Gilles Blanchard | Nicole Mücke
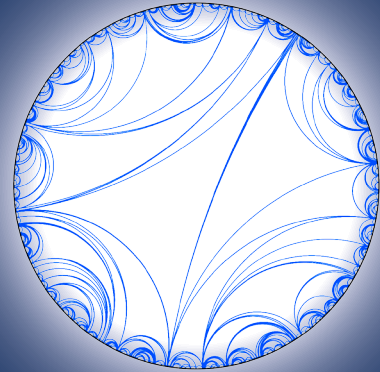
Optimal Rates for Regularization of Statistical Inverse Learning Problems

Gilles Blanchard | Nicole Mücke

Optimal Rates for Regularization of Statistical Inverse Learning Problems

OPTIMAL RATES FOR REGULARIZATION OF STATISTICAL INVERSE LEARNING PROBLEMS

GILLES BLANCHARD AND NICOLE MÜCKE

ABSTRACT. We consider a statistical inverse learning problem, where we observe the image of a function f through a linear operator A at i.i.d. random design points X_i, superposed with an additive noise. The distribution of the design points is unknown and can be very general. We analyze simultaneously the direct (estimation of Af) and the inverse (estimation of f) learning problems. In this general framework, we obtain strong and weak minimax optimal rates of convergence (as the number of observations n grows large) for a large class of spectral regularization methods over regularity classes defined through appropriate source conditions. This improves on or completes previous results obtained in related settings. The optimality of the obtained rates is shown not only in the exponent in n but also in the explicit dependency of the constant factor in the variance of the noise and the radius of the source condition set.

1. INTRODUCTION

1.1. **Setting.** Let A be a known linear operator from a Hilbert space $\mathcal{H}_1$ to a linear space $\mathcal{H}_2$ of real-valued functions over some input space $\mathcal{X}$. In this paper we consider a random and noisy observation scheme of the form

$$(1.1) \qquad Y_i := g(X_i) + \varepsilon_i \,, \quad g = Af \,, \quad i = 1, \ldots, n$$

at i.i.d. data points $X_1, \ldots, X_n$ drawn according to a probability distribution ν on $\mathcal{X}$, where ε_i are independent centered noise variables. More precisely, we assume that the observed data $(X_i, Y_i)_{1 \le i \le n}$ are i.i.d. observations, with $\mathbb{E}[Y_i|X_i] = g(X_i)$, so that the distribution of ε_i may depend on X_i, while satisfying $\mathbb{E}[\varepsilon_i|X_i] = 0$. This is also commonly called a *statistical learning* setting, in the sense that the data (X_i, Y_i) are generated by some external random source and the *learner* aims to infer from the data some reconstruction $\widehat{f}_n$ of f, without having influence on the underlying sampling distribution ν. For this reason we call model (1.1) an *inverse statistical learning problem*. The special case $A = I$ is just non-parametric regression under random design (which we also call the direct problem). Thus, introducing a general A gives a unified approach to the direct and inverse problem.

In the statistical learning context, the relevant notion of convergence and associated reconstruction rates to recover f concern the limit $n \to \infty$. More specifically, let $\widehat{f}_n$ be an estimator of f based on the observed data $(X_i, Y_i)_{1 \le i \le n}$. The usual notion of estimation

---

error in the statistical learning context is the averaged squared loss for the prediction of $g(X)$ at a new independent sample point X:

$$(1.2) \qquad \mathbb{E}_{X \sim \nu}[(g(X) - A\widehat{f}_n(X))^2] = \left\| A(f - \widehat{f}_n) \right\|_{L^2(\nu)}^2 .$$

In this paper, we are interested as well in the *inverse* reconstruction problem, that is, the reconstruction error for f itself in the input space norm, i.e.

$$\left\| f - \widehat{f}_n \right\|_{\mathcal{H}_1}^2 .$$

Estimates in $L^2(\nu)$-norm are standard in the learning context, while estimates in $\mathcal{H}_1$-norm are standard for inverse problems, and our results will present convergence results for a family of norms interpolating between these two. We emphasize that $\left\| A(f - \widehat{f}_n) \right\|_{L^2(\nu)}^2$ as well as $\left\| f - \widehat{f}_n \right\|_{\mathcal{H}_1}^2$ are random variables, depending on the observations. Thus the error rates above can be estimated either in expectation or in probability. In this paper we will present convergence rates for these different criteria, as n tends to infinity, both in expectation (for moments of all orders) and with high probability.

1.2. **Overview of the results.** In this section we present a short, informal overview of the results which will allow a comparison to other existing results in the next section. We start to show that, under appropriate assumptions, we can endow $\mathrm{Im}(A)$ with an appropriate reproducing kernel Hilbert space (RKHS) structure $\mathcal{H}_K$ with reproducing kernel K, such that A is a partial isometry from $\mathcal{H}_1$ onto $\mathcal{H}_K$. Through this partial isometry the initial problem (1.1) can be formally reduced to the problem of estimating the function $g \in \mathcal{H}_K$ by some $\widehat{g}$; control of the error $(g - \widehat{g})$ in $L^2(\nu)$-norm corresponds to the direct (prediction) problem, while control of this difference in $\mathcal{H}_K$-norm is equivalent to the inverse (reconstruction) problem. In particular, the kernel K completely encapsulates the information about the operator A. This equivalence also allows a direct comparison to previous existing results for convergence rates of statistical learning using a RKHS formalism (see next section). Let $L : g \in \mathcal{H}_K \mapsto \int g(x) K(x, .) d\nu(x) \in \mathcal{H}_K$ denote the kernel integral operator associated to K and the sampling measure ν. The rates of convergence presented in this paper will be governed by a *source condition* assumption on g of the form $\|L^r g\| \leq R$ for some constants $r, R > 0$ as well as by the *ill-posedness* of the problem, as measured by an assumed power decay of the eigenvalues of L with exponent $b > 1$. Our main upper bound result establishes that for a broad class of estimators defined via spectral regularization methods, for $s \in [0, \frac{1}{2}]$ it holds both with high probability as well as in the sense of p-th moment expectation that

$$\|L^s(g - \widehat{g}_{\lambda_n})\|_{\mathcal{H}_K} \lesssim R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{(r+s)}{2r+1+1/b}} ,$$

for an appropriate choice of the regularization parameter λ_n. (Note that $s = 0$ corresponds to reconstruction error, and $s = \frac{1}{2}$ to the prediction error i.e. $L^2(\nu)$ norm). Here σ^2 denotes

noise variance (classical Bernstein moment conditions are assumed to hold for the noise.) The symbol $\lesssim$ means that the inequality holds up to a multiplicative constant that can depend on various parameters entering in the assumptions of the result, but not on n, σ, nor R. An important assumption is that the inequality $q \geq r + s$ should hold, where q is the *qualification* of the regularization method, a quantity defined in the classical theory of inverse problems (see Section 2.5 for a precise definition).

This result is complemented by a minimax lower bound which matches the above rate not only in the exponent in n, but also in the precise behavior of the multiplicative constant in function of R and the noise variance σ^2. The obtained lower bounds come in two flavors, which we call weak and a strong asymptotic lower bound (see Section 3).

1.3. **Related work.** The analysis of inverse problems, discretized via (noisy) observations at a finite number of points, has a long history, which we will not attempt to cover in detail here. The introduction of reproducing kernel Hilbert space based methods was a crucial step forward in the end of the 1970s. Early references have focused, mostly, on spline methods on $[0,1]^d$; on observation point designs either deterministic regular, or random with a sampling probability comparable to Lebesgue; and on assumed regularity of the target function in terms of usual differentiability properties. We refer to [30] and references therein for a general overview. An early reference establishing convergence rates in a random design setting for (possibly nonlinear) inverse problems in a setup similar to those delineated above and a Tykhonov-type regularization method is [23]. Analysis of the convergence of fairly general regularization schemes for statistical inverse problems under the *white noise* model were established in [5]. The white noise model is markedly different from the setting considered in the present paper, in particular because it does not involve randomly sampled observation points, though as a general rule one expects a correspondence between optimal convergence rates in both settings.

We henceforth focus our attention on the more recent thread of literature concerning the *statistical learning* setting, whose results are more directly comparable to ours. In this setting, the emphasis is on general input spaces, and "distribution-free" results, which is to say, random sampling whose distribution ν is unknown, quite arbitrary and out of the control of the user. The use of reproducing kernel methods have enjoyed a wide popularity in this context since the 1990s, mainly for the direct learning problem. The connections between (the direct problem of) statistical learning using reproducing kernel methods, and inverse problem methodology, were first noted and studied in [10, 15, 11]. In particular, in [15] it was proposed to use general form regularization methods from the inverse problem literature for kernel-based statistical learning. There is a vast recent literature relating learning to regularization techniques for inverse problems (see [22], [31], [16] to mention just a few), confirming the strong conceptual analogy of certain learning algorithms with regularization algorithms. For example, Tikhonov regularization is known as regularized least-squares algorithm or ridge regression, while Landweber iteration is related to L^2-boosting or gradient descent, see e.g. [32] and [6].

In [11], the more general setting of the random discretization of an inverse problem defined by a Carleman operator is considered. This is essentially the setting we adopt in the

present work. More precisely, we start with the assumption that the map $(f, x) \mapsto (Af)(x)$ is continuous in f and measurable in x, which implies that A can be seen as a Carleman operator from $\mathcal{H}_1$ to $L^2(\nu)$. Moreover, as mentioned in the previous section, we observe that $\text{Im}(A)$ can be endowed with a RKHS structure $\mathcal{H}_K$ such that A is a partial isometry from $\mathcal{H}_1$ onto $\mathcal{H}_K$. While we do not expect this result to considered a novelty, it was not explicitly mentioned in [11] and in our opinion helps cement the equivalence between inverse statistical learning and direct learning with reproducing kernels. In particular, it makes a direct comparison possible between our results and previous results for the direct (kernel) learning problem.

Concerning the history of upper rates of convergence in a RKHS setting, covering number techniques were used in [8] to obtain (non-asymptotic) upper rates. In [10], [25], [26] these techniques were replaced by estimates on integral operators via concentration inequalities, and this is the path we follow in this paper.

We shall now briefly review previous results which are directly comparable to ours: Smale and Zhou [26], Bauer et al. [1], Yao et al. [32], Caponnetto and De Vito [9] and Caponnetto [7]. For convenience, we have tried to condense the most essential points in Table 1. Compared with our more general setting, all of these previous references only consider the special case $A = I$, but assume from the onset that $\mathcal{H}_1$ is a RKHS with given kernel. Thus, in the first column of Table 1, A is the identity and $g = f$, and in the second column $\mathcal{H}_1 = \mathcal{H}_K$. The more complicated form given in Table 1 is the reinterpretation in our setting (see Section 2). The first three references ([26], [1], [32]) do not analyze lower bounds and their upper bounds do not take into account the behaviour of the eigenvalues of the integral operator L corresponding to the assumed RKHS structure. But all three derive estimates on the error both in $L^2(\nu)$-norm and RKHS-norm. Only [1] considers a general class of spectral regularization methods.

The last two papers [9] and [7] obtain fast upper rates (depending on the eigenvalues of L) which are minimax optimal. The estimates, however, are only given in $L^2(\nu)$-norm. Furthermore, only [7] goes beyond Tikhonov regularization to handle a general class of spectral regularization methods. A closer look at Table 1 reveals that in treating general spectral regularization methods, the results of [7] require for certain parameter configurations ($r < 1/2 - 1/2b$) the availablility of additional unlabeled data from the sampling distribution ν. This appears somewhat suboptimal, since this does not reproduce the previously obtained result for Tikhonov in [9] which does not require unlabeled data.

To obtain these finer results ("fast rates" taking into account the spectral structure of L), a crucial technical tool is to consider the *effective dimension* $\mathcal{N}(\lambda) = \text{tr}((L + \lambda)^{-1}L)$, which determines the optimal choice of the regularization parameter. This idea of [9] and [7] is fundamental for our approach, which extends and refines these previous results.

Furthermore, we recall from [9] that the effective dimension $\mathcal{N}(\lambda)$ seems to be just the right parameter to establish an important connection between the operator theoretic and spectral methods and the results obtained via entropy methods (see [12], [28]) since $\mathcal{N}(\lambda)$ encodes via L crucial properties of the marginal distribution ν.

As delineated in Section 1.2, the main question adressed in this paper is that of minimax optimal rates of convergence as n grows to infinity. Our contribution is to improve on

	$\left\|A(\widehat{f}_n - f)\right\|_{L^2(\nu)}$	$\left\|\widehat{f}_n - f\right\|_{\mathcal{H}_K}$	Assumptions (q: qualification)	Method
Smale/ Zhou [26]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	$r \le \frac{1}{2}$	Tikhonov
Bauer et al. [1]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	$r \le q - \frac{1}{2}$	General
Yao et al. [32]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+3}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+\frac{5}{2}}}$	$q = \infty$	Landweber Iteration
Caponnetto, De Vito [9]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+\frac{1}{b}}}$	N/A	$r \le \frac{1}{2}$	Tikhonov
Caponnetto [7]	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+\frac{1}{b}}}$	N/A	$r \le q - \frac{1}{2}$ +unlabeled data if $2r + \frac{1}{b} < 1$	General

TABLE 1.   Upper rates available from earlier literature (for their applicability to the inverse learning setting considered in the present paper, see Section 2.4).

and extend the existing results presented above, aiming to present a complete picture. We consider a unified approach which allows to simultaneously treat the direct and the inverse learning problem, derive upper bounds (non-asymptotic and asymptotic) as well as lower bounds, both for the $L^2(\nu)$ and the $\mathcal{H}_1$ norm (as well as intermediate norms) for a general class of regularization methods, without requiring additional unlabeled data. In this generality, this is new. In addition, we present a refined analysis of (both strong and weak) minimax optimal rates also investigating their dependence on the complexity of the source condition and on the variance of the noise (our lower bounds come in slightly different strong and weak versions leading to the natural notion of weak and strong minimax optimality). To the best of our knowledge, this has never been done before.

We conclude this review by mentioning the recent work [19], which also concerns inverse statistical learning (see also [20]), albeit in a quite different setting. In that work, the main focus is on classification (Y only can take finitely many values or "classes"), and the inverse problem is that the sampling distribution for X is transformed via a linear operator A. The method analyzed there is empirical risk minimization using a modified loss which implicitly includes an estimation of the original class-conditional distributions from the transformed ones. In the present paper, we consider an (inverse) regression setting with a continuous output variable, the nature of the inverse problem is different since the transformation is applied to the regression function, and we also use a different methododological approach.

The outline of the rest of the paper is as follows. In Section 2, we fix notation and describe our setting in more detail. In particular, we adopt the theory of Carleman operators from the direct problem to our more general setting, including the inverse learning problem. We describe the source conditions, the assumptions on the noise and prior classes, and finally the general class of spectral regularization methods. Granted these preliminaries, we then

present in Section 3 our main results (Theorem 3.4, Theorem 3.5 and Corollary 3.6). In Section 4, we present a concluding discussion on some further aspects of the results. Section 5 contains the proofs of the upper bounds, Section 6 is devoted to the proof of lower bounds. In the Appendix we establish the concentration inequalities and a perturbation result needed in Section 5 and give some supplementary technical lemmata needed in section Section 6.

