Learning the optimal Tikhonov regularizer for inverse problems

Alberti, Giovanni S.

unspecified
publishedVersion

Learning the optimal Tikhonov regularizer for inverse problems

Giovanni S. Alberti[1], Ernesto De Vito[1], Matti Lassas[2], Luca Ratti[1], Matteo Santacesaria[1]

[1] MaLGa Center, Department of Mathematics, University of Genoa,
Via Dodecaneso 35, 16146 Genova, Italy
giovanni.alberti,ernesto.devito,luca.ratti,matteo.santacesaria@unige.it
[2] Department of Mathematics and Statistics, University of Helsinki,
Gustaf Hällströmin katu 2, 00014 Helsinki, Finland
matti.lassas@helsinki.fi

Abstract

In this work, we consider the linear inverse problem $y = Ax + \varepsilon$, where $A\colon X \to Y$ is a known linear operator between the separable Hilbert spaces X and Y, x is a random variable in X and ε is a zero-mean random process in Y. This setting covers several inverse problems in imaging including denoising, deblurring and X-ray tomography. Within the classical framework of regularization, we focus on the case where the regularization functional is not given a priori, but learned from data. Our first result is a characterization of the optimal generalized Tikhonov regularizer, with respect to the mean squared error. We find that it is completely independent of the forward operator A and depends only on the mean and covariance of x. Then, we consider the problem of learning the regularizer from a finite training set in two different frameworks: one supervised, based on samples of both x and y, and one unsupervised, based only on samples of x. In both cases we prove generalization bounds, under some weak assumptions on the distribution of x and ε, including the case of sub-Gaussian variables. Our bounds hold in infinite-dimensional spaces, thereby showing that finer and finer discretizations do not make this learning problem harder. The results are validated through numerical simulations.

1 Introduction

The aim of an inverse problem is to recover information about a physical quantity from indirect measurements. Virtually all imaging problems and modalities fall within this framework, including denoising, deblurring [14], computed tomography [41] and magnetic resonance imaging [15]. Classical and general approaches to solve inverse problems consist in studying a variational (minimization) problem and can be divided into two classes.

The first is based on the so-called regularization theory [14]. The aim is to recover a single, deterministic, unknown $x^\dagger$ from noisy data $y = F(x^\dagger) + \varepsilon$ by solving a minimization problem

$$\min_x d_Y(F(x), y) + J(x) \tag{1}$$

for a fidelity term $d_Y\colon Y \times Y \to \mathbb{R}$ and a regularization functional $J\colon X \to [0, +\infty)$. The latter is chosen in order to mitigate the ill-posedness of the map F, and represent some a-priori knowledge on x. For instance, in the classical Tikhonov regularization we have $J(x) = \lambda\|x\|_X^2$ for $\lambda > 0$.

The second approach considers the unknown as a random variable and is based on statistical/Bayesian methods [23, 45]. In this case one can recover the unknown using point estimators

such as the maximum a posteriori (MAP) estimator, or extract richer information on the probability distribution of the unknown. In practice, the MAP estimator is found by solving a minimization problem of the same form as (1). The main difference is that the fidelity term and the regularizer are tailored to the statistical properties of the unknown and the noise, which are usually assumed to be known.

In recent years, machine learning techniques, and especially deep learning, have shaken the field of inverse problems by providing the basis for data-driven methods that have outperformed the state-of-the-art in most imaging modalities [3, 43]. The most successful methods take inspiration from regularization theory [10, 1, 26, 2, 36, 21, 22, 27, 34, 39, 8, 17, 40]: the physical model given by the forward map F is assumed to be known while the regularizer (or the gradient updates related to it) is learned from a training set. While these approaches have shown impressive results in applications, a solid theory behind their successes is lacking. In view of the many sensitive applications where these methods are already being employed, e.g. in medical imaging, it is of utmost importance to fill this theoretical gap to better understand the strengths and limits of data-driven imaging modalities. Moreover, many inverse problems are naturally formulated in infinite-dimensional spaces [14, 41, 23, 15, 45, 37], and their discretization must be carefully treated due to their ill-posedness [25]. Hence, it is of main interest to provide a theoretical analysis in the infinite-dimensional setting.

In this work, we consider the problem of learning a regularizer for a linear inverse problem in the framework of statistical learning theory, which is the natural setting to derive precise theoretical guarantees. This is part of the growing research area of learning an operator between infinite dimensional spaces [46, 13, 32, 42, 31]. We study the case where the measurements are modeled by a linear, possibly ill-posed, forward map, and the penalty term is a generalized Tikhonov regularizer [47].

More precisely, let $A\colon X \to Y$ be a bounded linear operator between the separable real Hilbert spaces X and Y. We consider the inverse problem

$$y = Ax + \varepsilon, \tag{2}$$

which consists of the reconstruction of x from the knowledge of y, where ε represents noise. We assume that x is a random variable on X with mean μ and covariance Σ_x, and the noise ε, independent of the variable x, is a zero-mean random process on Y with covariance Σ_ε, (see Section 2 for more details). The operator A is typically injective but its inverse may be unbounded: typical examples include denoising (A is the identity) and deblurring (A is a convolution operator).

We aim to recover the unknown via generalized Tikhonov regularization. For a quadratic fidelity term $d_Y\colon Y \times Y \to \mathbb{R}$, the minimization problem

$$\min_x d_Y(Ax, y) + \|B^{-1}(x - h)\|_X^2, \tag{3}$$

has a unique solution $R_{h,B}(y)$, called the generalized Tikhonov reconstruction. Here the pair (h, B), where $h \in X$ and $B\colon X \to X$ is a positive bounded operator, is considered as a free parameter that we call *regularization pair*. For example, the operator B can be a smoothing operator in L^2, as a negative power of the Laplacian, so that B^{-1} is a differential operator, yielding a classical regularization in Sobolev spaces. We want to characterize and learn the optimal pair (h, B) with respect to the expected, or mean squared, error

$$L(h, B) = \mathbb{E}_{x,y}\|R_{h,B}(y) - x\|_X^2.$$

Our first contribution is a complete characterization of the minimizers of L. In particular we find that $(\mu, \Sigma_x^{1/2})$ is a global minimizer (and is unique if A is injective), which shows that the

best regularizer is completely independent of the forward operator A and depends only on the mean and covariance of x. This is consistent with the known linearized minimum mean squared error estimator in the finite-dimensional case [24], but it is usually not taken into account in the machine learning approaches to inverse problems mentioned above. The extension to the infinite-dimensional case is not straightforward due to the presence of unbounded operators in inverse problems.

Since the computation of the expected error requires the full distribution ρ of x and y, we study how this can be approximated from a finite training set. We suppose to have access to a sample of m pairs $\mathbf{z} = \{(x_j, y_j)\}_{j=1}^m$ drawn independently from ρ. In view of the results on the expected error, we consider two alternative ways to learn a regularizer pair $(\widehat{h}_{\mathbf{z}}, \widehat{B}_{\mathbf{z}})$ from a training set $\mathbf{z}$: either by minimizing the empirical risk [12] (supervised learning), or by using the empirical mean and covariance of $\{x_j\}$ (unsupervised learning). In both cases, we prove generalization bounds for the sample error $|L(\widehat{h}_{\mathbf{z}}, \widehat{B}_{\mathbf{z}}) - L(\mu, \Sigma_x^{1/2})|$. Under some natural compactness assumptions on the class of regularization pairs, we prove that the sample error has the asymptotic behavior

$$|L(\widehat{h}_{\mathbf{z}}, \widehat{B}_{\mathbf{z}}) - L(\mu, \Sigma_x^{1/2})| \lesssim \frac{1}{\sqrt{m}}, \tag{4}$$

with high probability, in both the supervised and the unsupervised approaches. We stress the point that these bounds hold in the infinite-dimensional setting, or, in other words, they do not depend on the discretization of the signal and of the measurements.

Finally, we complement our theoretical findings with some numerical experiments. For a 1D denoising problem (i.e. A is the identity operator) we replicate the asymptotic bound (4) at different discretization scales. Moreover, we find that the unsupervised approach, despite yielding the same rate (4), clearly outperforms the supervised one.

The paper is organized as follows. In Section 2 we introduce the main notation and technical assumptions that will be used throughout the paper, including several examples. Section 3 presents the main results for the minimization of the expected error, while Section 4 is devoted to the study of the sample error. Numerical experiments are the subject of Section 5. Concluding remarks and discussions are reserved for Section 6.

2 Setting the stage

2.1 The random objects x and ε

As mentioned in the introduction, we formulate (2) as a statistical inverse problem, where x and ε are not deterministic but random. Let us start with the description of the prior on x.

Assumption 2.1. Let x be a random variable on a probability space $(\Omega, \mathbb{P})$ taking values in X. More precisely, x is square-integrable, so that its expectation $\mu \in X$ and its covariance $\Sigma_x \colon X \to X$ is a trace-class operator. We assume that Σ_x is injective.

Without loss of generality we can always assume that Σ_x is injective (i.e., x is non-degenerate), since otherwise it would be enough to consider the inverse problem only in $(\ker \Sigma_x)^\perp \subsetneq X$.

Let us consider some common examples of priors arising in inverse problems.

Example 2.2 (Gaussian random variables)**.** A general class of priors arises when considering Gaussian random variables. We recall that x is a Gaussian random variable if for all $v \in X$, $\langle x, v \rangle$ is a real Gaussian random variable and, by Fernique's theorem, x is square-integrable [7]. Since

3

$\Sigma_x \colon X \to X$ is self-adjoint, positive and trace-class, we can write its singular value decomposition (SVD) as

$$\Sigma_x v = \sum_k \sigma_k^2 \langle v, e_k \rangle_X e_k, \qquad v \in X,$$

where $\{e_k\}_k$ is an orthonormal basis of X, $\sum_k \sigma_k^2 < +\infty$ and $\langle x, e_k \rangle \sim \mathcal{N}(\mu_k, \sigma_k)$, where $\mu = \sum_k \mu_k e_k$ is the mean of x. In other words,

$$x = \mu + \sum_k \sigma_k a_k e_k, \tag{5}$$

where a_k are i.i.d. standard Gaussian variables. This shows that, in infinite dimension, since $\sigma_k \to 0$, the variations of x along the direction e_k become smaller and smaller as $k \to +\infty$.

This abstract construction reduces to a smoothness prior by suitably choosing the covariance operator.

Example 2.3 (Smoothing priors). Let $X = L^2(\mathbb{T}^d)$, where $\mathbb{T}^d = \mathbb{R}^d / \mathbb{Z}^d$ is the d-dimensional torus and $d \geq 1$. Let Δ denote the Laplace-Beltrami operator on $\mathbb{T}^d$, which is simply the classical Laplace operator on $[0,1]^d$ with periodic boundary conditions. For $s > \frac{d}{2}$, the operator

$$(I - \Delta)^{-s} \colon L^2(\mathbb{T}^d) \to L^2(\mathbb{T}^d)$$

is trace class, and can be used to define the Gaussian distribution $\mathcal{N}(0, (I - \Delta)^{-s})$. In the notation of Example 2.2, the SVD of $(I - \Delta)^{-s}$ is given by $\sigma_k^2 = (1 + 4\pi |k|^2)^{-s}$ and $e_k(t) = e^{2\pi i k \cdot t}$ with $k \in \mathbb{Z}^d$. This enforces a smoothness prior on x, depending on the parameter s, which controls the decay of the Fourier coefficients of x (see [33, Appendix B] and [7]).

Let us now discuss the model for the noise ε.

Assumption 2.4. Let $\varepsilon = (\varepsilon)_{v \in Y}$ be a (linear) random process on Y with zero mean and such that its covariance $\Sigma_\varepsilon \colon Y \to Y$ defined by $\mathbb{E}[\varepsilon_v \varepsilon_w] = \langle \Sigma_\varepsilon v, w \rangle_Y$ is bounded and injective.

Notice that the injectivity of Σ_ε implies that noise is present in all directions of Y, whereas the boundedness of Σ_ε allows us to regard the random process ε as a bounded (linear) operator from Y into $L^2(\Omega, \mathbb{P})$. The reader is referred to [16] and to Appendix A.1 for additional details on random processes in Hilbert spaces. We mention here some basic properties that will be needed for the following discussion.

Remark 2.5. Even if ε may not belong to Y almost surely (see Example 2.7 below), it is always possible to view it as an element of a larger space, as we now discuss. Let K be a separable Hilbert space and $\iota \colon K \to Y$ be an injective linear map such that $\iota(K)$ is dense in Y and

$$\iota^* \circ \Sigma_\varepsilon \circ \iota \colon K \to K^* \text{ is trace-class,}[1] \tag{6}$$

where we identify $Y^* = Y$, but we do not identify K^* with K, and we regard ι^* as the canonical embedding $Y \to K^*$. The restriction of ε to K is a Hilbert-Schmidt operator from K into $L^2(\Omega, \mathbb{P})$, hence there exists a unique square-integrable random vector ε taking values in K^* such that $\varepsilon_v = \langle \varepsilon, v \rangle_{K^* \times K}$ for $v \in K$. It is easy to show that the random vector ε has zero mean and its covariance operator is $\iota^* \circ \Sigma_\varepsilon \circ \iota \colon K \to K^*$, since

$$\mathbb{E}[\langle \varepsilon, v \rangle_{K^* \times K} \langle \varepsilon, w \rangle_{K^* \times K}] = \mathbb{E}[\varepsilon_{\iota(v)} \varepsilon_{\iota(w)}] = \langle \Sigma_\varepsilon \iota v, \iota w \rangle_Y, \qquad v, w \in K. \tag{7}$$

---

[1]It is worth observing that K and ι always exist: it is enough to choose them, independently of ε, so that the embedding $\iota \colon K \to Y$ is Hilbert-Schmidt, which implies that $\iota^* \circ \Sigma_\varepsilon \circ \iota$ is trace class, since Σ_ε is bounded.

A random variable is always a random process, as we now describe.

Example 2.6. A simple example of this abstract construction consists of considering a random variable ε. In this case Σ_ε is trace-class itself, so that we can choose $K = Y$ and $\iota = I$ and $\varepsilon \in Y$ almost surely. As discussed in Example 2.2 for Gaussian variables, this means that, since $\sigma_k \to 0$, the expected amplitude of the noise in the direction e_k goes to 0 as $k \to \infty$. For instance, the choice of $\Sigma_\varepsilon = (I - \Delta)^{-s}$ as in Example 2.3 corresponds to smaller noise levels for higher Fourier modes.

A random process allows for considering noise that is uniformly distributed in all directions.

Example 2.7 (White noise)**.** The Gaussian white noise ε is a random process on Y such that for any $v \in Y$ it holds that ε_v is a standard Gaussian variable (mean 0 and variance 1), so that $\Sigma_\varepsilon = I$. Heuristically, in the notation of Example 2.2, this corresponds to $\sigma_k = 1$ for every k, and so by (5)

$$\varepsilon = \sum_k a_k e_k, \qquad a_k \sim \mathcal{N}(0, 1),$$

so that $\varepsilon \notin Y$ with probability 1 whenever Y is infinite dimensional (see, e.g., [16]). In view of Remark 2.5, it is possible to consider a larger space K^* so that $\varepsilon \in K^*$ almost surely. For concreteness of explanation, we focus on the case when $Y = L^2(\mathbb{T}^d)$, a typical framework in imaging. A possible choice for the space K is the Sobolev space $H^s(\mathbb{T}^d)$ with $s > d/2$ (see [25]), so that the canonical embedding $\iota \colon H^s(\mathbb{T}^d) \to L^2(\mathbb{T}^d)$ is a Hilbert-Schmidt operator, hence (6) is satisfied and ε can naturally be seen as an element of $H^{-s}(\mathbb{T}^d) = H^s(\mathbb{T}^d)^*$.

2.2 The new formulation of the inverse problem and of the regularization

As a consequence of Assumptions 2.4, since ε may not belong to Y, the inverse problem (2) must be interpreted from a different perspective, namely considering y as the stochastic process $y_v = \langle Ax, v \rangle_Y + \varepsilon_v$ on Y or by formulating the problem as an equation in K^*:

$$y = \iota^* Ax + \varepsilon,$$

i.e. $\langle y, v \rangle_{K^* \times K} = \langle Ax, \iota(v) \rangle_Y + \langle \varepsilon, v \rangle_{K^* \times K}$ for $v \in K$, where $\iota^* \colon Y \to K^*$ is the natural embedding. We denote the joint probability distribution of (x, y) on $X \times K^*$ by ρ.

We now provide a consistent formulation of the quadratic functional appearing in (3). The goal is to replicate what would be the natural choice in a finite dimensional context, i.e.

$$\min_x \|\Sigma_\varepsilon^{-1/2}(Ax - y)\|_Y^2 + \|B^{-1}(x - h)\|_X^2. \tag{8}$$

Unfortunately, if Y is infinite dimensional, the first factor is in general not well-defined, since for example for Gaussian processes $\varepsilon \in \operatorname{Im} \Sigma_\varepsilon^{1/2}$ with probability 0 [7]. Thus, we need to write this minimization problem in a different formulation. We start by stating the assumptions we make on B.

Assumption 2.8. Let us assume that $B \colon X \to X$ is a bounded positive operator such that

$$\operatorname{Im}(AB) \subseteq \operatorname{Im}(\Sigma_\varepsilon \iota). \tag{9}$$

It is worth observing that, whenever Y is infinite dimensional, then $\operatorname{Im}(\Sigma_\varepsilon \iota) \subsetneq Y$ since $\Sigma_\varepsilon \iota$ is compact. Furthermore, (9) requires, in some sense, the operator $AB \colon X \to Y$ to be at least as

"smoothing" as the operator $\Sigma_\varepsilon \iota \colon K \to Y$. For instance, in the case when AB and $\Sigma_\varepsilon \iota$ have the same left-singular vectors, this condition means that the singular values of AB should go to 0 at least as fast as the singular values of $\Sigma_\varepsilon \iota$.

