ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Learning Sparsity-Promoting Regularizers for Linear Inverse Problems

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Alberti, G.S., De Vito, E., Helin, T., Lassas, M., Ratti, L., Santacesaria, M. (2026). Learning Sparsity-Promoting Regularizers for Linear Inverse Problems. SIAM JOURNAL ON MATHEMATICS OF DATA SCIENCE, 8(1), 167-199 [10.1137/24M1719785].

Availability:
This version is available at: https://hdl.handle.net/11585/1064291 since: 2026-05-20

Published:
DOI: http://doi.org/10.1137/24M1719785

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

10 September 2026

arXiv:2412.16031v2 [stat.ML] 2 Mar 2026

Learning sparsity-promoting regularizers for linear inverse problems

Giovanni S. Alberti* Ernesto De Vito* Tapio Helin† Matti Lassas‡
Luca Ratti§ Matteo Santacesaria*

Abstract

This paper introduces a novel approach to learning sparsity-promoting regularizers for solving linear inverse problems. We develop a bilevel optimization framework to select an optimal synthesis operator, denoted as B, which regularizes the inverse problem while promoting sparsity in the solution. The method leverages statistical properties of the underlying data and incorporates prior knowledge through the choice of B. We establish the well-posedness of the optimization problem, provide theoretical guarantees for the learning process, and present sample complexity bounds. The approach is demonstrated through theoretical infinite-dimensional examples, including compact perturbations of a known operator and the problem of learning the mother wavelet, and through extensive numerical simulations. This work extends previous efforts in Tikhonov regularization by addressing non-differentiable norms and proposing a data-driven approach for sparse regularization in infinite dimensions.

Keywords: inverse problems, statistical learning, bilevel optimization, operator learning, sparsity-promoting regularization.

1 Problem formulation and main contributions

Consider a linear inverse problem

$$y = Ax + \varepsilon, \tag{1}$$

where we assume that $A\colon X \to Y$ is a linear bounded operator between the real and separable Hilbert spaces X, Y, whereas the inverse A^{-1} (if it exists) can be in general an unbounded operator.

We introduce a variational strategy [24] to regularize the inverse problem (1), and together promote the sparsity of the solution with respect to a suitable basis or frame. To do so, we consider an operator $B \in \mathcal{L}(\ell^2, X)$ (where $\mathcal{L}(\ell^2, X)$ denotes the space of bounded linear operators from the sequence space $\ell^2(\mathbb{N})$ to the Hilbert space X, equipped with the operator norm) and define the following minimization problem:

$$\hat{x}_B = B\hat{u}_B, \quad \hat{u}_B = \operatorname*{arg\,min}_{u \in \ell^2} \left\{ \frac{1}{2} \|\Sigma_\varepsilon^{-1/2} ABu\|_Y^2 - \langle y, \Sigma_\varepsilon^{-1} ABu \rangle_Y + \|u\|_{\ell^1} \right\}, \tag{2}$$

where Σ_ε is the covariance of the noise ε (see Assumption 1.1 below). In Section 2, we introduce a set of assumptions that guarantee the well-definedness and well-posedness of problem (2).

To better interpret the proposed regularization strategy, we remark that in a finite-dimensional setup (let, e.g., $X = \mathbb{R}^n$ and $Y \in \mathbb{R}^m$), if the matrix $B \in \mathbb{R}^{n \times n}$ is invertible, the minimization problem (2) admits the following formulation:

$$\hat{x}_B = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Ax - y\|_{\Sigma_\varepsilon}^2 + \|B^{-1}x\|_1 \right\},$$

*MaLGa Center, University of Genoa, Italy ({giovanni.alberti, ernesto.devito, matteo.santacesaria}@unige.it).
†Computational Engineering, Lappeenranta-Lahti University of Technology, Finland (tapio.helin@lut.fi).
‡Department of Mathematics and Statistics, University of Helsinki, Finland (matti.lassas@helsinki.fi).
§Department of Mathematics, University of Bologna, Italy (luca.ratti5@unibo.it).

where $\|\cdot\|_{\Sigma_\varepsilon} = \|\Sigma_\varepsilon^{-1/2}\cdot\|$ is a norm on $\mathbb{R}^m$ that leverages the knowledge of the noise covariance Σ_ε to whiten the residual error. It is worth observing that this argument, as well as all the results of this work, can be extended to complex Hilbert spaces X and Y, provided that the scalar product in (2) is replaced by its real part; to simplify the exposition we have decided to consider the real case only. Further details about this simplified formulation can be found in Section 5. The main focus of this paper is related to the choice of the synthesis operator $B\colon \ell^2 \to X$ within (2). In particular, let us introduce the map

$$R_B\colon Y \to X, \qquad R_B(y) = \hat{x}_B \quad \text{as in (2).} \tag{3}$$

For a fixed B satisfying the theoretical requirements expressed in Section 2, the map R_B is a *stable* reconstruction operator, i.e., it provides a continuous approximation of the inverse map A^{-1}. In particular, R_B promotes the sparsity of the regularized solution in terms of the synthesis operator B: namely, $R_B(y) = B\hat{u}_B$ is chosen so that it balances a good data fidelity ($AR_B(y) \approx y$) and only a few components of $\hat{u}_B$ are different from 0. We observe that the regularization parameter is not explicit in the above formulation, since it is absorbed by B.

The choice of the synthesis operator B encodes crucial information regarding the prior distribution of x. Consider, for simplicity, a signal processing problem (such as denoising, or deblurring), which, after a discretization of the spaces of signals, can be formulated as a linear inverse problem in $X = Y = \mathbb{R}^n$. Then, promoting the sparsity of the reconstructed signals under the choice $B = \mathrm{Id}$ encodes the prior information that the ground truths are expected to have few, isolated, spikes. Choosing B to be a basis of (discretized) sines and cosines promotes band-limited signals in the Fourier domain. Setting B to a wavelet transform, instead, might promote signals showing few jump discontinuities and smooth anywhere else. In general, the synthesis operator B should be carefully chosen to incorporate, in the reconstruction operator R_B, any significant prior knowledge, or prior belief, on the ground truths.

In this paper, we follow a statistical learning approach for the selection of B. In particular, we assume that x and y are random objects with a joint probability distribution ρ. We then leverage the (partial) knowledge of such a probability distribution to define a data-driven rule for selecting B. To do so, we must introduce some assumptions on the probability distribution ρ, which are carefully detailed in Section 3. Before delving into the details, though, let us briefly sketch the overall idea of the learning-based choice of B, introducing some minimal requirements on the random objects x, y.

Assumption 1.1. *Let $(x, y) \sim \rho$, being $y = Ax + \varepsilon$, and where:*

- *x is a square-integrable random variable on X;*
- *ε is a zero-mean square-integrable random variable on Y, independent of x, with known and injective covariance $\Sigma_\varepsilon\colon Y \to Y$.*

We recall that assuming ε to be a zero-mean square-integrable random variable (i.e., $\mathbb{E}[\|\varepsilon\|_Y^2] < +\infty$) implies that its covariance operator Σ_ε is positive and trace-class (or nuclear, see [9, Theorem 1.7]), namely, that the sequence of eigenvalues is summable. For this reason, such an assumption is in principle quite restrictive when Y is infinite-dimensional and, for example, the case of white noise requires careful treatment. In its most common formulation, white noise is modeled as a random process on a Hilbert space (for example, $Y = L^2(\Omega)$) with zero mean and covariance equal to the identity. Such a noise model clearly fails in satisfying the hypotheses whenever Y is infinite dimensional, because the covariance $\Sigma_\varepsilon = \mathrm{Id}$ is not trace-class, but only bounded. However, it is possible to represent such a process as a square-integrable random variable by carefully selecting the space Y, as described in Appendix A.

For a specific choice of synthesis operator B, the quality of the reconstruction operator R_B can be evaluated through the *expected loss* $L\colon \mathcal{L}(\ell^2, X) \to \mathbb{R}$:

$$L(B) = \mathbb{E}_{(x,y)\sim\rho}[\|R_B(y) - x\|_X^2]. \tag{4}$$

Let us now consider a suitable class $\mathcal{B} \subset \mathcal{L}(\ell^2, X)$ of *admissible* operators, namely, that satisfy the assumptions reported in Section 2 guaranteeing the existence and uniqueness of the solution of (2).

We define the optimal regularizer $R^\star = R_{B^\star}$, where $B^\star$ is a minimizer of the expected loss over the set of admissible operators $\mathcal{B}$, whose existence follows under suitable conditions on $\mathcal{B}$. In particular, putting together the expressions of (4) and (2), we obtain:

$$\begin{aligned} B^\star &\in \operatorname*{arg\,min}_{B\in\mathcal{B}} L(B), \\ R_B(y) &= B \operatorname*{arg\,min}_{u\in\ell^2} \Big\{ \frac{1}{2}\|\Sigma_\varepsilon^{-1/2}ABu\|_Y^2 - \langle y, \Sigma_\varepsilon^{-1}ABu\rangle_Y + \|u\|_{\ell^1} \Big\}, \end{aligned} \tag{5}$$

which is usually referred to as a *bilevel* optimization problem. Note that the optimality of $B^\star$ strongly depends on the choice of the class $\mathcal{B}$. We moreover stress that the optimal target $B^\star$ can only be computed if the joint probability distribution ρ of x and y is known. This is not verified in practical applications, but in many contexts, we may suppose to have access to a sample of m pairs $\mathbf{z} = \{(x_j, y_j)\}_{j=1}^m$ such that the family is independent and identically distributed as (x, y). In this case, following the paradigm of *supervised learning*, we can approximate the expected loss by an empirical average, also known as *empirical risk*, namely

$$\widehat{L}(B) = \frac{1}{m}\sum_{j=1}^m \|R_B(y_j) - x_j\|_X^2. \tag{6}$$

A natural estimator of $B^\star$ is then given by any minimizer $\widehat{B}$ of the empirical loss:

$$\widehat{B} \in \operatorname*{arg\,min}_{B\in\mathcal{B}} \widehat{L}(B). \tag{7}$$

We stress that both $\widehat{L}(B)$ and $\widehat{B}$ are random variables depending on the sample $\mathbf{z}$. We simply denote this dependence by $\widehat{\cdot}$. The task of *sample error estimates* is to quantify the dependence of the empirical target $\widehat{B}$ on the sample $\mathbf{z}$, in particular by bounding (either in probability or in expectation) the *excess risk* $L(\widehat{B}) - L(B^\star)$ in terms of the sample size m.

Motivation

The theoretical framework developed in this paper allows significant flexibility in choosing the class $\mathcal{B}$ of admissible operators, enabling various practical applications. In this paper we provide two examples, detailed in Section 4, of our theory in an infinite-dimensional setting. In particular, we establish concrete guarantees on the sample complexity of the learning problem by deriving explicit bounds on the covering numbers of the respective operator classes. In the first example, we assume that some prior knowledge about an effective synthesis operator B_0 is available (e.g., from physical understanding of the problem). In this setting, we construct a class $\mathcal{B}$ of compact perturbations around this reference operator. This approach maintains the desirable properties of B_0 while allowing for data-driven refinements. The second example is in the framework of signal processing applications where wavelets provide powerful sparse representations for many natural signals exhibiting multi-scale features. Our framework enables learning an optimal mother wavelet from data, rather than selecting from predefined wavelet families.

The examples are complemented by the numerical studies of Section 5, demonstrating how the abstract mathematical formulation directly connects to practical applications in finite-dimensional spaces.

Main contribution of this paper

In our analysis, we pursue the following main goals:

1. studying the theoretical properties of the inner minimization problem (2), and in particular its well-posedness for fixed B and the continuous dependence of the minimizer $\hat{x}_B$ in terms of $B \in \mathcal{B}$ (see Theorems 2.3 and 2.5 in Section 2);

2. studying the approximation properties of the empirical target $\widehat{B}$ (constructed leveraging the knowledge of A, Σ_ε, and of the sample $\mathbf{z}$) with respect to the optimal target $B^\star$, deriving sample error estimates for the excess risk (see Theorem 3.2 in Section 3);

3. formulate some relevant applications which satisfy the assumptions on x, y, A, and $\mathcal{B}$ introduced in Sections 2 and 3 (see Section 4);

4. discuss numerical strategies to provide an approximation of the empirical target $\widehat{B}$ and showcase their effectiveness on some benchmark applications (see Section 5).

Comparison with existing literature

This work originated from the analysis carried out by a subset of the authors in [5], where we considered the problem of learning (also) the optimal operator B in generalized Tikhonov regularization, i.e., when a quadratic penalty term is considered within the inner problem (2) instead of the ℓ^1 norm. The main difficulty associated with the proposed extensions resides in the lack of strong convexity and differentiability of the ℓ^1 norm. Moreover, unlike in the Tikhonov case, the inner minimization problem (2) does not possess a closed-form solution: unfortunately, this also prevents the explicit computation of the optimal regularizer $B^\star$, which was one of the main results in [5]. As a final consequence, it is not possible to formulate here an *unsupervised* strategy to learn $B^\star$, i.e., based only on a training set of ground truths $\{x_j\}_{j=1}^m$: indeed, the unsupervised strategy proposed in [5] extensively leveraged the explicit expression of the optimal operator $B^\star$.

Other extensions of the work [5] have been carried out in [44, 15, 19, 11, 6] even though a statistical learning approach for sparse optimization in infinite dimension was not considered yet. See also [30] for a connection between generalized Tikhonov and linear plug-and-play denoisers. Let us just mention that bilevel approaches for inverse problems in imaging have been studied, from numerical and optimization points of view since many years [16, 23]. In particular, the works [31, 32, 27, 26] study the bilevel learning of ℓ^1 regularizers, focusing on the finite-dimensional case and mostly on the algorithmic aspects, without considering the generalization issue from the theoretical point of view. Moreover, our work is also related to statistical inverse learning for inverse problems [8, 1, 14] and, also, more generally to the growing field of operator learning [42, 36, 35, 22, 10]. Let us also mention some of the main works on using machine learning techniques for solving inverse problems that had motivated this work [3, 34, 4, 38, 7, 37, 40, 41].

It is worth observing that the problem of learning an operator B yielding a sparse representation of a dataset $\{x_j\}_{j=1}^m$ is deeply connected with the well-known task of Dictionary Learning. Although it is possible to provide a formulation of a dictionary learning problem very close to the bilevel problem (5), the two problems pursue two distinct aims and may lead to different results, as shown in Appendix B. Despite this, the sample complexity bounds we obtain in this work are comparable to those derived for dictionary learning [29, 47].

2 Theoretical results on the deterministic optimization problem (2)

The goal of this section is to study the well-posedness of the minimization problem formulated in (2) for a fixed $B \in \mathcal{L}(\ell^2, X)$. Moreover, we study the stability of the minimization with respect to perturbations of B. To do so, we need to introduce a suitable set of hypotheses.

For the rest of the paper, we use the convention $J_B(u) = F_B(u) + \Phi(u)$, where

$$F_B(u) = \frac{1}{2}\|\Sigma_\varepsilon^{-1/2} ABu\|_Y^2 - \langle y, \Sigma_\varepsilon^{-1} ABu\rangle_Y \quad \text{and} \quad \Phi(u) = \|u\|_{\ell^1}. \tag{8}$$

Assumption 2.1 (Compatibility assumption between Σ_ε and A)**.** *Let the covariance matrix Σ_ε satisfy Assumption 1.1, and assume $\mathrm{Im}(A) \subset \mathrm{Im}(\Sigma_\varepsilon)$. Moreover, we assume that*

$$\Sigma_\varepsilon^{-1} A \colon X \to Y \text{ is a compact operator.} \tag{9}$$

Assumption 2.1 is a technical requirement that is needed to guarantee that the functional J_B is well-defined for a fixed $B \in \mathcal{L}(\ell^2, X)$. Indeed, the first term in (8) can be rewritten as $\frac{1}{2}\langle \Sigma_\varepsilon^{-1} ABu, ABu\rangle_Y$, and $\Sigma_\varepsilon^{-1} ABu$ exists and is unique since $\operatorname{Im}(AB) \subset \operatorname{Im} \Sigma_\varepsilon = \operatorname{dom}(\Sigma_\varepsilon^{-1})$, and analogously for the second term. Finally, since the minimization problem (2) is set in the Hilbert space $\ell^2 \supset \ell^1$, the third term in (2) should be interpreted as $\|u\|_{\ell^1}$ if $u \in \ell^1$ and $+\infty$ if $u \in \ell^2 \setminus \ell^1$. We moreover remark that Assumption 2.1 is always satisfied in finite dimension when the covariance Σ_ε is invertible.

