CONVEX REGULARIZATION IN STATISTICAL INVERSE LEARNING PROBLEMS

Published Version:
CONVEX REGULARIZATION IN STATISTICAL INVERSE LEARNING PROBLEMS / Bubba T.A.; Burger M.; Helin T.; Ratti L.. - In: INVERSE PROBLEMS AND IMAGING. - ISSN 1930-8337. - ELETTRONICO. - 17:6(2023), pp. 1193-1225. [10.3934/ipi.2023013]

Availability:
This version is available at: https://hdl.handle.net/11585/960680 since: 2024-02-28

Published:
DOI: http://doi.org/10.3934/ipi.2023013

Convex regularization in statistical inverse learning problems

Tatiana A. Bubba[1], Martin Burger[2,3], Tapio Helin[4] and Luca Ratti[5*†]

[1]Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, Bath, UK.
[2] Computational Imaging Group and Helmholtz Imaging, Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, 22607 Hamburg, Germany.
[3] Department of Mathematics, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany.
[4] School of Engineering Science, Lappeenranta-Lahti University of Technology, Yliopistokatu 34, 53850 Lappeenranta, Finland.
[5] Department of Mathematics, University of Bologna, Piazza di Porta S. Donato, 5, 40126 Bologna, Italy.

Abstract

We consider a statistical inverse learning problem, where the task is to estimate a function f based on noisy point evaluations of Af, where A is a linear operator. The function Af is evaluated at i.i.d. random design points u_n, $n = 1, ..., N$ generated by an unknown general probability distribution. We consider Tikhonov regularization with general convex and p-homogeneous penalty functionals and derive concentration rates of the regularized solution to the ground truth measured in the symmetric Bregman distance induced by the penalty functional. We derive concrete rates for Besov norm penalties and numerically demonstrate the correspondence with the observed rates in the context of X-ray tomography.

Keywords: Variational regularization; statistical learning; error estimates; Bregman distances; computed tomography.

AMS Subject Classification: 62G08, 62G20, 65J22, 68Q32.

1 Introduction

Inverse problems study how indirect observational data can be processed into information about objects of interest in a robust manner. The literature of inverse problems often adopts the perspective that the observational process can be designed or controlled to a sufficient degree. However, for many large-scale inverse problems in modern science and engineering massive data sets arise from poorly controllable experimental conditions. Such problems are closely connected to statistical learning setting, where the objective is to approximate a function $g : U \to V$ through a set of pairs $(u_n, v_n)_{n=1}^N$ drawn from an unknown probability measure on $U \times V$.

*e-mails: tab73@bath.ac.uk, martin.burger@fau.de, tapio.helin@lut.fi, luca.ratti5@unibo.it

†The first author did most of this work while at the Department of Mathematics and Statistics, University of Helsinki, Finland. The fourth author did most of this work while at the Machine Learning Genoa Center, Department of Mathematics, University of Genoa, Italy.

The framework, where observational data is limited to a finite set of random and noisy point evaluations of the output, has also a tradition in inverse problems [25, 4]. In particular, statistical inverse learning problems have recently gained attention and we give an overview below. Our interest lies in deriving convergence rates, for general regularization schemes, of the expected reconstruction error, namely, the distance (in a suitable metric) between the solutions of the inverse problem and the regularized one. In this regard, the state of the art was recently improved by Blanchard and Mücke [5], who derive minimax optimal convergence rates for the general spectral regularization approach in Hilbert spaces under certain classes of sampling measure.

This paper aims at blending inverse learning theory together with recent developments in convex regularization techniques in the context of inverse problems [3, 8]. Although there is a body of work studying methods such as Lasso and generalized approaches in learning theory (see, e.g., [17]), to the best our knowledge, general convex regularization has not been considered before for inverse statistical learning problems. Here, we focus on variational regularization schemes utilizing p-homogeneous penalties, in particular, focusing on the case $1 < p \leq 2$, and derive a framework for establishing convergence rates in expected symmetric Bregman distance. Our work is aligned with the common assumption in learning theory that the design measure, i.e., the probability distribution generating the evaluation points $(u_n)_{n=1}^N$, is unknown.

Let us consider a linear inverse problem

$$g = Af, \tag{1}$$

where $A : X \to Y$ is a bounded linear operator between a separable Banach space X and a Hilbert space Y. Furthermore, we assume that Y is a function space from a compact subset $U \subset \mathbb{R}^d$ to a Hilbert space V. We observe noisy point evaluations of g at given points $\{u_i\}_{i=1}^N \subset U$ according to

$$g_i^\delta = g^\dagger(u_i) + \delta\epsilon_i \tag{2}$$

for $i = 1, ..., N$, where ϵ_i are i.i.d. and have suitable distribution. Moreover, the noiseless observation $g^\dagger = Af^\dagger$ corresponds to our ground truth $f^\dagger \in X$. In the following we study properties of a regularized solutions $f_{\alpha,N}^\delta$ defined via the variational problem

$$f_{\alpha,N}^\delta := \arg\min_{f\in X} \left\{ \frac{1}{2N} \sum_{i=1}^N \left\| (Af)(u_i) - g_i^\delta \right\|_V^2 + \alpha R(f) \right\}, \tag{3}$$

where $R : X \to \mathbb{R} \cup \{\infty\}$ is a convex functional satisfying certain technical properties listed below (see Assumption 2.2).

Our main contributions in the case of a p-homogeneous R for $1 < p \leq 2$ are as follows:

- In Theorem 4.3 we derive an upper bound to the reconstruction error, i.e., the distance between $f_{\alpha,N}^\delta$ and $f^\dagger$ measured in the symmetric Bregman distance induced by R. This upper bound is composed of terms that generalize the approximation and sample error terms observed in the spectral regularization setting in Hilbert spaces.

- In addition to the standard framework usually developed for a fixed noise level δ, we discuss an interesting regime where the noise level is small compared to the number of design points, i.e., $\delta \simeq N^{-\rho}$, $\rho > 1$. Such a setup requires modified estimates and the corresponding analysis is developed in parallel to the standard framework. We conjecture that such alternative estimates can improve the standard estimates under specific sampling schemes discussed in Remark 4.12.

- We derive convergence rates for a penalty $R(f) = \frac{1}{p} \|f\|_X^p$ under suitable assumptions in the standard framework in Theorem 4.10 and for the small noise regime in Theorem 4.11. In terms of optimality, we compare our rates with the minimax-optimal rates in the Hilbert space setting for $p = 2$ obtained in [5]. Restricted to Hilbert spaces and classical Tikhonov regularization scheme, our method yields convergence rates that coincide with [5] only under restrictive assumptions. However, the underlying discrepancy between [5] and our approach is highlighted, and we propose a modification to the method that can provide improved rates.

- As an application of our theory, we prove a concrete convergence rate for the above case when X is a Besov space $B^s_{pp}(\mathbb{R}^d)$ in Section 5.1. Moreover, in Section 5.2 we discuss how to derive convergence rates if a continuous embedding of the Banach space X to some Hilbert space X_0 is available. If the embedding has suitable approximation properties, this approach can provide useful convergence rates.

- We study the classical inverse problem of X-ray tomography [24] under random sampling of the imaging angles and using Besov space penalties. We demonstrate that the Radon transform has suitable spectral properties making our work aligned with the optimal rates in the $p = 2$ setup. Moreover, we observe the convergence rates predicted by our results in numerical simulations also for $1 < p \leq 2$.

We emphasize that we do not prove minimax-optimality of our results and, in particular, no lower bounds are derived. However, we point out that similar techniques can lead to minimax-optimal rates in inverse problems when considered against suitable source conditions [35, 6].

Our methodology has close connections to the reproducing kernel methods, which is a popular field with a vast body of literature. Let us note that connections of kernel regression methods to regularization theory were first studied in [13, 34, 20] and the line of research has since become widely popular. Early work on upper rates of convergence in a reproducing kernel Hilbert space was carried out by Smale and Cucker in [11], where they utilized a covering number technique. After the initial success, there has been a long line of subsequent work [34, 29, 30, 2, 36, 9] providing convergence rates comparable to [5]. Let us also point out that there is an avenue of research [22, 32] considering penalties of type $R(f) = \frac{1}{p} \|f\|_X^p$. Notice that the notion of convergence in the usual learning context and the inverse problem setting is different and are not directly comparable: in learning theory the convergence rates are derived in $L^2(\mu)$ norm, where μ is the unknown sampling measure generating data points. However, since the solution and data space, i.e. X and Y, differ for the inverse problem, it is natural to consider modes of convergence in X. For a related discussion and brief overview on relevant convergence rate literature, see [23].

In terms of inverse learning problems, we mention that Tikhonov regularization of non-linear inverse problems is considered in [27] and adaptive parameter choice rules are studied in [21]. Moreover, for distributed learning of inverse problems, see [16] and references therein.

This paper is organized as follows. In Section 2 we provide preliminaries of the mathematical setting and assumptions in our work. Our assumptions on the sampling setup are closely aligned with previous literature such as [5]. In Section 3 we derive general bounds on the symmetric Bregman distance between $f^\delta_{\alpha,N}$ and $f^\dagger$. In Section 4 we develop these bounds further in the case $R(f) = \frac{1}{p}\|f\|_X^p$, which enables the use of Lemma 4.1 - a key ingredient of the convergence analysis. Section 5.1 discusses the Besov space regularization and derives concetration bounds. Numerical simulations of random angle X-ray tomography are presented in Section 6: the Radon transform is introduced in Section 6.1 and discretized formulation is specified in Section 6.2. In Section 6.4 we provide the numerical experiments related to the convergence rates.

2 Preliminaries

Let us briefly define some notation. For two functions $f, g : X \to \mathbb{R}$, we write $f \lesssim g$ if there exists a universal constant $C > 0$ such that $f(x) \leq Cg(x)$ for all $x \in X$. Similarly, we write $f \simeq g$ if it holds that $g \lesssim f \lesssim g$. Below, $C > 0$ will denote a generic constant unless otherwise specified in the context.

2.1 Sampling operator

Consider $Y = L^2(U; V)$, the Bochner space of functions defined on U and taking values in the Hilbert space V, where $U \subset \mathbb{R}^d$ is compact, i.e., the functions $g\colon U \to V$ such that

$$\int_U \|g(u)\|_V^2 du < \infty.$$

Here, we assume that the range of the operator A is contained in $Z = \mathcal{C}(U; V)$, the space of continuous functions on U taking values on V, equipped with the supremum norm

$$\|g\|_Z = \|g\|_\infty := \max_{u \in U} \|g(u)\|_V.$$

Moreover, we assume that $A : X \to Z$ is a bounded linear operator, so that point evaluations of $g^\dagger$ in equation (2) are well-defined. Notice that since U is compact, $Z \subset Y$, and with a slight abuse of notation we identify the functions in Z also as elements of Y.

Following previous work [9, 5], for any $u \in U$ we define the sampled operator $A_u \in \mathcal{L}(X, V)$ as

$$A_u f = (Af)(u).$$

We make the following Assumption:

Assumption 2.1. Assume that $\|A\|_{X \to Z} \leq 1$, where $\|A\|_{X \to Z}$ is the norm of A as an operator from X to Z.

Thanks this Assumption, it also holds that for all $u \in U$ and $f \in X$,

$$\|A_u f\|_V \leq \|f\|_X.$$

Moreover, by the continuity of A on Z, we also have that the mapping $u \mapsto (Af)(u)$ is measurable for all $f \in X$. Notice that assuming $\|A\|_{X \to Z} \leq 1$ is not restrictive; for larger values of the norm one can renormalize the problem in the spirit of [9, 5]. For every $u \in U$, we can interpret the operator $A_u\colon X \to V$ as a composition $A_u = S_u A$, where

$$S_u\colon Z \to V \quad \text{s.t.} \quad S_u(g) = g(u), \quad g \in Z$$

is referred to as the sampling operator, and clearly satisfies $\|S_u\|_{Z\to V} \le 1$ for all $u \in U$.

Next we consider sampling at multiple design points $\{u_i\}_{i=1}^N \subset U$. Let us introduce the following notation for the product spaces $U_N = \oplus_{i=1}^N U$ with the usual topology and $V_N = \oplus_{i=1}^N V$ with the inner product

$$\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle_{V_N} = \frac{1}{N} \sum_{j=1}^N \langle v_j, \tilde{v}_j \rangle_V,$$

where $\mathbf{v} = (v_j)_{j=1}^N, \tilde{\mathbf{v}} = (\tilde{v}_j)_{j=1}^N \in V_N$. The multiple sampling operator $A_{\mathbf{u}} \in \mathcal{L}(Z, V_N)$ is defined by

$$A_{\mathbf{u}} f = (A_{u_i} f)_{i=1}^N \in V_N$$

for $\mathbf{u} = \{u_i\}_{i=1}^N$, and can be equivalently understood as the composition $A_{\mathbf{u}} = S_{\mathbf{u}} A$, where

$$S_{\mathbf{u}} \colon Z \to V_N \quad \text{s.t.} \quad S_{\mathbf{u}}(g) = (g(u_i))_{i=1}^N, \quad g \in Z.$$

Notice carefully that

$$A_{\mathbf{u}}^* \mathbf{v} = \frac{1}{N} \sum_{i=1}^N A_{u_i}^* v_i \quad \text{for any } \mathbf{v} = (v_i)_{i=1}^N \in V_N.$$

Moreover, in the following it is convenient to introduce the following notation for the *normal sampling operator* $B_{\mathbf{u}} := A_{\mathbf{u}}^* A_{\mathbf{u}} \in \mathcal{L}(X, X^*)$, where X^* is the continuous dual of X.

In the following, we will consider the design points $\{u_i\}_{i=1}^N$ as a random sample drawn from a probability distribution μ. Thus, let μ be a probability measure on U and define the corresponding weighted space $Y_\mu = L^2(U, \mu; V)$ as a Hilbert space induced by the inner product

$$\langle g_1, g_2 \rangle_{Y_\mu} := \int_U \langle g_1(u), g_2(u) \rangle_V \mu(du).$$

Clearly, $Z \subset \mathcal{C}(U; V) \subset Y_\mu$ and we denote

$$A_\mu = \iota A : X \to Y_\mu,$$

where $\iota : Z \to Y_\mu$ is the canonical injection map. As a simple example, one can consider a uniform distribution on a bounded domain U. Obviously, in such a case, the inner products of Y_μ and Y coincide up to a constant.

2.2 Problem setting

For the ground truth $f^\dagger \in X$ we define a noise-free observations $\mathbf{g}_N$ by

$$\mathbf{g}_N := A_{\mathbf{u}} f^\dagger, \tag{4}$$

and the noisy observation $\mathbf{g}_N^\delta$ as

$$\mathbf{g}_N^\delta := A_{\mathbf{u}} f^\dagger + \delta \epsilon_N, \tag{5}$$

where $\delta > 0$ is the noise level, $\epsilon_N = (\epsilon_N^i)_{i=1}^N \in V_N$ is a random variable such that $\epsilon_N^i \sim \epsilon$ i.i.d. where ϵ is independent of μ, zero-mean and $\|\epsilon\|_V$ is sub-Gaussian. We briefly recall (see also [33, Definition 2.5.6]) that a real-valued random variable ξ is called sub-Gaussian if

$$\mathbb{P}(|\xi| \ge t) \le 2 \exp(-Kt^2) \qquad \forall t > 0$$

for some constant $K > 0$. The sub-Gaussian norm of such variables is defined as follows:

$$\|\xi\|_{\mathrm{sG}} = \inf\left\{c \geq 0 \mid \mathbb{E}\exp\left(\frac{\xi^2}{c^2}\right) \leq 2\right\}.$$

Therefore, we require that

$$\Big\| \|\epsilon\|_V \Big\|_{\mathrm{sG}} = M < \infty \tag{6}$$

It is always possible to assume that $M = 1$, by rescaling ϵ to ϵ/M and δ to δM. Notice that here $\delta > 0$ plays the role of standard deviation. To build intuition, we point out that a normally distributed ϵ in $V = \mathbb{R}$ with standard deviation σ satisfies (6) with $M = C\sigma$, being C an absolute constant (see [33, Example 2.5.8]).

