DOI: 10.1007/978-1-4419-9569-8_10.

[50] Combettes, P. L. and Wajs, V. R. “Signal Recovery by Proximal Forward-Backward Splitting”. In: *Multiscale Modeling & Simulation* 4.4 (2005), pp. 1168–1200. DOI: 10.1137/050626090.

[51] Condat, L. “A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms”. In: *Journal of Optimization Theory and Applications* 158.2 (2013), pp. 460–479. DOI: 10.1007/s10957-012-0245-9.

[52] Cristinelli, G., Iglesias, J. A., and Walter, D. “Conditional gradients for total variation regularization with PDE constraints: a graph cuts approach”. In: *ArXiv* abs/2310.19777 (2023).

[53] Crockett, C. and Fessler, J. A. “Bilevel Methods for Image Reconstruction”. In: *Foundations and Trends® in Signal Processing* 15.2-3 (2022), pp. 121–289. ISSN: 1932-8346. DOI: 10.1561/2000000111.

[54] Cucker, F. and Smale, S. “On the mathematical foundations of learning”. In: *Bull. Amer. Math. Soc. (N.S.)* 39.1 (2002), pp. 1–49. ISSN: 0273-0979,1088-9485. DOI: 10.1090/S0273-0979-01-00923-5.

[55] Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. “Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering”. In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095. DOI: 10.1109/TIP.2007.901238.

[56] Daubechies, I., Defrise, M., and De Mol, C. “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint”. In: *Communications on Pure and Applied Mathematics* 57.11 (2004), pp. 1413–1457. DOI: https://doi.org/10.1002/cpa.20042.

[57] De Vito, E., Fornasier, M., and Naumova, V. “A machine learning approach to optimal Tikhonov regularization I: Affine manifolds”. In: *Analysis and Applications* 20.02 (2022), pp. 353–400. DOI: 10.1142/S0219530520500220.

[58] Debarre, T. J. “Variational Methods For Continuous-Domain Inverse Problems: the Quest for the Sparsest Solution”. In: (2022), p. 325. DOI: https://doi.org/10.5075/epfl-thesis-9287.

[59] Deng, L. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142. DOI: 10.1109/MSP.2012.2211477.

[60] Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*. Vol. 31. Applications of Mathematics (New York). Springer-Verlag, New York, 1996, pp. xvi+636. ISBN: 0-387-94618-7. DOI: 10.1007/978-1-4612-0711-5.

[61] Dong, W., Wang, P., Yin, W., Shi, G., Wu, F., and Lu, X. “Denoising Prior Driven Deep Neural Network for Image Restoration”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.10 (2019), pp. 2305–2318. DOI: 10.1109/TPAMI.2018.2873610.

[62] Donoho, D. L. and Johnstone, I. M. “Adapting to Unknown Smoothness via Wavelet Shrinkage”. In: *Journal of the American Statistical Association* 90.432 (1995), pp. 1200–1224. DOI: 10.1080/01621459.1995.10476626.

[63] Duval, V. “Faces and extreme points of convex sets for the resolution of inverse problems”. Habilitation à diriger des recherches. Ecole doctorale SDOSE, 2022.

[64] Engl, H. W., Hanke, M., and Neubauer, A. *Regularization of inverse problems*. Vol. 375. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1996, pp. viii+321. ISBN: 0-7923-4157-0. DOI: 10.1007/978-3-540-70529-1_52.

[65] Farmer, J. D. “Extreme points of the unit ball of the space of Lipschitz functions”. In: *Proc. Amer. Math. Soc.* 121.3 (1994), pp. 807–813. ISSN: 0002-9939.

[66] Fazel, M., Hindi, H., and Boyd, S. “Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices”. In: *Proceedings of the 2003 American Control Conference, 2003.* Vol. 3. 2003, 2156–2162 vol.3. DOI: 10.1109/ACC.2003.1243393.

[67] Fazel, M. “Matrix rank minimization with applications”. PhD thesis. Stanford University, 2002.

[68] Folland, G. B. *Real analysis*. Second. Pure and Applied Mathematics (New York). Modern techniques and their applications, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999, pp. xvi+386. ISBN: 0-471-31716-0.