Since J_B is non-differentiable and not strictly convex, further assumptions on the interplay between A and B are needed to guarantee the well-posedness of the minimization task. Our analysis is confined to the case where AB is assumed to be finite basis injective.

Assumption 2.2 (Finite Basis Injectivity (FBI) of AB)**.** *For all $I \subset \mathbb{N}$ with $\operatorname{card}(I) < \infty$, the operator $(AB)|_{\ell^2_I}$ is injective, where we denoted $\mathbb{N} = \{1, 2, \dots\}$ and $\ell^2_I := \operatorname{span}\{e_i : i \in I\}$.*

Note that while this assumption is satisfied when B is injective (provided A is injective), it also holds for more general structures (e.g. FBI frames, see [12]). We remark that, in order to guarantee the well-posedness of (2), there exist some alternatives to Assumption 2.2, such as the Restricted Isometry Property (see [17]), the source conditions (see [28]), or the strict sparsity pattern (see [12]). Compared to these alternatives, Assumption 2.2 allows us to consider more general operators A, to provide a clearer interpretation, and to avoid introducing requirements on the unknown ground truth, respectively.

The well-posedness of the minimization problem (2) is now characterized by the following theorem.

Theorem 2.3. *Let $A \in \mathcal{L}(X, Y), \Sigma_\varepsilon \in \mathcal{L}(Y, Y)$, and $B \in \mathcal{L}(\ell^2, X)$ satisfy Assumptions 2.1 and 2.2, and let $y \in Y$. There exists a unique minimizer $\hat{u}_B = \hat{u}_B(y)$ of J_B, and, consequently, a unique $\hat{x}_B = B\hat{u}_B$ in (2).*

The proof is presented in two parts: the existence is proved at the end of Section 2.1 while the uniqueness follows as a consequence of Theorem 2.5 at the end of Section 2.2.

Let us next consider the stability of the minimizer with respect to perturbing B in a suitable class of operators $\mathcal{B}$. Towards that end, let us introduce the following assumption, which is also strongly motivated by the statistical learning results presented in Section 3, and has also been considered in [5].

Assumption 2.4 (Requirements on $\mathcal{B}$)**.** *Let $\mathcal{B} \subset \mathcal{L}(\ell^2, X)$ be a compact set of operators such that every $B \in \mathcal{B}$ satisfies Assumption 2.2.*

A first consequence of the compactness of $\mathcal{B}$ is that the norms $\|B\|$, $B \in \mathcal{B}$, are uniformly bounded by $|\mathcal{B}| := \sup_{B\in\mathcal{B}} \|B\| < \infty$. Moreover, as we show in the following subsections, Assumption 2.4 allows us to provide a uniform expression of several properties of the solutions of problem (2), namely, independently of the choice of B. The most relevant consequence of this discussion is the following result, providing a global stability estimate of the minimizers $\hat{x}_B$ with respect to perturbations of B within $\mathcal{B}$.

Theorem 2.5. *Let A, Σ_ε and $\mathcal{B}$ satisfy Assumptions 1.1, 2.1 and 2.4, and let $\|y\|_Y \le Q$. Then, there exists a constant $c_{\mathrm{ST}} = c_{\mathrm{ST}}(\mathcal{B}, A, \Sigma_\varepsilon, Q)$ such that for every $B_1, B_2 \in \mathcal{B}$,*

$$\|R_{B_1}(y) - R_{B_2}(y)\|_X \le c_{\mathrm{ST}} \|B_1 - B_2\|, \tag{10}$$

2.1 Existence of the minimizer with fixed $B \in \mathcal{B}$

To establish the existence of a minimizer $\hat{u}_B$ in (2), we first show that the sublevel sets of J_B are bounded, which ensures the weak convergence of any minimizing sequence.

Lemma 2.6. *Let A, Σ_ε and $\mathcal{B}$ satisfy Assumptions 2.1 and 2.4. There exists a monotonically increasing function $h\colon \mathbb{R}_+ \to \mathbb{R}_+$ depending on A, Σ_ε, and $|\mathcal{B}|$ such that the following holds: if $\|w\|_X \le |\mathcal{B}|$ and $\left\|\Sigma_\varepsilon^{-1} Aw\right\|_Y \ge \gamma > 0$, then $\left\|\Sigma_\varepsilon^{-1/2} Aw\right\|_Y \ge h(\gamma) > 0$.*

Proof. Let us prove existence of a general h by contradiction: assume that for all n there exists $w_n \in X$ satisfying $\|w_n\|_X \leq |\mathcal{B}|$, $\left\|\Sigma_\varepsilon^{-1} A w_n\right\|_Y \geq \gamma$ and $\left\|\Sigma_\varepsilon^{-1/2} A w_n\right\|_Y < \frac{1}{n}$. Since $\{w_n\}_{n=1}^\infty \subset X$ is bounded, we can find a subsequence $\{w_{n_k}\}_{k=1}^\infty$ that weakly converges to some $w \in X$. Due to the weak lower semicontinuity of the norm (see, e.g., [13, Section 1.4 and Corollary 3.9]), we have that $\|\Sigma_\varepsilon^{-1/2} A w\|_Y \leq \liminf_k \|\Sigma_\varepsilon^{-1/2} A w_{n_k}\|_Y = 0$, and also, by the injectivity of $\Sigma_\varepsilon^{-1/2}$, that $w \in \ker(A)$. Therefore, by the compactness of $\Sigma_\varepsilon^{-1} A$ we have that $\|\Sigma_\varepsilon^{-1} A w_{n_k}\|_Y \to 0$. This contradicts our assumption and, therefore, a lower bound $h(\gamma) > 0$ must exist.

The existence of a monotonically increasing h can also be demonstrated by contradiction: suppose that there exist $\gamma_1 < \gamma_2$ such that for any h satisfying the statement, we have $h(\gamma_1) > h(\gamma_2)$. Then we immediately see that $\tilde{h}$ defined by

$$\tilde{h}(\gamma) = \begin{cases} h(\gamma) & \gamma \neq \gamma_1 \\ h(\gamma_2) & \gamma = \gamma_1, \end{cases}$$

also satisfies the claim yielding the contradiction. □

Proposition 2.7. *Let $M \in \mathbb{R}$, and suppose Assumptions 1.1, 2.1, 2.2 and 2.4 hold. Let $\|y\|_Y \leq Q$. Then, there exists a constant $c_{\mathrm{UB}}(M) = c_{\mathrm{UB}}(M; A, \Sigma_\varepsilon, Q, |\mathcal{B}|)$ such that, for every $B \in \mathcal{B}$, we have*

$$\{u \in \ell^2 \mid J_B(u) \leq M\} \subset \{u \in \ell^2 \mid \|u\|_{\ell^1} \leq c_{\mathrm{UB}}(M)\}. \tag{11}$$

Proof. The condition $J_B(u) \leq M$ reads as

$$\frac{1}{2}\|\Sigma_\varepsilon^{-1/2} A B u\|_Y^2 + \|u\|_{\ell^1} \leq M + \langle y, \Sigma_\varepsilon^{-1} A B u\rangle_Y.$$

Let us next make a change of variables with $w = Bu/\|u\|_{\ell^1}$ and $\tau = \|u\|_{\ell^1}$, which leads to

$$\frac{1}{2}\|\Sigma_\varepsilon^{-1/2} A w\|_Y^2 \tau^2 + \tau \leq M + \tau \|y\|_Y \left\|\Sigma_\varepsilon^{-1} A w\right\|_Y. \tag{12}$$

In particular, we have

$$1 \leq \frac{M}{\tau} + \|y\|_Y \left\|\Sigma_\varepsilon^{-1} A w\right\|_Y.$$

Hence, either $\|u\|_{\ell^1} = \tau \leq \max\{2M, 0\}$ or for any $\tau \geq \max\{2M, 0\}$ it holds that

$$\left\|\Sigma_\varepsilon^{-1} A w\right\|_Y \geq \frac{1}{2\|y\|_Y}.$$

Since $\|Bu\|_X \leq |\mathcal{B}| \|u\|_{\ell^2}$ for all $u \in \ell^2$ and $B \in \mathcal{B}$, we have

$$\|u\|_{\ell^1} \geq \|u\|_{\ell^2} \geq \frac{1}{|\mathcal{B}|}\|Bu\|_X \qquad \forall u \in \ell^2, \forall B \in \mathcal{B},$$

which also implies that $\|w\|_X \leq |\mathcal{B}|$. Next, by Lemma 2.6 it follows that

$$\left\|\Sigma_\varepsilon^{-1/2} A w\right\|_Y \geq h,$$

where we abbreviate $h = h\left(\frac{1}{2\|y\|_Y}\right)$ for convenience. Modifying (12) we obtain

$$\frac{1}{2}h^2\tau^2 \leq M + \tau\|y\|_Y \left\|\Sigma_\varepsilon^{-1} A\right\| |\mathcal{B}| \leq M + \frac{1}{4}h^2\tau^2 + \frac{\|y\|_Y^2 \left\|\Sigma_\varepsilon^{-1} A\right\|^2 |\mathcal{B}|^2}{h^2}$$

and solving for τ yields

$$\tau^2 \leq \frac{4M}{h^2} + \frac{4\|y\|_Y^2 \left\|\Sigma_\varepsilon^{-1} A\right\|^2 |\mathcal{B}|^2}{h^4}. \tag{13}$$

Note that, by the monotonicity of h, from $\|y\|_Y \le Q$ it follows that $h \ge h\left(\frac{1}{2Q}\right)$, hence

$$\tau^2 \le \frac{4M}{h\left(\frac{1}{2Q}\right)^2} + \frac{4Q^2 \left\|\Sigma_\varepsilon^{-1} A\right\|^2 |\mathcal{B}|^2}{h\left(\frac{1}{2Q}\right)^4}. \tag{14}$$

Now the desired bound for $\tau = \|u\|_{\ell^1}$ follows for any $u \in \ell^2$, $B \in \mathcal{B}$. □

Proof of Theorem 2.3 (existence). The existence of a minimizer $\hat{u}_B$ in (2) is now guaranteed by standard arguments: any minimizing sequence $\{u_j\}_{j=1}^\infty \subset \ell^2$ belongs to some sublevel set and, therefore, to some ℓ^1-ball according to Proposition 2.7. By the Banach–Alaoglu theorem, the sequence has a weak-* converging subsequence in ℓ^1, namely, there exists $\{u_{j_l}\}_{l=1}^\infty \subset \ell^1$ such that $u_{j_l} \overset{*}{\rightharpoonup} \bar{u}$. Since F_B is continuous and the ℓ^1-norm is lower semicontinuous with respect to the weak-* topology (i.e., $u_{j_l} \overset{*}{\rightharpoonup} \bar{u}$ implies $\|\bar{u}\|_{\ell^1} \le \liminf_{l\to\infty} \|u_{j_l}\|_{\ell^1}$), one can show that the limit is a minimizer of J_B. □

Let us note that since $J_B(\hat{u}_B) \le J_B(0) = 0$, denoting by c_{UB} the constant $c_{\mathrm{UB}}(0)$ in (11) we have the following bound for such minimizers, uniformly in $B \in \mathcal{B}$.

$$\|\hat{u}_B\|_{\ell^2} \le \|\hat{u}_B\|_{\ell^1} \le c_{\mathrm{UB}}. \tag{15}$$

The uniqueness of such a minimizer is a consequence of Proposition 2.10, as we will show later.

2.2 Stability in $\mathcal{B}$ and uniqueness of the minimizer

The key result of this section is Proposition 2.10, as it directly enables the uniqueness of the minimizer in Theorem 2.3, and its stability with respect to B. Such a result is in the spirit of [12, Theorem 2], and its proof is partially based on the one of [12, Lemma 3]. The main contribution of Proposition 2.10 is that the constant appearing in (19) does not depend on the choice of B, provided that we consider an operator class $\mathcal{B}$ satisfying Assumption 2.4. The proof of Proposition 2.10 requires some preliminary lemmas.

Let us first note that $F_B \colon \ell^2 \to \mathbb{R}$ is convex and differentiable with gradient

$$F_B'(u) = B^* \left(\Sigma_\varepsilon^{-1} A\right)^* (ABu - y),$$

where we have used that $A^*(\Sigma_\varepsilon^{-1} A) = \left(\Sigma_\varepsilon^{-1} A\right)^* A$, since $\langle \Sigma_\varepsilon^{-1} Av, Aw\rangle_Y = \langle Av, \Sigma_\varepsilon^{-1} Aw\rangle_Y = \langle (\Sigma_\varepsilon^{-1} A)^* Av, w\rangle$, for all $v, w \in X$ as Σ_ε^{-1} is self-adjoint and $\mathrm{Im}(A) \subset \mathrm{dom}(\Sigma_\varepsilon^{-1})$. Notice that F_B' is affine in u and Lipschitz continuous, and the Lipschitz constant is uniformly bounded by $L_{F'} = |\mathcal{B}|^2 \|\Sigma_\varepsilon^{-1/2} A\|^2$ for $B \in \mathcal{B}$. Since both F_B and Φ are convex, $\hat{u}_B$ being a minimizer of J_B is equivalent to 0 belonging to the subgradient set of J_B at $\hat{u}_B$, i.e. $0 \in \partial J_B(\hat{u}_B)$. Now, due to the differentiability of F_B, an equivalent optimality condition is given by $\hat{w}_B \in \partial\Phi(\hat{u}_B)$, where

$$\hat{w}_B = -F_B'(\hat{u}_B) = B^* \left(\Sigma_\varepsilon^{-1} A\right)^* (y - AB\hat{u}_B). \tag{16}$$

Moreover, thanks to the explicit expression of the subdifferential of the ℓ^1-norm, we can specify that $\hat{w}_B \in \partial\Phi(\hat{u}_B)$ is equivalent to

$$\hat{w}_{B,k} = \begin{cases} \mathrm{sign}(\hat{u}_{B,k}) & \text{if } \hat{u}_{B,k} \neq 0 \\ \xi \in [-1, 1] & \text{if } \hat{u}_{B,k} = 0. \end{cases} \tag{17}$$

The following lemmas explore properties that are valid uniformly in the compact operator class $\mathcal{B}$. For $N > 0$, we introduce the orthogonal projection operator onto ℓ^2 by setting

$$P_N \colon \ell^2 \to \ell^2 : \ P_N u = (u_1, \ldots, u_N, 0, 0, \ldots), \qquad P_N^\perp = \mathrm{Id} - P_N.$$

Lemma 2.8. *Suppose that Assumptions 2.1, 2.2, and 2.4 are satisfied. Let* $\mathbb{B}_Y(Q) = \{y \in Y : \|y\|_Y \leq Q\}$. *Then the elements* $\hat{w}_B \in \ell^2$ *corresponding to the minimizers* $\hat{u}_B \in \ell^2$ *of* J_B *via* (16) *satisfy*

$$\lim_{N\to\infty} \sup\left\{\|P_N^{\perp}\hat{w}_B\|_{\ell^2} : \ B \in \mathcal{B}, \ y \in \mathbb{B}_Y(Q)\right\} = 0.$$

Proof. By contradiction, suppose there exist $\eta > 0$ and a diverging sequence N_M, $M = 1, 2, ...,$ such that

$$\sup\left\{\left\|P_{N_M}^{\perp} B^*\left(\Sigma_\varepsilon^{-1}A\right)^*\left(y - AB\hat{u}_B\right)\right\|_{\ell^2} : B \in \mathcal{B}, \ y \in \mathbb{B}_Y(Q)\right\} \geq 2\eta \quad \forall M.$$

Consider in particular a sequence $\{B_M\}_{M=1}^{\infty} \subset \mathcal{B}$ and a sequence $\{y_M\}_{M=1}^{\infty} \subset \mathbb{B}_Y(Q)$ satisfying

$$\left\|P_{N_M}^{\perp} B_M^*\left(\Sigma_\varepsilon^{-1}A\right)^*\left(y_M - AB_M\hat{u}_{B_M}\right)\right\|_{\ell^2} \geq \eta \quad \forall M,$$

being $\hat{u}_{B_M}$ the solution of (2) with $B = B_M$ and $y = y_M$. By the compactness of $\mathcal{B}$, there exists $B \in \mathcal{B}$ such that $B_M \to B$ in the strong operator topology, up to a subsequence. Moreover, consider the sequence $h_M = AB_M\hat{u}_{B_M}$, $M = 1, 2, ...$ Since $\|\hat{u}_{B_M}\|$ is bounded uniformly with respect to $B \in \mathcal{B}$ and $y \in \mathbb{B}_Y(Q)$ by (15), $\|B_M\| \leq |\mathcal{B}|$ independently of M, and A is bounded, the sequence $\{h_M\}_{M=1}^{\infty} \subset Y$ is bounded, thus weakly convergent (up to a subsequence) to an element $h \in Y$. As a consequence of the compactness of $\left(\Sigma_\varepsilon^{-1}A\right)^*$ by the Schauder theorem, we have that $(\Sigma_\varepsilon^{-1}A)^*h_M \to (\Sigma_\varepsilon^{-1}A)^*h$ in X. Similarly, the bounded sequence $\{y_M\}_{M=1}^{\infty}$ admits a weak limit $y \in Y$, and $(\Sigma_\varepsilon^{-1}A)^*y_M \to (\Sigma_\varepsilon^{-1}A)^*y$.