We are now ready to rewrite the functional in (8). The penalty term involving $B^{-1}(x-h)$ suggests the change of variables $x = h + Bx'$. The corresponding minimization problem for x' reads

$$\min_{x' \in X} \|\Sigma_\varepsilon^{-1/2}(A(h + Bx') - y)\|_Y^2 + \|x'\|_X^2. \tag{10}$$

This expression does not require the injectivity of B. By formally expanding the first factor we obtain

$$\|\Sigma_\varepsilon^{-1/2}ABx'\|_Y^2 - 2\langle \Sigma_\varepsilon^{-1/2}(y - Ah), \Sigma_\varepsilon^{-1/2}ABx'\rangle_Y + \|\Sigma_\varepsilon^{-1/2}(y - Ah)\|_Y^2.$$

Let us analyze these three terms separately:

1. Since Σ_ε is self-adjoint we have $\|\Sigma_\varepsilon^{-1/2}ABx'\|_Y^2 = \langle \Sigma_\varepsilon^{-1}ABx', AB'x\rangle_Y$, which is well-defined because $\mathrm{Im}(AB) \subseteq \mathrm{Im}(\Sigma_\varepsilon)$ thanks to (9).

2. The second factor is formally equivalent to $-2\langle y - Ah, \Sigma_\varepsilon^{-1}ABx'\rangle_Y$, which is not well-defined as scalar product in Y since y may not belong to Y. However, since $y \in K^*$ and $\Sigma_\varepsilon^{-1}ABx' \in \iota(K)$ by (9), this scalar product can be interpreted as the duality pairing $-2\langle y - \iota^*Ah, \iota^{-1}\Sigma_\varepsilon^{-1}ABx'\rangle_{K^* \times K}$.

3. The third factor $\|\Sigma_\varepsilon^{-1/2}(y - Ah)\|_Y^2$ is independent of x, and so it irrelevant for the minimization task: thus, we remove it. This is a key step in infinite dimension, since, as mentioned above, $\|\Sigma_\varepsilon^{-1/2}y\|_Y^2 = +\infty$ almost surely. See [45, Remark 3.8] for additional details on this aspect.

This discussion motivates the introduction of the following functional, formally equivalent to (10). For $y \in K^*$, we define the regularized solution of the inverse problem as $\hat{x} = h + B\hat{x}'$, where

$$\hat{x}' = \arg\min_{x' \in X} \|\Sigma_\varepsilon^{-1/2}ABx'\|_Y^2 - 2\langle y - \iota^*Ah, (\Sigma_\varepsilon \iota)^{-1}ABx'\rangle_{K^* \times K} + \|x'\|_X^2. \tag{11}$$

The minimizer exists and is unique, and gives the following expression for $\hat{x}$:

$$R_{h,B}(y) := h + B\hat{x}' = h + B(BA^*\Sigma_\varepsilon^{-1}AB + I)^{-1}((\Sigma_\varepsilon \iota)^{-1}AB)^*(y - \iota^*Ah), \tag{12}$$

where $R_{h,B} \colon K^* \to X$ is a bounded affine map. See Proposition A.2 for all the details. Note that $((\Sigma_\varepsilon \iota)^{-1}AB)^* \colon K^* \to X$ is well-defined thanks to (9).

3 The optimal regularizer

The regularization approach described so far is based on the choice of h and B. In classical regularization theory, these are chosen depending on the prior knowledge of the problem under consideration. In the data-driven approach we consider in this work, h and B are learned from training data. In this section, we let learning come into play and consider the problem of determining the optimal h and B, under the assumptions that the distributions of x and ε are fully known. More precisely, this allows for the explicit computation of the expected error

$$L(h, B) = \mathbb{E}_{(x,y)\sim\rho}\|R_{h,B}(y) - x\|_X^2 = \mathbb{E}_{x,\varepsilon}\|R_{h,B}(\iota^*Ax + \varepsilon) - x\|_X^2,$$

which quantifies the mean square error that our regularization functional (11) yields. Optimal choices of $h^\star$ and $B^\star$ are those that minimize $L(h, B)$, and are characterized in the following result.

6

Theorem 3.1. *Let X and Y and be separable real Hilbert spaces, $A \colon X \to Y$ be a bounded linear operator, x and ε satisfy Assumptions 2.1 and 2.4 and be independent, and K and ι be as in Remark 2.5. Suppose $B = \Sigma_x^{1/2}$ satisfies Assumption 2.8.*

Consider the minimization problem

$$\min_{h,B}\{\mathbb{E}_{(x,y)\sim\rho}\left[\|R_{h,B}(y) - x\|_X^2\right]\}, \tag{13}$$

where the minimum is taken over all B satisfying Assumption 2.8 and over all $h \in X$. Then $(B^\star, h^\star)$ is a global minimizer of (13) if and only if

$$h^\star = \mu \qquad and \qquad B^2|_{(\ker A)^\perp} = \Sigma_x|_{(\ker A)^\perp}.$$

In particular, $B^\star = \Sigma_x^{1/2}$ is always a global minimizer, and is unique if A is injective. Furthermore, for every minimizer $(h^\star, B^\star)$, the corresponding reconstruction map is independent of $B^\star$ and, for all $y \in K^*$, is given by

$$R^\star(y) = \mu + \Sigma_x^{1/2}(\Sigma_x^{1/2}A^*\Sigma_\varepsilon^{-1}A\Sigma_x^{1/2} + I_X)^{-1}((\Sigma_\varepsilon\iota)^{-1}A\Sigma_x^{1/2})^*(y - \iota^*A\mu) \tag{14}$$

$$= \mu + \Sigma_x A^*(\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}(y - \iota^*A\mu). \tag{15}$$

The proof is in Appendix A.3. Some comments on this result are in order.

- By assumption $\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)$ is an injective compact operator from Y to K^*, so that its inverse is not bounded, however it is possible to prove that $\Sigma_x A^*(\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}$ extends to a bounded operator from K^* into X. With a slight abuse of notation, we denote this extension in the same way, so that (15) makes sense for all $y \in K^*$.

- To prove this result, we first consider the minimization in (13) over all possibile affine maps, which yields the so-called Linearized Minimum Mean Square Error (LMMSE) estimator of x. Then, it is possible to show that such optimal affine functional is of the form $R_{h^\star,B^\star}$, for suitable $B^\star$ and $h^\star$. In a finite-dimensional context, such a result is a direct consequence of the expression of the LMMSE estimator (see, e.g., [24, Theorem 12.1]). Theorem 3.1 generalizes this result to the infinite-dimensional case.

- In the case of Gaussian random variables, the expression of the optimal regularizer $R^\star$ coincides with the maximum a posteriori (MAP) estimator. Nevertheless, our result does not require any assumptions on x and ε being Gaussian (see the discussion in [19, 20]).

- The minimum expected loss can be computed as

$$L(h^\star, B^\star) = \operatorname{tr}\left(\Sigma_x^{1/2}(\Sigma_x^{1/2}A^*\Sigma_\varepsilon^{-1}A\Sigma_x^{1/2} + I_X)^{-1}\Sigma_x^{1/2}\right), \tag{16}$$

as it is reported in Appendix A.3.

- It is worth observing that the optimal regularization parameters $B^\star = \Sigma_x^{1/2}$ and $h^\star = \mu$ are independent of A and ε, and depend only on the mean and the covariance of x.

4   Finding the optimal regularizer: the sample error

The computation of the optimal regularizer proposed in the previous section through the minimization of the expected loss L requires the knowledge of the joint probability distribution ρ of x and y. In this section, we suppose that ρ is unknown, [2] but we have access to a sample $\mathbf{z} = \{(x_j, y_j)\}_{j=1}^m$

---

[2]More precisely, we only assume that Σ_ε is known.

of m pairs $(x_j, y_j) \in Z = X \times K^*$ drawn independently from the joint probability distribution ρ, and we study how to learn an estimator $(\widehat{h}_{\mathbf{z}}, \widehat{B}_{\mathbf{z}})$ of the optimal parameters $(h^\star, B^\star)$. We propose two alternative ways to learn an estimator based on a training sample $\mathbf{z}$. For the ease of notation, from now on we omit the dependence on $\mathbf{z}$.

1. *Supervised learning*: $(\widehat{h}_S, \widehat{B}_S)$ is determined by minimizing the empirical risk $\widehat{L}$, namely

$$(\widehat{h}_S, \widehat{B}_S) = \underset{(h,B) \in \Theta}{\text{argmin}} \ \widehat{L}(h, B), \qquad \widehat{L}(h, B) = \frac{1}{m} \sum_{j=1}^{m} \| R_{h,B}(y_j) - x_j \|_X^2, \qquad (17)$$

where Θ is a suitable subset of $X \times \mathcal{L}(X, X)$.

2. *Unsupervised learning*: since the best parameters are $h^\star = \mu$ and $B^\star = \Sigma_x^{1/2}$, a natural estimator is provided by means of the sample $\{x_j\}$ alone as follows:

$$\widehat{h}_U = \widehat{\mu} = \frac{1}{m} \sum_{j=1}^{m} x_j, \qquad \widehat{B}_U = \widehat{\Sigma}_x^{1/2}, \quad \widehat{\Sigma}_x = \frac{1}{m} \sum_{j=1}^{m} (x_j - \widehat{\mu}) \otimes (x_j - \widehat{\mu}). \qquad (18)$$

In both cases, we evaluate the quality of $(\widehat{h}, \widehat{B})$ in terms of its *excess error* $L(\widehat{h}, \widehat{B}) - L(h^\star, B^\star)$.

4.1 Supervised learning: empirical risk minimization

There exist several techniques to show the convergence of the empirical risk minimizer to the optimal parameter, involving tools such as the VC dimension and the Rademacher complexity (see, e.g., [44]), which require some compactness assumption on Θ. Here, we fix a Hilbert space H with a compact embedding $j \colon H \to X$ having dense range. For $\varrho_1 > 0$, set

$$\Theta_1 = \{ j(\bar{h}) \colon \bar{h} \in H, \|\bar{h}\|_H \le \varrho_1 \}, \ \Theta_2 = \{ j\bar{B}j^* \colon \bar{B} \in \mathrm{HS}(H^*, H), \|\bar{B}\|_{\mathrm{HS}(H^*, H)} \le \varrho_1 \}, \qquad (19)$$

and define Θ as the set of pairs $\{(h, B) \in \Theta_1 \times \Theta_2 : B \text{ is positive}\}$. Here, $\mathrm{HS}(H^*, H)$ denotes the space of Hilbert-Schmidt operators from H^* to H. We further assume that

a) the map j can be decomposed as $j = j_2 \circ j_1$, where $j_1 \colon H \to X$ and $j_2 \colon X \to X$ are compact and satisfy

$$s_k(j_1) \lesssim k^{-s}, \qquad s > 0, \qquad \text{being } s_k(j_1) \text{ the singular values of } j_1; \qquad (20)$$

whereas j_2 is such that

$$\mathrm{Im}(A j_2) \subseteq \mathrm{Im}(\Sigma_\varepsilon \iota). \qquad (21)$$

b) The optimal parameter $(h^\star, B^\star) = (\mu, \Sigma_x^{1/2})$ belongs to Θ.

Assumption a), and in particular (20), allows us to explicitly compute the covering numbers, whereas (21) ensures that Assumption 2.8 holds uniformly for each positive operator $B \in \Theta_2$. For example, when $H = H^{\sigma_1}(\mathbb{T}^1)$ and $X = H^{\sigma_2}(\mathbb{T}^1)$ are Sobolev spaces on the one-dimensional torus, assumption a) is satisfied if $s = \sigma_1 - \sigma_2 > 0$. As a consequence, assumption b) can be interpreted as an *a priori* regularity assumption on the problem. Such hypothesis can be relaxed by introducing the approximation error, namely, the rate at which the space Θ approximates $X \times \mathcal{L}(X, X)$ as the radius ϱ_1 grows to ∞. Such an analysis, which easily follows from the range-density property of j, is not treated here.

Finally, we assume that both the inputs and the outputs are bounded, *i.e.*

$$\mathrm{supp}(\rho) \subset B_Z(\varrho_2), \text{ a ball of } Z = X \times K^* \text{ of radius } \varrho_2. \qquad (22)$$

Theorem 4.1. *Under the above conditions, let $(\widehat{h}_S, \widehat{B}_S)$ be defined by (17) and take $\tau > 0$. We have*

$$|L(\widehat{h}_S, \widehat{B}_S) - L(h^\star, B^\star)| \leq \left(\frac{c_1 + c_2\sqrt{\tau}}{\sqrt{m}}\right)^{1-\frac{1}{2s'+1}}, \qquad m \geq m_0, \tag{23}$$

with probability exceeding $1 - e^{-\tau}$, being $0 < s' < s$, where c_1, c_2, m_0 are independent of m and τ.

The proof of Theorem 4.1 is reported in the Appendices A.4 and A.5. The approach is inspired by [12, Proposition 4] and is suited for a much broader class of learning problems: by adapting Lemma A.6, it is possible to extend the current approach to non-quadratic regularization functionals.

A prominent example of H satisfying (20) comes from Sobolev spaces. Consider, e.g., $X = L^2(\mathbb{T}^d)$, where $\mathbb{T}^d$ is the d-dimensional torus, and $H = H^\sigma(\mathbb{T}^d)$. If $\sigma > 0$, the embedding of H in X is compact, and its singular values show a polynomial decay (20) with $s = \sigma/d$.

4.2 Unsupervised learning: empirical mean and covariance

As pointed out in (18), it is possible to recover an approximation of the optimal parameter $(h^\star, B^\star)$ only by taking advantage of a sample of the output variable $\{x_j\}_{j=1}^m$. Since this technique does not require matched couples of inputs and outputs, we refer to it as an unsupervised learning approach. In order to prove a bound in probability for the sample error, in this section we assume that x is a κ-sub-Gaussian random vector, i.e.,

$$\forall v \in X, \langle x, v \rangle_X \text{ is a real sub-Gaussian r.v., i.e. } \|\langle x, v \rangle_X\|_p \leq \kappa\sqrt{p}\|\langle x, v \rangle_X\|_2, \ \forall p > 1, \tag{24}$$

where $\|\langle x, v \rangle\|_p^p = \mathbb{E}[|\langle x, v \rangle|^p]$. It its known [48] that $\mathbb{E}[\|x\|^p]$ is finite for all $p > 0$, so that x has finite mean and its covariance operator Σ_x is trace-class. Gaussian random variables are a particular instance of sub-Gaussian random variables by Fernique's theorem [7]. Note that, in infinite-dimensional spaces, bounded random vectors in general are not sub-Gaussian.

We further assume that the injective operator Σ_ε has a bounded inverse, thus $\Sigma_\varepsilon + A\Sigma_x A^*$ is invertible. This is satisfied for example if ε is the white-noise, since $\Sigma_\varepsilon = I$. We also require that $A^*(\iota^*(\Sigma_\varepsilon + A\Sigma_x A^*))^{-1}$, defined on $\iota^*(Y) \subset K^*$, extends to a bounded operator from K^* into X.

Theorem 4.2. *Under the above conditions, let $(\widehat{h}_U, \widehat{B}_U)$ be defined by (18) and take $\tau > 0$. Then,*

$$|L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star)| \leq \frac{c_3 + c_4\sqrt{\tau}}{\sqrt{m}}, \qquad m \geq m_0, \tag{25}$$

with probability exceeding $1 - e^{-\tau}$, where m_0, c_3 and c_4 depend only on Σ_x, Σ_ε and A.

The proof of Theorem 4.2 is based on several concentration estimates reported in Appendix A.6. The rates we obtain can be meaningfully compared with recent results in supervised learning: see [6, 35].

5 Numerical simulations

We report some numerical results obtained from the supervised and unsupervised strategies for a denoising problem, using synthetic data. The goal of these experiments is twofold: on one hand, we want to study the asymptotic properties of the regularizers learned with the techniques proposed in Section 4 as the sample size m grows, verifying Theorems 4.1 and 4.2. On the other hand, we want to assess that those properties, obtained in an infinite-dimensional setting, do not suffer from the curse of dimensionality. We do so by introducing finer and finer discretizations, and showing that the theoretical bounds do not degrade as the dimension of the problem increases.

5.1 Problem formulation and discretization

We consider a denoising problem on $X = Y = L^2(\mathbb{T}^1)$, being $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ the one-dimensional torus, which consists in determining a signal x from the noisy measurement $y = x + \varepsilon$ and thus corresponds to the case $A = I$. We define a statistical model both for ε and for x, which we use for the generation of the training data. In the learning process, though, we do not take advantage of the knowledge of the introduced probability distributions, apart from the covariance operator of the noise Σ_ε. In accordance with Assumption 2.4, we assume that ε is a random process on Y, and in particular we consider a white noise process, i.e. with zero mean and $\Sigma_\varepsilon = \sigma^2 I$. We consider a noise level of 5%, namely, the standard deviation σ is set to the 5% of the peak value of the average signal. In different tests, we employ different white noise processes with different distributions, including the Gaussian (cfr. Example 2.7) and the uniform distributions. Regarding x, we assume a Gaussian distribution with fixed mean μ and covariance Σ_x, where $\mu = 1 - |2x - 1|$ and $\Sigma_x^{1/2}$ is a convolution operator.

In order to discretize the described problem, we fix $N > 0$ and approximate the space X by means of the $N-$dimensional space generated by a 1D-pixel basis. As a consequence, the functions in X and Y are approximated by vectors in $\mathbb{R}^N$, and the linear operators by matrices in $\mathbb{R}^{N \times N}$. More details on the discretizations and on the random process generation are reported in Appendix A.7.

5.2 Implementation and results

We denote $\theta = (h, B)$. The workflow of the numerical experiments is described as follows:

1. Fix the discretization size N, define the optimal regularizer $R_{\theta^\star}$. Compute $L(\theta^\star)$.

2. For each selected value of the sample size m,

   - generate the samples $\{x_j\}_{j=1}^m$, $\{\varepsilon_j\}_{j=1}^m$;
   - minimize the empirical risk $\widehat{L}(\theta)$ to find $\widehat{\theta}_S$;
   - compute the empirical mean and covariance $\widehat{\mu}$, $\widehat{\Sigma}_x$ to find $\widehat{\theta}_U$;
   - compute the excess risks $|L(\widehat{\theta}_S) - L(\theta^\star)|$ and $|L(\widehat{\theta}_U) - L(\theta^\star)|$.

