On learning the optimal regularization parameter in inverse problems

OPEN ACCESS
IOP Publishing
Inverse Problems

Inverse Problems **40** (2024) 125004 (39pp) https://doi.org/10.1088/1361-6420/ad8a84

On learning the optimal regularization parameter in inverse problems

Jonathan Chirinos-Rodríguez[1,*], Ernesto De Vito[1], Cesare Molinari[1], Lorenzo Rosasco[2,3,4] and Silvia Villa[1]

[1] MaLGa, DIMA, Dipartimento di Eccellenza 2023-2027, Università degli Studi di Genova, Genoa, Italy
[2] MaLGa, DIBRIS, Università degli Studi di Genova, Genoa, Italy
[3] Center for Brains, Minds and Machines, MIT, Cambridge, MA, United States of America
[4] Istituto Italiano di Tecnologia, Genoa, Italy

E-mail: c.rodriguez@dima.unige.it, ernesto.devito@unige.it, molinari@dima.unige.it, lorenzo.rosasco@unige.it and silvia.villa@unige.it

Received 17 May 2024; revised 9 October 2024
Accepted for publication 23 October 2024
Published 4 November 2024

Abstract

Selecting the best regularization parameter in inverse problems is a classical and yet challenging problem. Recently, data-driven approaches based on supervised learning have become popular to tackle this challenge. These approaches are appealing since they do require less *a priori* knowledge, but their theoretical analysis is limited. In this paper, we propose and study a statistical machine learning approach, based on empirical risk minimization. Our main contribution is a theoretical analysis, showing that, provided with enough data, this approach can reach sharp rates while being essentially adaptive to the noise and smoothness of the problem. Numerical simulations corroborate and illustrate the theoretical findings. Our results are a step towards grounding theoretically data-driven approaches based on supervised learning to inverse problems.

Keywords: supervised learning, inverse problems, stochastic inverse problems, parameter selection methods, cross-validation

* Author to whom any correspondence should be addressed.

1. Introduction

Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ and $(\mathcal{Y}, \langle\cdot,\cdot\rangle_{\mathcal{Y}})$ be real separable Hilbert spaces and $A\colon \mathcal{X} \to \mathcal{Y}$ a forward operator. Given A and a datum $y \in \mathcal{Y}$, the corresponding inverse problem is to find $x^* \in \mathcal{X}$ solving

$$A(x^*) = y.$$

In practice, only perturbed data are typically available, that is

$$\widehat{y} = y + \varepsilon, \qquad \|\varepsilon\|_{\mathcal{Y}} \leqslant \tau,$$

where we considered a deterministic noise model. The above problem is often ill-posed and, in particular, solutions might not depend continuously on the data. Regularization theory provides a principled approach towards finding stable solutions, see e.g. [11, 29]. First, a family of regularization operators is defined for every $\lambda \in (0,+\infty)$: $f_\lambda : \mathcal{Y} \to \mathcal{X}$. A classical choice for such class of operators is to consider solutions of the following variational problem

$$\min_x \ell(Ax, \widehat{y}) + \lambda J(x),$$

for some $\lambda \in (0,+\infty)$. The term $\ell(A\cdot,\widehat{y})$ is known as the *data-fidelity* term, and constraints the solution to remain close to the available measurements. Natural choices for ℓ are the squared distance, $\ell(x,x') = \|x-x'\|_2^2$, common in regression tasks; the hinge loss, $\ell(x,x') = \max\{0, 1-xx'\}$, typical in classification problems, or the Bregman distance with respect to a given function f, $\ell(x,x') = D_f(x,x')$, which arises naturally when considering convex variational problems, see for instance [12, 53]. The function $J\colon \mathcal{X} \to (0,+\infty]$, referred to as the *regularization function*, incorporates prior knowledge about the solution into the problem formulation. We recall that common choices are $J(x) = \|x\|_2^2$, $J(x) = \|x\|_1$, or $J(x) = \mathrm{TV}(x)$. Finally, a choice is specified for the regularization parameter λ. Ideally, for some given discrepancy ℓ, such a choice should allow to optimally control the error $\ell(f_\lambda(\widehat{y}), x^*)$. Classical strategies for choosing the regularization parameter are divided in *a priori*, where $\lambda = \lambda(\tau, x^*)$ and *a posteriori*, where $\lambda = \lambda(\tau)$. *A priori* choices are primarily of theoretical interest. The reason is that they allow to derive sharp error estimates that can be shown to match corresponding lower bounds, see e.g. [29]. However, they are usually impractical since they depend on the unknown solution x^*—or rather on its regularity properties expressed by some smoothness parameters. *A posteriori* choices, such as the classic Morozov discrepancy principle [49] are adaptive to the knowledge of the regularity properties of x^*, but still require the noise level τ. Since in many practical scenarios this information might not be available, a number of alternative strategies have been proposed, including generalized cross-validation [34, 64], quasi-optimality criterion [7, 61], L-curve method [37], and methods based on an estimation of the mean squared error, see e.g. [24] and references therein.

In recent years, data-driven approaches based on supervised learning to solve inverse problems have received much attention since they seem to provide improved results, while circumventing some limitations of classical approaches, see [2] and references therein. The starting point of supervised learning approaches is the assumption that a finite set of pairs of data and exact solutions $(\widehat{y}_1, x_1^*), \ldots, (\widehat{y}_n, x_n^*)$ is available. This *training set* can be used to define, or refine, a regularization strategy to be used on any future datum $\widehat{y}$ for which an exact solution is not known. This perspective has been already considered to provably learn a regularization parameter choice. For example, in [1] a general approach is analyzed to learn a regularizer in Tikhonov-like regularization schemes for linear inverse problems. Indeed, these results can be

adapted to learn the best regularization parameter in some cases. Another learning approach was analyzed in [3], where some convergence results where provided in the case where the forward operator A is unknown. Additionally, an unsupervised learning approach was studied in [23] and [44]. A bilevel optimization perspective is taken in [33], where some theoretical results are also given.

In this paper, we consider one of the most classical machine learning approaches, namely empirical risk minimization (ERM). We study the regularization parameter choice defined by the following problem,

$$\min_{\lambda\in\Lambda}\frac{1}{n}\sum_{i=1}^{n}\ell\left(f_\lambda\left(\widehat{y}_i\right),x_i^*\right)$$

where Λ is a suitable finite set of candidate values for λ. Our main contribution is characterizing the error performance of the above approach. Towards this end, we consider a statistical inverse problems framework and tackle the question with the aid of tools from statistical learning theory [20, 63]. The theory of ERM is well established, and the class of functions we need to consider is parameterized by just one parameter—the regularization parameter. However, the dependence on such a parameter is non linear/nonsmooth and possibly hard to characterize, making the application of standard ERM results not straightforward. To circumvent this challenge we borrow ideas from the literature of model selection in statistics and machine learning [26, 36] and in particular, we adapt ideas from [15]. Our theoretical analysis shows that the ERM approach for learning the best regularization parameter can essentially achieve the same performance of an ideal a-priori choice. As we will see, this is true up to an error term, which decreases fast with the size of the training set. General results are illustrated considering several inverse problems scenarios. In particular, we discuss the case of linear inverse problems with spectral regularization methods and Tikhonov regularization with general convex regularizers in sections 3 and 5 respectively. Also, we consider non-linear inverse problems in Hilbert spaces and the corresponding Tikhonov regularization in section 4. The theoretical results are illustrated through numerical experiments in section 6 for spectral regularization methods and sparsity promoting norms.

Notation

In the following, given an element x in a Hilbert space $(\mathcal{X},\langle\cdot,\cdot\rangle_{\mathcal{X}})$, $\|x\|_{\mathcal{X}}$ denotes the corresponding norm, i.e. $\|x\|_{\mathcal{X}}=\sqrt{\langle x,x\rangle_{\mathcal{X}}}$. Moreover, if $(\mathcal{Y},\langle\cdot,\cdot\rangle_{\mathcal{Y}})$ is also a Hilbert space, given a linear and bounded operator $A\colon\mathcal{X}\to\mathcal{Y}$, we denote by A^* its adjoint operator and, if A is injective, by A^{-1} its inverse. With $\|\cdot\|_{\mathrm{op}}$ we denote the operator norm.

We assume that $(\Omega,\mathcal{A},P)$ is a probability space. For classical notions from integration theory in infinite dimensional space, we refer to [27]; in particular, measurability [27, chapter 1, definition 3] (since we consider separable Hilbert spaces, note that the various notions of measurability are equivalent, see [27, chapter 1, proposition 22]), the Bochner integral [27, chapter 1, definition 35] and the conditional expectation [27, chapter 2, definition 38]. Random variables will be denoted in capital letters. Finally, the subdifferential of a proper, convex and lower semicontinuous function $f\colon\mathcal{X}\mapsto\mathbb{R}\cup\{+\infty\}$ is the set-valued operator $\partial f\colon\mathcal{X}\to 2^{\mathcal{X}}$ defined by

$$x\mapsto\{u\in\mathcal{X}\mid\text{for every } y\in\mathcal{X},\, f(x)+\langle y-x,\,u\rangle_{\mathcal{X}}\leqslant f(y)\}.$$

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 (Y,X) be a pair of random variables with values in $\mathcal{Y}\times\mathcal{X}$, with unknown joint probability distribution μ on $\mathcal{Y}\times\mathcal{X}$, and let $(Y_i,X_i)_{i=1}^n$ be n identical and independent copies of (Y,X). For $\lambda\in(0,+\infty)$, let $f_\lambda:\mathcal{Y}\to\mathcal{X}$ be a family of measurable functions parametrized by λ. Given a measurable loss function $\ell:\mathcal{X}\times\mathcal{X}\to[0,+\infty)$, for all measurable functions $f:\mathcal{Y}\to\mathcal{X}$ consider the expected risk

$$L(f):=\mathbb{E}\left[\ell\left(f(Y),X\right)\right]=\int_{\mathcal{Y}\times\mathcal{X}}\ell\left(f(y),x\right)\,\mathrm{d}\mu\left(y,x\right).$$

and the empirical risk

$$\widehat{L}(f)=\frac{1}{n}\sum_{i=1}^{n}\ell\left(f(Y_i),X_i\right).$$

Moreover, for some $N\in\mathbb{N}$, define Λ, the finite grid of regularization parameters, as

$$\Lambda=\{\lambda_1,\dots,\lambda_N\} \tag{2.1}$$

with $0<\lambda_1\leq\lambda_2\cdots\leq\lambda_N<\infty$. Considering the empirical risk minimization (ERM), we let

$$\widehat{\lambda}\in\operatorname*{arg\,min}_{\lambda\in\Lambda}\widehat{L}(f_\lambda). \tag{2.2}$$

Remark 1. The approach we propose, namely learning a unique parameter $\widehat{\lambda}$ for all the data, is natural in the setting of our problem, where the data are generated by a unique model. Another possible algorithm would be to first transform the dataset (Y_i,X_i) to (Y_i,λ_i), where each λ_i is the best parameter for the pair (Y_i,X_i); and then to learn the map from the data to the best parameter, namely $\lambda=\lambda(Y)$. In the case where the Hilbert space $\mathcal{Y}$ is a functional space, this may be done by polynomial or nonlinear functional data regression; see [41, 47, 58]. This study would require a considerable modification of our analysis, and it is left for future work.

We aim at characterizing $L(f_{\widehat{\lambda}})$, namely the expected risk corresponding to the regularization parameter chosen accordingly to the rule in (2.2). An idea would be to compare it directly to $\min_{\lambda\in(0,+\infty)}L(f_\lambda)$. Instead, as discussed next, we assume that a suitable error bound $\min_\lambda L(f_\lambda)\leq U(\lambda_*)$ is available, and then we compare $L(f_{\widehat{\lambda}})$ to $U(\lambda_*)$. Next, we list and comment the main assumptions.

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 and 4, and Bregman divergences associated with convex functionals, see section 5. While none one of these examples are bounded, since we will assume X to be almost surely bounded, a bounded loss will be obtained by composing the discrepancy with suitable truncation operators.

Assumption 2. There exists $U:(0,+\infty)\to(0,+\infty)$ such that, for every $\lambda\in(0,+\infty)$,

$$L(f_\lambda)\leq U(\lambda). \tag{2.3}$$

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

$$\lambda_* \in \operatorname*{arg\,min}_{\lambda \in (0,+\infty)} U(\lambda). \tag{2.4}$$

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

$$U(q\lambda_*) \leq C(q)\, U(\lambda_*). \tag{2.5}$$

The main reason for the above assumption is to avoid to require the Lipschitzianity of f_λ with respect to λ, which is typically assumed in classic studies of ERM, see e.g. [20]. This assumption might seem unusual for a learning setting but, as shown in sections 3–5, it is naturally satisfied in the context of inverse problems, where U is often called profile function, see e.g. [40]. Moreover, this is the usual strategy to design *a priori* choices of the regularization parameter, since in this latter setting it is often possible to derive tight bounds, in the sense that the two quantities, $L(f_\lambda)$ and $U(\lambda)$, have the same behavior with respect to λ and the noise level, and therefore $\min_{\lambda\in(0,+\infty)} L(f_\lambda)$ is comparable to $U(\lambda_*)$ (see e.g. [29, chapter 4]). To derive such optimal bounds, it is well-known that U should depend on the degrees of smoothness of the underlying exact solution. Related to this, note that the minimizer λ_* of U is relevant only if the upper bound U in (2.3) is sharp. We refer to the subsequent assumption 5 and theorem 2 for a concrete example.

We make one last assumption on how *large* is the set of candidate values Λ.