In conclusion, we have that

$$\begin{aligned}
\eta &\leq \|P_{N_M}^{\perp} B_M^*(\Sigma_\varepsilon^{-1}A)^*(y_M - h_M)\|_{\ell^2} \\
&\leq \|P_{N_M}^{\perp} B_M^*(\Sigma_\varepsilon^{-1}A)^*(y_M - y)\|_{\ell^2} + \|P_{N_M}^{\perp} B_M^*(\Sigma_\varepsilon^{-1}A)^*(h_M - h)\|_{\ell^2} \\
&\quad + \|P_{N_M}^{\perp} B_M^*(\Sigma_\varepsilon^{-1}A)^*(y - h)\|_{\ell^2} \\
&\leq |\mathcal{B}|\|(\Sigma_\varepsilon^{-1}A)^*(y_M - y)\|_X + |\mathcal{B}|\|(\Sigma_\varepsilon^{-1}A)^*(h_M - h)\|_X \\
&\quad + \|P_{N_M}^{\perp}(B_M - B)^*(\Sigma_\varepsilon^{-1}A)^*(y - h)\|_{\ell^2} + \|P_{N_M}^{\perp} B^*(\Sigma_\varepsilon^{-1}A)^*(y - h)\|_{\ell^2} \\
&\leq |\mathcal{B}|\|(\Sigma_\varepsilon^{-1}A)^*(y_M - y)\|_X + |\mathcal{B}|\|(\Sigma_\varepsilon^{-1}A)^*(h_M - h)\|_X \\
&\quad + \|B_M - B\|_{\ell^2\to X}\|(\Sigma_\varepsilon^{-1}A)^*(y - h)\|_X + \|P_{N_M}^{\perp} B^*(\Sigma_\varepsilon^{-1}A)^*(y - h)\|_{\ell^2}.
\end{aligned}$$

As $M \to \infty$, all the terms on the right-hand side converge to 0, which entails a contradiction. □

Lemma 2.9. *Suppose that Assumptions 2.1, 2.2, and 2.4 are satisfied. Take* $N \in \mathbb{N}$. *Then, there exists a constant* $c_{\mathrm{LB}} = c_{\mathrm{LB}}(A, \mathcal{B}, N) > 0$ *such that*

$$\|\Sigma_\varepsilon^{-1/2}ABP_Nu\|_Y^2 \geq c_{\mathrm{LB}}\|P_Nu\|_{\ell^2}^2$$

for all $B \in \mathcal{B}$ *and* $u \in \ell^2$.

Proof. We prove this claim by contradiction. Assume that there exist two sequences $\{u_M\}_{M=1}^{\infty} \subset \ell^2$ and $\{B_M\}_{M=1}^{\infty} \subset \mathcal{B}$ such that

$$\frac{\|\Sigma_\varepsilon^{-1/2}AB_MP_Nu_M\|_Y^2}{\|P_Nu_M\|_{\ell^2}^2} < \frac{1}{M} \quad \forall M. \tag{18}$$

Let us write $p_M = \frac{P_Nu_M}{\|P_Nu_M\|_{\ell^2}}$ to obtain that a sequence $\{p_M\}_{M=1}^{\infty} \subset H_N$, where $H_N = \{u \in \ell^2 : u_k = 0 \ \forall k > N, \|u\|_{\ell^2} = 1\}$. Notice that $H_N \subset X$ is a finite-dimensional, bounded and closed subset and hence compact. Rephrasing (18) we have

$$\|\Sigma_\varepsilon^{-1/2}AB_Mp_M\|_Y^2 < \frac{1}{M} \quad \forall M.$$

Since both $\{p_M\}_{M=1}^{\infty} \subset H_N$ and $\{B_M\}_{M=1}^{\infty} \subset \mathcal{B}$ belong to compact sets, up to a subsequence, the sequences converge to limit points $p \in H$ and $B \in \mathcal{B}$, respectively. By the continuity of $\Sigma_\varepsilon^{-1/2}A$ and by the equi-continuity of the operators in $\mathcal{B}$, we deduce that $\|\Sigma_\varepsilon^{-1/2}ABp\|_Y = 0$, which by the injectivity of $\Sigma_\varepsilon^{-1/2}$ and by Assumption 2.2 implies that $p = 0$. This yields a contradiction with the assumption that $\|p\| = 1$ and completes the proof. □

We can finally state and prove the main result of this subsection:

Proposition 2.10. *Let A, Σ_ε and $\mathcal{B}$ satisfy Assumptions 1.1, 2.1 and 2.4, and let $\|y\|_Y \le Q$. Let $\hat{u}_B \in \ell^2$ be a minimizer of J_B being $B \in \mathcal{B}$ and $y \in \mathbb{B}_Y(Q)$. For every M, there exists a constant $\tilde{c}_{\mathrm{ST}} = \tilde{c}_{\mathrm{ST}}(M; A, \Sigma_\varepsilon, \mathcal{B}, Q)$ such that for any $B \in \mathcal{B}$ we have*

$$\{v \in \ell^2 \mid J_B(v) \le M\} \subset \{v \in \ell^2 \mid \|v - \hat{u}_B\|_{\ell^2}^2 \le \tilde{c}_{\mathrm{ST}}(J_B(v) - J_B(\hat{u}_B))\}. \tag{19}$$

Proof. For clarity, the proof is divided into several steps.

Step 1. Let us show that there exists $N_0 = N_0(A, \Sigma_\varepsilon, \mathcal{B}, Q)$ such that $P_{N_0}^\perp \hat{u}_B = 0$. Rephrasing Lemma 2.8, we have that there exists a sequence δ_N, $N = 1, 2, ...$, converging to zero such that

$$\|P_N^\perp \hat{w}_B\|_{\ell^2} \le \delta_N \quad \forall B \in \mathcal{B},\ y \in \mathbb{B}_Y(Q).$$

Therefore, it is possible to pick N_0 such that

$$\|P_{N_0}^\perp \hat{w}_B\|_{\ell^2} \le \frac{1}{2}$$

for all $B \in \mathcal{B}$ and $y \in \mathbb{B}_Y(Q)$. Notice that such N_0 is independent of the specific choice of B, and depends on y only through Q: thus, we denote it as $N_0(A, \Sigma_\varepsilon, \mathcal{B}, Q)$. Since $\|P_{N_0}^\perp \hat{w}_B\|_{\ell^\infty} \le \|P_{N_0}^\perp \hat{w}_B\|_{\ell^2}$, we also deduce that

$$|\hat{w}_{B,k}| := |[\hat{w}_B]_k| \le \frac{1}{2} \quad \text{for } k > N_0, \tag{20}$$

and by the optimality condition satisfied by $\hat{w}_B$, together with the characterization of the subdifferential in (17) we get

$$\hat{u}_{B,k} := [\hat{u}_B]_k = 0 \quad \text{for } k > N_0, \tag{21}$$

whence $P_{N_0}^\perp \hat{u}_B = 0$, or $P_{N_0} \hat{u}_B = \hat{u}_B$, independently on B.

Step 2. We show that

$$\|P_{N_0}^\perp (v - \hat{u}_B)\|_{\ell^2}^2 \le 2 c_{\mathrm{UB}}(M) \left(J_B(v) - J_B(\hat{u}_B) - \frac{1}{2} \|\Sigma_\varepsilon^{-1/2} AB(v - \hat{u}_B)\|_Y^2 \right), \tag{22}$$

where $c_{\mathrm{UB}}(M) = c_{\mathrm{UB}}(M; A, \Sigma_\varepsilon, Q, |\mathcal{B}|)$ is given in Proposition 2.7. Let us denote

$$r_B(v) = J_B(v) - J_B(\hat{u}_B) - \frac{1}{2} \|\Sigma_\varepsilon^{-1/2} AB(v - \hat{u}_B)\|_Y^2.$$

By direct computations, we have

$$\begin{aligned} r_B(v) &= \Phi(v) - \Phi(\hat{u}_B) + \frac{1}{2}\|\Sigma_\varepsilon^{-1/2} ABv\|_Y^2 - \frac{1}{2}\|\Sigma_\varepsilon^{-1/2} AB\hat{u}_B\|_Y^2 - \frac{1}{2}\|\Sigma_\varepsilon^{-1/2} AB(v - \hat{u}_B)\|_Y^2 \\ &\quad - \langle y, \Sigma_\varepsilon^{-1} AB(v - \hat{u}_B)\rangle_Y \\ &= \Phi(v) - \Phi(\hat{u}_B) + \langle \Sigma_\varepsilon^{-1/2} AB\hat{u}_B, \Sigma_\varepsilon^{-1/2} AB(v - \hat{u}_B)\rangle_Y - \langle B^*(\Sigma_\varepsilon^{-1} A)^* y, v - \hat{u}_B\rangle_{\ell^2} \\ &= \Phi(v) - \Phi(\hat{u}_B) + \langle B^*(\Sigma_\varepsilon^{-1} A)^* (AB\hat{u}_B - y), v - \hat{u}_B\rangle_{\ell^2}. \end{aligned}$$

Using the definition of $\hat{w}_B$ in (16), we obtain the expression

$$r_B(v) = \Phi(v) - \Phi(\hat{u}_B) - \langle \hat{w}_B, v - \hat{u}_B\rangle_{\ell^2} = \sum_{k \in \mathbb{N}} \left(|v_k| - |\hat{u}_{B,k}| - \hat{w}_{B,k}(v_k - \hat{u}_{B,k}) \right).$$

Moreover, thanks to (17), it is easy to verify that each term of the previous summation is non-negative. Consider now N_0 introduced in Step 1: by (20) and (21),

$$\begin{aligned} r_B(v) &\ge \sum_{k > N_0} \left(|v_k| - |\hat{u}_{B,k}| - \hat{w}_{B,k}(v_k - \hat{u}_{B,k}) \right) = \sum_{k > N_0} \left(|v_k| - \hat{w}_{B,k} v_k \right) \\ &\ge \sum_{k > N_0} \left(|v_k| - \frac{1}{2}|v_k| \right) = \frac{1}{2}\|P_{N_0}^\perp v\|_{\ell^1} \ge \frac{1}{2}\|P_{N_0}^\perp v\|_{\ell^2} = \frac{1}{2}\|P_{N_0}^\perp (v - \hat{u}_B)\|_{\ell^2}. \end{aligned}$$

In conclusion,

$$\|P_{N_0}^{\perp}(v-\hat{u}_B)\|_{\ell^2}^2 \le 2r_B(v)\|P_{N_0}^{\perp}(v-\hat{u}_B)\|_{\ell^2} \le 2r_B(v)\|v\|_{\ell^2}.$$

Since $J_B(v) \le M$, by Proposition 2.7 we have $\|v\|_{\ell^2} \le c_{\mathrm{UB}}(M)$, hence (22) holds.

Step 3. Consider the projection operators associated with the choice $N = N_0$ as in Step 1:

$$\|v-\hat{u}_B\|_{\ell^2}^2 = \|P_{N_0}(v-\hat{u}_B)\|_{\ell^2}^2 + \|P_{N_0}^{\perp}(v-\hat{u}_B)\|_{\ell^2}^2. \tag{23}$$

We can bound the first term on the right-hand side of (23) by means of Lemma 2.9 and get, for any $B \in \mathcal{B}$, that

$$\begin{aligned}\|P_{N_0}(v-\hat{u}_B)\|_{\ell^2}^2 &\le \frac{1}{c_{\mathrm{LB}}}\|\Sigma_\varepsilon^{-1/2}ABP_{N_0}(v-\hat{u}_B)\|_Y^2 \\ &\le \frac{2}{c_{\mathrm{LB}}}\|\Sigma_\varepsilon^{-1/2}AB(v-\hat{u}_B)\|_Y^2 + \frac{2L_{F'}}{c_{\mathrm{LB}}}\|P_{N_0}^{\perp}(v-\hat{u}_B)\|_{\ell^2}^2,\end{aligned}$$

where the constant c_{LB} depends on N_0, which is nevertheless independent of B. Collecting now the terms and observing that (22) trivially holds with the larger constant

$$C = \max\left\{2c_{\mathrm{UB}}(M), \frac{4}{c_{\mathrm{LB}}}\left(1+\frac{2L_{F'}}{c_{\mathrm{LB}}}\right)^{-1}\right\},$$

by (23) we obtain

$$\begin{aligned}\|v-\hat{u}_B\|_{\ell^2}^2 &\le \frac{2\|\Sigma_\varepsilon^{-1/2}AB(v-\hat{u}_B)\|_Y^2}{c_{\mathrm{LB}}} + \left(1+\frac{2L_{F'}}{c_{\mathrm{LB}}}\right)\|P_{N_0}^{\perp}(v-\hat{u}_B)\|_{\ell^2}^2 \\ &\le \frac{2\|\Sigma_\varepsilon^{-1/2}AB(v-\hat{u}_B)\|_Y^2}{c_{\mathrm{LB}}} + C\left(1+\frac{2L_{F'}}{c_{\mathrm{LB}}}\right)\left(J_B(v) - J_B(\hat{u}_B) - \frac{1}{2}\|\Sigma_\varepsilon^{-1/2}AB(v-\hat{u}_B)\|_Y^2\right).\end{aligned}$$

Since $\frac{2}{c_{\mathrm{LB}}} - \frac{C}{2}\left(1+\frac{2L_{F'}}{c_{\mathrm{LB}}}\right) \le 0$ we have that (19) holds with

$$\tilde{c}_{\mathrm{ST}} = C\left(1+\frac{2L_{F'}}{c_{\mathrm{LB}}}\right) = \max\left\{2c_{\mathrm{UB}}(M)\left(1+\frac{2L_{F'}}{c_{\mathrm{LB}}}\right), \frac{4}{c_{\mathrm{LB}}}\right\}. \tag{24}$$

This completes the proof. □

Now the uniqueness of the minimizer $\hat{u}_B$ follows directly from Proposition 2.10.

Proof of Theorem 2.3 (uniqueness). Let $\hat{u}$ and $\hat{u}'$ be two minimizers of J_B. By Proposition 2.10, choosing $v = \hat{u}'$ and $M = 0$ (indeed, $J_B(\hat{u}) = J_B(\hat{u}') \le J_B(0) = 0$) we would have

$$\|\hat{u}-\hat{u}'\|_{\ell^2}^2 \le \tilde{c}_{\mathrm{ST}}(J_B(\hat{u}) - J_B(\hat{u}')) = 0.$$

□

Finally, we obtain a stability estimate for the perturbation of $\hat{u}_B$ with respect to modifications of B within the class $\mathcal{B}$.