3. Show the decay of both computed quantities as m increases.

We compute the mean squared errors $L(\theta^\star)$, $L(\widehat{\theta}_S)$ and $L(\widehat{\theta}_U)$ according to the definition of L, thus avoiding the use of a test set. Moreover, we perform the minimization of the empirical risk analytically, thanks to the explicit expression of the regularization functional provided by (12) (see Appendix A.7). As a final remark, the generalization bounds in Theorems 4.1 and 4.2 can be reformulated in expectation. Thus, to verify the expected decay, we repeat the same experiment 30 times, with different training samples for each size m and taking the average in each repetition.[3]

In Figure 1, we present the outcome of the numerical experiments, conducted under different statistical models for x and ε. The sample size ranges between $3 \cdot 10^3$ and $3 \cdot 10^5$. In all the presented scenarios, the decay of the excess risk both in the supervised and in the unsupervised cases agrees with the theoretical estimates, showing a decay of the order $1/\sqrt{m}$. Finally, in Figure 2, we show that the theoretical results are equivalently matched by numerics when the discretization size is increased.

---

[3]All computations were implemented with Matlab R2019a, running on a laptop with 16GB of RAM and 2.2 GHz Intel Core i7 CPU. All the codes are available at `https://github.com/LearnTikhonov/Code`
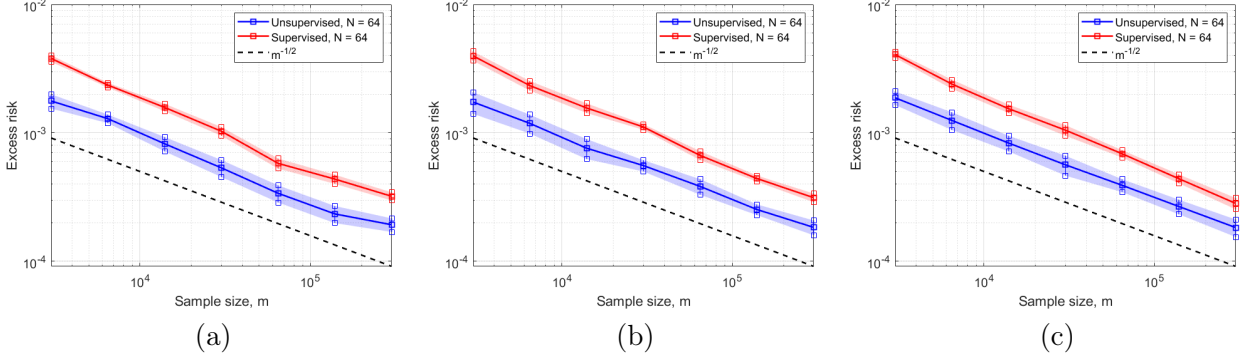
Figure 1: Decay of the excess risks $|L(\widehat{\theta}_S) - L(\theta^\star)|$ and $|L(\widehat{\theta}_U) - L(\theta^\star)|$ in three different cases: Gaussian variable x and (a) Gaussian white noise ε, (b) uniform white noise ε, and (c) white noise ε uniformly distributed w.r.t. the Haar wavelet transform. We also report standard deviation error bars.

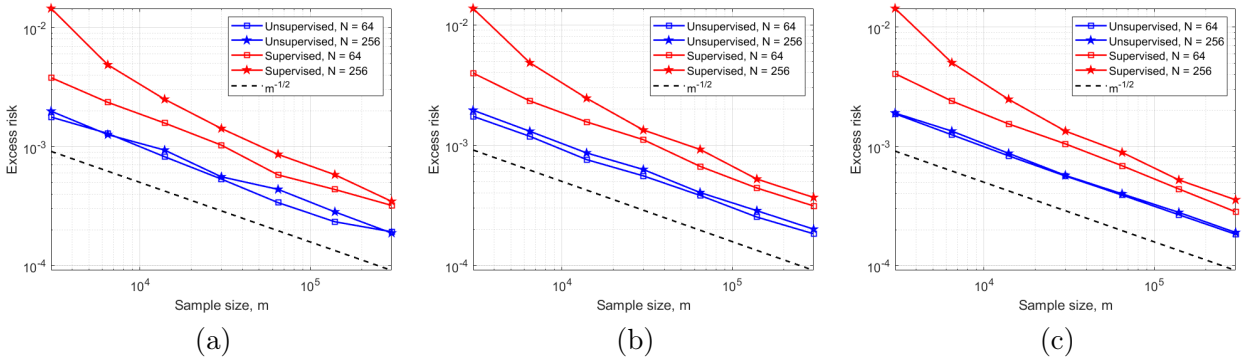

Figure 2: Comparison of decay the excess risks $|L(\widehat{\theta}_S) - L(\theta^\star)|$ and $|L(\widehat{\theta}_U) - L(\theta^\star)|$ with $N = 64$ and $N = 256$, in same three statistical models as in Figure 1.

11

Additional details regarding the results of the numerical experiments are reported in Appendix A.7. Moreover, in Appendix A.8 we replicate the presented numerical study for a different example, namely, a deconvolution problem for 1D signals. In this case, A is a convolution operator with respect to a continuous kernel, whose inverse is in general unbounded.

6   Conclusions and limitations

We studied the problem of learning a regularization functional for an inverse problem between infinite dimensional spaces. This problem has received huge interest in recent years due to the successes in several imaging modalities. Our work provides theoretical support to machine learning approaches in sensitive applications such as medical imaging.

We have considered the case of a linear inverse problem that is solved via generalized Tikhonov regularization. This involves an unknown operator B and a signal h, both to be learned from data. We proposed two learning strategies, one supervised and one unsupervised. Surprisingly, we found that the regularizer learned with the unsupervised strategy has the same (or slightly better) generalization bounds than the supervised one. Furthermore, the unsupervised approach does not need the knowledge of the forward operator A nor that of the distribution of the noise ε. This motivates the development of more advanced unsupervised approaches to the problem, e.g. with deep learning methods (see [26, 27, 34, 36, 39]).

The analysis presented here was possible thanks to the simple form of the regularizer. Our work does not cover, for instance, the case of sparsity promoting regularization functionals [18] or more general convex or non-convex penalty terms arising from deep learning methods. Some results regarding optimal (non-quadratic) regularizers associated with different priors can be found e.g. in [19, 20, 9], which nevertheless deal with a finite-dimensional setting. Extensions to more general regularizers will be the subject of future studies.

Acknowledgments

References

[1] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[2] J. Adler and O. Öktem. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37(6):1322–1332, 2018.

[3] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[4] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[5] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[6] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

[7] V. I. Bogachev. *Gaussian measures*. Number 62. American Mathematical Soc., 1998.

[8] T. A. Bubba, M. Galinier, M. Lassas, M. Prato, L. Ratti, and S. Siltanen. Deep neural networks for inverse problems with pseudodifferential operators: An application to limited-angle tomography. *SIAM Journal on Imaging Sciences*, 14(2):470–505, 2021.

[9] M. Burger and F. Lucka. Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper bayes estimators. *Inverse Problems*, 30(11):114004, 2014.

[10] L. Calatroni, C. Cao, J. C. De los Reyes, C.-B. Schönlieb, and T. Valkonen. Bilevel approaches for learning of variational imaging models. In *Variational methods*, volume 18 of *Radon Ser. Comput. Appl. Math.*, pages 252–290. De Gruyter, Berlin, 2017.

[11] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators*, volume 98 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1990. ISBN 0-521-33011-4. doi: 10.1017/CBO9780511897467. URL https://doi.org/10.1017/CBO9780511897467.

[12] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

[13] M. V. de Hoop, M. Lassas, and C. A. Wong. Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. *arXiv preprint arXiv:1912.11090*, 2019.

[14] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[15] C. L. Epstein. *Introduction to the mathematics of medical imaging*. SIAM, 2007.

[16] J. N. Franklin. Well-posed stochastic extensions of ill-posed linear problems. *Journal of mathematical analysis and applications*, 31(3):682–716, 1970.

[17] D. Gilton, G. Ongie, and R. Willett. Deep equilibrium architectures for inverse problems in imaging. *arXiv preprint arXiv:2102.07944*, 2021.

[18] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with lq penalty term. *Inverse Problems*, 24(5):055020, 2008.

[19] R. Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

[20] R. Gribonval and P. Machart. Reconciling" priors" &" priors" without prejudice? *Advances in Neural Information Processing Systems*, 26:2193–2201, 2013.

[21] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[22] A. Hauptmann, J. Adler, S. R. Arridge, and O. Oktem. Multi-scale learned iterative reconstruction. *IEEE Transactions on Computational Imaging*, 2020.

[23] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[24] S. M. Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[25] H. Kekkonen, M. Lassas, and S. Siltanen. Analysis of regularized inversion of data corrupted by white gaussian noise. *Inverse problems*, 30(4):045009, 2014.

[26] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock. Variational networks: connecting variational methods and deep learning. In *Pattern recognition*, volume 10496 of *Lecture Notes in Comput. Sci.*, pages 281–293. Springer, Cham, 2017. doi: 10.1007/978-3-319-66709-6. URL https://doi.org/10.1007/978-3-319-66709-6.

[27] E. Kobler, A. Effland, K. Kunisch, and T. Pock. Total deep variation for linear inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] H. Koch, A. Rüland, and M. Salo. On instability mechanisms for inverse problems. *arXiv preprint arXiv:2012.01855*, 2020.

[29] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[30] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. ISSN 1350-7265. doi: 10.3150/15-BEJ730. URL https://doi.org/10.3150/15-BEJ730.

[31] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021.

[32] S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions. *arXiv preprint arXiv:2102.09618*, 2021.

[33] M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009. ISSN 1930-8337. doi: 10.3934/ipi.2009.3.87. URL https://doi.org/10.3934/ipi.2009.3.87.

[34] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. Nett: Solving inverse problems with deep neural networks. *Inverse Problems*, 36(6):065005, 2020.

[35] J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

[36] S. Lunz, O. Öktem, and C.-B. Schönlieb. Adversarial regularizers in inverse problems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8516–8525, 2018.

[37] F. Monard, R. Nickl, and G. P. Paternain. Consistent inversion of noisy non-abelian x-ray transforms. *Communications on Pure and Applied Mathematics*, 74(5):1045–1099, 2021.

[38] J. L. Mueller and S. Siltanen. *Linear and nonlinear inverse problems with practical applications*. SIAM, 2012.

[39] S. Mukherjee, S. Dittmer, Z. Shumaylov, S. Lunz, O. Öktem, and C.-B. Schönlieb. Learned convex regularizers for inverse problems. *arXiv preprint arXiv:2008.02839*, 2020.

[40] S. Mukherjee, O. Öktem, and C.-B. Schönlieb. Adversarially learned iterative reconstruction for imaging inverse problems. *arXiv preprint arXiv:2103.16151*, 2021.

[41] F. Natterer. *The mathematics of computerized tomography*. SIAM, 2001.

[42] N. H. Nelsen and A. M. Stuart. The random feature model for input-output maps between banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021.

[43] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. doi: 10.1109/JSAIT.2020.2991563.

[44] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[45] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. ISSN 0962-4929. doi: 10.1017/S0962492910000061. URL https://doi.org/10.1017/S0962492910000061.

[46] P. Tabaghi, M. de Hoop, and I. Dokmanić. Learning schatten–von neumann operators. *arXiv preprint arXiv:1901.10076*, 2019.

[47] A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.

[48] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

A   Appendix

A.1   Random processes

The notion of random vector in infinite-dimensional vector spaces is not general enough to describe many models of noise, as for example the white noise described in Example 2.7. To overcome this problem, a possibility is to consider the noise as a *random process* on Y (see the approach in [16]). A random process is a collection $\{\varepsilon_v\}_{v \in Y}$ of real random variables ε_v, each of them defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and labelled by vectors $v \in Y$. Here we assume that the random process is linear, with zero mean and bounded. This means that

$$\varepsilon_{\alpha v + \beta w} = \alpha \varepsilon_v + \beta \varepsilon_w, \qquad v, w \in Y,$$

and for each $v \in Y$, ε_v has zero-mean with finite bounded variance

$$\mathbb{E}[\varepsilon_v^2] \leq C_\varepsilon \|v\|^2, \tag{26}$$

where $C_\varepsilon > 0$ is a suitable constant independent of v. The existence of C_ε is equivalent to assuming that the covariance operator $\Sigma_\varepsilon \colon Y \to Y$, given by

$$\mathbb{E}[\varepsilon_v \varepsilon_w] = \langle \Sigma_\varepsilon v, w \rangle_Y, \qquad v, w \in Y,$$

is bounded from Y to Y. It is easy to show that if $\varepsilon \colon \Omega \to Y$ is a square-integrable random vector, then the collection of real random variables $\{\varepsilon_v\}_{v \in Y}$

$$\varepsilon_v(\omega) = \langle \varepsilon(\omega), v \rangle_Y \tag{27}$$

is a linear bounded random process. However, the converse is not true as shown by the following example.

Example A.1. The Gaussian white noise ε on Y is a random process such that for any $v \in Y$ it holds that $\varepsilon(v)$ is a zero mean Gaussian variable, and $\Sigma_\varepsilon = I$, i.e., $\mathbb{E}[\varepsilon(v_i)\varepsilon(v_j)] = \langle v_i, v_j \rangle_Y$. Suppose now that $\varepsilon \in Y^*$: then, by Riesz representation theorem there should exist $\hat{\varepsilon} \in Y$ s.t. $\varepsilon(v) = \langle \hat{\varepsilon}, v \rangle_Y$. Nevertheless, this leads to a contradiction, since, letting $\{\phi_i\}_i$ be an orthonormal basis of Y,

$$\mathbb{E}[\|\varepsilon\|_{Y^*}^2] = \mathbb{E}[\|\hat{\varepsilon}\|_Y^2] = \sum_{i,j \in \mathbb{N}} \mathbb{E}[\langle \hat{\varepsilon}, \phi_i \rangle_Y \langle \hat{\varepsilon}, \phi_j \rangle_Y] = \sum_{i,j \in \mathbb{N}} \mathbb{E}[\varepsilon(\phi_i)\varepsilon(\phi_j)] = \sum_{i,j \in \mathbb{N}} \langle \phi_i, \phi_j \rangle_Y,$$

which is a divergent sum. It is moreover easy to show that $\mathbb{P}(\|\varepsilon\|_{Y^*} < \infty) = 0$ (see, e.g. [16]).

However, given a random process $\{\varepsilon_v\}_{v \in Y}$, it is always possible to define a Gelfand triple $K \subseteq Y \subseteq K^*$ and random vector ε taking value in K^* such that (27) holds true for all $v \in K$.

Indeed, let K be a Hilbert space with a continuous embedding $\iota \colon K \to Y$ such that $\iota(K)$ is dense in Y and the linear map

$$K \ni v \mapsto \varepsilon_{\iota(v)} \in L^2(\Omega, \mathbb{P})$$

is a Hilbert-Schmidt operator. Observe that (26) implies that the linear map

$$Y \ni v \mapsto \varepsilon_v \in L^2(\Omega, \mathbb{P})$$

is always bounded, so that it is enough to assume that ι is itself a Hilbert-Schmidt operator. To construct the Gelfand triple, we identify Y^* with Y, but we do not identify K^* with K. Hence,

since $\iota(Y)$ is dense in Y, then $\iota^*\colon Y \to K^*$ is injective and Y can be regarded as a (dense) subspace of K^*, so that $K \subseteq Y \subseteq K^*$.

For a fixed ι, the canonical identification

$$\mathrm{HS}(K, L^2(\Omega, \mathbb{P})) \simeq L^2(\Omega, \mathbb{P}) \otimes K^* \simeq L^2(\Omega, \mathbb{P}, K^*)$$

implies that there exists $\varepsilon \in L^2(\Omega, \mathbb{P}, K^*)$, *i.e.* a square-integrable random vector ε in K^*, such that

$$\varepsilon_v(\omega) = \langle \varepsilon(\omega), v \rangle_{K^* \times K}$$

almost surely. It is easy to show that the random vector $\varepsilon \in K^*$ has zero mean and its covariance operator $\widetilde{\Sigma}_\varepsilon$, defined as

$$\widetilde{\Sigma}_\varepsilon\colon K \to K^* \qquad \langle \widetilde{\Sigma}_\varepsilon v, w \rangle_{K^* \times K} = \mathbb{E}[\langle \varepsilon, v \rangle_{K^* \times K} \langle \varepsilon, w \rangle_{K^* \times K}] \qquad v, w \in K,$$

is given by $\widetilde{\Sigma}_\varepsilon = \iota^* \circ \Sigma_\varepsilon \circ \iota$, as shown by (7).

For example, for the white noise on the Hilbert space $Y = L^2(D)$, $D \subset \mathbb{R}^d$, a possible choice is $K = H^s(D)$ with $s > d/2$, so that the embedding is a Hilbert-Schmidt operator (see [25]). Furthermore, it is possible to show that ε is a Gaussian random vector in K^* and the space $Y \subseteq K^*$ is the corresponding Cameron-Martin space, so that $\mathbb{P}[\varepsilon \in Y] = 0$ (see [7]).

Finally, observe that if the random process is already a square-integrable random vector of Y, then Σ_ε is a trace-class operator and we can simply choose $K = Y$. Hence the random process setting extends the usual formalism of random variables. Clearly, if Y is finite dimensional, the two approaches are equivalent.

A.2 The solution of the regularization problem with fixed h and B

Throughout this section and the following ones, we denote the adjoint of an operator between Hilbert spaces $F\colon \mathcal{H}_1 \to \mathcal{H}_2$ as $F^*\colon \mathcal{H}_2^* \to \mathcal{H}_1^*$ such that $\langle F^* u, v \rangle_{\mathcal{H}_1^* \times \mathcal{H}_1} = \langle u, Fv \rangle_{\mathcal{H}_2^* \times \mathcal{H}_2}$ for all $v \in \mathcal{H}_1$, $u \in \mathcal{H}_2^*$. Notice that we identify the spaces X and Y with their dual spaces, so that, e.g., A^* is intended as an operator from Y to X. We do not identify K with K^* nor H with H^*, so that, e.g., $\iota^*\colon Y \to K^*$.