2. Notation and Preliminaries

In this section, we specify the mathematical setting and assumptions for the model (1.1) and reduce it to an equivalent model.

2.1. **Inverse Problems induced by Carleman Operators.** We assume that the input space $\mathcal{X}$ is a standard Borel space endowed with a probability measure ν, and the output space $\mathcal{Y}$ is equal to $\mathbb{R}$. Let $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ be a linear operator, were $\mathcal{H}_1$ is a infinite-dimensional real separable Hilbert space and $\mathcal{H}_2$ some vector space of functions $g : \mathcal{X} \longrightarrow \mathbb{R}$. We don't assume any specific structure on $\mathcal{H}_2$ for now. However, as will become clear shortly, the image $\mathrm{Im}(A) \subset \mathcal{H}_2$ will be endowed with a natural Hilbert space structure as a consequence of following key assumption:

Assumption 2.1. The evaluation functionals at a given point $x \in \mathcal{X}$:

$$
\begin{aligned}
S_x : \mathcal{H}_1 &\longrightarrow \mathbb{R} \\
f &\longmapsto (S_x)(f) := (Af)(x)
\end{aligned}
$$

are uniformly (w.r.t. $x \in \mathcal{X}$) bounded, i.e., there exists a constant $\kappa < \infty$ such that for any $x \in \mathcal{X}$

$$
|S_x(f)| \leq \kappa \, \|f\|_{\mathcal{H}_1} \, .
$$

For all x, the fact that S_x is continuous implies, by Riesz's representation theorem, the existence of an element $F_x \in \mathcal{H}_1$ such that

$$
(Af)(x) = \langle f, F_x \rangle_{\mathcal{H}_1}
$$

with

$$
\|F_x\|_{\mathcal{H}_1} = \|S_x\| \leq \kappa \, ,
$$

for any $x \in \mathcal{X}$. Define the map

$$
\begin{aligned}
K : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\
(x_1, x_2) &\longmapsto K(x_1, x_2) := \langle F_{x_1}, F_{x_2} \rangle_{\mathcal{H}_1} \, ,
\end{aligned}
$$

which is by construction a positive semidefinite (p.s.d.) kernel over $\mathcal{X}$ associated to the so-called feature space $\mathcal{H}_1$, and the feature map $F : x \in \mathcal{X} \mapsto F_x \in \mathcal{H}_1$. Observe that for any $x \in \mathcal{X}$, we have the bound $K(x, x) = \|F_x\|_{\mathcal{H}_1}^2 \leq \kappa^2$. A fundamental result (see [27], Theorem 4.21) is that to every p.s.d. kernel can be associated a unique reproducing kernel Hilbert space (RKHS). We reproduce this result here, adapted to the considered context:

Proposition 2.2. *(Unique RKHS associated to a psd kernel) The real-valued function space*

$$\mathcal{H}_K := \{g : \mathcal{X} \longrightarrow \mathbb{R} \mid \exists\, f \in \mathcal{H}_1 \text{ with } g(x) = \langle f, F_x \rangle_{\mathcal{H}_1} = (Af)(x) \,\forall x \in \mathcal{X}\}$$
$$= \mathrm{Im}(A) \subset \mathcal{H}_2,$$

equipped with the norm

$$\|g\|_{\mathcal{H}_K} := \inf\left\{ \|f\|_{\mathcal{H}_1} \ : \ f \in \mathcal{H}_1 \text{ s.t. } \forall x \in \mathcal{X} : g(x) = \langle f, F_x \rangle_{\mathcal{H}_1} = (Af)(x) \right\}$$
$$= \inf_{f \in A^{-1}(\{g\})} \|f\|_{\mathcal{H}_1}$$

is the unique RKHS for which K is a reproducing kernel. Moreover, the operator A is a partial isometry from $\mathcal{H}_1$ to $\mathcal{H}_K$ (i.e. an isometry on the orthogonal of its kernel), and

$$\mathcal{H}_K = \overline{\mathrm{Span}}\{K(x, .), x \in \mathcal{X}\}\, .$$

From now on, we can therefore forget about the space $\mathcal{H}_2$ and consider A as an operator from $\mathcal{H}_1$ onto $\mathcal{H}_K = \mathrm{Im}(A)$. As a consequence of A being a partial isometry onto $\mathcal{H}_K$, note that this RKHS is separable, since we have assumed that $\mathcal{H}_1$ is. Additionally, we assume

Assumption 2.3. For any $f \in \mathcal{H}_1$, the map $x \mapsto (Af)(x) = \langle f, F_x \rangle_{\mathcal{H}_1}$ is measurable.

Equivalently, it is assumed that all functions $g \in \mathcal{H}_K$ are measurable. Furthermore, Assumption 2.1 implies that $\|Af\|_\infty \le \kappa \|f\|_{\mathcal{H}_1}$ for all $f \in \mathcal{H}_1$, so that all functions in $\mathcal{H}_K$ are bounded in supremum norm. Therefore, $\mathcal{H}_K$ is a subset of $L^2(\mathcal{X}, \nu)$; let ι denote the associated canonical injection map $\mathcal{H}_K \hookrightarrow L^2(\mathcal{X}, \nu)$.

Together, Assumptions 2.3 and 2.1 thus imply that the map $F_. : \mathcal{X} \longrightarrow \mathcal{H}_1$ is a bounded *Carleman map* [18]. We define the associated *Carleman operator*, as

$$S_\nu : \mathcal{H}_1 \longrightarrow L^2(\mathcal{X}, \nu)$$
$$f \longmapsto S_\nu f := \iota(Af)\, .$$

The operator S_ν is bounded and satisfies $\|S_\nu\| \le \kappa$, since

$$\|S_\nu f\|^2_{L^2(\nu)} = \int_{\mathcal{X}} |(Af)(x)|^2 \, \nu(dx) = \int_{\mathcal{X}} |\langle f, F_x \rangle_{\mathcal{H}_1}|^2 \, \nu(dx) \le \kappa^2 \|f\|^2_{\mathcal{H}_1}\, .$$

We give an illustrative example which is very classical.

Example 2.4. (Differentiating a real function) We consider estimation of a derivative of a real function. To this end, we let $\mathcal{H}_1 := \{f \in L^2[0,1] : \mathbb{E}[f] = 0\}$, the subspace of $L^2([0,1], dt)$ consisting of functions with mean zero and $\mathcal{H}_2 := C[0,1]$, the space of continuous functions on $[0,1]$. Define $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ by

$$[Af](x) = \int_0^x f(t)\, dt\, .$$

Then $Af = g$ if and only if $f = g'$. It is easily checked that Assumption 2.1 is satisfied. To identify the kernel of $\mathrm{Im}(A)$, the reader can easily convince himself that

$$[Af](x) = \langle f, F_x \rangle_{L^2}\, ,$$

where $F_x(t) = \mathbb{1}_{[0,x]}(t) - x$. Thus, by definition $K(x,t) = \langle F_x, F_t \rangle_{L^2} = x \wedge t - xt$ and $\text{Im}(A)$ coincides with the real Sobolev space $H_0^1[0,1]$, consisting of absolutely continuous functions g on $[0,1]$ with weak derivatives of order 1 in $L^2[0,1]$, with boundary condition $g(0) = g(1) = 0$. The associated Carleman operator is given by $S = \iota \circ A$ with $\iota : H_0^1[0,1] \hookrightarrow L^2[0,1]$ and with marginal distribution $\nu = dt$, the Lebesgue measure on $[0,1]$.

We complete this section by defining $B_\nu := S_\nu^\star S_\nu : \mathcal{H}_1 \longrightarrow \mathcal{H}_1$. Then B_ν is positive, selfadjoint and satisfies $\|B_\nu\| \leq \kappa^2$. The following Proposition summarizes the main properties of the operators $S_\nu, S_\nu^\star$ and B_ν. Its proof can be found in the Appendix of [11] (Proposition 19).

Proposition 2.5. *Under Assumptions 2.1 and 2.3, the Carleman operator $S_\nu : \mathcal{H}_1 \longrightarrow L^2(\mathfrak{X}, \nu)$ is a Hilbert-Schmidt operator with nullspace*

$$\ker(S_\nu) = \text{Span}\{F_x \ : \ x \in \text{support}(\nu)\}^\perp .$$

The adjoint operator $S_\nu^\star : L^2(\mathfrak{X}, \nu) \longrightarrow \mathcal{H}_1$ is given by

$$S_\nu^\star g = \int_{\mathfrak{X}} g(x) F_x \, \nu(dx) ,$$

for any $g \in L^2(\mathfrak{X}, \nu)$ and where the integral converges in $\mathcal{H}_1$- norm.
Furthermore, if $F_x \otimes F_x^\star$ denotes the operator $f \in \mathcal{H}_1 \mapsto \langle f, F_x \rangle_{\mathcal{H}_1} F_x \in \mathcal{H}_1$, then

$$B_\nu = \int_{\mathfrak{X}} F_x \otimes F_x^\star \, \nu(dx) ,$$

where the integral converges in trace norm.

It is natural to consider the inverse problem $S_\nu f = g$ (rather than $Af = g$) as the idealized population version (i.e. noise and discretization-free) of (1.1), since since the former views the output of the operator in the geometry of $L^2(\mathfrak{X}, \nu)$, which is the natural population geometry when the sampling measure is ν. Multiplying on both sides by $S_\nu^\star$, we obtain the inverse problem $B_\nu f = S_\nu^\star g$ (called "normal equation" in the inverse problem literature).

Since B_ν is self-adjoint and compact, the spectral theorem ensures the existence of an orthonormal set $\{e_j\}_{j \geq 1}$ such that

$$(2.1) \qquad\qquad B_\nu = \sum_{j=1}^{\infty} \mu_j \langle \cdot, e_j \rangle_{\mathcal{H}_1} e_j$$

and

$$\mathcal{H}_1 = \ker(B_\nu) \oplus \overline{\text{Span}}\{e_j \ : \ j \geq 1\} .$$

The numbers μ_j are the positive eigenvalues of B_ν in decreasing order, satisfying $0 < \mu_{j+1} \leq \mu_j$ for all $j > 0$ and $\mu_j \searrow 0$. In the special case where B_ν has finite rank, the above set of positive eigenvalue and eigenvectors is finite, but to simplify the notation we will always assume that they are countably infinite; formally, we can accomodate for this special situation by allowing that the decreasing sequence of eigenvalues is equal to zero from a certain index on.

Remark 2.6. The considered operators depend on the sampling measure ν and thus also the eigenvalues $(\mu_j)_{j \geq 1}$. For the sake of reading ease, we omit this dependence in the notation; we will also denote henceforth $S = S_\nu$ and $B = B_\nu$.

2.2. **Discretization by random sampling.** For discretization, we consider a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), ..., (x_n, y_n)) \in (\mathcal{X} \times \mathbb{R})^n$ and introduce the associated *sampling operator*

$$S_{\mathbf{x}} : \mathcal{H}_1 \longrightarrow \mathbb{R}^n$$
$$f \longmapsto S_{\mathbf{x}} f,$$

with $(S_{\mathbf{x}} f)_j = \langle f, F_{x_j} \rangle_{\mathcal{H}_1}$, $j = 1, ..., n$ and where $\mathbb{R}^n$ is equipped with the inner product of the empirical L^2 structure,

$$\langle \mathbf{y}, \mathbf{y}' \rangle_{R^n} = \frac{1}{n} \sum_{j=1}^{n} y_j y_j'.$$

Formally, $S_{\mathbf{x}}$ is the counterpart of S_ν when replacing the sampling distribution ν by the empirical distribution $\widehat{\nu} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, and identifying $L^2(\mathcal{X}, \widehat{\nu})$ with $\mathbb{R}^n$ endowed with the above inner product. Additionally, the sampled vector $S_{\mathbf{x}} f$ is corrupted by noise $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)$ to yield the vector of observed values $\mathbf{y} = (y_1, ..., y_n) \in \mathbb{R}^n$:

(2.2)
$$y_j = g(x_j) + \varepsilon_j = (S_{\mathbf{x}} f)_j + \varepsilon_j, \quad j = 1, ..., n,$$

which can be interpretet as the discretized and noisy counterpart of the population problem $S_\nu f = g$. Replacing the measure ν with the empirical measure $\widehat{\nu}$ in Proposition 2.5 gives the following Corollary:

Corollary 2.7. *The sampling operator* $S_{\mathbf{x}} : \mathcal{H}_1 \longrightarrow \mathbb{R}^n$ *is a Hilbert-Schmidt operator with nullspace*

$$\ker(S_{\mathbf{x}}) = \mathrm{Span}\{F_{x_j} \; : \; j = 1, ..., n\}^\perp.$$

Furthermore, the adjoint operator $S_{\mathbf{x}}^\star : \mathbb{R}^n \longrightarrow \mathcal{H}_1$ *is given by*

$$S_{\mathbf{x}}^\star \mathbf{y} = \frac{1}{n} \sum_{j=1}^{n} y_j F_{x_j},$$

and the operator $B_{\mathbf{x}} := S_{\mathbf{x}}^\star S_{\mathbf{x}} : \mathcal{H}_1 \longrightarrow \mathcal{H}_1$ *is given by*

$$B_{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^{n} F_{x_j} \otimes F_{x_j}^\star.$$

With this notation, the normal equation associated to (2.2), obtained by multiplying both sides by $S_{\mathbf{x}}^\star$, reads $S_{\mathbf{x}}^\star \mathbf{y} = B_{\mathbf{x}} f + S_{\mathbf{x}}^\star \boldsymbol{\epsilon}$; it is the discretized and noisy counterpart of the population normal equation introduced in the previous section. The advantage of looking at the normal equations is that both the population and the empirical version act on the same space $\mathcal{H}_1$, so that the latter can be considered as a perturbation of the former (both for the operator and the noise term), an observation which is central to the theory [10].

2.3. **Statistical model, noise assumption, and prior classes.** We recall the considered setting of inverse learning, the sampling is assumed to be random i.i.d., where each observation point (X_i, Y_i) follows the model $Y = Af(X) + \varepsilon$. More precisely, (X_i, Y_i) are i.i.d. with Borel probability distribution ρ on $\mathcal{X} \times \mathbb{R}$. For (X, Y) having distribution ρ, denoting ν the marginal distribution of X, we assume:

Assumption 2.8. The conditional expectation wrt. ρ of Y given X exists and it holds for ν-almost all $x \in X$:

$$(2.3) \qquad \mathbb{E}_\rho[Y|X = x] = Af_\rho(x) = S_x f_\rho, \text{ for some } f_\rho \in \mathcal{H}_1.$$

Furthermore, we will make the following Bernstein-type assumption on the observation noise distribution:

Assumption 2.9. There exists $\sigma > 0$ and $M > 0$ such that for any integer $m \geq 2$:

$$(2.4) \qquad \mathbb{E}[\, |Y - Af_\rho(X)|^m \mid X \,] \leq \frac{1}{2} m! \, \sigma^2 M^{m-2} \quad \nu - \text{a.s.}$$

It is a generally established fact that given any estimator $\widehat{f}$ of f_ρ, one can construct a probability measure ρ on $\mathcal{X} \times \mathbb{R}$ such that the rate of convergence of $\widehat{f}$ to f_ρ can arbitrarily slow (see e.g. [17]). Thus, to derive nontrivial rates of convergence, we concentrate our attention on specific subsets (also called *models*) of the class of probability measures. We will work with the same type of assumptions as considered by [9] and introduce two sets of conditions concerning, on the one hand, the marginal distribution ν of X, and on the other hand, the conditional distribution $\rho(.|.)$ of Y given X.

Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X}$. We define classes of sampling distributions by introducing decay conditions on the eigenvalues μ_i of the operator B_ν defined in Section 2.1.

For $b > 1$ and $\alpha, \beta > 0$, we define

$$\mathcal{P}^<(b, \beta) := \{\nu \in \mathcal{P} : \ \mu_j \leq \beta/j^b \ \forall j \geq 1\},$$

$$\mathcal{P}^>(b, \alpha) := \{\nu \in \mathcal{P} : \ \mu_j \geq \alpha/j^b \ \forall j \geq 1\}$$

and

$$\mathcal{P}^>_{strong}(b, \alpha) := \{\nu \in \mathcal{P}^>(b, \alpha) : \ \exists \gamma > 0, j_0 \geq 1 \text{ s.th. } \frac{\mu_{2j}}{\mu_j} \geq 2^{-\gamma} \ \forall j \geq j_0\}.$$

In the inverse problem literature, such eigenvalue decay assumptions are related to the so-called degree of ill-posedness of the inverse problem $B_\nu f = S^\star g$. In the present setting, the ill-posedness of the problem is reflected by the eigenvalues of B_ν and depends both of the fixed operator A and the sampling distribution ν.

Example 2.10. Coming back to our example 2.4 the degree of ill-posedness is determined by the decay of the eigenvalues $(\mu_j)_j$ of the positive selfadjoint integral operator $L_K = SS^\star : L^2[0, 1] \longrightarrow L^2[0, 1]$

$$[L_K f](x) = \int_0^1 K(x, t) f(t) \, dt.$$

Elementary calculations show that the SVD basis is given by $e_j(x) = \sqrt{2}\sin(\pi j x)$ with corresponding singular values $\mu_j = \frac{1}{\pi^2 j^2}$. Thus, $b = 2$ and $\mathcal{P}^<(2, \frac{1}{\pi^2}) \cap \mathcal{P}^>(2, \frac{1}{\pi^2})$ as well as $\mathcal{P}^<(2, \frac{1}{\pi^2}) \cap \mathcal{P}^>_{strong}(2, \frac{1}{\pi^2})$ are not empty.

For a subset $\Omega \subseteq \mathcal{H}_1$, we let $\mathcal{K}(\Omega)$ be the set of regular conditional probability distributions $\rho(\cdot|\cdot)$ on $\mathcal{B}(\mathbb{R}) \times \mathcal{X}$ such that (2.3) and (2.4) hold for some $f_\rho \in \Omega$. (It is clear that these conditions only depend on the conditional $\rho(.|.)$ of Y given X.) We will focus on a *Hölder-type source condition*, which is a classical smoothness assumption in the theory of inverse problems. Given $r > 0, R > 0$ and $\nu \in \mathcal{P}$, we define

$$(2.5) \qquad \Omega_\nu(r, R) := \{f \in \mathcal{H}_1 : f = B_\nu^r h, \ \|h\|_{\mathcal{H}_1} \le R\}.$$

Note that for any $r \le r_0$ we have $\Omega_\nu(r_0, R) \subseteq \Omega_\nu(r, \kappa^{2(r_0 - r)}R)$, for any $\nu \in \mathcal{P}$. Since B_ν is compact, the source sets $\Omega_\nu(r, R)$ are precompact sets in $\mathcal{H}_1$.

Then the class of models which we will consider will be defined as

$$(2.6) \qquad \mathcal{M}(r, R, \mathcal{P}') := \{\, \rho(dx, dy) = \rho(dy|x)\nu(dx) \ : \ \rho(\cdot|\cdot) \in \mathcal{K}(\Omega_\nu(r, R)), \ \nu \in \mathcal{P}' \,\},$$

with $\mathcal{P}' = \mathcal{P}^<(b, \beta)$, $\mathcal{P}' = \mathcal{P}^>(b, \alpha)$ or $\mathcal{P}' = \mathcal{P}^>_{strong}(b, \alpha)$.

As a consequence, the class of models depends not only on the smoothness properties of the solution (reflected in the parameters $R > 0$, $r > 0$), but also essentially on the decay of the eigenvalues of B_ν.

2.4. **Equivalence with classical kernel learning setting.** With the notation and setting introduced in the previous sections, we point out that the "inverse learning" problem (1.1) can, provided Assumptions (2.1) and (2.3) are met, be reduced to a classical learning problem (hereafter called "direct" learning) under the setting and assumptions of reproducing kernel based estimation methods. In the direct learning setting, the model is given by (1.1) (i.e. $Y_i = g(X_i) + \varepsilon_i$) and the goal is to estimate the function g. Kernel methods *posit* that g belongs to some reproducing kernel Hilbert space[1] $\mathcal{H}_K$ with kernel K and construct an estimate $\widehat{g} \in \mathcal{H}_K$ of g based on the observed data. The reconstruction error $(\widehat{g} - g)$ can be analyzed in $L^2(\nu)$ norm or in $\mathcal{H}_K$-norm.

Coming back to the inverse learning setting $(Y_i = (Af)(X_i) + \varepsilon_i)$, let $\mathcal{H}_K$ be defined as in the previous sections and assume $f \in \ker(A)^\perp$ (we cannot hope to recover the part of f belonging to $\ker A$ anyway, and might as well make this assumption. It is also implied by any form of source condition as introduced in Section 2.3).

Consider applying a direct learning method using the reproducing kernel K; this returns some estimate $\widehat{g} \in \mathcal{H}_K$ of g. Now defining $\widehat{f} := A^{-1}\widehat{g}$, we have

$$\left\|\widehat{f} - f\right\|_{\mathcal{H}_1}^2 = \left\|A^{-1}\widehat{g} - f\right\|_{\mathcal{H}_1}^2 = \|\widehat{g} - Af\|_{\mathcal{H}_K}^2 = \|\widehat{g} - g\|_{\mathcal{H}_K}^2 ,$$

by the partial isometry property of A as an operator $\mathcal{H}_1 \mapsto \mathcal{H}_K$ (Proposition 2.2). Note that $\widehat{f}$ is, at least in principle, accessible to the statistician, since A (and therefore A^{-1}) is assumed to be known. Hence, a bound established for the direct learning setting in

---

[1]This can be extended to the case where g is only approximated in $L^2(\nu)$ by a sequence of functions in $\mathcal{H}_K$. For the sake of the present discussion, only the case where it is assumed $g \in \mathcal{H}_K$ is of interest.

the sense of the $\mathcal{H}_K$-norm reconstruction $\|\widehat{g} - g\|^2_{\mathcal{H}_K}$ also applies to the inverse problem reconstruction error $\left\|\widehat{f} - f\right\|^2_{\mathcal{H}_1}$. Furthermore, it is easy to see that the eigenvalue decay conditions and the source conditions involving the operator B_ν introduced in Section 2.3 are, via the same isometry, equivalent to similar conditions involving the kernel integral operator in the direct learning setting, as considered for instance in [1, 9, 7, 26]. It follows that estimates in $\mathcal{H}_K$-norm available from those references are directly applicable to the inverse learning setting. However, as summarized in Table 1, for the direct learning problem the results concerning $\mathcal{H}_K$-norm rates of convergence are far less complete than in $L^2(\nu)$-norm. In particular, such rates have not been established under consideration of simultaneous source and eigenvalue decay conditions, and neither have the corresponding lower bounds. In this sense, the contribution of the present paper is to complete the picture in Table 1, with the inverse learning setting as the underlying motivation.

2.5. **Regularization.** In this section, we introduce the class of linear regularization methods based on spectral theory for self-adjoint linear operators. These are standard methods for finding stable solutions for ill-posed inverse problems, see e.g. [13] or [15].

Definition 2.11 (Regularization function). Let $g : (0,1] \times [0,1] \longrightarrow \mathbb{R}$ be a function and write $g_\lambda = g(\lambda, \cdot)$. The family $\{g_\lambda\}_\lambda$ is called *regularization function*, if the following conditions hold:

  (i) There exists a constant $D < \infty$ such that
$$\sup_{0 < t \leq 1} |t g_\lambda(t)| \leq D,$$
    for any $0 < \lambda \leq 1$.
  (ii) There exists a constant $E < \infty$ such that
$$(2.7) \qquad \sup_{0 < t \leq 1} |g_\lambda(t)| \leq \frac{E}{\lambda},$$
    for any $0 < \lambda \leq 1$.
  (iii) Defining the *residual*
$$(2.8) \qquad r_\lambda(t) = 1 - g_\lambda(t) t \,,$$
    there exists a constant $\gamma_0 < \infty$ such that
$$\sup_{0 < t \leq 1} |r_\lambda(t)| \leq \gamma_0,$$
    for any $0 < \lambda \leq 1$.

Definition 2.12 (Qualification). The *qualification* of the regularization $\{g_\lambda\}_\lambda$ is the maximal q such that for any $0 < \lambda \leq 1$
$$\sup_{0 < t \leq 1} |r_\lambda(t)| t^q \leq \gamma_q \lambda^q.$$
for some constant $\gamma_q > 0$.

The next lemma provides a simple inequality (see e.g. [21], Proposition 3) that shall be used later.

Lemma 2.13. Let $\{g_\lambda\}_\lambda$ be a regularization function with qualification q. Then, for any $r \leq q$ and $0 < \lambda \leq 1$:

$$\sup_{0<t\leq 1} |r_\lambda(t)| t^r \leq \gamma_r \lambda^r,$$

where $\gamma_r := \gamma_0^{1-\frac{r}{q}} \gamma_q^{\frac{r}{q}}$.

We give some examples which are common both in classical inverse problems [13] and in learning theory [1].

Example 2.14. **(Spectral Cut-off)** A very classical regularization method is *spectral cut-off* (or truncated singular value decomposition), defined by

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

In this case, $D = E = \gamma_0 = \gamma_q = 1$. The qualification q of this method can be arbitrary.

Example 2.15. **(Tikhonov Regularization)** The choice $g_\lambda(t) = \frac{1}{\lambda+t}$ corresponds to *Tikhonov regularization*. In this case we have $D = E = \gamma_0 = 1$. The qualification of this method is $q = 1$ with $\gamma_q = 1$.

Example 2.16. **(Landweber Iteration)** The *Landweber Iteration* (gradient descent algorithm with constant stepsize) is defined by

$$g_k(t) = \sum_{j=0}^{k-1} (1 - t)^j \text{ with } k = 1/\lambda \in \mathbb{N} \,.$$

We have $D = E = \gamma_0 = 1$. The quailfication q of this algorithm can be arbitrary with $\gamma_q = 1$ if $0 < q \leq 1$ and $\gamma_q = q^q$ if $q > 1$.

Given the sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathbb{R})^n$, we define the regularized approximate solution $f_\mathbf{z}^\lambda$ of problem (2.2), for a suitable a-priori parameter choice $\lambda = \lambda_n$, by

$$(2.9) \qquad f_\mathbf{z}^{\lambda_n} := g_{\lambda_n}(\kappa^{-2} B_\mathbf{x}) \kappa^{-2} S_\mathbf{x}^\star \mathbf{y} = g_{\lambda_n}(\bar{B}_\mathbf{x}) \bar{S}_\mathbf{x}^\star \mathbf{y} \,,$$

where we have introduced the shortcut notation $\bar{B}_x := \kappa^{-2} B_\mathbf{x}$ and $\bar{S}_\mathbf{x} := \kappa^{-2} S_\mathbf{x}$. Note that $g_\lambda(\bar{B}_\mathbf{x})$ is well defined since $\|\bar{B}_\mathbf{x}\| \leq 1 \,.$

3. Main results: upper and lower bounds on convergence rates

Before stating our main results, we recall some basic definitions in order to clarify what we mean by asymptotic upper rate, lower rate and minimax rate optimality. We want to track the precise behavior of these rates not only for what concerns the exponent in the number of examples n, but also in terms of their scaling (multiplicative constant) as a function of some important parameters (namely the noise variance σ^2 and the complexity radius R in the source condition). For this reason, we introduce a notion of a family of rates

over a family of models. More precisely, in all the forthcoming definitions, we consider an indexed family $(\mathcal{M}_\theta)_{\theta \in \Theta}$, where for all $\theta \in \Theta$, $\mathcal{M}_\theta$ is a class of Borel probability distributions on $\mathcal{X} \times \mathbb{R}$ satisfying the basic general assumption 2.8. We consider rates of convergence in the sense of the p-th moments of the estimation error, where $p > 0$ is a fixed real number.

Definition 3.1. (Upper Rate of Convergence)
A family of sequences $(a_{n,\theta})_{(n,\theta) \in \mathbb{N} \times \Theta}$ of positive numbers is called **upper rate of convergence in L^p** for the interpolation norm of parameter $s \in [0, \frac{1}{2}]$, over the family of models $(\mathcal{M}_\theta)_{\theta \in \Theta}$, for the sequence of estimated solutions $(f_{\mathbf{z}}^{\lambda_{n,\theta}})_{(n,\theta) \in \mathbb{N} \times \Theta}$, using regularization parameters $(\lambda_{n,\theta})_{(n,\theta) \in \mathbb{N} \times \Theta}$, if

$$\sup_{\theta \in \Theta} \limsup_{n \to \infty} \sup_{\rho \in \mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}} \left[\| B_\nu^s (f_\rho - f_{\mathbf{z}}^{\lambda_{n,\theta}}) \|_{\mathcal{H}_1}^p \right]^{\frac{1}{p}}}{a_{n,\theta}} < \infty.$$

Definition 3.2. (Weak and Strong Minimax Lower Rate of Convergence)
A family of sequences $(a_{n,\theta})_{(n,\theta) \in \mathbb{N} \times \Theta}$ of positive numbers is called **weak minimax lower rate of convergence in L^p** for the interpolation norm of parameter $s \in [0, \frac{1}{2}]$, over the family of models $(\mathcal{M}_\theta)_{\theta \in \Theta}$, if

$$\inf_{\theta \in \Theta} \limsup_{n \to \infty} \inf_{f_\bullet} \sup_{\rho \in \mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}} \left[\| B_\nu^s (f_\rho - f_{\mathbf{z}}) \|_{\mathcal{H}_1}^p \right]^{\frac{1}{p}}}{a_{n,\theta}} > 0,$$

where the infimum is taken over all estimators, i.e. measurable mappings $f_\bullet : (\mathcal{X} \times \mathbb{R})^n \longrightarrow \mathcal{H}_1$. It is called a **strong minimax lower rate of convergence in L^p** if

$$\inf_{\theta \in \Theta} \liminf_{n \to \infty} \inf_{f_\bullet} \sup_{\rho \in \mathcal{M}_\theta} \frac{\mathbb{E}_{\rho^{\otimes n}} \left[\| B_\nu^s (f_\rho - f_{\mathbf{z}}) \|_{\mathcal{H}_1}^p \right]^{\frac{1}{p}}}{a_{n,\theta}} > 0.$$

The difference between weak and strong lower rate can be summarily reformulated in the following way: if r_n denotes the sequence of minimax errors for a given model and reconstruction error, using n observations, then $a_n = \mathcal{O}(r_n)$ must hold if a_n is a strong lower rate, while a_n being a weak lower means that $r_n = o(a_n)$ is excluded.