Assumption 3. Let Λ be defined as in (2.1). Assume that

$$\lambda_* \in [\lambda_1, \lambda_N] \tag{2.6}$$

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

$$Q = \left(\frac{\lambda_N}{\lambda_1}\right)^{\frac{1}{N-1}}. \tag{2.7}$$

The above assumption states that we can choose a sufficiently large interval for our discretization so that the best regularization parameter λ_* in (2.4) according to the available upper bound U 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 $\widehat{\lambda}$ achieves an error close to that of λ_*.

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

$$L\left(f_{\widehat{\lambda}}\right) \leq 2C(Q)\, U(\lambda_*) + \frac{13M}{2n} \log \frac{2N}{\eta}.$$

The above result shows that $\widehat{\lambda}$ 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 $1/n$.

From the expression (2.7), 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. In the following, we provide concrete examples in the context of inverse problems that illustrate and instantiate the above results.

We first provide the proof of theorem 1.

2.1. 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 $\widehat{\lambda}$ to that of

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

which is the ideal regularization parameter choice when restricting the search to Λ. Indeed, we prove in lemma 1 that with high probability

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

for some constant $c>0$. Then, in lemma 2 we show that there exists $1\leq q<Q$ such that $q\lambda_* \in \Lambda$ and so

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

Combining the above results and using condition (2.5), we get with high probability that

$$L\left(f_{\widehat{\lambda}}\right) \lesssim 2L\left(f_{q\lambda_*}\right) + \frac{\log(2N)}{n} \lesssim 2C(Q)\,U(\lambda_*) + \frac{\log(2N)}{n},$$

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

Lemma 1. *Under assumption 1, for $\eta\in(0,1)$ we have that, with probability at least $1-\eta$,*

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

The proof is based on a classic union bound argument and the following concentration inequality, see proposition 11 in [15], which we report for simplicity.

Proposition 1. *Let $Z_1,\dots Z_n$ be a sequence of i.i.d. real random variables with mean μ, such that $|Z_i|\leq B$ a.s. and $\mathbb{E}[|Z_i-\mu|^2]\leq\sigma^2$. Then for all $\alpha,\varepsilon>0$*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq \varepsilon + \alpha\sigma^2\right\} \leq 2e^{-\frac{6n\alpha\varepsilon}{3+4\alpha B}}. \tag{2.8}$$

The idea of the proof is adapted from [15].

Proof of lemma 1. For $\lambda\in\Lambda$, let $Z_i(\lambda)=\ell(f_\lambda(Y_i),X_i)$, $i=1,\dots,n$. Then,

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

and

$$\mathbb{E}[Z_i(\lambda)] = L(f_\lambda).$$

Moreover, since the loss is bounded by assumption 1, then $Z_i(\lambda)\leq M$ and this implies

$$\mathbb{E}\left[|Z_i(\lambda)|^2\right] = \mathbb{E}\left[\ell(f_\lambda(Y_i),X_i)\,\ell(f_\lambda(Y_i),X_i)\right] \leq ML(f_\lambda).$$

Now, we apply (2.8) with $B = M$ and, by recalling that $\mathbb{E}[|Z_i(\lambda) - \mathbb{E}[Z_i(\lambda)]|^2] \leqslant \mathbb{E}[|Z_i(\lambda)|^2]$, we fix $\sigma^2 = ML(f_\lambda)$. We then get, for each $\lambda \in \Lambda$ and for all $\alpha, \varepsilon > 0$,

$$P\left\{|\widehat{L}(f_\lambda) - L(f_\lambda)| \geq \varepsilon + \alpha ML(f_\lambda)\right\} \leq 2e^{-\frac{6n\alpha\varepsilon}{3+4\alpha M}}.$$

Moreover, since the probability of a union of events is less or equal than the sum of their probabilities, we have that, for all $\alpha, \varepsilon > 0$,

$$P\left(\bigcup_{\lambda\in\Lambda}\left\{|\widehat{L}(f_\lambda) - L(f_\lambda)| \geq \varepsilon + \alpha ML(f_\lambda)\right\}\right) \leq 2|\Lambda|e^{-\frac{6n\alpha\varepsilon}{3+4\alpha M}}.$$

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

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

and

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

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

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

The result follows by plugging in the expression of ε and by recalling that $|\Lambda| = N$. □

Note that the above result holds under minimal assumptions. Indeed, the structural assumptions we introduced are used to prove the following lemma.

Lemma 2. *Let assumptions 2 and 3 be satisfied and consider λ_* as in assumption 2. Then, there exists $1 \leqslant q \leqslant Q$ such that $q\lambda_* \in \Lambda$ and so*

$$L\left(f_{\lambda_\Lambda}\right) \leqslant L\left(f_{q\lambda_*}\right).$$

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

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

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

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

so that

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

Finally, by the definition of λ_Λ, we get

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

concluding the proof. □

We add one final remark.

Remark 2 (comparison with union bound combined with Hoeffding). A slightly different estimate can be obtained using a union bound argument and a different concentration result, namely Hoeffiding inequality (2.10). Indeed, if we let $\eta \in (0,1)$, the following bound holds with probability at least $1-\eta$:

$$L\left(f_{\widehat{\lambda}}\right) \leq L\left(f_{\lambda_\Lambda}\right) + 2\sqrt{\frac{M}{n}\log\frac{2N}{\eta}}. \tag{2.9}$$

Compared to the estimate obtained in lemma 1, 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. By using inequality (2.9) in place of lemma 1, it is possible to derive a result analogous to theorem 1. Again, this allows to improve the bound by a factor of 2 while achieving a much worse dependence on the number of data points. For completeness, we report the proof of inequality (2.9), which is based on Hoeffding's inequality:

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq \varepsilon\right\} \leq 2e^{-\frac{n\varepsilon^2}{2B}}, \tag{2.10}$$

where B is an upper bound on the random variables Z_i, as in proposition 1. Indeed, by adding the subtracting the empirical risks we have that,

$$\begin{aligned} L\left(f_{\widehat{\lambda}}\right) - L\left(f_{\lambda_\Lambda}\right) &= L\left(f_{\widehat{\lambda}}\right) - \widehat{L}\left(f_{\widehat{\lambda}}\right) + \widehat{L}\left(f_{\widehat{\lambda}}\right) - \widehat{L}\left(f_{\lambda_\Lambda}\right) + \widehat{L}\left(f_{\lambda_\Lambda}\right) - L\left(f_{\lambda_\Lambda}\right) \\ &\leq L\left(f_{\widehat{\lambda}}\right) - \widehat{L}\left(f_{\widehat{\lambda}}\right) + \widehat{L}\left(f_{\lambda_\Lambda}\right) - L\left(f_{\lambda_\Lambda}\right) \\ &\leq 2\sup_{\lambda\in\Lambda} |L(f_\lambda) - \widehat{L}(f_\lambda)|, \end{aligned}$$

using the fact that the term $\widehat{L}(f_{\widehat{\lambda}}) - \widehat{L}(f_{\lambda_\Lambda})$ is negative by definition of $\widehat{\lambda}$. Then, combining (2.10) and a union bound, we get

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

Inequality (2.9) follows by setting $\eta = 2Ne^{-(n\varepsilon^2)/M}$ and deriving the expression for ε.

3. Spectral regularization for linear inverse problems

In this section, we illustrate the general results considering spectral regularization methods 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 so that assumption 2 holds. Let $\mathcal{X}, \mathcal{Y}$ be real and separable Hilbert spaces, let $A\colon \mathcal{X} \to \mathcal{Y}$ be a linear and bounded operator and assume that $\|A\|_{\text{op}} \leq 1$. Then, let X, ε be a pair of random variables with values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, and

$$Y = AX + \varepsilon, \quad \text{a.s.} \tag{3.1}$$

Here we recall some classical linear inverse problems. A large list of examples can be found in [56].

Example 3.1 (denoising). Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^{N\times N}$, $1 \leqslant N < +\infty$. The image denoising problem consists in recovering a clean image $x^* \in \mathbb{R}^{N\times N}$ from noisy measurements $Y \in \mathbb{R}^{N\times N}$. It writes as

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

where $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$, $\tau > 0$, is some additive Gaussian noise. In this case, the forward operator A coincides with the identity map.

Example 3.2 (deconvolution). Let $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{R}^N)$, $1 \leqslant N < +\infty$. Given $f \in L^1(\mathbb{R}^N)$, we consider A as the linear convolution operator,

$$x \mapsto Ax := f * x; \quad (f*x)(t) := \int_{\mathbb{R}^N} f(s-t)\, x(s)\, \mathrm{d}s.$$

Then, the general deconvolution problem [29] consists in recovering a ground-truth term $x^* \in L^2(\mathbb{R}^N)$, from noisy measurements $Y \in L^2(\mathbb{R}^N)$. It reads as

$$Y = f * x^* + \varepsilon,$$

where $\varepsilon \in L^2(\mathbb{R}^N)$ models the possible presence of noise.

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

Assumption 4. We assume that

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

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

$$\mathbb{E}\left[\|\varepsilon\|_{\mathcal{Y}}^2 | X\right] \leq \tau^2.$$

The above condition is a simple and natural stochastic extension of the classical bounded variance assumption. We also assume that X satisfies the following stochastic extension of the classical Hölder source conditions [29].

Assumption 5. The random variable X is such that $\|X\|_{\mathcal{X}} \leqslant 1$ a.s. and there exist a random variable Z with values in $\mathcal{X}$, and β, $s > 0$ such that,

$$X = (A^*A)^s Z,$$

and

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

In this setting, a corresponding Tikhonov regularized estimator is defined as

$$X_\lambda = \operatorname*{arg\,min}_{x\in\mathcal{X}} \|Ax - Y\|_{\mathcal{Y}}^2 + \lambda \|x\|_{\mathcal{X}}^2. \tag{3.2}$$

Clearly, $X_\lambda = X_\lambda(Y)$, but we omit the dependence for conciseness. A more explicit expression is given by

$$X_\lambda = (A^*A + \lambda I)^{-1} A^* Y. \tag{3.3}$$

More generally, the class of spectral regularization methods is given by

$$X_\lambda = g_\lambda(A^*A)A^*Y, \tag{3.4}$$

defined by a suitable function $g_\lambda : (0,1] \to \mathbb{R}$ using spectral calculus. Note that the above expression ensures that X_λ is measurable, since it is the image of a linear operator applied to Y.

The following assumption characterizes the key properties required on g_λ.

Assumption 6. There exists a 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,

$$\sup_{\sigma\in(0,1]} |(1 - g_\lambda(\sigma)\sigma)\sigma^s| \leq C_2\lambda^\alpha. \tag{3.5}$$

Assumption 6 is satisfied by a large class of filter functions such as Tikhonov regularization, the Landweber iteration, that is gradient descent on the least squares error, spectral cut-off, heavy-ball methods and the ν-method [29], or Nesterov acceleration [51]. We add some remarks regarding this assumption.

Note that the first assumption implies that the norm of the regularization operator $g_\lambda(A^*A)A^*$ is always bounded and controlled by λ. The second 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. For many spectral regularization methods, there is $q > 0$ such that

$$\sup_{\sigma\in(0,1]} |(1 - g_\lambda(\sigma)\sigma)\sigma^\nu| \leq C_2\lambda^\nu, \quad \text{for every } \nu \leqslant q.$$

The number q is called qualification parameter and depends on the regularization method g_λ; see [6]. Therefore, assumption 6 is satisfied for $\alpha = \min(q, s)$. Both of the above assumptions allow us to derive suitable error bounds and corresponding *a priori* regularization parameter choice, extending classical results in the deterministic setting.

Theorem 2. *Under assumptions 4–6, the following bound holds for all* $\lambda \in (0, +\infty)$,

$$\mathbb{E}\left[\|X_\lambda - X\|_{\mathcal{X}}^2\right] \leq C_1^2\frac{\tau^2}{\lambda} + C_2^2\beta^2\lambda^{2\alpha}. \tag{3.6}$$

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

$$\mathbb{E}\left[\|X_{\lambda_*} - X\|_{\mathcal{X}}^2\right] \leq (2\alpha+1)\left[\left(\frac{C_1^2}{2\alpha}\right)^{2\alpha} C_2^2\right]^{1/(2\alpha+1)} \left(\frac{\tau^{2\alpha}}{\beta}\right)^{2/(2\alpha+1)}. \tag{3.7}$$

Proof. To relate X_λ and X, we observe that

$$\mathbb{E}[X_\lambda|X] = \mathbb{E}[g_\lambda(A^*A)A^*Y|X] = \mathbb{E}[g_\lambda(A^*A)A^*AX|X] = g_\lambda(A^*A)A^*AX,$$

where we used the definition of Y and assumption 4. Then, we can decompose the deviation of X_λ to X as

$$\begin{aligned} X_\lambda - X &= X_\lambda - \mathbb{E}[X_\lambda|X] + \mathbb{E}[X_\lambda|X] - X \\ &= g_\lambda(A^*A)A^*(Y - AX) + (g_\lambda(A^*A)A^*A - I)X \\ &= g_\lambda(A^*A)A^*\varepsilon + (g_\lambda(A^*A)A^*A - I)(A^*A)^s Z. \end{aligned} \tag{3.8}$$

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

$$\|g_\lambda(A^*A)A^*\|_{\text{op}} \leq \frac{C_1}{\sqrt{\lambda}}, \quad \|(I - g_\lambda(A^*A)A^*A)(A^*A)^s\|_{\text{op}} \leq C_2\lambda^\alpha, \tag{3.9}$$

see e.g. [29]. If we take the expectation of the squared norm in (3.8) and develop the square, we get

$$\mathbb{E}\left[\|X_\lambda - X\|_{\mathcal{X}}^2\right] = \mathbb{E}\left[\|g_\lambda(A^*A)A^*\varepsilon\|_{\mathcal{Y}}^2\right] + \mathbb{E}\left[\|(g_\lambda(A^*A)A^*A - I)X\|_{\mathcal{X}}^2\right],$$

since, by assumption 4, we have

$$\begin{aligned} &\mathbb{E}\left[\langle g_\lambda(A^*A)A^*\varepsilon, (g_\lambda(A^*A)A^*A - I)X\rangle_{\mathcal{X}}\right] \\ &\quad = \mathbb{E}\left[\langle g_\lambda(A^*A)A^*\mathbb{E}[\varepsilon|X], (g_\lambda(A^*A)A^*A - I)X\rangle\right] = 0. \end{aligned}$$

Then, using again assumptions 4–6 as well as the estimates (3.9), we derive

$$\begin{aligned} \mathbb{E}\left[\|X_\lambda - X\|_{\mathcal{X}}^2\right] &\leq \|g_\lambda(A^*A)A^*\|_{\text{op}}^2 \mathbb{E}\left[\|\varepsilon\|_{\mathcal{Y}}^2\right] + \|(I - g_\lambda(A^*A)A^*A)(A^*A)^s\|_{\text{op}}^2 \mathbb{E}\left[\|Z\|_{\mathcal{X}}^2\right] \\ &\leq C_1^2\frac{\tau^2}{\lambda} + C_2^2\beta^2\lambda^{2\alpha}. \end{aligned}$$

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

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

which is the inequality that we were aiming for. □

The expression given in (3.6) 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 (3.6), we obtain the bound in (3.7). These results are analogous to the ones usually obtained in the deterministic setting (see for instance corollary 4.4 in [29]), and are known to be optimal in the sense of definition 3.17 in [29].