In the following we consider regularized solutions $f^\delta_{\alpha,N}$ to problems (4) and (5) given by

$$f^\delta_{\alpha,N} \in \operatorname*{arg\,min}_{f\in X} J^\delta_{\alpha,N}(f) := \operatorname*{arg\,min}_{f\in X}\left\{\frac{1}{2}\left\|A_{\mathbf{u}}f - \mathbf{g}^\delta_N\right\|^2_{V_N} + \alpha R(f)\right\}. \tag{7}$$

A regularized solution for the noise-free data is denoted by $f_{\alpha,N}$. Below we introduce a set of Assumptions to guarantee its existence, together with some additional properties. Notice that we do not require the minimizers of such problems to be unique at this stage.

Assumption 2.2. There exists a topology τ on X, with respect to which the sampling operator $S_{\mathbf{u}}$ is lower semicontinuous. The convex functional $R : X \to \mathbb{R}\cup\{\infty\}$ satisfies the following four condition:

(R1) the functional R is lower semicontinuous in the topology τ;

(R2) the sublevel sets $M_r = \{R \leq r\}$ are sequentially compact in the topology τ;

(R3) the convex conjugate $R^\star$ is finite on a ball in X^* centered at zero;

(R4) $R(-f) = R(f)$ for all $f \in X$.

The results contained in the next Section are set in a deterministic framework. However, in the following we consider more specific functionals R that ensure uniqueness of $f^\delta_{\alpha,N}$ and comment on the measurability of the learning method. Notice that the symmetry condition (R4) is not necessary, but is employed to make the results more accessible.

3 Bounds on the Bregman distance

The optimality criterion associated with (7) is given by

$$A^*_{\mathbf{u}}(A_{\mathbf{u}}f^\delta_{\alpha,N} - \mathbf{g}^\delta_N) + \alpha r^\delta_{\alpha,N} = 0 \tag{8}$$

for $r^\delta_{\alpha,N} \in \partial R(f^\delta_{\alpha,N})$, where ∂R denotes the subdifferential:

$$\partial R(f) = \{r \in X^* \mid R(f) - R(\tilde{f}) \leq \langle r, f - \tilde{f}\rangle_{X^*\times X} \text{ for all } \tilde{f} \in X\}.$$

Moreover, for $r_f \in \partial R(f)$ and $r_{\tilde{f}} \in \partial R(\tilde{f})$ we define the symmetric Bregman distance between f and $\tilde{f}$ as

$$D_R^{r_f,r_{\tilde{f}}}(f,\tilde{f}) = \langle r_f - r_{\tilde{f}}, f - \tilde{f}\rangle_{X^*\times X}.$$

When the subdifferential elements $r_f \in \partial R(f)$ are unique, we will drop the dependence on the subgradients in the notation of the symmetric Bregman distance and write simply $D_R(f,\tilde{f})$.

Proposition 3.1 (A-priori estimates)**.** *Let R satisfy the Assumption 2.2. Then the functional $J^\delta_{\alpha,N}$ has a minimizer. Any minimizer $f^\delta_{\alpha,N} \in X$ of $J^\delta_{\alpha,N}$ satisfies*

$$R(f^\delta_{\alpha,N}) \le R(f^\dagger) + \frac{\delta^2}{2\alpha} \|\epsilon_N\|^2_{V_N} . \tag{9}$$

In addition, if R is p-homogeneous with $p > 1$ we have for some constant $C > 0$ that

$$R(f^\delta_{\alpha,N}) \le C \left(R(f^\dagger) + \left(\frac{\delta}{\alpha}\right)^{\frac{p}{p-1}} R^\star(A^*_{\mathbf{u}}\epsilon_N) \right). \tag{10}$$

Proof. Consider the sublevel set $M = \{f \in X \mid J^\delta_{\alpha,N}(f) \le J^\delta_{\alpha,N}(f^\dagger)\}$. Now, any $f \in M$ satisfies

$$\begin{aligned} \alpha R(f) &\le J^\delta_{\alpha,N}(f^\dagger) - \frac{1}{2} \left\| A_{\mathbf{u}} f - \mathbf{g}^\delta_N \right\|^2_{V_N} \\ &= \frac{\delta^2}{2} \|\epsilon_N\|^2_{V_N} + \alpha R(f^\dagger) - \frac{1}{2} \left\| A_{\mathbf{u}}(f - f^\dagger) - \delta\epsilon_N \right\|^2_{V_N} . \end{aligned} \tag{11}$$

This immediately leads to estimate (9), discarding the non-positive term on the right-hand side. In order to get (10), we first expand (11) as follows:

$$\alpha R(f) \le \alpha R(f^\dagger) - \frac{1}{2} \left\| A_{\mathbf{u}}(f - f^\dagger) \right\|^2_{V_N} - \langle A_{\mathbf{u}}(f - f^\dagger), \delta\epsilon_N \rangle_{V_n}.$$

Then, by the generalized Fenchel–Young's inequality we get

$$\begin{aligned} \langle A_{\mathbf{u}}(f - f^\dagger), \delta\epsilon_N \rangle_{V_N} &= \langle f - f^\dagger, \delta A^*_{\mathbf{u}}\epsilon_N \rangle_{V_N} \\ &\le R\left(\beta\alpha^{\frac{1}{p}}(f - f^\dagger)\right) + R^\star\left(\frac{\delta}{\beta\alpha^{\frac{1}{p}}} A^*_{\mathbf{u}}\epsilon_N\right) \\ &\le C\beta^p\alpha \left(R(f) + R(f^\dagger)\right) + \frac{\delta^{\frac{p}{p-1}}}{\beta^{\frac{p}{p-1}}\alpha^{\frac{1}{p-1}}} R^\star(A^*_{\mathbf{u}}\epsilon_N), \end{aligned} \tag{12}$$

where $\beta > 0$ is an arbitrary constant and the triangle inequality for R follows due to convexity, (R4) in Assumption 2.2 and homogeneity with some constant $C > 0$ depending on p. Applying inequality (12) to the second to last expression in (11) and setting $\beta^p = \frac{1}{2C}$ yields the a priori estimate (10) after a division by α.

The existence of the minimizer follows by standard arguments. Assume that $\{f_j\}_{j=1}^\infty \subset M$ is a minimizing sequence of $J^\delta_{\alpha,N}$. Since $R(f_j) \le \frac{1}{\alpha} J^\delta_{\alpha,N}(f_j) \le \frac{1}{\alpha} J^\delta_{\alpha,N}(f^\dagger)$ for each j, the sequence is contained in a suitable subset of R, hence by (R2) in Assumption 2.2 we can extract a τ-converging subsequence $f_{j_k} \to \tilde{f} \in X$. Finally, Assumption 2.2 ensures that both the terms in $J^\delta_{\alpha,N}$ are sequentially lower continuous in the topology τ, whence $\tilde{f}$ is a minimizer of $J^\delta_{\alpha,N}$. □

Proposition 3.2. *Suppose Assumption 2.2 is satisfied. Then any regularized solution $f^\delta_{\alpha,N}$ given by* (7) *satisfies*

$$\begin{aligned} D_R^{r^\delta_{\alpha,N}, r^\dagger}(f^\delta_{\alpha,N}, f^\dagger) \le \inf_{\bar{w}\in V_N} &\left(R^\star\left((\Gamma_1^{-1})^*(r^\dagger - A^*_{\mathbf{u}}\bar{w})\right) + \frac{\alpha}{2}\|\bar{w}\|^2_{V_N}\right) + R(\Gamma_1(f^\dagger - f^\delta_{\alpha,N})) \\ &+ \frac{1}{\alpha}\left(R^\star\left(\delta(\Gamma_2^{-1})^* A^*_{\mathbf{u}}\epsilon_N\right) + R\left(\Gamma_2(f^\dagger - f^\delta_{\alpha,N})\right)\right), \end{aligned} \tag{13}$$

where $r^\delta_{\alpha,N} \in \partial R(f^\delta_{\alpha,N})$, $r^\dagger \in \partial R(f^\dagger)$ and $\Gamma_1, \Gamma_2 : X \to X$ are arbitrary linear invertible operators.

Proof. Let us apply the data-generating distribution of $\mathbf{g}^\delta_N$ given in (5) to the optimality criterion (8) and substract $r^\dagger$ on both sides to obtain

$$B_{\mathbf{u}}(f^\delta_{\alpha,N} - f^\dagger) + \alpha(r^\delta_{\alpha,N} - r^\dagger) = -\alpha r^\dagger + \delta A^*_{\mathbf{u}}\epsilon_N.$$

Now taking dual pairing with $f^\delta_{\alpha,N} - f^\dagger$ on both sides yields

$$\begin{aligned}\left\|A_{\mathbf{u}}(f^\delta_{\alpha,N} - f^\dagger)\right\|^2_{V_N} &+ \alpha D_R^{r^\delta_{\alpha,N}, r^\dagger}(f^\delta_{\alpha,N}, f^\dagger) \\ &= \alpha\langle r^\dagger, f^\dagger - f^\delta_{\alpha,N}\rangle_{X^*\times X} + \delta\langle A^*_{\mathbf{u}}\epsilon_N, f^\delta_{\alpha,N} - f^\dagger\rangle_{X^*\times X} \quad (14)\end{aligned}$$

Applying both standard Young's and Fenchel–Young's inequalities to the first term on the right hand side now yields, for any $\bar{w} \in V_N$

$$\begin{aligned}&\alpha\langle r^\dagger, f^\dagger - f^\delta_{\alpha,N}\rangle_{X^*\times X} \\ &\quad = \alpha\langle (\Gamma_1^{-1})^*(r^\dagger - A^*_{\mathbf{u}}\bar{w}), \Gamma_1(f^\dagger - f^\delta_{\alpha,N})\rangle_{X^*\times X} + \alpha\langle \bar{w}, A_{\mathbf{u}}(f^\dagger - f^\delta_{\alpha,N})\rangle_{V_N} \\ &\quad \le \alpha R^\star\left((\Gamma_1^{-1})^*(r^\dagger - A^*_{\mathbf{u}}\bar{w})\right) + \alpha R(\Gamma_1(f^\dagger - f^\delta_{\alpha,N})) + \frac{\alpha^2}{2}\|\bar{w}\|^2_{V_N} + \frac{1}{2}\left\|A_{\mathbf{u}}(f^\delta_{\alpha,N} - f^\dagger)\right\|^2_{V_N},\end{aligned}$$

where we introduced an arbitrary invertible linear operator $\Gamma_1 : X \to X$. Similarly, the second term on the right hand side of (14) can be bounded by

$$\begin{aligned}\delta\langle A^*_{\mathbf{u}}\epsilon_N, f^\dagger - f^\delta_{\alpha,N}\rangle_{X^*\times X} &= \delta\langle (\Gamma_2^{-1})^* A^*_{\mathbf{u}}\epsilon_N, \Gamma_2(f^\delta_{\alpha,N} - f^\dagger)\rangle_{X^*\times X} \\ &\le R^\star((\delta\Gamma_2^{-1})^* A^*_{\mathbf{u}}\epsilon_N) + R(\Gamma_2(f^\dagger - f^\delta_{\alpha,N})) \quad (15)\end{aligned}$$

Now dividing by α on both sides yields the claim. □

An alternative bound for the Bregman distance between $f^\delta_{\alpha,N}$ and $f^\dagger$ is provided in the following result.

Proposition 3.3. *Suppose Assumption 2.2 is satisfied. Then any regularized solution $f^\delta_{\alpha,N}$ given by* (7) *satisfies*

$$\begin{aligned}&D_R^{r^\delta_{\alpha,N}, r^\dagger}(f^\delta_{\alpha,N}, f^\dagger) \\ &\quad \le \inf_{\bar{w}\in V_N}\left(R^\star\left((\Gamma^{-1})^*(r^\dagger - A^*_{\mathbf{u}}\bar{w})\right) + \frac{\alpha}{2}\|\bar{w}\|^2_{V_N}\right) + R(\Gamma(f^\dagger - f^\delta_{\alpha,N})) + \frac{\delta^2}{2\alpha}\|\epsilon_N\|^2_{V_N}, \quad (16)\end{aligned}$$

where $r^\delta_{\alpha,N} \in \partial R(f^\delta_{\alpha,N})$, $r^\dagger \in \partial R(f^\dagger)$ and $\Gamma : X \to X$ is an arbitrary invertible linear operator.

Proof. The proof is identical to the previous Proposition. However, we apply the bound

$$\delta\langle A^*_{\mathbf{u}}\epsilon_N, f^\dagger - f^\delta_{\alpha,N}\rangle_{X^*\times X} \le \frac{\delta^2}{2}\|\epsilon_N\|^2_{V_N} + \frac{1}{2}\left\|A_{\mathbf{u}}(f^\delta_{\alpha,N} - f^\dagger)\right\|^2_{V_N}. \quad (17)$$

instead of (15). □

Remark 3.4. *Let us note that in what follows the Propositions 3.2 and 3.3 yield different convergence rates. This will result in two different error bounds, which are reported in Theorem 4.6 and Theorem 4.9 for the case $p = 2$ and in Theorem 4.10 and Theorem 4.11 for the case $1 < p < 2$. At this stage, this can be seen by observing the variance terms appearing in the above propositions, which are $\frac{1}{\alpha} R^\star(\delta(\Gamma_2^{-1})^* A_{\mathbf{u}}^* \epsilon_N)$ and $\frac{\delta^2}{2\alpha} \|\epsilon_N\|_{V_N}^2$, respectively. The expectation of the former decays w.r.t. N, whereas the expectation of the latter is independent of N. Therefore, for a fixed noise level δ one cannot expect convergence with the bounds developed based on Proposition 3.3. On the other hand, the rest of the terms of the upper bound (generalized approximation error) are rather similar in both propositions, which will imply different balancing properties for the estimate and therefore different convergence regimes.*

4 The p-homogeneous regularizer and convergence rates

From here on, we assume that the space X is a $2-$convex and $p-$smooth Banach space, with $1 < p \leq 2$ (see Definitions 2.32-2.33 in [28]). Notice that, by Theorem 2.50 in [28], this also ensures that X is reflexive, and by Remark 2.38 in [28] the norm $\|\cdot\|_X$ is Fréchet differentiable. Then, we consider the p-homogenous regularizer

$$R(f) = \frac{1}{p} \|f\|_X^p . \tag{18}$$

This definition entails several properties of the regularizer:

- Firstly, such an R satisfies Assumption 2.2 choosing τ to be the weak topology of X. In particular, condition (R2) is guaranteed by the Banach-Alaoglu Theorem via the reflexivity of X.
- Secondly, the differentiability of $\|\cdot\|_X$ and the fact that $p > 1$ ensure that ∂R is single-valued, and therefore, in the following, we drop the related notation from the Bregman distance.
- Finally, the 2-convexity of X ensures that X is also strictly convex (see Theorem 2.50 in [28]), which finally implies, since $p > 1$, the R is also a stricly convex functional; hence, the regularized solution $f_{\alpha,N}^\delta$ of (7) is unique.

In addition, the mapping

$$(f, (u_i, g_i^\delta)_{i=1}^N) \mapsto \frac{1}{2} \left\| A_{\mathbf{u}} f - \mathbf{g}_N^\delta \right\|_{V_N}^2 + \alpha R(f)$$

is continuous and the measurability of $(\mathbf{u}, \epsilon_N) \mapsto f_{\alpha,N}^\delta$ with respect to the universal completion of the product σ-algebra of $(U \times V)^N$ follows by Aumann's measurable selection principle (analogous to [31, Lemma 6.23]). In particular, following the Definition [31, Definition 6.2] the introduced learning method is measurable.

4.1 General bounds

We start by introducing the following notations. Let us abbreviate

$$E_{\beta,\mathbf{u}}(\bar{w}; r^\dagger) := R^\star\left(r^\dagger - A_{\mathbf{u}}^* \bar{w}\right) + \frac{\beta}{2} \|\bar{w}\|_{V_N}^2 .$$

and, denoting by $r^\dagger$ the only element of $\partial R(f^\dagger)$, set

$$\mathscr{R}(\beta, \mathbf{u}; f^\dagger) = \inf_{\bar{w} \in V_N} E_{\beta,\mathbf{u}}(\bar{w}; r^\dagger).$$

As we will observe, the decay of the term $\mathscr{R}$ w.r.t. N and β will be instrumental in deriving the convergence rate of the reconstruction error as $\mathscr{R}$ relates to the underlying source condition regaring the ground truth and the mapping properties of A_μ. In this Section we work under a general assumption on the decay, see e.g. inequality (35). In Section 5.1 we specify standard source conditions under which the decay can be bounded in the case of Besov penalties (Proposition 5.4). Next, we develop Propositions 3.2 and 3.3 further by combining them with a priori bounds and the following technical Lemma that applies to p-homogeneous functionals.

