Machine Learning
https://doi.org/10.1007/s10994-022-06284-8

Inverse learning in Hilbert scales

Abhishake Rastogi[1] · **Peter Mathé**[2]

Received: 28 February 2020 / Revised: 18 October 2021 / Accepted: 16 November 2022

Abstract
We study linear ill-posed inverse problems with noisy data in the framework of statistical learning. The corresponding linear operator equation is assumed to fit a given Hilbert scale, generated by some unbounded self-adjoint operator. Approximate reconstructions from random noisy data are obtained with general regularization schemes in such a way that these belong to the domain of the generator. The analysis has thus to distinguish two cases, the regular one, when the true solution also belongs to the domain of the generator, and the 'oversmoothing' one, when this is not the case. Rates of convergence for the regularized solutions will be expressed in terms of certain distance functions. For solutions with smoothness given in terms of source conditions with respect to the scale generating operator, then the error bounds can then be made explicit in terms of the sample size.

Keywords Statistical inverse problem · Spectral regularization · Hilbert Scales · Reproducing kernel Hilbert space · Minimax convergence rates

Mathematics Subject Classification Primary: 62G20 · Secondary: 62G08 · 65J15 · 65J20 · 65J22

1 Introduction

We consider learning in linear inverse problems in Hilbert space. Within the classical framework of supervised learning, we are given data $\{(x_i, y_i)\}_{i=1}^m$ which follow the model

Editor: Steve Hanneke.

✉ Abhishake Rastogi
abhishake@tu-berlin.de

Peter Mathé
peter.mathe@wias-berlin.de

[1] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany

[2] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \dots, m, \tag{1}$$

where ε_i is the observational noise, and m denotes the sample size. The function g is unknown, belonging to some reproducing kernel Hilbert space, say $\mathcal{H}'$. The goal is to learn it from the given data. To be more precise, we assume that the random observations $\{(x_i, y_i)\}_{i=1}^m$ are independent and follow some unknown probability distribution ρ, defined on the sample space $Z = X \times Y$. Further, we assume that the input space X is a Polish space, and that the output space Y is a real separable Hilbert space.

In inverse learning, the function g from (1) is driven by some element f in a Hilbert space $\mathcal{H}$ via a mapping $A : \mathcal{H} \to \mathcal{H}'$ as

$$A(f) = g, \qquad \text{for} \quad f \in \mathcal{H} \quad \text{and} \quad g \in \mathcal{H}'. \tag{2}$$

In the present study, this mapping is assumed to be a bounded linear (smoothing) mapping, and it is also assumed to be injective to have the correspondence g to f unique. The unique solution of (2) is denoted by f_ρ. Literature for this setup is scarce, and we mention (Blanchard & Mücke, 2018), and a related study Rastogi et al. (2020), in which the underlying mapping A is assumed to be non-linear.

Often the sought for element f_ρ is known to have additional features, as e.g. smoothness and the standard approaches for reconstruction of an approximation of f_ρ do not take this into account. Therefore, we shall analyze such inverse learning problems in scales of Hilbert spaces. This topic has a long history within the classical setup of regularization theory, starting from (Natterer, 1984), see also the monograph (Engl et al., 1996, Chapt. 8). In most cases, the scale of Hilbert spaces is assumed to be a scale of Sobolev spaces, and the smoothing properties of the underlying operator A are measured with respect to this scale. This allows for a mathematical analysis, even if the singular value decomposition of A cannot be used to design a regularization scheme. Also, solution smoothness, i.e., the smoothness of f_ρ is described by assuming that it belongs to some space within this scale. Recently, regularization in Hilbert scales gained interest in statistical inverse problems, especially for the Bayesian approach to such problems, where we mention the studies (Gugushvili et al., 2020), and more recently (Agapiou & Mathé, 2022). To the best of our knowledge, inverse learning problems in scales have not been studied, yet.

Here we highlight the following prototypical example.

Example Let $A : \mathscr{L}_0^2(0,1) \to \mathcal{H}_0^1(0,1)$ be the integration operator

$$(Af)(x) := \int_0^x f(t)\, dt, \quad x \in (0,1), \tag{3}$$

where $\mathcal{H}_0^1(0,1)$ denotes the Sobolev space of abs. continuous functions g with $g(0) = g(1) = 0$, and $\mathscr{L}_0^2(0,1)$ consists of elements which integrate to zero, i. e., $(Af)(1) = 0$. Thus, we are looking for finding the derivative of a given function, one of the most classical inverse problems. In the above formulation, the operator A is injective. Moreover, it is known that the Sobolev space $\mathcal{H}' := \mathcal{H}_0^1(0,1)$ is a reproducing kernel Hilbert space. Details are given in Blanchard and Mücke (2018). Therefore, a suitable scale of Hilbert spaces is the class of Sobolev spaces $\mathcal{H}_0^s(0,1)$, $s \in [0,p]$ for some $p \geq 1$. For such analysis to work we assume that the given operator A 'fits the scale', which will be expressed in terms of a link condition. For the above example, the operator A has step one, meaning that elements from $\mathscr{L}_0^2(0,1)$ are mapped to $\mathcal{H}_0^1(0,1)$. Moreover, in this context, smoothness is also given relative to this scale, as e.g., $f_\rho \in \mathcal{H}_0^s(0,1)$ for some $0 < s \leq p$.

This is significantly different from other works, where smoothness is relative to the underlying covariance operator, and hence cannot be verified.

Further examples of Hilbert Scales relevant for learning in Reproducing Kernel Hilbert Spaces can be find in Mücke and Reiss (2020).

More generally, in the present study we shall assume that there is an unbounded self-adjoint operator $L : \mathcal{D}(L) \subset \mathcal{H} \to \mathcal{H}$, which generates a scale of Hilbert spaces $\mathcal{H}_s := \mathcal{D}(L^s)$, $s \geq 0$. Both, the operator Eq. (2), and the solution smoothness are assumed to fit this scale by assumptions, made below.

We highlight one specific means of reconstruction, often called *penalized least squares*. In this standard approach the estimator $f_{\mathbf{z},\lambda}$ is the minimizer of

$$\frac{1}{m}\sum_{i=1}^{m}\left\|A(f)(x_i) - y_i\right\|_Y^2 + \lambda\|f\|_{\mathcal{H}}^2, \tag{4}$$

where λ is a positive regularization parameter which balances the error term and the penalty $\|f\|_{\mathcal{H}}^2$. This penalty will control the norm (in $\mathcal{H}$) of the minimizer, but it cannot incur additional properties. Here, we implement such additional properties by assuming that all considered minimizers $f_{\mathbf{z},\lambda}$, which are taken into account belong to $\mathcal{D}(L)$. In the analysis of inverse problems this setup has a long history, starting from the above-mentioned study (Natterer, 1984), and it has since then been frequently considered both for linear (Böttcher et al., 2006; Mair, 1994; Mathé & Tautenhahn, 2006, 2007; Nair, 1999, 2002; Nair et al., 2005; Neubauer, 1988; Tautenhahn, 1996), and for non-linear mappings A Hofmann and Mathé (2018, 2020). The additional information $f_{\mathbf{z},\lambda} \in \mathcal{D}(L)$ is taken into account by replacing the above minimization problem (4) by

$$\frac{1}{m}\sum_{i=1}^{m}\left\|A(f)(x_i) - y_i\right\|_Y^2 + \lambda\|Lf\|_{\mathcal{H}}^2, \tag{5}$$

with minimzer $f_{\mathbf{z},\lambda} \in \mathcal{D}(L)$, and we may formally introduce $u_{\mathbf{z},\lambda} := Lf_{\mathbf{z},\lambda} \in \mathcal{H}$.

In the regular case, when $f_\rho \in \mathcal{D}(L)$, then we let $u_\rho := Lf_\rho \in \mathcal{H}$. With this notation we can rewrite (2) as

$$g = Af = AL^{-1}u, \quad u \in \mathcal{D}(L).$$

Then the Tikhonov minimization problem (5) would reduce to the standard one

$$\frac{1}{m}\sum_{i=1}^{m}\left\|(AL^{-1})(u)(x_i) - y_i\right\|_Y^2 + \lambda\|u\|_{\mathcal{H}}^2, \tag{6}$$

albeit for a different operator AL^{-1}. Accordingly, the error bounds relate as

$$\left\|f_\rho - f_{\mathbf{z},\lambda}\right\|_{\mathcal{H}} = \left\|L^{-1}(u_\rho - u_{\mathbf{z},\lambda})\right\|_{\mathcal{H}}.$$

Therefore, error bounds for $u_\rho - u_{\mathbf{z},\lambda}$ in the weak norm (in $\mathcal{H}_{-1}$) yield bounds for $f_\rho - f_{\mathbf{z},\lambda}$. The latter bounds (in the weak norm) are not known from previous studies. In the oversmoothing cases, i.e., when $f_\rho \notin \mathcal{D}(L)$, then such one-to-one correspondence cannot be established, and additional efforts are required.

For 'classical' inverse problems the fundamental features of regularization in Hilbert scales are known. The questions that we address here try to answer whether these features retain in inverse learning.

- Do regularization schemes which are known to provide optimal rates of reconstruction (as the noise level tends to zero) have analogs here with similar results in inverse learning (as the sample size tends to infinity)?
- Are optimal rates of reconstruction obtained when the true solution does not belong to $\mathcal{D}(L)$ (oversmoothing case)?
- Will the use of a smoothness promoting operator L^{-1} delay saturation?