Next, we show that the regularization parameter on the grid learned from data, namely $\widehat{\lambda}$ defined in (2.2), achieves a similar performance to the one of λ_*. Indeed, with the aid of the previous results, and in combination with theorem 1, we obtain a sharp error bound for the

regularized solution with $\widehat{\lambda}$. Toward this end, we consider the truncation operator $T:\mathcal{X}\to\mathcal{X}$ defined for all $x\in\mathcal{X}$ as

$$Tx=\begin{cases} x, & \|x\|_{\mathcal{X}}\leq 1,\\ \dfrac{x}{\|x\|_{\mathcal{X}}}, & \|x\|_{\mathcal{X}}>1. \end{cases} \tag{3.10}$$

To apply the result in section 2, we consider the loss function defined, for every $(x,x')\in\mathcal{X}^2$, as

$$\ell(x,x')=\|Tx-Tx'\|_{\mathcal{X}}^2. \tag{3.11}$$

Then, the corresponding expected risk is, for every measurable function f,

$$L(f)=\mathbb{E}\left[\|Tf(Y)-TX\|_{\mathcal{X}}^2\right]. \tag{3.12}$$

Under assumption 3, for every $\lambda\in(0,+\infty)$ let $f_\lambda(Y)=X_\lambda$, where X_λ is defined as in (3.4). Now, we next study the error obtained in this context by choosing λ with ERM.

Consider a finite set of independent and identical copies (Y_i,X_i), $i=1,\dots,n$, of the pair (Y,X) distributed as in (3.1) and let $X_\lambda^i:=f_\lambda(Y_i)$. Finally, the associated ERM is given by

$$\widehat{\lambda}\in\arg\min_{\lambda\in\Lambda}\frac{1}{n}\sum_{i=1}^{n}\|TX_\lambda^i-X_i\|_{\mathcal{X}}^2, \tag{3.13}$$

where we used that $X_i=TX_i$ a.s.. since $\|X\|_{\mathcal{X}}\leqslant 1$ almost surely.

The following corollary provides the desired error estimates.

Corollary 1. *Let assumption 3 be satisfied with λ_* as in theorem 2. Suppose that assumptions 4–6 hold, and choose the loss as in (3.11). Let $\eta\in(0,1)$. Then, with probability at least $1-\eta$,*

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

In this setting, assumption 1 is trivially satisfied. The proof will therefore consist in verifying that also assumption 2 holds, so that theorem 1 can be applied.

Proof. In this case, assumption 1 is satisfied with $M=4$. We just need to show that assumption 2 is satisfied for $f_\lambda=X_\lambda$ and L defined as in (3.12). Since T is a projection, it is 1-Lipschitz. Then, for all measurable functions $f:\mathcal{Y}\to\mathcal{X}$,

$$L(f)=\mathbb{E}\left[\|Tf(Y)-TX\|_{\mathcal{X}}^2\right]\leq\mathbb{E}\left[\|f(Y)-X\|_{\mathcal{X}}^2\right].$$

Then, if we define $U(\lambda)$ as the right hand side of equation (3.6), (2.3) holds. In addition, λ_* defined as in theorem 2 is the minimizer of U. 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 (3.7), we derive, for any $q\in[1,+\infty)$, that

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

Hence, assumption 2 is satisfied. The result follows by applying theorem 1. □

Corollary 1 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 $\widehat{\lambda}$, 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. (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 $h \in \mathcal{X}$. 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 X, 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 assumptions 5 and 6, making a direct comparison 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 [15] for the cross-validation method.

4. Tikhonov regularization for non linear inverse problems

Next, we consider the problem of selecting the regularization parameter for Tikhonov regularization in the setting of non linear inverse problems [29]. Let $\mathcal{X}$, $\mathcal{Y}$ be real and separable Hilbert spaces, and $A : \mathrm{dom}(A) \subseteq \mathcal{X} \to \mathcal{Y}$ be a (non linear) operator whose domain has nonempty interior. Let X, ε be a pair of random variables with values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, and let

$$Y = A(X) + \varepsilon, \quad \text{a.s.} \tag{4.1}$$

with $X \in \mathrm{int}(\mathrm{dom}(A))$ almost surely. First, we provide some classic examples of non linear inverse problems.

Example 4.1 (phase recovery). Let $\mathcal{X} = \mathbb{C}^p$, $1 \leqslant p < +\infty$, $\mathcal{Y} = \mathbb{C}^n$, $1 \leqslant n < +\infty$, and let $B \in \mathbb{C}^{p\times n}$ be a linear operator. The phase recovery problem [32, 65] consists in recovering a signal $x^* \in \mathcal{X}$ by having access only to noisy measurements of its amplitude. In this case, the operator A is given by

$$A(x) = |Bx|^2.$$

Example 4.2 (electrical impedance tomography). The so-called Electrical Impedance Tomography problem [14, 56, 59] is defined as follows. Let $\Omega \subset \mathbb{R}^n$, $1 \leqslant n < \infty$, be a bounded domain describing a conducting body. Let $\gamma \in L^\infty(\Omega)$ be a positive function, typically known as conductivity. Consider the differential operator

$$L_\gamma x := \mathrm{div}(\gamma \nabla x).$$

Now, given $g \in L^2(\partial\Omega)$ defined on the boundary of Ω, we aim at finding potentials $x \in H^1(\Omega)$ solving the Dirichlet problem

$$\begin{cases} L_\gamma x = 0 & \text{in } \Omega, \\ x|_{\partial\Omega} = g. \end{cases}$$

For every γ, it is possible to consider the Dirichlet-to-Neumann (or current-to-voltage) map, defined as

$$\Lambda_\gamma : x \in H^{1/2}(\partial\Omega) \mapsto \gamma \frac{\partial x}{\partial \nu} \in H^{-1/2}(\partial\Omega),$$

where ν denotes the outer normal to $\partial\Omega$, and that associates each solution x of the Dirichlet problem to its trace $\gamma\, \partial x/\partial\nu$. In this setting, the non linear inverse problem consists in recovering γ by having access to the Dirichlet-to-Neumann map Λ_γ.

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

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

$$\mathbb{E}\left[\|\varepsilon\|_{\mathcal{Y}}^2 | X\right] \leq \tau^2 \quad \text{a.s.}$$

Using Jensen's inequality for the conditional expectation [66, 9.7 (h)], we derive from the previous assumption that

$$\mathbb{E}\left[\|\varepsilon\|_{\mathcal{Y}} | X\right] \leq \tau \text{a.s.} \tag{4.2}$$

Next we impose fairly standard conditions on the operator A.

Assumption 8. The operator $A : \mathrm{dom}(A) \to \mathcal{Y}$ 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 x and $x' \in \mathrm{int}(\mathrm{dom}(A))$,

$$\|A'(x) - A'(x')\|_{\mathrm{op}} \leq C_0 \|x - x'\|_{\mathcal{X}}. \tag{4.3}$$

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

$$\|A(x') - A(x) - A'(x)(x' - x)\|_{\mathcal{Y}} \leq \frac{C_0}{2} \|x' - x\|_{\mathcal{X}}^2,$$

so that, by the triangle inequality,

$$\|A'(x)(x' - x)\|_{\mathcal{Y}} \leq \|A(x') - A(x)\|_{\mathcal{Y}} + \frac{C_0}{2} \|x' - x\|_{\mathcal{X}}^2. \tag{4.4}$$

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 [19]. For non linear 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 X_0 with values in $\mathcal{X}$. In addition, the set of minimizers

$$\underset{x \in \mathrm{dom}(A)}{\arg\min} \|A(x) - Y(\omega)\|_{\mathcal{Y}}^2 + \lambda \|x - X_0(\omega)\|_{\mathcal{X}}^2$$

is nonempty for every $\omega \in \Omega$ thanks to assumption 8, see [19, theorem 10.1]. A corresponding Tikhonov regularized estimator is a random variable X_λ defined by setting, for almost all $\omega \in \Omega$

$$X_\lambda(\omega) \in \underset{x \in \mathrm{dom}(A)}{\arg\min} \|A(x) - Y(\omega)\|_{\mathcal{Y}}^2 + \lambda \|x - X_0(\omega)\|_{\mathcal{X}}^2. \tag{4.5}$$

Note that X_λ depends on Y and X_0, but we will omit this dependence for the sake of simplicity. The existence of a random variable X_λ taking values in the set of minimizers is ensured under some additional assumptions, see e.g. Filippov's Implicit function theorem [39, 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 X is such that $\|X - X_0\|_{\mathcal{X}} \leqslant 1$ and, under assumption 8, there exists a random variable Z with values in $\mathcal{Y}$, $\beta > 0$ such that almost surely

$$X - X_0 = A'(X)^* Z,$$

and

$$\|Z\|_{\mathcal{Y}} \leq \beta \text{a.s.}, \quad \text{with} \quad \beta C_0 < 1,$$

where C_0 is the Lipschitz constant introduced in assumption 8.

The above assumption can be seen as a non linear version of the source condition considered in assumption 5 (for $s = 1$). The next example is taken from [45, section 4], where the authors show that assumptions 8 and 9 are satisfied.

Example 4.3. Consider the boundary value problem

$$\begin{cases} -(\gamma x')' + cx = f & \text{in } (0,1), \\ x(0) = x(1) = 0, \end{cases}$$

where $f \in L^2(0,1)$, $f \neq 0$, and $c \in L^2(0,1)$ is nonnegative almost everywhere. The non linear operator $A\colon \mathrm{dom} A \to L^2(0,1)$ is the parameter-to-solution map, associating to γ the correspondent solution x. In this case, the ill-posed inverse problem consists in determining the coefficient γ from noisy measurements of x.

Other examples in which the source condition in assumption 9 is satisfied can be found in [30] in a deterministic setting, and in [21, section 7], where the authors study parameter identification problems for stochastic PDEs. Further generalizations of this assumption have also been provided, see [31, 43, 46]. On the other hand, the second part of assumption 9 is known as smallness condition, and generally it is hard to verify in practice [29, 30]. For this reason, several works have tried to either drop this condition or relax it, see for instance [21, 30, 42].

In the next result, which is analogous to theorem 2, we derive a bound on the error of the Tikhonov regularized solution, leading to *a priori* parameter choices.

Theorem 3. *Suppose that assumptions 7–9 are satisfied. Then the following bound holds: for all* $\lambda \in (0, +\infty)$,

$$\mathbb{E}\left[\|X_\lambda - X\|_{\mathcal{X}}^2\right] \leq \frac{(\tau + \beta\lambda)^2}{(1 - \beta C_0)\lambda}. \tag{4.6}$$

In particular, setting $\lambda_* = \tau/\beta$,

$$\mathbb{E}\left[\|X_{\lambda_*} - X\|_{\mathcal{X}}^2\right] \leq 4\,(1-\beta C_0)^{-1}\,\tau\beta.$$

The proof is a modification of the one in the deterministic setting, see e.g. [19, 29].