Proof of Theorem 2.5. Notice that, thanks to the uniform bound on the minimizers (15) it holds that, independently of $B_1, B_2 \in \mathcal{B}$,

$$\begin{aligned}J_{B_2}(\hat{u}_1) &= \frac{1}{2}\|\Sigma_\varepsilon^{-1/2}AB_2\hat{u}_1\|_Y^2 - \langle y, \Sigma_\varepsilon^{-1}AB_2\hat{u}_1\rangle_Y + \|\hat{u}_1\|_{\ell^1} \\ &\le \frac{1}{2}L_{F'}c_{\mathrm{UB}} + Q\|\Sigma_\varepsilon^{-1}A\||\mathcal{B}|c_{\mathrm{UB}} + c_{\mathrm{UB}}.\end{aligned} \tag{25}$$

Then, we can apply Proposition 2.10 to J_{B_2} for the choice $v = \hat{u}_1$ with a M-level set specified by the upper bound in (25). We obtain

$$\|\hat{u}_1-\hat{u}_2\|_{\ell^2}^2 \le \tilde{c}_{\mathrm{ST}}(J_{B_2}(\hat{u}_1) - J_{B_2}(\hat{u}_2)),$$

where the explicit expression of the constant $\tilde{c}_{\mathrm{ST}}$ is given in (24). Now, since $J_{B_1}(\hat{u}_1) \leq J_{B_1}(\hat{u}_2)$, we have

$$\|\hat{u}_1 - \hat{u}_2\|_{\ell^2}^2 \leq \tilde{c}_{\mathrm{ST}}\big(J_{B_2}(\hat{u}_1) - J_{B_2}(\hat{u}_2) + J_{B_1}(\hat{u}_2) - J_{B_1}(\hat{u}_1)\big).$$

Consider now the terms $J_{B_2}(\hat{u}_1) - J_{B_1}(\hat{u}_1)$: by direct computations, we get

$$\begin{aligned} J_{B_2}(\hat{u}_1) - J_{B_1}(\hat{u}_1) &= \|\Sigma_\varepsilon^{-1/2} A B_2 \hat{u}_1\|_Y^2 - \|\Sigma_\varepsilon^{-1/2} A B_1 \hat{u}_1\|_Y^2 - \langle y, \Sigma_\varepsilon^{-1} A (B_2 - B_1)\hat{u}_1\rangle_Y \\ &= \langle \Sigma_\varepsilon^{-1} A (B_2 + B_1)\hat{u}_1, A(B_2 - B_1)\hat{u}_1\rangle_Y - \langle \big(\Sigma_\varepsilon^{-1} A\big)^* y, (B_2 - B_1)\hat{u}_1\rangle. \end{aligned}$$

Analogously, it holds

$$J_{B_1}(\hat{u}_2) - J_{B_2}(\hat{u}_2) = \langle \Sigma_\varepsilon^{-1} A (B_1 + B_2)\hat{u}_2, A(B_1 - B_2)\hat{u}_2\rangle_Y - \langle \big(\Sigma_\varepsilon^{-1} A\big)^* y, (B_1 - B_2)\hat{u}_2\rangle.$$

Summing the two expressions, we obtain

$$\begin{aligned} \|\hat{u}_1 - \hat{u}_2\|_{\ell^2}^2 \leq \tilde{c}_{\mathrm{ST}}\big(&\langle \Sigma_\varepsilon^{-1} A (B_1 + B_2)\hat{u}_1, A(B_2 - B_1)\hat{u}_1\rangle_Y - \langle \Sigma_\varepsilon^{-1} A (B_1 + B_2)\hat{u}_2, A(B_2 - B_1)\hat{u}_2\rangle_Y \\ &- \langle \big(\Sigma_\varepsilon^{-1} A\big)^* y, (B_2 - B_1)\hat{u}_1\rangle + \langle \big(\Sigma_\varepsilon^{-1} A\big)^* y, (B_2 - B_1)\hat{u}_2\rangle\big) \\ = \tilde{c}_{\mathrm{ST}}\big(&\langle \Sigma_\varepsilon^{-1} A (B_1 + B_2)\hat{u}_1, A(B_2 - B_1)(\hat{u}_1 - \hat{u}_2)\rangle_Y \\ &+ \langle \Sigma_\varepsilon^{-1} A (B_1 + B_2)(\hat{u}_1 - \hat{u}_2), A(B_2 - B_1)\hat{u}_2\rangle_Y \\ &- \langle \big(\Sigma_\varepsilon^{-1} A\big)^* y, (B_2 - B_1)(\hat{u}_1 - \hat{u}_2)\rangle\big) \\ \leq \tilde{c}_{\mathrm{ST}}\big(&4\|\Sigma_\varepsilon^{-1} A\||\mathcal{B}|c_{\mathrm{UB}}\|A\| + \|\Sigma_\varepsilon^{-1} A\|Q\big)\|B_1 - B_2\|\|\hat{u}_1 - \hat{u}_2\|_{\ell^2}. \end{aligned}$$

In conclusion, letting $C = \|\Sigma_\varepsilon^{-1} A\|\big(4|\mathcal{B}|c_{\mathrm{UB}}\|A\| + Q\big)$, we have

$$\|\hat{u}_1 - \hat{u}_2\|_{\ell^2} \leq C\tilde{c}_{\mathrm{ST}}\|B_1 - B_2\|,$$

and, since $\hat{x}_i = B_i\hat{u}_i$, it follows that

$$\begin{aligned} \|\hat{x}_1 - \hat{x}_2\|_X &\leq \|B_1 - B_2\|\|\hat{u}_1\|_{\ell^2} + \|B_2\|\|\hat{u}_1 - \hat{u}_2\|_{\ell^2} \\ &\leq (c_{\mathrm{UB}} + |\mathcal{B}|C\tilde{c}_{\mathrm{ST}})\|B_1 - B_2\| \end{aligned}$$

which concludes the proof with the choice $c_{\mathrm{ST}} = c_{\mathrm{UB}} + |\mathcal{B}|C\tilde{c}_{\mathrm{ST}}$. □

3 Statistical learning framework

As shown in the previous section, for every choice of $B \in \mathcal{B}$ and any noisy output $y \in Y$, we can associate a solution $x = R_B(y) \in X$ of the inverse problem $Ax = y$, which depends on B. As discussed in Section 1 (see in particular Assumption 1.1), the output y and hence the solution $R_B(y)$ are random variables, and the optimal $B^\star$ is defined as the minimizer of the expected risk L defined by (4). Since L depends on the unknown distribution of the pair (x, y), $B^\star$ is estimated by its empirical version $\widehat{B}$, given by (7). In this section, we provide a finite sample bound on the discrepancy

$$L(\widehat{B}) - L(B^\star),$$

under the assumption that both x and ε are bounded random variables, and $\mathcal{B}$ is compact. In fact, boundedness is assumed for convenience; for similar techniques with sub-Gaussian random variables, see e.g. [44].

Assumption 3.1. *There exists $Q_0 > 0$ such that $\|x\|_X \leq Q_0$ and $\|\varepsilon\|_Y \leq Q_0$ almost surely.*

Assumption 3.1 directly implies a bound $\|y\|_Y \leq Q = (\|A\| + 1)Q_0$, which was used in Section 2.

The following result shows the existence of $B^\star$ and $\widehat{B}$, and provides a finite sample bound on $L(\widehat{B}) - L(B^\star)$ in probability. It is proved analogously to Proposition 4 in [20] (see also [5, Lemma A.5]). We recall that, for any $r > 0$, $\mathcal{N}(\mathcal{B}, r)$ denotes the *covering number* of $\mathcal{B}$, i.e., the minimum number of balls of radius r (in the strong operator norm on $\mathcal{L}(\ell^2, X)$) whose union contains $\mathcal{B}$.

Theorem 3.2. *Let Assumptions 1.1, 2.1, 2.2, 2.4, and 3.1 hold. There exist a minimizer $B^\star$ of L and, with probability 1, a minimizer $\widehat{B}$ of $\widehat{L}$ over $\mathcal{B}$. Furthermore, for all $\eta > 0$,*

$$\mathbb{P}_{\mathbf{z}\sim\rho^m}\left[L(\widehat{B}) - L(B^\star) \leq \eta\right] \geq 1 - 2\,\mathcal{N}(\mathcal{B}, C\eta)e^{-C'm\eta^2} \tag{26}$$

for some $C, C' > 0$ depending only on A, Σ_ε, Q_0 and $\mathcal{B}$.

In the above statement, the quantity $\widehat{B}$ depends on $\mathbf{z}$ and is therefore a random variable. For ease and convenience, in the rest of the section, we denote by $\mathbb{P}$ and $\mathbb{E}$ the probability and the expected value computed with respect to the sample $\mathbf{z}$, respectively.

Proof. Consider two different elements $B_1, B_2 \in \mathcal{B}$. Then, ρ-a.e. in $X \times Y$

$$\begin{aligned}\left|\|R_{B_1}(y) - x\|_X^2 - \|R_{B_2}(y) - x\|_X^2\right| &= |\langle R_{B_1}(y) - R_{B_2}(y), R_{B_1}(y) - x + R_{B_2}(y) - x\rangle| \\ &\leq \|R_{B_1}(y) - R_{B_2}(y)\|_X(\|R_{B_1}(y)\|_X + \|R_{B_2}(y)\|_X + 2Q_0) \\ &\leq 2(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}}\|B_1 - B_2\|,\end{aligned}$$

here we have used Theorem 2.5 and (15). By integrating with respect to the probability distribution ρ, the above bound holds for L. Indeed,

$$\begin{aligned}|L(B_1) - L(B_2)| &= \left|\mathbb{E}[\|R_{B_1}(y) - x\|_X^2] - \mathbb{E}[\|R_{B_2}(y) - x\|_X^2]\right| \\ &\leq \mathbb{E}\left[|\|R_{B_1}(y) - x\|_X^2 - \|R_{B_2}(y) - x\|_X^2|\right] \\ &\leq 2(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}}\|B_1 - B_2\|,\end{aligned} \tag{27}$$

and, by replacing ρ with its empirical counterpart, with probability 1,

$$\begin{aligned}|\widehat{L}(B_1) - \widehat{L}(B_2)| &= \left|\frac{1}{m}\sum_{j=1}^m \|R_{B_1}(y_j) - x_j\|_X^2 - \frac{1}{m}\sum_{j=1}^m \|R_{B_2}(y_j) - x_j\|_X^2\right| \\ &\leq \frac{1}{m}\sum_{j=1}^m \left|\|R_{B_1}(y_j) - x_j\|_X^2 - \|R_{B_2}(y_j) - x_j\|_X^2\right| \\ &\leq 2(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}}\|B_1 - B_2\|.\end{aligned} \tag{28}$$

Since both L and $\widehat{L}$ are continuous functionals on $\mathcal{B}$ and $\mathcal{B}$ is compact with respect to the operator topology, the corresponding minimizers $B^\star$ and $\widehat{B}$ over $\mathcal{B}$ exist (almost surely for $\widehat{B}$).

Next, we notice that

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

which ultimately implies that

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

since the central difference is negative by the definition of $\widehat{B}$. Thus,

$$\mathbb{P}\left[L(\widehat{B}) - L(B^\star) \leq \eta\right] \geq \mathbb{P}\left[\sup_{B\in\mathcal{B}} |\widehat{L}(B) - L(B)| \leq \frac{\eta}{2}\right].$$

We now provide a lower bound for the latter term. In view of (27) and (28), by using the reverse triangle inequality, for every $B_1, B_2 \in \mathcal{B}$,

$$\begin{aligned}\left||\widehat{L}(B_1) - L(B_1)| - |\widehat{L}(B_2) - L(B_2)|\right| &\leq |\widehat{L}(B_1) - \widehat{L}(B_2)| + |L(B_1) - L(B_2)| \\ &\leq 4(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}}\|B_1 - B_2\|.\end{aligned}$$

Let now $N = \mathcal{N}\left(\mathcal{B}, \frac{\eta}{16(|\mathcal{B}|c_{\mathrm{UB}}+Q_0)c_{\mathrm{ST}}}\right)$ and consider a discrete set $B_1, \dots, B_N$ such that the balls $\mathcal{B}_k$ centered at B_k with radius $r = \frac{\eta}{16(|\mathcal{B}|c_{\mathrm{UB}}+Q_0)c_{\mathrm{ST}}}$ cover the entire $\mathcal{B}$. In each ball $\mathcal{B}_k$, for every $B \in \mathcal{B}_k$ it holds

$$\left||\widehat{L}(B) - L(B)| - |\widehat{L}(B_k) - L(B_k)|\right| \leq 4(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}}\|B - B_k\| \leq \frac{\eta}{4}.$$

Therefore, the event $|\widehat{L}(B) - L(B)| > \frac{\eta}{2}$ implies the event $|\widehat{L}(B_k) - L(B_k)| > \frac{\eta}{4}$, and a bound (in probability) of this term can be provided by standard concentration results. Indeed, $\widehat{L}(B_k)$ is the sample average of m realizations of the random variable $\|R_{B_k}(y) - x\|_X^2$, whose expectation is $L(B_k)$. Moreover, such random variable is bounded by $(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)^2$ by assumption, and therefore via Hoeffding's inequality

$$\mathbb{P}\left[\sup_{B\in\mathcal{B}_k} |\widehat{L}(B) - L(B)| > \frac{\eta}{2}\right] \leq \mathbb{P}\left[|\widehat{L}(B_k) - L(B_k)| > \frac{\eta}{4}\right] \leq 2e^{-\frac{m\eta^2}{8(|\mathcal{B}|c_{\mathrm{UB}}+Q_0)^4}}.$$

Notice that this inequality holds uniformly in k. Finally, we obtain

$$\begin{aligned}
\mathbb{P}\left[\sup_{B\in\mathcal{B}} |\widehat{L}(B) - L(B)| \leq \frac{\eta}{2}\right] &= 1 - \mathbb{P}\left[\sup_{B\in\mathcal{B}} |\widehat{L}(B) - L(B)| > \frac{\eta}{2}\right] \\
&\geq 1 - \sum_{k=1}^{N} \mathbb{P}\left[\sup_{B\in\mathcal{B}_k} |\widehat{L}(B) - L(B)| > \eta\right] \\
&\geq 1 - 2Ne^{-\frac{m\eta^2}{8(|\mathcal{B}|c_{\mathrm{UB}}+Q_0)^4}}.
\end{aligned}$$

This proves (26), with $C = (16(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}})^{-1}$ and $C' = (8(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)^4)^{-1}$. □

We now provide a more explicit expression for the sample error estimate under the assumption that the covering numbers of the set $\mathcal{B}$ have a specific decay rate. It is possible to obtain similar bounds, for example, whenever the singular values of the compact embedding of $\mathcal{B}$ into its ambient space show a polynomial decay.

Corollary 3.3. *Let Assumptions 1.1, 2.1, 2.2, 2.4, and 3.1 hold. Assume further that*

$$\log(\mathcal{N}(\mathcal{B}, r)) \leq Cr^{-1/s}, \tag{29}$$

holds for some constants $C, s > 0$. Suppose that $\tau > 0$. Then, for sufficiently large values of m,

$$\mathbb{P}\left[L(\hat{B}) - L(B^\star) \leq \left(\frac{\alpha_1 + \alpha_2\sqrt{\tau}}{\sqrt{m}}\right)^{1-\frac{1}{1+2s}}\right] \geq 1 - e^{-\tau} \tag{30}$$

for some $\alpha_1, \alpha_2 > 0$ depending only on A, $\mathcal{B}$, Σ_ε, and Q_0.