In order to answer these questions we shall discuss rates of convergence for general (spectral) regularization schemes in Hilbert scales, and under quite general noise condition, see Assumption 2. As mentioned before, in order to treat regularization in Hilbert scales we shall link the given operator A to the scale, which is done in Assumption 4. Then we pursue a novel approach. Instead of assuming smoothness of the sought for f_ρ we shall measure the violation of smoothness relative to a fixed benchmark smoothness, as this will be expressed in terms of a distance function, introduced in Definitions 6 and 7, respectively. Later, in Sect. 4 we shall see how smoothness relative to the given Hilbert scale translates to the behavior of the distance function, and hence, which are the resulting convergence rates.

The paper is organized as follows. The basic definitions, assumptions, and notation required in our analysis are presented in Sect. 2. In Sect. 3, we discuss the bounds of the reconstruction error in the direct learning setting and inverse problem setting by means of distance functions. This section comprises of two main results: The first result is devoted to convergence rates in the oversmoothing case, while the second result focuses on the regular case. When specifying smoothness in terms of source conditions, and this program is performed in Sect. 4, then we can bound the distance functions, and this in turn yields convergence rates in terms of the sample size m. In case that both, the smoothness as well as the link condition are of power type we establish the optimality of the obtained error bounds in the regular case in Sect. 5. Proofs will be given in the appendices. Also, we recall and prove probabilistic estimates which provide the tools for obtaining the error bounds.

2 Notation and assumptions

In this section, we introduce some basic concepts, definitions, notation, and assumptions required in our analysis.

We assume that X is a Polish space, therefore the probability distribution ρ allows for disintegration as

$$\rho(x, y) = \rho(y|x)\nu(x),$$

where $\rho(y|x)$ is the conditional probability distribution of y given x, and $\nu(x)$ is the marginal probability distribution. We consider random observations $\{(x_i, y_i)\}_{i=1}^m$ which follow the model $y = A(f)(x) + \varepsilon$ with centered noise ε. We assume throughout the paper that the operator A is injective.

Assumption 1 *(The true solution)* The conditional expectation w.r.t. ρ of y given x exists (a.s.), and there exists $f_\rho \in \mathcal{H}$ such that

$$\int_Y y d\rho(y|x) = g_\rho(x) = A(f_\rho)(x), \text{ for all } x \in X.$$

The element f_ρ is the true solution which we aim at estimating.

Assumption 2 *(Noise condition)* There exist some constants M, Σ such that for almost all $x \in X$,

$$\int_Y \left(e^{\left\|y - A(f_\rho)(x)\right\|_Y / M} - \frac{\left\|y - A(f_\rho)(x)\right\|_Y}{M} - 1 \right) d\rho(y|x) \leq \frac{\Sigma^2}{2M^2}.$$

This is usually referred to as a *Bernstein-type assumption*.

We recall the unbounded operator $L : \mathcal{D}(L) \subset \mathcal{H} \to \mathcal{H}$, which is assumed to be unbounded and self-adjoint. By spectral theory, the operator $L^s : \mathcal{D}(L^s) \to \mathcal{H}$ is well-defined for $s \in \mathbb{R}$, and the spaces $\mathcal{H}_s := \mathcal{D}(L^s), s \geq 0$ equipped with the inner product $\langle f, g \rangle_{\mathcal{H}_s} = \langle L^s f, L^s g \rangle_{\mathcal{H}}$, $f, g \in \mathcal{H}_s$ are Hilbert spaces. For $s < 0$, the space $\mathcal{H}_s$ is defined as completion of $\mathcal{H}$ under the norm $\|f\|_s := \langle f, f \rangle_s^{1/2}$. The collection $\left\{\mathcal{H}_s,\ s \in \mathbb{R}\right\}$ of Hilbert spaces is called the Hilbert scale induced by L. The following interpolation inequality is an important tool for the analysis:

$$\|f\|_{\mathcal{H}_r} \leq \|f\|_{\mathcal{H}_t}^{\frac{s-r}{s-t}} \|f\|_{\mathcal{H}_s}^{\frac{r-t}{s-t}}, \qquad f \in \mathcal{H}_s, \tag{7}$$

which holds for any $t < r < s$, see e.g. (Engl et al., 1996, Chapt. 8).

2.1 Reproducing Kernel Hilbert spaces and related operators

We start with the concept of reproducing kernel Hilbert spaces, see the seminal study (Aronszajn, 1950), which can be characterized by a symmetric, positive semidefinite kernel and each of its functions satisfies the reproducing property. We consider vector-valued reproducing kernel Hilbert spaces, following (Micchelli & Pontil, 2005), which are the generalization of real-valued reproducing kernel Hilbert spaces.

Definition 1 *(Vector-valued reproducing kernel Hilbert space)* For a non-empty set X and a real separable Hilbert space $(Y, \langle \cdot, \cdot \rangle_Y)$, a Hilbert space $\mathcal{H}$ of functions from X to Y is said to be the vector-valued reproducing kernel Hilbert space, if linear functional $F_{x,y} : \mathcal{H} \to \mathbb{R}$, defined by

$$F_{x,y}(f) = \langle y, f(x) \rangle_Y \qquad \forall f \in \mathcal{H},$$

is continuous for every $x \in X$ and $y \in Y$.

Definition 2 *(Operator-valued positive semi-definite kernel)* Suppose $\mathcal{L}(Y) : Y \to Y$ is the Banach space of bounded linear operators. A function $K : X \times X \to \mathcal{L}(Y)$ is said to be an operator-valued positive semi-definite kernel if

(i) $K(x,x')^* = K(x',x) \qquad \forall\, x,x' \in X.$

(ii) $\sum_{i,j=1}^{N} \langle y_i, K(x_i,x_j)y_j\rangle_Y \geq 0 \qquad \forall\, \{x_i\}_{i=1}^N \subset X \text{ and } \{y_i\}_{i=1}^N \subset Y.$

For a given operator-valued positive semi-definite kernel $K : X \times X \to \mathcal{L}(Y)$, we can construct a unique vector-valued reproducing kernel Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of functions from X to Y as follows:

(i) We define the linear function

$$K_x : Y \to \mathcal{H} : y \mapsto K_x y,$$

where $K_x y : X \to Y : x' \mapsto (K_x y)(x') = K(x',x)y$ for $x, x' \in X$ and $y \in Y$.

(ii) The span of the set $\{K_x y : x \in X, y \in Y\}$ is dense in $\mathcal{H}$.

(iii) **Reproducing property**

$$\langle f(x), y\rangle_Y = \langle f, K_x y\rangle_{\mathcal{H}}, \qquad x \in X,\ y \in Y,\ \forall f \in \mathcal{H},$$

in other words, $f(x) = K_x^* f$.

Moreover, there is a one-to-one correspondence between operator-valued positive semi-definite kernels and vector-valued reproducing kernel Hilbert spaces, see (Micchelli & Pontil, 2005).

Assumption 3 The space $\mathcal{H}'$ is assumed to be a vector-valued reproducing kernel Hilbert space of functions $f : X \to Y$ corresponding to the kernel $K : X \times X \to \mathcal{L}(Y)$ such that

(i) $K_x : Y \to \mathcal{H}'$ is a Hilbert–Schmidt operator for $x \in X$ with

$$\kappa'^2 := \sup_{x\in X} \|K_x\|_{HS}^2 = \sup_{x\in X} \operatorname{tr}(K_x^* K_x) < \infty.$$

(ii) For $y, y' \in Y$, the real-valued function $\varsigma : X \times X \to \mathbb{R} : (x,x') \mapsto \langle K_x y, K_{x'} y'\rangle_{\mathcal{H}'}$ is measurable.

Example In case that the set Y is a bounded subset of $\mathbb{R}$ then the reproducing kernel Hilbert space becomes real-valued reproducing kernel Hilbert space. The corresponding kernel will then be symmetric, positive semi-definite $K : X \times X \to \mathbb{R}$ with the reproducing property $f(x) = \langle f, K_x\rangle_{\mathcal{H}}$. In this case Assumption 3 simplifies to the condition that the kernel is measurable and $\kappa'^2 := \sup_{x\in X} \|K_x\|_{\mathcal{H}'}^2 = \sup_{x\in X} K(x,x) < \infty$.

Now we introduce some relevant operators used in the convergence analysis. We introduce the notation for the vectors $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_m)$, $\mathbf{z} = (z_1, \ldots, z_m)$. The product Hilbert space Y^m is equipped with the inner product $\langle \mathbf{y}, \mathbf{y}'\rangle_m = \frac{1}{m}\sum_{i=1}^m \langle y_i, y_i'\rangle_Y$, and the corresponding norm $\|\mathbf{y}\|_m^2 = \frac{1}{m}\sum_{i=1}^m \|y_i\|_Y^2$. We define the *sampling operator* $S_{\mathbf{x}} : \mathcal{H}' \to Y^m : g \mapsto (g(x_1), \ldots, g(x_m))$, then the adjoint $S_{\mathbf{x}}^* : Y^m \to \mathcal{H}'$ is given by

$$S_{\mathbf{x}}^* \mathbf{y} = \frac{1}{m}\sum_{i=1}^m K_{x_i} y_i.$$

We observe that under Assumption 3 we have

$$\|f\|^2_{\mathscr{L}^2(X,\nu;Y)} = \int_X \|f(x)\|^2_Y d\nu(x) = \int_X \left\|K_x^* f\right\|^2_Y d\nu(x) \leq \kappa'^2 \|f\|^2_{\mathcal{H}'},$$

and

$$\left\|S_{\mathbf{x}} f\right\|^2_m = \frac{1}{m}\sum_{i=1}^m \|f(x_i)\|^2_Y = \frac{1}{m}\sum_{i=1}^m \left\|K_{x_i}^* f\right\|^2_Y \leq \kappa'^2 \|f\|^2_{\mathcal{H}'}.$$

In particular, the canonical injection map $I_\nu : \mathcal{H}' \to \mathscr{L}^2(X,\nu;Y)$ is norm bounded by κ', and so is the empirical version $S_{\mathbf{x}}$.