Proof. The expressions below are all intended to hold almost surely. By definition of X_λ, X and ε, it follows that

$$\begin{aligned}\|A\,(X_\lambda) - Y\|_{\mathcal{Y}}^2 + \lambda\|X_\lambda - X_0\|_{\mathcal{X}}^2 &\leq \|A\,(X) - Y\|_{\mathcal{Y}}^2 + \lambda\|X - X_0\|_{\mathcal{X}}^2 \\ &= \|\varepsilon\|_{\mathcal{Y}}^2 + \lambda\|X - X_0\|_{\mathcal{X}}^2.\end{aligned} \tag{4.7}$$

Since

$$\|X_\lambda - X_0\|_{\mathcal{X}}^2 = \|X_\lambda - X\|_{\mathcal{X}}^2 + \|X - X_0\|_{\mathcal{X}}^2 + 2\langle X_\lambda - X, X - X_0\rangle_{\mathcal{X}}, \tag{4.8}$$

inequality (4.7) implies

$$\|A\,(X_\lambda) - Y\|_{\mathcal{Y}}^2 + \lambda\|X_\lambda - X\|_{\mathcal{X}}^2 \leq \|\varepsilon\|_{\mathcal{Y}}^2 - 2\lambda\langle X_\lambda - X, X - X_0\rangle_{\mathcal{X}}.$$

Then, assumption 9 and Cauchy–Schwartz inequality yield

$$\|A\,(X_\lambda) - Y\|_{\mathcal{Y}}^2 + \lambda\|X_\lambda - X\|_{\mathcal{X}}^2 \leq \|\varepsilon\|_{\mathcal{Y}}^2 + 2\lambda\|A'\,(X)\,(X_\lambda - X)\,\|_{\mathcal{Y}}\|Z\|_{\mathcal{Y}}. \tag{4.9}$$

Since $X \in \mathrm{int}(\mathrm{dom}(A))$ and $X_\lambda \in \mathrm{dom}(A)$, and $\mathrm{dom}(A)$ is convex by assumption, inequality (4.4) with $x = X$ and $x' = X_\lambda$ yields

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

so that, by adding and subtracting Y in the first term of the right hand side, we obtain

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

Plugging the above inequality into (4.9), we get

$$\begin{aligned}\|A\,(X_\lambda) - Y\|_{\mathcal{Y}}^2 + \lambda\|X_\lambda - X\|_{\mathcal{X}}^2 \leq \|\varepsilon\|_{\mathcal{Y}}^2 + 2\lambda\|Z\|_{\mathcal{Y}}\,\big(&\|A\,(X_\lambda) - Y\|_{\mathcal{Y}} \\ &+\|\varepsilon\|_{\mathcal{Y}} + \frac{C_0}{2}\|X_\lambda - X\|_{\mathcal{X}}^2\Big).\end{aligned}$$

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

$$\begin{aligned}\left(\|A\,(X_\lambda) - Y\|_{\mathcal{Y}} - \lambda\|Z\|_{\mathcal{Y}}\right)^2 + \lambda\|X_\lambda - X\|_{\mathcal{X}}^2 \leqslant \|\varepsilon\|_{\mathcal{Y}}^2 + 2\lambda\|Z\|_{\mathcal{Y}}\,\big(&\|\varepsilon\|_{\mathcal{Y}} \\ &+\frac{C_0}{2}\|X_\lambda - X\|_{\mathcal{X}}^2\Big) + \lambda^2\|Z\|_{\mathcal{Y}}^2.\end{aligned}$$

Next, we take expectations on both sides. First, recall that assumption 7 implies (4.2), i.e. $\mathbb{E}[\|\varepsilon\|_{\mathcal{Y}}] \leqslant \tau$ and therefore, with assumption 9,

$$\mathbb{E}\left[\|Z\|_{\mathcal{Y}}\|\varepsilon\|_{\mathcal{Y}}\right] \leqslant \beta\tau.$$

Assumption 9 implies also that

$$\mathbb{E}\left[\|Z\|_{\mathcal{Y}}\|X_\lambda - X\|_{\mathcal{X}}^2\right] \leqslant \beta\mathbb{E}\left[\|X_\lambda - X\|_{\mathcal{X}}^2\right].$$

We then get that

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

In particular,

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

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

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

and the corresponding error bound is

$$\mathbb{E}\left[\|X_{\lambda_*} - X\|_{\mathcal{X}}^2\right] \leq 4(1-\beta C_0)^{-1}\tau\beta,$$

which proves the result. □

To apply theorem 1, we consider the problem obtained with a truncated square loss:

$$\ell(x,x') = \|T(x-X_0) - T(x'-X_0)\|_{\mathcal{X}}^2, \tag{4.10}$$

where T is the truncation operator defined in (3.10). The corresponding expected risk is given by

$$L(f) = \mathbb{E}\left[\|T(f(Y)-X_0) - T(X-X_0)\|_{\mathcal{X}}^2\right].$$

We focus on Tikhonov regularization, where, for every $\lambda \in (0,+\infty)$, $f_\lambda(Y) = X_\lambda$ is given by (4.5), and analyze the error corresponding to the choice of the regularization parameter with ERM. Consider independent and identical copies (Y_i, X_i), $i = 1,\ldots,n$, of the pair of random variables (Y,X) as in (4.1). The ERM problem is given by

$$\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^n \|T\left(X_\lambda^i - X_0\right) - (X_i - X_0)\|_{\mathcal{X}}^2. \tag{4.11}$$

where $X_\lambda^i := f_\lambda(Y_i)$. In the following result we derive an upper bound corresponding to the expected risk.

Corollary 2. *Suppose that assumptions 7–9 hold. Let assumption 3 be satisfied with $\lambda_* = \tau/\beta$ and choose the loss as in (4.10). Let $\eta \in (0,1)$. Then, with probability at least $1-\eta$,*

$$L\left(X_{\widehat{\lambda}}\right) \leq \frac{(1+Q)^2}{2Q(1-\beta C_0)}\tau\beta + \frac{26}{n}\log\frac{2N}{\eta}.$$

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 (4.10) is bounded by 4. Moreover, since T defined in (3.10) is the projection on a convex and closed set, it is $1-$Lipschitz, so that theorem 3 implies

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

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

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

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 1. □

Corollary 2 establishes an upper bound on the expected risk of $X_{\widehat{\lambda}}$, 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 $\widehat{\lambda}$ 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.

5. Variational regularization with convex regularizers for linear inverse problems

In this section, we consider the linear inverse problem setting in section 3, with assumption 4 on the noise. We study variational regularization with a general function $J : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ instead of the squared norm,

$$X_\lambda(\omega) \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \frac{1}{2}\|Ax - Y(\omega)\|_{\mathcal{Y}}^2 + \lambda J(x). \tag{5.1}$$

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

$$x \mapsto \|Ax - Y(\omega)\|_{\mathcal{Y}}^2/2 + \lambda J(x)$$

is nonempty for almost every $\omega \in \Omega$, and that $\omega \mapsto X_\lambda(\omega)$ is a measurable selection of the set of minimizers. This setting includes various examples of sparsity-inducing regularizers beyond Hilbertian norms, see e.g. [11] for references. We discuss specific examples in sections 5.1 and 5.2. For this class of regularization schemes, a natural error metric is given by the Bregman divergence, defined for every $x, x' \in \mathcal{X}$ as

$$D_J(x,x') = \begin{cases} J(x) - J(x') - \langle s_J(x'), x - x'\rangle_{\mathcal{X}}, & \text{if } x' \in \operatorname{int}(\operatorname{dom} J), \\ +\infty, & \text{elsewhere,} \end{cases} \tag{5.2}$$

where $s_J(x')$ is an element of $\partial J(x')$, which is nonempty as long as $x' \in \operatorname{int}(\operatorname{dom} J)$ [9, theorem 9.23]. If x and x' belong to $\operatorname{int}(\operatorname{dom} J)$, we can consider also the symmetric Bregman distance, that is

$$d_J(x,x') = D_J(x,x') + D_J(x',x) = \langle s_J(x) - s_J(x'), x - x'\rangle_{\mathcal{X}}.$$

Of course, if J is not differentiable, both the Bregman divergence and the symmetric one depend on the choice of the specific subgradient $s_J(x)$ (and $s_J(x')$). To derive an error bound we consider the following assumptions.

Assumption 10. The function $J : \mathcal{X} \to \mathbb{R}$ is proper, convex, lower semicontinuous and satisfies $\mathrm{dom}(\partial J) = \mathrm{int}(\mathrm{dom}(J))$.

The previous assumption is satisfied, in particular, in two settings, which are discussed in the following: the one where $\mathrm{dom} J = \mathbb{R}^d$ and the one where J is essentially smooth.

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

$$\mathbb{E}\left[\|Z\|_{\mathcal{Y}}^2\right] \leq \beta^2.$$

Assumption 11 can be seen as a generalization of the source condition for the squared norm regularization in assumption 5, in the case $s = 1$. In the following, we will analyze the behavior of $d_J(X_\lambda, X)$. We first show that this quantity is well-defined. From the optimality condition for the Tikhonov problem (5.1) we derive that, almost surely,

$$\frac{1}{\lambda} A^* (Y - AX_\lambda) \in \partial J(X_\lambda). \tag{5.3}$$

In particular we know that $X_\lambda \in \mathrm{dom}\,\partial J$ and so, by assumption 10, that $X_\lambda \in \mathrm{int}(\mathrm{dom} J)$. Moreover, from assumption 11 we have that $X \in \mathrm{int}(\mathrm{dom} J)$ almost surely, and

$$A^*Z \in \partial J(X).$$

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

$$d_J(X_\lambda, X) = \langle \frac{1}{\lambda} A^* (Y - AX_\lambda) - A^*Z, X_\lambda - X \rangle_{\mathcal{X}}. \tag{5.4}$$

The Bregman divergences 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 4. *Under assumptions 4, 10 and 11 the following bound holds, for all* $\lambda \in (0, +\infty)$,

$$\mathbb{E}[d_J(X_\lambda, X)] \leq \frac{\tau^2}{2\lambda} + \frac{\beta^2 \lambda}{2}. \tag{5.5}$$

In particular, taking $\lambda_* = \tau/\beta$, *we have*

$$\mathbb{E}[d_J(X_{\lambda_*}, X)] \leq \beta\tau. \tag{5.6}$$

Proof. The identities and inequalities below are intended to hold almost surely. By assumption 11,

$$\begin{aligned}\lambda d_J(X_\lambda, X) + \|A(X_\lambda - X)\|_{\mathcal{Y}}^2 &= \langle A^*(Y - AX_\lambda) - \lambda A^* Z, X_\lambda - X\rangle_{\mathcal{X}} \\ &\quad + \|A(X_\lambda - X)\|_{\mathcal{Y}}^2 \\ &= \langle Y - AX_\lambda - \lambda Z + AX_\lambda - AX, A(X_\lambda - X)\rangle_{\mathcal{Y}} \\ &= \langle Y - AX - \lambda Z, A(X_\lambda - X)\rangle_{\mathcal{Y}} \\ &\leqslant \frac{1}{2}\|Y - AX - \lambda Z\|_{\mathcal{Y}}^2 + \frac{1}{2}\|A(X_\lambda - X)\|_{\mathcal{Y}}^2.\end{aligned}$$

Rearranging the terms, we obtain

$$\lambda d_J(X_\lambda, X) + \frac{1}{2}\|A(X_\lambda - X)\|_{\mathcal{Y}}^2 \leq \frac{1}{2}\|Y - AX - \lambda Z\|_{\mathcal{Y}}^2.$$

Taking the conditional expectation with respect to X, we get

$$\begin{aligned}\lambda \mathbb{E}[d_J(X_\lambda, X)|X] + \frac{1}{2}\mathbb{E}\left[\|A(X_\lambda - X)\|_{\mathcal{Y}}^2|X\right] &\leq \frac{1}{2}\mathbb{E}\left[\|Y - AX\|_{\mathcal{Y}}^2|X\right] + \frac{\lambda^2}{2}\mathbb{E}\left[\|Z\|_{\mathcal{Y}}^2|X\right] \\ - \lambda \mathbb{E}[\langle Y - AX, Z\rangle_{\mathcal{Y}}|X].\end{aligned}$$

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

$$\begin{aligned}\lambda \mathbb{E}[d_J(X_\lambda, X)] + \frac{1}{2}\mathbb{E}\left[\|A(X_\lambda - X)\|_{\mathcal{Y}}^2\right] &\leq \frac{1}{2}\mathbb{E}\left[\|Y - AX\|_{\mathcal{Y}}^2\right] + \frac{\lambda^2}{2}\mathbb{E}\left[\|Z\|_{\mathcal{Y}}^2\right] \\ &\leqslant \frac{\tau^2}{2} + \frac{\beta^2\lambda^2}{2},\end{aligned}$$

by assumptions 4 and 11. Therefore,

$$\mathbb{E}[d_J(X_\lambda, X)] \leqslant \frac{\tau^2}{2\lambda} + \frac{\beta^2\lambda}{2}. \tag{5.7}$$

The value of λ minimizing the above upper bound is

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

and the theorem follows. □

Remark 4. Following [13], the above analysis can be extended considering $\mathcal{X}$ to be a Banach space embedded in a Hilbert space. In this case, the inner product in $\mathcal{X}$ needs to be replaced by the corresponding duality pairing.

In the rest of the section, we will apply theorem 1 to different loss functions, all based on the Bregman divergence. To perform the analysis, additional assumptions are needed on J to ensure that the hypotheses of theorem 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 $J(x) = |Gx|$, 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 J of Legendre type.

5.1. Sparsity inducing regularizers

In this section, we focus on the finite-dimensional setting, where $\mathcal{X} = \mathbb{R}^d$, $1 \leqslant d < +\infty$. We study sparsity-inducing regularizers such as the $\ell 1$ norm [4]. 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 $G\colon (\mathcal{X}, \|\cdot\|) \to (\mathbb{R}^m, |\cdot|)$. We will consider the following structural assumption.