[69] Fortune, S. “Voronoĭ diagrams and Delaunay triangulations”. In: *Computing in Euclidean geometry*. Vol. 1. Lecture Notes Ser. Comput. World Sci. Publ., River Edge, NJ, 1992, pp. 193–233. DOI: 10.1142/9789814355858_0006.

[70] Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. “Bilevel Programming for Hyperparameter Optimization and Meta-Learning”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, 2018, pp. 1568–1577.

[71] Frank, M. and Wolfe, P. “An algorithm for quadratic programming”. In: *Naval Research Logistics Quarterly* 3.1-2 (1956), pp. 95–110. DOI: https://doi.org/10.1002/nav.3800030109.

[72] Garrigos, G., Rosasco, L., and Villa, S. “Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry.” In: *Mathematical Programming* 198.1 (2023), pp. 937–996. ISSN: 00255610. DOI: 10.1007/s10107-022-01809-4.

[73] Golub, G. H. and Matt, U. von. “Generalized Cross-Validation for Large-Scale Problems”. In: *Journal of Computational and Graphical Statistics* 6.1 (1997), pp. 1–34. DOI: 10.2307/1390722.

[74] Grasmair, M., Haltmeier, M., and Scherzer, O. “Sparse regularization with lq penalty term”. In: *Inverse Problems* 24.5 (2008), p. 055020. DOI: 10.1088/0266-5611/24/5/055020.

[75] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002, pp. xvi+647. ISBN: 0-387-95441-4. DOI: 10.1007/b97848.

[76] Hadamard, J. *Lectures on Cauchy’s problem in linear partial differential equations*. Vol. 15. Yale university press, 1923.

[77] Hämarik, U. and Tautenhahn, U. “On the Monotone Error Rule for Parameter Choice in Iterative and Continuous Regularization Methods”. In: *BIT* 41 (Dec. 2001), pp. 1029–1038. DOI: 10.1023/A:1021945429767.

[78] Hannukainen, A., Korotov, S., and Křížek, M. “On numerical regularity of the face-to-face longest-edge bisection algorithm for tetrahedral partitions”. In: *Science of Computer Programming* 90 (2014). Special issue on Numerical Software: Design, Analysis and Verification, pp. 34–41. ISSN: 0167-6423. DOI: https://doi.org/10.1016/j.scico.2013.05.002.

[79] Hansen, P. C. “Analysis of Discrete Ill-Posed Problems by Means of the L-Curve”. In: *SIAM Review* 34.4 (1992), pp. 561–580. DOI: 10.1137/1034115.

[80] Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN: 1498712169.

[81] He, J., Yang, Y., Wang, Y., Zeng, D., Bian, Z., Zhang, H., Sun, J., Xu, Z., and Ma, J. "Optimizing a Parameterized Plug-and-Play ADMM for Iterative Low-Dose CT Reconstruction". In: *IEEE Transactions on Medical Imaging* 38.2 (2019), pp. 371–382. DOI: 10.1109/TMI.2018.2865202.

[82] Heide, F., Steinberger, M., Tsai, Y.-T., Rouf, M., Pajkak, D., Reddy, D., Gallo, O., Liu, J., Heidrich, W., Egiazarian, K., Kautz, J., and Pulli, K. "FlexISP: A Flexible Camera Image Processing Framework". In: *ACM Trans. Graph.* 33 (Nov. 2014), 231:1–231:13. DOI: 10.1145/2661229.2661260.

[83] Heinonen, J. *Lectures on Lipschitz Analysis*. Tech. rep. Jyväskylä: University of Jyväskylä, 2005.

[84] Helmberg, G. *Introduction to spectral theory in Hilbert space*. North-Holland Series in Applied Mathematics and Mechanics, Vol. 6. North-Holland Publishing Co., Amsterdam-London; Wiley-Interscience [John Wiley & Sons], New York, 1969, pp. xiii+346.

[85] Hertrich, J., Neumayer, S., and Steidl, G. "Convolutional proximal neural networks and Plug-and-Play algorithms". In: *Linear Algebra and its Applications* 631 (2021), pp. 203–234. ISSN: 0024-3795. DOI: https://doi.org/10.1016/j.laa.2021.09.004.

[86] Himmelberg, C. J. "Measurable relations". In: *Fundamenta Mathematicae* 87 (1975), pp. 53–72.