We denote the population operators $B_\nu := I_\nu A L^{-1} : \mathcal{H} \to \mathscr{L}^2(X,\nu;Y)$, $T_\nu := B_\nu^* B_\nu : \mathcal{H} \to \mathcal{H}$, $L_\nu := A^* I_\nu^* I_\nu A : \mathcal{H} \to \mathcal{H}$, and their empirical versions $B_{\mathbf{x}} = S_{\mathbf{x}} A L^{-1} : \mathcal{H} \to Y^m, T_{\mathbf{x}} = B_{\mathbf{x}}^* B_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}, L_{\mathbf{x}} = A^* S_{\mathbf{x}}^* S_{\mathbf{x}} A : \mathcal{H} \to \mathcal{H}$. The operators T_ν, $T_{\mathbf{x}}$, L_ν, $L_{\mathbf{x}}$ are positive, self-adjoint and depend on the kernel. Under Assumption 3, the operators $B_{\mathbf{x}}$, B_ν are bounded by $\kappa := \kappa' \left\|AL^{-1}\right\|_{\mathcal{H}\to\mathcal{H}'}$ and the operators $L_{\mathbf{x}}$, L_ν are bounded by $\tilde{\kappa}^2$ for $\tilde{\kappa} := \kappa' \|A\|_{\mathcal{H}\to\mathcal{H}'}$, i.e., $\left\|B_{\mathbf{x}}\right\|_{\mathcal{H}\to Y^m} \leq \kappa$, $\left\|B_\nu\right\|_{\mathcal{H}\to\mathscr{L}^2(X,\nu;Y)} \leq \kappa$, $\left\|L_{\mathbf{x}}\right\|_{\mathcal{L}(\mathcal{H})} \leq \kappa^2$ and $\left\|L_\nu\right\|_{\mathcal{L}(\mathcal{H})} \leq \tilde{\kappa}^2$.

2.2 Link condition

The subsequent analysis will frequently use the notion of an index function.

Definition 3 *(Index function)* A function $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is said to be an index function if it is continuous and strictly increasing with $\varphi(0) = 0$.

An index function is called sub-linear whenever the mapping $t \mapsto t/\varphi(t)$, $t > 0$, is nondecreasing. Further, we require this index function to belong to the following class of functions.

$$\begin{aligned}\mathcal{F} = \{\varphi = \varphi_1\varphi_2 : \varphi_1, \varphi_2 : [0,\kappa^2] \to [0,\infty), \varphi_1 \text{ nondecreasing continuous sub-linear,} \\ \varphi_2 \text{ nondecreasing Lipschitz, } \varphi_1(0) = \varphi_2(0) = 0\}.\end{aligned} \tag{8}$$

The representation $\varphi = \varphi_2\varphi_1$ is not unique, therefore φ_2 can be assumed as a Lipschitz function with Lipschitz constant 1. We shall also rely upon the following important result for such Lipschitz continuous index functions φ_2, needed in our analysis (Peller, 2016, Corollary 1.2.2):

$$\left\|\varphi_2(T_{\mathbf{x}}) - \varphi_2(T_\nu)\right\|_{HS} \leq \left\|T_{\mathbf{x}} - T_\nu\right\|_{HS}.$$

Example Power-type functions $\varphi(t) = t^r$ with $r > 0$, and logarithmic functions $\varphi(t) = t^p \log^{-\nu}\left(\frac{1}{t}\right)$, $p, \nu \geq 0$, are examples of functions in the class $\mathcal{F}$.

The following assumption is used to relate smoothness in terms of the operator L to the covariance operator T_ν.

Assumption 4 *(link condition)* There exist a power $q > 1$ and an index function ϱ, for which the function ϱ^2 is sub-linear. There is a constant $1 \leq \beta < \infty$ such that

$$\|L^{-q}u\|_{\mathcal{H}} \le \left\|\varrho^q(T_\nu)u\right\|_{\mathcal{H}} \le \beta^q \|L^{-q}u\|_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

The function $t \mapsto \varphi(t) := \varrho^{q-1}(t)$ belongs to the class $\mathcal{F}$.

Only the left inequality will be used for the regular case. For the oversmoothing case, when we need to relate the effective dimensions we require the other side as well. Also, to show the optimality of the rates both side inequalities are used.

As shown in Böttcher et al. (2006), Assumption 4 implies the range identity $\mathcal{R}(L^{-q}) = \mathcal{R}(\varrho^q(T_\nu))$. In the context of a comparison of operators we mention the well-known Heinz Inequality, see (Engl et al., 1996, Prop. 8.21). This asserts that for every exponent $0 < a \le 1$ it holds true

$$\|Gu\|_{\mathcal{H}} \le \|Hu\|_{\mathcal{H}},\ u \in \mathcal{H} \quad \text{yields} \quad \|G^a u\|_{\mathcal{H}} \le \|H^a u\|_{\mathcal{H}},\ u \in \mathcal{H}. \tag{9}$$

Applying this to the above link condition we obtain the following:

Proposition 1 *Under Assumption 4 we have*

$$\left\|L^{-1}u\right\|_{\mathcal{H}} \le \left\|\varrho(T_\nu)u\right\|_{\mathcal{H}} \le \beta\left\|L^{-1}u\right\|_{\mathcal{H}}, \quad u \in \mathcal{H}$$

and

$$\left\|L^{-(q-1)}u\right\|_{\mathcal{H}} \le \left\|\varrho^{q-1}(T_\nu)u\right\|_{\mathcal{H}} \le \beta^{(q-1)}\left\|L^{-(q-1)}u\right\|_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

Moreover, we have that

$$\left\|\varrho(T_\nu)\left(T_\nu + \lambda I\right)^{-1/2}\right\|_{\mathcal{L}(\mathcal{H})} \le \frac{\varrho(\lambda)}{\sqrt{\lambda}}, \quad 0 < \lambda \le 1. \tag{10}$$

Remark 1 It is heuristically clear that the function ϱ^2 cannot increase faster than linearly, because the operator $T_\nu = L^{-1}L_\nu L^{-1}$ has L^{-2} in it. Therefore, requiring sub-linearity is not a strong restriction. More details will be given in Sect. 5.

Link conditions as in Assumption 4 imply decay rates for the singular numbers of the operators, known as Weyl's Monotonicity Theorem (Bhatia, 1997, Cor. III.2.3). In our case, this yields that $s_j(\varrho(T_\nu)) = \varrho(s_j(T_\nu)) \asymp s_j(L^{-1})$. For classical spaces, as e.g. Sobolev spaces, when $L := (I + \Delta)^{-1/2}$, then $s_j(L^{-1}) \asymp 1/j$ (one spatial dimension). For the above index function ϱ this means that $s_j(T_\nu) \asymp \varrho^{-1}(1/j)$.

Example **(Finitely smoothing covariance operators)** In case that the function ϱ, and hence its inverse are of power type then this implies a power type decay of the singular numbers of T_ν. In this case, the operator T_ν is called finitely smoothing.

Example **(Infinitely smoothing covariance operators)** If, on the other hand, the function ϱ is logarithmic, as e.g., $\varrho(t) = \left(\log \frac{1}{t}\right)^{-\frac{1}{\mu}}$, then $s_j(T_\nu) \asymp e^{-j^\mu}$. In this case, the operator T_ν is called infinitely smoothing.

2.3 Effective dimension

The concept of the effective dimension, as introduced in Zhang (2002), proved to be important for deriving fast rates of convergence under Hölder's source condition, see (Blanchard & Mücke, 2018; Caponnetto & De Vito, 2007; Guo et al., 2017), and also for general source conditions, see (Lin et al., 2020; Shuai et al., 2020; Rastogi & Sampath, 2017). For the trace–class operator T_ν its effective dimension is defined as,

$$\mathcal{N}_{T_\nu}(\lambda) := \mathrm{Tr}\big((T_\nu + \lambda I)^{-1} T_\nu\big), \text{ for } \lambda > 0.$$

It is known that the function $\lambda \to \mathcal{N}_{T_\nu}(\lambda)$ is continuous and decreasing from ∞ to zero for $0 < \lambda < \infty$ for an infinite dimensional operator T_ν (see for details Blanchard and Mathé, 2012; Blanchard and Mücke, 2020; Lin et al., 2015; Shuai et al., 2020; Zhang, 2002). However, we shall use, and this follows from spectral calculus, that the function $\lambda \mapsto \lambda \mathcal{N}_{T_\nu}(\lambda)$ is increasing.

We have the trivial bound

$$\mathcal{N}_{T_\nu}(\lambda) \le \left\|(T_\nu + \lambda I)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \mathrm{Tr}\big(T_\nu\big) \le \frac{\kappa^2}{\lambda}.$$

In the subsequent analysis, we shall need a relationship between the effective dimensions $\mathcal{N}_{T_\nu}(\lambda)$ and $\mathcal{N}_{L_\nu}(\lambda)$ of the operators T_ν and L_ν, respectively. For this, the following assumption, introduced in Lin et al. (2015), will be used. There, it was shown that it is satisfied for both moderately ill-posed and severely ill-posed operators.