Definition 3.3. (Minimax Optimal Rate of Convergence)
The sequence of estimated solutions $(f_{\mathbf{z}}^{\lambda_{n,\theta}})_n$ using the regularization parameters $(\lambda_{n,\theta})_{(n,\theta) \in \mathbb{N} \times \Theta}$ is called **weak/strong minimax optimal in L^p** for the interpolation norm of parameter $s \in [0, \frac{1}{2}]$, over the model family $(\mathcal{M}_\theta)_{\theta \in \Theta}$, with **rate of convergence** given by the sequence $(a_{n_\theta})_{(n,\theta) \in \mathbb{N} \times \Theta}$, if the latter is a weak/strong minimax lower rate as well as an upper rate for $(f_{\mathbf{z}}^{\lambda_{n,\theta}})_{n,\theta}$.

We now formulate our main theorems.

Theorem 3.4. *Consider the model $\mathcal{M}_{\sigma,M,R} := \mathcal{M}(r, R, \mathcal{P}^<(b, \beta))$ (as defined in Section 2.3), where $r > 0$, $b > 1$ and $\beta > 0$ are fixed, and $(R, M, \sigma) \in \mathbb{R}_+^3$ (remember that (σ, M) are the parameters in the Bernstein moment condition (2.4), in particular σ^2 is a bound on*

the noise variance.) Given a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathfrak{X} \times \mathbb{R})^n$, define $f_{\mathbf{z}}^\lambda$ as in (2.9), using a regularization function of qualification $q \geq r + s$, with the parameter sequence

$$(3.1) \qquad \lambda_{n,(\sigma,R)} = \min\left(\left(\frac{\sigma^2}{R^2 n} \right)^{\frac{b}{2br+b+1}}, 1 \right).$$

Then for any $s \in [0, \frac{1}{2}]$, the sequence

$$(3.2) \qquad a_{n,(\sigma,R)} = R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{b(r+s)}{2br+b+1}}$$

is an upper rate of convergence in L^p for all $p > 0$, for the interpolation norm of parameter s, for the sequence of estimated solutions $(f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})$ over the family of models $(\mathcal{M}_{\sigma,M,R})_{(\sigma,M,R) \in \mathbb{R}_+^3}$.

Theorem 3.5. *Let $r > 0, R > 0, b > 1$ and $\alpha > 0$ be fixed. Let ν be a distribution on $\mathfrak{X}$ belonging to $\mathcal{P}^>(b, \alpha)$. Then the sequence $(a_{n,(\sigma,R)})$ defined in (3.2) is a weak minimax lower rate of convergence in L^p for all $p > 0$, for the model family $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r, R, \{\nu\})$, $(R, M, \sigma) \in \mathbb{R}_+^3$. If ν belongs to $\mathcal{P}^>_{strong}(b, \alpha)$, then the sequence $a_{n,(\sigma,R)}$ is a strong minimax lower rate of convergence in L^p for all $p > 0$, for the model family $\mathcal{M}_{R,M,\sigma}$.*

Finally, we have as a direct consequence:

Corollary 3.6. *Let $r > 0, b > 1$, $\beta \geq \alpha > 0$ be fixed and assume $\mathcal{P}' = \mathcal{P}^<(b, \beta) \cap \mathcal{P}^>(b, \alpha) \neq \emptyset$. Then the sequence of estimators $f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}}$ as defined in (2.9) is strong minimax optimal in L^p for all $p > 0$, under the assumptions and parameter sequence (3.1) of Theorem 3.4, over the class $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r, R, \mathcal{P}')$, $(R, M, \sigma) \in \mathbb{R}_+^3$.*

4. Discussion

We conclude by briefly discusssing some specific points related to our results.

Non-asymptotic, high probability bounds. The results presented in Section 3 are asymptotic in nature and concern moments of the reconstruction error. However, the main underlying technical result is an exponential deviation inequality which holds non-asymptotically. For simplicity of the exposition we have chose to relegate this result to the Appendix (Proposition 5.8 there). Clearly, this is thanks to such a deviation inequality that we are able to handle moments of all orders of the error. Furthermore, while the asymptotics considered in the previous section always assume that all parameters are fixed as $n \to \infty$, going back to the deviation inequality one could in principle analyze asymptotics of other nonstandard regimes where some parameters are allowed to depend on n.

Adaptivity. For our results we have assumed that the crucial parameters b, r, R concerning the eigenvalue decay of the operator B_ν as well as the regularity of the target function are known, and so is the noise variance σ; these parameters are used in the choice of regulatizing constant λ_n. This is, of course, very unrealistic. Ideally, we would like to have a procedure doing almost as good without knowledge of these parameters in advance – this is the question of adaptivity. While this topic is outside of the scope of the present paper,

in work in progress we study such an adaptive procedure based on Lepski's principle for the oracle selection of a suitable regularizing constant λ – this is again a situation where an exponential deviation inequality is a particularly relevant tool.

Weak and strong lower bounds. The notion of strong and weak lower bounds introduced in this work (corresponding respectively to a lim inf and lim sup in n) appear to be new. They were motivated by the goal to consider somewhat minimal assumptions on the eigenvalue behavior, i.e. only a one-sided power decay bound, to obtain lower minimax bounds under source condition regularity. It turns out a one-sided power decay bound is the main driver for minimax rates, but excluding arbitrarly abrupt relative variations μ_{2j}/μ_j appears to play a role in distinguishing the weak and strong versions. Such a condition is also called one-sided regular variation, see [4] for extensive considerations on such issues. We believe that this type of assumption can be relevant for the analysis of certain inverse problems when the eigenvalues do not exhibit a two-sided power decay.

Smoothness and source conditions. In considering source conditions (2.5) in terms of the operator B_ν as measure of regularity of the target f, we have followed the general approach adopted in previous works on statistical learning using kernels, itself inspired by the setting considered in the (deterministic) inverse problem literature. It is well-established in the latter literature that representing the target function in terms of powers of the operator to be inverted is a very natural way to measure its regularity; it can be seen as a way to relate noise and signal in a geometry that is appropriate for the considered ill-posed problem. In our setting, one can however wonder why a measure of regularity of the target function should depend on the sampling distribution ν. A high-level answer is that the sampling can itself be seen as a source of noise (or uncertainty), and that it is natural that it enters in the ill-posedness of the problem. For instance, regions in space with sparser sampling will result in more uncertainty. On the other hand, if, say, the support of ν is contained in a low-dimensional manifold, the problem becomes intrinsically lower-dimensional, being understood that we must abandon any hope of estimating outside of the support, and this should also be reflected in the measure of regularity. A more detailed analysis of such issues, and relations to more common notions of regularity, is out of the scope of the present work but certainly an interesting future perspective.

5. Proof of Upper Rate

We recall the shortcut notation $\bar{B}_x := \kappa^{-2} B_{\mathbf{x}}$, $\bar{S}_{\mathbf{x}} := \kappa^{-2} S_{\mathbf{x}}$ and similarly define $\bar{B} := \kappa^{-2} B$. Recall that we denote $\|A\|$ the spectral norm of an operator A between Hilbert spaces; additionally we will denote $\|A\|_{\mathrm{HS}}$ the Hilbert-Schmidt norm of A (assuming it is well-defined).

All along the proof, we will use the notation C_a to denote a positive factor only depending on the quantity a. The exact expression of this factor depends on the context and can potentially change from line to line.

5.1. **Concentration Inequalities.** We introduce the *effective dimension* $\mathcal{N}(\lambda)$, appearing in [9] in a similar context. For $\lambda \in (0, 1]$ we set

(5.1) $$\mathcal{N}(\lambda) = \operatorname{tr}(\ (\bar{B} + \lambda)^{-1}\bar{B}\ ) \ .$$

Since by Proposition 2.5 the operator B is trace-class, $\mathcal{N}(\lambda) < \infty$. Moreover, we have the following estimate (see [9], Proposition 3):

Lemma 5.1. *Assume that the marginal distribution ν of X belongs to $\mathcal{P}^<(b, \beta)$ (with $b > 1$ and $\beta > 0$). Then the effective dimension $\mathcal{N}(\lambda)$ satisfies*

$$\mathcal{N}(\lambda) \leq \frac{\beta b}{b - 1}(\kappa^2 \lambda)^{-\frac{1}{b}} \ .$$

Furthermore, for $\lambda \leq ||\bar{B}||$, since $\bar{B}$ is positive

$$\mathcal{N}(\lambda) = \sum_{\mu_j \geq \kappa^2 \lambda} \frac{\mu_j}{\mu_j + \kappa^2 \lambda} + \sum_{\mu_j < \kappa^2 \lambda} \frac{\mu_j}{\mu_j + \kappa^2 \lambda} \geq \min_{\mu_j \geq \kappa^2 \lambda} \left\{ \frac{\mu_j}{\mu_j + \kappa^2 \lambda} \right\} \geq \frac{1}{2} \ ,$$

since the first sum has at least one term. The following propositions summarize important concentration properties of the empirical quantities involved. The proofs are given in Appendix A.

Proposition 5.2. *For $n \in \mathbb{N}$, $\lambda \in (0, 1]$ and $\eta \in (0, 1]$, it holds with probability at least $1 - \eta$:*

$$\left\| (\bar{B} + \lambda)^{-\frac{1}{2}} \left(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y} \right) \right\|_{\mathcal{H}_1} \leq 2 \log(2\eta^{-1})\kappa^{-1} \left(\frac{M}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) \ .$$

Also, it holds with probability at least $1 - \eta$:

$$\left\| \bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y} \right\|_{\mathcal{H}_1} \leq 2 \log(2\eta^{-1})\kappa^{-1} \left(\frac{M}{n} + \sqrt{\frac{\sigma^2}{n}} \right) \ .$$

Proposition 5.3. *For any $n \in \mathbb{N}$, $\lambda \in (0, 1]$ and $\eta \in (0, 1)$, it holds with probability at least $1 - \eta$:*

$$\left\| (\bar{B} + \lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}}) \right\|_{\mathrm{HS}} \leq 2 \log(2\eta^{-1}) \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \right) \ .$$

Proposition 5.4. *Let $\eta \in (0, 1)$. Assume that $\lambda \in (0, 1]$ satisfies*

(5.2) $$\sqrt{n\lambda} \geq 8 \log(2\eta^{-1}) \sqrt{\max(\mathcal{N}(\lambda), 1)} \ .$$

Then, with probability at least $1 - \eta$:

(5.3) $$\left\| (\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda) \right\| \leq 2 \ .$$

Proposition 5.5. *For any $n \in \mathbb{N}$ and $0 < \eta < 1$ it holds with probability at least $1 - \eta$:*

$$\left\| \bar{B} - \bar{B}_{\mathbf{x}} \right\|_{\mathrm{HS}} \leq 6 \log(2\eta^{-1}) \frac{1}{\sqrt{n}} \ .$$

5.2. **Some operator perturbation inequalities.**

Proposition 5.6. *Let B_1, B_2 be two non-negative self-adjoint operators on some Hilbert space with $\|B_j\| \leq a$, $j = 1, 2$, for some non-negative a.*

(i) *If $0 \leq r \leq 1$, then*
$$\|B_1^r - B_2^r\| \leq C_r \|B_1 - B_2\|^r,$$

for some $C_r < \infty$.

(ii) *If $r > 1$, then*
$$\|B_1^r - B_2^r\| \leq C_{a,r} \|B_1 - B_2\|,$$

for some $C_{a,r} < \infty$.

Proof.   (i) Since $t \mapsto t^r$ is operator monotone if $r \in [0, 1]$, this result follows from Theorem X.1.1 in [2] for positive matrices, but it the proof applies as well to positive operators on a Hilbert space.

(ii) The proof follows from Proposition B.1 in Appendix B.

$\square$

Proposition 5.7 ([2], Theorem IX.2.1-2). *Let A, B be to self-adjoint, positive operators on a Hilbert space. Then for any $s \in [0, 1]$:*

(5.4)
$$\|A^s B^s\| \leq \|AB\|^s.$$

Note: this result is stated for positive matrices in [2], but it is easy to check that the proof applies as well to positive operators on a Hilbert space.

5.3. **Proof of Theorem 3.4.** The following proposition is our main error bound and the convergence rate will follow.

Proposition 5.8. *Let $s \in [0, \frac{1}{2}]$, $r > 0$, $R > 0$, $M > 0$. Suppose $f_\rho \in \Omega_\nu(r, R)$ (defined in (2.5)). Let $f_{\mathbf{z}}^\lambda$ be defined as in (2.9) using a regularization function of qualification $q \geq r+s$ and put $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. Then, for any $\eta \in (0, 1)$, $\lambda \in (0, 1]$ and $n \in \mathbb{N}$ satisfying*

(5.5)
$$n \geq 64\lambda^{-1} \max(\mathcal{N}(\lambda), 1) \log^2(8/\eta),$$

we have with probability at least $1 - \eta$:
(5.6)
$$\left\|B^s(f_\rho - f_{\mathbf{z}}^\lambda)\right\|_{\mathcal{H}_1} \leq C_{r,s,D,E} \, \log(8\eta^{-1})\kappa^{2s}\lambda^s \left(\bar{\gamma}\kappa^{2r}R\left(\lambda^r + \frac{1}{\sqrt{n}}\right) + \kappa^{-1}\left(\frac{M}{n\lambda} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\lambda}}\right)\right).$$

Proof. We start with a preliminary inequality. Assumption (5.5) implies that (5.2) holds with η replaced by $\eta/4$. We can therefore apply Proposition 5.4 and obtain that, with probability at least $1 - \eta/4$, inequality (5.3) holds. Combining this with (5.4), we get for any $u \in [0, 1]$:
(5.7)
$$\left\|\bar{B}^u(\bar{B}_{\mathbf{x}} + \lambda)^{-u}\right\| = \left\|\bar{B}^u(\bar{B} + \lambda)^{-u}(\bar{B} + \lambda)^u(\bar{B}_{\mathbf{x}} + \lambda)^{-u}\right\| \leq \left\|(\bar{B} + \lambda)(\bar{B}_{\mathbf{x}} + \lambda)^{-1}\right\|^u \leq 2.$$

From this we deduce readily that, with probability at least $1 - \eta/4$, we have

$$(5.8) \qquad \left\| B^s(f_\rho - f_{\mathbf{z}}^\lambda) \right\|_{\mathcal{H}_1} \leq 2\kappa^{2s} \left\| (\bar{B}_{\mathbf{x}} + \lambda)^s (f_\rho - f_{\mathbf{z}}^\lambda) \right\|_{\mathcal{H}_1}.$$

We now consider the following decomposition

$$\begin{aligned}
f_\rho - f_{\mathbf{z}}^\lambda &= f_\rho - g_\lambda(\bar{B}_{\mathbf{x}})\bar{S}_{\mathbf{x}}^\star \mathbf{y} \\
&= (f_\rho - g_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}f) + g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}}f - \bar{S}_{\mathbf{x}}^\star \mathbf{y}) \\
(5.9) \qquad &= r_\lambda(\bar{B}_{\mathbf{x}})f + g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}}f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}),
\end{aligned}$$

where r_λ is given by (2.8). We now upper bound $\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s (f_\rho - f_{\mathbf{z}}^\lambda) \right\|_{\mathcal{H}_1}$ by treating separately the two terms corresponding to the above decomposition.