Proposition A.2. *Let*

- *X, Y and K be separable real Hilbert spaces;*
- *$A\colon X \to Y$ be a bounded map;*
- *$\Sigma_\varepsilon\colon Y \to Y$ satisfy Assumption 2.4;*
- *$\iota\colon K \to Y$ be an injective linear map satisfying (6);*
- *B satisfy Assumption 2.8 and $h \in X$.*

For $y \in K^*$, the problem

$$\hat{x}' = \underset{x' \in X}{\arg\min} \|\Sigma_\varepsilon^{-1/2} ABx'\|_Y^2 - 2\langle y - \iota^* Ah, (\Sigma_\varepsilon \iota)^{-1} ABx' \rangle_{K^* \times K} + \|x'\|_X^2 \tag{28}$$

admits a unique solution, which is given by the bounded affine function $R'_{h,B}\colon K^* \to X$ defined as

$$R'_{h,B}(y) = (BA^* \Sigma_\varepsilon^{-1} AB + I)^{-1}((\Sigma_\varepsilon \iota)^{-1} AB)^*(y - \iota^* Ah). \tag{29}$$

Proof. We have

$$\hat{x}' = \underset{x' \in X}{\arg\min}\, G(x'), \quad G(x') = \|\Sigma_\varepsilon^{-1/2} ABx'\|_Y^2 - 2\langle y - \iota^* Ah, (\Sigma_\varepsilon \iota)^{-1} ABx'\rangle_{K^* \times K} + \|x'\|_X^2.$$

Since the functional G is strictly convex and differentiable, we can find its unique minimum by computing the first-order optimality condition. The Gateaux derivative of G in x' along the direction w reads as

$$\begin{aligned}
G'(x')[w] &= 2\langle BA^*\Sigma_\varepsilon^{-1} ABx', w\rangle_X - 2\langle y - \iota^* Ah, (\Sigma_\varepsilon \iota)^{-1} ABw\rangle_{K^* \times K} + 2\langle x', w\rangle_X \\
&= 2\langle BA^*\Sigma_\varepsilon^{-1} ABx', w\rangle_X - 2\langle((\Sigma_\varepsilon \iota)^{-1} AB)^*(y - \iota^* Ah), w\rangle_X + 2\langle x', w\rangle_X.
\end{aligned}$$

Imposing that $G'(\hat{x}')[w] = 0$ for all $w \in X$ therefore implies that

$$(BA^*\Sigma_\varepsilon^{-1} AB + I_X)\hat{x}' = ((\Sigma_\varepsilon \iota)^{-1} AB)^*(y - \iota^* Ah).$$

The operator $BA^*\Sigma_\varepsilon^{-1} AB + I_X$ is invertible since it is a perturbation of the identity by a self-adjoint, non-negative operator, which leads to the expression of the minimizer

$$\hat{x}' = (BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1}(((\Sigma_\varepsilon \iota)^{-1} AB)^*(y - \iota^* Ah)),$$

as in (29), where we also use that B is self-adjoint since it is positive.

We now show that $R'_{h,B} \colon K^* \to X$ is bounded. We need to show that

$$(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1}(((\Sigma_\varepsilon \iota)^{-1} AB)^* \colon K^* \to X$$

is bounded. As observed above, since $BA^*\Sigma_\varepsilon^{-1} AB$ is self-adjoint and non-negative, we have that $\|(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1}\|_{X \to X} \le 1$. Thus, it remains to show that $((\Sigma_\varepsilon \iota)^{-1} AB)^* \colon K^* \to X$ is bounded. Recall that this composition is well defined thanks to Assumption 2.8 (see (9)). We prove that

$$(\Sigma_\varepsilon \iota)^{-1} AB \colon X \to K \tag{30}$$

is bounded. By assumption, the map $\Sigma_\varepsilon \iota \colon K \to Y$ is bounded, hence closed. Therefore, $(\Sigma_\varepsilon \iota)^{-1} \colon \iota(K) \subseteq Y \to K$ is closed too. By assumption, $AB \colon X \to Y$ is bounded, hence closed. Thus, the composition (30) is closed, hence bounded thanks to the closed graph theorem. $\square$

In view of this result, the regularized solution $\hat{x} = h + B\hat{x}'$ to the inverse problem may be written as in (12):

$$\hat{x} = R_{h,B}(y) = W_B y + b_{h,B}, \tag{31}$$

where

$$\begin{aligned}
W_B &= B(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1}((\Sigma_\varepsilon \iota)^{-1} AB)^*, \\
b_{h,B} &= h - B(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1}((\Sigma_\varepsilon \iota)^{-1} AB)^* \iota^* Ah,
\end{aligned} \tag{32}$$

and $W_B \colon K^* \to X$ is bounded.

We now wish to derive an alternative expression for W_B.

Proposition A.3. *Assume that the hypotheses of Proposition A.2 hold true. The operator $B^2 A^*(\iota^*(AB^2 A^* + \Sigma_\varepsilon))^{-1}$ extends to a bounded linear operator from K^* to X, which coincides with W_B. With an abuse of notation, we have*

$$W_B = B^2 A^*(\iota^*(AB^2 A^* + \Sigma_\varepsilon))^{-1}. \tag{33}$$

18

Proof. First, observe that

$$((\Sigma_\varepsilon \iota)^{-1} AB)^* \iota^* \Sigma_\varepsilon = ((\Sigma_\varepsilon \iota)^{-1} AB)^* (\Sigma_\varepsilon \iota)^* = ((\Sigma_\varepsilon)(\Sigma_\varepsilon \iota)^{-1} AB)^* = (AB)^* = BA^*,$$

so that

$$((\Sigma_\varepsilon \iota)^{-1} AB)^* \iota^*|_{\operatorname{Im} \Sigma_\varepsilon} = BA^* \Sigma_\varepsilon^{-1}. \tag{34}$$

This identity will be used below and in the following.

In order to prove the result, since the operator $\iota^*(AB^2 A^* + \Sigma_\varepsilon)$ is injective, it is enough to show that W_B satisfies

$$W_B \iota^*(AB^2 A^* + \Sigma_\varepsilon) = B^2 A^*.$$

Replacing the expression of W_B given in (32), we obtain

$$B(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1} ((\Sigma_\varepsilon \iota)^{-1} AB)^* \iota^*(AB^2 A^* + \Sigma_\varepsilon) = B^2 A^*.$$

Since B satisfies Assumption 2.8, we have $\operatorname{Im}(AB^2 A^*) \subseteq \operatorname{Im}(\Sigma_\varepsilon \iota) \subseteq \operatorname{Im}\Sigma_\varepsilon$. Thus, by (34), the above identity is equivalent to

$$B(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1} BA^*\Sigma_\varepsilon^{-1}(AB^2 A^* + \Sigma_\varepsilon) = B^2 A^*.$$

In order to prove this identity, it is enough to show that

$$(BA^*\Sigma_\varepsilon^{-1} AB + I_X)^{-1} BA^*\Sigma_\varepsilon^{-1}(AB^2 A^* + \Sigma_\varepsilon) = BA^*.$$

Since $BA^*\Sigma_\varepsilon^{-1} AB + I_X$ is invertible with bounded inverse, this identity is equivalent to

$$BA^*\Sigma_\varepsilon^{-1}(AB^2 A^* + \Sigma_\varepsilon) = (BA^*\Sigma_\varepsilon^{-1} AB + I_X)BA^*.$$

A quick visual inspection shows that this is always true, concluding the proof. □

A.3 Proof of Theorem 3.1 and of (16)

The proof of Theorem 3.1 is based on the following observation.

Lemma A.4. *Assume that the hypotheses of Theorem 3.1 hold true. Let $B_1, B_2 \colon X \to X$ satisfy Assumption 2.8 and suppose that B_1 is injective. Then $W_{B_1} = W_{B_2}$ if and only if*

$$B_1^2 = B_2^2 \quad \text{in } (\ker A)^\perp.$$

Proof. Using (32) for W_{B_1} and (33) for W_{B_2}, the condition $W_{B_1} = W_{B_2}$ reads

$$B_1(B_1 A^*\Sigma_\varepsilon^{-1} AB_1 + I_X)^{-1}((\Sigma_\varepsilon \iota)^{-1} AB_1)^* = B_2^2 A^*(\iota^*(AB_2^2 A^* + \Sigma_\varepsilon))^{-1}.$$

Since B_1 is injective, this is equivalent to

$$((\Sigma_\varepsilon \iota)^{-1} AB_1)^* \iota^*(AB_2^2 A^* + \Sigma_\varepsilon) = (B_1 A^*\Sigma_\varepsilon^{-1} AB_1 + I_X)B_1^{-1} B_2^2 A^*.$$

Since B_2 satisfies Assumption 2.8, we have $\operatorname{Im}(AB_2^2 A^*) \subseteq \operatorname{Im}(\Sigma_\varepsilon \iota) \subseteq \operatorname{Im}\Sigma_\varepsilon$. Thus, by (34),

$$B_1 A^*\Sigma_\varepsilon^{-1}(AB_2^2 A^* + \Sigma_\varepsilon) = (B_1 A^*\Sigma_\varepsilon^{-1} AB_1 + I_X)B_1^{-1} B_2^2 A^*.$$

We readily derive

$$B_1 A^*\Sigma_\varepsilon^{-1} AB_2^2 A^* + B_1 A^* = B_1 A^*\Sigma_\varepsilon^{-1} AB_2^2 A^* + B_1^{-1} B_2^2 A^*,$$

yielding $(B_1^2 - B_2^2)A^* = 0$. By continuity of B_1 and B_2, this is equivalent to having $B_1^2 - B_2^2 = 0$ in $\overline{\operatorname{Im} A^*} = (\ker A)^\perp$, as desired. □

Proof of Theorem 3.1. Recall that $R_{h,B}$ is given by (31).

Step 1: arbitrary affine estimators. We first consider the case of an arbitrary affine estimator $y \mapsto Wy + b$, where $W \colon K^* \to X$ is a bounded linear operator and $b \in X$. Thanks to the independence of x and ε and the fact that $\mathbb{E}\,x = \mu$ and $\mathbb{E}\,\varepsilon = 0$, the corresponding expected error can be expressed as

$$
\begin{aligned}
\mathbb{E}_{x,y}[\|Wy + b - x\|_X^2] &= \mathbb{E}_{x,\varepsilon}[\|(W(\iota^* Ax + \varepsilon) + b - x\|_X^2] \\
&= \mathbb{E}_{x,\varepsilon}[\|(W\iota^* A - I_X)x + W\varepsilon + b\|_X^2] \\
&= \mathbb{E}_x[\|(W\iota^* A - I_X)(x - \mu)\|_X^2] + \|(W\iota^* A - I_X)\mu + b\|_X^2 + \mathbb{E}_\varepsilon[\|W\varepsilon\|_X^2] \\
&= \operatorname{tr}[(W\iota^* A - I_X)\Sigma_x(W\iota^* A - I_X)^*] + \operatorname{tr}[W\iota^* \Sigma_\varepsilon \iota W^*] + \|(W\iota^* A - I_X)\mu + b\|_X^2,
\end{aligned}
$$

where the last step is a consequence of the definition of the covariance operators, e.g.

$$
\begin{aligned}
\mathbb{E}_\varepsilon[\|W\varepsilon\|_X^2] &= \sum_i \mathbb{E}_\varepsilon[\langle W\varepsilon, \varphi_i \rangle_X^2] \\
&= \sum_i \mathbb{E}_\varepsilon[\langle \varepsilon, W^* \varphi_i \rangle_{K^* \times K}^2] \\
&= \sum_i \langle \Sigma_\varepsilon \iota W^* \varphi_i, \iota W^* \varphi_i \rangle_Y \\
&= \sum_i \langle W\iota^* \Sigma_\varepsilon \iota W^* \varphi_i, \varphi_i \rangle_X \\
&= \operatorname{tr}[W\iota^* \Sigma_\varepsilon \iota W^*],
\end{aligned}
$$

where $\{\varphi_i\}$ is an orthonormal basis of X and the third identity follows from (7).

The minimization of the mean square error easily decouples in a minimization in b, yielding

$$
b = (I_X - W\iota^* A)\mu, \tag{35}
$$

and in finding W that minimizes

$$
J(W) = \operatorname{tr}\left[(W\iota^* A - I_X)\Sigma_x(W\iota^* A - I_X)^* + W\iota^* \Sigma_\varepsilon \iota W^*\right].
$$

It is worth observing that, under the introduced hypotheses, such a functional is well-defined. Indeed, since $\iota^* \Sigma_\varepsilon \iota$ is trace-class (cfr. eq. (6)) and W is a bounded operator, the composition $W\iota^* \Sigma_\varepsilon \iota W^*$ defines a trace-class operator, and analogously with the first term, since Σ_x is trace-class.

Step 2: the optimal B. Let us consider the minimization of J. Note that J is convex and differentiable, hence its minimizer can be found by imposing the following first-order optimality condition

Fix an operator $V : K^* \to X$, by imposing that the Gateaux derivative of $J(W)$ along V is zero, we get

$$
\operatorname{tr}\left[V\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)\iota W^* + W\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)\iota V^* - V\iota^* A\Sigma_x - \Sigma_x A^* \iota V^*\right] = 0. \tag{36}
$$

Choose $V = w \otimes v : K^* \to X$, where $v \in K$ and $w \in X$, then

$$
\langle \iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)\iota W^* w, v \rangle_{K^* \times K} + \langle W\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)\iota v, w \rangle_X = \langle \iota^* A\Sigma_x w, v \rangle_{K^* \times K} + \langle \Sigma_x A^* \iota v, w \rangle_X,
$$

so that

$$
\langle W\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)\iota v, w \rangle_X = \langle \Sigma_x A^* \iota v, w \rangle_X.
$$

Since v and w are arbitrary, we get

$$W(\iota^* A \Sigma_x A^* \iota + \iota^* \Sigma_\varepsilon \iota) = \Sigma_x A^* \iota. \tag{37}$$

It is easy to show that, if W satisfies (37), equality (36) holds true for all V. Since $\iota(K)$ is dense in Y and W, A, Σ_x and Σ_ε are bounded, (37) is also equivalent to

$$W \iota^* (A \Sigma_x A^* + \Sigma_\varepsilon) = \Sigma_x A^*. \tag{38}$$

Observe that the operator $A\Sigma_x A^* + \Sigma_\varepsilon \colon Y \to Y$ is positive and injective, hence it has dense range. Further, ι is injective, and so $\iota^* \colon Y \to K^*$ has dense range. Thus, $\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon)$ has dense range. This shows that there exists at most one bounded operator $W \colon K^* \to X$ satisfying (38). Furthermore, Proposition A.3 gives that $W_{\Sigma_x^{1/2}}$ satisfies (38), so that $W_{\Sigma_x^{1/2}}$ is the unique global minimizer of J. Since Σ_x is injective, by Lemma A.4 we have that the B's such that $W_B = W_{\Sigma_x^{1/2}}$ are those satisfying $B^2 = \Sigma_x$ in $(\ker A)^\perp$, as desired.

Step 3: the optimal h. Let us consider (35). It is evident that b is uniquely determined by W, and we know that $W = W_{\Sigma_x^{1/2}}$. We now show that, in the case $b = b_{\Sigma_x^{1/2}, h}$ and $W = W_{\Sigma_x^{1/2}}$, equation (35) reduces to $h = \mu$. Indeed, we have

$$h - \Sigma_x^{1/2}(\Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2} + I_X)^{-1}((\Sigma_\varepsilon \iota)^{-1} A \Sigma_x^{1/2})^* \iota^* A h = (I_X - W_{\Sigma_x^{1/2}} \iota^* A)\mu$$

$$\iff \Sigma_x^{1/2}(\Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2} + I_X)^{-1}((\Sigma_\varepsilon \iota)^{-1} A \Sigma_x^{1/2})^* \iota^* A h = h - \mu + W_{\Sigma_x^{1/2}} \iota^* A \mu$$

$$\iff ((\Sigma_\varepsilon \iota)^{-1} A \Sigma_x^{1/2})^* \iota^* A (h - \mu) = (\Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2} + I_X) \Sigma_x^{-1/2}(h - \mu)$$

$$\iff \Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A (h - \mu) = \Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A (h - \mu) + \Sigma_x^{-1/2}(h - \mu)$$

$$\iff \Sigma_x^{-1/2}(h - \mu) = 0.$$

Therefore the optimal value is $h^\star = \mu$. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

Proof of (16). We provide an expression for the minimum value of the expected loss, $L(h^\star, B^\star)$. We have

$$L(h^\star, B^\star) = J(W^\star) = \operatorname{tr}\left[(W^\star \iota^* A - I_X)\Sigma_x(W^\star \iota^* A - I_X)^* + W^\star \iota^* \Sigma_\varepsilon \iota W^{\star*}\right],$$

where the optimal linear functional $W^\star$ satisfies (38), namely, $W^\star \iota^* = \Sigma_x A^* (\Sigma_\varepsilon + A\Sigma_x A^*)^{-1}$.

$$\begin{aligned}
J(W^\star) &= \operatorname{tr}\left[(W^\star \iota^*)(A\Sigma_x A^* + \Sigma_\varepsilon)\iota(W^\star)^* - \Sigma_x A^* \iota(W^\star)^* - W^\star \iota^* A \Sigma_x + \Sigma_x\right] \\
&= \operatorname{tr}\left[\Sigma_x A^* \iota(W^\star)^* - \Sigma_x A^* \iota(W^\star)^* - W^\star \iota^* A \Sigma_x + \Sigma_x\right] = \\
&= \operatorname{tr}\left[\Sigma_x - \left(\Sigma_x^{1/2}(\Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2} + I_X)^{-1}((\Sigma_\varepsilon \iota)^{-1} A \Sigma_x^{1/2})^* \iota^* A \Sigma_x\right)\right] \\
&= \operatorname{tr}\left[\Sigma_x^{1/2}\left(I_X - ((\Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2} + I_X)^{-1} \Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2})\right) \Sigma_x^{1/2}\right] \\
&= \operatorname{tr}\left(\Sigma_x^{1/2}(\Sigma_x^{1/2} A^* \Sigma_\varepsilon^{-1} A \Sigma_x^{1/2} + I_X)^{-1} \Sigma_x^{1/2}\right),
\end{aligned}$$

where the second line is a consequence of (38) the third line is due to (32) and the forth line holds true by Assumption 2.8 with $B = \Sigma_x^{1/2}$.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

A.4 Proof of Theorem 4.1

In order to prove Theorem 4.1, we adapt the classical result on empirical risk minimization to the present discussion, in particular we follow the simple approach in [12]. We postpone to a future work the use of more refined techniques [4, 29]. We consider the parameter space $\Theta \subset X \times \mathcal{L}(X, X)$ as in (19) and assume in particular that it satisfies (21). We recall that every $B \in \Theta_2$ can be written as $j\bar{B}j^*$, being $\bar{B} \in \mathrm{HS}(H^*, H)$; moreover, since $j = j_2 \circ j_1$, we can also denote it as $B = j_2 \bar{\bar{B}} j_2^*$, where $\bar{\bar{B}} \colon X \to X$, $\bar{\bar{B}} = j_1 \bar{B} j_1^*$. Notice that, using that j_1 and j_2 are injective and have dense range, given B, then $\bar{B}$ and $\bar{\bar{B}}$ are uniquely determined. We can therefore define the following norms on Θ:

$$\|\theta\|_* = \|(h, B)\|_* = \max\left\{\|h\|_X, \|\bar{\bar{B}}\|_{\mathcal{L}(X,X)}\right\},$$

$$\|\theta\|_{**} = \|(h, B)\|_{**} = \max\left\{\|\bar{h}\|_H, \|\bar{B}\|_{\mathrm{HS}(H^*,H)}\right\},$$

where $\bar{h} = j^{-1}(h)$. Notice that, according to (19), the set Θ can be seen as a closed subset of the ball of radius ϱ_1 with respect to $\|\cdot\|_{**}$. Nevertheless, the first result we prove does not require that Θ is chosen as in (19), nor that the functional $R_\theta = R_{h,B}$ is as in (12).