Proof. Let $a_1 = C\big(16(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)c_{\mathrm{ST}}\big)^{1/s}$ and $a_2 = (8(|\mathcal{B}|c_{\mathrm{UB}} + Q_0)^4)^{-1}$: then, by (26) and (29), for $\eta \in (0, 1]$ and m sufficiently large (namely, such that $a_2 m\eta^{2+1/s} \geq a_1$) we have

$$\mathbb{P}\left[L(\widehat{B}) - L(B^\star) \leq \eta\right] \geq 1 - e^{-\eta^{-1/s}(a_2 m\eta^{2+1/s} - a_1)} \geq 1 - e^{-(a_2 m\eta^{2+1/s} - a_1)} = 1 - e^{-\tau}, \tag{31}$$

being $\tau = a_2 m\eta^{2+1/s} - a_1$. We can rephrase (31) by expressing η as a function of τ and m:

$$1 - e^{-\tau} \leq \left[L(\widehat{B}) - L(B^\star) \leq \left(\frac{a_1 + \tau}{a_2 m}\right)^{\frac{1}{2+1/s}}\right] \leq \mathbb{P}\left[L(\hat{B}) - L(B^\star) \leq \left(\frac{\alpha_1 + \alpha_2\sqrt{\tau}}{\sqrt{m}}\right)^{1-\frac{1}{1+2s}}\right],$$

where $\alpha_1 = \sqrt{a_1/a_2}$ and $\alpha_2 = \sqrt{1/a_2}$, as $\sqrt{a_1 + \tau} \leq \sqrt{a_1} + \sqrt{\tau}$. This concludes the proof. □

Note that, besides the bound in probability (30), we can also provide a bound in expectation for the excess risk. Indeed, by (31),

$$\mathbb{P}\left[L(\widehat{B}) - L(B^\star) < \eta\right] \leq \min\left\{1, e^{a_1\eta^{-1/s} - a_2 m\eta^2}\right\},$$

and by the tail integral formula (notice that $e^{a_1\eta^{-1/s} - a_2 m\eta^2} = 1$ for $\eta = \hat{\eta} = k_1 m^{-\frac{s}{2s+1}}$)

$$\begin{aligned}\mathbb{E}\big[L(\hat{B}) - L(B^\star)\big] &= \int_0^\infty \mathbb{P}\left[L(\widehat{B}) - L(B^\star) < \eta\right] d\eta \\ &\leq \hat{\eta} + e^{a_1\hat{\eta}^{-1/s}} \int_{\hat{\eta}}^\infty e^{-a_2 m\eta^2} d\eta \leq k_1 m^{-\frac{s}{2s+1}} + k_2 m^{-\frac{1}{2}},\end{aligned}$$

which allows us to conclude that, for sufficiently large m,

$$\mathbb{E}\big[L(\hat{B}) - L(B^\star)\big] \lesssim m^{-\frac{s}{2s+1}}. \tag{32}$$

The bound (32) should be compared with the one in [44, Corollary 7.2] setting $\alpha = 1$ and $q = 2$, namely

$$\mathbb{E}\big[L(\hat{B}) - L(B^\star)\big] \lesssim m^{-\frac{1}{2}} \tag{33}$$

provided that $s > \frac{1}{2}$, compare with the assumptions of [44, Proposition 7.3]. The rate in (32) is worse than the rate in (33), obtained via a chaining argument. On the other side, the bound in Corollary 3.3 holds in probability, whereas the bound (33) holds only in expectation.

4 On the choice of the parameter class $\mathcal{B}$

4.1 Compact perturbations of a known operator

We first consider an example of a class $\mathcal{B}$ consisting of certain compact perturbations of a reference operator B_0. Under rather general assumptions, we can provide an estimate of the covering numbers of $\mathcal{B}$. The proofs of this section are standard, and are postponed to Appendix C.

Assume that A is injective. We fix an injective operator $B_0 \colon \ell^2 \to X$ and two compact operators $E_1, E_2 \colon \ell^2 \to \ell^2$ such that $\|B_0\|_{\mathcal{L}(\ell^2,X)} \leq 1$ and $\|E_i\|_{\mathcal{L}(\ell^2)} \leq 1$ for $i = 1, 2$ (here, $\mathcal{L}(X) := \mathcal{L}(X, X)$). For every finite set $I \subset \mathbb{N}$, let $c_I > 0$. We define

$$\mathcal{B} = \{B_0(\mathrm{Id} + K) : K \in \mathcal{H}\}, \tag{34}$$

where the set of operators $\mathcal{H}$ is defined as

$$\begin{aligned}\mathcal{H} = \{K = E_1 T E_2 : T \in \mathcal{L}(\ell^2),\ \|T\|_{\mathcal{L}(\ell^2)} \leq 1 \text{ and} \\ \|(\mathrm{Id} + K)u\|_{\ell^2} \geq c_I \|u\|_{\ell^2} \text{ for every finite } I \subset \mathbb{N} \text{ and } u \in \ell^2_I\},\end{aligned} \tag{35}$$

where $\ell^2_I := \mathrm{span}\{e_i : i \in I\}$.

Lemma 4.1. *The set $\mathcal{B}$ defined in (34)-(35) satisfies Assumption 2.4.*

The theoretical results of Section 3 rely on the covering numbers of $\mathcal{B}$. We now derive an estimate of the form (29) in the case when $\mathcal{B}$ is given by (34). For the sake of clarity, we indicate the norm used to perform the covering explicitly.

Proposition 4.2. *Let $\mathcal{B}$ be defined as in (34)-(35). Suppose that $E_1 = E_2 = E$, being $E \in \mathcal{L}(\ell^2)$ a positive operator whose eigenvalues λ_n satisfy*

$$\lambda_n \leq c n^{-s}, \qquad n \in \mathbb{N}$$

for some $c, s > 0$. Take $s' < s$. Then there exists $C > 0$ such that

$$\log \mathcal{N}(\mathcal{B}, r; \|\cdot\|_{\mathcal{L}(\ell^2,X)}) \leq C r^{-\frac{2s+1}{2ss'}}, \qquad r > 0. \tag{36}$$

With this estimate on the covering numbers, it is possible to apply Corollary 3.3 and obtain a finite sample bound in the case when $\mathcal{B}$ is defined as in (34). It is worth observing that the most relevant approaches in the literature focusing on learning sparsifying dictionaries through the solution of a bilevel problem, such as [31, 32, 27, 26], are formulated in a finite-dimensional setup, and so naturally satisfy the assumptions of this example.

4.2 Learning the mother wavelet

We consider here the problem of learning the mother wavelet. In other words, the set $\mathcal{B}$ will consist of synthesis operators associated to a family of wavelets.

Consider the Hilbert space $X = L^2(\mathbb{R}^d)$ and, for $\psi \in L^2(\mathbb{R}^d)$, define

$$\psi_{j,k}(x) = 2^{\frac{jd}{2}} \psi(2^j x - k), \qquad \text{for a.e. } x \in \mathbb{R}^d,\ j \in \mathbb{Z},\ k \in \mathbb{Z}^d.$$

For $f \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ (the space $L^1(\mathbb{R}^d)$ is added just to simplify the exposition, thanks to the continuity of $\hat{f}$), define

$$\|f\|_W^2 = \sup_{\xi \in \mathbb{R}^d} \sum_{j,k} |\hat{f}(2^{-j}\xi + 2\pi k)|^2, \qquad \hat{f}(s) = \int_{\mathbb{R}^d} f(x) e^{-ix\cdot s}\, dx,$$

and consider

$$W = \{f \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) : \|f\|_W < +\infty\}.$$

It is easy to verify that $\|\cdot\|_W$ is a norm on W. It is well known that, for $\psi \in W$, the family $\{\psi_{j,k}\}_{j\in\mathbb{Z},k\in\mathbb{Z}^d}$ is a Bessel sequence of $L^2(\mathbb{R}^d)$, namely, the synthesis operator

$$B_\psi \colon \ell^2(\mathbb{Z}\times\mathbb{Z}^d) \to L^2(\mathbb{R}^d), \qquad (c_{j,k}) \mapsto \sum_{j,k} c_{j,k}\psi_{j,k},$$

is well defined and bounded (see, e.g., [21, Section 3.3] and [33, Theorem 3]).

We now construct a compact family of "mother wavelets" ψ in W. (Note that the term mother wavelet here is used even though the family $\psi_{j,k}$ need not be a frame of $L^2(\mathbb{R}^d)$.)

Lemma 4.3. *Let $\Psi \subseteq W$ be a compact set (with respect to $\|\cdot\|_W$) and $a > 0$. Set*

$$\Psi_a = \{\psi \in \Psi : \|B_\psi x\|_{L^2(\mathbb{R}^d)} \geq a\|x\|_{\ell^2}\ \forall x \in \ell^2\}.$$

Then

$$\mathcal{B}_{\Psi_a} = \{B_\psi : \psi \in \Psi_a\}$$

is a compact subset of $\mathcal{L}(\ell^2, L^2(\mathbb{R}^d))$ with respect to the operator norm, and

$$\mathcal{N}\left(\mathcal{B}_{\Psi_a}, \delta; \|\cdot\|_{\mathcal{L}(\ell^2,L^2(\mathbb{R}^d))}\right) \leq \mathcal{N}\left(\Psi, (2\pi)^{\frac{3}{2}d}\,\delta; \|\cdot\|_W\right), \qquad \delta > 0.$$

Proof. By the estimates in the proof of Theorem 3 in [33], we have

$$\sum_{j,k} |\langle f, \psi_{j,k}\rangle_{L^2(\mathbb{R}^d)}|^2 \leq (2\pi)^{-3d}\|\psi\|_W^2\|f\|_{L^2(\mathbb{R}^d)}^2, \qquad \psi \in W,\ f \in L^2(\mathbb{R}^d).$$

In other words, we have the following bound on the norm of the analysis operator: $\|B_\psi^*\| \leq (2\pi)^{-\frac{3}{2}d}\|\psi\|_W$. As a consequence, we have

$$\|B_\psi\| \leq (2\pi)^{-\frac{3}{2}d}\|\psi\|_W, \qquad \psi \in W.$$

In other words, the linear map

$$\zeta \colon W \to \mathcal{L}(\ell^2, L^2(\mathbb{R}^d)), \qquad \psi \mapsto B_\psi$$

is linear and bounded, with $\|\zeta\| \leq (2\pi)^{-\frac{3}{2}d}$.

Since ζ is bounded and Ψ is compact, we have that $\{B_\psi : \psi \in \Psi\} = \zeta(\Psi)$ is compact. Thus, $\mathcal{B}_\Psi$ is the intersection of a compact set and a closed set, and so it is compact.

Finally, the estimate on the covering numbers of $\mathcal{B}_{\Psi_a}$ immediately follows:

$$\mathcal{N}\left(\mathcal{B}_{\Psi_a}, \delta\right) \leq \mathcal{N}\left(\zeta(\Psi), \delta\right) \leq \mathcal{N}\left(\Psi, \|\zeta\|^{-1}\delta\right) \leq \mathcal{N}\left(\Psi, (2\pi)^{\frac{3}{2}d}\,\delta\right).$$

□

We can now combine this result with Corollary 3.3 and obtain the following corollary, in which we compare the loss corresponding to the optimal mother wavelet ψ^* with the loss corresponding to the mother wavelet $\widehat{\psi}$ obtained by minimizing the empirical risk.

Corollary 4.4. *Consider the settings of Corollary 3.3 and of Lemma 4.3. Suppose that*

$$\log \mathcal{N}(\Psi, r; \|\cdot\|_W) \leq C r^{-1/s}, \qquad r > 0,$$

for some $C, s > 0$. There exist $\alpha_1, \alpha_2 > 0$ such that the following is true. Take $\tau > 0$ and

$$\psi^* \in \operatorname*{arg\,min}_{\psi \in \Psi_a} L(B_\psi), \qquad \widehat{\psi} \in \operatorname*{arg\,min}_{\psi \in \Psi_a} \widehat{L}(B_\psi).$$

Then, with probability larger than $1 - e^{-\tau}$, we have

$$L(B_{\widehat{\psi}}) - L(B_{\psi^*}) \leq \left(\frac{\alpha_1 + \alpha_2 \sqrt{\tau}}{\sqrt{m}} \right)^{1 - \frac{1}{1+2s}}.$$

Proof. By Lemma 4.3, we have

$$\log \mathcal{N}\left(\mathcal{B}_{\Psi_a}, r\right) \leq \log \mathcal{N}\left(\Psi, (2\pi)^{\frac{3}{2}d} r\right) \leq C (2\pi)^{-\frac{3d}{2s}} r^{-1/s}.$$

The result is now a direct consequence of Corollary 3.3. □

We remark that the task of finding an optimally sparsifying wavelet transform by learning the mother wavelet has been recently studied in relation to applications both on 1D and 2D signals (see [46, 45, 25]).

5 Numerical implementation and experiments

In this section, we discuss numerical strategies through which the bilevel problem defined by (5) can be approximately solved, and present some related experiments aiming at validating the theoretical findings of the paper, showcasing a competitive comparison with alternative techniques (such as dictionary learning), and illustrating applications to challenging ill-posed problems.

Throughout this section, we assume that the spaces X and Y are finite-dimensional (and, in particular, we let $X = \mathbb{R}^n$). As a consequence, since by Assumption 1.1 the covariance Σ_ε is injective, the term $\frac{1}{2}\|\Sigma_\varepsilon^{-1/2} y\|_Y^2$ in (2) (which is irrelevant for optimization purposes) can be added to (2), obtaining the following simplified expression of the inner minimization problem:

$$R_B(y) = B\hat{u}_B = B \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\Sigma_\varepsilon^{-1/2}(ABu - y)\|_Y^2 + \|u\|_1 \right\}. \tag{37}$$

When $\Sigma_\varepsilon = \sigma^2 I$, this reduces to the familiar form of Lasso regression with ℓ^1 regularization parameter $1/\sigma^2$. Problem (37) makes clear that the regularized solution $\hat{x}_B = R_B(y)$ of Problem (2) is sparse with respect to the basis associated with the invertible matrix B.

In particular, the expression in (37) is known as the *synthesis* formulation of such a problem, as opposed to the *analysis* one in which the transform operator appears within the 1 norm.

5.1 On the numerical resolution of the bilevel problem

We now consider the bilevel problem of minimizing the empirical loss (6) for a fixed dataset $\{(x_j, y_j)\}_{j=1}^m$ subject to the inner problem (37). Such a problem is extremely challenging, since the inner minimization problem does not have an explicit formula for the solution, and the first-order optimality conditions associated with it are non-differentiable with respect to the parameter B. Among the most effective strategies developed for the solution of this bilevel problem, [31] proposed a local sensitivity analysis of the inner problem's solution, leading to an effective algorithm, of which

we consider a simple but fruitful relaxation. On the other side, in our numerical experiments, more recent methods based on automatic differentiation and neural networks proved to be unsuccessful for the general formulation of the problem (leading to hard-to-train unrolled or deep equilibrium architectures) and show good results only for the denoising problem, as we discuss in detail below. For this reason, in this subsection, we first propose a specific strategy suited for the denoising problem, and then discuss an approximate method to tackle the general formulation of the problem.

An exact formula for a denoising problem Let us consider the case in which $X = Y = \mathbb{R}^n$, the noise covariance is simply $\Sigma_\varepsilon = \sigma^2 I$, and the forward operator is $A = I$, so that the inverse problem reduces to a denoising task. Assume moreover that the minimization of the empirical loss in (7) is performed on the space $\mathcal{B} = \{B = \alpha Q : \ 0 < \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \ Q \in \mathbb{R}^{n\times n} \text{ orthogonal}\}$ for fixed $0 < \underline{\alpha} \leq \overline{\alpha}$. In this case, the inner problem (37) admits the following explicit solution:

$$R_B(y) = R_{\alpha Q}(y) = Q S_{\frac{\sigma^2}{\alpha}}(Q^T y), \tag{38}$$

where $S_\beta(u)$ denotes the (component-wise) soft-thresholding operator $[S_\beta(u)]_k = \text{sign}(u_k)(|u_k| - \beta)^+$. Although this expression is non-differentiable with respect to α and Q, it can be successfully treated via subgradient-based schemes. As a result, we can employ automatic differentiation tools to iteratively update the parameters α and Q in (38) so as to minimize the following loss:

$$\mathcal{L}_1(\alpha Q) = \widehat{L}(\alpha Q) + \frac{\mu_{\text{reg}}}{2}(\|Q^T Q - I\|^2_{\text{Fro}} + \|QQ^T - I\|^2_{\text{Fro}}),$$

where $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm and the regularization term is designed to promote the orthogonality of the matrix Q, instead of employing a projection on the Stiefel manifold at each iteration.