Step 1: First term: since $f_\rho \in \Omega_\nu(r, R)$, we have

$$\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})f_\rho \right\|_{\mathcal{H}_1} \leq \kappa^{2r} R \left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}^r \right\|.$$

We now concentrate on the operator norm appearing in the RHS of the above bound, and distinguish between two cases. The first case is $r \geq 1$, for which we write

$$(5.10) \qquad (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}^r = (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^r + (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}^r - \bar{B}_{\mathbf{x}}^r).$$

The operator norm of the first term is estimated via

$$\begin{aligned}
\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}_{\mathbf{x}}^r \right\| &\leq \sup_{t \in [0,1]} (t + \lambda)^s t^r r_\lambda(t) \\
&\leq \sup_{t \in [0,1]} t^{s+r} r_\lambda(t) + \lambda^s \sup_{t \in [0,1]} t^r r_\lambda(t) \\
(5.11) \qquad &\leq 2\bar{\gamma}\lambda^{s+r},
\end{aligned}$$

by applying (twice) Lemma 2.13 and the assumption that the qualification q of the regularization is greater than $r + s$; we also introduced $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. The second term in equation (5.10) is estimated via

$$\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}^r - \bar{B}_{\mathbf{x}}^r) \right\| \leq \left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}}) \right\| \left\| \bar{B}^r - \bar{B}_{\mathbf{x}}^r \right\| \leq 2\bar{\gamma}C_r \lambda^s \left\| \bar{B} - \bar{B}_{\mathbf{x}} \right\|.$$

For the first factor we have used the same device as previously for the first term based on Lemma 2.13, and for the second factor we used Proposition 5.6 (ii). Finally using Proposition 5.5 to upper bound $\left\| \bar{B} - \bar{B}_{\mathbf{x}} \right\|$, collecting the previous estimates we obtain with probability at least $1 - \eta/2$:

$$(5.12) \qquad \left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})f_\rho \right\|_{\mathcal{H}_1} \leq \bar{\gamma}C_r \kappa^{2r} R \log\left(4\eta^{-1}\right) \left(\lambda^r + \frac{1}{\sqrt{n}}\right) \lambda^s.$$

We turn to the case $r < 1$, for which we want to establish a similar inequality. Instead of (5.10) we use:

$$\begin{aligned}
\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})\bar{B}^r \right\| &= \left\| (\bar{B}_{\mathbf{x}} + \lambda)^s r_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} + \lambda)^r (\bar{B}_{\mathbf{x}} + \lambda)^{-r}\bar{B}^r \right\| \\
&\leq 2 \left\| (\bar{B}_{\mathbf{x}} + \lambda)^{r+s} r_\lambda(\bar{B}_{\mathbf{x}}) \right\| \\
&\leq 8\bar{\gamma}\lambda^{r+s},
\end{aligned}$$

where we have used the (transposed version of) inequality (5.7) (valid with probability at least $1 - \eta/4$); and, for the last inequality, an argument similar the one leading to (5.11) (using this time that $(t + \lambda)^{r+s} \leq 2(t^{r+s} + \lambda^{r+s})$ for all $t \geq 0$ since $r + s \leq 2$ in the case we are considering). This implies *a fortiori*, that inequality (5.12) holds in the case $r < 1$ as well (also with probability at least $1 - \eta/2$).

Step 2: Bound on $\left\| (\bar{B}_{\mathbf{x}} + \lambda)^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}) \right\|_{\mathcal{H}_1}$.

We further split by writing

$$(5.13) \qquad (\bar{B}_{\mathbf{x}} + \lambda)^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y}) \;\; = \;\; H_{\mathbf{x}}^{(1)} \cdot H_{\mathbf{x}}^{(2)} \cdot h_{\mathbf{z}}^\lambda$$

with

$$\begin{aligned}
H_{\mathbf{x}}^{(1)} &:= (\bar{B}_{\mathbf{x}} + \lambda)^s g_\lambda(\bar{B}_{\mathbf{x}})(\bar{B}_{\mathbf{x}} + \lambda)^{\frac{1}{2}}, \\
H_{\mathbf{x}}^{(2)} &:= (\bar{B}_{\mathbf{x}} + \lambda)^{-\frac{1}{2}}(\bar{B} + \lambda)^{\frac{1}{2}}, \\
h_{\mathbf{z}}^\lambda &:= (\bar{B} + \lambda)^{-\frac{1}{2}}(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y})
\end{aligned}$$

and proceed by bounding each factor separately.

For the first term, we have (for any $\lambda \in (0, 1]$ and $\mathbf{x} \in \mathcal{X}^n$), and remembering that $s \leq 1/2$:

$$\begin{aligned}
\left\| H_{\mathbf{x}}^{(1)} \right\| &\leq \sup_{t \in [0,1]} (t + \lambda)^{s + \frac{1}{2}} g_\lambda(t) \\
&\leq \lambda^{s + \frac{1}{2}} \sup_{t \in [0,1]} g_\lambda(t) + \sup_{t \in [0,1]} \left| t^{s + \frac{1}{2}} g_\lambda(t) \right| \\
&\leq E \lambda^{s - \frac{1}{2}} + \left(\sup_{t \in [0,1]} |t g_\lambda(t)| \right)^{s + \frac{1}{2}} \left(\sup_{t \in [0,1]} |g_\lambda(t)| \right)^{\frac{1}{2} - s} \\
(5.14) \qquad &\leq E \lambda^{s - \frac{1}{2}} + D^{s + \frac{1}{2}} E^{\frac{1}{2} - s} \lambda^{s - \frac{1}{2}} = C_{s,D,E} \lambda^{s - \frac{1}{2}},
\end{aligned}$$

where we have used Definition 2.11 (i), (ii).

The probabilistic bound on $H_{\mathbf{x}}^{(2)}$ follows from Proposition 5.4, which we can apply using assumption (5.5), combined with Proposition 5.7. This ensures with probability at least $1 - \eta/4$

$$(5.15) \qquad \left\| H_{\mathbf{x}}^{(2)} \right\| \leq 2 .$$

Finally, the probabilistic bound on $h_{\mathbf{z}}^\lambda$ follows from Proposition 5.2: with probability at least $1 - \eta/4$, we have

$$(5.16) \qquad \left\| h_{\mathbf{z}}^\lambda \right\|_{\mathcal{H}_1} \leq 2 \log(8\eta^{-1}) \kappa^{-1} \left(\frac{M}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) .$$

As a result, combining (5.14), (5.15) and (5.16) with (5.13) gives with probability at least $1 - \eta/2$

$$(5.17) \quad \|(B_{\mathbf{x}} + \lambda)^s g_\lambda(B_{\mathbf{x}})(B_{\mathbf{x}}f - S_{\mathbf{x}}^\star \mathbf{y})\|_{\mathcal{H}_1} \leq C_{s,D,E} \log(8\eta^{-1})\kappa^{-1}\lambda^s \left(\frac{M}{n\lambda} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n\lambda}} \right) ,$$

We end the proof by collecting (5.8), (5.9), (5.12) and (5.17) and finally obtain bound (5.6) holding with probability at least $1 - \eta$. $\qquad\square$

Corollary 5.9. *Let* $s \in [0, \frac{1}{2}]$, $\sigma > 0, M > 0, r > 0, R > 0, \beta > 0, b > 1$ *and assume the generating distribution of* (X, Y) *belongs to* $\mathcal{M}(r, R, \mathcal{P}^<(b, \beta))$ *(defined in Section 2.3). Let* $f_{\mathbf{z}}^\lambda$ *be the estimator defined as in (2.9) using a regularization function of qualification* $q \geq r + s$ *and put* $\bar{\gamma} := \max(\gamma_0, \gamma_q)$. *Then, there exists* $n_0 > 0$ *(depending on the above parameters), so that for all* $n \geq n_0$, *if we set*

$$(5.18) \qquad\qquad \lambda_n = \min\left(\left(\frac{\sigma}{R\sqrt{n}} \right)^{\frac{2b}{2br+b+1}}, 1 \right) ,$$

then with probability at least $1 - \eta$:

$$\left\| B^s(f - f_{\mathbf{z}}^{\lambda_n}) \right\|_{\mathcal{H}_1} \leq C_{r,s,b,\beta,\bar{\gamma},D,E,\kappa} \ \log(8\eta^{-1})R \left(\frac{\sigma}{R\sqrt{n}} \right)^{\frac{2b(r+s)}{2br+b+1}} ,$$

provided $\log \eta^{-1} \leq C_{b,\beta,\kappa,\sigma,R} n^{\frac{br}{2br+b+1}}$.

Remark: In the above corollary, n_0 can possibly depend on all parameters, but the constant in front of the upper bound does not depend on R, σ, nor M. In this sense, this result tracks precisely the effect of these important parameters on the scaling of the rate, but remains asymptotic in nature: it cannot be applied if, say, R, σ of M also depend on n (because the requirement $n \geq n_0$ might then lead to an impossiblity.) If some parameters are allowed to change with n, one should go back to the nonasymptotic statement of Theorem 5.8 for an analysis of the rates.

Proof. We check that the assumptions of Proposition 5.8 are satisfied provided n is big enough. Concerning assumption (5.5), let us recall that by Lemma 5.1:

$$(5.19) \qquad\qquad \mathcal{N}(\lambda) \leq C_{b,\beta,\kappa} \lambda^{-1/b} ,$$

for some $C_{b,\beta,\kappa} > 0$. Consequently, (5.5) is ensured by the sufficient condition

$$(5.20) \qquad n \geq C_{b,\beta,\kappa} \log^2(8\eta^{-1})\lambda^{-\frac{1}{b}-1} \ \Leftarrow \ \log \eta^{-1} \leq C_{b,\beta,\kappa,\sigma,R} n^{\frac{br}{2br+b+1}} .$$

Applying Proposition 5.8, Lemma 5.1 again, and folding the effect of the parameters we do not intend to track precisely into a generic multiplicative constant, we obtain that with probability $1 - \eta$:
(5.21)

$$\left\| B^s(f_\rho - f_{\mathbf{z}}^\lambda) \right\|_{\mathcal{H}_1} \leq C_{r,s,\kappa,\bar{\gamma},D,E,b,\beta} \ \log(8\eta^{-1})\lambda^s \left(R\left(\lambda^r + \frac{1}{\sqrt{n}} \right) + \left(\frac{M}{n\lambda} + \frac{\sigma}{\sqrt{n}}\lambda^{-\frac{b+1}{2b}} \right) \right) .$$

Observe that the choice (5.18) implies that $n^{-\frac{1}{2}} = o(\lambda_n^r)$. Therefore, up to requiring n large enough and multiplying the front factor by 2 , we can disregard the term $1/\sqrt{n}$ in the second factor of the above bound. Similarly, by comparing the exponents in n, one can readily check that

$$\frac{M}{n\lambda_n} = o\left(\sqrt{\frac{1}{n}\lambda_n^{-\frac{b+1}{b}}}\right),$$

so that we can also disregard the term $(n\lambda_n)^{-1}$ for n large enough (again, up multiplying the front factor by 2) and concentrate on the two remaining main terms of the upper bound in (5.21), which are $R\lambda^r$ and $\sigma\lambda^{-\frac{b+1}{2b}}n^{-\frac{1}{2}}$. The proposed choice of λ_n balances precisely these two terms and easy computations lead to the announced conclusion.   □

Proof of Theorem 3.4. We would like to "integrate" the bound of Corollary 5.9 over η to obtain a bound in L^p norm (see Lemma C.1 in the Appendix), unfortunately the condition on η prevents this since very large deviations are excluded. To alleviate this, we first derive a much coarser "fallback" upper bound which will be valid for all $\eta \in (0,1)$. To this aim, we revisit shortly the proof of Proposition 5.8 . We recall the decomposition (5.9)

$$B^s(f_\rho - f_\mathbf{z}^\lambda) = B^s\left(r_\lambda(\bar{B}_\mathbf{x})f + g_\lambda(\bar{B}_\mathbf{x})(\bar{B}_\mathbf{x}f_\rho - \bar{S}_\mathbf{x}^\star\mathbf{y})\right).$$

A rough bound on the first term using (2.8), and $f_\rho \in \Omega(r,R)$ is

(5.22) $$\left\|B^s r_\lambda(\bar{B}_\mathbf{x})f_\rho\right\|_{\mathcal{H}_1} \le \gamma_0\kappa^{2(r+s)}R.$$

For the second term, using (2.7) and the second part or Proposition 5.2 , we obtain that with probability at least $1 - \eta$,
(5.23)

$$\left\|B^s g_\lambda(\bar{B}_\mathbf{x})(\bar{B}_\mathbf{x}f_\rho - \bar{S}_\mathbf{x}^\star\mathbf{y})\right\|_{\mathcal{H}_1} \le 2\kappa^{2s-1}\log(2\eta^{-1})\frac{E}{\lambda}\left(\frac{M}{n} + \sqrt{\frac{\sigma^2}{n}}\right) \le C_{\kappa,E,M,\sigma}\frac{1}{\lambda\sqrt{n}}\log(2\eta^{-1}).$$

For the rest of this proof, to simplify notation and argument we will adopt the following conventions:

- the dependence of multiplicative constants C on various parameters will (generally) be omitted, except for σ, M, R, η and n which we want to track precisely .
- the expression "for n big enough" means that the statement holds for $n \ge n_0$, with n_0 potentially depending on all model parameters (including σ, M and R), but not on η.

From (5.22) and (5.23) , we conclude that

$$\mathbb{P}\left[\left\|B^s(f_\rho - f_\mathbf{z}^\lambda)\right\|_{\mathcal{H}_1} \ge a' + b'\log\eta^{-1}\right] \le \eta,$$

for all $\eta \in (0,1)$, with $a' := C_{\sigma,M,R}\max\left(\frac{1}{\lambda\sqrt{n}}, 1\right)$ and $b' := \frac{C_{\sigma,M}}{\lambda\sqrt{n}}$. On the other hand, Corollary 3.6, ensured that

$$\mathbb{P}\left[\left\|B^s(f_\rho - f_\mathbf{z}^{\lambda_{n,(\sigma,R)}})\right\|_{\mathcal{H}_1} \ge a + b\log\eta^{-1}\right] \le \eta, \quad \text{for } \log\eta^{-1} \le \log\eta_0^{-1} := C_{\sigma,R}n^{\frac{br}{2br+b+1}},$$

with $a = b := CR\left(\frac{\sigma}{R\sqrt{n}}\right)^{\frac{2b(r+s)}{2br+b+1}} = Ca_{n,(\sigma,R)}$, provided that n is big enough.