The following result is a restatement of Proposition 4 in [12].

Lemma A.5. *Fix a compact subset Θ of $X \times \{j_2 \bar{\bar{B}} j_2^* : \bar{\bar{B}} \colon X \to X \text{ bounded}\}$, endowed with the norm $\|\cdot\|_*$, and a family of functions $R_\theta \colon K^* \to X$ labelled by $\theta \in \Theta$ satisfying, for a.e. $(x, y) \in X \times K^*$:*

a) $\|R_\theta(y) - x\|_X \leq M_1$ for every $\theta \in \Theta$;

b) $\|R_{\theta_1}(y) - R_{\theta_2}(y)\|_X \leq M_2 \|\theta_1 - \theta_2\|_*$, for every $\theta_1, \theta_2 \in \Theta$.

Then, with probability 1 there exist minimizers of L and $\widehat{L}$ over Θ

$$\theta^\star = \operatorname*{argmin}_{\theta \in \Theta} L(\theta), \qquad \widehat{\theta}_S = \operatorname*{argmin}_{\theta \in \Theta} \widehat{L}(\theta),$$

and, for all $\eta > 0$,

$$\mathbb{P}_{\mathbf{z} \sim \rho^m}\left[|L(\widehat{\theta}_S) - L(\theta^\star)| \leq \eta\right] \geq 1 - 2\mathcal{N}\left(\Theta, \frac{\eta}{16M_1 M_2}\right) e^{-\frac{m\eta^2}{8M_1^4}},$$

where $\mathcal{N}(\Theta, r)$ denotes the covering number of Θ, i.e., the minimum number of balls of radius r (in norm $\|\cdot\|_*$) whose union contains Θ.

Proof. For ρ-almost all $(x, y) \in X \times K^*$

$$\left|\|R_{\theta_1}(y) - x\|_X^2 - \|R_{\theta_2}(y) - x\|_X^2\right| = |\langle R_{\theta_1}(y) - R_{\theta_2}(y), R_{\theta_1}(y) - x + R_{\theta_2}(y) - x\rangle|$$
$$\leq 2M_1 M_2 \|\theta_1 - \theta_2\|_*.$$

By integrating with respect to the probability distribution ρ or the empirical measure $\widehat{\rho}$, the above bound holds for $L, \widehat{L}$. Indeed,

$$\begin{aligned}
|L(\theta_1) - L(\theta_2)| &= \left|\mathbb{E}[\|R_{\theta_1}(y) - x\|_X^2] - \mathbb{E}[\|R_{\theta_2}(y) - x\|_X^2]\right| \\
&\leq \mathbb{E}\left[\left|\|R_{\theta_1}(y) - x\|_X^2 - \|R_{\theta_2}(y) - x\|_X^2\right|\right] \leq 2M_1 M_2 \|\theta_1 - \theta_2\|_*,
\end{aligned} \tag{39}$$

and, with probability 1,

$$
\begin{aligned}
|\widehat{L}(\theta_1) - \widehat{L}(\theta_2)| &= \left| \frac{1}{m} \sum_{j=1}^{m} \|R_{\theta_1}(y_j) - x_j\|_X^2 - \frac{1}{m} \sum_{j=1}^{m} \|R_{\theta_2}(y_j) - x_j\|_X^2 \right| \\
&\leq \frac{1}{m} \sum_{j=1}^{m} \left| \|R_{\theta_1}(y_j) - x_j\|_X^2 - \|R_{\theta_2}(y_j) - x_j\|_X^2 \right| \\
&\leq 2M_1 M_2 \|\theta_1 - \theta_2\|_*.
\end{aligned}
\tag{40}
$$

Since both L and $\widehat{L}$ are Lipschitz continuous and Θ is compact, the corresponding minimizers $\theta^\star$ and $\widehat{\theta}_S$ exist almost surely.

Next, we notice that the event $\{|L(\widehat{\theta}_S) - L(\theta^\star)| \leq \eta\}$ is a superset of the event $\left\{ \sup_{\theta \in \Theta} |\widehat{L}(\theta) - L(\theta)| \leq \eta/2 \right\}$. Indeed,

$$
\sup_{\theta \in \Theta} |\widehat{L}(\theta) - L(\theta)| \leq \frac{\eta}{2} \quad \Rightarrow \quad L(\widehat{\theta}_S) - \widehat{L}(\widehat{\theta}_S) \leq \frac{\eta}{2} \quad \text{and} \quad \widehat{L}(\theta^\star) - L(\theta^\star) \leq \frac{\eta}{2},
$$

and ultimately it also holds that

$$
0 \leq L(\widehat{\theta}_S) - L(\theta^\star) = \left(L(\widehat{\theta}_S) - \widehat{L}(\widehat{\theta}_S) \right) + \left(\widehat{L}(\widehat{\theta}_S) - \widehat{L}(\theta^\star) \right) + \left(\widehat{L}(\theta^\star) - L(\theta^\star) \right) \leq \eta,
$$

where we also used the fact that the central difference is negative by definition of $\widehat{\theta}_S$. Thus,

$$
\mathbb{P}_{\mathbf{z} \sim \rho^m} \left[|L(\widehat{\theta}_S) - L(\theta^\star)| \leq \eta \right] \geq \mathbb{P}_{\mathbf{z} \sim \rho^m} \left[\sup_{\theta \in \Theta} |\widehat{L}(\theta) - L(\theta)| \leq \frac{\eta}{2} \right].
$$

We now provide a lower bound for the latter term. In view of (39) and (40), by using the reverse triangle inequality, for every $\theta_1, \theta_2 \in \Theta$,

$$
\left| |\widehat{L}(\theta_1) - L(\theta_1)| - |\widehat{L}(\theta_2) - L(\theta_2)| \right| \leq 4M_1 M_2 \|\theta_1 - \theta_2\|_*.
$$

Let now $N = \mathcal{N}\left(\Theta, \frac{\eta}{8M_1 M_2} \right)$ and consider a discrete set $\theta_1, \ldots, \theta_N$ such that the balls B_k centered at θ_k with radius $\frac{\eta}{8M_1 M_2}$ cover the entire Θ. In each ball B_k, for every $\theta \in B_k$ it holds

$$
\left| |\widehat{L}(\theta) - L(\theta)| - |\widehat{L}(\theta_k) - L(\theta_k)| \right| \leq 4M_1 M_2 \|\theta - \theta_k\|_* \leq \frac{\eta}{2}.
$$

Therefore, the event $|\widehat{L}(\theta) - L(\theta)| > \eta$ is a subset of $|\widehat{L}(\theta_k) - L(\theta_k)| > \frac{\eta}{2}$, and a bound (in probability) of this term can be provided by standard concentration results. Indeed, $\widehat{L}(\theta_k)$ is the sample average of m realization of the random variable $\|R_{\theta_k}(y) - x\|_X^2$, whose expectation is $L(\theta_k)$. Moreover, such random variable is bounded by M_1^2 by assumption, and therefore via Hoeffding's inequality

$$
\mathbb{P}_{\mathbf{z} \sim \rho^m} \left[\sup_{\theta \in B_k} |\widehat{L}(\theta) - L(\theta)| > \eta \right] \leq \mathbb{P}_{\mathbf{z} \sim \rho^m} \left[|\widehat{L}(\theta_k) - L(\theta_k)| > \frac{\eta}{2} \right] \leq 2e^{-\frac{m\eta^2}{2M_1^4}}.
$$

Notice that this inequality holds uniformly in k. Finally, since Θ is covered by the union of the balls $B_1, \ldots, B_N$, with $N = \mathcal{N}\left(\Theta, \frac{\eta}{8M_1 M_2}\right)$, we finally obtain

$$
\begin{aligned}
\mathbb{P}_{\mathbf{z} \sim \rho^m}\left[\sup_{\theta \in \Theta} |\widehat{L}(\theta) - L(\theta)| \leq \eta\right] &= 1 - \mathbb{P}_{\mathbf{z} \sim \rho^m}\left[\sup_{\theta \in \Theta} |\widehat{L}(\theta) - L(\theta)| > \eta\right] \\
&\geq 1 - \sum_{k=1}^{N} \mathbb{P}_{\mathbf{z} \sim \rho^m}\left[\sup_{\theta \in B_k} |\widehat{L}(\theta) - L(\theta)| > \eta\right] \\
&\geq 1 - 2Ne^{-\frac{m\eta^2}{2M_1^4}}. \qquad \square
\end{aligned}
$$

Lemma A.5 provides a very general result: in order to apply it to our current framework, we have to first show that the functional R_θ defined as in (12) satisfies the assumptions $a)$ and $b)$ in the statement. This is the subject of the following result.

Lemma A.6. *Under the assumptions of Section 4.1, let Θ be as in (19). then the family of functions $R_\theta\colon K^* \to X$, defined by (12), satisfies the assumptions of Lemma A.5.*

Proof. Without loss of generality we assume that $\|j_1\|_{\mathcal{L}(H,X)} \leq 1$ and $\|j_2\|_{\mathcal{L}(X,X)} \leq 1$. We first notice that, thanks to (21), for any $B_1, B_2 \in \Theta$,

$$
\begin{aligned}
\|((\Sigma_\varepsilon \iota)^{-1} A B_1)^* - ((\Sigma_\varepsilon \iota)^{-1} A B_2)^*\|_{\mathcal{L}(K^*,X)} &\leq \|(\Sigma_\varepsilon \iota)^{-1} A (B_1 - B_2)\|_{\mathcal{L}(X,K)} \\
&= \|(\Sigma_\varepsilon \iota)^{-1} A j_2 (\bar{\bar{B}}_1 - \bar{\bar{B}}_2) j_2^*\|_{\mathcal{L}(X,K)} \\
&\leq \|(\Sigma_\varepsilon \iota)^{-1} A j_2\|_{\mathcal{L}(X,K)} \|\bar{\bar{B}}_1 - \bar{\bar{B}}_2\|_{\mathcal{L}(X,X)}.
\end{aligned}
$$

Denote by $\varrho_3 = \|(\Sigma_\varepsilon \iota)^{-1} A j_2\|_{\mathcal{L}(X,K)}$. Note that, thanks to (21), arguing as in the proof of Proposition A.2 we have that $(\Sigma_\varepsilon \iota)^{-1} A j_2\colon X \to K$ is bounded. Then, by (19), for every $B_1, B_2 \in \Theta_2$,

$$
\begin{aligned}
\|((\Sigma_\varepsilon \iota)^{-1} A B_1)^* - ((\Sigma_\varepsilon \iota)^{-1} A B_2)^*\|_{\mathcal{L}(X,K)} &\leq \varrho_3 \|\bar{\bar{B}}_1 - \bar{\bar{B}}_2\|_{\mathcal{L}(X,X)}, \\
\|((\Sigma_\varepsilon \iota)^{-1} A B_1)^*\|_{\mathcal{L}(X,K)} &\leq \varrho_1 \varrho_3.
\end{aligned}
\tag{41}
$$

Assumption $a)$ in Lemma A.5 requires that $\|R_\theta(y) - x\|_X \leq M_1$ for a.e. $(x,y) \in X \times K^*$ and for all θ. Notice that, by the expression of $R_\theta = R_{h,B}$ in (12),

$$
\begin{aligned}
\|R_\theta(y) - x\|_X &\leq \|h\|_X + \|B\|_{\mathcal{L}(X,X)} \|(BA^* \Sigma_\varepsilon^{-1} AB + I_X)^{-1}\|_{\mathcal{L}(X,X)} \|((\Sigma_\varepsilon \iota)^{-1} AB)^*\|_{\mathcal{L}(K^*,X)} \\
&\quad \cdot (\|y\|_{K^*} + \|\iota^* A\|_{\mathcal{L}(X,K^*)} \|h\|_X) + \|x\|_X \\
&\leq \varrho_1 + \varrho_1^2 \varrho_3 (\varrho_2 + \|\iota^* A\|_{\mathcal{L}(X,K^*)} \varrho_1) + \rho_2 =: M_1,
\end{aligned}
$$

where we have also used (19), (22), (41) and the fact that the norm of $(BA^* \Sigma_\varepsilon^{-1} AB + I_X)^{-1}$ is less than or equal to 1.

Assumption $b)$ requires instead that $\|R_{\theta_1}(y) - R_{\theta_2}(y)\|_X \leq M_2 \|\theta_1 - \theta_2\|_*$. According to the definition of $\|\cdot\|_*$, we can decouple the perturbation of θ and study separately the perturbation of h and of B. We observe that

$$
R_{h_1,B}(y) - R_{h_2,B}(y) = (h_1 - h_2) - B(BA^* \Sigma_\varepsilon^{-1} AB + I_X)^{-1} ((\Sigma_\varepsilon \iota)^{-1} AB)^* \iota^* A (h_1 - h_2),
$$

hence again by (41), (22) and (19) we get

$$
\|R_{h_1,B}(y) - R_{h_2,B}(y)\| \leq (1 + \varrho_1^2 \varrho_3 \|\iota^* A\|_{\mathcal{L}(X,K^*)}) \|h_1 - h_2\|_X.
$$

24

The treatment of the perturbations of B is slightly more delicate. Let $C_i = (B_i A^* \Sigma_\varepsilon^{-1} A B_i + I_X)^{-1}$. Then we have

$$
\begin{aligned}
R_{h,B_1}(y) &- R_{h,B_2}(y) \\
&= (B_1 C_1 ((\Sigma_\varepsilon \iota)^{-1} A B_1)^* - B_2 C_2 ((\Sigma_\varepsilon \iota)^{-1} A B_2)^*)(y - \iota^* A h) \\
&= (B_1 - B_2) C_1 ((\Sigma_\varepsilon \iota)^{-1} A B_1)^*(y - \iota^* A h) + B_2 (C_1 - C_2)((\Sigma_\varepsilon \iota)^{-1} A B_1)^*(y - \iota^* A h) \\
&\quad + B_2 C_2 (((\Sigma_\varepsilon \iota)^{-1} A B_1)^* - ((\Sigma_\varepsilon \iota)^{-1} A B_2)^*)(y - \iota^* A h)
\end{aligned}
$$

In the latter summation, by means of (41), (22) and (19) we easily get that the first and the third terms are both bounded by $\varrho_1 \varrho_3 (\varrho_2 + \|\iota^* A\|_{\mathcal{L}(X,K^*)} \varrho_1) \|\bar{\bar{B}}_1 - \bar{\bar{B}}_2\|_{\mathcal{L}(X,X)}$. The second term can be reformulated taking into account that

$$
\begin{aligned}
C_1 - C_2 &= (I_X + B_1 A^* \Sigma_\varepsilon^{-1} A B_1)^{-1} - (I_X + B_2 A^* \Sigma_\varepsilon^{-1} A B_2)^{-1} \\
&= (I_X + B_1 A^* \Sigma_\varepsilon^{-1} A B_1)^{-1} (B_2 A^* \Sigma_\varepsilon^{-1} A B_2 - B_1 A^* \Sigma_\varepsilon^{-1} A B_1)(I_X + B_2 A^* \Sigma_\varepsilon^{-1} A B_2)^{-1},
\end{aligned}
$$

and its norm can be bounded by $2\varrho_1^3 \varrho_3^2 \|\iota^* A\|_{\mathcal{L}(X,K^*)} (\varrho_2 + \|\iota^* A\|_{\mathcal{L}(X,K^*)} \varrho_1) \|\bar{\bar{B}}_1 - \bar{\bar{B}}_2\|_{\mathcal{L}(X,X)}$ using similar arguments. $\qquad \square$

Now that the assumptions $a)$ and $b)$ of Lemma A.5 are guaranteed, we have to show the compactness of the parameter class Θ. The following lemma only assumes that Θ is defined as in (19), by means of a Hilbert space H and a compact, dense-range operator $j \colon H \to X$.

Lemma A.7. *The set Θ defined as in (19) is a compact subset of $X \times \{j_2 \bar{\bar{B}} j_2^* \colon \bar{\bar{B}} \colon X \to X$ bounded$\}$ with respect to the topology induced by the norm $\|\cdot\|_*$.*

Proof. We first show that $\Theta_1 \times \Theta_2$ is compact. Set

$$
\bar{\bar{\Theta}}_2 = \{j_1 \bar{B} j_1^* \colon \ \bar{B} \in \mathrm{HS}(H^*, H), \|\bar{B}\|_{\mathrm{HS}(H^*,H)} \le \varrho_1\} \tag{42}
$$

so that $\Theta_2 = \{j_2 \bar{\bar{B}} j_2^* \colon \ \bar{\bar{B}} \in \bar{\bar{\Theta}}_2\}$. The definition of the norm $\|\cdot\|_*$ implies that $\Theta_1 \times \Theta_2$ is compact with respect to the topology induced by the norm $\|\cdot\|_*$ if and only if $\Theta_1 \times \bar{\bar{\Theta}}_2$ is compact as subset of $X \times \mathrm{HS}(X,X)$ endowed with the product topology. Hence, it is enough to show that Θ_1 and $\bar{\bar{\Theta}}_2$ are compact in X and $\mathrm{HS}(X,X)$, respectively. By definition, since j is compact and Θ_1 is the image of the closed ball of radius ρ_1 in H, then Θ_1 is compact.