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

$$J(x) = |Gx|, \tag{5.8}$$

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

The above condition describes the class of sparsity inducing regularizers we consider, including Lasso [60] (G equal to the identity and $|\cdot|$ the ℓ^1 norm), Graph-Lasso [48], penalties for multitask learning [50], group lasso [55], ℓq penalties [35], and Total Variation regularization [54], among others (see [38] and references therein). For these regularization functions J, the subdifferential can be written as

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

which is nonempty at every point $x \in \mathcal{X}$. In addition, recall that the subdifferential of the norm can be computed as [4, remark 1.1]

$$\partial |\cdot| (x) = \{\eta \in \mathbb{R}^m : \quad \langle \eta, x \rangle = |x|,\ |\eta|_* \leqslant 1\}.$$

In this section, we consider the loss function defined by the Bregman divergence for every x and $x' \in \mathbb{R}^d$:

$$\ell(x, x') = D_J(x, x') \tag{5.9}$$

where D_J is defined as in (5.2), for some subgradient $s_J(x') \in \partial J(x')$. As before, if we let $f_\lambda(Y) = X_\lambda$, then the corresponding expected risk is given by

$$L(X_\lambda) = \mathbb{E}\left[D_J(X, X_\lambda)\right]. \tag{5.10}$$

In this case, and as in section 3, we also assume that the random variable X is such that $\|X\| \leqslant 1$ a.s.. Finally, if we denote $X_\lambda^i := f_\lambda(Y_i)$, then the ERM is given by

$$\widehat{\lambda} \in \operatorname*{arg\,min}_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} D_J\left(X_i, X_\lambda^i\right). \tag{5.11}$$

The latter approach has been already applied in practice. In particular, in the case of R being the Total Variation regularizer [18]. We can now state the probabilistic error estimates for this setting.

Corollary 3. *In the setting of this subsection, let assumptions 4, 11 and 12 be satisfied, let assumption 3 be satisfied with $\lambda_* = \tau/\beta$ as in theorem 4 and choose the loss as in (5.9). Let $\eta \in (0,1)$. Then, with probability at least $1 - \eta$,*

$$L\left(X_{\widehat{\lambda}}\right) \leqslant \frac{1+Q^2}{Q} \beta\tau + \frac{13R}{n} \log \frac{2N}{\eta}. \tag{5.12}$$

Proof. To apply theorem 1, we need to check that assumptions 1 and 2 are satisfied. For every $x \in \mathbb{R}^d$ with $\|x\| \leqslant 1$ and $z \in \mathbb{R}^d$, we have

$$\begin{aligned} D_J(x,x') &= |Gx| - |Gx'| - \langle G^* s_{|\cdot|}(Gx'), x - x' \rangle_{\mathbb{R}^m} \\ &= |Gx| - |Gx'| - \langle s_{|\cdot|}(Gx'), Gx - Gx' \rangle_{\mathbb{R}^m} \\ &= |Gx| - \langle s_{|\cdot|}(Gx'), Gx \rangle_{\mathbb{R}^m} \\ &\leqslant \left(1 + |s_{|\cdot|}(Gx')|_*\right) |Gx| \\ &\leqslant 2\|G\|_{\mathrm{op}} \|x\| \\ &\leqslant 2R. \end{aligned}$$

Hence, the loss function is bounded on the cylinder $\{(x,x') \in \mathbb{R}^{d\times d} : \|x\| \leqslant 1\}$, and assumption 1 is therein satisfied with $M = 2R$. We are left to show that assumption 2 is satisfied for $f_\lambda(Y) = X_\lambda$ and L defined as in (5.12). From the inequality

$$D_J(X, X_\lambda) \leqslant d_J(X, X_\lambda)$$

and theorem 4, we derive that

$$L(X_\lambda) \leqslant U(\lambda),$$

where $U(\lambda) = \tau^2/(2\lambda) + \beta^2\lambda/2$. The latter is minimized by $\lambda_* = \tau/\beta$ and satisfies

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

where the multiplicative factor depending on q is a nondecreasing function for $q \geqslant 1$. The statement then follows from theorem 1. □

5.2. Legendre regularizers

In this section, we consider Legendre regularizers. We start by recalling some definitions, see [8] for more details. A proper, convex and lower semicontinuous function $J\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be essentially smooth if ∂J is locally bounded and single valued on its domain. The function J is essentially strictly convex if $(\partial J)^{-1}$ is locally bounded on its domain and J is strictly convex on every convex subset of $\mathrm{dom}\,\partial J$. A function J is Legendre if it is proper, lower semicontinuous and it is both essentially smooth and essentially strictly convex. In this section, we will rely on the following assumption.

Assumption 13. The function $J\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is Legendre.

In particular, assumption 13 implies assumption 10 by [8, theorem 5.6].

Now, consider $x_0 \in \mathrm{int}(\mathrm{dom}\,J)$ and $r > 0$ such that the ball centered at x_0 with radius r, $B := \{x \in \mathcal{X} : \|x - x_0\| \leqslant r\}$, is a subset of $\mathrm{int}(\mathrm{dom}\,J)$. Again, by assumption 13, it is possible to define the projection onto B with respect to the Bregman divergence for every $x \in \mathcal{X}$ (see [8, corollary 7.9]), by setting

$$\pi_B(x) := \operatorname*{arg\,min}_{z\in B} D_J(z,x). \tag{5.13}$$

In this setting, the Bregman projection is univocally defined, meaning that it does not depend on the choice of the subgradient. Indeed, if $x \notin \mathrm{int}(\mathrm{dom}\,J)$, then $D_J(z,x) = +\infty$. Otherwise, $x \in$

$\mathrm{int}(\mathrm{dom}J) = \mathrm{dom}(\partial J)$, where the subdifferential of J is single valued. Moreover, by definition, $\pi_B(x) \in B \subseteq \mathrm{int}(\mathrm{dom}J)$. Recalling that it always holds $\mathrm{int}(\mathrm{dom}J) \subseteq \mathrm{dom}(\partial J)$, we know that the subdifferential of J is non empty at each point of B. In particular, under assumption 13, the subdifferential of J is single valued on B. Then, for every $x \in B$ we denote by $\nabla J(x)$ the subdifferential of J at $x \in B$. We need an additional assumption on the function J on the set B, namely a uniform upper bound for the norm of its gradient; i.e. of ∇J.

Assumption 14. There exists $R > 0$ such that

$$\sup_{x\in B} \|\nabla J(x)\| \leqslant R.$$

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

$$\ell(x,x') = D_J(\pi_B(x), \pi_B(x')), \tag{5.14}$$

which is univocally defined since $\pi_B(x') \in B$, and the subdifferential of J is non empty and single valued on B. We consider also the corresponding expected risk, defined as

$$L(f) = \mathbb{E}[D_J(\pi_B(X), \pi_B(f(Y)))].$$

In this case, and in opposition with the other sections where we assumed that $\|X\| \leqslant 1$, we assume that X is such that $X \in B$ a.s.. As in the previous sections, we want to bound the expected risk of the regularization method $f_\lambda(Y) = X_\lambda$ defined as in (5.1), when λ is selected by ERM,

$$\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} D_J(\pi_B(X_i), \pi_B(X_\lambda^i))),$$

where $X_\lambda^i := f_\lambda(Y_i)$. The corresponding error bound is given in the following corollary.

Corollary 4. *Let assumptions 4, 11, 13 and 14 be satisfied, let assumption 3 be satisfied with $\lambda_* = \tau/\beta$ as in theorem 4 and choose the loss as in (5.14). Let $\eta \in (0,1)$. Then, with probability at least $1-\eta$,*

$$L(X_{\widehat{\lambda}}) \leqslant \frac{1+Q^2}{Q}\beta\tau + \frac{26Rr}{n}\log\frac{2N}{\eta}.$$

Proof. To prove the statement, we will rely again on theorem 1. Therefore we just need to show that assumptions 1 and 2 hold. We first show that assumption 1 is satisfied. Since both $\pi_B(x)$ and $\pi_B(x')$ belong to B, and by assumption 14, recalling that ∂J is single valued on B, it follows that

$$\begin{aligned} 0 \leqslant \ell(x,x') &= D_J(\pi_B(x), \pi_B(x')) \leqslant D_J(\pi_B(x), \pi_B(x')) + D_J(\pi_B(x'), \pi_B(x)) \\ &= \langle \nabla J(\pi_B(x)) - \nabla J(\pi_B(x')), \pi_B(x) - \pi_B(x')\rangle \leqslant 4Rr. \end{aligned}$$

Then, the considered loss function (5.14) is bounded and assumption 1 is satisfied with $M = 4Rr$. Next, we check that assumption 2 is satisfied. First, observe that both X and X_λ belong

to $\mathrm{dom}(\partial J)$ almost surely since $X \in B$ by assumption and by the optimality condition stated in (5.3). Then, the subdifferential of J is not empty (and so single valued) at X, X_λ and

$$d_J(X, X_\lambda) \geq D_J(X, X_\lambda) \geq D_J(X, \pi_B(X_\lambda)) + D_J(\pi_B(X_\lambda), X_\lambda),$$

by the first order optimality conditions of problem (5.13) and the fact that $X \in B$.

Again, since $X \in B$ almost surely, we have that $\pi_B(X) = X$ almost surely. Then, the previous inequality implies that

$$L(X_\lambda) = \mathbb{E}[D_J(\pi_B(X), \pi_B(X_\lambda))] = \mathbb{E}[D_J(X, \pi_B(X_\lambda))] \leq \mathbb{E}[d_J(X, X_\lambda)]. \tag{5.15}$$

Theorem 4 gives the bound $\mathbb{E}[d_J(X, X_\lambda)] \leqslant U(\lambda)$, where $U(\lambda) = \tau^2/(2\lambda) + \beta^2\lambda/2$. So, together with (5.15), this implies that

$$L(X_\lambda) \leqslant U(\lambda).$$

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

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

for any $q \geqslant 1$, where the multiplicative term $(1+q^2)/(2q)$ is a non decreasing function for $q \geq 1$. Hence, assumption 2 is satisfied and we can apply theorem 1 to obtain the desired result. □

6. Numerical results

In this section, we provide an empirical validation of the theoretical results discussed in the previous sections. We consider different numerical settings and, for each of them, analyze the behavior of the expected risk for the proposed data driven parameter choice. First, we consider the setting of linear inverse problems with squared norm regularization. In this case, we focus on Tikhonov regularization and the Landweber method. For both of them we compare the proposed supervised learning procedure with the so-called quasi-optimality criterion [7]. 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 experiments, the expected risk L is always approximated empirically using 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 details: All of the simulations have been implemented in Python on a laptop with 32 GB of RAM and 2.2 GHz Intel Core I7 CPU. In section 6.2.2 we also use the library Numerical Tours by Peyré [52]. The code is available at https://github.com/TraDE-OPT/Learning-the-Regularization-Parameter.

6.1. Spectral regularization methods

In this section, we empirically analyze the proposed supervised parameter selection strategy for Tikhonov regularization and the Landweber method to solve an instance of a linear inverse problem as in section 3. We consider a problem of the form

$$Y = AX + \varepsilon,$$

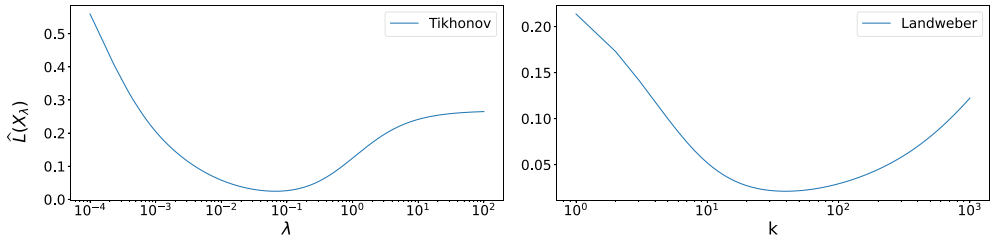

Figure 1. Empirical risk trajectories, of the Tikhonov and Landweber regularization methods, with respect to the regularization parameter λ.

which we describe next. The operator A is a 70×70 matrix, with Gaussian entries $a_{i,j} \sim N(0,1)$, $1 \leqslant i, j \leqslant 70$, 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 $X \in \mathbb{R}^{70}$ as

$$X = (A^*A)^s Z,$$

with $s > 0$ to be fixed later and Z is sampled uniformly in the unit ball. This, jointly with $\|A\|_2 \leqslant 1$, ensures that $\|X\| \leqslant 1$ almost surely. 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 sampling n independent pairs (y_i, x_i) 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 [62].

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

6.1.1. Illustration of the supervised learning parameter choice. We start considering the problem described in section 6.1 with noise level $\tau = 10^{-2}$ and source condition $s = 0.5$. Starting from a training set $\{(y_i, x_i)\}_{i=1}^{50}$, for every $\lambda \in \Lambda$, we define the empirical risk for the Tikhonov regularized solution as

$$\widehat{L}(X_\lambda) = \frac{1}{50} \sum_{i=1}^{50} \|TX_\lambda^i - x_i\|^2, \tag{6.1}$$

where $X_\lambda^i := (A^*A + \lambda I)^{-1} A^* y_i$ (see section 3). The empirical risk for the Landweber method is defined analogously, where in this case $X_\lambda^i = (I - \gamma A^*A)^{\lfloor 1/\lambda \rfloor} A^* y_i$ 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 2, the regularization parameter learned by our approach is $\widehat{\lambda}$, a minimizer of (6.1) on the grid Λ. In figure 1, the function $\lambda \in \Lambda \mapsto \widehat{L}(X_\lambda)$ is plotted for Tikhonov regularization. For Landweber, we plot the function in terms of number of iterations k.