[87] Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms. I*. Vol. 305. Springer-Verlag, Berlin, 1993, pp. xviii+417. ISBN: 3-540-56850-6.

[88] Holler, G., Kunisch, K., and Barnard, R. C. "A bilevel approach for parameter learning in inverse problems". In: *Inverse Problems* 34.11 (2018), p. 115012. DOI: 10.1088/1361-6420/aade77.

[89] Hurault, S., Leclaire, A., and Papadakis, N. "Proximal Denoiser for Convergent Plug-and-Play Optimization with Nonconvex Regularization". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 9483–9505.

[90] Iglesias, J. A. and Walter, D. *Extremal points of total generalized variation balls in 1D: characterization and applications*. arXiv:2112.06846. 2022.

[91] Jaggi, M. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". In: *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. Proceedings of Machine Learning Research 1. 2013, pp. 427–435. URL: https://proceedings.mlr.press/v28/jaggi13.html.

[92] Johanis, M. "Approximation of Lipschitz mappings". In: *Serdica Math. J.* 29.2 (2003), pp. 141–148. URL: http://eudml.org/doc/219548.

[93] Kamilov, U. S., Bouman, C. A., Buzzard, G. T., and Wohlberg, B. "Plug-and-Play Methods for Integrating Physical and Learned Models in Computational Imaging: Theory, algorithms, and applications". In: *IEEE Signal Processing Magazine* 40.1 (2023), pp. 85–97. DOI: 10.1109/MSP.2022.3199595.

[94] Kereta, Z. and Naumova, V. "On an unsupervised method for parameter selection for the elastic net". In: *Mathematics in Engineering* 4.6 (2022), pp. 1–36. DOI: 10.3934/mine.2022053.

[95] Kirszbraun, M. "Über die zusammenziehende und Lipschitzsche Transformationen". ger. In: *Fundamenta Mathematicae* 22.1 (1934), pp. 77–108. URL: http://eudml.org/doc/212681.

[96] Kobler, E., Effland, A., Kunisch, K., and Pock, T. "Total Deep Variation for Linear Inverse Problems". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7546–7555. DOI: 10.1109/CVPR42600.2020.00757.

[97] Kunisch, K. and Pock, T. "A Bilevel Optimization Approach for Parameter Learning in Variational Models". In: *SIAM Journal on Imaging Sciences* 6.2 (2013), pp. 938–983. DOI: 10.1137/120882706.

[98] Křížek, M. and Strouboulis, T. "How to Generate Local Refinements of Unstructured Tetrahedral Meshes Satisfying a Regularity Ball Condition". In: *Numerical Methods for Partial Differential Equations* 13.2 (1997), pp. 201–214. DOI: https://doi.org/10.1002/(SICI)1098-2426(199703)13:2<201::AID-NUM5>3.0.CO;2-T.

[99] Lepskii, O. V. "On a Problem of Adaptive Estimation in Gaussian White Noise". In: *Theory of Probability & Its Applications* 35.3 (1991), pp. 454–466.

[100] Lions, P. and Mercier, B. "Splitting Algorithms for the Sum of Two Nonlinear Operators". In: *SIAM Journal on Numerical Analysis* 16.6 (1979), pp. 964–979. DOI: 10.1137/0716071.

[101] Megginson, R. E. *An introduction to Banach space theory*. Vol. 183. Graduate Texts in Mathematics. Springer-Verlag, New York, 1998, pp. xx+596. ISBN: 0-387-98431-3. DOI: 10.1007/978-1-4612-0603-3.

[102] Meinhardt, T., Moeller, M., Hazirbas, C., and Cremers, D. "Learning Proximal Operators: Using Denoising Networks for Regularizing Inverse Imaging Problems". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2017, pp. 1799–1808. DOI: 10.1109/ICCV.2017.198.

[103] Meinshausen, N. and Bühlmann, P. "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3 (2006), pp. 1436–1462. DOI: 10.1214/009053606000000281.

[104] Morozov, V. A. “On the solution of functional equations by the method of regularization”. In: *Doklady Akademii Nauk*. Vol. 167. Russian Academy of Sciences. 1966, pp. 510–512.