Assumption 5 There exists a constant C such that for $0 < t \le \|L_\nu\|_{\mathcal{L}(\mathcal{H})}$ we have

$$t^{-1} \sum_{s_j(L_\nu) < t} s_j(L_\nu) < C \#\big\{j, \quad s_j(L_\nu) \ge t\big\}.$$

Proposition 2 *Suppose Assumptions 4 and 5 hold true. Suppose that the function ϱ from the link condition, Assumption 4, is such that the function $t \mapsto \left(\varrho^{2q}\right)^{-1}(t)$ is operator concave, and that there is some $n \in \mathbb{N}$ for which the function $t \mapsto \varrho^{-1}(t)/t^n$ is concave. Then, there is $\widetilde{C}$ for which we have that*

$$\mathcal{N}_{L_\nu}\left(\frac{\lambda}{\varrho^2(\lambda)}\right) \le 2\beta^{n+1}\widetilde{C}\mathcal{N}_{T_\nu}(\lambda), \quad 0 < \lambda \le \|T_\nu\|_{\mathcal{L}(\mathcal{H})}.$$

Remark 2 For a power type function $\varrho(t) := t^a$ the above concavity assumptions hold true whenever $2aq \ge 1$ and $n \le 1/a \le n + 1$. In particular, the number n is uniquely determined.

2.4 Regularization schemes

General regularization schemes were introduced and discussed in ill-posed inverse problems and learning theory (See Shuai & Pereverzev, 2013, Sect. 2.2 and Bauer et al., 2007, Sect. 3.1) for brief discussion). By using the notation from Sect. 2.1, the Tikhonov regularization scheme from (5) can be re-expressed as follows:

$$f_{\mathbf{z},\lambda} = \underset{f \in \mathcal{D}(L)}{\mathrm{argmin}} \left\{ \|S_{\mathbf{x}} A(f) - \mathbf{y}\|_m^2 + \lambda \|Lf\|_{\mathcal{H}}^2 \right\},$$

with minimizer given as

$$f_{\mathbf{z},\lambda} = L^{-1}(T_{\mathbf{x}} + \lambda I)^{-1} B_{\mathbf{x}}^* \mathbf{y}.$$

The following definition extends this by replacing the operator $(T_{\mathbf{x}} + \lambda I)^{-1}$ by some operator function $g_\lambda(T_{\mathbf{x}})$.

Definition 4 *(Spectral regularization)* We say that a family of functions $g_\lambda : [0, \kappa^2] \to \mathbb{R}$, $0 < \lambda \leq a$, is a regularization scheme if there exists D, B, γ such that

- $\sup_{t\in[0,\kappa^2]} \left|tg_\lambda(t)\right| \leq D.$
- $\sup_{t\in[0,\kappa^2]} \left|g_\lambda(t)\right| \leq \frac{B}{\lambda}.$
- $\sup_{t\in[0,\kappa^2]} \left|r_\lambda(t)\right| \leq \gamma \qquad \text{for} \quad r_\lambda(t) = 1 - g_\lambda(t)t.$
- For some constant γ_p (independent of λ), the maximal p satisfying the condition:

$$\sup_{t\in[0,\kappa^2]} \left|r_\lambda(t)\right| t^p \leq \gamma_p \lambda^p$$

is said to be the qualification of the regularization scheme g_λ.

Definition 5 The qualification p covers the index function φ if the function $t \to \frac{t^p}{\varphi(t)}$ is nondecreasing.

We mention the following result.

Proposition 3 *Suppose φ is a nondecreasing index function and the qualification, say $p \geq 1$, of the regularization g_λ covers φ. Then*

$$\sup_{t\in[0,\kappa^2]} \left|r_\lambda(\sigma)\right| \varphi(\sigma) \leq c_p \varphi(\lambda), \quad c_p = \max(\gamma, \gamma_p).$$

Also, we have that

$$\sup_{t\in[0,\kappa^2]} \left|r_\lambda(\sigma)\right| \varphi(\lambda + \sigma) \leq 2^p c_p \varphi(\lambda).$$

Most of the linear (spectral) regularization schemes (Tikhonov regularization, Landweber iteration or spectral cut-off) satisfy the properties of general regularization. Inspired by the representation for the minimizer of the Tikhonov functional (5) we consider a general regularized solution in Hilbert scales corresponding to the above regularization g_λ in the form

$$f_{\mathbf{z},\lambda} = L^{-1} g_\lambda(T_{\mathbf{x}}) B_{\mathbf{x}}^* \mathbf{y}, \tag{11}$$

where by spectral calculus the real-valued function g_λ is applied to the self-adjoint operator $T_{\mathbf{x}}$.

3 Convergence analysis

The analysis will distinguish between two cases, the 'regular' one, when $f_\rho \in \mathcal{D}(L)$, and the 'low smoothness' case, when $f_\rho \notin \mathcal{D}(L)$. In either case, we shall first utilize the concept of *distance functions*. This will later give rise to establish convergence rates in a more classical style.

For the asymptotical analysis, we shall require the standard assumption relating the sample size m and the parameter λ such that

$$\mathcal{N}_{T_\nu}(\lambda) \leq m\lambda \qquad \text{and} \qquad 0 < \lambda \leq 1. \tag{12}$$

It will be seen, that asymptotically the condition (12) is always satisfied for the parameter which is optimally chosen under known smoothness.

Since the function $\mathcal{N}_{T_\nu}(\lambda)$ is decreasing in λ, for $\lambda \leq 1$ we have that $\mathcal{N}_{T_\nu}(1) \leq \mathcal{N}_{T_\nu}(\lambda)$. Hence condition (12) yields that

$$\mathcal{N}_{T_\nu}(1) \leq m\lambda. \tag{13}$$

Several probabilistic quantities will be used to express the error bounds. Precisely, for an index function ζ we let

$$\Xi^\zeta = \Xi^\zeta(\lambda) := \left\| \left(\frac{1}{\zeta}\right)(T_{\mathbf{x}} + \lambda I)\zeta(T_\nu + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})}, \tag{14}$$

$$\Lambda = \Lambda(\lambda) := \left\| (L_\nu + \lambda I)^{-1/2}(L_\nu - L_{\mathbf{x}}) \right\|_{HS}, \tag{15}$$

$$\Upsilon = \Upsilon(\lambda) := \left\| (T_\nu + \lambda I)^{-1/2}(T_\nu - T_{\mathbf{x}}) \right\|_{HS}, \tag{16}$$

and

$$\Psi = \Psi(\lambda) := \left\| (T_\nu + \lambda I)^{-1/2} B_{\mathbf{x}}^*(S_{\mathbf{x}} A f_\rho - \mathbf{y}) \right\|_{\mathcal{H}}. \tag{17}$$

In case that $\zeta(t) = t^r$ we abbreviate Ξ^{t^r} by Ξ^r and Ξ^t by Ξ, not to be confused with the power. High probability bounds for these quantities are known, and these are given correspondingly in Propositions 4 and 5 in Appendix C.

3.1 The oversmoothing case

As mentioned before, we shall use distance functions, which measure the violation of a benchmark smoothness. Here the benchmark will be $f_\rho \in \mathcal{D}(L)$.

Definition 6 We define the distance function $d : [0, \infty) \to [0, \infty)$ by

$$d(R) = \inf \left\{ \left\| f_\rho - f \right\|_{\mathcal{H}} : f = L^{-1}v \text{ and } \|v\|_{\mathcal{H}} \leq R \right\}, \quad R > 0. \tag{18}$$

The distance function is positive, decreasing, convex and continuous for all $0 \leq R < 1$. It tends to 0 as $R \to \infty$, see (Hofmann, 2006). Hence, the unique minimizer exists and will be denoted by f_ρ^R.

Notice the following: If $f_\rho \in \mathcal{D}(L)$ then for some R the minimizer f_ρ^R of the distance function will obey $f_\rho^R = f_\rho$.

Remark 3 In a rudimentary form, this approach was given in (Baumeister, 1987, Thm. 6.8). It was then introduced in regularization theory in Hofmann (2006). Within learning theory, such a concept was also used in the study (Smale & Zhou, 2003).

Theorem 1 *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* ρ. *Suppose the Assumptions* 1–5 *hold true. Let* g_λ *be a regularization with corresponding regularized solution* $f_{\mathbf{z},\lambda}$ (*see* (11)). *Suppose that the qualification* p *of the regularization* g_λ *covers the function* ϱ (*from Assumption* 4) *and that the functions* $\varrho^{-1}(t)/t^n$, *and* $\left(\varrho^{2q}\right)^{-1}(t)$ *are concave, or operator concave, for some* $n \geq 1$, *respectively. Then for all* $0 < \eta < 1$, *and for* λ *satisfying the condition* (12) *the following upper bound holds true with confidence* $1 - \eta$:

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq C\{d(R) + 2R\varrho(\lambda)\} \log^4\left(\frac{4}{\eta}\right), \quad R \geq \Sigma + \kappa M/\mathcal{N}_{T_v}(1),$$

where C *depends on* $B, D, c_p, \kappa, n, \beta, \widetilde{C}$.

The bound from Theorem 1 is valid for all $R \geq \Sigma + \kappa M/\mathcal{N}_{T_v}(1)$, and we shall now optimize the bound from Theorem 1 with respect to the choice of $R \geq \Sigma + \kappa M/\mathcal{N}_{T_v}(1)$.

First, if $f_\rho \in \mathcal{D}(L)$ then there is $\bar{R} \geq \Sigma + \kappa M/\mathcal{N}_{T_v}(1)$ such that $d(\bar{R}) = 0$, and

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq C\bar{R}\,\varrho(\lambda) \log^4\left(\frac{4}{\eta}\right),$$

where C depends on $B, D, c_p, \kappa, n, \beta, \widetilde{C}$.