In order to prove that $\bar{\bar{\Theta}}_2$ is compact, we identify $\mathrm{HS}(H^*, H)$ and $\mathrm{HS}(X,X)$ with $H \otimes H$ and $X \otimes X$, respectively, so that for all $v, w \in H$, $v \otimes w \colon H^* \to H$ is the rank one operator

$$
(v \otimes w)(z) = \langle z, w \rangle_{H^*, H} \, v, \qquad z \in H^*.
$$

With this identification, since

$$
j_1 (v \otimes w) j_1^* = (j_1 v) \otimes (j_1 w),
$$

the map $\bar{B} \mapsto j_1 \bar{B} j_1^*$ is given by

$$
j_1 \otimes j_1 \colon H \otimes H \to X \otimes X,
$$

which is compact, since j_1 is so. As above, $\bar{\bar{\Theta}}_2$ is the image of the closed ball of radius ρ_1 in $H \otimes H$, so that it is a compact subset of $\mathrm{HS}(X,X)$.

The compactness of Θ follows from the fact that the subset of positive operators $\bar{\bar{B}} \colon X \to X$ is closed in $\mathrm{HS}(X,X)$ and $B = j_2 \bar{\bar{B}} j_2^*$ is positive if and only if $\bar{\bar{B}}$ is positive.

$\qquad \square$

To conclude the proof of Theorem 4.1, we need to provide an explicit expression for the covering numbers of the set Θ in the $\|\cdot\|_*$ norm. This is possible, e.g. by assuming the polynomial decay of the singular values of j_1 as in (20), by means of some tools that are presented in the next section.

A.5   Entropy numbers, singular values and covering numbers

Let $\mathcal{H}$ and $\mathcal{X}$ be real Hilbert spaces and let $\mathcal{B}$ denote the unit closed ball in $\mathcal{H}$. We use instead the notation $B(v, \varepsilon)$ to denote the closed ball in $\mathcal{X}$ with center v and radius ε. For any compact operator $T \colon \mathcal{H} \to \mathcal{X}$ we can define the following quantities.

1. **Entropy numbers:** for each $k \in \mathbb{N}$, $k \geq 1$,
$$\varepsilon_k(T) = \inf\{\varepsilon > 0 \mid \exists v_1, \ldots, v_k \in \mathcal{X} \text{ such that } \cup_{i=1}^k B(v_i, \varepsilon) \supseteq T(\mathcal{B})\};$$

2. **Singular values:** $s_k(T) = \lambda_k(|T|)$, where $\lambda_k(|T|)$ is the k-th non-zero eigenvalue of $|T|$, which are counted with their multiplicity and ordered in a non-increasing way. If $|T|$ has less than K non-zero eigenvalues, then $s_k(T) = 0$ for $k \geq K$.

3. **Covering numbers** of T: the covering numbers of the set $T(\mathcal{B})$; namely, for $r > 0$,
$$\mathcal{N}_r(T) = \mathcal{N}(T(\mathcal{B}), r) = \inf\{k \in \mathbb{N}_* \mid \exists v_1, \ldots, v_k \in \mathcal{X} \text{ such that } \cup_{i=1}^k B(v_i, r) \supseteq T(\mathcal{B})\}.$$

Properties of covering and entropy numbers have recently been used in the study of instability in inverse problems [28]. We have the following results (see [11]):
$$\varepsilon_k(T) \leq r \qquad \Longleftrightarrow \qquad \mathcal{N}_r(T) \leq k; \tag{43}$$

and for all $k \in \mathbb{N}$, $k \geq 1$
$$\sup_{1 \leq \ell < \infty} \left(k^{-1/\ell} \left(\Pi_{i=1}^{\ell} s_i(T) \right)^{1/\ell} \right) \leq \varepsilon_k(T) \leq 14 \sup_{1 \leq \ell < \infty} \left(k^{-1/\ell} \left(\Pi_{i=1}^{\ell} s_i(T) \right)^{1/\ell} \right). \tag{44}$$

We now use these properties to quantify the covering numbers $\mathcal{N}(\Theta, r)$ appearing in Lemma A.5. We assume, for simplicity, that $\varrho_1 = 1$. For $\varrho_1 \neq 1$, we can rescale the covering numbers by the formula $\mathcal{N}(\varrho B, r) = \mathcal{N}(B, r/\varrho)$, where $\varrho B = \{\varrho b : b \in B\}$. By the definition of Θ, it is evident that $\mathcal{N}(\Theta, r) \leq \mathcal{N}(\Theta_1, r)\mathcal{N}(\Theta_2, r)$. The following two lemmas take care of estimating the covering numbers of Θ_1 and Θ_2, respectively.

Lemma A.8. *Under Assumption* (20) *we have*
$$\ln(\mathcal{N}(\Theta_1, r)) \leq Cr^{-\frac{1}{s}}, \qquad r > 0, \tag{45}$$

where $C > 0$ is independent of r.

Proof. Observe that $\mathcal{N}(\Theta_1, r) = \mathcal{N}_r(j) \leq \mathcal{N}_{\frac{r}{\|j_2\|_{\mathcal{L}(X,X)}}}(j_1)$. Condition (20) yields
$$\Pi_{i=1}^{\ell} s_i(j_1) \lesssim (\ell!)^{-s} \lesssim \ell^{-s\ell} e^{s\ell},$$

where the last bound is a consequence of the fact that $(\ell)! \geq e\ell^{\ell} e^{-\ell}$. Estimate (44) implies that
$$\varepsilon_k(j_1) \lesssim \sup_{1 \leq \ell < \infty} \left(k^{-1/\ell} \left(\ell^{-s\ell} e^{s\ell} \right)^{1/\ell} \right) = \sup_{1 \leq \ell < \infty} \left(k^{-1/\ell} \ell^{-s} e^{s} \right).$$

Let $t = 1/\ell$, since the function $e^{-t \ln k} t^s$ takes its maximum at $t = s/\ln k$, then
$$\varepsilon_k(j_1) \lesssim (\ln k)^{-s},$$

where the constant in $\lesssim$ depends on s. Eq. (43) yields $\mathcal{N}_{(\ln k)^{-s}}(j_1) \lesssim k$ and ultimately the thesis. $\qquad\square$

Lemma A.9. *Under assumption (20), for every $s' \in (0, s)$ we have*

$$\ln(\mathcal{N}(\Theta_2, r)) \leq C r^{-\frac{1}{s'}}, \qquad r > 0, \tag{46}$$

where $C > 0$ is independent of r.

Proof. Observe that $\mathcal{N}(\Theta_2, r) = \mathcal{N}(\bar{\bar{\Theta}}_2, r) = \mathcal{N}_r(j_1 \otimes j_1)$, since $j_1 \otimes j_1$ represents the (compact) embedding of $\mathrm{HS}(H^*, H)$ into $\mathrm{HS}(X, X)$, see the proof of Lemma A.7 and (42).

We bound the singular values of $j_1 \otimes j_1$. Let $f \colon (0, +\infty) \to \mathbb{N}$ be defined by

$$f(t) = \#\{(k_1, k_2) \in \mathbb{N}_* \times \mathbb{N}_* : s_{k_1, k_2}(j_1 \otimes j_1) = s_{k_1}(j_1) s_{k_2}(j_1) \geq t\}.$$

Then, for all $k \in \mathbb{N}_*$ we have

$$s_k(j_1 \otimes j_1) = \sup\{t \in (0, +\infty) : f(t) \geq k\}.$$

By the polynomial decay of the singular values (20)

$$f(t) \leq \#\{(k_1, k_2) \in \mathbb{N}_* \times \mathbb{N}_* : (k_1 k_2)^{-s} \geq Ct\} = g(t),$$

where C is a suitable constant. Then,

$$s_k(j_1 \otimes j_1) \leq \sup\{t \in (0, +\infty) : g(t) \geq k\}.$$

We now estimate $g(t)$. Let $\tau = (Ct)^{-1/s}$ and $N = [\tau]$ be the integer part of τ. Fix $0 < \varepsilon < 1$

$$g(t) = \#\{(k_1, k_2) \in \mathbb{N}_* \times \mathbb{N}_* : k_1 k_2 \leq (Ct)^{-1/s} = \tau\}$$

$$= \sum_{k_1=1}^{N} \left[\frac{\tau}{k_1}\right] \leq \sum_{k_1=1}^{N} \frac{\tau}{k_1} \leq \tau + \int_1^N \frac{\tau}{x} dx$$

$$\leq \tau(1 + \ln(\tau)) \leq \frac{1}{\varepsilon} \tau^{1+\varepsilon} \lesssim \frac{1}{\varepsilon} t^{-(1+\varepsilon)/s}$$

where $(1 + \ln x) \leq \varepsilon^{-1} x^\varepsilon$ provided that $0 < \varepsilon < 1$. Set $s/2 \leq s' = s/(1 + \varepsilon) < s$, then

$$g(t) \lesssim \frac{s}{s - s'} t^{-\frac{1}{s'}}$$

so that

$$s_k(j_1 \otimes j_1) \lesssim k^{-s'}. \tag{47}$$

Clearly, the above bound holds true also when $0 < s' < s/2$. The proof follows by repeating the argument of the proof of Lemma A.8. $\square$

We are now able to prove the main result of section 4.1.

Proof of Theorem 4.1. By Lemma A.8 and A.9, we conclude that

$$\ln(\mathcal{N}(\Theta, r)) \leq \ln(\mathcal{N}(\Theta_1, r)) + \ln(\mathcal{N}(\Theta_2, r)) \lesssim r^{-\frac{1}{s}} + r^{-\frac{1}{s'}} \lesssim r^{-\frac{1}{s'}}.$$

Substituting this result in Lemma A.5 allows to conclude that

$$\mathbb{P}_{\mathbf{z} \sim \rho^m}\left[|L(\widehat{\theta}_{\mathbf{z}}) - L(\theta^*)| \leq \eta\right] \geq 1 - e^{\tilde{c}_1 \eta^{-1/s'} - \tilde{c}_2 m \eta^2} = 1 - e^{-\tau}.$$

27

We can express η as a function of m and τ when $\eta < 1$ by the following estimate:

$$\tau \geq \tilde{c}_2 m \eta^2 - \tilde{c}_1 \eta^{-1/s'} \quad \Rightarrow \quad \tilde{c}_2 m \eta^{2+\frac{1}{s'}} \leq \tilde{c}_1 + \tau \eta^{1/s'} \leq \tilde{c}_1 + \tau,$$

and therefore (with constants c_1, c_2 independent of m, τ, η)

$$|L(\widehat{\theta}_{\mathbf{z}}) - L(\theta^*)| \leq \eta \leq \left(\frac{\tilde{c}_1 + \tau}{\tilde{c}_2 m} \right)^{\frac{1}{2+1/s'}} \leq \left(\frac{c_1 + c_2 \sqrt{\tau}}{\sqrt{m}} \right)^{1 - \frac{1}{2s'+1}},$$

with probability larger than or equal to $1 - e^{-\tau}$. $\qquad\square$

A.6 Proof of Theorem 4.2

In the unsupervised setting, the regularizer $R_{\widehat{h}_U, \widehat{B}_U}$ is given by

$$R_{\widehat{h}_U, \widehat{B}_U}(y) = \widehat{h}_U + \widehat{B}_U^2 A^* (\iota^* (A \widehat{B}_U^2 A^* + \Sigma_\varepsilon))^{-1} (y - \iota^* A \widehat{h}_U), \qquad y \in K^*,$$

where

$$\widehat{h}_U = \hat{\mu} = \frac{1}{m} \sum_{j=1}^m x_j, \qquad \widehat{B}_U^2 = \widehat{\Sigma}_x = \frac{1}{m} \sum_{j=1}^m (x_j - \hat{\mu}) \otimes (x_j - \hat{\mu}).$$

Hence, in order to analyze the statistical properties of $R_{\widehat{h}_U, \widehat{B}_U}$, we first provide two concentration inequalities for $\hat{\mu}$ and $\widehat{\Sigma}_x$, which are known since x is a sub-Gaussian random vector in X. We include the proofs for the sake of completeness.

Lemma A.10. *Let x be a κ sub-Gaussian vector as in (24). Fix $\tau > 0$, then, with probability exceeding $1 - 2e^{-\tau}$,*

$$\|\hat{\mu} - \mu\| \leq c\kappa \left(\sqrt{\frac{\operatorname{tr}(\Sigma_x)}{m}} + 2\sqrt{\frac{\tau \|\Sigma_x\|}{m}} \right), \tag{48}$$

where $c > 0$ is a universal constant.

Proof. Define $\xi = x - \mu$ where $\mu = \mathbb{E}[x]$. It is easy to show that ξ is a zero mean κ sub-Gaussian vector and its covariance matrix is $\Sigma_x = \mathbb{E}[\xi \otimes \xi]$. The assumption that ξ is a sub-Gaussian random variable (24) can be equivalently expressed by requiring that

$$\|\langle \xi, v \rangle_X\|_{\psi_2} \leq \kappa \|\langle \xi, v \rangle_X\|_2,$$

where $\|\langle \xi, v \rangle_X\|_{\psi_2} = \sup_{p \geq 2} \frac{\|\langle \xi, v \rangle_X\|_p}{\sqrt{p}}$. For each v in the unit ball B_1 of X, we set $\xi_v = \langle \xi, v \rangle_X$ and we regard $(\xi_v)_{v \in B_1}$ as a random process on B_1, viewed as metric space with respect to the metric $\mathrm{d}(v, w) = \|\xi_v - \xi_w\|_2$. Since $B_1 = -B_1$, then

$$\|\xi\|_X = \sup_{v \in B_1} |\langle \xi, v \rangle| = \sup_{v \in B_1} \langle \xi, v \rangle$$

and a standard result of random processes – see Exercise 8.6.5 and Theorem 8.5.5 in [48] – gives that

$$\mathbb{P} \left(\sup_{v \in B_1} |\xi_v| \leq c\kappa (W(B_1) + t \operatorname{diam}(B_1)) \right) \geq 1 - 2e^{-t^2}, \qquad t > 0,$$

where c is a universal constant, W is the width of the process

$$W(B_1) = \mathbb{E}[\sup_{v \in B_1} \langle \xi, v \rangle_X] = \mathbb{E}[\sup_{v \in B_1} |\langle \xi, v \rangle_X|] = \mathbb{E}[\|\xi\|_X],$$

28

and $\mathrm{diam}(B_1)$ is the diameter of B_1 with respect to the metric $\mathrm{d}(v, w)$

$$\mathrm{diam}(B_1) = \sup_{v,w \in B_1} \mathrm{d}(v, w) = \sup_{v,w \in B_1} \|\langle \xi, v \rangle - \langle \xi, w \rangle\|_2 = \sup_{v,w \in B_1} \mathbb{E}[\langle \xi, v - w \rangle^2]^{1/2}$$

$$= \sup_{v,w \in B_1} (\langle \Sigma_\xi(v - w), v - w \rangle)^{1/2} \leq 2\|\Sigma_\xi\|_{\mathcal{L}(X,X)}^{1/2}.$$

Hölder's inequality implies that

$$\mathbb{E}[\|\xi\|_X] \leq \mathbb{E}[\|\xi\|_X^2]^{1/2} = (\mathrm{tr}(\Sigma_x))^{1/2},$$

so that we get

$$\mathbb{P}\left(\|\xi\|_X \leq c\kappa \left(\sqrt{\mathrm{tr}\,\Sigma_x} + 2t\sqrt{\|\Sigma_x\|}\right)\right) \geq 1 - 2e^{-t^2}. \tag{49}$$

Define $\xi_1 = x_1 - \mu, \ldots, \xi_m = x_m - \mu$, which are i.i.d. as ξ. We claim that there exists an absolute constant d such that $\widehat{\xi} = \frac{1}{m}\sum_{j=1}^m \xi_j$ is $d\kappa$ sub-Gaussian. Indeed,

$$\|\langle \sum_{j=1}^m \xi_j, v \rangle\|_{\psi_2}^2 = \|\sum_{j=1}^m \langle \xi_j, v \rangle\|_{\psi_2}^2$$

$$\leq d^2 \sum_{j=1}^m \|\langle \xi_j, v \rangle\|_{\psi_2}^2 \leq d^2\kappa^2 \sum_{j=1}^m \|\langle \xi_j, v \rangle\|_2^2 = d^2\kappa^2 \|\sum_{j=1}^m \langle \xi_j, v \rangle\|_2^2,$$

where the first inequality is due to the rotational invariance property of sub-Gaussian real random variables [48, Proposition 2.6.1], and the last equality is a consequence of the independence of the variables ξ_j. Thus $\widehat{\xi} = \widehat{\mu} - \mu$ is $d\kappa$ sub-Gaussian. Notice that the covariance of $\widehat{\xi}$ is given by

$$\mathbb{E}[\widehat{\xi} \otimes \widehat{\xi}] = \frac{1}{m^2}\sum_{i,j} \mathbb{E}[\xi_i \otimes \xi_j] = \frac{1}{m}\sum_j \mathbb{E}[\xi_j \otimes \xi_j] = \frac{1}{m}\Sigma_x.$$

Setting $\tau = t^2$, by applying (49) to $\widehat{\xi}$ we can finally deduce that

$$\mathbb{P}\left(\|\widehat{\xi}\|_X \leq cd\kappa \left(\sqrt{\frac{\mathrm{tr}\,\Sigma_x}{m}} + 2\sqrt{\frac{\tau\|\Sigma_x\|}{m}}\right)\right) \geq 1 - 2e^{-\tau},$$

which provides the claimed bound by redefining the universal constant c. $\qquad \square$

The following lemma is a restatement of a fundamental result in [30]. We include in the statement also the previous inequality.