Lemma 4.1. *Suppose R is of the form* (18) *and let $f, \tilde{f} \in X$. It follows that for $p = 2$ we have*

$$R(f - \tilde{f}) = \frac{1}{2} D_R(f, \tilde{f})$$

and for $1 < p < 2$ it holds that

$$\gamma^p R(f - \tilde{f}) \leq C \left(1 - \frac{p}{2}\right) \gamma^{\frac{2p}{2-p}} \max\left\{R(f), R(\tilde{f})\right\} + \frac{p}{2} D_R(f, \tilde{f}),$$

for some $C > 0$ depending on p with any $\gamma > 0$.

Proof. The case $p = 2$ is trivial. For $1 < p < 2$ consider the Xu–Roach inequality II [28, Thm. 2.40(b)] in X that yields

$$D_R(f, \tilde{f}) \geq C \max\left\{\|f\|_X, \left\|\tilde{f}\right\|_X\right\}^{p-2} \left\|f - \tilde{f}\right\|_X^2 = Cp \max\left\{\|f\|_X, \left\|\tilde{f}\right\|_X\right\}^{p-2} R(f - \tilde{f})^{\frac{2}{p}},$$

and, therefore,

$$R(f - \tilde{f}) \leq \frac{C}{p} \gamma^{\frac{p^2}{2}} \max\left\{\|f\|_X, \left\|\tilde{f}\right\|_X\right\}^{\frac{p(2-p)}{2}} \cdot \gamma^{-\frac{p^2}{2}} D_R(f, \tilde{f})^{\frac{p}{2}}$$

for any $\gamma > 0$. Next, applying Young's inequality to the right-hand side with Hölder conjugates $\left(\frac{2}{2-p}, \frac{2}{p}\right)$ yields

$$R(f - \tilde{f}) \leq \left(1 - \frac{p}{2}\right) \left(\frac{C}{p}\right)^{\frac{2}{2-p}} \gamma^{\frac{p^2}{2-p}} \max\left\{\|f\|_X, \left\|\tilde{f}\right\|_X\right\}^p + \frac{p}{2} \gamma^{-p} D_R(f, \tilde{f}).$$

Finally, multiplying both sides by γ^p and using the fact that

$$\max\left\{\|f\|_X, \left\|\tilde{f}\right\|_X\right\}^p \leq \max\left\{\|f\|_X^p, \left\|\tilde{f}\right\|_X^p\right\} = p \max\left\{R(f), R(\tilde{f})\right\},$$

we obtain the claim. □

Now we are ready to establish more concrete bounds based on Propositions 3.2 and 3.3. Let us first consider the quadratic case $p = 2$.

Theorem 4.2. *Suppose that Assumption 2.2 holds and R is of the form* (18) *with $p = 2$. Then the regularized solution $f^\delta_{\alpha,N}$ given by* (7) *satisfies*

$$D_R(f^\delta_{\alpha,N}, f^\dagger) \leq \min\left\{4\mathscr{R}\left(\frac{\alpha}{2}, N; f^\dagger\right) + \frac{2\delta^2}{\alpha^2} \left\|A^*_{\mathbf{u}} \epsilon_N\right\|^2_{X^*}, 2\mathscr{R}\left(\alpha, N; f^\dagger\right) + \frac{\delta^2}{\alpha} \left\|\epsilon_N\right\|^2_{V_N}\right\}. \quad (19)$$

Proof. In the statement of Proposition 3.2, consider $\Gamma_1 = \gamma_1 I$ and $\Gamma_2 = \gamma_2 I$. By the homogeneity of R, and applying the first estimate in Lemma 4.1, we have that

$$\left(1 - \frac{\gamma_1^2}{2} - \frac{\gamma_2^2}{2\alpha}\right) D_R(f_{\alpha,N}^\delta, f^\dagger) \le \gamma_1^{-2}\mathscr{R}(\alpha\gamma_1^2, N; r) + \frac{\delta^2}{2\alpha\gamma_2^2}\left\|A_{\mathbf{u}}^*\epsilon_N\right\|_{X^*}^2$$

Now setting $\gamma_1^2 = \frac{\gamma_2^2}{\alpha} = \frac{1}{2}$ yields

$$\frac{1}{2} D_R(f_{\alpha,N}^\delta, f^\dagger) \le 2\mathscr{R}\left(\frac{\alpha}{2}, N; f^\dagger\right) + \frac{\delta^2}{\alpha^2}\left\|A_{\mathbf{u}}^*\epsilon_N\right\|_{X^*}^2 .$$

Similarly, starting from the statement of Proposition 3.3, setting $\Gamma = \gamma I$, via Lemma 4.1 we have

$$\left(1 - \frac{\gamma^2}{2}\right) D_R(f_{\alpha,N}^\delta, f^\dagger) \le \gamma^{-2}\mathscr{R}(\alpha\gamma^2, N; f^\dagger) + \frac{\delta^2}{2\alpha}\left\|\epsilon_N\right\|_{V_N}^2$$

Setting $\gamma^2 = 1$ yields the second part of the claim. This completes the proof. □

Theorem 4.3. *Suppose that Assumption 2.2 holds and R is of the form* (18) *with $1 < p < 2$. Then the regularized solution $f_{\alpha,N}^\delta$ given by* (7) *satisfies the following two inequalities:*

(i) It holds that

$$D_R(f_{\alpha,N}^\delta, f^\dagger) \le \widetilde{C}_p\left[\gamma_1^{-q}\mathscr{R}(\alpha\gamma_1^q, \mathbf{u}; f^\dagger) + H(\alpha,\delta,\gamma_1,\gamma_2) R^\star(A_{\mathbf{u}}^*\epsilon_N) + \left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{2}{2-p}} R(f^\dagger)\right] \quad (20)$$

for arbitrary $\gamma_1, \gamma_2 > 0$, where $\widetilde{C}_p > 0$ is a constant dependent on p,

$$H(\alpha,\delta,\gamma_1,\gamma_2) = \frac{\delta^q}{\alpha\gamma_2^q} + \left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{2}{2-p}}\left(\frac{\delta}{\alpha}\right)^q . \quad (21)$$

(ii) We have

$$D_R(f_{\alpha,N}^\delta, f^\dagger) \le C_p\left(\gamma^{-q}\mathscr{R}(\alpha\gamma^q, \mathbf{u}; f^\dagger) + \frac{\delta^2}{\alpha}\left(1 + \gamma^{\frac{2p}{2-p}}\right)\left\|\epsilon_N\right\|_{V_N}^2 + \gamma^{\frac{2p}{2-p}} R(f^\dagger)\right) \quad (22)$$

for arbitrary $\gamma > 0$, where $C_p > 0$ is a constant dependent on p and (p, q) are Hölder conjugates.

Proof. Consider the first claim. By choosing $\Gamma_1 = \gamma_1 I$ and $\Gamma_2 = \gamma_2 I$ in Proposition 3.2, we get

$$D_R(f_{\alpha,N}^\delta, f^\dagger) \le \gamma_1^{-q}\mathscr{R}(\alpha\gamma_1^q, \mathbf{u}; f^\dagger) + \frac{\delta^q}{\alpha\gamma_2^q} R^\star(A_{\mathbf{u}}^*\epsilon_N) + \left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right) R(f^\dagger - f_{\alpha,N}^\delta).$$

Next, by applying Lemma 4.1 to the last term, with $\gamma = \left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{1}{p}}$, we get

$$\begin{aligned} D_R(f_{\alpha,N}^\delta, f^\dagger) \le &\gamma_1^{-q}\mathscr{R}(\alpha\gamma_1^q, \mathbf{u}; f^\dagger) + \frac{\delta^q}{\alpha\gamma_2^q} R^\star(A_{\mathbf{u}}^*\epsilon_N) + \frac{p}{2} D_R(f_{\alpha,N}^\delta, f^\dagger) \\ &+ C\left(1 - \frac{p}{2}\right)\left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{2}{2-p}} \max\{R(f_{\alpha,N}^\delta), R(f^\dagger)\}. \end{aligned}$$

In the last term on the right-hand side, we bound the maximum of the two quantities with their sum, and we use the second claim of Proposition 3.1 to get

$$\left(1-\frac{p}{2}\right) D_R(f_{\alpha,N}^\delta, f^\dagger) \leq \gamma_1^{-q} \mathscr{R}(\alpha\gamma_1^q, \mathbf{u}; f^\dagger) + \left[\frac{\delta^q}{\alpha\gamma_2^q} + C\left(1-\frac{p}{2}\right)\left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{2}{2-p}} \left(\frac{\delta}{\alpha}\right)^q\right] R^\star(A_{\mathbf{u}}^* \epsilon_N)$$
$$+ C\left(1-\frac{p}{2}\right)\left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{2}{2-p}} R(f^\dagger).$$

For the second inequality we deduce similarly applying Lemma 4.1 and the first a priori bound in Proposition 3.1 to Proposition 3.3 that

$$\begin{aligned}
&\left(1-\frac{p}{2}\right) D_R(f_{\alpha,N}^\delta, f^\dagger) \\
&\quad \leq \gamma^{-q} \mathscr{R}(\alpha\gamma^q, \mathbf{u}; f^\dagger) + \frac{\delta^2}{2\alpha} \|\epsilon_N\|_{V_N}^2 + C\left(1-\frac{p}{2}\right)\gamma^{\frac{2p}{2-p}} \max\{R(f_{\alpha,N}^\delta), R(f^\dagger)\} \\
&\quad \leq \gamma^{-q} \mathscr{R}(\alpha\gamma^q, \mathbf{u}; f^\dagger) + \frac{\delta^2}{2\alpha}\left(1 + C\left(1-\frac{p}{2}\right)\gamma^{\frac{2p}{2-p}}\right) \|\epsilon_N\|_{V_N}^2 + C\left(1-\frac{p}{2}\right)\gamma^{\frac{2p}{2-p}} R(f^\dagger),
\end{aligned}$$

which yields inequality (22) after dividing by $1-\frac{p}{2}$. This completes the proof. □

4.2 The case $p = 2$ in Hilbert spaces

In this Section we compare our technique developed above and the convergence rates it implies to the optimal convergence rates known in the case when X is a Hilbert space and $p = 2$ (observe that a Hilbert space is always 2−convex and 2−smooth, see e.g. Example 2.46 in [28]). Although optimal rates are known for general spectral regularization schemes [5], our key message below can be demonstrated by considering classical Tikhonov regularization

$$R(f) = \frac{1}{2}\|f\|_X^2. \tag{23}$$

We note that this setting implies that if $r \in \partial R(f)$, then $r = f$. Moreover, recall that $X^* = X$.

Lemma 4.4. *Suppose X is an Hilbert space and R satisfies (23). Then we have*

$$\mathscr{R}(\beta, \mathbf{u}; f^\dagger) = \frac{\beta}{2}\left\|(B_{\mathbf{u}} + \beta I)^{-\frac{1}{2}} f^\dagger\right\|_X^2.$$

Proof. Recall that

$$\mathscr{R}(\beta, \mathbf{u}; f^\dagger) = \inf_{\bar{w}\in V_N}\left(\frac{1}{2}\left\|f^\dagger - A_{\mathbf{u}}^* \bar{w}\right\|_X^2 + \frac{\beta}{2}\|\bar{w}\|_{V_N}^2\right).$$

The minimizing element on the right hand side naturally satisfies

$$\bar{w}_{\inf} = (A_{\mathbf{u}} A_{\mathbf{u}}^* + \beta I)^{-1} A_{\mathbf{u}} f^\dagger$$

Next, we have that

$$
\begin{aligned}
2E_{\beta,N}(\bar{w}_{\text{inf}}, f^\dagger) &= \left\| f^\dagger - A_{\mathbf{u}}^* \bar{w}_{\text{inf}} \right\|_X^2 + \beta \left\| \bar{w}_{\text{inf}} \right\|_{V_N}^2 \\
&= \left\| f^\dagger \right\|_X^2 - 2\langle f^\dagger, A_{\mathbf{u}}^* \bar{w}_{\text{inf}} \rangle + \langle (A_{\mathbf{u}} A_{\mathbf{u}}^* + \beta I) \bar{w}_{\text{inf}}, \bar{w}_{\text{inf}} \rangle \\
&= \left\| f^\dagger \right\|_X^2 - \langle f^\dagger, A_{\mathbf{u}}^* \bar{w}_{\text{inf}} \rangle \\
&= \langle f^\dagger, \left(I - A_{\mathbf{u}}^* (A_{\mathbf{u}} A_{\mathbf{u}}^* + \beta I)^{-1} A_{\mathbf{u}} \right) f^\dagger \rangle \\
&= \langle f^\dagger, \left(I - (\beta I + B_{\mathbf{u}})^{-1} B_{\mathbf{u}} \right) f^\dagger \rangle \\
&= \beta \langle f^\dagger, (\beta I + B_{\mathbf{u}})^{-1} f^\dagger \rangle.
\end{aligned}
$$

This yields the result. □

The previous Lemma enables us to prove sharp concentration results for the $\mathscr{R}$ term based on techniques developed in previous inverse learning theory literature. Towards this end, let us introduce so-called *effective dimension* which is defined by

$$
\mathcal{N}(\alpha) = \operatorname{Tr}\left[(B_\mu + \alpha)^{-1} B_\mu \right] \tag{24}
$$

in the Hilbert space setting. Moreover, the source condition is typically characterized by restricting the ground truth to the subset

$$
\widehat{\Omega}(s, L) = \{ f \in X \mid f = B_\mu^s w, \ \|w\|_X \le L \} \subset X.
$$

Later on, we will focus to a more specific set

$$
\widetilde{\Omega}(L) = \{ f \in X \mid f = A_\mu^* \tilde{w}, \ \|\tilde{w}\|_{Y_\mu} \le L, \ \tilde{w} = \iota(w), \ w \in Z \} \subset X. \tag{25}
$$

Now we are ready to prove the following result.

Proposition 4.5. *Let us define*

$$
\mathcal{B}_N(\beta) := 1 + \left(\frac{2}{N\beta} + \sqrt{\frac{\mathcal{N}(\beta)}{N\beta}} \right)^2 \tag{26}
$$

for any $\beta > 0$ and $N \in \mathbb{N}$. We assume that $f^\dagger \in \widehat{\Omega}(s, L)$ for $s \le \frac{1}{2}$. It follows that

$$
\mathbb{E}\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \le C L^2 \beta^{2s} \mathcal{B}_N(\beta)^{2s} \tag{27}
$$

for some constant $C > 0$ independent of β, N and L. Also, if $f^\dagger \in \widetilde{\Omega}(L)$, it holds that

$$
\mathbb{E}\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \le C L^2 \left(\beta + \frac{1}{N} \right) \tag{28}
$$

for some $C > 0$ with any $\beta > 0$.

Proof. By Lemma 4.4 we have that, for any $\gamma \in (0,1]$, with probability larger than $1-\eta$

$$\begin{aligned}
\mathscr{R}(\beta, \mathbf{u}; f^\dagger) &= \frac{\beta}{2}\left\|(\beta I + B_{\mathbf{u}})^{-\frac{1}{2}} B_\mu^s w\right\|_X^2 \\
&\leq \frac{\beta}{2}\left\|(\beta I + B_{\mathbf{u}})^{s-\frac{1}{2}}\right\|^2 \left\|(\beta I + B_{\mathbf{u}})^{-s} B_\mu^s\right\|^2 L^2 \\
&\leq \frac{\beta}{2}\left\|(\beta I + B_{\mathbf{u}})^{s-\frac{1}{2}}\right\|^2 \left\|(\beta I + B_{\mathbf{u}})^{-1}(\beta I + B_\mu)(\beta I + B_\mu)^{-1} B_\mu\right\|^{2s} L^2 \\
&\leq \frac{\beta}{2} L^2 \left\|(\beta I + B_{\mathbf{u}})^{s-\frac{1}{2}}\right\|^2 \left\|(\beta I + B_{\mathbf{u}})^{-1}(\beta I + B_\mu)\right\|^{2s} \\
&\leq \frac{\beta}{2} L^2 \left\|(I + \beta^{-1} B_{\mathbf{u}})^{s-\frac{1}{2}}\right\|^2 \beta^{2s-1} \mathcal{B}_N(\beta)^{2s} \log^{4s}\left(\frac{2}{\eta}\right) \\
&\leq \frac{1}{2} L^2 \beta^{2s} \mathcal{B}_N(\beta)^{2s} \log^{4s}\left(\frac{2}{\eta}\right),
\end{aligned}$$

where we used that $\|(B_\mu + \beta I)^{-1} B_\mu\| \leq 1$ and we applied Propositions A.2 and A.4. Now the claim follows by Lemma A.3.