An approximate solver based on ℓ^1-norm relaxation We now consider a strategy to treat the bilevel problem for general forward operators, which may also be applied when the matrix A is not square. One way to overcome the non-differentiability of the outer loss with respect to B is to consider a slightly modified version of the inner problem, namely:

$$R^\nu_B(y) = B\hat{u}^\nu_B(y) = B \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \Big\{ \frac{1}{2}\|\Sigma_\varepsilon^{-1/2}(ABu - y)\|^2_Y + \|u\|_{1,\nu} \Big\}, \tag{39}$$

being $\|u\|_{1,\nu} = \sum_{i=1}^n \sqrt{u_i^2 + \nu^2}$ and $\nu > 0$ a small parameter such that the solution of Problem (39) is close to the exact solution $R_B(y)$. For the ease of notation, we assume here that $\Sigma_\varepsilon = \sigma^2 I$. Thanks to the proposed relaxation, it is possible to characterize $R^\nu_B(y)$ by the first-order optimality condition of (39), namely

$$B^T A^T (AB\hat{u}^\nu_B(y) - y) + \sigma^2 w_\nu(\hat{u}^\nu_B(y)) = 0,$$

being $[w_\nu(u)]_k = \frac{u_k}{\sqrt{u_k^2+\nu^2}}$. Finally, by differentiating such an expression, we can compute the first-order variation of $\hat{u}^\nu_B$ with respect to B along a direction $\Theta \in \mathbb{R}^{n\times n}$ as follows:

$$(\hat{u}^\nu_B)'[\Theta] = (B^T A^T AB + \sigma^2 W_\nu(\hat{u}^\nu_B))^{-1}(\Theta^T A^T (y - AB\hat{u}^\nu_B) - B^T A^T A\Theta \hat{u}^\nu_B)$$

with $W_\nu(u) = \text{diag}\left(\frac{\nu^2}{(u_k^2+\nu^2)^{3/2}}\right)$. This allows us to compute the gradient of the following regularized loss:

$$\mathcal{L}_2(B) = \widehat{L}(B) + \frac{\mu_{\text{reg}}}{2}\|B\|^2_{\text{Fro}} \ ,$$

$$\nabla\mathcal{L}_2(B) = \frac{2}{m}\sum_{j=1}^m \Big((B\hat{u}^\nu_B(y_j) - x_j - A^T AB\beta_j)\hat{u}^\nu_B(y_j)^T - (A^T(AB\hat{u}^\nu_B(y_j) - y_j))\beta_j^T \Big) + \mu_{\text{reg}} B$$

being $\beta_j = (B^T A^T AB + \sigma^2 W_\nu(\hat{u}^\nu_B(y_j)))^{-1} B^T (B\hat{u}^\nu_B(y_j) - x_j)$, and can be employed by any first-order scheme for the minimization of $\mathcal{L}_2$. In this case, regularization is not needed to enforce desirable properties of the matrix B, such as orthogonality. Instead, we include the quadratic term $\frac{\mu_{\text{reg}}}{2}\|B\|^2_{\text{Fro}}$ to guarantee the stability of the optimization scheme and reduce the risk of overfitting.

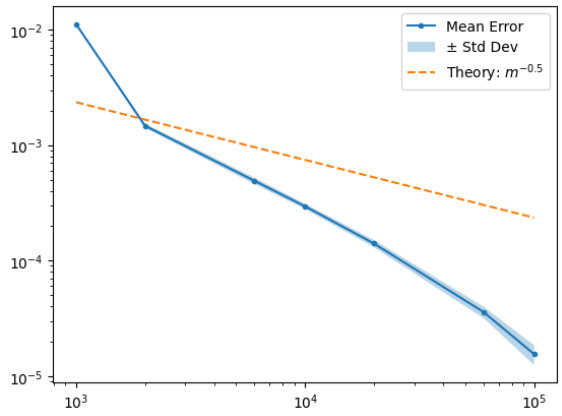

Figure 1: Decay of the average sample error for increasing sample size.

5.2 Numerical experiments

We present three numerical experiments on synthetic datasets, each with a specific goal. First, a 1D denoising problem is used to numerically validate the theoretical sample error decay from Section 3. Second, we compare our supervised strategy against dictionary learning on a 2D denoising task. Third, we demonstrate the effectiveness of the relaxation-based algorithm from Section 5.1 for 1D deblurring.

Example 1: a 1D-denoising problem For this experiment, we construct datasets of 1D discrete signals $x_j \in \mathbb{R}^n$ with $n = 100$, such that each of them has only $s = \{8, 9, 10\}$ (randomly selected) non-zero entries with values in the range $[-1, -0.5] \cup [0.5, 1]$, so that each signal x_j is sparse with respect to the canonical basis of $\mathbb{R}^n$. The observed vectors are $y_i = x_j + \varepsilon_j$, where $\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma = 0.25$, resulting in a rather low average value of the PSNR between y_j and x_j (approx. 18.10).

In this simplified scenario, the ground truth signals $\{x_j\}$ are sparse with respect to the canonical basis by construction. This provides strong prior knowledge, allowing us to assume that an optimal synthesis operator $B^\star$ (a minimizer of the expected loss L) should belong to the class of scaled identity matrices, $B = \alpha I$. This ansatz simplifies the problem to finding the optimal regularization parameter for a standard Lasso denoiser. We can therefore establish a strong benchmark by finding the optimal operator within this simplified class, $B^\star = \alpha^\star I$. We identify the optimal $\alpha^\star$ by minimizing the expected loss $L(\alpha I)$, which we approximate by performing a line search on the empirical loss $\widehat{L}(\alpha I)$ computed over a very large, separate dataset (distinct from the training sets used to find $\widehat{B}$). This $B^\star$ serves as a "near-optimal" benchmark to measure the excess risk $L(\widehat{B}) - L(B^\star)$.

In order to visualize the decay of the sample error as the size of the training set grows, we first select the sizes $m \in \{10^3, 2 \cdot 10^3, 6 \cdot 10^3, 10^4, 2 \cdot 10^4, 6 \cdot 10^4, 10^5\}$ and, for each size, we construct $K = 10$ different training sets. We then minimize the loss $\mathcal{L}_1$ associated to each set as discussed in Section 5.1, initializing the algorithm with a random orthogonal matrix and performing a maximum of `n_epochs` $= 10000$ iterations of Adam with a stepsize `learning_rate` $= 10^{-3}$ and choosing $\mu_{\text{reg}} = 10^{-5}$. We obtain a different $\widehat{B}$ for each experiment, and compute the sample error $L(\widehat{B}) - L(B^\star)$, approximating the expected loss with the empirical average on a large test set. Finally, for each value of m we average the sample error across the K repeated experiments, and illustrate the dependence of the resulting quantity on m in Figure 1.

As shown in Figure 1, as the sample size increases, the sample error (blue line) decreases faster than the theoretical bound (30) (the orange line when $s \to \infty$). This behavior demonstrates the effectiveness of the approach proposed in this work. The improvement of the decay rate in the numerical experiments might be the result of many factors, including the finite-dimensional (and simple) nature of the considered problem, or some specific advantages induced by the use of the

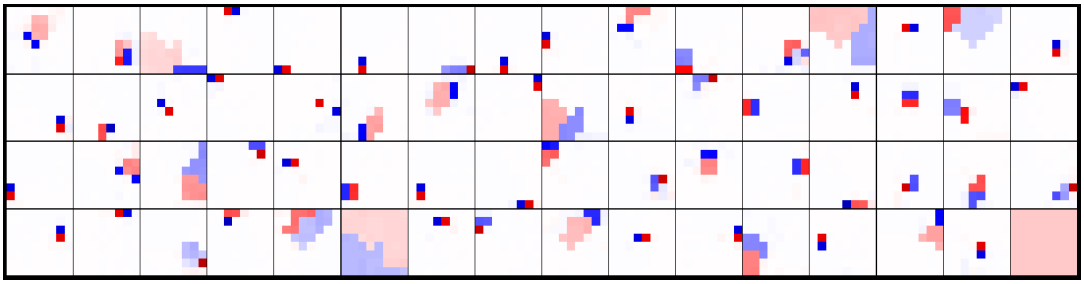

Figure 2: Learned sparsifying transform. Each column of B is represented as an 8×8 patch.

quadratic MSE loss. This result opens the way for further theoretical investigations, possibly in the spirit of [20, Theorem C*].

Example 2: a 2D-denoising problem In this experiment, we consider the task of denoising images $y_j = x_j + \varepsilon_j$ synthetically generated by adding independent noise realizations $\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I)$ (with $\sigma = 0.1$) to images of size 128×128 containing ellipses with random positions, aspect ratios, orientations and intensities. To reduce the computational burden of learning an operator $B \in \mathbb{R}^{n \times n}$ with $n = 128^2$, we consider a patch-based denoiser. Namely, from each image we extract non-overlapping patches of size $p \times p$ (with $p = 8$), and apply our supervised strategy to minimize $\mathcal{L}_1$ on the resulting dataset of patches, thus learning an operator $B \in \mathbb{R}^{p^2 \times p^2}$. We consider a training dataset of $m = 1000$ images (i.e., 256000 patches), computing a maximum of `n_epochs` $= 20000$ iterations of Adam with a stepsize `learning_rate` $= 10^{-3}$ and choosing $\mu_{\text{reg}} = 10^{-3}$. The denoised patches are then recollected to form the denoised images. In Figure 2 we report the learned operator Q, being $B = \alpha Q$. To ease the visualization, we represent each column of Q (i.e., each element of the learned basis) not as a vector in $\mathbb{R}^{p^2}$ but as a patch in $\mathbb{R}^{p \times p}$. The learned basis resembles a wavelet one, reflecting the cartoon-like nature of the ellipses' images.

Since in this case we do not have an insight into an optimal $B^\star$ to use as a benchmark, we compare our strategy with Dictionary Learning (see Appendix B for a theoretical comparison). In particular, employing the online algorithm developed in [39], we construct a sparsifying dictionary D_{DL} leveraging the dataset of ground truth patches extracted from $\{x_j\}_{i=1}^m$ and later select $B = \alpha D_{\text{DL}}$. Here, α controls the intensity of the regularization depending on the noise level and is selected by a line-search technique, minimizing the empirical risk of the denoising task. Compared with our strategy, this Dictionary-Learning-based technique requires to specify two parameters: in addition to α, indeed, we must also select the internal regularization parameter μ_{DL} of the online algorithm of [39] (which controls the level of desired sparsity induced by the dictionary D_{DL}). Tuning μ_{DL}, which can be done again via line-search, is computationally expensive and not straightforward, as the Dictionary Learning stage is not aware of the physics of the inverse problems we aim at solving, but only of the properties of the ground-truth images. Instead, our technique does not require tuning any parameters, learns simultaneously the basis and the regularization parameter, and achieves slightly better performance than the one associated with the best choice of parameters in Dictionary Learning (in this example, $\mu_{\text{DL}} = 0.06$ and $\alpha = 0.1537$). The results are reported in Table 1 and in Figure 3.

	Noisy data	Dictionary Learning	Our Method
MSE	1.0001E-02	1.78891E-03	**1.6403E-03**

Table 1: Comparison of the average MSE on the test set between the ground truths and the original noisy data, the reconstructions given by Dictionary Learning, and by our supervised strategy.

Example 3: a 1D-deblurring problem We now consider the same 1D signals $\{x_j\}_{j=1}^m$ as in Example 1, but generate the measurements $\{y_j\}_{j=1}^m$ according to a different model: namely,

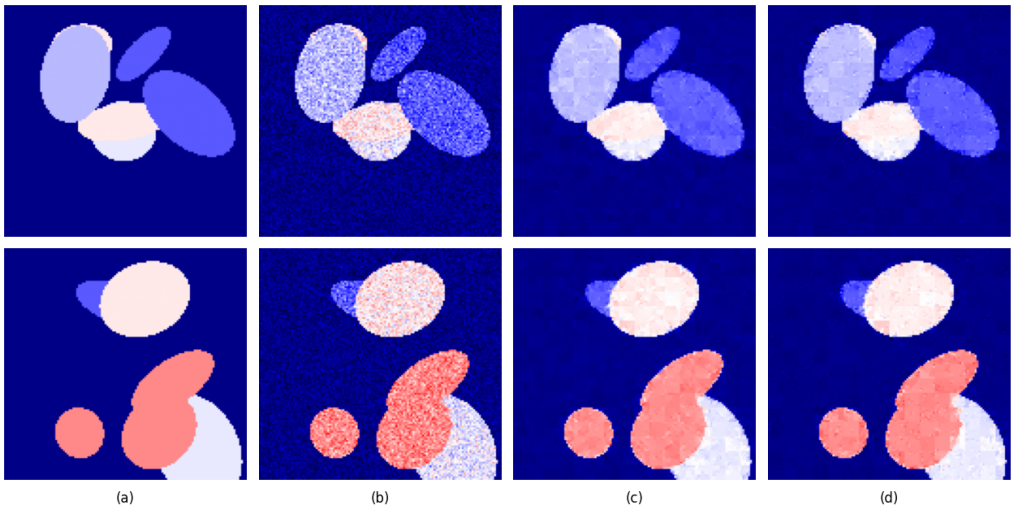

Figure 3: Visual comparison between some reconstructions of the images of the test set. (a): ground truths; (b): noisy images; (c): denoised images via Dictionary Learning; (d): denoised images via our strategy.

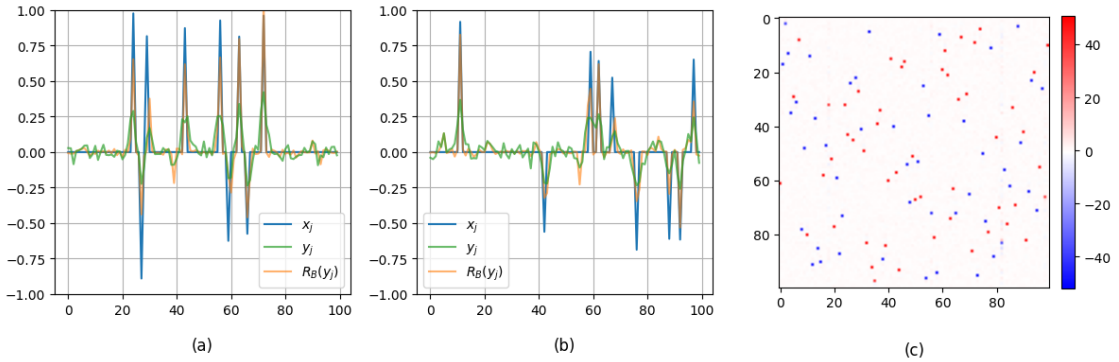

Figure 4: Results of the proposed technique for the 1D deblurring problem. Columns (a) and (b): comparison between the ground truth signals x_j, their blurred and noisy versions y_j, and their reconstructions $R_B(y_j)$ for two different examples. Column (c): the learned matrix $\widehat{B}$.

$y_j = k * x_j + \varepsilon_j$, being $*$ the 1D discrete convolution operation with reflecting boundary conditions, k a Gaussian filter of standard deviation $\sigma_{\text{blur}} = 1.2$, and $\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I)$ i.i.d. with $\sigma = 0.05$. This is a deblurring inverse problem, where the forward operator $A \in \mathbb{R}^{n \times n}$ is the circulant matrix associated with the filter k. We apply the relaxation-based algorithm proposed in Section 5.1, choosing a small relaxation parameter, namely $\nu = 10^{-3}$. The size of the training set is $m = 20000$. We employ the Adam algorithm, initialized with a random orthogonal matrix, and leverage the explicit expression of the gradient of the loss, specifying a maximum number `n_epochs` $= 5000$ of iterations, a stepsize `learning_rate` $= 1$, and $\mu_{\text{reg}} = 0.03$. The results of the training process are reported in Figure 4, where we show two examples of signals from the test set, together with their blurry, noisy observations and the reconstructions via our learned operator $R_{\widehat{B}}$. In the same figure, we report the learned matrix $\widehat{B}$: as expected, the learned basis $\widehat{B}$ is just a scaled version of the canonical basis (the identity matrix), but with its columns randomly shuffled (permuted) and flipped (sign swaps). This is the natural result because the signals x_j were designed to be sparse in that exact basis. The learned scaling factor simply adjusts the effective strength of the ℓ^1 regularization.