Lemma A.11. *Let x be a κ sub-Gaussian vector as in (24). Fix $\tau > 1$, then, with probability exceeding $1 - 3e^{-\tau}$,*

$$\|\widehat{\Sigma}_x - \Sigma_x\| \leq c\kappa^2\|\Sigma_x\| \max\left\{\sqrt{\frac{\mathrm{tr}\,\Sigma_x}{m\|\Sigma_x\|}}, \frac{\mathrm{tr}\,\Sigma_x}{m\|\Sigma_x\|}, \sqrt{\frac{\tau}{m}}, \frac{\tau}{m}\right\}, \tag{50}$$

$$\|\widehat{\mu} - \mu\| \leq c\kappa \left(\sqrt{\frac{\mathrm{tr}(\Sigma_x)}{m}} + 2\sqrt{\frac{\tau\|\Sigma_x\|}{m}}\right), \tag{51}$$

where c is a universal constant.

Proof. We first introduce the operator

$$\widehat{\Sigma}_\xi = \frac{1}{m} \sum_{j=1}^{m} \xi_j \otimes \xi_j = \frac{1}{m} \sum_{j=1}^{m} (x_j - \mu) \otimes (x_j - \mu).$$

Since $\mathbb{E}[\widehat{\Sigma}_\xi] = \Sigma_x$, Theorem 9 of [30] gives that

$$\|\widehat{\Sigma}_\xi - \Sigma_x\| \leq c'\|\Sigma_x\| \max\left\{ \sqrt{\frac{\operatorname{tr}\Sigma_x}{m\|\Sigma_x\|}}, \frac{\operatorname{tr}\Sigma_x}{m\|\Sigma_x\|}, \sqrt{\frac{\tau}{m}}, \frac{\tau}{m} \right\}, \tag{52}$$

with probability greater than $1 - e^{-\tau}$. As usual, it holds that

$$\widehat{\Sigma}_x = \frac{1}{m} \sum_{j=1}^{m} (x_j - \widehat{\mu}) \otimes (x_j - \widehat{\mu}) = \frac{1}{m} \sum_{j=1}^{m} (x_j - \mu + \mu - \widehat{\mu}) \otimes (x_j - \mu + \mu - \widehat{\mu})$$

$$= \frac{1}{m} \sum_{j=1}^{m} (x_j - \mu) \otimes (x_j - \mu) + (\mu - \widehat{\mu}) \otimes \left(\frac{1}{m} \sum_{j=1}^{m} (x_j - \mu) \right)$$

$$+ \left(\frac{1}{m} \sum_{j=1}^{m} (x_j - \mu) \right) \otimes (\mu - \widehat{\mu}) + (\widehat{\mu} - \mu) \otimes (\widehat{\mu} - \mu)$$

$$= \widehat{\Sigma}_\xi - (\widehat{\mu} - \mu) \otimes (\widehat{\mu} - \mu).$$

As a consequence,

$$\|\widehat{\Sigma}_x - \Sigma_x\| \leq \|\widehat{\Sigma}_\xi - \Sigma_x\| + \|\widehat{\mu} - \mu\|^2.$$

By (52) and (48), with probability exceeding $1 - 3e^{-\tau}$, we have both (51) and

$$\|\widehat{\Sigma}_x - \Sigma_x\| \leq c'\|\Sigma_x\| \max\left\{ \sqrt{\frac{\operatorname{tr}\Sigma_x}{m\|\Sigma_x\|}}, \frac{\operatorname{tr}\Sigma_x}{m\|\Sigma_x\|}, \sqrt{\frac{\tau}{m}}, \frac{\tau}{m} \right\}$$

$$+ c\kappa^2 \left(\frac{\operatorname{tr}(\Sigma_x)}{m} + 4\frac{\sqrt{\tau\|\Sigma_x\| \operatorname{tr}(\Sigma_x)}}{m} + 4\frac{\tau\|\Sigma_x\|}{m} \right),$$

which provides the claimed bounds by redefining the constants c. $\qquad\square$

The following lemma shows that the excess risk $L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star)$ is bounded by $\|\widehat{\Sigma}_x - \Sigma_x\|$ and $\|\widehat{\mu} - \mu\|$. Note that Lemma A.6 would provide a bound in terms of $\|\widehat{\Sigma}_x^{\frac{1}{2}} - \Sigma_x^{\frac{1}{2}}\|$. Since the square root is a monotone increasing function, it holds true that

$$\|\widehat{\Sigma}_x^{\frac{1}{2}} - \Sigma_x^{\frac{1}{2}}\| \leq \sqrt{\|\widehat{\Sigma}_x - \Sigma_x\|},$$

see Theorem X.1.1 of [5], which would provide a worse bound.

Lemma A.12. *Assume that $A\Sigma_x A^* + \Sigma_\varepsilon \colon Y \to Y$ has a bounded inverse, that the operator*

$$A^*(\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1} \colon \iota^*(Y) \subseteq K^* \to X$$

extends to a bounded operator from K^* to X, and that

$$\|(A\Sigma_x A^* + \Sigma_\varepsilon)^{-1} A(\widehat{\Sigma}_x - \Sigma_x)A^*\| \leq 1/2. \tag{53}$$

Then

$$|L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star)| = \mathrm{O}\left(\|\widehat{\Sigma}_x - \Sigma_x\|\right) + \mathrm{O}\left(\|\widehat{\mu} - \mu\|\right), \tag{54}$$

where the constant in O *only depends on* $A, \Sigma_x, \Sigma_\varepsilon, \iota$ *and* μ.

Proof. Let

$$\widehat{x}_U = R_{\widehat{h}_U, \widehat{B}_U}(y), \qquad x^\star = R_{h^\star, B^\star}(y),$$

so that

$$L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star) = \mathbb{E}\left[\|\widehat{x}_U - x\|^2\right] - \mathbb{E}\left[\|x^\star - x\|^2\right].$$

Since $x^\star$ minimizes the mean square error, clearly $L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star) \geq 0$. We now prove the upper bound. Since

$$\begin{aligned}
\|\widehat{x}_U - x\|^2 - \|x^\star - x\|^2 &= \|\widehat{x}_U - x^\star\|^2 + 2\langle \widehat{x}_U - x^\star, x^\star - x\rangle \\
&\leq \|\widehat{x}_U - x^\star\|^2 + 2\|\widehat{x}_U - x^\star\|\|x^\star - x\|,
\end{aligned}$$

then, by Hölder inequality,

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{x}_U - x\|^2\right] - \mathbb{E}\left[\|x^\star - x\|^2\right] &\leq \mathbb{E}\left[\|x^\star - \widehat{x}_U\|^2\right] + 2\sqrt{\mathbb{E}\left[\|\widehat{x}_U - x^\star\|^2\right]\mathbb{E}\left[\|x^\star - x\|^2\right]} \\
&= \mathrm{O}\left(\sqrt{\mathbb{E}\left[\|\widehat{x}_U - x^\star\|^2\right]}\right), \tag{55}
\end{aligned}$$

where the constant in O only depends on A, Σ_x and Σ_ε. By (15) and the definition of $\widehat{x}_U$

$$x^\star = \Sigma_x A^* (\iota^* (A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}(y - \iota^* A\mu) + \mu = \widetilde{W}y + \widetilde{b}$$
$$\widehat{x}_U = \widehat{\Sigma}_x A^* (\iota^* (A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1}(y - \iota^* A\widehat{\mu}) + \widehat{\mu} = \widehat{W}y + \widehat{b},$$

where $\widetilde{W}$ and $\widehat{W}$ are given in (33), for $B^2 = \Sigma_x$ and $B^2 = \widehat{\Sigma}_x$, respectively, and $\widetilde{b}, \widehat{b}$ in (35). As a consequence,

$$\widehat{x}_U - x^\star = (\widehat{W} - \widetilde{W})y + \widehat{b} - \widetilde{b} = Wy + b = W\iota^* A(x - \mu) + W\varepsilon + W\iota^* A\mu + b,$$

where

$$W = \widehat{W} - \widetilde{W}, \qquad b = \widehat{b} - \widetilde{b}.$$

Hence, taking into account that $x - \mu$ and ε are zero mean random variables, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{x}_U - x^\star\|^2\right]^{\frac{1}{2}} &= \left(\mathrm{tr}\left[W(\iota^* A\Sigma_x A^* \iota + \iota^* \Sigma_\varepsilon \iota)W^*\right] + \|W\iota^* A\mu + b\|^2\right]\right)^{\frac{1}{2}} \\
&\leq \mathrm{tr}\left[W(\iota^* A\Sigma_x A^* \iota + \iota^* \Sigma_\varepsilon \iota)W^*\right]^{\frac{1}{2}} + \|W\iota^* A\mu + b\| \\
&\leq (\mathrm{tr}\,(\iota^* A\Sigma_x A^* \iota + \iota^* \Sigma_\varepsilon \iota))^{\frac{1}{2}}\|W\| + \|W\iota^* A\mu + b\|. \tag{56}
\end{aligned}$$

We now bound the norm of W where

$$W = \widehat{\Sigma}_x A^* (\iota^* (A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1} - \Sigma_x A^* (\iota^* (A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}.$$

A delicate issue is that $(\iota^* (A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1}$ and $(\iota^* (A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}$ do not have bounded inverses, see the remark after Theorem 3.1. We first prove that $A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon$ has a bounded inverse. Indeed, let $\Delta = A(\widehat{\Sigma}_x - \Sigma_x)A^*$, then

$$A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon = (A\Sigma_x A^* + \Sigma_\varepsilon)\left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right).$$

By assumption (53), $\|(A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\| \le 1/2 < 1$, so that using Neumann series we have that $A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon$ is invertible and

$$(A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon)^{-1} - (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}$$
$$= \left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right)^{-1} (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1} - (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}$$
$$= \left(\left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right)^{-1} - I\right) (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}$$
$$= \left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right)^{-1} (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta(A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}.$$

Then, on $\iota^*(Y) \subseteq K^*$

$$(i^*(A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1} - (i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}$$
$$= \left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right)^{-1} (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}A(\widehat{\Sigma}_x - \Sigma_x)A^*(i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}.$$

The density of $\iota^*(Y) \subset K^*$ and the assumption that $A^*(i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}$ extends to a bounded operator from K^* to X implies that

$$\|(i^*(A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1} - (i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}\|$$
$$\le \| \left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right)^{-1} \| \, \|(A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}A\| \|\widehat{\Sigma}_x - \Sigma_x\| \|A^*(i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}\|$$
$$\le 2\|(A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}A\| \|\widehat{\Sigma}_x - \Sigma_x\| \|A^*(i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}\|,$$

where we used (53) to bound $\| \left(I + (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1}\Delta\right)^{-1} \|$ with 2, so that

$$\|(i^*(A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1} - (i^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}\| \lesssim \|\widehat{\Sigma}_x - \Sigma_x\|, \tag{57}$$

where the constant in $\lesssim$ only depends on A, Σ_x and Σ_ε. Since

$$W = \widehat{\Sigma}_x A^* \left((\iota^*(A\widehat{\Sigma}_x A^* + \Sigma_\varepsilon))^{-1} - (\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1}\right)$$
$$+ (\widehat{\Sigma}_x - \Sigma_x)A^*(\iota^*(A\Sigma_x A^* + \Sigma_\varepsilon))^{-1},$$

eq. (57) and the fact that $\|\widehat{\Sigma}_x\| \le \|\widehat{\Sigma}_x - \Sigma_x\| + \|\Sigma_x\|$ both imply

$$\|W\| \lesssim \|\widehat{\Sigma}_x - \Sigma_x\| + \|\widehat{\Sigma}_x - \Sigma_x\|^2 = \mathrm{O}\left(\|\widehat{\Sigma}_x - \Sigma_x\|\right), \tag{58}$$

where the constants in $\lesssim$ and O only depend on A, Σ_x and Σ_ε. We now observe that

$$b = (\widehat{\mu} - \mu) - (\widehat{W}\iota^*A\widehat{\mu} - \widetilde{W}\iota^*A\mu) = (\widehat{\mu} - \mu) - \widehat{W}\iota^*A(\widehat{\mu} - \mu) - W\iota^*A\mu,$$

so that

$$W\iota^*A\mu + b = (I - \widehat{W}\iota^*A)(\widehat{\mu} - \mu),$$

and

$$\|W\iota^*A\mu + b\| \le \|(I - \widehat{W}\iota^*A)\| \, \|\widehat{\mu} - \mu\| \le (\|(I - \widetilde{W}\iota^*A)\| + \|W\iota^*A\|) \, \|\widehat{\mu} - \mu\|.$$

Eq. (58) implies that

$$\|W\iota^*A\mu + b\| \lesssim \|\widehat{\mu} - \mu\| + \|\widehat{\mu} - \mu\| \|\widehat{\Sigma}_x - \Sigma_x\| = \mathrm{O}\left(\|\widehat{\mu} - \mu\|\right), \tag{59}$$

where the constants in $\lesssim$ and O only depend on $A, \Sigma_x, \Sigma_\varepsilon$ and μ. Eqs. (55) and (56) with (59) give (54). $\square$

We are now able to prove the main result of section 4.2.

Proof of Theorem 4.2 . Since the map $C \mapsto (A\Sigma_x A^* + \Sigma_\varepsilon)^{-1} ACA^*$ is continuous from $\mathcal{L}(X,X)$ into $\mathcal{L}(Y,Y)$, there exists $\delta > 0$ such that

$$\|(A\Sigma_x A^* + \Sigma_\varepsilon)^{-1} ACA^*\| \leq 1/2 \qquad \forall C \in \mathcal{L}(X,X) \quad \|C\| \leq \delta.$$

Set $m_0 \in \mathbb{N}$ such that

$$c\kappa^2 \|\Sigma_x\| \max \left\{ \sqrt{\frac{\operatorname{tr}\Sigma_x}{m_0 \|\Sigma_x\|}}, \frac{\operatorname{tr}\Sigma_x}{m_0 \|\Sigma_x\|}, \sqrt{\frac{\tau}{m_0}}, \frac{\tau}{m_0} \right\} \leq \delta,$$

where c is the constant in Lemma A.11. Eq. (50) implies that for all $m \geq m_0$ condition (53) is satisfies with probability exceeding $1 - 4e^{-\tau}$. Possibly redefining m_0, by (50) and (51) we can assume that on the same event

$$\max\{\|\widehat{\mu} - \mu\|, \|\widehat{\Sigma}_x - \Sigma_x\|\} \leq \min\{1, \frac{c_1 + c_2\sqrt{\tau}}{\sqrt{m}}\}, \tag{60}$$

where c_1 and c_2 are suitable constants independent of m and τ. Hence, eq. (54) implies that on the same event

$$|L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star)| = O\left(\|\widehat{\Sigma}_x - \Sigma_x\|\right) + O\left(\|\widehat{\mu} - \mu\|\right) \leq C\frac{c_1 + c_2\tau}{\sqrt{m}},$$

where the last inequality is a consequence of (60). Eq. (25) is now clear. $\qquad\square$

A.7 Numerical results: further details

A.7.1 Experimental setup

In Section 5, we set $X = L^2(\mathbb{T}^1)$, being $\mathbb{T}^1$ the one-dimensional torus. For any $N > 0$, we can introduce the partition $\{I_{N,i}\}_{i=1}^N$ of the interval $(0,1)$, being $I_{N,i} = \left(\dfrac{i-1}{N}, \dfrac{i}{N}\right)$ and define the 1D-pixel basis $\{\varphi_{N,i}\}_{i=1}^N$ as follows:

$$\varphi_{N,i}(t) = \sqrt{N}\chi_{N,i}(t), \qquad \chi_{N,i}(t) = \begin{cases} 1 & t \in I_{N,i}, \\ 0 & \text{otherwise.} \end{cases}$$

The functions $\{\varphi_{N,i}\}_{i=1}^N$ form an orthogonal set, and we define X_N as the linear space generated by them. Each element $u \in X_N$ can be uniquely represented by a vector $\boldsymbol{u} \in \mathbb{R}^N$ as follows:

$$\boldsymbol{u}_i = \frac{1}{|I_{N,i}|} \int_{I_{N,i}} u = \sqrt{N}\langle u, \varphi_{N,i}\rangle_X, \qquad u(t) = \sum_{i=1}^N \langle u, \varphi_{N,i}\rangle_X \varphi_{N,i}(t) = \sum_{i=1}^N \frac{1}{\sqrt{N}}\boldsymbol{u}_i \varphi_{N,i}(t).$$

As a consequence, for $u \in X_N$, we can compute $\|u\|_X^2 = \sum_{i=1}^N \langle u, \varphi_{N,i}\rangle^2 = \frac{1}{N}\sum_{i=1}^N \boldsymbol{u}_i^2$. The representation of a linear operator $B\colon X_N \to X_N$ can be done via a matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ as follows:

$$\boldsymbol{B}_{i,j} = \langle B\varphi_{N,j}, \varphi_{N,i}\rangle_X; \qquad v = Bu \iff \boldsymbol{v} = \boldsymbol{B}\boldsymbol{u}.$$