For the purpose of the second claim, let us recall the point evaluation operator $S_{\mathbf{u}} : Z \to V_N$ such that

$$S_{\mathbf{u}} f = (f(u_n))_{n=1}^N \in V_N, \tag{29}$$

which allows us to write $A_{\mathbf{u}} = S_{\mathbf{u}} A$. The term $\mathscr{R}$ can be bounded by setting $\bar{w} = S_{\mathbf{u}} \tilde{w}$, which yields

$$\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \leq \left\|(A_\mu^* - A_{\mathbf{u}}^* S_{\mathbf{u}})\tilde{w}\right\|_X^2 + \frac{\beta}{2}\left\|S_{\mathbf{u}}\tilde{w}\right\|_{V_N}^2.$$

We first observe that

$$\mathbb{E}\left\|S_{\mathbf{u}} w\right\|_{V_N}^2 = \frac{1}{N}\sum_{n=1}^N \mathbb{E}\left\|w(u_n)\right\|_V^2 = \frac{1}{N}\sum_{n=1}^N \|w\|_{Y_\mu}^2 = \|w\|_{Y_\mu}^2. \tag{30}$$

Moreover, since the design points are independent, we have

$$\mathbb{E}\left\|(A_\mu^* - A_{\mathbf{u}}^* S_{\mathbf{u}})\tilde{w}\right\|_X^2 = \frac{1}{N}\mathbb{E}\left\|(A_\mu^* - A_u^* S_u)\tilde{w}\right\|_X^2 \leq \frac{C}{N}\|\tilde{w}\|_{Y_\mu}^2,$$

where the last inequality follows due to Assumption 2.1 and the argument in (30). $\square$

By applying Propositions 4.5 and the fact that $\mathbb{E}\left\|A_{\mathbf{u}}^* \epsilon_N\right\|_X^2 < C/N$ as $A_{\mathbf{u}}^* \in \mathcal{L}(V_N, X^*)$ to Theorem 4.2 we deduce the convergence rate of the expected error. Notice that Theorem 4.2 proposes two alternative bounds, which will lead to two different estimates: the outcome of the first one will be denoted as *standard bound*, whereas the outcome of the second one as an *alternative bound*. Further comments on the comparison between them is provided in Remark 4.12. For the next result, we assume that the effective dimension $\mathcal{N}(\alpha)$, defined in (24), can be expressed as $\mathcal{N}(\alpha) = \alpha^{-\frac{1}{b}}$ for $b > 1$.

Theorem 4.6 (*Standard* estimate)**.** *i) Let $f^\dagger \in \widehat{\Omega}(s, L)$, $0 < s \leq \frac{1}{2}$. Assume that $\delta > 0$ is a constant (independent of N). Then, we have*

$$\mathbb{E}\left\|f_{\alpha,N}^\delta - f^\dagger\right\|_X^2 \lesssim L^2\left(\frac{\delta^2}{L^2 N}\right)^{\frac{2s}{2s+2}} \quad \textit{for} \quad \alpha \simeq \left(\frac{\delta^2}{L^2 N}\right)^{\frac{1}{2s+2}}. \tag{31}$$

ii) Let $f^\dagger \in \widetilde{\Omega}(L)$. Suppose that, as $N \to \infty$, it holds $\frac{\delta^2}{N} \to 0$ and $N\delta \to \infty$, then the rate (31) *holds, whereas when $N\delta$ is bounded the optimal rate is N^{-1} and is achieved by $\alpha \simeq N^{-1}$.*

Proof. To prove the first statement, we consider the expected value of the first term in (19) and plug in equation (27), getting

$$\begin{aligned}\mathbb{E}\left\|f^\delta_{\alpha,N} - f^\dagger\right\|_X^2 &\lesssim L^2\alpha^{2s}\mathcal{B}_N(\alpha)^{2s} + \frac{\delta^2}{\alpha^2 N}\\ &\lesssim L^2\alpha^{2s}\left(1 + \frac{1}{N\alpha^{1+\frac{1}{b}}} + \frac{1}{N^2\alpha^2}\right)^{2s} + \frac{\delta^2}{\alpha^2 N}.\end{aligned}$$

In order for the right-hand side to vanish, we need to require that $\alpha \to 0$ and $\alpha^2 N \to \infty$, i.e., $N = o(\alpha^{-2})$. As a consequence,

$$\begin{aligned}\mathbb{E}\left\|f^\delta_{\alpha,N} - f^\dagger\right\|_X^2 &\lesssim L^2\alpha^{2s} + L^2N^{-2s}\alpha^{-\frac{2s}{b}} + L^2N^{-4s}\alpha^{-2s} + \delta^2N^{-1}\alpha^{-2}\\ &= L^2\alpha^{2s} + N^{-1}\alpha^{-2}\left(L^2N^{1-2s}\alpha^{2-\frac{2s}{b}} + L^2N^{1-4s}\alpha^{2-2s} + \delta^2\right)\\ &\lesssim L^2\alpha^{2s} + \frac{\delta^2}{N\alpha^2}.\end{aligned}$$

In the last step, we have used that $N^{1-2s}\alpha^{2-\frac{2s}{b}} = o\left(\alpha^{-2+4s}\alpha^{2-\frac{2s}{b}}\right) = o\left(\alpha^{2s(2-\frac{1}{b})}\right)$ which vanishes if $b > \frac{1}{2}$, and that $N^{1-4s}\alpha^{2-2s} = o\left(\alpha^{-2+8s}\alpha^{2-2s}\right) = o(\alpha^{6s})$, which also vanishes. The optimal choice of α is the one that balances the two remaining terms, hence leading to (31).

For the second estimate, we use instead the bound (27) due to the different source condition, leading to

$$\mathbb{E}\left\|f^\delta_{\alpha,N} - f^\dagger\right\|_X^2 \lesssim L^2\alpha + \frac{L^2}{N} + \frac{\delta^2}{\alpha^2 N} \lesssim \alpha\left(L^2 + \frac{L^2}{\alpha N} + \frac{\delta^2}{\alpha^3 N}\right).$$

In order to ensure convergence, δ need not to be fixed, but still we need to require $\frac{\delta^2}{N} \to 0$. The optimal choice of α is the one ensuring that the third term in the last summation is bounded and asymptotically equivalent to L^2 (hence concluding what is reported in (31)), provided that the second term is vanishing, i.e.,

$$\frac{1/N}{(\delta^2/N)^{\frac{1}{3}}} \to 0 \quad \Rightarrow \quad \frac{1}{\delta N} \to 0$$

If this is not the case (i.e., when δN is bounded), the optimal α is the one balancing the second term, namely, $\alpha \simeq \frac{1}{N}$. □

We remark that, in the limit case $s = \frac{1}{2}$, the achieved convergence rate in the presence of fixed noise level δ is of the order $N^{-1/3}$. Notice that we could extend also the first statement in order to treat the case of non-fixed noise level δ. Nevertheless, for the purpose of this work, statement *i)* is mainly intended to compare the results carried out via the presented technique with the optimal estimates of the statistical learning literature, in which δ is typically a constant.

On the contrary, in statement *ii)* we admit the possibility for δ to vary as $N \to \infty$, which is more common from an inverse problems perspective. To get a more clear interpretation of such statement, suppose that $\delta \simeq N^{-\beta}$: then, (31) shows that the convergence rate is $N^{-\frac{2\beta+1}{3}}$ when $-1/2 < \beta \leq 1$ (so, even if the noise is mildly growing), whereas if the noise decay is faster ($\beta > 1$) the convergence rate gets saturated at N^{-1}.

Remark 4.7. *The result in Theorem 4.6 is comparable to the rates derived in [5], where it is proven that the weak or strong minimax optimal rate for* $\sqrt{\mathbb{E}\left\|f^\delta_{\alpha,N} - f^\dagger\right\|^2_X}$ *is given by*

$$L\left(\frac{\delta^2}{L^2 N}\right)^{\frac{s}{2s+1+\frac{1}{b}}}$$

for $b > 1$ *under certain assumptions on the design measure* μ *that imply* $\mathcal{N}(\alpha) \leq C\alpha^{-\frac{1}{b}}$, *i.e., our assumption regarding the effective dimension. Our setup yields asymptotically the same rate only in the limit* $b = 1$.

Let us briefly explain why such discrepancy emerges: in the Hilbert space setup, the error term can be explicitly solved by

$$\begin{aligned} f^\dagger - f^\delta_{\alpha,N} &= f^\dagger - (B_{\mathbf{u}} + \alpha)^{-1} A^*_{\mathbf{u}} \mathbf{g}^\delta_N \\ &= f^\dagger - (B_{\mathbf{u}} + \alpha)^{-1}(B_{\mathbf{u}} f^\dagger + \delta A^*_{\mathbf{u}} \epsilon_N) \\ &= \alpha(B_{\mathbf{u}} + \alpha)^{-1} f^\dagger - \delta(B_{\mathbf{u}} + \alpha)^{-1} A^*_{\mathbf{u}} \epsilon_N \\ &=: E_{appr} + E_{sample}, \end{aligned} \tag{32}$$

where terms E_{appr} *and* E_{sample} *are called the approximation and sample error, respectively.*

The result in [5] is developed by applying the triangle inequality to identity (32) *and estimating* $\|E_{appr}\|$ *and* $\|E_{sample}\|$ *separately. First, the norm of the approximation error bound shown in [5] essentially coincides with Proposition 4.5 (the difference being the applicable qualification regime of the regularization scheme, which is more limited here). Second, the sample error bound given in [5] is inherently sharper; [5, Prop. 5.8] yields a rate of order* $\mathcal{N}(\alpha)/\alpha N = 1/\alpha^{1+\frac{1}{b}} N$ *compared to*

$$\frac{1}{\alpha^2}\mathbb{E}\left\|A^*_{\mathbf{u}}\epsilon_N\right\|^2_X \leq \frac{C}{\alpha^2 N}.$$

Remark 4.8 (Is it possible to obtain optimal rates?)**.** *As noted in the previous remark our approach developed above can yield suboptimal convergence rates. This feature can be traced back to the choice of utilizing operators* $\Gamma_1 = \gamma_1 I$ *and* $\Gamma_2 = \gamma_2 I$ *when applying Proposition 3.2 in Section 4. Instead, we can set*

$$\Gamma_2 = (B_{\mathbf{u}} + \alpha)^{\frac{1}{2}} \tag{33}$$

and obtain a variance (corresponding to a square of the sample error) term

$$\frac{1}{\alpha}R^\star(\delta(\Gamma_2^{-1})^* A^*_{\mathbf{u}}\epsilon_N) = \frac{\delta^2}{\alpha}\left\|(B_{\mathbf{u}} + \alpha)^{-\frac{1}{2}} A^*_{\mathbf{u}}\epsilon_N\right\|^2_X$$

on the right hand side of the standard bound in Theorem 4.2. With this modification the analysis of the sample error is aligned with [5] and we can apply [5, Prop. 5.2] in order to obtain

$$\frac{1}{\alpha^2}\mathbb{E}\left\|A^*_{\mathbf{u}}\epsilon_N\right\|^2_X \leq \frac{1}{\alpha^{1+\frac{1}{b}} N}.$$

Unfortunately, the choice of Γ_2 *in* (33) *implies that we cannot apply Lemma 4.1 in order to prove Theorem 4.3 and we end up with an extra term*

$$\frac{1}{\alpha}R(\Gamma_2(f^\dagger - f^\delta_{\alpha,N})) = \frac{1}{\alpha}\left\|(B_{\mathbf{u}}+\alpha)^{\frac{1}{2}}(f^\dagger - f^\delta_{\alpha,N})\right\|_X^2$$

on the right hand side of the fixed-noise error upper bound in (19).

Note that while this extra term is unsatisfactory, in principle, by applying identity (32) *and the triangle inequality followed by the technique utilized in [5] one could hope to achieve optimal rates. Clearly, such an argument provides limited insight but demonstrates that the approach could be further developed towards optimality of the rates. It remains part of future work to consider implications of general operators* Γ_1 *and* Γ_2 *in an arbitrary* p*-homogeneous case.*

By considering the second estimate proposed in (19), we can derive an alternative bound for the error. The slightly modified source condition enables the use of a stronger estimate in Proposition 4.5. However, as we will see, the obtained rate is weaker than in Theorem 4.2, which is discussed below.

Proposition 4.9 (*Alternative* estimate)**.** *Let* $f^\dagger \in \widetilde{\Omega}(L)$*. Suppose that, as* $N \to \infty$*, it holds* $\delta \to 0$ *and* $N\delta \to \infty$*, then*

$$\mathbb{E}\left\|f^\delta_{\alpha,N} - f^\dagger\right\|_X^2 \lesssim L\delta \quad \textit{for} \quad \alpha \simeq \frac{\delta}{L}; \tag{34}$$

whereas if $N\delta$ *is bounded the optimal rate is* N^{-1} *and is achieved by* $\alpha \simeq N^{-1}$*.*

Proof. Applying inequality (28) to the second estimate in (19) we obtain

$$\mathbb{E}\left\|f^\delta_{\alpha,N} - f^\dagger\right\|_X^2 \lesssim L^2\left(\alpha + \frac{1}{N}\right) + \frac{\delta^2}{\alpha} \lesssim \alpha\left(L^2 + \frac{L^2}{\alpha N} + \frac{\delta^2}{\alpha^2}\right).$$

As in the proof of Theorem 4.6, the optimal rate is obtained by selecting α so that the third term in the summation is bounded and asymptotically equivalent to L^2, i.e., $\alpha \simeq \frac{\delta}{L}$, provided that the second one is vanishing ($\frac{1}{N\delta} \to 0$); otherwise, saturation occurs on the rate N^{-1}. □

We immediately notice that the obtained rate is weaker than Theorem 4.2. This can be seen by an application of Young's inequality

$$\left(\frac{\delta^2}{N}\right)^{\frac{1}{3}} \lesssim \delta + \frac{1}{N} \leq \max\left\{\delta, \frac{1}{N}\right\},$$

where Hölder conjugates 3/2 and 3 were applied. What we observe is that the saturation point $1/N$ in both convergence rates is due to the Monte Carlo-type estimate in Proposition 4.5. If a faster concentration bound of type

$$\mathbb{E}\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \leq CL^2\left(\beta + \frac{1}{N^\rho}\right)$$

for $\rho > 1$ could be derived, it can be seen that the alternative scheme is preferable in small noise regime such that $\delta \lesssim N^{-\rho}$. This motivates us to consider a general concentration bound for $\mathscr{R}$ in the p-homogeneous case in the next Section.

4.3 Convergence rates for $1 < p \leq 2$

In this Subsection we derive general convergence rate results for the p-homogeneous case in equation (18) with $1 < p \leq 2$. We develop the results under assumptions on the concentration of expectations of the random terms appearing in upper bounds of Theorem 4.3 that generalize the usual bias and variance terms. Such conditions are then proved for specific cases in later Sections.

Theorem 4.10 (p-homogeneous case, standard estimate)**.** *Consider the p-homogeneous regularization functional defined in* (18) *applied to the direct problem introduced in equations* (1) *and* (2)*. Suppose that Assumptions 2.1 and 2.2 are satisfied and that $R(f^\dagger) \leq L$. Moreover, we assume that there exists a constant $Q > 0$ such that*

$$\mathbb{E}\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \leq D_1 \beta + D_2 N^{-Q} \tag{35}$$

and

$$\mathbb{E}R^\star(A^*_{\mathbf{u}}\epsilon_N) \leq D_3 N^{-\frac{q}{2}} \tag{36}$$

for some fixed values $D_1, D_2, D_3 > 0$. Suppose that, as $N \to \infty$, it holds $\frac{\delta^2}{N} \to 0$ and $\delta N^{\frac{3Q}{q} - \frac{1}{2}} \to \infty$: then

$$\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) \lesssim \left(D_1^2 D_3^{\frac{2}{q}} L^{\frac{q-2}{q}}\right)^{\frac{1}{3}} \left(\frac{\delta^2}{N}\right)^{\frac{1}{3}} \quad \textit{for} \quad \alpha \simeq \left(\frac{D_3^{\frac{2}{q}} L^{\frac{q-2}{q}}}{D_1}\right)^{\frac{1}{3}} \left(\frac{\delta^2}{N}\right)^{\frac{1}{3}}; \tag{37}$$

whereas if $\delta N^{\frac{3Q}{q} - \frac{1}{2}}$ is bounded the optimal rate is $N^{-\frac{2Q}{q}}$ and is achieved by $\alpha \simeq N^{-\frac{2Q}{q}}$.