Otherwise, in the low smoothness case, $f_\rho \notin \mathcal{D}(L)$, we introduce the following function

$$\Gamma(R) := \frac{d(R)}{R}, \qquad R \geq \Sigma + \kappa M/\mathcal{N}_{T_v}(1),$$

which is non-vanishing decreasing function, and hence the inverse Γ^{-1} exists, and it is decreasing. Given $\lambda > 0$, by letting $R = R(\lambda)$ solve the equation $\Gamma(R) = \varrho(\lambda)$ we find that

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq CR(\lambda)\varrho(\lambda) \log^4\left(\frac{4}{\eta}\right), \tag{19}$$

where C depends on $B, D, c_p, \kappa, n, \beta, \widetilde{C}$.

The above dependency $\lambda \to R(\lambda)$ can be made explicit when assuming that f_ρ has some smoothness measured in terms of a source condition, see Sect. 4, below. For Theorem 1 we get the error bound (19) but the parameter λ has to obey (12).

3.2 The regular case

Here we analyze the rates of convergence in the case when the underlying true solution f_ρ belongs to the domain of the operator L. Again, we shall choose a benchmark smoothness,

here in the form of $f_\rho \in \mathcal{D}(L^q)$ for some $q \geq 1$. This benchmark smoothness is determined by the user. With respect to this benchmark we introduce the following distance function.

Definition 7 Given $q \geq 1$ we define the distance function $d_q : [0,\infty) \to [0,\infty)$ by

$$d_q(R) = \inf\left\{ \left\| L(f - f_\rho) \right\|_{\mathcal{H}} : f = L^{-q} v \text{ and } \|v\|_{\mathcal{H}} \leq R \right\}. \tag{20}$$

Theorem 2 *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* ρ*. Suppose the Assumptions 1–4 hold true. Let* g_λ *be a regularization with corresponding regularized solution* $f_{\mathbf{z},\lambda}$ *(see (11)). Let* ζ *be any index function, such that* $\frac{1}{2}$ *covers* ζ*. Suppose that the qualification* p *of the regularization* g_λ *covers the function* $\zeta\varphi$ *(with* φ *from Assumption 4). Then for all* $0 < \eta < 1$*, and for* λ *satisfying the condition (12), the following upper bound holds true with confidence* $1 - \eta$*:*

$$\left\| \zeta(T_v) L\left(f_{\mathbf{z},\lambda} - f_\rho\right) \right\|_{\mathcal{H}} \leq C\zeta(\lambda) \left\{ d_q(R) + R\left(\varphi(\lambda) + \frac{1}{\sqrt{m}} \right) + C' \sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{m\lambda}} \right\} \times \log^4\left(\frac{4}{\eta}\right),$$

Consequently, we find that

$$\left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}} \leq C\varrho(\lambda) \left\{ d_q(R) + R\left(\varphi(\lambda) + \frac{1}{\sqrt{m}} \right) + C' \sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{m\lambda}} \right\} \log^4\left(\frac{4}{\eta}\right)$$

and

$$\left\| I_v A(f_{\mathbf{z},\lambda} - f_\rho) \right\|_{\mathcal{L}^2(X,\nu;Y)} \leq C\sqrt{\lambda} \left\{ d_q(R) + R\left(\varphi(\lambda) + \frac{1}{\sqrt{m}} \right) + C' \sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{m\lambda}} \right\} \times \log^4\left(\frac{4}{\eta}\right),$$

where C *depends on* B, D, c_p, κ*, and* $C' = 2\kappa M + \Sigma$*.*

The bound from Theorem 2 is valid for all $R \geq 1$, and we shall now optimize the bound from Theorem 2 with respect to the choice of $R \geq 1$.

First, if $f_\rho \in \mathcal{R}(L^{-q})$ then $d_q(\bar{R}) = 0$ for some $\bar{R}$, we find that

$$\left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}} \leq C\varrho(\lambda) \left\{ \bar{R}\left(\varphi(\lambda) + \frac{1}{\sqrt{m}} \right) + C' \sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{m\lambda}} \right\} \log^4\left(\frac{4}{\eta}\right).$$

Otherwise, in case that $f_\rho \notin \mathcal{R}(L^{-q})$ we introduce the following function

$$\Gamma_q(R) := \frac{d_q(R)}{R}, \qquad R \geq 1, \tag{21}$$

which is non-vanishing decreasing function, and hence the inverse Γ_q^{-1} exists and it is decreasing. We finally get the main result, by letting $R = R(\lambda)$ solving the equation $\Gamma_q(R) = \varphi(\lambda)$, and we find that

$$\left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}} \leq C\varrho(\lambda)\left\{ R(\lambda)\left(\varphi(\lambda) + \frac{1}{\sqrt{m}} \right) + C'\sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{m\lambda}} \right\} \log^4\left(\frac{4}{\eta}\right).$$

4 Smoothness in terms of source-wise representation

So far convergence results were established in terms of distance functions. We will now specify the smoothness of the true solution in terms of the bounded, injective and self-adjoint operator L^{-1}. This is genuine for regularization in Hilbert scales.

Assumption 6 *(General source condition)* For an index function θ, the true solution f_ρ belongs to the class $\Omega(\theta, R^\dagger)$ with

$$\Omega(\theta, R^\dagger) := \left\{ f \in \mathcal{H} : f = \theta(L^{-1})v \text{ and } \|v\|_{\mathcal{H}} \leq R^\dagger \right\}.$$

Notice that elements from $\Omega(\theta, R^\dagger)$ belong to the range of $\theta(L^{-1})$ which coincides with the domain of $\theta(L)$, since L^{-1} was assumed to be bounded.

We aim at bounding the distance functions $d(R)$ and $d_q(R)$ from the oversmoothing and regular cases, respectively.

For a better understanding, we shall highlight the obtained general bounds when the considered index functions are of power type, and we specify the function $\theta(t) := t^r$, which represents the smoothness, as well as $\varrho(t) = t^a$, representing the link, for this purpose. It will be seen that the index function $t \mapsto \theta(\varrho(t))$, $t > 0$ is relevant in the subsequent corollaries, which here reads as $\theta(\varrho(t)) = t^{ar}$, $t > 0$. Also, in the regular case with benchmark smoothness $f_\rho \in \mathcal{R}(L^{-q})$, the function $t \mapsto \frac{t^q}{\theta}(t)$ appears, and this is required to be an index function. Within the power type context, this reads as $r < q$, and it simply means that the actual smoothness is not beyond the benchmark.

Finally, we emphasize that the rates will depend on the decay of the effective dimension of the covariance operator T_v, which was introduced in Sect. 2.3. Therefore, we will highlight the obtained bounds under specified decay rates for the effective dimension in Sect. 4.3. The obtained overall rates will be highlighted in Tables 1 and 2, respectively.

4.1 The oversmoothing case

Here the benchmark source condition $f_\rho \in \mathcal{R}(L^{-1})$ $(q = 1)$ is linear, represented by the identity function $\iota : t \mapsto t$, and we shall thus assume that the index function θ is sub-linear. The obtained bounds will rely on the results from (Hofmann & Mathé, 2007, Theorem 5.9). Under Assumption 6 we find that

Table 1 (under Assumption 7, and for $a \leq \frac{1}{2}$, $aq \leq p$): convergence rates of the regularized solution $f_{\mathbf{z},\lambda}$

Case	Convergence rates	Parameter $\lambda^* = \mathcal{O}(\cdot)$	True Smooth.	Conditions
Oversmooth.	$\left(\frac{1}{\sqrt{m}}\right)^{\frac{2ar}{b+1}}$	$\left(\frac{1}{\sqrt{m}}\right)^{\frac{2}{b+1}}$	$r \leq 1$	$a \geq \frac{1}{n+1}$
Regular	$\left(\frac{1}{\sqrt{m}}\right)^{\frac{r}{q-1}}$	$\left(\frac{1}{\sqrt{m}}\right)^{\frac{1}{a(q-1)}}$	$r \geq 1$	$q \geq r + \frac{b+1}{2a}$
	$\left(\frac{1}{\sqrt{m}}\right)^{\frac{2ar}{2ar+b+1-2a}}$	$\left(\frac{1}{\sqrt{m}}\right)^{\frac{2}{2ar+b+1-2a}}$		$r \leq q \leq r + \frac{b+1}{2a}$

Table 2 (under Assumption 8, and for $a \leq \frac{1}{2}$, $aq \leq p$): convergence rates of the regularized solution $f_{\mathbf{z},\lambda}$

Case	Convergence rates	Parameter $\lambda^* = \mathcal{O}(\cdot)$	True Smooth.	Conditions
Oversmooth.	$\left(\frac{\log m}{m}\right)^{\frac{ar}{b+1}}$	$\left(\frac{\log m}{m}\right)^{\frac{1}{b+1}}$	$r \leq 1$	$a \geq \frac{1}{n+1}$
Regular	$\left(\frac{\log m}{m}\right)^{\frac{r}{2(q-1)}}$	$\left(\frac{\log m}{m}\right)^{\frac{1}{2a(q-1)}}$	$r \geq 1$	$q \geq r + \frac{b+1}{2a}$
	$\left(\frac{\log m}{m}\right)^{\frac{ar}{2ar+b+1-2a}}$	$\left(\frac{\log m}{m}\right)^{\frac{1}{2ar+b+1-2a}}$		$r \leq q \leq r + \frac{b+1}{2a}$

$$d(R) \leq R\left(\left(\frac{\iota}{\theta}\right)^{-1}\left(\frac{R^\dagger}{R}\right)\right), \quad R > 0.$$

In order to minimize the bound from Theorem 1, we balance $d(R) = R\varrho(\lambda)$, resulting in

$$R(\lambda) = R^\dagger \frac{\theta(\varrho(\lambda))}{\varrho(\lambda)}, \quad \lambda > 0. \tag{22}$$

Thus, for this value of $R(\lambda)$ under the condition (12), the bound (19) reduces to

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq CR(\lambda)\varrho(\lambda)\log^4(4/\eta) \leq CR^\dagger\theta(\varrho(\lambda))\log^4(4/\eta). \tag{23}$$

The following corollary is the consequence of Theorem 1 which explicitly provide us with an error bound in terms of the sample size m.