As in Example 1, we look for a benchmark B based on the ansatz $B^\star = \alpha^\star I$, selecting $\alpha^\star$ via

line-search. In Table 2, we compare the values of $L(\widehat{B})$ and $L(B^\star)$ (approximated on a sufficiently large test set). We observe that our technique achieves a comparable accuracy without requiring any prior information on the basis with respect to which the signals are sparse.

	Noisy, blurry data	Optimal choice, $B^\star$	Our Method, $\widehat{B}$
MSE	3.4108E-02	**1.7026E-02**	1.7312E-02

Table 2: Comparison of the average MSE on the test set between the ground truths and the original noisy and blurry data, the reconstruction given by $B^\star$, and our supervised strategy.

Appendix A White noise

As we mentioned in Section 1, the assumption on Σ_ε does not allow us, in principle, to consider white noise in infinite dimensions, for which the covariance $\Sigma_\varepsilon = \mathrm{Id}$ is not trace-class, but only bounded. However, this situation can be considered by embedding Y into a larger space. Instead of providing a general and abstract construction, we prefer to discuss two particular examples. For more details, see [5, Section A.1].

Example A.1. *Let Ω be a connected bounded open subset of $\mathbb{R}^d$, and $\{\varepsilon_y\}_{y\in L^2(\Omega)}$ be a Hilbert random process such that*

$$\mathbb{E}[\varepsilon_y] = 0 \qquad \mathbb{E}[\varepsilon_y \varepsilon_{y'}] = \langle y, y'\rangle_{L^2(\Omega)}, \qquad y, y' \in L^2(\Omega),$$

which is the classical model for white noise, since the covariance operator of ε is the identity. Assume that Ω satisfies a cone condition and fix $s > d/2$. Then the Rellich-Kondrachov theorem [2, Theorem 6.3] implies that the Sobolev space $H^s(\Omega)$ is embedded into $C_b(\Omega)$, the space of bounded continuous functions on Ω, and the canonical inclusion $\iota\colon H^s(\Omega) \to L^2(\Omega)$ is a compact operator. The above embedding implies that $H^s(\Omega)$ is a reproducing kernel Hilbert space with bounded reproducing kernel $K\colon \Omega\times\Omega \to \mathbb{R}$. An easy calculation shows that

$$\mathrm{tr}(\iota^* \circ \iota) = \int_U K(x,x)dx < +\infty.$$

This fact implies that in the Gelfand triple

$$H^s(\Omega) \xrightarrow{\iota} L^2(\Omega) \xrightarrow{\iota^*} H^{-s}(\Omega), \tag{40}$$

both ι and ι^* are Hilbert-Schmidt operators. Hence, it is possible to define a random variable $\widehat{\varepsilon}$ taking values in $H^{-s}(\Omega)$ such that

$$\langle \widehat{\varepsilon}, y\rangle_{H^{-s},H^s} = \varepsilon_{\iota(y)}, \qquad y \in H^s(\Omega).$$

It is immediate to check that

$$\mathbb{E}[\langle \widehat{\varepsilon}, y\rangle_{H^{-s},H^s}\ \langle \widehat{\varepsilon}, y'\rangle_{H^{-s},H^s}] = \langle \iota(y), \iota(y')\rangle_{L^2(\Omega)},$$

so that the covariance operator of $\widehat{\varepsilon}$ is $\iota^*\circ\iota$, which is a trace class operator. Thus, $\widehat{\varepsilon}$ is a square-integrable random variable in $H^{-s}(\Omega)$. To apply the results of our paper, it is enough to set $Y = H^{-s}(\Omega)$ and to lift the inverse problem (1) to $H^{-s}(\Omega)$:

$$\iota^*(y) = (\iota^* \circ A)x + \widehat{\varepsilon}.$$

This requires identifying $H^s(\Omega)$ and $H^{-s}(\Omega)$ using the Riesz lemma. Note that this identification is not standard, since it is not compatible with the double embedding of (40). However, the intermediate space $L^2(\Omega)$ does not matter once $\widehat{\varepsilon}$ is defined.

In the next example, we consider the sequence space ℓ^2, as a prototypical Hilbert space (once an orthonormal basis has been fixed).

Example A.2. *Let X be a Hilbert space and set $Y = \ell^2$. Let $A\colon X \to \ell^2$ be a bounded and linear map, and consider the inverse problem* (1)

$$y = Ax + \varepsilon. \tag{41}$$

Let $\{e_n\}_{n\in\mathbb{N}}$ be the canonical basis of ℓ^2, defined by $(e_n)_i = \delta_{i,n}$. We consider the case when ε is a white Gaussian noise with variance σ^2, namely,

$$\varepsilon = \sum_{n\in\mathbb{N}} \varepsilon_n e_n,$$

where the random variables ε_n are i.i.d. scalar Gaussians with mean zero and variance σ^2, i.e. $\varepsilon_n \sim \mathcal{N}(0,\sigma^2)$. The above expression for ε is only formal, because the series is divergent in ℓ^2 with probability 1. As a consequence, (41) *is not well-defined in ℓ^2. However, writing* (41) *in components with respect to the orthonormal basis $\{e_n\}_{n\in\mathbb{N}}$ yields*

$$y_n = (Ax)_n + \varepsilon_n, \qquad n \in \mathbb{N}, \tag{42}$$

where we wrote $Ax = \sum_n (Ax)_n e_n$. This is a family of well-defined scalar equations.

Let us now see how it is possible to reformulate this as a problem in a Hilbert space. Equivalently, we can rewrite (42) *as*

$$\frac{y_n}{n^s} = \frac{(Ax)_n}{n^s} + \frac{\varepsilon_n}{n^s}, \qquad n \in \mathbb{N}, \tag{43}$$

for some $s > \frac{1}{2}$. Let us introduce the embedding $\iota^*\colon \ell^2 \to \ell^2$ defined by $\iota^*(e_n) = e_n/n^s$ and the random variable

$$\widehat{\varepsilon} = \sum_{n\in\mathbb{N}} \frac{\varepsilon_n}{n^s} e_n.$$

Note that $\mathbb{E}[\|\widehat{\varepsilon}\|_2^2] < +\infty$, and $\widehat{\varepsilon} \in \ell^2$ with probability 1. Thus, we can rewrite (43) *as*

$$\iota^*(y) = \iota^*(Ax) + \widehat{\varepsilon}.$$

This equation is meaningful in ℓ^2, and has the same form as the original inverse problem (41).

Appendix B Connections with Dictionary Learning and unsupervised strategies

Although our approach shares the same aim of dictionary learning, i.e., promoting the sparse representation of some ground truths by selecting a suitable synthesis operator, it is possible to outline some substantial differences. The key observation is that the optimal operator sought in dictionary learning is independent of the forward operator and the noise distribution, and depends only on the distribution of x. On the contrary, the optimal $\widehat{B}$ in (5) depends on both ε and A, yielding, in general, a smaller MSE and, consequently, better statistical guarantees for the solution of the inverse problem.

Let us briefly introduce the standard dictionary learning framework [48]. If, instead of the training dataset $\mathbf{z} = \{(x_j, y_j)\}_{j=1}^m$ employed in (7) to discretize (5), only a collection of ground truths $\{x_j\}_{j=1}^m$ is available, i.i.d. sampled from the (unknown) marginal ρ_x, we may consider the following unsupervised technique: let $\widetilde{R}_B$ a sparsity-promoting regularizer of the form (2)-(3) associated with $A = \mathrm{Id}$ and without assuming the knowledge of the covariance Σ_ε, namely:

$$\widetilde{R}_B(x) = B\widetilde{u}_B(x), \quad \widetilde{u}_B(x) = \operatorname*{arg\,min}_{u\in\ell^1} \left\{ \frac{1}{2}\|Bu - x\|^2 + \|u\|_{\ell^1} \right\}$$

$$\widehat{B}_{DL} \in \operatorname*{arg\,min}_{B\in\mathcal{B}} \left\{ \frac{1}{m}\sum_{j=1}^m \|x_j - \widetilde{R}_B(x_j)\|^2 \right\}.$$

This problem yields a bilevel formulation of the well-known Dictionary Learning problem [43, 50]. Our supervised strategy resembles dictionary learning, indeed, let us recall that, according to (7),

$$\widehat{B} \in \operatorname*{arg\,min}_{B\in\mathcal{B}} \left\{ \frac{1}{m}\sum_{j=1}^{m} \|x_j - R_B(Ax_j+\varepsilon_j)\|^2 \right\},$$

being R_B as in (3). However, since R_B is a nonlinear map, differently from our previous work on quadratic regularizers [5], it is easy to show that the two problems are not equivalent in general. In particular, while $\widehat{B}_{DL}$ is independent of the forward operator A and on the noise ε by construction, $\widehat{B}$ will in general depend on both. We illustrate this with a simple 1D example.

Let $\sigma^2 = \mathbb{E}(\varepsilon^2) > 0$ denote the variance of the noise (with zero mean) and consider the 1D problem

$$Ax = ax \quad \text{with } a \in (0,\sigma]$$

and the regularization $B^{-1}x = bx$, $b > 0$. Given an unknown $x^\dagger$ and data $y = Ax^\dagger + \varepsilon = ax^\dagger + \varepsilon$, the Lasso reconstruction is given by

$$\hat{x} = \operatorname*{arg\,min}_x \frac{1}{2}|\sigma^{-1}(Ax-y)|^2 + |bx| = \operatorname*{arg\,min}_x \frac{1}{2}|x-y/a|^2 + \frac{\sigma^2 b}{a^2}|x|.$$

Setting $\gamma = \frac{a}{\sigma}$, this may be rewritten by using the soft-thresholding operator S_λ as

$$\hat{x} = S_{\frac{b}{\gamma^2}}\left(x^\dagger + \gamma^{-1}\tilde{\varepsilon}\right),$$

where $\tilde{\varepsilon} = \varepsilon/\sigma$ satisfies $\mathbb{E}(\tilde{\varepsilon}^2) = 1$. Now consider the mean squared error

$$\mathrm{MSE} = \mathbb{E}_{x,\tilde{\varepsilon}}\left[\left| S_{\frac{b}{\gamma^2}}\left(x+\gamma^{-1}\tilde{\varepsilon}\right) - x\right|^2\right].$$

Note that we consider the minimization of the expected risk, and not of the empirical risk, as it is more significant. For simplicity, we choose the following zero-mean independent distributions

$$\mathbb{P}(x=\pm 1) = \frac{1}{2}, \quad \mathbb{P}(\tilde{\varepsilon}=\pm 1) = \frac{1}{2}.$$

After a series of elementary computations, we can show that

$$\mathrm{MSE} \;= \begin{cases} \frac{1}{\gamma^4}(b-\gamma)^2 & b \in \left(0, \gamma-\gamma^2\right), \\ \frac{1}{2}\left[1+\frac{1}{\gamma^4}(b-\gamma)^2\right] & b \in \left[\gamma-\gamma^2, \gamma+\gamma^2\right), \\ 1 & b \geqslant \gamma+\gamma^2. \end{cases}$$

Therefore, the optimal value for the MSE is achieved at $b = \gamma = \frac{a}{\sigma}$ and depends on both a and σ, namely, on both the forward operator and the noise level. Any unsupervised choice for b would be independent of a and σ and give rise, in general, to larger values of the MSE.

Appendix C Proofs of Section 4.1

Proof of Lemma 4.1. Every element $B \in \mathcal{B}$ satisfies Assumption 2.2 because A is injective and $B|_{\ell^2_I}$ is injective for every finite subset $I \subset \mathbb{N}$ by (35). It remains to show that $\mathcal{B}$ is compact. Since the map

$$\mathcal{L}(\ell^2) \ni U \longmapsto B_0(\mathrm{Id}+U) \in \mathcal{L}(\ell^2, X) \tag{44}$$

is continuous, it is enough to show that $\mathcal{H} \subset \mathcal{L}(\ell^2)$ is compact. Write $\mathcal{H} = \mathcal{K} \cap \mathcal{C}$, where

$$\mathcal{K} = \{E_1 T E_2 : T \in \mathcal{L}(\ell^2),\ \|T\|_{\mathcal{L}(\ell^2)} \leq 1\},$$
$$\mathcal{C} = \{K \in \mathcal{L}(\ell^2) : \|(\mathrm{Id}+K)u\| \geq c_I\|u\| \ \text{ for every finite } I \subset \mathbb{N} \text{ and } u \in \ell^2_I\}.$$

The set $\mathcal{K}$ is compact by [49, Theorem 3], because $\left\{T \in \mathcal{L}(\ell^2) : \|T\|_{\mathcal{L}(\ell^2)} \leq 1\right\}$ is closed. The set $\mathcal{C}$ is closed, because convergence in operator norm implies pointwise convergence. Hence, $\mathcal{H} = \mathcal{K} \cap \mathcal{C}$ is compact. □

Proof of Proposition 4.2. Note that

$$\mathcal{N}(\mathcal{B}, r; \|\cdot\|_{\mathcal{L}(\ell^2,X)}) \leq \mathcal{N}(\mathcal{H}, r; \|\cdot\|_{\mathcal{L}(\ell^2)}), \tag{45}$$

since the map in (44) is Lipschitz with constant 1 (recall that $\|B_0\|_{\mathcal{L}(\ell^2,X)} \leq 1$). Using the notation of the proof of Lemma 4.1, we have $\mathcal{H} = \mathcal{K} \cap \mathcal{C}$, so that $\mathcal{H} \subseteq \mathcal{K}$. Thus

$$\mathcal{N}(\mathcal{H}, r; \|\cdot\|_{\mathcal{L}(\ell^2)}) \leq \mathcal{N}(\mathcal{K}, r; \|\cdot\|_{\mathcal{L}(\ell^2)}). \tag{46}$$

To further bound the right-hand side of (46), we consider the singular value decomposition of E. Since E is self-adjoint and compact, we can write its spectral decomposition in terms of its eigenvalues $\{\lambda_n\}$ and eigenvectors $\{e_n\}$, and define its rank-N approximation E_N as follows:

$$E = \sum_{n=1}^{\infty} \lambda_n e_n \otimes e_n, \qquad E_N = \sum_{n=1}^{N} \lambda_n e_n \otimes e_n = P_N E = E P_N, \tag{47}$$

where we denote by $u \otimes v$ the rank-1 operator such that $(u \otimes v)x = \langle v, x\rangle_{\ell^2} u$, being $\langle\cdot,\cdot\rangle_{\ell^2}$ the inner product in ℓ^2, and $P_N = \sum_{n=1}^{N} e_n \otimes e_n$ is the orthogonal projection onto $\mathrm{span}\{e_1, \ldots, e_N\}$. Assuming that the sequence $\{\lambda_n\}$ is decreasingly ordered, we know that $\|E - E_N\|_{\mathcal{L}(\ell^2)} \leq \lambda_{N+1}$. We now introduce the spaces

$$\mathcal{K}_N = \left\{K = E_N T E_N : T \in \mathcal{L}(\ell^2),\ \|T\|_{\mathcal{L}(\ell^2)} \leq 1\right\},$$

and prove that the following bound holds:

$$\mathcal{N}(\mathcal{K}, \rho + 2\lambda_{N+1}; \|\cdot\|_{\mathcal{L}(\ell^2)}) \leq \mathcal{N}(\mathcal{K}_N, \rho; \|\cdot\|_{\mathcal{L}(\ell^2)}) =: \hat{\mathcal{N}}. \tag{48}$$

In order to prove (48), consider a ρ-covering $\{K_1, \ldots, K_{\hat{\mathcal{N}}}\}$ of $\mathcal{K}_N$. For any $K = ETE \in \mathcal{K}$, let $K^N = E_N T E_N \in \mathcal{K}_N$, and let K_i be such that $\|K^N - K_i\| \leq \rho$. Then,

$$\begin{aligned}
\|K - K_i\|_{\mathcal{L}} &\leq \|K - K^N\|_{\mathcal{L}} + \|K^N - K_i\|_{\mathcal{L}} \\
&\leq \|ETE - ETE_N\|_{\mathcal{L}} + \|ETE_N - E_N T E_N\|_{\mathcal{L}} + \rho \\
&\leq \|E\|_{\mathcal{L}} \|T\|_{\mathcal{L}} \|E - E_N\|_{\mathcal{L}} + \|E - E_N\|_{\mathcal{L}} \|T\|_{\mathcal{L}} \|E_N\|_{\mathcal{L}} + \rho \\
&\leq 2\lambda_{N+1} + \rho,
\end{aligned}$$

which shows that $\{K_1, \ldots, K_{\hat{\mathcal{N}}}\}$ is a $(2\lambda_{N+1} + \rho)$-covering of $\mathcal{K}$.