We can now apply Corollary C.2 in the appendix, which encapsulates some tedious computations, to conclude that for any $p \leq \frac{1}{2} \log \eta_0^{-1}$, and n big enough:

$$\mathbb{E}\left[\left\|B^s(f_\rho - f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})\right\|_{\mathcal{H}_1}^p\right] \leq C_p\left(a_{n,(\sigma,R)}^p + \eta_0\left((a')^p + 2(b' \log \eta_0^{-1})^p\right)\right) .$$

Now for fixed σ, M, R, and p, the quantities a', b' are powers of n, while $\eta_0 = \exp(-C_{\sigma,r} n^{\nu_{b,r}})$ for $\nu_{b,r} > 0$. The condition $p \leq \frac{1}{2} \log \eta_0^{-1}$ is thus satisfied for n large enough and we have

$$\limsup_{n\to\infty} \sup_{\rho \in \mathcal{M}_{\sigma,M,R}} \frac{\mathbb{E}_{\rho^{\otimes n}}\left[\left\|B^s(f_\rho - f_{\mathbf{z}}^{\lambda_{n,(\sigma,R)}})\right\|_{\mathcal{H}_1}^p\right]^{\frac{1}{p}}}{a_{n,(\sigma,R)}} \leq C ,$$

(where we reiterate that the constant C above may depend on all parameters including p, but not on σ, M nor R.). Therefore taking the supremum over (σ, M, R) yields the desired conclusion. $\square$

6. Proof of Lower Rate

Consider a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures on a measurable space $(\mathbf{Z}, \mathcal{A})$, indexed by Θ. Additionally, let $d : \Theta \times \Theta \longrightarrow [0, \infty)$ be a (semi-) distance.

For two probability measures P_1, P_2 on some common measurable space $(\mathbf{Z}, \mathcal{A})$, we recall the definition of the *Kullback-Leibler divergence* between P_1 and P_2

$$\mathcal{K}(P_1, P_2) := \int_{\mathcal{X}} \log\left(\frac{dP_1}{dP_2}\right) dP_1 ,$$

if P_1 is absolutely continuous with respect to P_2. If P_1 is not absolutely continuous with respect to P_2, then $\mathcal{K}(P_1, P_2) := \infty$. One easily observes that

$$\mathcal{K}(P_1^{\otimes n}, P_2^{\otimes n}) = n\, \mathcal{K}(P_1, P_2) .$$

In order to obtain minimax lower bounds we briefly recall the general reduction scheme, presented in Chapter 2 of [29]. The main idea is to find N_ε parameters $\theta_1, ..., \theta_{N_\varepsilon} \in \Theta$, depending on $\varepsilon < \varepsilon_0$ for some $\varepsilon_0 > 0$, with $N_\varepsilon \to \infty$ as $\varepsilon \to 0$, such that any two of these parameters are ε-separated with respect to the distance d, but that the associated distributions $P_{\theta_j} =: P_j \in \mathcal{P}$ have small Kullback-Leibler divergence to each other and are therefore statistically close. It is then clear that

$$(6.1) \quad \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{P}} \mathbb{E}_\theta[d^p(\hat{\theta}, \theta)]^{\frac{1}{p}} \geq \varepsilon \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{P}} \mathbb{P}_\theta[d(\hat{\theta}, \theta) \geq \varepsilon] \geq \varepsilon \inf_{\hat{\theta}} \max_{1 \leq j \leq N_\varepsilon} \mathbb{P}_j[d(\hat{\theta}, \theta_j) \geq \varepsilon] ,$$

where the infimum is taken over all estimators $\hat{\theta}$ of θ. The above RHS is then lower bounded through the following proposition which is a consequence of Fano's lemma, see [29], Theorem 2.5:

Proposition 6.1. *Assume that $N \geq 2$ and suppose that Θ contains $N+1$ elements $\theta_0, ..., \theta_N$ such that:*

(i) *For some $\varepsilon > 0$, and for any $0 \leq i < j \leq N$, $d(\theta_i, \theta_j) \geq 2\varepsilon$;*

(ii) *For any $j = 1, ..., N$, P_j is absolutely continuous with respect to P_0, and*

(6.2)
$$\frac{1}{N} \sum_{j=1}^{N} \mathcal{K}(P_j, P_0) \leq \omega \, \log(N) \, ,$$

for some $0 < \omega < 1/8$.

Then

$$\inf_{\hat{\theta}} \max_{1 \leq j \leq N} P_j \big(\, d(\hat{\theta}, \theta_j) \geq \varepsilon \, \big) \; \geq \; \frac{\sqrt{N}}{1 + \sqrt{N}} \left(1 - 2\omega - \sqrt{\frac{2\omega}{\log(N)}} \right) > 0 \, ,$$

where the infimum is taken over all estimators $\hat{\theta}$ of θ.

6.1. **Proof of Theorem 3.5.** We will apply the above general result to our target distance $d_s : \Omega_\nu(r, R) \times \Omega_\nu(r, R) \longrightarrow \mathbb{R}_+$, given by

$$d_s(f_1, f_2) \; = \; \| B^s(f_1 - f_2) \|_{\mathcal{H}_1} \, ,$$

with $s \in [0, \frac{1}{2}]$ and $\nu \in \mathcal{P}^>(b, \alpha)$. We will establish the lower bounds in the particular case where the distribution of Y given X is Gaussian with variance σ^2 (which satisfies the Bernstein moment condition (2.4) with $M = \sigma$). The main effort is to construct a finite subfamily belonging to the model of interest and suitably satisfying the assumptions of Proposition 6.1; this is the goal of the forthcoming propositions and lemmata.

Proposition 6.2. *Let $\nu \in \mathcal{P}^>(b, \alpha)$, for $b > 1$, $\alpha > 0$. Assume that $r > 0, R > 0$. To each $f \in \Omega_\nu(r, R)$ and $x \in \mathcal{X}$ we associate the following measure:*

(6.3)
$$\rho_f(dx, dy) := \rho_f(dy|x)\nu(dx) \, , \quad \text{where} \quad \rho_f(dy|x) := \mathcal{N}(Af(x), \sigma^2) \, .$$

Then:

(i) *The measure ρ_f belongs to the class $\mathcal{M}(r, R, \mathcal{P}^>(b, \alpha))$, defined in (2.6).*
(ii) *Given $f_1, f_2 \in \Omega_\nu(r, R)$, the Kullback-Leibler divergence between ρ_1 and ρ_2 satisfies*

$$\mathcal{K}(\rho_1, \rho_2) = \frac{1}{2\sigma^2} \big\| \sqrt{B}(f_1 - f_2) \big\|_{\mathcal{H}_1}^2 \, .$$

Proof. Point (i) follows directly from the definition of the class $\mathcal{M}(r, R, \mathcal{P}^>(b, \alpha))$. For point (ii), note that the Kullback-Leibler divergence between two Gaussian distributions with identical variance σ^2 and mean difference Δ is $\Delta/2\sigma^2$. Since ρ_1, ρ_2 have the same X-marginal ν, it holds

$$\mathcal{K}(\rho_1, \rho_2) = \mathbb{E}[\mathcal{K}(\rho_1(.|X), \rho_2(.|X))] = \frac{1}{2\sigma^2} \int \big(A(f_1 - f_2)(x) \big)^2 d\nu(x)$$

$$= \frac{1}{2\sigma^2} \| S(f_1 - f_2) \|_{L^2(\nu)}^2 = \frac{1}{2\sigma^2} \big\| \sqrt{B}(f_1 - f_2) \big\|_{\mathcal{H}_1}^2 \, .$$

$\square$

The following lemma is a variant from [9], Proposition 6, which will be useful in the subsequent proposition.

Lemma 6.3. *For any $m \geq 28$ there exist an integer $N_m > 3$ and $\pi_1, ..., \pi_{N_m} \in \{-1, +1\}^m$ such that for any $i, j \in \{1, ..., N_m\}$ with $i \neq j$ it holds*

$$(6.4) \qquad \log(N_m - 1) > \frac{m}{36} > \frac{2}{3} \,,$$

and

$$(6.5) \qquad \sum_{l=1}^{m} (\pi_i^l - \pi_j^l)^2 \geq m \,,$$

where $\pi_i = (\pi_i^1, ..., \pi_i^m)$.

Proposition 6.4. *Assume $\nu \in \mathcal{P}^>(b, \alpha)$. Let $0 \leq s \leq 1/2$, $R > 0$ and $r > 0$. For any $\varepsilon_0 > 0$ there exist $\varepsilon \leq \varepsilon_0$, $N_\varepsilon \in \mathbb{N}$ and functions $f_1, ..., f_{N_\varepsilon} \in \mathcal{H}_1$ satisfying*

(i) *$f_i \in \Omega_\nu(r, R)$ for any $i = 1, ..., N_\varepsilon$ and*

$$\| B^s (f_i - f_j) \|_{\mathcal{H}_1}^2 > \varepsilon^2 \,,$$

    for any $i, j = 1, ..., N_\varepsilon$ with $i \neq j$.

(ii) *Let $\rho_i := \rho_{f_i}$ be given by (6.3). Then it holds*

$$\mathcal{K}(\rho_i, \rho_j) \leq C_{b,r,s} \, R^2 \sigma^{-2} \left(\frac{\varepsilon}{R} \right)^{\frac{2r+1}{r+s}} \,,$$

    for any $i, j = 1, ..., N_\varepsilon$ with $i \neq j$.

(iii) *$\log(N_\varepsilon - 1) \geq C_{\alpha,b,r,s} \left(\frac{R}{\varepsilon} \right)^{\frac{1}{b(r+s)}}$.*

If ν belongs to the subclass $\mathcal{P}^>_{strong}(b, \alpha)$, then the assertions from (i), (ii) and (iii) are valid for all $\varepsilon > 0$ small enough (depending on the parameters r, R, s, α, b as well as j_0, γ coming from the choice of ν in $\mathcal{P}^>_{strong}(b, \alpha)$; the multiplicative constants in (ii), (iii) then also depend on γ.)

Proof. We first prove the proposition under the stronger assumption that ν belongs to $\mathcal{P}^>_{strong}(b, \alpha)$. We recall from (2.1) that we denote $(e_l)_{l \geq 1}$ an orthonormal family of $\mathcal{H}_1$ of eigenvectors of B corresponding to the eigenvalues $(\mu_l)_{l \geq 1}$, which satisfy by definition of $\mathcal{P}^>_{strong}(b, \alpha)$:

$$(6.6) \qquad \forall l \geq 0 \,:\, \mu_l \geq \alpha l^{-b}$$

and

$$(6.7) \qquad \forall l \geq l_0 \,:\, \mu_{2l} \geq 2^{-\gamma} \mu_l \,,$$

for some $l_0 \in \mathbb{N}$ and for some $\gamma > 0$. For any given $\varepsilon < R 2^{-\gamma(r+s)} \left(\frac{\alpha^{1/b}}{\max(28, l_0)} \right)^{b(r+s)}$ we pick $m = m(\varepsilon) := \max\{ l \geq 1 \,:\, \mu_l \geq 2^\gamma (\varepsilon R^{-1})^{\frac{1}{r+s}} \}$. Note that $m \geq \max(28, l_0)$, following from the choice of ε and from (6.6).

Let $N_m > 3$ and $\pi_1, ..., \pi_{N_m} \in \{-1, +1\}^m$ be given by Lemma 6.3 and define

$$(6.8) \qquad g_i := \frac{\varepsilon}{\sqrt{m}} \sum_{l=m+1}^{2m} \pi_i^{(l-m)} \left(\frac{1}{\mu_l} \right)^{r+s} e_l \,.$$

We have by (6.7) and from the definition of m

$$\|g_i\|_{\mathcal{H}_1}^2 = \frac{\varepsilon^2}{m} \sum_{l=m+1}^{2m} \left(\frac{1}{\mu_k}\right)^{2(r+s)} \leq \varepsilon^2 \mu_{2m}^{-2(r+s)} \leq \varepsilon^2 2^{2\gamma(r+s)} \mu_m^{-2(r+s)} \leq R^2 \,.$$

For $i = 1, ..., N_m$ let $f_i := B^r g_i \in \Omega_\nu(r, R)$, with g_i as in (6.8). Then

$$\| B^s(f_i - f_j) \|_{\mathcal{H}_1}^2 = \| B^{r+s}(g_i - g_j) \|_{\mathcal{H}_1}^2$$
$$= \frac{\varepsilon^2}{m} \sum_{l=m+1}^{2m} (\pi_i^{l-m} - \pi_j^{l-m})^2 \left(\frac{1}{\mu_l}\right)^{2(r+s)} \mu_l^{2(r+s)} \geq \varepsilon^2 \,,$$

by (6.5), and the proof of (i) is finished. For $i = 1, ..., N_\varepsilon$, let $\rho_i = \rho_{f_i}$ be defined by (6.3). Then, using the definition of m, the Kullback-Leibler divergence satisfies

$$\mathcal{K}(\rho_i, \rho_j) = \frac{1}{2\sigma^2} \left\| \sqrt{B}(f_i - f_j) \right\|_{\mathcal{H}_1}^2$$
$$= \frac{1}{2\sigma^2} \left\| B^{r+1/2}(g_i - g_j) \right\|_{\mathcal{H}_1}^2$$
$$= \frac{\varepsilon^2}{2\sigma^2 m} \sum_{l=m+1}^{2m} (\pi_i^{l-m} - \pi_j^{l-m})^2 \left(\frac{1}{\mu_l}\right)^{2(r+s)} \mu_l^{2r+1}$$
$$\leq 2\sigma^{-2} \mu_{m+1}^{1-2s} \varepsilon^2$$
$$\leq 2^{1+\gamma(1-2s)} \sigma^{-2} R^2 \left(\frac{\varepsilon}{R}\right)^{\frac{1+2r}{r+s}} \,,$$

which shows (ii). Finally, (6.4), (6.6), (6.7) and the definition of m imply

$$\log(N_m - 1) \geq \frac{m}{36} \geq \frac{\alpha^{1/b}}{36} \mu_m^{-1/b} \geq \frac{\alpha^{1/b}}{36} 2^{-\gamma/b} \mu_{2m}^{-1/b} \geq \frac{\alpha^{1/b}}{36} 2^{-2\gamma/b} \left(\frac{R}{\varepsilon}\right)^{\frac{1}{b(r+s)}} \,,$$

thus (iii) is established.

We now assume that ν belongs to $\mathcal{P}^>(b, \alpha)$ and only satisfies condition (6.6). Let any $\varepsilon_0 > 0$ be given. We pick $m \in \mathbb{N}$ satisfying $m \geq 28$ and the two following conditions:

$$(6.9) \qquad\qquad \mu_m \leq 2^{b+1}(R^{-1}\varepsilon_0)^{\frac{1}{r+s}} \,,$$

$$(6.10) \qquad\qquad \frac{\mu_{2m}}{\mu_m} \geq 2^{-b-1} \,.$$

Since the sequence of eigenvalues (μ_m) converges to 0, condition (6.9) must be satisfied for any m big enough, say $m \geq m_0(\varepsilon_0)$. Subject to that condition, we argue by contradiction that there must exist m satisfying (6.10). If that were not the case, we would have by immediate recursion for any $l > 0$, introducing $m' := 2^l m_0(\varepsilon_0)$:

$$\mu_{m'} < 2^{-l(b+1)} \mu_{m_0(\varepsilon_0)} = \left(\frac{m'}{m_0(\varepsilon_0)}\right)^{-b-1} \mu_{m_0(\varepsilon_0)} = C_{\varepsilon_0}(m')^{-(b+1)} \,,$$

which would (eventually, for l big enough) contradict (6.6). Therefore, there must exist an $m > m_0$ satisfying the required conditions. Now put

$$(6.11) \qquad \varepsilon := 2^{-(b+1)(r+s)} R \mu_m^{r+s} \leq \varepsilon_0 \,,$$

where the inequality is from requirement (6.9). For $i = 1, ..., N_m$, we define g_i as in (6.8). Then $\|g_i\|_{\mathcal{H}_1} \leq R$. Again, let $f_i := B^r g_i \in \Omega_\nu(r, R)$ and the same calculations as above (with γ replaced by $b+1$) lead to $(i), (ii)$ and (iii). $\qquad \square$

Now we are in the position to prove the minimax lower rate.