Corollary 1 *Suppose that the unknown f_ρ obeys Assumption 6 for a sub-linear function θ. Under the same assumptions of Theorem 1 and with the a-priori choice of the regularization parameter $\lambda^* = \lambda^*(m)$ from solving the equation $\mathcal{N}_{T_v}(\lambda^*) = m\lambda^*$, for all $0 < \eta < 1$, the following error estimates holds with confidence $1 - \eta$:*

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq C\theta(\varrho(\lambda^*))\log^4\left(\frac{4}{\eta}\right),$$

where C depends on B, D, c_p, κ, n, β, $\widetilde{C}$, M, Σ, and $R^\dagger$.

Evidently, the above parameter choice satisfies condition (12).

4.2 The regular case

In this case the benchmark is given by the index function $\imath^q$, and we shall assume that the given smoothness, measured in terms of θ, is such that the function $\imath^q/\theta$ for $0 < t \leq \kappa^2$, is an index function. However, the definition of the distance function $R \mapsto d_q(R)$ is non-standard. The target norm is $\left\|L(f - f_\rho)\right\|_{\mathcal{H}}$, and, in order to apply the result from (Hofmann & Mathé, 2007, Theorem 5.9) we have to 'rescale' the given smoothness (in terms of the operator L^{-1}) by factor L^{-1}. If Assumption 6 holds true with index function θ, for which the quotient $\imath^q/\theta$ is an index function, and so will be the function $\imath^{q-1}/(\theta/\imath)$, then this results in the bound

$$d_q(R) \leq R \left[\left(\frac{\imath^q}{\theta} \right)^{-1} \left(\frac{R^\dagger}{R} \right) \right]^{q-1}, \quad R > 0. \tag{24}$$

According to Theorem 2 we balance

$$d_q(R) = R\varphi(\lambda).$$

This yields

$$R(\lambda) = R^\dagger \frac{\theta(\varrho(\lambda))}{\varrho^q(\lambda)}, \quad R > 0.$$

Inserting this bound into Theorem 2 we find that

$$\begin{aligned} &\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \\ &\leq C\varrho(\lambda)\left\{ R^\dagger \frac{\theta(\varrho(\lambda))}{\varrho(\lambda)} \left(1 + \frac{1}{\sqrt{m}\varphi(\lambda)} \right) + C' \sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{m\lambda}} \right\} \log^4\left(\frac{4}{\eta}\right) \\ &= C\varrho(\lambda)\left\{ R^\dagger \frac{\theta(\varrho(\lambda))}{\varrho(\lambda)} + \frac{1}{\sqrt{m}} \left(R^\dagger \frac{\theta(\varrho(\lambda))}{\varrho^q(\lambda)} + C' \sqrt{\frac{\mathcal{N}_{T_v}(\lambda)}{\lambda}} \right) \right\} \log^4\left(\frac{4}{\eta}\right) \end{aligned} \tag{25}$$

provided that (12) holds.

The optimization of the bound in the inequality (25) depends on which term is dominant in the last two summands. Then we can balance the remaining (two) terms. This results in the following corollaries for the different choices of the regularization parameter:

Corollary 2 *Suppose that the unknown f_ρ obeys Assumption 6 for an index function θ, and that the related functions $\frac{\imath^q}{\theta}(t)$ and $\frac{\imath^q}{\theta}(\varrho(t))\sqrt{\frac{\mathcal{N}_{T_v}(t)}{t}}$ are index functions. Under the same assumptions of Theorem 2, and for the a-priori choice of the regularization parameter $\lambda^* = \varphi^{-1}\left(\frac{1}{\sqrt{m}}\right)$, for all $0 < \eta < 1$, the following upper bound holds with confidence $1 - \eta$:*

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq C\theta(\varrho(\lambda^*)) \log^4\left(\frac{4}{\eta}\right),$$

where C depends on B, D, c_p, κ, M, Σ, and $R^\dagger$.

Corollary 3 *Suppose that the unknown f_ρ obeys Assumption 6 for an index function θ, and that the related functions $\frac{t^q}{\theta}(t)$ and $\frac{t^q}{\theta}(\varrho(t))\sqrt{\frac{\mathcal{N}_{T_v}(t)}{t}}$ are index functions. Under the same assumptions of Theorem 2, and for the a-priori choice of the regularization parameter λ^* as solution to the equation $\frac{\Theta^2(\varrho(\lambda^*))}{\varrho^2(\lambda^*)}\lambda^* m = \mathcal{N}_{T_v}(\lambda^*)$, for all $0 < \eta < 1$, the following upper bound holds with confidence $1-\eta$:*

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\mathcal{H}} \leq C\theta(\varrho(\lambda^*))\log^4\left(\frac{4}{\eta}\right),$$

where C depends on B, D, c_p, κ, M, Σ, and $R^\dagger$.

Since by assumption the function $t \mapsto \frac{\Theta^2(\varrho(\lambda^*))}{\varrho^2(\lambda^*)}$ is an index function we will have that condition (12) holds for m large enough.

4.3 Taking the behavior of effective dimension into account

Below, to be specific, we consider the following two behaviors of the decay of the effective dimensions of the covariance operator T_v, say power-type and logarithmic type, known to hold true in many situations.

Assumption 7 *(Polynomial decay)* There exists some positive constant $c > 0$ such that

$$\mathcal{N}_{T_v}(\lambda) \leq c\lambda^{-b}, \quad \text{for } 0 \leq b < 1,\ \forall \lambda > 0.$$

Assumption 8 *(Logarithmic decay)* There exists some positive constant $c > 0$ such that

$$\mathcal{N}_{T_v}(\lambda) \leq c\log\left(\frac{1}{\lambda}\right), \quad \forall \lambda > 0.$$

Remark 4 We mention that a polynomial decay of the eigenvalues of the covariance operator T_v yields a power-type behavior of the effective dimension, see (Caponnetto & De Vito, 2007). In some situations this behavior is not evident. Shuai et al. (2020) showed that for Gaussian kernel $K_1(x,x') = xx' + e^{-8(x-x')^2}$ with the uniform sampling on [0, 1], the effective dimension exhibits a log-type behavior (Assumption 8), on the other hand, the kernel $K_2(x,x') = \min\{x,x'\} - xt$ exhibits a power-type behavior (Assumption 7).

Below, we shall summarize the convergence rates under the specific behavior of the effective dimension, Assumptions 7 and 8, respectively, in the Tables 1 and 2. We confine to the power type case, when both the link condition as well as the source condition are of power type, i.e., $\varrho(t) = t^a$ and $\theta(t) = t^r$ for parameters $a, r > 0$. The qualification of the regularization is denoted by p as before. Also, the benchmark smoothness is q, where either $q = 1$ (oversmoothing case) or $q > 1$ (regular case). Notice, that due to the sub-linearity condition for ϱ^2 we must have that $0 < a \leq 1/2$. Also, throughout the analysis, we assume that the qualification covers the given smoothness, i.e., $aq \leq p$. The bounds presented in the tables are consequences of Corollaries 1–3, respectively. Therefore, Assumptions 1–6 are assumed to be satisfied.

The tables are structured as follows. In the first column we present the rates of convergence $\varepsilon(m)$ for the error estimates of the form:

$$\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\left\|f_{\mathbf{z},\lambda}-f_\rho\right\|_{\mathcal{H}} \le C\varepsilon(m)\log^4\left(\frac{4}{\eta}\right)\right\} \ge 1-\eta.$$

In the second column, the corresponding order of the regularization parameter choice λ^* in terms of m is indicated. In the third column, we highlight the smoothness of the true solution f_ρ. In fourth column, we emphasize additional constraints, specifically on the benchmark smoothness.

The first row corresponds to the oversmoothing case, and the last two rows correspond to the regular case. In the regular case, we observe that the validity of the rates of the convergence depends on the benchmark smoothness through aq. At the intersection point, when $aq = ar + \frac{b+1}{2}$, both rates coincide. As we will see in the next section the rates of convergence in the regular case ($q > 1$) are optimal provided that the benchmark smoothness is chosen appropriately.

5 Optimality of the error bounds

We shall discuss the optimality of the previously obtained error bounds, in the regular case, and we shall use the known optimality results from (Blanchard & Mücke, 2018). However, at present the smoothness is measured with respect to the operator T_ν, whereas in Blanchard and Mücke (2018) this was done with respect to the operator $L_\nu := A^* I_\nu^* I_\nu A = L T_\nu L$. Therefore, the following 'recipe' will be used.

1. Transfer smoothness as given in terms of L^{-1} to smoothness in terms of L_ν, and
2. Knowing the decay of the singular numbers of the operator T_ν inherent in Assumption 7, find the decay of the singular numbers of L_ν.

In order to keep the analysis simple and transparent, we confine to power type smoothness $\theta(t) = t^r$, $0 < r \le q$ in Assumption 6, as well as to power type link in Assumption 4 with $\varrho(t) := t^a$ for some $a > 0$.