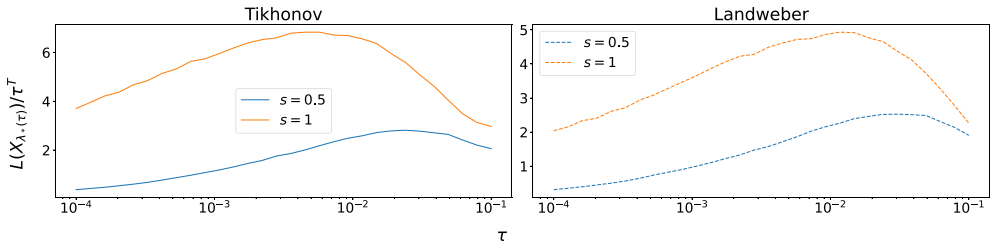

Figure 2. Behavior of $L(X_{\lambda_*})$ with respect to the rate τ^T, $T = (4s)/(2s+1)$, obtained in theorem 2, for different smoothness parameters s for Tikhonov and Landweber. It can be seen that each trajectory is upper bounded, as suggested by the rate in theorem 2. The horizontal axes are shown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

6.1.2. Illustration of theorem 2. In this section we investigate the dependence on the noise level τ of the error $L(X_{\lambda_*})$, see equation (3.7) in theorem 2. 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,

$$\lambda_*(\tau) \in \underset{\lambda \in (0,+\infty)}{\arg\min}\ L(X_\lambda). \tag{6.2}$$

As stated in theorem 2, $L(X_{\lambda_*(\tau)})$ goes to zero when τ vanishes. The parameter α in assumption 6 plays an important role in the bound, since $L(X_{\lambda_*(\tau)}) \lesssim \tau^{4\alpha/(2\alpha+1)}$. In particular, we expect $L(X_{\lambda_*(\tau)})$ 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 2 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 insights into the behavior of the expected risk with respect to the deterministic rate obtained in theorem 2. In figure 2, we illustrate the quantity $L(X_{\lambda_*(\tau)})/\tau^{(4s)/(2s+1)}$, 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 ([29] and theorem 2): 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.

6.1.3. Illustration of error bounds. In this section, we discuss some numerical experiments supporting the error bound stated in corollary 1, both for Tikhonov and Landweber regularization methods. By corollary 1, with high probability, there exist constants c_2, $c_3 > 0$ such that

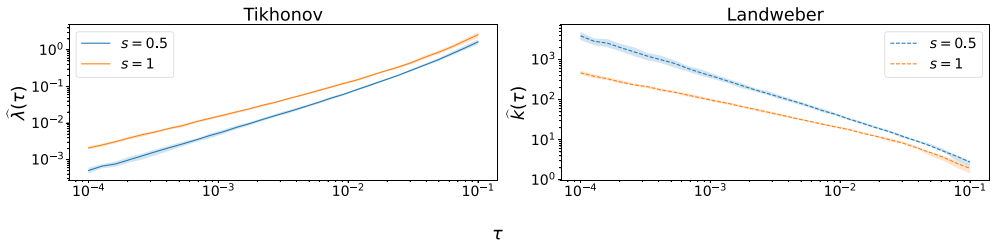

Figure 3. Value of $\widehat{\lambda}, \widehat{k}$ when varying the noise level for both Tikhonov and Landweber methods. Both parameters have been selected over a training set of 10 points, constructed with different smoothness parameters as shown in the plot. Solid lines represent the mean value, while the shaded regions represent the 5th-percentiles and 95th-percentiles over 30 trials. Both axis are shown in logarithmic scale.

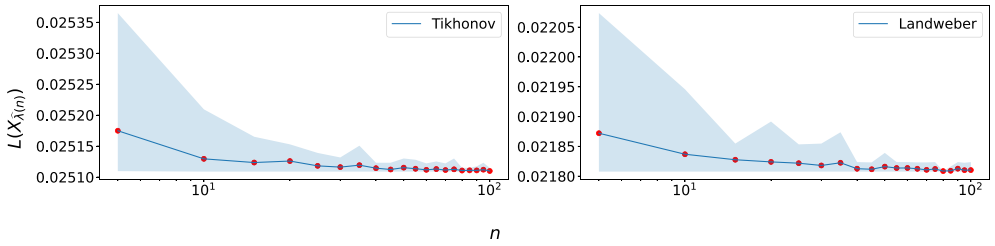

Figure 4. Behavior of $L(X_{\widehat{\lambda}(n)})$, both for Tikhonov and Landweber regularization, as a function of n. The solid lines represent the mean value, while the shaded regions represent the 5th-percentiles and 95th-percentiles over 30 trials. The x-axis is shown in logarithmic scale.

$$L\left(X_{\widehat{\lambda}}\right) \leqslant c_2 \tau^{4\alpha/(2\alpha+1)} + \frac{c_3}{n}.$$

Therefore $L(X_{\widehat{\lambda}})$ with fixed noise level, behaves as $L(X_{\lambda_*})$ up to an additive constant. The same holds for fixed n, and $\tau \to 0$.

We consider the same setting as for figure 1 with noise level $\tau = 0.01$ and smoothness parameter $s = 0.5$. We define the empirical risk, $\widehat{L}(X_\lambda)$, for every $n \in \{5, 10, \ldots, 100\}$, where we sample 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 4 we show that $L(X_{\widehat{\lambda}(n)})$ goes to a certain constant, that depends on the noise level, when n increases, see figure 4.

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 $\widehat{\lambda}(\tau)$, $\widehat{k}(\tau)$ as the minimizers of the empirical risk $\widehat{L}(X_\lambda)$, constructed with $n = 5$ training points. In figure 5 we plot, for every noise level τ, the values $L(X_{\lambda_*(\tau)})$ and

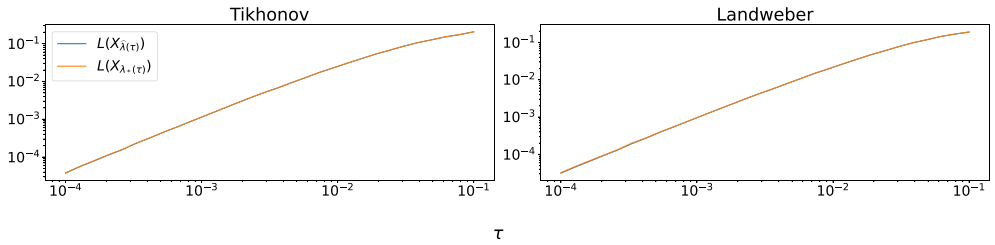

Figure 5. Comparison between $L(X_{\lambda_*(\tau)})$, in orange, and $L(X_{\widehat{\lambda}(\tau)})$, in blue, when varying the noise level τ both for Tikhonov and Landweber regularization. As it can be observed, in such a scale, the lines almost coincide.

$L(X_{\widehat{\lambda}(\tau)})$ in the cases of Tikhonov and Landweber, showing that their behavior with respect to τ is comparable.

6.1.4. Comparison with the quasi-optimality criterion. In this section we compare our supervised learning approach to the quasi-optimality criterion [62]. 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 [7]. 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 $\{(y_i, x_i)\}_{i=1}^{n_{\text{test}}}$, $n_{\text{test}} \in \mathbb{N}$, and a regularization method X_λ, it outputs the best regularization parameter $\widehat{\lambda}_i$ for each y_i, $i = 1, \dots, n_{\text{test}}$. 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 $i = 1, \dots, n_{\text{test}}$, 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 denote by $X^i_{\lambda_j}$ the solution of the regularized problem for the parameter λ_j and datum y_i, $i \in \{1, \dots, n_{\text{test}}\}$. We fix $n_{\text{test}} = 50$. For each pair (y_i, x_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

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

Our method instead provides a unique $\widehat{\lambda}$, 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

$$L^{\text{qo}} = \frac{1}{50} \sum_{i=1}^{50} \|X^i_{\lambda_i^{\text{qo}}} - x_i\|^2.$$

For the Landweber iteration, we fix a grid of $N = 800$ regularization parameters $\Lambda \subseteq [1/1000, 1]$, with $Q \approx 1.0087$ we follow the implementation of the quasi-optimality criterion

Table 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.

$L^{\text{learn}} - L^{\text{qo}}$, 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 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.

$L_{\text{learn}} - L^{\text{qo}}$, 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

proposed in [5], and we define $\lambda_i^{\text{qo}} = \lambda_{j_*}$, where $j_*(i)$ is defined as

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

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 1 and 2 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 supervised learning 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 table 2 that the learned regularization parameter performs better for lower values of the noise level.

6.2. *Sparsity inducing regularizers*

In this section, we explore the theoretical results in section 5.1 for three different examples: denoising and deblurring of a sparse signal, and Total Variation regularization for image denoising. In particular, we will focus on illustrating, experimentally, corollary 3. To do so, we will perform the same experiments that we did for the spectral case in section 6.1.3: first, we show that the expected risk for $\widehat{\lambda}$ with fixed noise level τ, converges when the number of training points goes to infinity. Second, we show that the expected risk, when evaluated at the best empirical parameter $\widehat{\lambda}$ for a fixed number of training points, has a comparable behavior to the expected risk evaluated at its minimum.

We start with the problem of denoising of a sparse signal.

6.2.1. Denoising of a sparse signal. Let $x^* \in \mathbb{R}^d$ be an s-sparse signal; i.e. a signal with s nonzero entries, and consider the white noise model $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$, with noise level $\tau > 0$. We consider the denoising problem,

$$y = x^* + \varepsilon, \tag{6.3}$$

where x^* is such that $\|x^*\|_2 \leqslant 1$ as required. The most classical approach to recover x^* having access only to y is to solve the variational problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_1. \tag{6.4}$$

where the ℓ^1 norm promotes sparsity [22]. In this case, it is easy to show that the solution admits a closed-form expression, that is

$$X_\lambda = \mathcal{S}_\lambda(y), \quad \lambda \in (0, +\infty),$$

where $\mathcal{S}_\lambda$ denotes the so-called soft-thresholding operator [28], and it is defined componentwise as

$$(\mathcal{S}_\lambda(y))_i := \begin{cases} y_i - \lambda \mathrm{sign}(y_i), & \text{if } |y_i| > \lambda, \\ 0, & \text{if } |y_i| \leqslant \lambda, \end{cases}$$

for every $i \leqslant d$.

As an illustrative example, we show in figure 6 the behavior of the empirical risk in this setting, where the training set $\{(y_i, x_i)\}_{i=1}^n$ is generated according to (6.3) with $d = 1024$, $s = 64$ and noise level $\tau = 0.25$,

$$\widehat{L}(\mathcal{S}_\lambda) = \frac{1}{n}\sum_{i=1}^n D_{\|\cdot\|_1}(x_i, \mathcal{S}_\lambda(y_i)),$$

for $n = 10$ and λ in a grid $\Lambda \subseteq [10^{-4}, 10]$ with $N = 1000$ and so $Q \approx 1.0116$. It can be seen that this behavior matches the one predicted by the theoretical results. We first recall the Bregman divergence for this case, that is

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

for every x, $y \in \mathbb{R}^d$. On the one hand, observe that, for every $i = 1, \ldots, n$, we have that $\mathrm{sign}(\mathcal{S}_\lambda(x_i)) \to \mathrm{sign}(y_i)$ as $\lambda \to 0$. This leads to

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

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

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

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

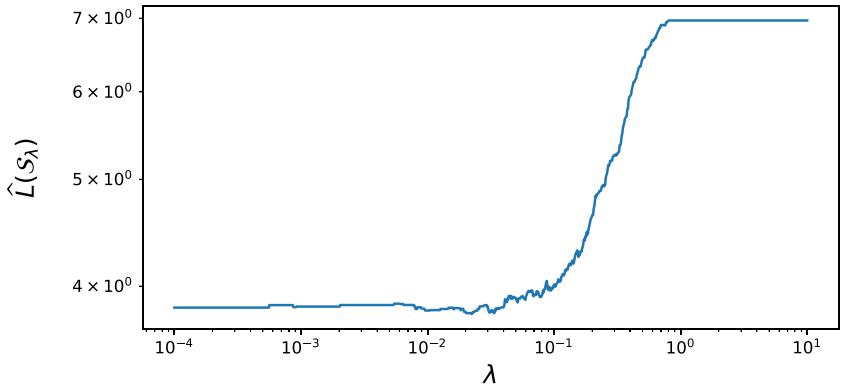

Figure 6. Behavior of $\widehat{L}(\mathcal{S}_\lambda)$ with respect to the regularization parameter λ for the signal denoising problem.

Next, we illustrate numerically corollary 3 for this setting. First, we show that the expected risk for the learned regularization parameter $\widehat{\lambda}$, converges 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 $n \in \{1, 2, \ldots, 20\}$ we define $\widehat{\lambda}(n)$ as a minimizer of the empirical risk,

$$\widehat{\lambda}(n) \in \arg\min_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} D_{\|\cdot\|_1}(x_i, \mathcal{S}_\lambda(y_i)).$$

where, for every n, we consider an independent set of training points $\{(y_i, x_i)\}_{i=1}^n$, generated according to (6.3). In figure 7, we plot the quantity $L(\mathcal{S}_{\widehat{\lambda}(n)})$ for different values of the dimension, $d = 512$, 1024 and 2048 and fixed sparsity $s = 16$, showing that it converges when the number of training points goes to infinity, and the convergence does not depend on the dimension of the problem.

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 $\tau \in [0.1, 1]$. 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 $Q \approx 1.0233$. Then, we find $\widehat{\lambda}_\Lambda$ as the minimizer of the empirical risk $\widehat{L}$, constructed with $n = 5$ training points. In figure 8, we show that the behavior of both $L(\mathcal{S}_{\lambda_*(\tau)})$ and $L(\mathcal{S}_{\widehat{\lambda}(\tau)})$ with respect to τ is comparable.

6.2.2. Deblurring of a sparse signal. In this section, we consider the deblurring of a sparse signal[5]. Our aim is to recover a sparse signal $x^* \in \mathbb{R}^{256}$ that has been corrupted via a convolution operator A and additive noise,

$$y = Ax^* + \varepsilon, \tag{6.5}$$

where x^* is an 8-sparse signal such that $\|x^*\|_2 \leqslant 1$ by assumption, and $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$ as pointed in assumption 4. Moreover, the forward mapping A is a linear convolution operator

$$x \in \mathbb{R}^{256} \mapsto Ax = h * x \in \mathbb{R}^{256},$$

[5] see www.numerical-tours.com/python/.