Proof. Let us first estimate

$$\left(\gamma_1^p + \frac{\gamma_2^p}{\alpha}\right)^{\frac{2}{2-p}} \leq C_p \left(\gamma_1^{\frac{2p}{2-p}} + \gamma_2^{\frac{2p}{2-p}} \alpha^{-\frac{2}{2-p}}\right)$$

which together with bounds (35) and (36) yields for the inequality (20) that

$$\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) \leq C \left(D_1 \alpha + Z_1 \gamma_1^{-q} + Z_2 \gamma_1^{\frac{2p}{2-p}} + Z_3 \gamma_2^{-q} + Z_4 \gamma_2^{\frac{2p}{2-p}}\right), \tag{38}$$

where

$$\begin{aligned} Z_1 &= D_2 N^{-Q}, \\ Z_2 &= D_3 \delta^q \alpha^{-q} N^{-\frac{q}{2}} + L, \\ Z_3 &= D_3 \delta^q \alpha^{-1} N^{-\frac{q}{2}} \quad \text{and} \\ Z_4 &= \alpha^{-\frac{2}{2-p}} \left(D_3 \delta^q \alpha^{-q} N^{-\frac{q}{2}} + L\right). \end{aligned}$$

In order to optimize γ_1 and γ_2, we record the following calculation: a function $\varphi(\gamma) = a\gamma^{-q} + b\gamma^{\frac{2p}{2-p}}$ is minimized at $\gamma_* = \left(\frac{a}{b}\right)^{\tilde{p}}$, where $\tilde{p} = \frac{(2-p)(p-1)}{p^2}$. At the minimizer the function φ obtains value

$$\varphi(\gamma_*) = 2b \left(\frac{a}{b}\right)^{\frac{2(p-1)}{p}} = 2a^{\frac{2}{q}} b^{\frac{2}{p}-1}. \tag{39}$$

The optimal choices of $\gamma_1, \gamma_2 > 0$ in inequality (38) is given by

$$\gamma_1 = \left(\frac{Z_1}{Z_2}\right)^{\tilde{p}} \quad \text{and} \quad \gamma_2 = \left(\frac{Z_3}{Z_4}\right)^{\tilde{p}}.$$

This yields a bound

$$\begin{aligned}
\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) &\leq C\left(D_1\alpha + Z_1\left(\frac{Z_1}{Z_2}\right)^{\frac{-q}{\frac{2p}{2-p}+q}} + Z_3\left(\frac{Z_3}{Z_4}\right)^{\frac{-q}{\frac{2p}{2-p}+q}}\right)\\
&= C\left(D_1\alpha + Z_1^{\frac{2}{q}} Z_2^{\frac{2}{p}-1} + Z_3^{\frac{2}{q}} Z_4^{\frac{2}{p}-1}\right).
\end{aligned}$$

Substituting the expressions of Z_1, Z_2, Z_3, Z_4, by direct computations we get

$$\begin{aligned}
\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) \lesssim D_1\alpha &+ D_3^{\frac{2}{q}} L^{\frac{q-2}{q}}(\delta\alpha^{-1}N^{-1/2})^2 + D_2^{\frac{2}{q}} L^{\frac{q-2}{q}} N^{-\frac{2Q}{q}}\\
&+ D_3(\delta\alpha^{-1}N^{-1/2})^q + D_2^{\frac{2}{q}} D_3^{\frac{q-2}{q}}(\delta\alpha^{-1}N^{-1/2})^{q-2}N^{-\frac{2Q}{q}},
\end{aligned}$$

from which we deduce that it is necessary that $(\delta\alpha^{-1}N^{-1/2}) \to 0$, and therefore we need to require $\delta N^{-1/2} \to 0$. Moreover, since $q > 2$, we can neglect the faster terms and get

$$\begin{aligned}
\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) &\lesssim D_1\alpha + D_3^{\frac{2}{q}} L^{\frac{q-2}{q}}(\delta\alpha^{-1}N^{-1/2})^2 + D_2^{\frac{2}{q}} L^{\frac{q-2}{q}} N^{-\frac{2Q}{q}}\\
&\lesssim \alpha\left(D_1 + D_3^{\frac{2}{q}} L^{\frac{q-2}{q}} \frac{\delta^2/N}{\alpha^3} + D_2^{\frac{2}{q}} L^{\frac{q-2}{q}} \frac{N^{-\frac{2Q}{q}}}{\alpha}\right).
\end{aligned}$$

The optimal rate is obtained by selecting α so that the second term in the summation is bounded and asymptotically equivalent to D_1 (which results in the choice described in (37)), provided that the third term is vanishing, i.e.,

$$\frac{N^{-\frac{2Q}{q}}}{\alpha} \to 0 \quad \Rightarrow \quad \frac{N^{-\frac{2Q}{q}}}{(\delta^2/N)^{\frac{1}{3}}} \to 0 \quad \Rightarrow \quad \frac{1}{\delta^{\frac{2}{3}} N^{\frac{2Q}{q}-\frac{1}{3}}} \to 0,$$

that is equivalent to requiring $\delta N^{\frac{3Q}{q}-\frac{1}{2}} \to \infty$. If instead such term is bounded, the third term dominates and the convergence rate cannot get better than $N^{-\frac{2Q}{q}}$, in accordance with the parameter choice $\alpha \simeq N^{-\frac{2Q}{q}}$. □

Theorem 4.11 (p-homogeneous case, alternative estimate)**.** *Consider the p-homogeneous regularization functional defined in* (18) *applied to the direct problem introduced in equations* (1) *and* (2)*. Let Assumptions 2.2 and 2.1 be satisfied and $R(f^\dagger) \leq L$. Moreover, assume that the inequality* (35) *holds for some $Q, D_1, D_2 > 0$. Suppose that, as $N \to \infty$, it holds $\delta \to 0$ and $\delta N^{\frac{2Q}{q}} \to \infty$: then*

$$\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) \lesssim D_1^{\frac{1}{2}}\delta \quad \textit{for} \quad \alpha \simeq D_1^{-\frac{1}{2}}\delta; \tag{40}$$

whereas if $\delta N^{\frac{2Q}{q}}$ is bounded the optimal rate is $N^{-\frac{2Q}{q}}$ and is achieved by $\alpha \simeq N^{-\frac{2Q}{q}}$.

Proof. We have

$$\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) \lesssim D_1\alpha + D_2 N^{-Q}\gamma^{-q} + \frac{\delta^2}{\alpha} + \gamma^{\frac{2p}{2-p}} L.$$

It follows from the calculation in equation (39) that

$$\mathbb{E}D_R(f^\delta_{\alpha,N}, f^\dagger) \lesssim D_1\alpha + \frac{\delta^2}{\alpha} + 2D_2^{\frac{2}{q}} N^{-\frac{2Q}{q}} L^{\frac{2}{p}-1} \lesssim \alpha\left(D_1 + \frac{\delta^2}{\alpha^2} + D_2^{\frac{2}{q}} L^{\frac{q-2}{q}} \frac{N^{-\frac{2Q}{q}}}{\alpha}\right),$$

from which we proceed analogously as in the proof of Theorem 4.10. □

Remark 4.12. *Let us compare the standard and alternative estimates, which read as*

$$\left(\frac{\delta^2}{N}\right)^{\frac{1}{3}} + N^{-\frac{2Q}{q}} \quad vs. \quad \max\left\{\delta, N^{-\frac{2Q}{q}}\right\}.$$

The term $N^{-2Q/q}$ sets the fastest possible rate and in the typical case (see Section 5.1 for the Besov case) we find $Q = \frac{q}{2}$, leading to $1/N$, as in the case $p = 2$. In such a regime the alternative estimate does not yield an improvement, only $-2Q/q < -1$ would yield a rate $\delta \simeq N^{-\rho}$, $\rho > 1$ for which the alternative estimates are better. In the Hilbert case this would need an improvement of Proposition 4.5, from whose proof we see that this is based on the general choice $\overline{w} = S_{\mathbf{u}}\tilde{w}$ (being $S_{\mathbf{u}}$ the evaluation operator defined as in (29)*). A potential improvement is possible only using a more optimal choice of $\overline{w}$ taking into account specific properties of the operator A^* or a more structured randomness taking again depending on A^* (similar to the proof of Theorem 2.1 in [7]). This consideration is outside the focus of this paper, but opens interesting questions for further studies about the optimal balance between approximation and sample errors in problems, where the observational noise is substantially smaller than the inverse of a feasible number of design points.*

5 Strategies for obtaining concentration rates

In this Section, we prove a concrete convergence rate for the special case when X is a Besov space $B^s_{pp}(\mathbb{R}^d)$ and discuss how to derive convergence rates if a continuous embedding of the Banach space X to some Hilbert space X_0 is available.

5.1 Hoeffding's inequality applied to Besov regularizers

Let $X = B^s_{pp}(\Omega) := B^s_p(\Omega)$ be a Besov space [35], being $\Omega \subset \mathbb{R}^d$ a bounded Lipschitz domain. Recall that these are 2−convex and p−smooth, see Example 2.47 in [28]. Consider the following regularization functional, based on the equivalent norm of Besov spaces (see [35]):

$$R(f) = \frac{1}{p}\|f\|^p_{B^s_p} := \frac{1}{p}\sum_{\lambda=1}^{\infty} c_{\lambda,p,s,d}|\langle f, \psi_\lambda\rangle|^p, \tag{41}$$

for some $1 < p < 2$, where

$$c_{\lambda,p,s,d} = 2^{|\lambda|d\left(p(\frac{s}{d}+\frac{1}{2})-1\right)}. \tag{42}$$

Here, $\psi_\lambda : \mathbb{R}^d \to \mathbb{R}$, with $\lambda = 1, ..., \infty$, are suitably regular functions that form an orthonormal wavelet basis for $L^2(\Omega)$ with global indexing λ. The notation $|\lambda|$ is used to denote the scale of the wavelet basis associated with the index λ (see [12]). Notice that when $s = d\left(\frac{1}{p} - \frac{1}{2}\right)$ the Besov norm (41) reduces to the ℓ_p-norm of the wavelet coefficients. It is easily seen (e.g., in [28]) that the convex conjugate of R satisfies

$$R^\star(g) = \frac{1}{q} \|g\|^q_{B_q^{-s}} ,$$

where p and q are Hölder conjugates. Clearly, Assumption 2.2 is satisfied by this choice and R is p-homogeneous.

As in Section 2.1, we consider $Y = L^2(U;V)$, and $Z = C(U;V)$, equipped with the supremum norm. We assume the following source condition.

Assumption 5.1. Let us define

$$\Omega_R(L) := \{f \in X \mid R(f) \le L\}$$

and

$$\Omega_\mu(L) := \{f \in X \mid r = \partial R(f) = A_\mu^* w \text{ for } \|w\|_Z \le L\}$$

The ground truth $f^\dagger \in X$ satisfies a *classical source condition* if

$$f^\dagger \in \Omega_R(L_1) \cap \Omega_\mu(L_2) \tag{43}$$

for some $0 < L_1, L_2 < \infty$.

In order to prove verify the assumptions of Theorem 4.10, we rely on concentration estimates for bounded random variables and for sub-Gaussian ones, and in particular to the Hoeffding's inequality, which can be stated as follows.

Proposition 5.2 (Hoeffding's inequality, [19]). *(1) Let $\xi_1, ..., \xi_n$ be zero-mean independent random variables bounded on the interval $[a, b]$ containing zero. It holds that*

$$\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

(2) Let $\xi_1, ..., \xi_n$ be zero-mean independent sub-Gaussian random variables. It holds that

$$\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i\right| \ge t\right) \le 2\exp\left(-\frac{ct^2}{\sum_{i=1}^n \|\xi_i\|^2_{sG}}\right),$$

where $c > 0$ is an absolute constant.

Let us further make the following technical Assumption.

Assumption 5.3. The wavelet basis satisfies

$$\sum_{\lambda=1}^\infty c_{\lambda,q,-s,d} \|A\psi_\lambda\|^q_\infty < \infty,$$

where $c_{\lambda,q,-s,d}$ is defined according to (42).

This requirement can be fulfilled by imposing a sufficiently strong decay of the coefficients $c_{\lambda,q,-s,d}$, or by some regularity assumptions on the operator A. For example, in Section 6.1 we show that it holds true for a particular example of a kernel operator A, associated with a sufficiently smooth kernel.

Proposition 5.4. *Under the Assumptions 5.1 and 5.3 it follows that*

$$\mathbb{E}\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \le C_{q,s,d} L_2^q N^{-\frac{q}{2}} + L_2^2 \beta$$

for

$$C_{q,s,d} = C_q \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d} \left\| A\psi_\lambda \right\|_\infty^q , \tag{44}$$

where $C_q > 0$ depends on q.

Proof. Utilizing the source condition $r^\dagger = A_\mu^* w$ for some $w \in Z$ such that $\|w\|_Z \le L_2$, we have

$$2\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \le \frac{1}{q} \left\| (A_\mu^* - A_\mathbf{u}^* S_\mathbf{u}) w \right\|_{B_q^{-s}}^q + \beta \left\| S_\mathbf{u} w \right\|_{V_N}^2 .$$

The expectation of the second term coincides with $\beta \|w\|_{Y_\mu}^2$ and can be bounded by βL_2^2 due to the continuous embedding of Z to Y_μ. For the first term, we can write

$$\langle (A_\mu^* - A_\mathbf{u}^* S_\mathbf{u}) w, \psi_\lambda \rangle = \frac{1}{N} \sum_{n=1}^{N} \langle (A_\mu^* - A_{u_n}^* S_{u_n}) w, \psi_\lambda \rangle =: \frac{1}{N} \sum_{n=1}^{N} \xi_n^\lambda,$$

where we have set

$$\xi_n^\lambda = \langle (A_\mu^* - A_{u_n}^* S_{u_n}) w, \psi_\lambda \rangle = \langle w, A_\mu \psi_\lambda \rangle_{Y_\mu} - \langle w(u_n), (A\psi_\lambda)(u_n) \rangle_V$$

It follows that random variables ξ_n^λ are zero-mean and also i.i.d. since the design points u_n are assumed to be i.i.d.. Furthermore, due to Assumption 2.1 and by applying Cauchy–Schwarz inequality we have that

$$\langle w(u), (A\psi_\lambda)(u) \rangle_V \le \|w\|_\infty \left\| A\psi_\lambda \right\|_\infty$$

for any $u \in U$. Therefore, for each λ the random variables ξ_n^λ, $i = n, ..., N$, are bounded uniformly according to

$$\langle w, A\psi_\lambda \rangle_{Y_\mu} - L_2 \left\| A\psi_\lambda \right\|_\infty \le \xi_i^\lambda \le \langle w, A\psi_\lambda \rangle_{Y_\mu} + L_2 \left\| A\psi_\lambda \right\|_\infty .$$

Now, by applying the first Hoeffding's inequality it follows that

$$\begin{aligned}
\mathbb{E} R^\star((A_\mu^* - A_\mathbf{u}^* S_\mathbf{u}) w) &= \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d} N^{-q} \mathbb{E} \left| \sum_{n=1}^{N} \xi_n^\lambda \right|^q \\
&= \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d} N^{-q} \int_0^\infty t^{q-1} \mathbb{P} \left(\left| \sum_{n=1}^{N} \xi_n^\lambda \right| > t \right) dt \\
&\le 2 \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d} N^{-q} \int_0^\infty t^{q-1} \exp \left(- \frac{t^2}{2 N L_2^2 \left\| A\psi_\lambda \right\|_\infty^2} \right) dt \\
&= 2 \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d} N^{-\frac{q}{2}} L_2^q \left\| A\psi_\lambda \right\|_\infty^q \int_0^\infty s^{q-1} \exp \left(-\frac{1}{2} s^2 \right) ds \\
&= C_{q,s,d} L_2 N^{-\frac{q}{2}},
\end{aligned}$$

where we applied a change of variables. Above, $c_{\lambda,q,-s,d}$ is defined according to (42) and the constant $C_{q,s,d}$ is given by (44) and is bounded due to Assumption 5.3. Above, $C_q > 0$ is only dependent on q.