[105] Mosci, S., Rosasco, L., Santoro, M., Verri, A., and Villa, S. “Solving Structured Sparsity Regularization with Proximal Methods”. In: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science*. 2010, pp. 418–433.

[106] Neubauer, A. “On Nesterov acceleration for Landweber iteration of linear ill-posed problems”. In: *Journal of Inverse and Ill-posed Problems* 25.3 (2017), pp. 381–390. DOI: doi:10.1515/jiip-2016-0060.

[107] Pesquet, J.-C., Repetti, A., Terris, M., and Wiaux, Y. “Learning Maximally Monotone Operators for Image Recovery”. In: *SIAM Journal on Imaging Sciences* 14.3 (2021), pp. 1206–1237. DOI: 10.1137/20M1387961.

[108] Peyré, G. “The Numerical Tours of Signal Processing”. In: *Computing in Science & Engineering* 13.4 (2011), pp. 94–97. DOI: 10.1109/MCSE.2011.71.

[109] Quarteroni, A. *Numerical models for differential problems*. Third. Vol. 16. MS&A. Modeling, Simulation and Applications. Springer, Cham, 2017, pp. xvii+681. ISBN: 978-3-319-49315-2; 978-3-319-49316-9. DOI: 10.1007/978-3-319-49316-9. URL: https://doi.org/10.1007/978-3-319-49316-9.

[110] Rao, V. and Roy, A. “Extreme Lipschitz functions”. In: *Mathematische Annalen* 189 (Mar. 1970), pp. 26–46. DOI: 10.1007/BF01350198.

[111] Reyes, J. C. De los, Schönlieb, C.-B., and Valkonen, T. “Bilevel parameter learning for higher-order total variation regularisation models”. In: *Journal of Mathematical Imaging and Vision* 57.1 (2017), pp. 1–25. DOI: 10.1007/s10851-016-0662-8.

[112] Rice, J. A. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[113] Rolewicz, S. “On extremal points of the unit ball in the Banach space of Lipschitz continuous functions”. In: *J. Austral. Math. Soc. Ser. A* 41.1 (1986), pp. 95–98. ISSN: 0263-6115.

[114] Roy, A. K. “Extreme points and linear isometries of the Banach space of Lipschitz functions”. In: *Canadian J. Math.* 20 (1968), pp. 1150–1164. ISSN: 0008-414X. DOI: 10.4153/CJM-1968-109-9.

[115] Rudin, L. I., Osher, S., and Fatemi, E. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259–268. ISSN: 0167-2789. DOI: https://doi.org/10.1016/0167-2789(92)90242-F.

[116] Ryu, E., Liu, J., Wang, S., Chen, X., Wang, Z., and Yin, W. “Plug-and-Play Methods Provably Converge with Properly Trained Denoisers”. In: *Proceedings of the 36th*

International Conference on Machine Learning. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 5546–5557.

[117] Salzo, S. and Villa, S. "Proximal Gradient Methods for Machine Learning and Imaging". In: *Harmonic and Applied Analysis: From Radon Transforms to Machine Learning*. Springer International Publishing, 2021, pp. 149–244. DOI: 10.1007/978-3-030-86664-8_4.

[118] Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. *Variational methods in imaging*. Vol. 167. Applied Mathematical Sciences. Springer, 2009, pp. xiv+320. ISBN: 978-0-387-30931-6. DOI: 10.1007/978-0-387-69277-7.

[119] Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001. ISBN: 0262194759.

[120] Sherry, F., Benning, M., Reyes, J., Graves, M., Maierhofer, G., Williams, G., Schönlieb, C.-B., and Ehrhardt, M. "Learning the Sampling Pattern for MRI". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4310–4321. DOI: 10.1109/TMI.2020.3017353.

[121] Sherry, F., Celledoni, E., Ehrhardt, M. J., Murari, D., Owren, B., and Schönlieb, C.-B. *Designing Stable Neural Networks using Convex Analysis and ODEs*. 2023. arXiv: 2306.17332 [cs.LG].

[122] Smarzewski, R. "Extreme points of unit balls in Lipschitz function spaces". In: *Proc. Amer. Math. Soc.* 125.5 (1997), pp. 1391–1397. ISSN: 0002-9939. DOI: 10.1090/S0002-9939-97-03866-5.