Proof of Theorem 3.5. Let the parameters $r, R, s, b, \alpha, \sigma$ be fixed for the rest of the proof, and the marginal distribution $\nu \in \mathcal{P}^>(b, \alpha)$ also be fixed.

Our aim is to apply Proposition 6.1 to the distance $d_s(f_1, f_2) := \|B^s(f_1 - f_2)\|_{\mathcal{H}_1}$ ($s \in [0, \frac{1}{2}]$), on the class $\Theta := \Omega_\nu(r, R)$, where for any $f \in \Omega_\nu(r, R)$, the associated distribution is $P_f := \rho_f^{\otimes n}$ with ρ_f defined as from Proposition 6.2 (i); more precisely, we will apply this proposition along a well-chosen sequence $(n_k, \varepsilon_k)_{k \geq 0}$. From Proposition 6.4, we deduce that there exists a decreasing null sequence $(\varepsilon_k) > 0$ such that for any ε belonging to the sequence, there exists N_ε and functions $f_1, ..., f_{N_\varepsilon}$ satisfying (i)-(ii)-(iii). In the rest of this proof, we assume $\varepsilon = \varepsilon_k$ is a value belonging to the null sequence. Point (i) gives requirement (i) of Proposition 6.1. We turn to requirement (6.2). Let $\rho_j = \rho_{f_j}$ be given by (6.3). Then by Proposition 6.4 (ii)-(iii):

$$\frac{1}{N_\varepsilon - 1} \sum_{j=1}^{N_\varepsilon - 1} \mathcal{K}(\rho_j^{\otimes n}, \rho_{N_\varepsilon}^{\otimes n}) = \frac{n}{N_\varepsilon - 1} \sum_{j=1}^{N_\varepsilon - 1} \mathcal{K}(\rho_j, \rho_{N_\varepsilon})$$

$$\leq n C_{b,r,s} \, R^2 \sigma^{-2} \left(\frac{\varepsilon}{R}\right)^{\frac{2r+1}{r+s}}$$

$$\leq n C_{\alpha,b,r,s} \, R^2 \sigma^{-2} \left(\frac{\varepsilon}{R}\right)^{\frac{2br+b+1}{b(r+s)}} \log(N_\varepsilon - 1)$$

$$=: \omega \, \log(N_\varepsilon - 1) \,.$$

Choosing $n := \left\lfloor \left(8 C_{\alpha,b,r,s} \, R^2 \sigma^{-2} \, (\varepsilon R^{-1})^{\frac{2br+2r+1}{b(r+s)}}\right)^{-1} \right\rfloor$ ensures $\omega \leq \frac{1}{8}$ and therefore requirement (6.2) is satisfied. Then Proposition 6.1 entails:

$$\inf_{\hat{f}_\bullet} \max_{1 \leq j \leq N_\varepsilon} \rho_j^{\otimes n} \left(\left\|B^s(\hat{f}_\bullet - f_j)\right\|_{\mathcal{H}_1} \geq \frac{\varepsilon}{2} \right) \geq \frac{\sqrt{N_\varepsilon - 1}}{1 + \sqrt{N_\varepsilon - 1}} \left(1 - 2\omega - \sqrt{\frac{2\omega}{\log(N_\varepsilon - 1)}}\right)$$

$$\geq \frac{1}{2}\left(\frac{3}{4} - \sqrt{\frac{3}{8}}\right)$$

$$> 0 \,.$$

This inequality holds for any (n_k, ε_k) for ε_k in the decreasing null sequence and n_k given by the above formula; we deduce that $n_k \to \infty$ with

$$\varepsilon_k \geq C_{\alpha,b,r,s} R \left(\frac{\sigma}{R\sqrt{n_k}} \right)^{\frac{2b(r+s)}{2br+b+1}} .$$

Thus, applying (6.1) and taking the limsup gives the result.

Now suppose that that ν belongs to $\mathcal{P}^>_{strong}(b, \alpha)$. Define $\varepsilon := R(8C_{\alpha,b,r,s})^{-\frac{b(r+s)}{2br+b+1}} \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{b(r+s)}{2br+b+1}}$, then for any n sufficiently large, points (i)-(ii)-(iii) of Proposition 6.4 will hold. The same calculations as above now hold for any n large enough; finally taking the liminf finishes the proof. $\qquad\square$

Proof of Corollary 3.6. The main point is only to ensure that the *strong* minimax lower bound applies, for this we simply check that $\mathcal{P}^>_{strong}(b, \alpha) \supset \mathcal{P}' = \mathcal{P}^<(b, \beta) \cap \mathcal{P}^>(b, \alpha)$. For any $\nu \in \mathcal{P}^<(b, \beta) \cap \mathcal{P}^>(b, \alpha)$, the eigenvalues of the operator B_ν satisfy $\alpha j^{-b} \leq \mu_j \leq \beta j^{-b}$ for all $j \geq 1$. It follows that for any $j \geq 1$:

$$\frac{\mu_{2j}}{\mu_j} \geq \frac{\alpha}{\beta} 2^{-b},$$

so that the conditions for $\nu \in \mathcal{P}^>_{strong}(b, \alpha)$ are met (with parameters $\gamma := b + \log_2 \frac{\beta}{\alpha}$, $l_0 = 1$). Since $\mathcal{P}'$ is assumed to be nonempty, for any $\nu \in \mathcal{P}'$ the strong lower minimax bound of Theorem 3.5 applies to the family $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r, R, \{\nu\})$ and a fortiori to the family $\mathcal{M}_{R,M,\sigma} := \mathcal{M}(r, R, \mathcal{P}')$ whose models are larger. On the other hand since $\mathcal{M}(r, R, \mathcal{P}') \subset \mathcal{M}(r, R, \mathcal{P}^<(b, \beta))$ the upper bound of Theorem 3.4 applies and we are done. $\qquad\square$

APPENDIX A. PROOF OF CONCENTRATION INEQUALITIES

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 $m \geq 2$*

$$(A.1) \qquad \mathbb{E}\big[\|\xi - \mathbb{E}[\xi]\|^m_{\mathcal{H}_1} \big] \leq \frac{1}{2} m! \sigma^2 L^{m-2}.$$

If the sample $z_1, ..., z_n$ is drawn i.i.d. from Z according to $\mathbb{P}$, then, for any $0 < \eta < 1$, with probability greater than $1 - \eta$

$$(A.2) \qquad \left\| \frac{1}{n} \sum_{j=1}^n \xi(z_j) - \mathbb{E}[\xi] \right\|_{\mathcal{H}} \leq \delta(n, \eta),$$

where

$$\delta(n, \eta) \;\; := \;\; 2\log(2\eta^{-1}) \left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}} \right).$$

In particular, (A.1) *holds if*

$$\|\xi(z)\|_{\mathcal{H}_1} \leq \frac{L}{2} \quad a.s. \,,$$

$$\mathbb{E}\big[\, \|\xi\|_{\mathcal{H}_1}^2\,\big] \leq \sigma^2.$$

Proof. See [7], from the original result of [24] (Corollary 1). $\qquad\square$

Proof of Proposition 5.2. Define $\xi_1 : \mathcal{X} \times \mathbb{R} \longrightarrow \mathcal{H}_1$ by

$$\xi_1(x,y) := (\bar{B} + \lambda)^{-1/2} \kappa^{-2} (y - Af(x)) F_x \,,$$

abusing notation we also denote ξ_1 the random variable $\xi_1(X,Y)$ where $(X,Y) \sim \rho$. The model assumption (2.3) implies

$$\mathbb{E}[\xi_1] = \kappa^{-2}(\bar{B} + \lambda)^{-1/2} \int_{\mathcal{X}} F_x \int_{\mathbb{R}} (y - Af_\rho(x))\, \rho(dy|x)\nu(dx)$$

$$= (B + \lambda)^{-1/2} \int_{\mathcal{X}} F_x (Af_\rho(x) - Af_\rho(x))\, \nu(dx)$$

$$= 0\,,$$

and therefore

$$\frac{1}{n} \sum_{j=1}^{n} \xi_1(x_j, y_j) - \mathbb{E}[\xi_1] = \frac{1}{n} \sum_{j=1}^{n} (\bar{B} + \lambda)^{-1/2} \kappa^{-2} (y_j - Af_\rho(x_j)) F_{x_j}$$

$$= (\bar{B} + \lambda)^{-1/2} \kappa^{-2} S_{\mathbf{x}}^\star\, (\mathbf{y} - S_{\mathbf{x}} f_\rho)\,.$$

$$= (\bar{B} + \lambda)^{-1/2} \big(\bar{S}_{\mathbf{x}}^\star \mathbf{y} - \bar{B}_{\mathbf{x}} f_\rho\big)\,.$$

Moreover, by assumption (2.4), for $m \geq 2$:

$$\mathbb{E}[\, \|\xi_1\|_{\mathcal{H}_1}^m\,] = \int_{\mathcal{X} \times \mathbb{R}} \big\|\kappa^{-2}(y - Af_\rho(x))(\bar{B} + \lambda)^{-1/2} F_x\big\|_{\mathcal{H}_1}^m \rho(dx, dy)$$

$$= \int_{\mathcal{X}} \kappa^{-2m} \big\|(\bar{B} + \lambda)^{-1/2} F_x\big\|_{\mathcal{H}_1}^m \int_{\mathbb{R}} |y - Af_\rho(x)|^m\, \rho(dy|x)\nu(dx)$$

$$\leq \frac{1}{2} m! \sigma^2 M^{m-2} \kappa^{-2m} \sup_{x \in \mathcal{X}} \big\|(\bar{B} + \lambda)^{-1/2} F_x\big\|_{\mathcal{H}_1}^{m-2} \int_{\mathcal{X}} \mathrm{tr}\left((\bar{B} + \lambda)^{-1} F_x \otimes F_x^\star\right)\, \nu(dx)$$

$$\leq \frac{1}{2} m! \kappa^{-m} \sigma^2 M^{m-2} \lambda^{-\frac{m-2}{2}} \mathrm{tr}\left((\bar{B} + \lambda)^{-1/2} \kappa^{-2} \int_{\mathcal{X}} F_x \otimes F_x^\star\, \nu(dx)\right)$$

$$= \frac{1}{2} m! \left(\kappa^{-1} \sigma \sqrt{\mathcal{N}(\lambda)}\right)^2 \left(\frac{\kappa^{-1} M}{\sqrt{\lambda}}\right)^{m-2}\,.$$

As a result, Proposition $A.1$ implies with probability at least $1 - \eta$

$$\big\|(\bar{B} + \lambda)^{-1/2}(\bar{B}_{\mathbf{x}} f_\rho - \bar{S}_{\mathbf{x}}^\star \mathbf{y})\big\|_{\mathcal{H}_1} \leq \delta_1(n, \eta)\,,$$

where

$$\delta_1(n, \eta) = 2 \log(2\eta^{-1}) \kappa^{-1} \left(\frac{M}{n\sqrt{\lambda}} + \frac{\sigma}{\sqrt{n}} \sqrt{\mathcal{N}(\lambda)}\right).$$

For the second part of the proposition, we introduce similarly $\xi_1'(x,y) := \kappa^{-2}(y - Af(x))F_x$, which satisfies

$$\mathbb{E}[\xi_1'] = 0\,; \qquad \frac{1}{n}\sum_{j=1}^n \xi_1'(x_j, y_j) - \mathbb{E}[\xi_1'] = \bar{S}_{\mathbf{x}}^\star \mathbf{y} - \bar{B}_{\mathbf{x}} f_\rho\,,$$

and

$$\mathbb{E}\big[\,\|\xi_1'\|_{\mathcal{H}_1}^m\,\big] \leq \kappa^{-m}\mathbb{E}\big[\,|y - Af(x)|^m\,\big] \leq \frac{1}{2}m!\left(\kappa^{-1}\sigma\right)^2\left(\kappa^{-1}M\right)^{m-2}\,.$$

Applying Proposition $A.1$ yields the result. $\qquad\square$

Proof of Proposition 5.3. We proceed as above by defining $\xi_2 : \mathfrak{X} \longrightarrow \mathrm{HS}(\mathcal{H}_1)$ (the latter denoting the space of Hilbert-Schmidt operators on $\mathcal{H}_1$) by

$$\xi_2(x) := (\bar{B} + \lambda)^{-1}\kappa^{-2}B_x\,,$$

where $B_x := F_x \otimes F_x^\star$. We also use the same notation ξ_2 for the random variable $\xi_2(X)$ with $X \sim \nu$. Then,

$$\mathbb{E}[\xi_2] = (\bar{B}+\lambda)^{-1}\kappa^{-2}\int_{\mathfrak{X}} B_x\,\nu(dx) = (\bar{B}+\lambda)^{-1}\bar{B}\,,$$

and therefore

$$\frac{1}{n}\sum_{j=1}^n \xi_2(x_j) - \mathbb{E}[\xi_2] = (\bar{B}+\lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}})\,.$$

Furthermore, since $\bar{B}_x$ is of trace class and $(\bar{B}+\lambda)^{-1}$ is bounded, we have using Assumption 2.1

$$\|\xi_2(x)\|_{\mathrm{HS}} \leq \big\|(\bar{B}+\lambda)^{-1}\big\|\,\kappa^{-2}\,\|B_x\|_{\mathrm{HS}} \leq \lambda^{-1} =: L_2/2\,,$$

uniformly for any $x \in \mathfrak{X}$. Moreover,

$$\mathbb{E}\big[\,\|\xi_2\|_{\mathrm{HS}}^2\,\big] = \kappa^{-4}\int_{\mathfrak{X}}\mathrm{tr}\left(B_x(\bar{B}+\lambda)^{-2}B_x\right)\nu(dx)$$

$$\leq \big\|(\bar{B}+\lambda)^{-1}\big\|\,\kappa^{-4}\int_{\mathfrak{X}}\|B_x\|\,\mathrm{tr}\left((\bar{B}+\lambda)^{-1}B_x\right)\nu(dx)$$

$$\leq \frac{\mathcal{N}(\lambda)}{\lambda} =: \sigma_2^2\,.$$

Thus, Proposition $A.1$ applies and gives with probability at least $1 - \eta$

$$\big\|(\bar{B}+\lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}})\big\|_{\mathrm{HS}} \leq \delta_2(n,\eta)$$

with

$$\delta_2(n,\eta) = 2\log(2\eta^{-1})\left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right)\,.$$

$\qquad\square$

Proof of Proposition 5.4. We write the Neumann series identity

$$(\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda) = (I - \bar{S}_{\mathbf{x}}(\lambda))^{-1} = \sum_{j=0}^{\infty} \bar{S}_{\mathbf{x}}^j(\lambda) \, ,$$

with

$$\bar{S}_{\mathbf{x}}(\lambda) = (\bar{B} + \lambda)^{-1}(\bar{B} - \bar{B}_{\mathbf{x}}) \, ;$$

it is well known that the series converges in norm provided that $\left\| \bar{S}_{\mathbf{x}}(\lambda) \right\| < 1$. In fact, applying Proposition 5.3 gives with probability at least $1 - \eta$:

$$\left\| \bar{S}_{\mathbf{x}}(\lambda) \right\| \; \leq \; 2\log(2\eta^{-1}) \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \right) \, .$$