□

Proposition 5.5. *Under the above Assumptions, it follows that*

$$\mathbb{E}R^\star(A^*_{\mathbf{u}}\epsilon_N) \leq \widetilde{C}_{q,s,d}N^{-q/2},$$

where the constant is given by

$$\widetilde{C}_{q,s,d} = \widetilde{C}_q \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d} \left\| A\psi_\lambda \right\|_\infty^q \tag{45}$$

and $\widetilde{C}_q$ is only dependent on q.

Proof. Similar to the previous proposition, we write

$$\langle A^*_{\mathbf{u}}\epsilon_N, \psi_\lambda\rangle = \frac{1}{N}\sum_{n=1}^{N}\langle \epsilon_N^n, A_{u_n}\psi_\lambda\rangle_V =: \frac{1}{N}\sum_{n=1}^{N}\tilde{\xi}_n^\lambda.$$

The random variables $\tilde{\xi}_n^\lambda = \langle \epsilon_N^n, A_{u_n}\psi_\lambda\rangle_V$ are independent and zero-mean, since ϵ_N^n is zero-mean and independent of u_n. Moreover, by Assumption 2.1 and by (6) it follows that each $\tilde{\xi}_n^\lambda$ is also sub-Gaussian: indeed,

$$\begin{aligned}
\left\|\tilde{\xi}_n^\lambda\right\|_{\mathrm{sG}} &= \inf\left\{t > 0 \mid \mathbb{E}\exp\left(\frac{(\tilde{\xi}_n^\lambda)^2}{t^2}\right) \leq 2\right\} \\
&\leq \inf\left\{t > 0 \mid \mathbb{E}\exp\left(\frac{\|A_{u_n}\psi_\lambda\|_V^2\|\epsilon_N^n\|_V^2}{t^2}\right) \leq 2\right\} \\
&= \|A_{u_n}\psi_\lambda\|_V \inf\left\{\tilde{t} > 0 \mid \mathbb{E}\exp\left(\frac{\|\epsilon_N^n\|_V^2}{\tilde{t}^2}\right) \leq 2\right\} \\
&\leq \|A\psi_\lambda\|_\infty \left\|\|\epsilon_N^n\|_V\right\|_{\mathrm{sG}}.
\end{aligned}$$

As a consequence, by applying the second Hoeffding's inequality we obtain

$$\begin{aligned}
\mathbb{E}R^\star(A^*_{\mathbf{u}}\epsilon_N) &= \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d}N^{-q}\mathbb{E}\left|\sum_{n=1}^{N}\tilde{\xi}_n^\lambda\right|^q \\
&= \sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d}N^{-q}\int_0^\infty t^{q-1}\mathbb{P}\left(\left|\sum_{n=1}^{N}\tilde{\xi}_n^\lambda\right| > t\right)dt \\
&\leq 2\sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d}N^{-q}\int_0^\infty t^{q-1}\exp\left(-\frac{Ct^2}{N\left\|A\psi_\lambda\right\|_\infty^2}\right)dt \\
&\leq 2\sum_{\lambda=1}^{\infty} c_{\lambda,q,-s,d}C^{-\frac{q}{2}}N^{-\frac{q}{2}}\left\|A\psi_\lambda\right\|_\infty^q\int_0^\infty s^{q-1}\exp\left(-\frac{1}{2}s^2\right)ds \\
&= \widetilde{C}_{q,s,d}N^{-\frac{q}{2}},
\end{aligned}$$

where the constant C combines the effect of the absolute constant in Proposition 5.2 and the uniform bound on $\|\epsilon_N^n\|_{sG}$. Moreover, the constant $\widetilde{C}_{q,s,d}$ is given by (45) and is finite due to Assumption 5.3. $\square$

By applying Propositions 5.4 and 5.5 to Theorem 4.3 we obtain the following result.

Corollary 5.6. *Consider the Besov regularizer of* (41) *applied to the direct problem introduced in equations* (1) *and* (2). *Suppose Assumptions 2.1, 5.1 and 5.3 hold.*

(1) Standard estimate: if $\frac{\delta^2}{N} \to 0$ *and* $N\delta \to \infty$,

$$\mathbb{E} D_R(f_{\alpha,N}^\delta, f^\dagger) \lesssim L_1^{\frac{2-p}{3p}} L_2^{\frac{4}{3}} \left(\frac{\delta^2}{N}\right)^{\frac{1}{3}} \quad \textit{for} \quad \alpha \simeq L_1^{\frac{2-p}{3p}} L_2^{-\frac{2}{3}} \left(\frac{\delta^2}{N}\right)^{\frac{1}{3}}; \tag{46}$$

if instead $N\delta$ *is bounded, the optimal rate is* N^{-1}, *associated with the choice* $\alpha \simeq N^{-1}$.

(2) Alternative estimate: if $\delta \to 0$ *and* $N\delta \to \infty$,

$$\mathbb{E} D_R(f_{\alpha,N}^\delta, f^\dagger) \lesssim L_2 \delta \quad \textit{for} \quad \alpha \simeq L_2^{-1}\delta; \tag{47}$$

if instead $N\delta$ *is bounded, the optimal rate is* N^{-1}, *associated with the choice* $\alpha \simeq N^{-1}$.

Proof. Reflecting the results of Propositions 5.4 and 5.5 according to the notation of Theorems 4.10 and 4.11, we have

$$Q = \frac{q}{2}, \quad D_1 = L_2^2, \quad D_2 = C_{q,s,d} L_2^q, \quad D_3 = \widetilde{C}_{q,s,d} \quad \text{and} \quad L = L_1.$$

Substituting such terms in the statements of Theorems 4.10 and 4.11 and without tracking the constants depending only on p, s, d, which immediately yields the claim. $\square$

5.2 Utilizing Hilbert space embeddings

Let us consider how Hilbert space embeddings of X can be utilized in deriving convergence rates for the symmetric Bregman distance. Suppose that the Banach space X can be embedded continuously to some Hilbert space X_0 and $\|f\|_{X_0} \le \|f\|_X$ for all $f \in X_0$. Due to the embedding property we also have that

$$R^\star(g) \le \frac{1}{q} \|g\|_{X_0}^q. \tag{48}$$

Below, we identify elements of X and X_0 in X^* through the following dependency

$$X \subset X_0 = X_0^* \subset X^*.$$

If the embedding is suitably tight, we can derive useful convergence rates as demonstrated in the following results.

Proposition 5.7. *Suppose* $r^\dagger = B_\mu^s w \in X_0$ *for some* $s \in (0, \frac{1}{2})$. *We have*

$$\mathscr{R}(\beta, \mathbf{u}; f^\dagger) \le \widehat{C}_{p,s} \beta^{rs} \|w\|_{X_0}^r + \frac{1}{q} \left\|(B_\mu^s - B_\mathbf{u}^s)w)\right\|_{X_0}^q,$$

where the constant $\widehat{C}_{p,s}$ *depends on* p *and* s *and*

$$r = \frac{p}{p-1+s(2-p)}. \tag{49}$$

Proof. The Fenchel dual of $E_{\beta,N}$ is given by

$$F_\beta(v;\mathbf{u},f^\dagger) = \frac{1}{2\beta}\|A_\mathbf{u} v\|_{V_N}^2 - \langle r^\dagger, v\rangle_{X^*\times X} + R(v)$$

and, therefore, the Fenchel duality Theorem yields

$$\mathscr{R}(\beta,\mathbf{u};f^\dagger) = -\inf_{v\in X} F_\alpha(v;\mathbf{u},f^\dagger).$$

Due to the embedding property and our Assumption on $r^\dagger$, we have a lower bound

$$\begin{aligned} F_\beta(v;\mathbf{u},f^\dagger) &= \frac{1}{2\beta}\left\|B_\mathbf{u}^{\frac{1}{2}} v\right\|_{X_0}^2 - \langle w, B_\mathbf{u}^s v\rangle_{X_0} - \langle (B_\mu^s - B_\mathbf{u}^s)w, v\rangle_{X^*\times X} + R(v) \\ &\geq \frac{1}{2\beta}\left\|B_\mathbf{u}^{\frac{1}{2}} v\right\|_{X_0}^2 - \|w\|_{X_0}\|B_\mathbf{u}^s v\|_{X_0} - c_p R^\star((B_\mu^s - B_\mathbf{u}^s)w) + \frac{1}{2p}\|v\|_{X_0}^p, \end{aligned}$$

where we applied the generalized Young's inequality. Interpolation of the norms yields

$$\|B_\mathbf{u}^s v\|_{X_0} \leq \left\|B_\mathbf{u}^{\frac{1}{2}} v\right\|_{X_0}^{2s}\|v\|_{X_0}^{1-2s}$$

and by Young's inequality we obtain

$$\|w\|_{X_0}\left\|B_\mathbf{u}^{\frac{1}{2}} v\right\|_{X_0}^{2s}\|v\|_{X_0}^{1-2s} \leq \frac{1}{2\beta}\left\|B_\mathbf{u}^{\frac{1}{2}} v\right\|_{X_0}^2 + \frac{1}{2p}\|v\|_{X_0}^p + \widehat{C}_{p,s}\beta^{rs}\|w\|_{X_0}^r,$$

where the constant $\widehat{C}_{p,s}$ depends on p and s, and r is defined by (49). In conclusion, we obtain

$$F_\beta(v;\mathbf{u},f^\dagger) \geq -\widehat{C}_{p,s}\beta^{rs}\|w\|_{X_0}^r - c_p R^\star((B_\mu^s - B_\mathbf{u}^s)w)$$

which yields the claim. □

Since we also have

$$R^\star(A_\mathbf{u}^*\epsilon_N) \leq \frac{1}{q}\|A_\mathbf{u}^*\epsilon_N\|_{X_0}^q,$$

it follows that we can bound $D_R(f_{\alpha,N}^\delta, f^\dagger)$ in cases (i) and (ii) of Theorem 4.3 involving the random terms $\|B_\mu^s - B_\mathbf{u}^s\|_{\mathcal{L}(X_0)}^q$ and $\|A_\mathbf{u}^*\epsilon_N\|_{X_0}^q$. Therefore, if the spectral properties of $B_\mathbf{u}$ are well-understood on X_0, Propositions 5.2 and 5.5 in [5] yield probabilistic bounds on the symmetric Bregman distance, when Theorem 4.3 and, in particular, the concentration Assumption in inequality (35) is generalized for arbitrary power of β. This generalization is technical and outside of the scope of this paper.

6 Random angle X-ray tomography

As an application of our theory, we study the case when the operator A is the semidiscrete Radon transform and we perform random sampling of the imaging angles.

6.1 Semidiscrete Radon transform

Before introducing the semidiscrete Radon transform, we start by recalling the classical definition of Radon transform $\mathcal{R}$:

$$\mathcal{R}f(\theta, s) = \int_{\mathbb{R}} f(s\theta + t\theta^{\perp})dt \qquad \theta \in S^1, s \in \mathbb{R}.$$

When considering the operator $\mathcal{R}$ acting on a function $f \in X = \{g \in L^2(\Omega) : \text{supp}(g) \subset \overline{\Omega}\}$, with $\Omega \subset \mathbb{R}^2$ bounded, then $\mathcal{R}f$ (the so-called *sinogram*) belongs to the space $L^2([0, 2\pi) \times (-\bar{s}, \bar{s}))$ for a suitable $\bar{s} > 0$ depending on Ω. We aim at defining the sampling operator as a function associating an angle $\theta \in U = [0, 2\pi)$ to the sinogram related to that direction, namely, $\mathcal{R}(\theta) = \mathcal{R}(\theta, \cdot) \in L^2(-\bar{s}, \bar{s})$. Unfortunately, the sinogram space $L^2([0, 2\pi) \times (-\bar{s}, \bar{s})) \cong L^2(U; L^2(-\bar{s}, \bar{s}))$ does not show sufficient regularity to perform pointwise evaluations with respect to the angles.

One way to overcome this difficulty is to rely on a semidiscrete version of the Radon transform. In particular, we set the variable s in a discrete space, which corresponds to modeling the X-ray attenuation measurements performed with a finite-accuracy detector, consisting of N_{dtc} cells. To this end, we introduce a uniform partition $\{I_1, \ldots, I_{N_{\text{dtc}}}\}$ of the interval $(-\bar{s}, \bar{s})$, where we denote by s_j the midpoint of each interval I_j and take a continuous positive function ρ of compact support within $(-1, 1)$ such that $\int_{-1}^{1} \rho = 1$. The semidiscrete Radon transform is a function $A : X \to L^2([0, 2\pi); \mathbb{R}^{N_{\text{dtc}}})$ such that, for any $f \in X$ and $\theta \in [0, 2\pi)$, each component of the vector $Af(\theta) \in \mathbb{R}^{N_{\text{dtc}}}$ can be written as

$$[(Af)(\theta)]_j = \int_{I_j} \mathcal{R}f(\theta, s)\rho\left(\frac{s - s_j}{|I_j|}\right) ds = \int_{I_j} \int_{\mathbb{R}} f(s\theta + t\theta^{\perp})\rho\left(\frac{s - s_j}{|I_j|}\right) dtds. \tag{50}$$

Notice carefully that, according to the formalism of Section 2, $X = \{f \in L^2(\Omega) : \text{supp}(f) \subset \overline{\Omega}\}$, $Y = L^2(U; V)$, $U = S^1 \cong [0, 2\pi)$ and $V = \mathbb{R}^{N_{\text{dtc}}}$.

We observe that each component of $Af(\theta)$ is a suitable average of $\mathcal{R}f(\theta, s)$ in a subinterval I_j. By the change of variables $x = s\theta + t\theta^{\perp}$ in equation (50) we observe that

$$[(Af)(\theta)]_j = \int_{\mathbb{R}^2} f(x)\rho_j(x, \theta)dx,$$

being $\rho_j(x, \theta) = \rho\left(\frac{x \cdot \theta - s_j}{|I_j|}\right)$. As a consequence, from the continuity of ρ we can deduce that for any $f \in X$ each component of $Af(\theta)$ is a continuous function of θ, hence we can consider $A : X \to Z$ being $Z = \mathcal{C}(U; V)$ and the sampled operator $A_\theta : X \to V$ is well defined for every $\theta \in U$. Moreover, the following bound holds uniformly in θ:

$$\|A_\theta f\|_V^2 = \|Af(\theta)\|_V^2 = \sum_{j=1}^{N_{\text{dtc}}} \left| \int_{\Omega} f(x)\rho_j(x, \theta)dx \right|^2 \leq N_{\text{dtc}}|\Omega| \, \|f\|_{L^2(\Omega)}^2 \, \|\rho\|_\infty^2 ,$$

and therefore we conclude that A is a bounded operator from $L^2(\Omega)$ to Z. We now verify that the semidiscrete Radon transform satisfies Assumption 5.3 for any choice of wavelet basis $\{\psi_\lambda\}$ and sufficiently regular Besov space B_p^s.