[123] Sreehari, S., Venkatakrishnan, S. V., Wohlberg, B., Buzzard, G. T., Drummy, L. F., Simmons, J. P., and Bouman, C. A. "Plug-and-Play Priors for Bright Field Electron Tomography and Sparse Interpolation". In: *IEEE Transactions on Computational Imaging* 2.4 (2016), pp. 408–423. DOI: 10.1109/tci.2016.2599778.

[124] Stein, C. M. "Estimation of the Mean of a Multivariate Normal Distribution". In: *The Annals of Statistics* 9.6 (1981), pp. 1135 –1151.

[125] Svaiter, B. F. "On weak convergence of the Douglas-Rachford method". In: *SIAM J. Control Optim.* 49.1 (2011), pp. 280–287. ISSN: 0363-0129. DOI: 10.1137/100788100.

[126] Tautenhahn, U and Hämarik, U. "The use of monotonicity for choosing the regularization parameter in ill-posed problems". In: *Inverse Problems* 15.6 (1999), pp. 1487–1505.

[127] Teodoro, A. M., Bioucas-Dias, J. M., and Figueiredo, M. A. T. "Image restoration and reconstruction using variable splitting and class-adapted image priors". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3518–3522. DOI: 10.1109/ICIP.2016.7533014.

[128] Teodoro, A. M., Bioucas-Dias, J. M., and Figueiredo, M. A. T. "Scene-Adapted Plug-and-Play algorithm with convergence guarantees". In: *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2017, pp. 1–6. DOI: 10.1109/MLSP.2017.8168194.

[129] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

[130] Tikhonov, A. N., Glasko, V. B., and Kriksin, Y. A. "On the question of quasi-optimal choice of a regularized approximation". In: *Doklady Akademii Nauk*. Vol. 248. Russian Academy of Sciences. 1979, pp. 531–535.

[131] Tikhonov, A. N. and Arsenin, V. Y. *Solutions of ill-posed problems*. Scripta Series in Mathematics. V. H. Winston & Sons, Washington, D.C.; John Wiley & Sons, New York-Toronto, Ont.-London, 1977, pp. xiii+258.

[132] Tsybakov, A. "On the best rate of adaptive estimation in some inverse problems". In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 330.9 (2000), pp. 835–840. ISSN: 0764-4442. DOI: https://doi.org/10.1016/S0764-4442(00)00278-0.

[133] Unser, M. "A Unifying Representer Theorem for Inverse Problems and Machine Learning". In: *Foundations of Computational Mathematics* 21 (Sept. 2020), pp. 1–20. DOI: 10.1007/s10208-020-09472-x.

[134] Unser, M., Fageot, J., and Ward, J. P. "Splines are universal solutions of linear inverse problems with generalized TV regularization". In: *SIAM Rev.* 59.4 (2017), pp. 769–793. ISSN: 0036-1445. DOI: 10.1137/16M1061199.

[135] Vainikko, G. "The discrepancy principle for a class of regularization methods". In: *USSR Computational Mathematics and Mathematical Physics* 22.3 (1982), pp. 1–19.

[136] Vandenberghe, L. and Boyd, S. "Semidefinite Programming". In: *SIAM Review* 38.1 (1996), pp. 49–95. DOI: 10.1137/1038003.

[137] Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

[138] Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. "Plug-and-Play priors for model based reconstruction". In: *2013 IEEE Global Conference on Signal and Information Processing*. 2013, pp. 945–948. DOI: 10.1109/GlobalSIP.2013.6737048.

[139] Wahba, G. "Practical Approximate Solutions to Linear Operator Equations When the Data are Noisy". In: *SIAM Journal on Numerical Analysis* 14.4 (1977), pp. 651–667. DOI: 10.1137/0714044.

[140] Weaver, N. *Lipschitz algebras*. Second. World Scientific, 2018, pp. xiv+458. ISBN: 978-981-4740-63-0.

[141] Williams, D. *Probability with Martingales*. Cambridge University Press, 1991. DOI: 10.1017/CBO9780511813658.

[142] Zhang, K., Zuo, W., Gu, S., and Zhang, L. "Learning Deep CNN Denoiser Prior for Image Restoration". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2808–2817. DOI: 10.1109/CVPR.2017.300.