In the subsequent subsections, we shall sketch the proof of the lower bounds step by step, reaching the optimality assertion at the end. In order to get there, additional assumptions have to be made, a lifting condition (Assumption 9), and a singular number decay condition (Assumption 10).

5.1 Relating smoothness

The link condition is crucial, and the subsequent arguments are of interpolation type, applying Heinz Inequality within the present context. To this end, we require that q is chosen such that $aq \ge 1/2$. In this case Assumption 4 yields, by applying Heinz Inequality (9) with exponent $1/(2aq) \le 1$ that

$$\left\|I_\nu A L^{-1} u\right\|_{\mathscr{L}^2(X,\nu;Y)} = \left\|T_\nu^{1/2} u\right\|_{\mathcal{H}} \asymp^1 \left\|L^{-\frac{1}{2a}} u\right\|_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

[1] Letting $v := L^{-1}u$ we find that

$$\left\|L_\nu^{1/2}v\right\|_{\mathcal{H}} = \left\|I_\nu Av\right\|_{\mathscr{L}^2(X,\nu;Y)} \asymp \left\|L^{-(\frac{1}{2a}-1)}v\right\|_{\mathcal{H}}, \quad v \in \mathcal{H}. \tag{26}$$

First, we see from this that $a < 1/2$, because otherwise L_ν would be continuously invertible. Also, the relation (26) would allow transferring smoothness r with respect to L^{-1} to L_ν as long as $0 < r \leq \frac{1}{2a} - 1$. In order to treat higher smoothness (in terms of L^{-1}) a lifting condition is unavoidable. This must be consistent with the link from (26). Thus we look for a factor z such that $t^{(\frac{1}{2a}-1)z} = t^q$, yielding $z := \frac{2aq}{1-2a}$.

Assumption 9 *(lifting condition)* We have that

$$\|L^{-q}u\|_{\mathcal{H}} \asymp \left\|L_\nu^{\frac{aq}{1-2a}}u\right\|_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

Remark 5 The strengthening of the original link condition, Assumption 4, towards a lifting condition has been discussed in more detail in Mathé (2019).

Having this lifting, and applying Heinz Inequality (9) (with exponent r/q) yields

$$\|L^{-r}v\|_{\mathcal{H}} \asymp \left\|L_\nu^{\frac{ar}{1-2a}}v\right\|_{\mathcal{H}}, \quad v \in \mathcal{H}, \tag{27}$$

and a source-wise representation as in Assumption 6 yields a corresponding source-wise representation with respect to the operator L_ν (with different constant).

5.2 Relating effective dimensions

Here we shall use the following consequence of Assumption 4. Indeed, turning from squared norms to quadratic forms we see that

$$\langle L^{-2q}u, u\rangle \asymp \langle T_\nu^{2aq}u, u\rangle, \quad u \in \mathcal{H}.$$

The Weyl Monotonicity Theorem (Bhatia, 1997, Cor. III.2.3) yields that then $s_j(L^{-2q}) \asymp s_j(T_\nu^{2aq})$, $j = 1, 2, \ldots$, or simplified that $s_j(L^{-1}) \asymp s_j^a(T_\nu)$, $j = 1, 2, \ldots$ by spectral calculus. Here $s_j(L^{-1})$ and $s_j(T_\nu)$ denote the singular numbers of the operators. Similarly, we obtain from (26) that $s_j(L_\nu) \asymp s_j^{\frac{1-2a}{a}}(L^{-1})$, and a fortiori that $s_j(L_\nu) \asymp s_j^{1-2a}(T_\nu)$.

5.3 Lower bound

In order to show the optimality of the error bounds as discussed in Table 1, we shall assure that the decay of the effective dimension cannot be faster than asserted in Assumption 7.

[1] We shall suppress the recalculations of the corresponding constants.

Assumption 10 (decay of singular number) There is a constant $c > 0$ such that the singular numbers of the operator T_ν obey

$$s_j(T_\nu) \geq cj^{-1/b}, \quad j = 1, 2, \ldots$$

Notice that this yields that $\mathcal{N}(\lambda) \geq c\lambda^{-b}$, such that this is the limiting case for which Assumption 7 holds. Hence, the assumed decay of the singular numbers of T_ν is best possible by order. The following is reported in Blanchard and Mücke (2018) for the problem (2): Under smoothness r with respect to the operator L_ν, and with the decay of the singular numbers $s_j(L_\nu)$ not faster than $j^{-1/b}$, the optimal rate is of the order $\left(\frac{1}{\sqrt{m}}\right)^{\frac{2r}{2r+b+1}}$. In the present context, we have to assign $r \leftarrow \frac{ar}{1-2a}$ and $b \leftarrow \frac{b}{1-2a}$. This yield a lower bound of the order

$$\left(\frac{1}{\sqrt{m}}\right)^{\frac{2ar/(1-2a)}{2ar/(1-2a)+b/(1-2a)+1}} = \left(\frac{1}{\sqrt{m}}\right)^{\frac{2ar}{2ar+b+1-2a}}$$

for the range $\frac{ar}{1-2a} \leq p$.

This corresponds to the upper bound for $a \leq \frac{1}{2}$, $aq \leq p$, $r \leq q \leq r + \frac{b+1}{2a}$, as discussed in the last row of Table 1, and it shows that the rate is of optimal order.

6 Conclusion

We investigated regularization schemes in Hilbert scales for linear inverse (learning) problems. Regularized solutions are constructed under the requirement that these belong to $\mathcal{D}(L)$, for the (unbounded) operator L, which generates the scale. Clearly, this may be extended to the case that the regularized solutions belong to $\mathcal{D}(L^s)$ for some $s > 0$, simply be considering L^s as a generator of the (same) scale.

We draw the following conclusions. Some arguments consider the cases of power type conditions, and for this, we recourse to Tables 1 and 2 for details.

Optimal rates: In the regular case, we can achieve the optimal rates of convergence provided that the benchmark smoothness q is chosen in the appropriate region (see Sect. 5.3). In contrast, in the mis-specified case (oversmoothing) we can only prove sub-optimal rates of convergence. By now no techniques are known which are capable to improve the rates in this case.

Saturation effects: In case $q = r$, we observe from the above analysis that optimal rates can be proven for the range $ar \leq p$, provided that the scheme has qualification p. For standard regularization schemes, this would hold for the range $\frac{ar}{1-2a} \leq p$, only. Hence, the saturation effect is delayed here.

Convergence rates without source condition: Typically, rates of convergence are shown under smoothness in terms of source conditions. Here we establish error bounds by using the concept of distance functions, measuring the violation of a benchmark source condition. When specifying smoothness as a source condition, we use known bounds of the considered distance function. This provides us with convergence rates in terms of the sample size.

Source conditions: When studying kernel methods, the smoothness of the true solution is measured in terms of the source condition with respect to the covariance operator, and

hence may hardly be checked. We consider source conditions in terms of the Hilbert scale. This has a clear meaning, and it is independent of the choice of kernel. However, the chosen kernel comes into play when requiring the validity of a link condition.

A Proofs of Sect. 2

Proof (Proof of Proposition 1) The first assertions are a consequence of Heinz Inequality (9) with $a := 1/q < 1$. For the last one, we argue as follows. Since ϱ^2 is assumed to be sub-linear. Hence we find that

$$\begin{aligned}\left\|\varrho(T_\nu)\left(T_\nu+\lambda I\right)^{-1/2}\right\|_{\mathcal{L}(\mathcal{H})} &= \frac{1}{\sqrt{\lambda}}\left\|\varrho(T_\nu)\left(\lambda\left(T_\nu+\lambda I\right)^{-1}\right)^{1/2}\right\|_{\mathcal{L}(\mathcal{H})}\\ &\le \frac{1}{\sqrt{\lambda}}\left\|\varrho^2(T_\nu)\left(\lambda\left(T_\nu+\lambda I\right)^{-1}\right)\right\|^{1/2}_{\mathcal{L}(\mathcal{H})}\\ &\le \frac{\varrho(\lambda)}{\sqrt{\lambda}},\end{aligned}$$

which completes the proof. □

For proving Proposition 2 we start with the following technical result.

Lemma 1 *Suppose that the function ϱ from the link condition, Assumption 4 is such that the function $t \mapsto \left(\varrho^{2q}\right)^{-1}(t)$ is operator concave, and that there is some $n \in \mathbb{N}$ for which the function $t \mapsto \varrho^{-1}(t)/t^n$ is concave. Under Assumption 4 we have that*

$$\frac{s_j\left(T_\nu\right)}{s_j\left(\varrho^2(T_\nu)\right)} \le \beta^{n-1} s_j\left(L_\nu\right) \le \beta^{2n}\frac{s_j\left(T_\nu\right)}{s_j\left(\varrho^2(T_\nu)\right)}, \quad j = 1, 2, \ldots$$

Proof The proof is based on two consequences of Assumption 4, which, in terms of the partial ordering for self-adjoint operators in Hilbert space can be restated as

$$\langle (L^{-1})^{2q}u, u\rangle_{\mathcal{H}} \le \langle \varrho^{2q}(T_\nu)u, u\rangle_{\mathcal{H}} \le \langle (\beta L^{-1})^{2q}u, u\rangle_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

Since the operator concave function $t \mapsto \left(\varrho^{2q}\right)^{-1}(t)$ respects the partial ordering we obtain[2] that

$$\langle \varrho^{-1}(L^{-1})u, u\rangle_{\mathcal{H}} \le \langle T_\nu u, u\rangle_{\mathcal{H}} \le \langle \varrho^{-1}(\beta L^{-1})u, u\rangle_{\mathcal{H}}.$$