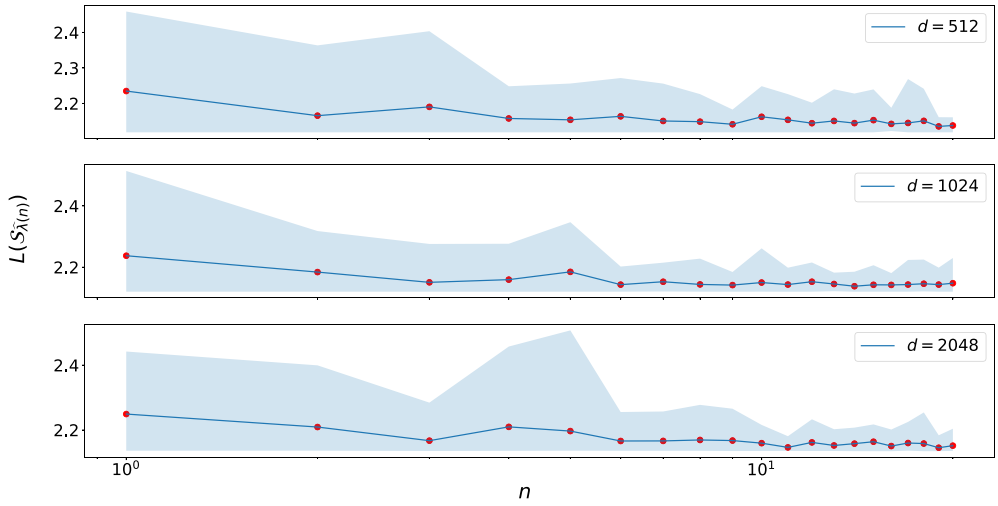

Figure 7. Behavior of $L(\mathcal{S}_{\hat{\lambda}(n)})$ as a function of n for different values of the dimension. The solid lines represent the mean value, while the shaded regions represent the values between the 5th-percentiles and 95th-percentiles over 30 trials. The x-axis is shown in logarithmic scale.

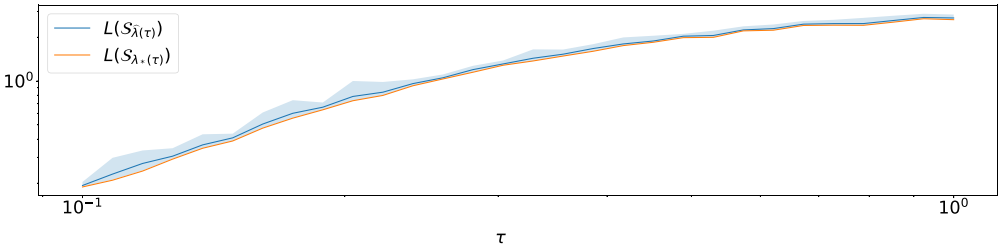

Figure 8. Behavior of the expected risk L with respect to the noise level τ. Recall that $L(\mathcal{S}_{\hat{\lambda}(\tau)})$ has been computed 30 times. We therefore report the mean value, in a solid line, and the values between the 5th-percentile and 95th-percentile, which corresponds to the shaded region. Both axis are shown in logarithmic scale.

with h the second derivative of a Gaussian. More precisely, let $\phi(x) = e^{-x^2/(2\pi^2)}$, then $h = \phi'' - \mu(\phi'')$, being $\mu(\phi'')$ the expectation of ϕ''. In order to recover x^*, we solve the Lasso problem [60]:

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1, \tag{6.6}$$

where $\lambda > 0$ is the regularization parameter, running FISTA with constant stepsize [10] until convergence; i.e. until the difference between iterates is smaller than 10^{-6}.

We now aim at illustrating corollary 3; 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

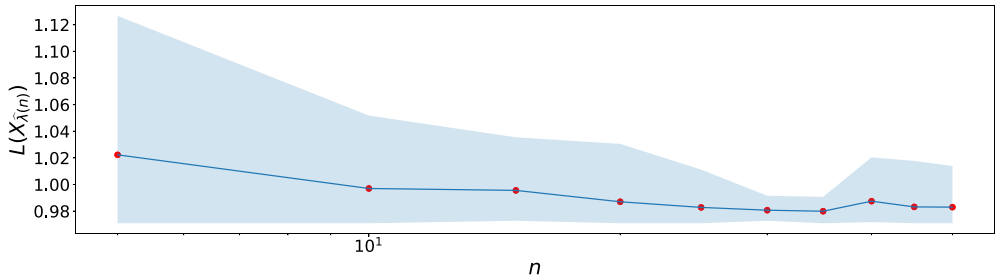

Figure 9. Behavior of $L(X_{\widehat{\lambda}(n)})$ for the signal deblurring problem, showing that it goes to a certain constant as n increases. The solid line represents the mean value, while the shaded region represents the value between 5th-percentile and the 95th-percentile over 30 trials. The x-axis is shown in logarithmic scale.

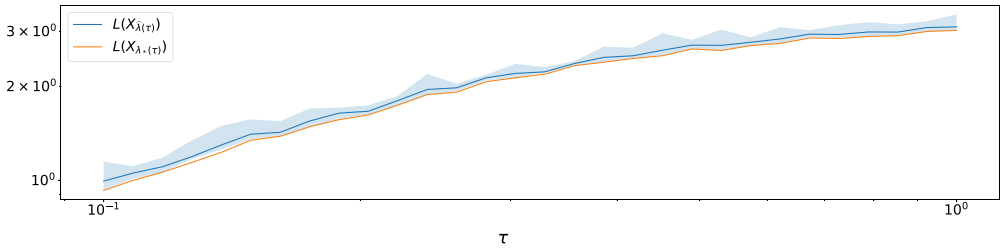

Figure 10. Behavior of the expected risk L, with respect to the noise level τ, for the signal deblurring problem. Recall that $L(X_{\widehat{\lambda}(\tau)})$, in blue has been computed 30 times. We therefore report the mean value, in a solid line, and the values between the 5th-percentile and 95th-percentile, which corresponds to the shaded region. Both axis are shown in logarithmic scale.

ERM in this case can be written as,

$$\widehat{\lambda}(n) \in \arg\min_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} D_{\|\cdot\|_1}\left(x_i, X_\lambda^i\right).$$

where $X_\lambda^i := X_\lambda(y_i)$ and, for every $n \in \{5, 10, \ldots, 50\}$, we consider independent sets of training points $\{(y_i, x_i)\}_{i=1}^n$, generated according to (6.5). From corollary 3, the expected risk evaluated at $\widehat{\lambda}(n)$, $L(X_{\widehat{\lambda}(n)})$, should converge to constant when $n \to \infty$. We plot this behavior in figure 9.

Next, we show, empirically, that the behavior of the learned regularization parameter and the optimal one is comparable with respect to the noise level τ. We 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 $n = 5$ training points for every value of the noise level τ. In figure 10, we plot the quantities $L(X_{\lambda_*(\tau)})$ and $L(X_{\widehat{\lambda}(\tau)})$, showing that their behavior is comparable with respect to the noise level τ.

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

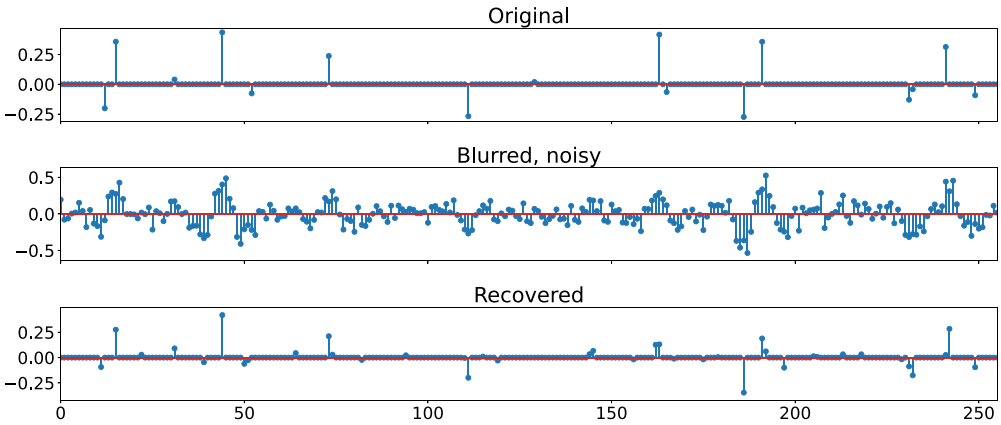

Figure 11. Deblurring of a sparse noisy, blurred signal with learned regularization parameter. In the first row, we show the original signal; in the second, its blurred and noisy version; and in the third row, the regularized solution with learned regularization parameter.

and corrupted signals with the same distribution as the test element that we want to reconstruct, with noise level $\tau = 0.1$. Then, the learned regularization parameter minimizing the empirical risk (5.11). We show in the third row of figure 11, the resulting regularized solution with the learned regularization parameter.

6.2.3. Total variation for image denoising. In this section, we use our supervised learning algorithm for choosing the regularization parameter of a Total Variation regularizer [17, 54]. To do so, we focus on the image denoising problem

$$y = x^* + \varepsilon \tag{6.7}$$

where x, $y \in \mathbb{R}^{d\times d}$, $d \in \mathbb{N}$, and $\varepsilon \sim N(0, \tau^2 \mathrm{Id})$ with noise level $\tau > 0$. A classical approach to solve (6.7) is to rely on the following variational approach [57]

$$\min_x \frac{1}{2}\|x - y\|_2^2 + \lambda \mathrm{TV}(x), \tag{6.8}$$

where $\lambda > 0$ is the regularization parameter and

$$\mathrm{TV}(x) = \|Dx\|_1,$$

and $Dx = (D_1 x, D_2 x) \in \mathbb{R}^{2d(d-1)}$ is the discrete derivative, defined as in [16]. Then, we propose as regularization method X_λ a solution of problem (6.8). Since (6.8) does not have a closed-form solution, we compute it by running FISTA on the dual problem of (6.8), until convergence, i.e. until the difference between iterates is smaller than 10^{-8}. To illustrate corollary 3, we first show the behavior of the expected risk, evaluated at the learned regularization parameter $\widehat{\lambda}$ for this example.

We consider the MNIST dataset [25] of 28×28 images of digits from 0 to 9, and corrupt them as in (6.7). To illustrate corollary 3, 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 $n \in \{5, 10, \ldots, 50\}$, we let $\widehat{\lambda}(n)$ be

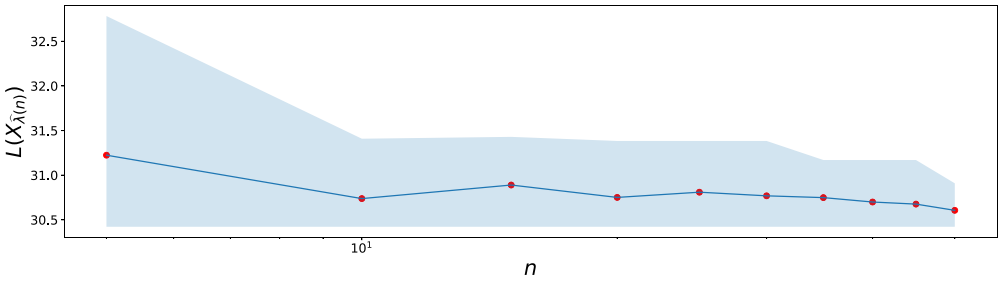

Figure 12. Behavior of $L(X_{\widehat{\lambda}(n)})$ as a function of n for image denoising problem, showing that it goes to a certain constant as n increases. The solid line represents the mean value, while the shaded region represents the value between the 5th-percentile and 95th-percentile over 30 trials. The x-axis is shown in logarithmic scale.

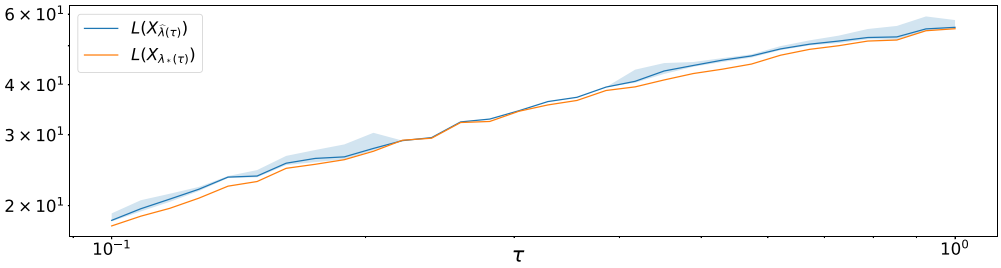

Figure 13. Comparison between $L(X_{\lambda_*(\tau)})$, in orange, and $L(X_{\widehat{\lambda}(\tau)})$, in blue, when varying the noise level τ for the Total Variation regularization. In the case of $L(X_{\widehat{\lambda}(\tau)})$, the solid line represents the mean value, while the shaded region represents the values between the 5th-percentile and 95th-percentile. Bot axes are shown in logarithmic scale.

a minimizer of the empirical risk,

$$\widehat{\lambda}(n) \in \arg\min_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{TV}}\left(x_i, X_\lambda^i\right)$$

where, for every n, we consider an independent training set of points $\{(y_i, x_i)\}_{i=1}^n$ randomly selected from a set of 3000 images. We therefore plot in figure 12 the behavior of the expected risk L for $\widehat{\lambda}(n)$, and show it converges as $n \to \infty$.