In order to generate a discrete version of the random process ε and of the random variable x, we first generate the vectors $\boldsymbol{\nu}_x, \boldsymbol{\nu}_\varepsilon$ such that each component $[\boldsymbol{\nu}_x]_i$ and $[\boldsymbol{\nu}_\varepsilon]_i$ is independently
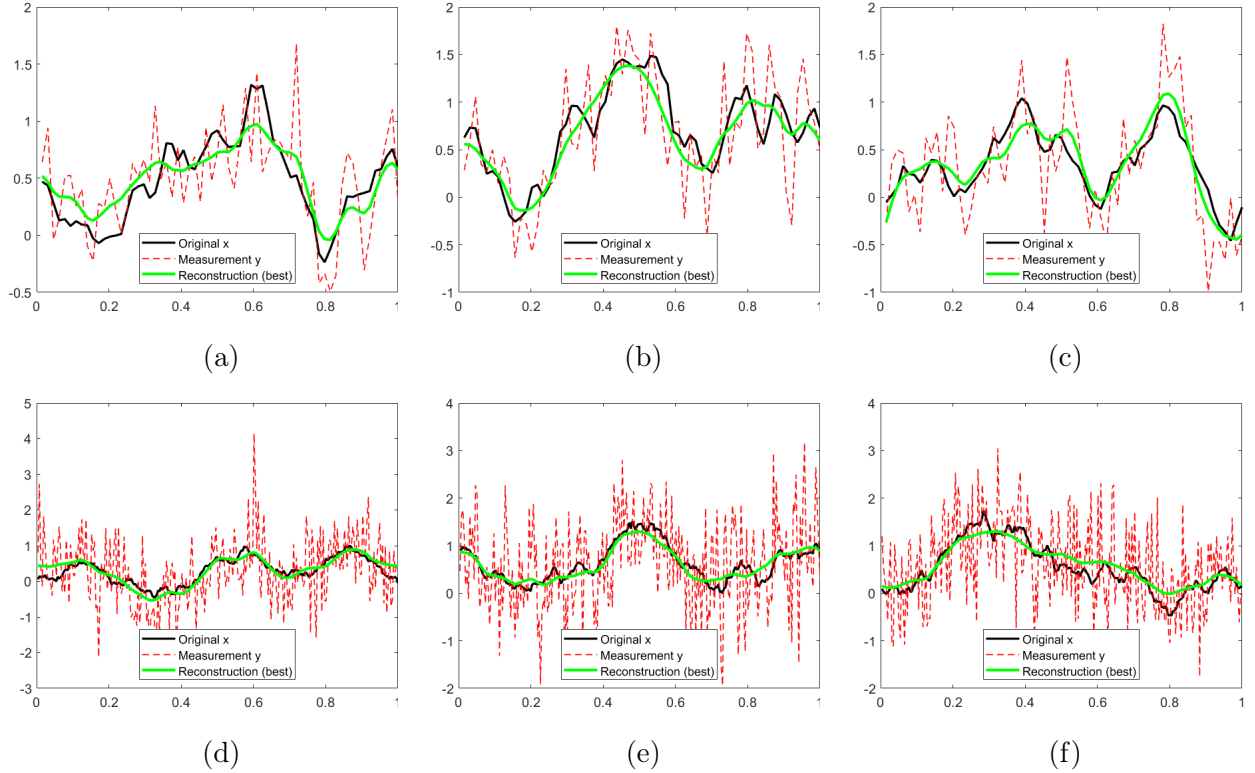
Figure 3: Signals drawn from the joint distribution in the case where both x and ε are Gaussian. (a),(b),(c): $N = 64$, (d),(e),(f): $N = 256$. We show in black the original signal x, in red the noisy datum $y = x + \varepsilon$ and in green the reconstruction $R_{\theta^\star}(y)$ associated with the optimal regularizer.

distributed with mean 0 and covariance 1. In the proposed tests, we either draw from a Gaussian distribution $\mathcal{N}(0, 1)$ or a uniform distribution $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$, taking advantage of the Matlab commands `randn` and `rand`. Then in order to approximate the white noise process ε, with zero mean and covariance operator $\Sigma_\varepsilon = \sigma^2 I$, we introduce $\boldsymbol{\varepsilon}$ such that

$$\mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j] = \mathbb{E}[\sqrt{N}\langle \varepsilon, \varphi_{N,i}\rangle \sqrt{N}\langle \varepsilon, \varphi_{N,j}\rangle] = \sigma^2 N \delta_{ij},$$

thus resulting in $\boldsymbol{\varepsilon} = \sigma\sqrt{N}\boldsymbol{\nu}_\varepsilon$. As an alternative, we also consider a random process whose components with respect to the Haar wavelet basis are randomly sampled as a white noise, i.e., $\boldsymbol{\varepsilon} = \sigma\sqrt{N}\boldsymbol{W}^T \boldsymbol{\nu}_\varepsilon$, where $\boldsymbol{W}$ is the discrete Haar wavelet transform and $\boldsymbol{W}^T$ its transpose.

The random variable $\boldsymbol{x}$ is instead computed as $\boldsymbol{x} = \boldsymbol{\mu} + \sqrt{N}\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\nu}_x$, being

$$\boldsymbol{\mu}_i = \sqrt{N}\langle \mu, \varphi_{N,i}\rangle_X, \qquad [\boldsymbol{\Sigma}_x^{1/2}]_{i,j} = \langle \Sigma_x^{1/2}\varphi_{N,j}, \varphi_{N,i}\rangle_X.$$

In the experiments, we picked $\mu(t) = 1 - |2t - 1|$ and $\Sigma_x^{1/2}$ s.t.

$$\Sigma_x^{1/2} u(t) = \int_{\mathbb{T}^1} k_{\Sigma_x}(t')u(t - t')dt, \quad k_{\Sigma_x}(t) = 1 - \exp(-(c/t)^4)\chi_{(-c,c)}(t),$$

being $c = 0.2$. Finally, we selected $\sigma = 0.05$. In Figure 3, we show some signals from the training sample, both in dimension $N = 64$ and $N = 256$.

A.7.2 Implementation aspects

As expressed in Section 5, it is possible to compute the mean squared error L associated with the optimal parameter $\theta^\star = (h^\star, B^\star)$ and the learned parameters $\widehat{\theta}_S = (\widehat{h}_S, \widehat{B}_S)$, $\widehat{\theta}_U = (\widehat{h}_U, \widehat{B}_U)$ with an explicit formula. Indeed, since the employed data are synthetically generated, we can take advantage of the knowledge of $\mu, \Sigma_x, \Sigma_\varepsilon$. We simulate the computations in Section 3 and in Appendix A.3, in a finite-dimensional context: here, since for any N the operator Σ_ε is invertible, Assumption 2.8 is satisfied with $K = Y$. The expression of the regularizer in (12) then reads as

$$R_{h,B}(y) = Wy + b,$$

being $W = B^2 A^* (\Sigma_\varepsilon + AB^2 A^*)^{-1}$ and $b = (I_X - WA)h$. Moreover, as in Appendix A.3, we can compute

$$L(h, B) = \operatorname{tr}[(WA - I_X)\Sigma_x(WA - I_X)^*] + \operatorname{tr}[W\Sigma_\varepsilon W^*] + \|(WA - I_X)\mu + b\|_X^2.$$

By this formula, it is possible to compute the mean squared error associated to any parameter θ, and in particular for $(h^\star, B^\star) = (\mu, \Sigma_x^{1/2})$ and $(\widehat{h}_U, \widehat{B}_U) = (\widehat{\mu}, \widehat{\Sigma}_x^{1/2})$.

In order to detect the empirical risk minimizer $(\widehat{h}_S, \widehat{B}_S)$, and in particular to compute the quantity $L(\widehat{h}_S, \widehat{B}_S)$, we rely on the same strategy adopted for the minimization of L. Therefore, we first look for the affine functional $Wy + b$ which minimizes the empirical risk, defined as

$$\widehat{L}_{b,W} = \frac{1}{m} \sum_{j=1}^{m} \|Wy_j + b - x_j\|_X^2;$$

then, if the optimal b and W can be written as $W = B^2 A^*(\Sigma_\varepsilon + AB^2 A^*)^{-1}$ and $b = (I_X - WA)h$, the pair (h, B) is a minimizer of $\widehat{L}(h, B)$. Thanks to the empirical mean and covariance matrices

$$\widehat{y} = \frac{1}{m} \sum_{j=1}^{m} y_j, \qquad \widehat{\Sigma}_y = \frac{1}{m} \sum_{j=1}^{m} (y_j - \widehat{y}) \otimes (y_j - \widehat{y}), \qquad \widehat{\Sigma}_{yx} = \frac{1}{m} \sum_{j=1}^{m} (y_j - \widehat{y}) \otimes (x_j - \widehat{\mu}),$$

it is also possible to provide a more explicit formula for $\widehat{L}_{b,W}$. Indeed,

$$\widehat{L}_{b,W} = \frac{1}{m} \sum_{j=1}^{m} \left(\|W(y_j - \widehat{y})\|_X^2 + \|x_j - \widehat{\mu}\|_X^2 - 2\langle W(y_j - \widehat{y}), x_j - \widehat{\mu}\rangle_X + \|W\widehat{y} - \widehat{\mu} + b\|_X^2 \right)$$

$$= \operatorname{tr}[W\widehat{\Sigma}_y W^*] + \operatorname{tr}[\widehat{\Sigma}_x] - 2\operatorname{tr}[W\widehat{\Sigma}_{yx}] + \|W\widehat{y} - \widehat{\mu} + b\|_X^2,$$

where we have used that $\sum_j (y_j - \widehat{y}) = 0$ and $\sum_j (x_j - \widehat{\mu}) = 0$. Thus, the minimizer of $\widehat{L}_{b,W}$ is the affine operator associated with $W = \widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1}$ and $b = \widehat{\mu} - W\widehat{y}$. Unfortunately, such W does not yield the optimal parameter $\widehat{B}_S$: indeed, W cannot be written in the form $W = B^2 A^*(\Sigma_\varepsilon + AB^2 A^*)^{-1}$, but rather $W = MA^*(\Sigma_\varepsilon + AMA^*)^{-1}$, where the resulting M is not symmetric. We overcome such issue by considering the symmetric part of M, which we denote by M'. Indeed, despite the operator W' associated with M' is possibly different from the minimizer of $\widehat{L}_{b,W}$ among the functionals of the form $W = B^2 A^*(\Sigma_\varepsilon + AB^2 A^*)^{-1}$, numerical evidence shows that the values of L evaluated in W' and W are very close. Since the former is an upper bound of $L(\widehat{h}_S, \widehat{B}_S)$ and the latter a lower bound, we conclude that the expected loss L evaluated in W' provides a sufficiently tight upper estimate of the value of $L(\widehat{h}_S, \widehat{B}_S)$, without explicitly requiring the computation of $\widehat{B}_S$ and $\widehat{h}_S$.

As a final remark, we show how the generalization bounds in probability obtained in Theorems 4.1 and 4.2 can be reformulated in expectation. Let us first consider the unsupervised case, in which

$$\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|\leq c_3\frac{1}{\sqrt{m}}+c_4\frac{\sqrt{\tau}}{\sqrt{m}}\right]\geq 1-e^{-\tau}.$$

Inverting $\eta=c_3\frac{1}{\sqrt{m}}+c_4\frac{\sqrt{\tau}}{\sqrt{m}}$ in terms of τ we get

$$\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|\leq\eta\right]\geq 1-\widetilde{c}_1 e^{-\widetilde{c}_2 m\eta^2}.$$

This can be translated into a bound in expectation by means of the following identity:

$$\mathbb{E}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|\right]=\int_0^\infty\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|>\eta\right]d\eta\lesssim\frac{1}{\sqrt{m}}.$$

In a similar way, in the supervised case we have (see the Appendix A.5)

$$\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{\theta}_{\mathbf{z}})-L(\theta^*)|\leq\eta\right]\geq 1-e^{c_1\eta^{-1/s'}-c_2 m\eta^2}.$$

Notice that, when $\eta\to 0$, such bound could be meaningless, as the term $e^{c_1\eta^{-1/s'}}$ blows up. We therefore substitute it with the following estimate:

$$\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{\theta}_{\mathbf{z}})-L(\theta^*)|\leq\eta\right]\geq 1-\min\{1,e^{c_1\eta^{-1/s'}-c_2 m\eta^2}\}.$$

As a consequence,

$$\mathbb{E}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|\right]=\int_0^\infty\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|>\eta\right]d\eta$$

$$\leq\int_0^\infty\min\{1,e^{c_1\eta^{-1/s'}-c_2 m\eta^2}\}d\eta.$$

Notice that $1\leq e^{c_1\eta^{-1/s'}-c_2 m\eta^2}$ when $c_1\eta^{-1/s'}\geq c_2 m\eta^2$, namely when $\eta\leq\widehat{\eta}(m)=\left(\frac{c_1}{c_2 m}\right)^{\frac{1}{2+1/s'}}$. Thus,

$$\mathbb{E}_{\mathbf{z}\sim\rho^m}\left[|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|\right]\leq\widehat{\eta}(m)+\int_{\widehat{\eta}(m)}^\infty e^{c_1\eta^{-1/s'}-c_2 m\eta^2}d\eta$$

$$\leq\widehat{\eta}(m)+\int_{\widehat{\eta}}^\infty e^{c_1\widehat{\eta}^{-1/s'}-c_2 m\eta^2}d\eta=\left[\begin{array}{c}c_1\widehat{\eta}^{-1/s'}-c_2 m\eta^2=-\beta^2\\ d\eta=\frac{1}{\sqrt{c_2 m}}\frac{\beta}{\sqrt{c_1\widehat{\eta}^{-1/s'}+\beta^2}}d\beta\leq\frac{1}{\sqrt{c_2 m}}d\beta\end{array}\right]$$

$$\leq\widehat{\eta}(m)+\frac{1}{\sqrt{c_2 m}}\int_0^\infty e^{-\beta^2}d\beta\lesssim\left(\frac{1}{m}\right)^{\frac{1}{2+1/s'}}+\frac{1}{\sqrt{m}},$$

and the leading order is $\left(\frac{1}{m}\right)^{\frac{1}{2+1/s'}}$, which can be rewritten as $\left(\frac{1}{\sqrt{m}}\right)^{1-\frac{1}{2s'+1}}$, and converges to $\frac{1}{\sqrt{m}}$ for large values of s' (namely, of s).

Finally, in Table 1 we report the numerical values of the excess risk $|L(\widehat{h}_U,\widehat{B}_U)-L(h^\star,B^\star)|$ and $|L(\widehat{h}_S,\widehat{B}_S)-L(h^\star,B^\star)|$ associated with all the studied cases.

Model (a)	Sample size, m						
	3000	6463	13925	30000	64633	139248	300000
$N = 64$, unsup.	0.00174	0.00119	0.00076	0.00056	0.00038	0.00025	0.00018
$N = 64$, sup.	0.00398	0.00233	0.00156	0.00111	0.00067	0.00044	0.00031
$N = 256$, unsup.	0.00195	0.00131	0.00086	0.00063	0.00040	0.00029	0.00020
$N = 256$, sup.	0.01369	0.00485	0.00246	0.00134	0.00092	0.00052	0.00037
Model (b)							
$N = 64$, unsup.	0.00177	0.00129	0.00082	0.00053	0.00034	0.00023	0.00019
$N = 64$, sup.	0.00380	0.00236	0.00158	0.00103	0.00058	0.00044	0.00032
$N = 256$, unsup.	0.00199	0.00126	0.00094	0.00056	0.00044	0.00028	0.00019
$N = 256$, sup.	0.01449	0.00487	0.00250	0.00142	0.00086	0.00058	0.00035
Model (c)							
$N = 64$, unsup.	0.00188	0.00125	0.00083	0.00056	0.00039	0.00027	0.00018
$N = 64$, sup.	0.00407	0.00240	0.00154	0.00105	0.00069	0.00044	0.00028
$N = 256$, unsup.	0.00190	0.00134	0.00088	0.00057	0.00040	0.00028	0.00019
$N = 256$, sup.	0.01434	0.00503	0.00248	0.00135	0.00089	0.00052	0.00036

Table 1: Tabulated values of the excess risks associated with Figure 2, computed at two discretization levels and in three different statistical setups: Gaussian variable x and (a) uniform white noise ε, (b) Gaussian white noise ε, and (c) white noise ε uniformly distributed w.r.t. the Haar wavelet transform.

A.8 An ill-posed inverse problem: deconvolution of 1D signals

We provide a numerical verification of the estimates of Theorems 4.1 and 4.2 for a 1D deconvolution problem, extending the experiments of section 5 to the case of an ill-posed operator A. We consider again $X = Y = L^2(\mathbb{T}^1)$, and introduce the convolution operator $(Ax)(t) = (k * x)(t) = \int_{\mathbb{T}} k(t - \tau)x(\tau)d\tau$. This operation can be used to describe the blurring of one-dimensional signals, the function k being the convolutional filter, or point spread function. In our experiments, we consider $k(t) = \chi_{[-L,L]}(t)$, the indicator function of the interval $[-L, L]$, and set $L = 0.02$. Such a kernel k can be referred to as the average filter. When discretizing the interval $\mathbb{T}^1$ with N 1D-pixels, the operator A reduces to a discrete (periodic) convolution with a constant vector $\mathbf{k}$, whose number of entries is LN. Both at a continuous and at a discrete level, the deconvolution problem is known to be ill-posed (see, e.g., [38]), and the smallest singular value of the discretized operator vanishes as N grows. Nevertheless, we expect to observe the same generalization bounds as in the denoising case.

We replicate the same experiments as in section 5, assuming that x is a random Gaussian variable (with mean μ and covariance Σ_x as reported in section 5) and ε is white uniform noise with covariance $\Sigma_\varepsilon = \sigma^2 I$. We fix a noise level of 2.5% by setting σ equal to the 2.5% of the peak value of the average signal. The results of the numerical experiments are reported in Figure 4. We observe that in both scenarios the decay of the excess risk is of the order $1/\sqrt{m}$, and the unsupervised technique still provides (slightly) better results, which in particular are not affected by the increased ill-posedness of the operator at a refined scale.
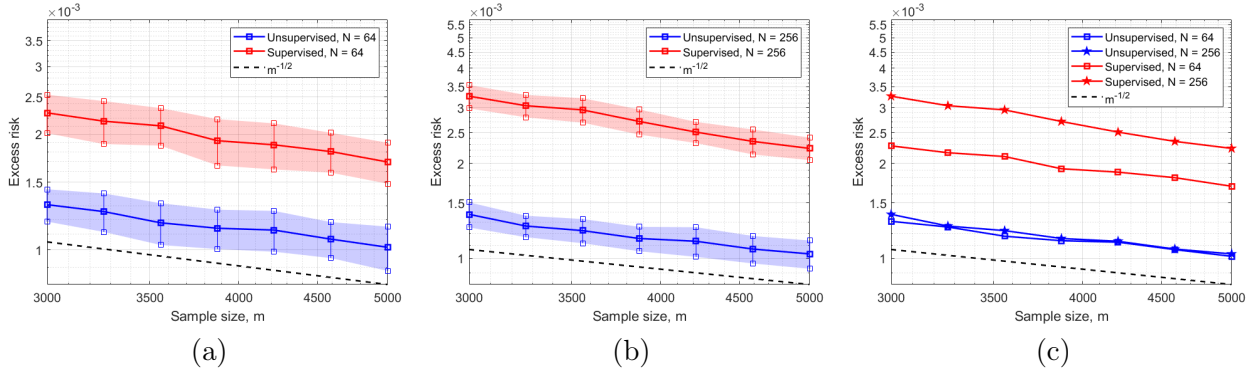
Figure 4: Decay of the excess risks $|L(\widehat{\theta}_S) - L(\theta^\star)|$ and $|L(\widehat{\theta}_U) - L(\theta^\star)|$ (with standard deviation error bars) with two different discretization sizes, $N = 64$ (a) and $N = 256$ (b), and comparison (c).