Proposition 6.1. *Let $\Omega \subset \mathbb{R}^d$ and $X = B_p^s(\Omega)$, with $1 < p \leq 2$ and s such that $\frac{s}{d} \geq \frac{1}{p} - \frac{1}{2}$. Let $\{\psi_\lambda\}$ be an orthonormal basis of $L^2(\Omega)$. Then, the semidiscrete Radon transform satisfies Assumption 5.3.*

Proof. Since, by hypothesis, $q \geq 2$ and $c_{\lambda,q,-s,d} \leq 1$, it is enough to prove that

$$\sum_{\lambda=1}^{\infty} \|A\psi_\lambda\|_\infty^2 < \infty.$$

By Sobolev embedding, for any $f \in L^2(\Omega)$,

$$\|Af\|_\infty^2 = \|Af\|^2_{C((0,2\pi);\mathbb{R}^{N_{dtc}})} \leq C_S \|Af\|^2_{H^1((0,2\pi);\mathbb{R}^{N_{dtc}})} = C_S \sum_{j=1}^{N_{dtc}} \|h_j\|^2_{H^1(0,2\pi)},$$

being $h_j(\theta) = \int_\Omega f(x)\rho_j(x,\theta)dx$. Denoting by $\langle \cdot, \cdot \rangle$ the scalar product in $L^2(\Omega)$ have that

$$\|h_j\|^2_{H^1(0,2\pi)} = \int_0^{2\pi} \langle f, \rho_j(\cdot,\theta)\rangle^2 d\theta + \int_0^{2\pi} \langle f, \nabla_\theta \rho_j(\cdot,\theta)\rangle^2 d\theta.$$

Therefore, by Parseval's identity,

$$\begin{aligned}
\sum_{\lambda=1}^{\infty} \|A\psi_\lambda\|_\infty^2 &\leq C_S \sum_{\lambda=1}^{\infty} \sum_{j=1}^{N_{dtc}} \left(\int_0^{2\pi} \langle \rho_j(\cdot,\theta), \psi_\lambda \rangle^2 d\theta + \int_0^{2\pi} \langle \nabla_\theta \rho_j(\cdot,\theta), \psi_\lambda \rangle^2 d\theta \right) \\
&= C_S \sum_{j=1}^{N_{dtc}} \int_0^{2\pi} \left(\sum_{\lambda=1}^{\infty} \langle \rho_j(\cdot,\theta), \psi_\lambda \rangle^2 + \sum_{\lambda=1}^{\infty} \langle \nabla_\theta \rho_j(\cdot,\theta), \psi_\lambda \rangle^2 \right) \\
&= C_S \sum_{j=1}^{N_{dtc}} \int_0^{2\pi} \left(\|\rho_j(\cdot,\theta)\|^2_{L^2} + \|\nabla_\theta \rho_j(\cdot,\theta)\|^2_{L^2} \right) \leq C(\Omega, \|\rho\|_{C^1}, \{|I_j|\}, N_{dtc}).
\end{aligned}$$

□

6.2 Discretization

In order to perform numerical simulation, we now introduce a fully discretized version of the sampled Radon transform. To this end, we replace the functional space X with $\mathbb{R}^{N_{\text{pxl}}}$ and consider the following discrete model:

$$\boldsymbol{g}_N^\delta = \boldsymbol{g}_N^\dagger + \delta \boldsymbol{\epsilon}_N = \boldsymbol{A}_{\boldsymbol{\theta}} \boldsymbol{f}^\dagger + \delta \boldsymbol{\epsilon}_N \tag{51}$$

where $\boldsymbol{f}^\dagger \in \mathbb{R}^{N_{\text{pxl}}}$ denotes the (unknown) discrete and vectorized image, $\boldsymbol{A}_{\boldsymbol{\theta}} \in \mathbb{R}^{N_{\text{dtc}}N \times N_{\text{pxl}}}$ represents the sampled version of the Radon operator corresponding to the N randomly sampled angles $\boldsymbol{\theta}$, $\boldsymbol{g}_N^\dagger \in \mathbb{R}^{N_{\text{dtc}}N}$ is the subsampled sinogram and $\boldsymbol{\epsilon}_N \in \mathbb{R}^{N_{\text{dtc}}N}$ is the noise. In the implementation, we consider a normal distribution for the noise vector, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N_{\text{dtc}}N})$, where $\boldsymbol{I}_{N_{\text{dtc}}N}$ is the identity matrix in $\mathbb{R}^{N_{\text{dtc}}N \times N_{\text{dtc}}N}$. A practical example is depicted in Figure 1.

The numerical experiments are conducted in the presence of the following regularization term:

$$\boldsymbol{R}(\boldsymbol{f}) = \frac{1}{p} \|\boldsymbol{W}\boldsymbol{f}\|_p^p \tag{52}$$

where $1 < p \leq 2$ and $\boldsymbol{W} \in \mathbb{R}^{N_{\text{pxl}} \times N_{\text{pxl}}}$ is an orthogonal matrix. Notice that this expression allows to consider two scenarios of interest:

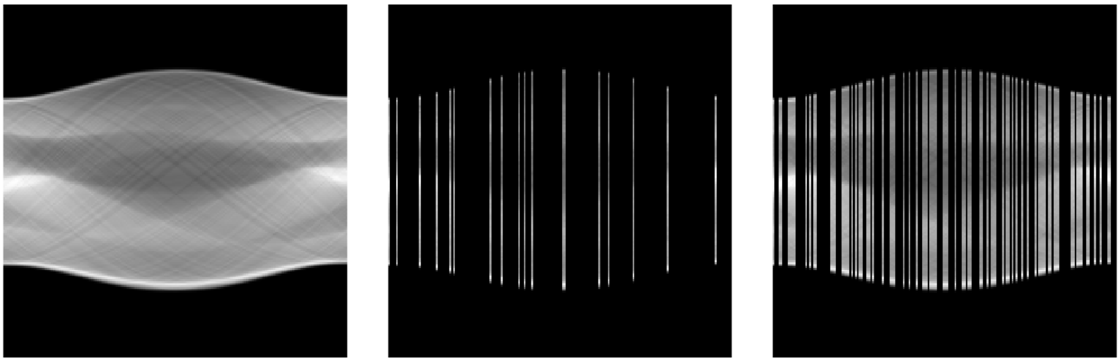


Figure 1: (a) Fully sampled sinogram. (b) Randomly subsampled sinogram with $N = 18$ projection angles. (c) Randomly subsampled sinogram with $N = 81$ projection angles. In each case, the random angles are sampled using Matlab's `rand`, and therefore are identically distributed and independent.

i) if $p = 2$ and $\boldsymbol{W} = \boldsymbol{I}_{N_{\text{pxl}}}$, the identity matrix in $\mathbb{R}^{N_{\text{pxl}} \times N_{\text{pxl}}}$, then (52) reduces to the standard Tikhonov regularization, analyzed in Subsection 4.2;

ii) if $1 < p < 2$ and $\boldsymbol{W}$ is the matrix representation of an orthonormal wavelet transform, then (52) represents a Besov norm. In particular, according to (41), $\boldsymbol{R}(\boldsymbol{f})$ is equivalent to the $B^s_p(\Omega)$ norm, provided that $s = d\left(\frac{1}{p} - \frac{1}{2}\right)$.

Finally, the discrete counterpart of (3) reads as:

$$\boldsymbol{f}^{\delta}_{\alpha,N} = \operatorname*{arg\,min}_{\boldsymbol{f} \in \mathbb{R}^{N_{\text{pxl}}}} \left\{ \frac{1}{2N} \left\| \boldsymbol{A}_{\boldsymbol{\theta}} \boldsymbol{f} - \boldsymbol{g}^{\delta}_{N} \right\|_2^2 + \alpha \boldsymbol{R}(\boldsymbol{f}) \right\}. \tag{53}$$

6.3 Proximal gradient descent algorithm

To solve the minimization problem (53), we use a proximal gradient descent (PGD) algorithm, adapting the forward-backward algorithm reported in [10, Algorithm 10.3]. In particular, by denoting $\Phi(\boldsymbol{f}) = \frac{1}{p} \|\boldsymbol{f}\|_p^p$, the $(k+1)$-th iteration of PGD for the minimization of (53) is given by:

$$\boldsymbol{f}^{(k+1)} = \boldsymbol{W}^{\mathrm{T}} \operatorname{prox}_{\tau_k \alpha \Phi} \left(\boldsymbol{W} \left(\boldsymbol{f}^{(k)} - \frac{\tau_k}{N} \boldsymbol{A}^{\mathrm{T}}_{\boldsymbol{\theta}} (\boldsymbol{A}_{\boldsymbol{\theta}} \boldsymbol{f}^{(k)} - \boldsymbol{g}^{\delta}_{N}) \right) \right) \tag{54}$$

where

$$\operatorname{prox}_{\alpha\tau\Phi}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{z} \in \mathbb{R}^{N_{\text{pxl}}}} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \alpha\tau\Phi(\boldsymbol{z}) \right\} \tag{55}$$

is the proximal operator of Φ and τ_k is a suitable step length, which we update according to the Barzilai-Borwein rule [1].

The expression of Φ allows to provide a more explicit formula for the associated proximal operator. In particular, since for $1 < p \leq 2$ the functional in (55) is differential and convex, by first-order optimality condition it holds that if $\bar{\boldsymbol{z}} \in \mathbb{R}^{N_{\text{pxl}}}$ is such that $\bar{\boldsymbol{z}} = \operatorname{prox}_{\alpha\tau\Phi}(\boldsymbol{x})$, then

$\bar{\boldsymbol{z}} - \boldsymbol{x} + \alpha\tau\nabla\Phi(\boldsymbol{z}) = 0$. Moreover, $\nabla\Phi(\boldsymbol{z}) = \boldsymbol{z}^{[p-1]}$, where $\boldsymbol{x}^{[n]}$ denotes the component-wise signed n-th power:

$$(\boldsymbol{x}^{[n]})_i = \text{sign}(x_i)|x_i|^n. \tag{56}$$

As a result, the optimality condition satisfied by $\bar{\boldsymbol{z}}$ reads as follows:

$$\bar{z}_i + \alpha\tau|\bar{z}_i|^{p-1}\text{sign}(\bar{z}_i) - x_i = 0 \qquad \forall i = 1, \dots, N_{\text{pxl}}, \tag{57}$$

where all the components are decoupled. Notice that $\text{sign}(x_i) = \text{sign}(\bar{z}_i)$, and so the solution of equation (57) is $\bar{z}_i = \text{sign}(x_i)z_i$ where z_i is the positive solution of

$$z_i + \alpha\tau z_i^{p-1} - |x_i| = 0. \tag{58}$$

Therefore, for any choice of $p \in (1, 2)$ the proximal $\bar{\boldsymbol{z}}$ can be efficiently computed by numerically solving N_{pxl} decoupled equations. Additionally, we remark that for the choice $p = 3/2$ and $p = 4/3$ (and, in principle $p = 5/4$) the solution of equation (58) has an explicit, analytic expression given by the formula for the solution of the second, third and fourth degree algebraic equations, respectively. For this reason, without loss of generality, we implement the cases $p = 3/2$ and $p = 4/3$.

6.4 Numerical Experiments

In the following, we present and discuss our numerical experiments. Computations were implemented with Matlab R2021a, running on a laptop with 16GB RAM and Apple M1 chip.

The aim is to verify the expected convergence rates proven in Theorem 4.6 for the Tikhonov case and in Corollary 5.6 for the Besov regularization. We test inequalities (31) and (46) in the following two scenarios:

- Fixed noise, i.e., $\delta > 0$ constant. Since $\delta N \to \infty$, we take $\alpha \simeq \delta^{2/3}N^{-1/3}$ and in particular we choose $\delta = c_\delta$ and $\alpha = c_\alpha/N^{1/3}$;
- Decreasing noise, e.g., $\delta \simeq N^{-1}$. In this case, the optimal parameter choice is $\alpha \simeq N^{-1}$: therefore, we choose $\delta = c_\delta N^{-1}$ and $\alpha = c_\alpha N^{-1}$.

The positive constants c_δ and c_α are specified in the following (see Table 1 and Subsection 6.4.2).

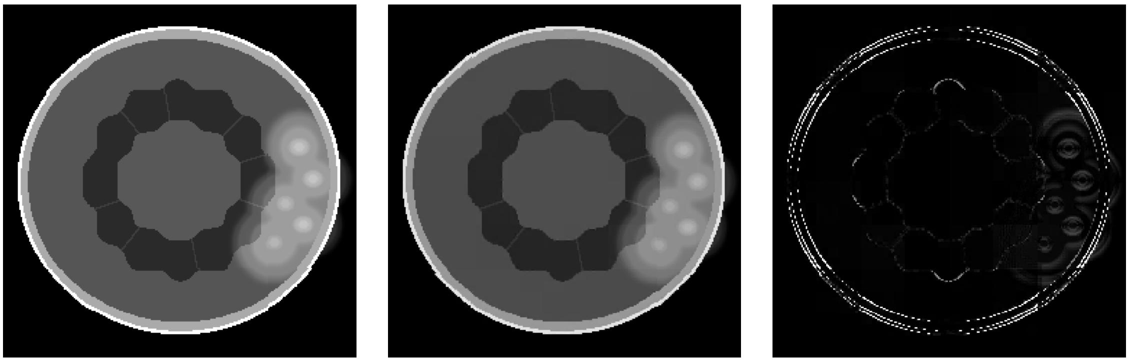
Notice that in the fixed noise regime, the only valid bounds are the standard estimates (31) and (46), whereas in the decreasing noise regime the alternative estimates (34) and (47) are also valid, although since $\delta \simeq N^{-1}$ they actually coincide with the standard ones.

6.4.1 Implementing the source condition

All the numerical tests are run on images $\boldsymbol{f}^\dagger$ satisfying the source condition (25) in the Tikhonov case and (43) in the Besov case. In the discrete setting, both (25) and (43) can be formulated as:

$$\exists\, \boldsymbol{w} \in \mathbb{R}^{N_{\text{dtc}}N_\theta} \qquad \text{s.t.} \quad \boldsymbol{W}^{\mathrm{T}}(\boldsymbol{W}\boldsymbol{f}^\dagger)^{[p-1]} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{w} \tag{59}$$

where $\boldsymbol{A} \in \mathbb{R}^{N_{\text{dtc}}N_\theta \times N_{\text{pxl}}}$ with a fixed $N_\theta \gg N$ is a matrix representing the Radon transform acting from $\mathbb{R}^{N_{\text{pxl}}}$ to $\mathbb{R}^{N_{\text{dtc}}N_\theta}$, a sufficiently refined discretization of the space Y of full sinograms. As in (56), $\boldsymbol{f}^{[p]}$ is the component-wise signed power. In practice, a generic phantom of interest $\boldsymbol{f}_0$ does not necessarily satisfy (59). Therefore, in order to guarantee that the



(a) $\boldsymbol{f}_0$ (b) $\boldsymbol{f}^\dagger$ (c) $\boldsymbol{f}^\dagger - \boldsymbol{f}_0$

Figure 2: (a) Original phantom. (b) Phantom satisfying the source condition in Assumption 5.1 with $p = 3/2$. (c) Difference between (a) and (b), with relative error 5%.

test images satisfy (59), we first determine a vector $\boldsymbol{w} \in \mathbb{R}^{N_{\mathrm{dtc}} N_\theta}$ solution of the regularized problem

$$\boldsymbol{w} = \underset{\widetilde{\boldsymbol{w}} \in \mathbb{R}^{N_{\mathrm{dtc}} N_\theta}}{\arg\min} \left\{ \frac{1}{2} \left\| \boldsymbol{A}^{\mathrm{T}} \widetilde{\boldsymbol{w}} - \boldsymbol{W}^{\mathrm{T}} (\boldsymbol{W} \boldsymbol{f}_0)^{[p-1]} \right\|_2^2 + \lambda_{SC} \left\| \widetilde{\boldsymbol{w}} \right\|_2^2 \right\}, \tag{60}$$

for a suitable $\lambda_{SC} > 0$. Regularization is needed since the inverse problem to determine $\boldsymbol{w}$ as in (59) is ill-posed. Then, we compute $\boldsymbol{f}^\dagger = \boldsymbol{W}^{\mathrm{T}} (\boldsymbol{W} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{w})^{[1/(p-1)]}$: as a result, $\boldsymbol{f}^\dagger$ satisfies the source condition associated with $\boldsymbol{w}$, and $\left\| \boldsymbol{f}^\dagger - \boldsymbol{f}_0 \right\|_2$ is expected to be small. An example for $p = 3/2$ is given in Figure 2. Notice that for $p = 2$ the source condition (59) reduces to

$$\exists\, \boldsymbol{w} \in \mathbb{R}^{N_{\mathrm{dtc}} N_\theta} \qquad \text{s.t.} \quad \boldsymbol{f}^\dagger = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{w}. \tag{61}$$

6.4.2 Numerical setup

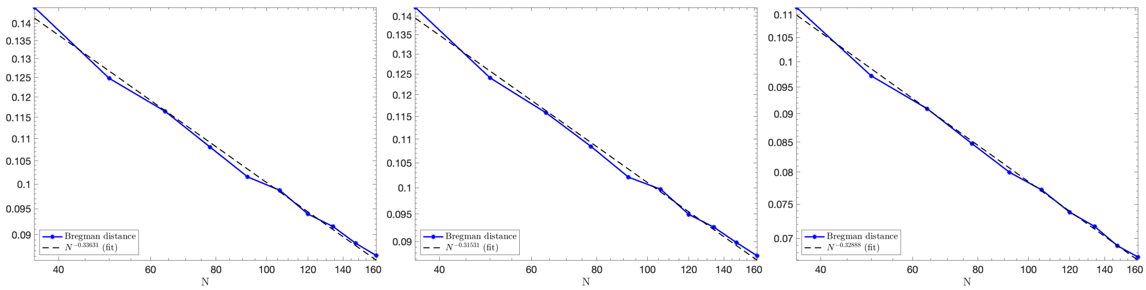
We use the Plant phantom, available on GitHub [18] (see also Figure 2(a)). The size of the phantom is 128×128, hence $N_{\mathrm{pxl}} = 128^2$. In order to generate a phantom satisfying the source condition, we follow the strategy proposed in Subsection 6.4.1, employing the operator $\boldsymbol{A}$ with $N_\theta = 360$. The forward operator and its adjoint are implemented using Matlab's `radon` and `iradon` routines, with suitable normalization. Reconstructions are computed with $N = 36, 50, 64, \ldots, 162$ in the interval $[0, \pi)$. To gain intuition on the subsampling rate associated with this choice, the endpoints $N_0 = 36$ and $N_1 = 162$ correspond to 10% and 45% of $N_\theta = 360$, which is typically used to provide a sufficiently fine discretization of the full sinogram. In each scenario, N random angles are sampled using Matlab's `rand` (thus ensuring that the random points are stochastically independent). The Gaussian noise $\boldsymbol{\epsilon}_N$ is created by the command `randn` and the constant c_δ appearing in the expression of the noise level δ is chosen depending on the noise scenario:

- Fixed noise ($\delta = c_\delta$): $c_\delta = 0.01 \left\| \boldsymbol{A} \boldsymbol{f}^\dagger \right\|_\infty$;
- Decreasing noise ($\delta = c_\delta N^{-1}$): $c_\delta = 0.02\, N_0 \left\| \boldsymbol{A} \boldsymbol{f}^\dagger \right\|_\infty$. Therefore, δ ranges between $0.02 \left\| \boldsymbol{A} \boldsymbol{f} \right\|_\infty$ and $0.02 N_0 / N_1 \left\| \boldsymbol{A} \boldsymbol{f} \right\|_\infty \approx 0.005 \left\| \boldsymbol{A} \boldsymbol{f} \right\|_\infty$.