Put $C_\eta := 2\log(2\eta^{-1}) > 1$ for any $\eta \in (0, 1)$. Assumption (5.2) reads $\sqrt{n\lambda} \geq 4C_\eta\sqrt{\max(\mathcal{N}(\lambda), 1)}$, implying $\sqrt{n\lambda} \geq 4C_\eta \geq 4$ and therefore $\frac{2}{n\lambda} \leq \frac{1}{2\sqrt{n\lambda}} \leq \frac{1}{8C_\eta}$, hence

$$C_\eta \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \right) \leq C_\eta \left(\frac{1}{8C_\eta} + \frac{1}{4C_\eta} \right) < \frac{1}{2} \, .$$

Thus, with probability at least $1 - \eta$:

$$\left\| (\bar{B}_{\mathbf{x}} + \lambda)^{-1}(\bar{B} + \lambda) \right\| \leq 2 \, .$$

$\square$

Proof of Proposition 5.5. Defining $\xi_3 : \mathcal{X} \longrightarrow \mathrm{HS}(\mathcal{H}_1)$ by

$$\xi_3(x) := \kappa^{-2} F_x \otimes F_x^\star = \kappa^{-2} B_x$$

and denoting also, as before, ξ_3 for the random variable $\xi_3(X)$ (with $X \sim \nu$), we have $\mathbb{E}[\xi_3] = \bar{B}$ and therefore

$$\frac{1}{n} \sum_{j=1}^{n} \xi_3(x_j) - \mathbb{E}[\xi_3] = (\bar{B}_{\mathbf{x}} - \bar{B}) \, .$$

Furthermore, by Assumption 2.1

$$\left\| \xi_3(x) \right\|_{\mathrm{HS}} = \kappa^{-2} \left\| F_x \right\|^2 \leq 1 =: \frac{L_3}{2} \qquad \text{a.s.} \, ,$$

also leading to $\mathbb{E}[\|\xi_2\|_{\mathrm{HS}}^2] \leq 1 =: \sigma_3^2$. Thus, Proposition $A.1$ applies and gives with probability at least $1 - \eta$

$$\left\| \bar{B} - \bar{B}_{\mathbf{x}} \right\|_{\mathrm{HS}} \; \leq \; 6\log(2\eta^{-1}) \frac{1}{\sqrt{n}} \, .$$

$\square$

APPENDIX B. PERTURBATION RESULT

The estimate of the following proposition is crucial for proving the upper bound in case the source condition is of Hölder type r with $r > 1$. We remark that for $r > 1$ the function $t \mapsto t^r$ is not operator monotone. One might naively expect estimate $(B.1)$ to hold for a constant C given by the Lipschitz-constant of the scalar function t^r. As shown in [3], this is false even for finite dimensional positive matrices. The point of Proposition B.1 is that $(B.1)$ still holds for some larger constant depending on r and the upper bound of the spectrum. We do not expect this result to be particularly novel, but tracking down a proof in the literature proved elusive, not to mention that occasionally erroneous statements about related issues can be found. For this reason we here provide a self-contained proof for completeness sake.

Proposition B.1. *Let B_1, B_2 be two non-negative self-adjoint operators on some Hilbert space with $||B_j|| \leq a$, $j = 1, 2$, for some $a > 0$. Assume the B_j belong to the Schatten class $\mathbb{S}^p$ for $1 \leq p \leq \infty$. If $1 < r$, then*

$$(B.1) \qquad ||B_1^r - B_2^r||_p \ \leq \ C_r a^{r-1} \, ||B_1 - B_2||_p \ ,$$

for some $C_r < \infty$. This inequality also holds in operator norm for non-compact bounded (non-negative and self-adjoint) B_j.

Proof. We extend the proof of [14], given there in the case $r = 3/2$ in operator norm. We also restrict ourselves to the case $a = 1$. On $\mathcal{D} := \{z : |z| \leq 1\}$, we consider the functions $f(z) = (1 - z)^r$ and $g(z) = (1 - z)^{r-1}$. The proof is based on the power series expansions

$$f(z) = \sum_{n \geq 0}^{\infty} b_n z^n \quad \text{and} \quad g(z) = \sum_{n \geq 0}^{\infty} c_n z^n \ ,$$

which converge absolutely on $\mathcal{D}$. To ensure absolute convergence on the boundary $|z| = 1$, notice that

$$c_n = \frac{1}{n!} g^{(n)}(0) = \frac{(-1)^n}{n!} \prod_{j=1}^{n} (r - j) \ ,$$

so that all coefficients c_n for $n \geq r$ have the same sign $s := (-1)^{\lfloor r \rfloor}$ (if r is an integer these coefficients vanish without altering the argument below) implying for any $N > r$:

$$\sum_{n=0}^{N} |c_n| = \sum_{n=0}^{\lfloor r \rfloor} |c_n| + s \sum_{n=\lfloor r \rfloor+1}^{N} c_n = \sum_{n=0}^{\lfloor r \rfloor} |c_n| + s \lim_{z \nearrow 1} \sum_{n=\lfloor r \rfloor+1}^{N} c_n z^n$$

$$\leq \sum_{n=0}^{\lfloor r \rfloor} |c_n| + s \lim_{z \nearrow 1} \left(g(z) - \sum_{n=0}^{\lfloor r \rfloor} c_n \right)$$

$$= 2 \sum_{i=0}^{\lfloor r/2 \rfloor} |c_{\lfloor r \rfloor - 2i}| \ .$$

A bound for $\sum_n |b_n|$ can be derived analogously. Since $f(1 - B_j) = B_j^r$, we obtain

$$\|B_1^r - B_2^r\|_p \leq \sum_{n=0}^{\infty} |b_n| \, \|(I - B_1)^n - (I - B_2)^n\|_p \, .$$

Using the algebraic identity $T^{n+1} - S^{n+1} = T(T^n - S^n) + (T - S)S^n$, the triangle inequality and making use of $\|TS\|_p \leq \|T\| \, \|S\|_p$ for $S \in \mathcal{S}^p$, T bounded, the reader can easily convince himself by induction that

- for $j = 1, 2$, $B_j \in \mathcal{S}^p$ imply $(I - B_1)^n - (I - B_2)^n \in \mathcal{S}^p$ and
- $\|(I - B_1)^n - (I - B_2)^n\|_p \leq n\|B_1 - B_2\|_p$.

From $f'(z) = -rg(z)$ we have the relation $|b_n| = \frac{r}{n}|c_{n-1}|$, $n \geq 1$. Collecting all pieces leads to

$$\|B_1^r - B_2^r\|_p \leq \|B_1 - B_2\|_p \sum_{n=0}^{\infty} n|b_n| = r\|B_1 - B_2\|_p \sum_{n=0}^{\infty} |c_n| \, .$$

$\square$

Appendix C. Auxiliary technical lemmata

Lemma C.1. *Let X be a nonnegative real random variable such that the following holds:*

(C.1) $$\mathbb{P}[X > F(t)] \leq t \, , \quad \text{for all } t \in (0, 1] \, ,$$

where F is a monotone nonincreasing function $(0, 1] \to \mathbb{R}_+$. Then

$$\mathbb{E}[X] \leq \int_0^1 F(u)du \, .$$

Proof. An intuitive, non-rigorous proof is as follows. Let G be the tail distribution function of X, then it is well known that $\mathbb{E}[X] = \int_{\mathbb{R}_+} G$. Now it seems clear that $\int_{\mathbb{R}_+} G = \int_0^1 G^{-1}$, where G^{-1} is the upper quantile function for X. Finally F is an upper bound on G^{-1}.

Now for a rigorous proof, we can assume without loss of generality that F is left continuous: replacing F by its left limit in all points of $(0, 1]$ can only make it larger since it is nonincreasing, hence (C.1) is still satisfied, moreover since a monotone function has an at most countable number of discontinuity points, this operation does not change the value of the integral $\int_0^1 F$. Define the following pseudo-inverse for $x \in \mathbb{R}_+$:

$$F^{\dagger}(x) := \inf \left\{ t \in (0, 1] : F(t) < x \right\} \, ,$$

with the convention $\inf \emptyset = 1$. Denote $\widetilde{U} := F^{\dagger}(X)$. From the definition of $F^{\dagger}$ and the monotonicity of F it holds that $F^{\dagger}(x) < t \Rightarrow x > F(t)$ for all $(x, t) \in \mathbb{R}_+ \times (0, 1]$. Hence for any $t \in (0, 1]$

$$\mathbb{P}[\widetilde{U} < t] \leq \mathbb{P}[X > F(t)] \leq t \, ,$$

implying that for all $t \in [0, 1]$, $\mathbb{P}[\widetilde{U} \leq t] \leq t$, i.e. $\widetilde{U}$ is stochastically larger than a uniform variable on $[0, 1]$. Furthermore, by left continuity of F, one can readily check that $F(F^{\dagger}(x)) \geq x$ if $x \leq F(0)$. Since $\mathbb{P}[X > F(0)] = 0$, we can replace X by $\widetilde{X} :=$

$\min(X, F(0))$ without changing its distribution (nor that of $\widetilde{U}$). With this modification, in then holds that $F(\widetilde{U}) = F(F^\dagger(\widetilde{X})) \geq \widetilde{X}$. Hence

$$\mathbb{E}[X] = \mathbb{E}[\widetilde{X}] \leq \mathbb{E}[F(\widetilde{U})] \leq \mathbb{E}[F(U)] = \int_0^1 F(u)du\,,$$

where U is a uniform variable on $[0,1]$, and the second equality holds since F is nonincreasing. $\qquad\square$

Corollary C.2. *Let X be a nonnegative random variable and $t_0 \in (0,1)$ such that the following holds:*

(C.2) $$\mathbb{P}[X > a + b \log t^{-1}] \leq t\,, \text{ for all } t \in (t_0, 1]\,, \text{ and}$$

(C.3) $$\mathbb{P}[X > a' + b' \log t^{-1}] \leq t\,, \text{ for all } t \in (0, 1]\,,$$

where a, b, a', b' are nonnegative numbers. Then for any $p \leq \frac{1}{2}\log t_0^{-1}$:

$$\mathbb{E}[X^p] \leq C_p \left(a^p + b^p \Gamma(p+1) + t_0 \left((a')^p + 2(b' \log t_0^{-1})^p\right)\right)\,,$$

with $C_p := \max(2^{p-1}, 1)$.

Proof. Let $F(t) := \mathbf{1}\{t \in (t_0, 1]\}(a + b \log t^{-1}) + \mathbf{1}\{t \in (0, t_0]\}(a' + b' \log t^{-1})$. Then F is nonnegative, nonincreasing on $(0,1]$ and

$$\mathbb{P}[X^p > F^p(t)] \leq t$$

for all $t \in (0,1]$. Applying Lemma C.1, we find

(C.4) $$\mathbb{E}[X^p] \leq \int_0^{t_0} (a' + b' \log t^{-1})^p dt + \int_{t_0}^1 (a + b \log t^{-1})^p dt\,.$$

Using $(x+y)^{p-1} \leq C_p(x^{p-1} + y^{p-1})$ for $x, y \geq 0$, where $C_p = \max(2^{p-1}, 1)$, we upper bound the second integral in (C.4) via

$$\int_{t_0}^1 (a + b \log t^{-1})^p dt \leq C_p \left(a^p + b^p \int_0^1 (\log t^{-1})^p dt\right) = C_p \left(a^p + b^p \Gamma(p+1)\right)\,.$$

Concerning the first integral in (C.4), we write similarly

$$\int_0^{t_0} (a' + b' \log t^{-1})^p dt \leq C_p \left(t_0(a')^p + (b')^p \int_0^{t_0} (\log t^{-1})^p dt\right)$$
$$= C_p \left(t_0(a')^p + (b')^p \Gamma(p+1, \log t_0^{-1})\right)\,,$$

by the change of variable $u = \log t^{-1}$, where Γ is the incomplete gamma function. We use the following coarse bound: it can be checked that $t \mapsto t^p e^{-\frac{t}{2}}$ is decreasing for $t \geq 2p$, hence, putting $x := \log t_0^{-1}$,

$$\Gamma(p, x) = \int_x^\infty t^p e^{-t} dt \leq x^p e^{-\frac{x}{2}} \int_x^\infty e^{-\frac{t}{2}} dt = 2x^p e^{-x} = 2t_0(\log t_0^{-1})^p\,,$$

provided $x = \log t_0^{-1} \geq 2p$. Collecting all the above pieces we get the conclusion. $\qquad\square$

REFERENCES

[1] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

[2] R. Bhatia. *Matrix Analysis.* Springer, 1997.

[3] R. Bhatia and J. Holbrook. Fréchet derivatives of the power function. *Indiana University Mathematics Journal*, 49 (3):1155–1173, 2000.

[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, 1987.

[5] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Analysis*, 45(6):2610–2636, 2007.

[6] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98(462):324–339, 2003.

[7] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MIT, 2006.

[8] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

[9] E. De Vito and A. Caponnetto. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

[10] E. De Vito, L.Rosasco, A. Caponnetto, and U. De Giovannini. Learning from examples as an inverse problem. *J. of Machine Learning Research*, 6:883–904, 2005.

[11] E. De Vito, L. Rosasco, and Caponnetto. Discretization error analysis for Tikhonov regularization. *Analysis and Applications*, 4(1):81–99, 2006.

[12] R. DeVore, G. Kerkyacharian, D. Picard, and V.Temlyakov. Mathematical methods for supervised learning. *Foundations of Computational Mathematics*, 6(1):3–58, 2006.

[13] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer Academic Publishers, 2000.

[14] K. Fukumizu, F. R. Bach, and G. A. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

[15] L. L. Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[16] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7(2):219–269, 1993.

[17] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression.* Springer, 2002.

[18] P. Halmos and V. Sunder. *Bounded Integral Operators on L^2-Spaces.* Springer, 1978.

[19] S. Loustau. Inverse statistical learning. *Electron. J. Statist.*, 7:2065–2097, 2013.

[20] S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. *Bernoulli*, 21(1):176–208, 02 2015.

[21] P. Mathé and S. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789, 2003.

[22] S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[23] F. O'Sullivan. Convergence characteristics of methods of regularization estimators for nonlinear operator equations. *SIAM J. Numer. Anal.*, 27(6):1635–1649, 1990.

[24] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large devia-tions. *Theory Probab. Appl.*, 30(1):143–148, 1985.

[25] S. Smale and D. Zhou. Shannon sampling II: Connections to learning theory. *Appl. Comput. Harmon. Analysis*, 19(3):285–302, 2005.

[26] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximation. *Constructive Approximation*, 26(2):153–172, 2007.
[27] I. Steinwart and A. Christman. *Support Vector Machines*. Springer, 2008.
[28] V. Temlyakov. Approximation in learning theory. *Constructive Approximation*, 27(1):33–74, 2008.
[29] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.
[30] G. Wahba. *Spline Models for Observational Data*, volume 59. SIAM CBMS-NSF Series in Applied Mathematics, 1990.
[31] C. Wang and D.-X. Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.
[32] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

INSTITUTE OF MATHEMATICS, UNIVERSITY OF POTSDAM, KARL-LIEBKNECHT-STRAE 24-25 14476 POTSDAM, GERMANY
   E-mail address: {blanchard,nmuecke}@uni-potsdam.de