Next, we illustrate the behavior of the expected risk. We consider the same experimental setting as we did for figure 10 for the signal deblurring problem, and we show in figure 13 that the behavior of both $L(X_{\lambda_*(\tau)})$ and $L(X_{\widehat{\lambda}(\tau)})$ is comparable with respect to τ.

Finally, as an illustrative example, we explore the performance of our parameter selection method on test images from the MNIST dataset. We compute four different supervised learning regularization parameters for four different training sets, each of 100 training points, and check the reconstruction results of the TV regularized solution for two different digits in the test set. The results are shown in figure 14. 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.

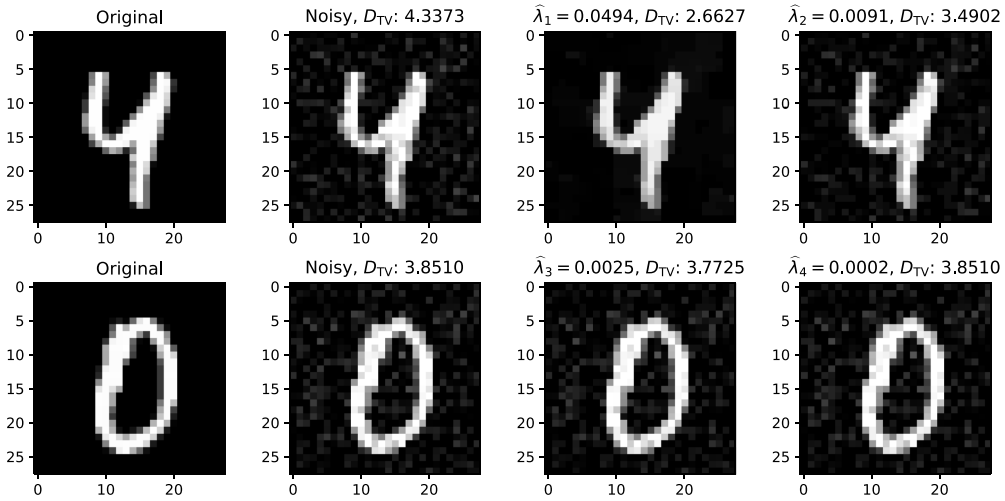

Figure 14. Total Variation denoising algorithm for two digit in the test set. From left to right, in every row, we plot the original image, its noisy version, and the recovery obtained with different regularization parameters. We also include, accordingly, the Bregman divergence with respect to the original image and the value of the regularization parameter that has been used for such recovery.

7. Conclusions

We studied the problem of learning the regularization parameter in statistical inverse problems. In particular we consider a data driven approach that we cast as an instance of empirical risk minimization, common in machine learning. Borrowing results from statistical learning theory, we derive general error guarantees, that we specialize considering different classes of inverse problems and regularization methods. Theoretical results are illustrated by extensive numerical experiments to illustrate.

Possible developments include considering a higher dimensional parameterization of the regularization approaches, as well as possible instance dependent parameter choices. We believe that our analysis provides useful insights into how to tackle these questions.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This project has been supported by the TraDE-OPT project, which received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No 861137. L R acknowledges the Center for Brains, Minds and Machines (CBMM), funded by NSF STC Award CCF-1231216. J C R and L R acknowledge the Ministry of Education, University and Research (Grant ML4IP R205T7J2KP). L R and S V acknowledge the European Research Council (Grant SLING

819789), the US Air Force Office of Scientific Research (FA8655-22-1-7034). The research by E D V, L R, C M and S V has been supported by the MIUR Grant PRIN 202244A7YL. The research by E D V, C M and S V has been supported by the MUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001. E D V, C M and S V are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). This work represents only the view of the authors. The European Commission and the other organizations are not responsible for any use that may be made of the information it contains.

ORCID iDs

Jonathan Chirinos-Rodríguez https://orcid.org/0000-0003-4107-5108
Ernesto De Vito https://orcid.org/0000-0002-4320-3292
Cesare Molinari https://orcid.org/0000-0003-0864-5682
Lorenzo Rosasco https://orcid.org/0000-0003-3098-383X
Silvia Villa https://orcid.org/0000-0002-6232-5631

References

[1] Alberti G S, De Vito E, Lassas M, Ratti L and Santacesaria M 2021 Learning the optimal Tikhonov regularizer for inverse problems *Advances in Neural Information Processing Systems* vol 34 (Curran Associates, Inc.) pp 25205–16
[2] Arridge S, Maass P, Öktem O and Schönlieb C-B 2019 Solving inverse problems using data-driven models *Acta Numer.* **28** 1–174
[3] Aspri A, Korolev Y and Scherzer O 2020 Data driven regularization by projection *Inverse Problems* **36** 125009
[4] Bach F *et al* 2012 Optimization with sparsity-inducing penalties *Foundam. Trends® Mach. Learn.* **4** 1–106
[5] Bauer F and Lukas M A 2011 Comparing parameter choice methods for regularization of ill-posed problems *Math. Comput. Simul.* **81** 1795–841
[6] Bauer F, Pereverzev S and Rosasco L 2007 On regularization algorithms in learning theory *J. Complexity* **23** 52–72
[7] Bauer F and Reiß M 2008 Regularization independent of the noise level: an analysis of quasi-optimality *Inverse Problems* **24** 055009
[8] Bauschke H H, Borwein J M and Combettes P L 2001 Essential smoothness, essential strict convexity and Legendre functions in Banach spaces *Commun. Contemp. Math.* **03** 615–47
[9] Bauschke H H and Combettes P L 2011 *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (*CMS Books in Mathematics/Ouvrages de MathéMatiques de la SMC*) (Springer)
[10] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202
[11] Benning M and Burger M 2018 Modern regularization methods for inverse problems *Acta Numer.* **27** 1–111
[12] Burger M and Osher S 2004 Convergence rates of convex variational regularization *Inverse Problems* **20** 1411
[13] Burger M, Resmerita E and He L 2007 Error estimation for Bregman iterations and inverse scale space methods in image restoration *Computing* **81** 109–35
[14] Calderón A 2006 On inverse boundary value problem *Comput. Appl. Math.* **25** 65–73
[15] Caponnetto A and Yao Y 2010 Cross-validation based adaptation for regularization operators in learning theory *Anal. Appl.* **08** 161–83
[16] Chambolle A 2004 An algorithm for total variation minimization and applications *J. Math. Imaging Vis.* **20** 89–97
[17] Chambolle A and Lions P-L 1997 Image recovery via total variation minimization and related problems *Numer. Math.* **76** 167–88

[18] Chenchene E, Hosseini A and Bredies K 2023 A hybrid proximal generalized conditional gradient method and application to total variation parameter learning *2023 European Control Conf. (ECC)* (IEEE) pp 1–6
[19] Clason C 2021 Regularization of inverse problems (arXiv:2001.00617)
[20] Cucker F and Smale S 2002 On the mathematical foundations of learning *Bull. Amer. Math. Soc.* **39** 1–49
[21] Dambrine M, Khan A A, Sama M and Starkloff H-J 2023 Stochastic elliptic inverse problems. Solvability, convergence rates, discretization and applications *J. Convex Anal.* **30** 851–85
[22] Daubechies I, Defrise M and De Mol C 2004 An iterative thresholding algorithm for linear inverse problems with a sparsity constraint *Commun. Pure Appl. Math.* **57** 1413–57
[23] De Vito E, Fornasier M and Naumova V 2022 A machine learning approach to optimal Tikhonov regularization I: affine manifolds *Anal. Appl.* **20** 353–400
[24] Deledalle C-A, Vaiter S, Fadili J and Peyré G 2014 Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection *SIAM J. Imaging Sci.* **7** 2448–87
[25] Deng L 2012 The MNIST database of handwritten digit images for machine learning research *IEEE Signal Process. Mag.* **29** 141–2
[26] Devroye L, Györfi L and Lugosi G 1996 *A Probabilistic Theory of Pattern Recognition* (*Applications of Mathematics* vol 31) (Springer)
[27] Dinculeanu N 2010 Vector integration and stochastic integration in Banach spaces *The Oxford Handbook of Functional Data Analysis* (Oxford University Press)
[28] Donoho D L and Johnstone I M 1995 Adapting to unknown smoothness via wavelet shrinkage *J. Am. Stat. Assoc.* **90** 1200–24
[29] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (*Mathematics and its Applications* vol 375) (Kluwer Academic Publishers Group)
[30] Engl H W, Kunisch K and Neubauer A 1989 Convergence rates for Tikhonov regularisation of non-linear ill-posed problems *Inverse Problems* **5** 523
[31] Engl H W and Zou J 1907 A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction *Inverse Problems* **16** 1907
[32] Fienup J R 1982 Phase retrieval algorithms: a comparison *Appl. Opt.* **21** 2758–69
[33] Franceschi L, Frasconi P, Salzo S, Grazzi R and Pontil M 2018 Bilevel programming for hyperparameter optimization and meta-learning *Proc. 35th Int. Conf. on Machine Learning* vol 80 (PMLR) pp 1568–77
[34] Golub G H and von Matt U 1997 Generalized cross-validation for large-scale problems *J. Comput. Graph. Stat.* **6** 1–34
[35] Grasmair M, Haltmeier M and Scherzer O 2008 Sparse regularization with lq penalty term *Inverse Problems* **24** 055020
[36] Györfi L, Kohler M, Krzyżak A and Walk H 2002 *A Distribution-Free Theory of Nonparametric Regression (Springer Series in Statistics)* (Springer)
[37] Hansen P C 1992 Analysis of discrete ill-posed problems by means of the l-curve *SIAM Rev.* **34** 561–80
[38] Hastie T, Tibshirani R and Wainwright M 2015 *Statistical Learning With Sparsity: The Lasso and Generalizations* (Chapman & Hall/CRC)
[39] Himmelberg C J 1975 Measurable relations *Fundam. Math.* **87** 53–72
[40] Hofmann B and Mathé P 2017 Tikhonov regularization with oversmoothing penalty for non-linear ill-posed problems in Hilbert scales *Inverse Problems* **34** 015007
[41] Holzleitner M and Pereverzyev S V 2024 On regularized polynomial functional regression *J. Complexity* **83** 101853
[42] Hao D N and Quyen T N T 2010 Convergence rates for Tikhonov regularization of coefficient identification problems in Laplace-type equations *Inverse Problems* **26** 125014
[43] Kaltenbacher B and Hofmann B 2010 Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces* *Inverse Problems* **26** 035007
[44] Kereta Z and Naumova V 2022 On an unsupervised method for parameter selection for the elastic net *Math. Eng.* **4** 1–36
[45] Kunisch K and Ring W 1993 Regularization of nonlinear illposed problems with closed operators *Numer. Funct. Anal. Optim.* **14** 389–404
[46] Kügler P and Sincich E 2009 Logarithmic convergence rates for the identification of a nonlinear Robin coefficient *J. Math. Anal. Appl.* **359** 451–63

[47] Luo R and Qi X 2023 Nonlinear function-on-scalar regression via functional universal approximation *Biometrics* **79** 3319–31
[48] Meinshausen N and Bühlmann P 2006 High-dimensional graphs and variable selection with the Lasso *Ann. Stat.* **34** 1436–62
[49] Morozov V A 1966 On the solution of functional equations by the method of regularization *Doklady Akademii Nauk* vol 167 (Russian Academy of Sciences) pp 510–2
[50] Mosci S, Rosasco L, Santoro M, Verri A and Villa S 2010 Solving structured sparsity regularization with proximal methods *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science* pp 418–33
[51] Neubauer A 2017 On Nesterov acceleration for Landweber iteration of linear ill-posed problems *J. Inverse Ill-Posed Problems* **25** 381–90
[52] Peyré G 2011 The numerical tours of signal processing *Comput. Sci. Eng.* **13** 94–97
[53] Resmerita E and Scherzer O 2006 Error estimates for non-quadratic regularization and the relation to enhancement *Inverse Problems* **22** 801
[54] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica* D **60** 259–68
[55] Salzo S and Villa S 2021 *Proximal Gradient Methods for Machine Learning and Imaging* (Springer) pp 149–244
[56] Scherzer O 2015 *Handbook of Mathematical Methods in Imaging* (Springer)
[57] Scherzer O, Grasmair M, Grossauer H, Haltmeier M and Lenzen F 2009 *Variational Methods in Imaging* vol 167 (Springer)
[58] Shi Z, Fan J, Song L, Zhou D-X, and Suykens J A 2024 Nonlinear functional regression by functional deep neural network with kernel embedding (arXiv:2401.02890)
[59] Sylvester J and Uhlmann G 1987 A global uniqueness theorem for an inverse boundary value problem *Ann. Math.* **125** 153–69
[60] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc.* B **58** 267–88
[61] Tikhonov A N and Arsenin V Y 1977 Solutions of ill-posed problems *Scripta Series in Mathematics* (V. H. Winston & Sons, Wiley)
[62] Tikhonov A N, Glasko V B and Kriksin Y A 1979 On the question of quasi-optimal choice of a regularized approximation *Doklady Akademii Nauk* vol 248 (Russian Academy of Sciences) pp 531–5
[63] Vapnik V 1999 *The Nature of Statistical Learning Theory* (Springer)
[64] Wahba G 1977 Practical approximate solutions to linear operator equations when the data are noisy *SIAM J. Numer. Anal.* **14** 651–67
[65] Waldspurger I, d'Aspremont A and Mallat S 2012 Phase recovery, maxcut and complex semidefinite programming *Math. Program.* **149** 47–81
[66] Williams D 1991 *Probability With Martingales* (Cambridge University Press)