Reconstruction are computed using the PGD algorithm as described in Subsection 6.3. The regularization parameter α depends on the value of c_α which is heuristically determined.

	fixed noise	reducing noise
$p = 3/2$	0.05	0.15
$p = 4/3$	0.04	0.15
$p = 2$	0.13	0.16

Table 1: Optimal values for c_α.



(a) (b) (c)

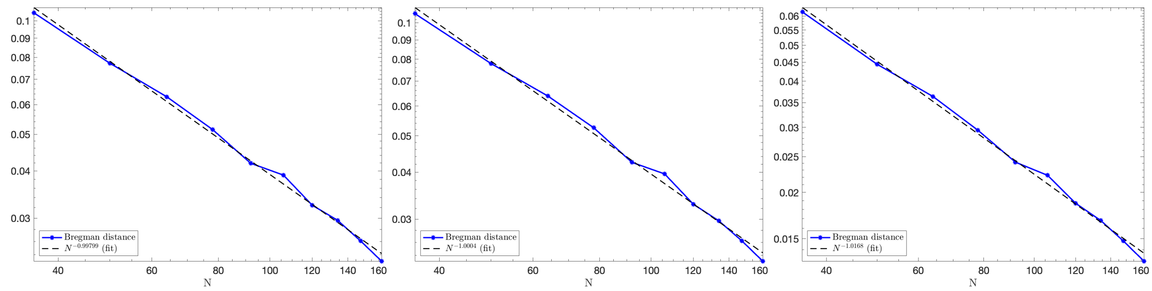
Figure 3: Estimates in the fixed noise case. (a) $p = 3/2$ (b) $p = 4/3$ (c) $p = 2$

Optimal values for c_α are reported in Table 1. The expected values appearing in Theorem 4.6 and in Corollary 5.6 are approximated by sample averages, computed using 30 random realizations. This means that, for each number of angles N, the reconstruction is performed 30 times, each time with a different set of N drawn angles and noise vector.

6.4.3 Numerical results

In Figures 4 and 3 we report the value of the expected Bregman distance $\mathbb{E}D_{\boldsymbol{R}}(\boldsymbol{f}^\delta_{\alpha,N}, \boldsymbol{f}^\dagger)$ as a function of N, both in the reducing noise and fixed noise regimes. We compare three different choices of functional $\boldsymbol{R}$: $p = 3/2$ and $p = 4/3$, associated with the choice of the Haar wavelet transform $\boldsymbol{W}$ (Besov regularization), and $p = 2$ with the identity matrix (Tikhonov regularization). According to Corollary 5.6 and Theorem 4.6, we should expect the same decay of $\mathbb{E}D_{\boldsymbol{R}}(\boldsymbol{f}^\delta_{\alpha,N}, \boldsymbol{f}^\dagger)$, independently of $\boldsymbol{R}$: as $N^{-1/3}$ in the fixed noise one, and as N^{-1} in the reducing noise scenario. We can see in Figures 3 and 4 that the theoretical behaviour is numerically verified.

In order to provide a quantitative assessment, we compare the theoretically predicted decay with the experimental one, which is obtained by computing the best monomial approximation $a_{\text{mono}}(N) = cN^\beta$ of the reported curves. In each plot, the value of the expected Bregman distance is indicated by a blue solid line and its monomial approximation by a black dashed line. We observe a good match with the theoretical previsions, as reviewed in Table 2. The results reported in this Section allow to conclude that, in the example of the discrete Radon transform, the decay of the expected Bregman distance reported in Corollary 5.6 and Theorem 4.6 is verified. We do not attempt to provide an expression for the constants appearing in such inequalities. Moreover, we do not aim at comparing the effectiveness of the three different regularization strategies. For example, it is not worth to compare the values in Figure 3 (a),(b) and (c) among each other, because the object of the plot, $\mathbb{E}D_{\boldsymbol{R}}(\boldsymbol{f}^\delta_{\alpha,N}, \boldsymbol{f}^\dagger)$, is different in each of them: the Bregman distance clearly depends on $\boldsymbol{R}$, but also $\boldsymbol{f}^\dagger$ subtly changes with $\boldsymbol{R}$, according to the proposed strategy to impose the source condition to the



(a) (b) (c)

Figure 4: Estimates in the decreasing noise case. (a) $p = 3/2$ (b) $p = 4/3$ (c) $p = 2$

scenario	theoretical	$p = 3/2$	$p = 4/3$	$p = 2$
reducing noise	-1	-0.99799	-1.0004	-1.0168
fixed noise	$-1/3$	-0.33631	-0.3153	-0.3289

Table 2: Approximate decay β of the expected Bregman distance.

phantom.

7 Conclusions

In this paper we developed a novel convergence study for a linear problem within the statistical inverse learning framework. We assume a regularization scheme with a general convex p-homogeneous penalty functional for $p > 1$ and derive concentration rates of the regularized solution to the ground truth measured in the symmetric Bregman distance induced by the penalty functional. We provide concrete rates for Besov-norm based penalties and observe these rates numerically, for $1 < p \leq 2$, in the case of X-ray tomography with randomly sampled imaging angles.

In the usual framework of statistical inverse learning, the noise level $\delta > 0$ is fixed. Here, we developed estimates also for the asymptotic regime, where the noise is small with respect to the number of design points, i.e., $\delta \simeq N^{-\rho}$ for some $\rho > 1$. More work is needed to clarify conditions, where such small noise estimates become preferable to the standard framework. The identity $Q = q/2$ in Theorem 4.9 as observed with the Besov penalties in Section 5.1 seems natural to the Monte Carlo type approximation error in learning theory. However, it is intriguing to understand if and when faster rates with $Q > q/2$ are possible in the small noise regime.

Finally, the results presented here produce two immediate questions for future studies: first, it would be valuable to understand whether optimal convergence rates can be achieved with the developed framework. Second, arguably the most interesting p-homogenous case $p = 1$ is not considered here. Enabling convergence studies for penalties such as the Total Variation functional is part of future study.

Acknowledgments

TAB was supported by the Academy of Finland through the postdoctoral grant decision number 330522 and by the Royal Society through the Newton International Fellowship grant n. NIF\R1\201695. TAB and LR acknowledge support by the Academy of Finland through the Finnish Centre of Excellence in Inverse Modelling and Imaging 2018-2025, decision number 312339. Also INdAM-GNCS and INdAM-GNAMPA are acknowledged. The work of MB has been supported by ERC via Grant EU FP7 ERC Consolidator Grant 615216 LifeInverse, by the German Ministry of Science and Technology (BMBF) under grant 05M2020 - Deleto, and by the EU under grant 2020 NoMADS - DLV-777826. TH was supported by the Academy of Finland through decision number 326961. LR was supported by the Air Force Office of Scientific Research under award number FA8655-20-1-7027, and acknowledges the support of Fondazione Compagnia di San Paolo.

A Technical Lemmas

Let us record here the technical Lemmas used in Section 4.2. The following concentration result was first shown in [26, Corollary 1].

Proposition A.1. *Let $(Z, \mathcal{B}, \mathbb{P})$ be a probability space and ξ a random variable on Z with values in a real separable Hilbert space $\mathcal{H}$. Assume that there are two positive constants L and σ such that for any $m \geq 2$ we have*

$$\mathbb{E} \|\xi - \mathbb{E}\xi\|_{\mathcal{H}}^m \leq \frac{1}{2} m! \sigma^2 L^{m-2}.$$

If the sample $z_1, ..., z_N$ drawn i.i.d. from Z according to $\mathbb{P}$, then, for any $0 < \eta < 1$ we have

$$\left\| \frac{1}{N} \sum_{j=1}^{N} \xi(z_j) - \mathbb{E}\xi \right\|_{\mathcal{H}} \leq 2 \log(2\eta^{-1}) \left(\frac{L}{N} + \frac{\sigma}{\sqrt{N}} \right)$$

with probability greater than $1 - \eta$.

Proposition A.2 (Cordes inequality [15, 14])**.** *Let T_1, T_2 be two self-adjoint, positive operators on a Hilbert space. Then for any $s \in [0, 1]$ we have*

$$\|T_1^s T_2^s\| \leq \|T_1 T_2\|^s.$$

The following two results are the basis for estimating expectation of the quadratic loss in Section 4.2.

Lemma A.3. *Let X be a nonnegative random variable with $\mathbb{P}\left(X > Z \log^\gamma \left(\frac{k}{\eta}\right)\right) < \eta$ for any $\eta \in (0, 1]$. It follows that*

$$\mathbb{E}X \leq Zk\gamma\Gamma(\gamma).$$

Proof. The result follows from identity $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt$ and changing variables in the probabilistic bound. $\square$

Proposition A.4. *[16, Prop. 1] For any $\beta > 0$ and $\eta \in (0,1]$ we have*

$$\left\|(B_{\mathbf{u}} + \beta I)^{-1}(B_\mu + \beta I)\right\| \leq C\mathcal{B}_N(\beta)\log^2\left(\frac{2}{\eta}\right)$$

for some constant $C > 0$ with probability at least $1-\eta$, where $\mathcal{B}_N$ is given by

$$\mathcal{B}_N(\beta) = 1 + \left(\frac{2}{N\beta} + \sqrt{\frac{\mathcal{N}(\beta)}{N\beta}}\right)^2.$$

References

[1] Jonathan Barzilai and Jonathan M. Borwein, *Two point step size gradient methods*, IMA journal of numerical analysis **8** (1988), 141–8.

[2] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco, *On regularization algorithms in learning theory*, Journal of complexity **23** (2007), no. 1, 52–72.

[3] Martin Benning and Martin Burger, *Modern regularization methods for inverse problems*, Acta Numerica **27** (2018), 1–111.

[4] Nicolai Bissantz, Thorsten Hohage, and Axel Munk, *Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise*, Inverse Problems **20** (2004), no. 6, 1773.

[5] Gilles Blanchard and Nicole Mücke, *Optimal rates for regularization of statistical inverse learning problems*, Foundations of Computational Mathematics **18** (2018), no. 4, 971–1013.

[6] Martin Burger, Tapio Helin, and Hanne Kekkonen, *Large noise in variational regularization*, Transactions of Mathematics and its Applications **2** (2018), no. 1, 1–45.

[7] Martin Burger and Andreas Neubauer, *Error bounds for approximation with neural networks*, Journal of Approximation Theory **112** (2001), no. 2, 235–250.

[8] Martin Burger and Stanley Osher, *Convergence rates of convex variational regularization*, Inverse problems **20** (2004), no. 5, 1411.

[9] Andrea Caponnetto and Ernesto De Vito, *Optimal rates for the regularized least-squares algorithm*, Foundations of Computational Mathematics **7** (2007), no. 3, 331–368.

[10] Patrick L. Combettes and Jean-Christophe Pesquet, *Proximal splitting methods in signal processing*, pp. 185–212, Springer New York, New York, NY, 2011.

[11] Felipe Cucker and Steve Smale, *Best choices for regularization parameters in learning theory: on the bias-variance problem*, Foundations of Computational Mathematics **2** (2002), no. 4, 413–428.

[12] Ingrid Daubechies, Michel Defrise, and Christine De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences **57** (2004), no. 11, 1413–1457.

[13] Ernesto De Vito, Lorenzo Rosasco, and Andrea Caponnetto, *Discretization error analysis for Tikhonov regularization*, Analysis and Applications **4** (2006), no. 01, 81–99.

[14] Jun Ichi Fujii and Masatoshi Fujii, *A norm inequality for operator monotone functions*, Mathematica Japonica **35** (1990), no. 2, 249–252.

[15] Takayuki Furuta, *Norm inequalities equivalent to Löwner-Heinz theorem*, Reviews in Mathematical Physics **1** (1989), no. 1, 135–137.

[16] Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou, *Learning theory of distributed spectral algorithms*, Inverse Problems **33** (2017), no. 7, 074009.

[17] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, CRC press, 2015.

[18] Tommi Heikkilä, *Plant phantom*, https://github.com/tommheik/PlantPhantom, 2020.

[19] Wassily Hoeffding, *Probability inequalities for sums of bounded random variables*, The Collected Works of Wassily Hoeffding, Springer, 1994, pp. 409–426.

[20] L. Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri, *Spectral algorithms for supervised learning*, Neural Computation **20** (2008), no. 7, 1873–1897.

[21] Shuai Lu, Peter Mathé, and Sergei V Pereverzev, *Balancing principle in supervised learning for a general regularization scheme*, Applied and Computational Harmonic Analysis **48** (2020), no. 1, 123–148.

[22] Shahar Mendelson and Joseph Neeman, *Regularization in kernel learning*, The Annals of Statistics **38** (2010), no. 1, 526–565.

[23] Nicole Mücke, *Direct and inverse problems in machine learning*, Doctoral thesis, Universität Potsdam, 2017, p. 159.

[24] Frank Natterer, *The mathematics of computerized tomography*, SIAM, 2001.

[25] Finbarr O'Sullivan, *Convergence characteristics of methods of regularization estimators for nonlinear operator equations*, SIAM Journal on Numerical Analysis **27** (1990), no. 6, 1635–1649.

[26] Iosif F Pinelis and Alexander I Sakhanenko, *Remarks on inequalities for large deviation probabilities*, Theory of Probability & Its Applications **30** (1986), no. 1, 143–148.

[27] Abhishake Rastogi, Gilles Blanchard, and Peter Mathé, *Convergence analysis of Tikhonov regularization for non-linear statistical inverse learning problems*, Electronic Journal of Statistics **14** (2020), no. 2, 2798–2841.

[28] Thomas Schuster, Barbara Kaltenbacher, Bernd Hofmann, and Kamil S Kazimierski, *Regularization methods in Banach spaces*, vol. 10, Walter de Gruyter, 2012.

[29] Steve Smale and Ding-Xuan Zhou, *Shannon sampling II: Connections to learning theory*, Applied and Computational Harmonic Analysis **19** (2005), no. 3, 285–302.

[30] ———, *Learning theory estimates via integral operators and their approximations*, Constructive approximation **26** (2007), no. 2, 153–172.

[31] Ingo Steinwart and Andreas Christmann, *Support vector machines*, Springer Science & Business Media, 2008.

[32] Ingo Steinwart, Don R Hush, Clint Scovel, et al., *Optimal rates for regularized least squares regression.*, COLT, 2009, pp. 79–93.

[33] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[34] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone, *Learning from examples as an inverse problem*, Journal of Machine Learning Research **6** (2005), no. May, 883–904.

[35] Frederic Weidling, Benjamin Sprung, and Thorsten Hohage, *Optimal convergence rates for Tikhonov regularization in Besov spaces*, SIAM Journal on Numerical Analysis **58** (2020), no. 1, 21–47.

[36] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto, *On early stopping in gradient descent learning*, Constructive Approximation **26** (2007), no. 2, 289–315.