In order to bound the covering numbers of $\mathcal{K}_N$, we claim that $\mathcal{K}_N \subset \mathcal{F}_N$, where

$$\mathcal{F}_N = \left\{K = ETE : T \in \mathrm{HS}(\ell^2),\ \|T\|_{\mathrm{HS}(\ell^2)} \leq \sqrt{N}\right\},$$

and $\mathrm{HS}(\ell^2)$ denotes the class of Hilbert-Schmidt operators from ℓ^2 to ℓ^2, namely, the ones for which the singular values are square-summable. In order to show this, take $K \in \mathcal{K}_N$, with $K = E_N T E_N$ for some $T \in \mathcal{L}(\ell^2)$ such that $\|T\|_{\mathcal{L}(\ell^2)} \leq 1$. Setting $T_N = P_N T P_N$, by (47) we get

$$K = E_N T E_N = E P_N T P_N E = E T_N E.$$

Note that T_N is a rank-N operator, thus belonging to the Hilbert-Schmidt class. Furthermore,

$$\|T_N\|_{\mathrm{HS}(\ell^2)} = \|P_N T P_N\|_{\mathrm{HS}(\ell^2)} \leq \|P_N\|_{\mathrm{HS}(\ell^2)} \|T P_N\|_{\mathcal{L}(\ell^2)} \leq \sqrt{N},$$

because $\|P_N\|_{\mathrm{HS}(\ell^2)} = \sqrt{N}$, $\|T P_N\|_{\mathcal{L}(\ell^2)} \leq 1$ and

$$\|P_N T P_N\|^2_{\mathrm{HS}(\ell^2)} = \sum_{n=1}^{\infty} \|P_N T P_N e_n\|^2_{\ell^2} \leq \sum_{n=1}^{\infty} \|P_N T\|^2_{\mathcal{L}(\ell^2)} \|P_N e_n\|^2_{\ell^2} = \|P_N T\|^2_{\mathcal{L}(\ell^2)} \|P_N\|^2_{\mathrm{HS}(\ell^2)}.$$

This shows that $K \in \mathcal{F}_N$, as claimed. By a simple scaling argument, we have

$$\mathcal{N}(\mathcal{K}_N, \rho; \|\cdot\|_{\mathcal{L}}) \leq \mathcal{N}(\mathcal{F}_N, \rho; \|\cdot\|_{\mathcal{L}}) = \mathcal{N}(\mathcal{F}_1, \rho N^{-1/2}; \|\cdot\|_{\mathcal{L}}). \tag{49}$$

We finally have to estimate the covering numbers $\mathcal{N}(\mathcal{F}_1, \varrho; \|\cdot\|_{\mathcal{L}})$. Let us define

$$F_1 = \left\{T \in \mathrm{HS}(\ell^2) : \ \|T\|_{\mathrm{HS}(\ell^2)} \leq 1\right\},$$

which entails that $\mathcal{F}_1 = j(F_1)$, where j is the (compact) embedding $j\colon \mathrm{HS}(\ell^2) \to \mathrm{HS}(\ell^2)$ defined as $j(T) = ETE$. Since $\|\cdot\|_{\mathcal{L}(\ell^2)} \leq \|\cdot\|_{\mathrm{HS}(\ell^2)}$, a $\varrho-$covering of $\mathcal{F}_1$ with respect to $\|\cdot\|_{\mathrm{HS}(\ell^2)}$ is also a ϱ-covering with respect to $\|\cdot\|_{\mathcal{L}(\ell^2)}$. Thus,

$$\mathcal{N}(\mathcal{F}_1, \varrho; \|\cdot\|_{\mathcal{L}(\ell^2)}) \leq \mathcal{N}(\mathcal{F}_1, \varrho; \|\cdot\|_{\mathrm{HS}(\ell^2)}). \tag{50}$$

The quantity $\mathcal{N}(\mathcal{F}_1, \varrho; \|\cdot\|_{\mathrm{HS}(\ell^2)})$ is the covering number of the image of the unit sphere F_1 of the Hilbert space $\mathrm{HS}(\ell^2)$ through the embedding j, and is linked to the *entropy numbers* $\varepsilon_k(j)$: indeed, according to the definitions in [18, Chapter 1], one clearly sees that

$$\mathcal{N}(\mathcal{F}_1, \varrho; \|\cdot\|_{\mathrm{HS}(\ell^2)}) \leq k \quad \iff \quad \varepsilon_k(j) \leq \varrho.$$

In order to estimate the entropy numbers of j, we can rely on its singular values, thanks to [18, Theorem 3.4.2]. In view of the decay $\lambda_n \lesssim n^{-s}$ of the eigenvalues of E, following the proof of [5, Lemma A.9] we can show that the singular values of j decay as $n^{-s'}$, for any $s' < s$. Then, by the same argument used in the proof of [5, Lemma A.8], we have that $\varepsilon_k(j) \lesssim (\log k)^{-s'}$, which implies $\mathcal{N}(\mathcal{F}_1, (\log k)^{-s'}; \|\cdot\|_{\mathrm{HS}(\ell^2)}) \leq k$ and ultimately

$$\log \mathcal{N}(\mathcal{F}_1, \varrho; \|\cdot\|_{\mathrm{HS}(\ell^2)}) \lesssim \varrho^{-1/s'}. \tag{51}$$

We can now conclude the proof: by (45), (46) and (48), setting $\rho = \frac{r}{2}$ and choosing N such that $\lambda_{N+1} = \frac{r}{4}$ (which implies $(N+1)^{-s} \gtrsim \frac{r}{4}$, hence $N \lesssim r^{-1/s}$), we get

$$\mathcal{N}(\mathcal{B}, r; \|\cdot\|_{\mathcal{L}(\ell^2, X)}) \lesssim \mathcal{N}(\mathcal{K}_N, r; \|\cdot\|_{\mathcal{L}(\ell^2)}).$$

Next, by (49), we obtain

$$\mathcal{N}(\mathcal{B}, r; \|\cdot\|_{\mathcal{L}(\ell^2, X)}) \lesssim \mathcal{N}(\mathcal{F}_1, rN^{-1/2}; \|\cdot\|_{\mathcal{L}(\ell^2)}),$$

and by (50) and (51) we finally obtain

$$\log \mathcal{N}(\mathcal{B}, r; \|\cdot\|_{\mathcal{L}(\ell^2, X)}) \lesssim \left(rN^{-1/2}\right)^{-1/s'} \lesssim r^{-\frac{2s+1}{2ss'}}.$$

This shows (36), and the proof is concluded. □

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award FA8655-23-1-7083 and was cofunded by the European Union (ERC, SAMPDE, 101041040, ERC, SLING, 819789, ERC, AdG project 101097198, and Next Generation EU - project FAIR, PE0000013, spoke 8). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. GSA, EDV, LR and MS are members of INdAM (Istituto Nazionale di Alta Matematica). TH and ML were supported by the Research Council of Finland, decision numbers 348504, 353094, 359182 and 359183 (FAME flagship).

References

[1] Abhishake, T. Helin, and N. Mücke, Statistical inverse learning problems with random observations random observations, Data-driven Models in Inverse Problems, 31 (2024), p. 201.

[2] R. A. Adams and J. J. Fournier, Sobolev spaces, Elsevier, 2003.

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

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

[5] G. S. Alberti, E. De Vito, M. Lassas, L. Ratti, and M. Santacesaria, Learning the optimal Tikhonov regularizer for inverse problems, Advances in Neural Information Processing Systems, 34 (2021).

[6] G. S. Alberti, L. Ratti, M. Santacesaria, and S. Sciutto, Learning a Gaussian mixture for sparsity regularization in inverse problems, IMA Journal of Numerical Analysis, (2025), p. draf037, https://doi.org/10.1093/imanum/draf037, https://doi.org/10.1093/imanum/draf037, https://arxiv.org/abs/https://academic.oup.com/imajna/advance-article-pdf/doi/10.1093/imanum/draf037/63442712/draf037.pdf.

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

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

[9] D. Bosq, Linear processes in function spaces: theory and applications, vol. 149, Springer Science & Business Media, 2000.

[10] N. Boullé and A. Townsend, A mathematical guide to operator learning, in Numerical Analysis Meets Machine Learning, S. Mishra and A. Townsend, eds., vol. 25 of Handbook of Numerical Analysis, Elsevier, 2024, pp. 83–125, https://doi.org/https://doi.org/10.1016/bs.hna.2024.05.003, https://www.sciencedirect.com/science/article/pii/S1570865924000036.

[11] C. Brauer, N. Breustedt, T. De Wolff, and D. A. Lorenz, Learning variational models with unrolling and bilevel optimization, Analysis and Applications, 22 (2024), pp. 569–617.

[12] K. Bredies and D. A. Lorenz, Linear convergence of iterative soft-thresholding, Journal of Fourier Analysis and Applications, 14 (2008), pp. 813–837.

[13] H. Brezis and H. Brézis, Functional analysis, Sobolev spaces and partial differential equations, vol. 2, Springer, 2011.

[14] T. Bubba, M. Burger, T. Helin, and L. Ratti, Convex regularization in statistical inverse learning problems, Inverse Problems and Imaging, 17 (2023), pp. 1193–1225.

[15] M. Burger and S. Kabri, Learned regularization for inverse problems, De Gruyter, Berlin, Boston, 2025, pp. 39–72, https://doi.org/doi:10.1515/9783111251233-002, https://doi.org/10.1515/9783111251233-002.

[16] L. Calatroni, C. Chung, J. C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, Bilevel approaches for learning of variational imaging models, Variational Methods: In Imaging and Geometric Control, 18 (2017), p. 2.

[17] E. J. Candes and T. Tao, Decoding by linear programming, IEEE transactions on information theory, 51 (2005), pp. 4203–4215.

[18] B. Carl and I. Stephani, Entropy, compactness and the approximation of operators, vol. 98 of Cambridge Tracts in Mathematics, Cambridge University Press, Cambridge, 1990, https://doi.org/10.1017/CBO9780511897467, https://doi.org/10.1017/CBO9780511897467.

[19] J. Chirinos-Rodríguez, E. De Vito, C. Molinari, L. Rosasco, and S. Villa, On learning the optimal regularization parameter in inverse problems, Inverse Problems, 40 (2024), p. 125004.

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

[21] I. Daubechies, Ten lectures on wavelets, vol. 61 of CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992, https://doi.org/10.1137/1.9781611970104, https://doi.org/10.1137/1.9781611970104.

[22] M. V. de Hoop, N. B. Kovachki, N. H. Nelsen, and A. M. Stuart, Convergence rates for learning linear operators from noisy data, SIAM/ASA Journal on Uncertainty Quantification, 11 (2023), pp. 480–513.

[23] J. C. De los Reyes, C.-B. Schönlieb, and T. Valkonen, Bilevel parameter learning for higher-order total variation regularisation models, Journal of Mathematical Imaging and Vision, 57 (2017), pp. 1–25.

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

[25] G. Frusque and O. Fink, Learnable wavelet packet transform for data-adapted spectrograms, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 3119–3123.

[26] A. Ghosh, M. McCann, M. Mitchell, and S. Ravishankar, Learning sparsity-promoting regularizers using bilevel optimization, SIAM Journal on Imaging Sciences, 17 (2024), pp. 31–60, https://doi.org/10.1137/22M1506547, https://doi.org/10.1137/22M1506547.

[27] A. Ghosh, M. T. Mccann, and S. Ravishankar, Bilevel learning of ℓ^1 regularizers with closed-form gradients (blorc), in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 1491–1495, https://doi.org/10.1109/ICASSP43922.2022.9747201.

[28] M. Grasmair, O. Scherzer, and M. Haltmeier, Necessary and sufficient conditions for linear convergence of ℓ1-regularization, Communications on Pure and Applied Mathematics, 64 (2011), pp. 161–182.

[29] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, Sample complexity of dictionary learning and other matrix factorizations, IEEE Transactions on Information Theory, 61 (2015), pp. 3469–3486, https://doi.org/10.1109/TIT.2015.2424238.

[30] A. Hauptmann, S. Mukherjee, C.-B. Schönlieb, and F. Sherry, Convergent regularization in inverse problems and linear plug-and-play denoisers, Foundations of Computational Mathematics, (2024), pp. 1–34.

[31] L. Horesh and E. Haber, Sensitivity computation of the l_1 minimization problem and its application to dictionary design of ill-posed problems, Inverse Problems, 25 (2009), pp. 095009, 20, https://doi.org/10.1088/0266-5611/25/9/095009, https://doi.org/10.1088/0266-5611/25/9/095009.

[32] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ_1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 6 (2012), pp. 447–464, https://doi.org/10.3934/ipi.2012.6.447, https://doi.org/10.3934/ipi.2012.6.447.

[33] Z. Jing, On the stability of wavelet and Gabor frames (Riesz bases), J. Fourier Anal. Appl., 5 (1999), pp. 105–125, https://doi.org/10.1007/BF01274192.

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

[35] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, Neural operator: Learning maps between function spaces with applications to pdes, Journal of Machine Learning Research, 24 (2023), pp. 1–97.

[36] S. Lanthaler, S. Mishra, and G. E. Karniadakis, Error estimates for DeepONets: a deep learning framework in infinite dimensions, Transactions of Mathematics and Its Applications, 6 (2022), p. tnac001.

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

[38] 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, 2018, pp. 8516–8525.

[39] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, Online dictionary learning for sparse coding, in Proceedings of the 26th annual international conference on machine learning, 2009, pp. 689–696.

[40] 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).

[41] S. Mukherjee, O. Öktem, and C.-B. Schönlieb, Adversarially learned iterative reconstruction for imaging inverse problems, in Scale Space and Variational Methods in Computer Vision, A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, and L. Simon, eds., Cham, 2021, Springer International Publishing, pp. 540–552.

[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 (2021), pp. A3212–A3243.

[43] G. Peyré and J. M. Fadili, Learning analysis sparsity priors, in International Conference on Sampling Theory and Applications, 2011.

[44] L. Ratti, Learned reconstruction methods for inverse problems: sample error estimates, De Gruyter, Berlin, Boston, 2025, pp. 163–200, https://doi.org/10.1515/9783111251233-005.

[45] D. Recoskie and R. Mann, Learning filters for the 2d wavelet transform, in 2018 15th Conference on Computer and Robot Vision (CRV), IEEE, 2018, pp. 198–205.

[46] D. Recoskie and R. Mann, Learning sparse wavelet representations, arXiv preprint arXiv:1802.02961, (2018).

[47] J. Sulam, C. You, and Z. Zhu, Recovery and generalization in over-realized dictionary learning, Journal of Machine Learning Research, 23 (2022), pp. 1–23, http://jmlr.org/papers/v23/20-1360.html.

[48] I. Tošić and P. Frossard, Dictionary learning, IEEE Signal Processing Magazine, 28 (2011), pp. 27–38, https://doi.org/10.1109/MSP.2010.939537.

[49] K. Vala, Compact set of compact operators, Ann. Acad. Sci. Fenn. Ser. A I, 351 (1964), pp. 1–9.

[50] J. Yang, Z. Wang, Z. Lin, X. Shu, and T. Huang, Bilevel sparse coding for coupled feature spaces, in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2360–2367, https://doi.org/10.1109/CVPR.2012.6247948.