Letting $u := Lv \in \mathcal{H}$, and since by construction $T_\nu = L^{-1}L_\nu L^{-1}$ we deduce that

$$\langle \varrho^{-1}(L^{-1})L^2 v, v\rangle_{\mathcal{H}} \le \langle L_\nu v, v\rangle_{\mathcal{H}} \le \langle \varrho^{-1}(\beta L^{-1})L^2 v, v\rangle_{\mathcal{H}}, \quad v \in \mathcal{D}(L).$$

The sub-linearity of ϱ^2 implies that the function $t \mapsto \varrho^{-1}(t)/t^2$ is non-decreasing, such that the operator $\varrho^{-1}(\beta L^{-1})L^2$ is bounded, and hence the above inequality extends to $v \in \mathcal{H}$. Next we apply the Weyl Monotonicity Theorem (Bhatia, 1997, Cor. III.2.3) to see that

[2] we use that $\left(\varrho^{2q}\right)^{-1}(t^{2q}) = (\varrho)^{-1}(t)$.

$$\frac{s_j\big(\varrho^{-1}(L^{-1})\big)}{s_j^2(L^{-1})} \le s_j\big(L_\nu\big) \le \frac{s_j\big(\varrho^{-1}(\beta L^{-1})\big)}{s_j^2(L^{-1})}, \quad j = 1, 2, \dots \tag{28}$$

Applying this theorem to the first inequality in Proposition 1 we also find that

$$s_j\big(\varrho^{-1}(L^{-1})\big) \le s_j\big(T_\nu\big) \le s_j\big(\varrho^{-1}(\beta L^{-1})\big),\ j = 1, 2, \dots$$

To proceed we shall use the sub-linearity of the function ϱ^2, and the concavity of the function $\varsigma(t) := \varrho^{-1}(t)/t^n$. This yields that $\varsigma(\beta t) \le \beta\varsigma(t),\ \beta \ge 1$ and overall, we find that

$$\begin{aligned}\frac{s_j\big(\varrho^{-1}(L^{-1})\big)}{s_j^2(L^{-1})} &\le \frac{s_j\big(T_\nu\big)}{s_j\big(\varrho^2(T_\nu)\big)} \le \frac{s_j\big(\varrho^{-1}(\beta L^{-1})\big)}{s_j^2(\beta L^{-1})} \\ &= \beta^{n-2} s_j^{n-2}(L^{-1}) \frac{s_j\big(\varrho^{-1}(\beta L^{-1})\big)}{s_j^n(\beta L^{-1})} \le \beta^{n-1} \frac{s_j\big(\varrho^{-1}(L^{-1})\big)}{s_j^2(L^{-1})}.\end{aligned}$$

This, together with the inequalities (28) gives

$$\frac{s_j\big(T_\nu\big)}{s_j\big(\varrho^2(T_\nu)\big)} \le \beta^{n-1} s_j\big(L_\nu\big) \le \beta^{2n} \frac{s_j\big(T_\nu\big)}{s_j\big(\varrho^2(T_\nu)\big)},$$

and the proof is complete. □

Proof (Proof of Proposition 2) Since the function $t \mapsto t/\varrho^2(t)$ is assumed to be an index function, we find from Lemma 1 that the assertion

$$\beta^{n+1} \frac{\lambda}{\varrho^2(\lambda)} \le s_j(L_\nu) \quad \text{implies} \quad \lambda \le s_j(T_\nu)$$

holds true. This yields

$$\#\left\{j, \quad s_j(L_\nu) \ge \beta^{n+1} \frac{\lambda}{\varrho^2(\lambda)}\right\} \le \#\big\{j, \quad s_j(T_\nu) \ge \lambda\big\}, \quad \lambda \le \big\|T_\nu\big\|_{\mathcal{L}(\mathcal{H})}. \tag{29}$$

As a consequence of (Lin et al., 2015, Prop. 6) there is $\widetilde{C}$ such that

$$\mathcal{N}_{L_\nu}(\lambda) \le \widetilde{C}\#\big\{j, \quad s_j(L_\nu) \ge \lambda\big\}.$$

This, together with (29), implies that

$$\begin{aligned}\mathcal{N}_{L_\nu}\left(\beta^{n+1} \frac{\lambda}{\varrho^2(\lambda)}\right) &\le \widetilde{C}\#\left\{j, \quad s_j(L_\nu) \ge \beta^{n+1} \frac{\lambda}{\varrho^2(\lambda)}\right\} \le \widetilde{C}\#\big\{j, \quad s_j(T_\nu) \ge \lambda\big\} \\ &= 2\widetilde{C} \sum_{s_j(T_\nu) \ge \lambda} \frac{1}{2} \le 2\widetilde{C} \sum_{j=1}^{\infty} \frac{s_j(T_\nu)}{\lambda + s_j(T_\nu)} = 2\widetilde{C}\mathcal{N}_{T_\nu}(\lambda), \quad \lambda \le \big\|T_\nu\big\|_{\mathcal{L}(\mathcal{H})}.\end{aligned}$$

Since the function $\lambda \mapsto \lambda\mathcal{N}_{L_\nu}(\lambda)$ is non-decreasing we continue to bound

$$\mathcal{N}_{L_\nu}\left(\frac{\lambda}{\varrho^2(\lambda)}\right) \le \beta^{n+1}\mathcal{N}_{L_\nu}\left(\beta^{n+1}\frac{\lambda}{\varrho^2(\lambda)}\right) \le 2\beta^{n+1}\widetilde{C}\mathcal{N}_{T_\nu}(\lambda), \quad \lambda \le \|T_\nu\|_{\mathcal{L}(\mathcal{H})}, \tag{30}$$

which completes the proof. □

Proof (Proof of Proposition 3) The first assertion is a restatement of (Mathé and Pereverzev, 2003, Proposition 3). For the second assertion, we stress that $(\lambda+\sigma)^p \le 2^{p-1}(\lambda^p+\sigma^p)$, which follows from convexity. This yields

$$\begin{aligned}|r_\lambda(\sigma)|\varphi(\lambda+\sigma) &\le |r_\lambda(\sigma)|(\lambda+\sigma)^p\frac{\varphi(\lambda+\sigma)}{(\lambda+\sigma)^p}\\ &\le 2^{p-1}|r_\lambda(\sigma)|(\lambda^p+\sigma^p)\frac{\varphi(\lambda)}{\lambda^p} \le 2^p c_p\lambda^p\frac{\varphi(\lambda)}{\lambda^p},\end{aligned}$$

which implies the second assertion and completes the proof. □

B Proofs of Sect. 3

Proof (Proof of Theorem 1) For the minimizer f_ρ^R of the distance function defined in (18), the error can be expressed as follows:

$$f_\rho - f_{\mathbf{z},\lambda} = L^{-1}\Big\{r_\lambda(T_\mathbf{x})L(f_\rho - f_\rho^R) + r_\lambda(T_\mathbf{x})Lf_\rho^R + g_\lambda(T_\mathbf{x})B_\mathbf{x}^*(S_\mathbf{x}Af_\rho - \mathbf{y})\Big\}.$$

By using Proposition 1 the error for the regularized solution can be bounded as

$$\left\|f_\rho - f_{\mathbf{z},\lambda}\right\|_\mathcal{H} \tag{31}$$

$$\le \left\|L^{-1}r_\lambda(T_\mathbf{x})L(f_\rho - f_\rho^R)\right\|_\mathcal{H} + \left\|L^{-1}r_\lambda(T_\mathbf{x})Lf_\rho^R\right\|_\mathcal{H} + \left\|L^{-1}g_\lambda(T_\mathbf{x})B_\mathbf{x}^*(S_\mathbf{x}Af_\rho - \mathbf{y})\right\|_\mathcal{H}$$

$$\le d(R)\underbrace{\left\|L^{-1}r_\lambda(T_\mathbf{x})L\right\|_{\mathcal{L}(\mathcal{H})}}_{I_1} + \underbrace{\left\|\varrho(T_\nu)r_\lambda(T_\mathbf{x})Lf_\rho^R\right\|_\mathcal{H}}_{I_2} + \underbrace{\left\|\varrho(T_\nu)g_\lambda(T_\mathbf{x})B_\mathbf{x}^*(S_\mathbf{x}Af_\rho - \mathbf{y})\right\|_\mathcal{H}}_{I_3}. \tag{32}$$

We shall bound each summand on the right in (31).

I_1: By Lemma 2 we find that

$$\left\|L^{-1}r_\lambda(T_\mathbf{x})L\right\|_{\mathcal{L}(\mathcal{H})} \le 1 + (B+D)\left(\Xi^\varrho\Xi^\upsilon + \Xi\varrho(\lambda)(\varrho(\lambda)+1)\frac{\Lambda}{\sqrt{\lambda}}\right)$$

with Ξ^ϱ, Λ as in (14), (15) and $\upsilon(t) := t/\varrho(t)$, $t > 0$. From the estimates of Propositions 4, 5 we get with confidence $1-\eta/2$ that

$$\left\|L^{-1}r_\lambda(T_{\mathbf{x}})L\right\|_{\mathcal{L}(\mathcal{H})} \leq 1 + (B+D)\Big\{(2\kappa+1)^8 + 2(2\kappa+1)^4(\varrho(\lambda)+1) \times\left(\frac{\tilde{\kappa}\varrho(\lambda)}{m\lambda} + \sqrt{\frac{\tilde{\kappa}\varrho^2(\lambda)\mathcal{N}_{L_v}(\lambda)}{m\lambda}}\right)\Big\}\log^4\left(\frac{4}{\eta}\right), \tag{33}$$

For $\vartheta(\lambda) := \frac{\lambda}{\varrho^2(\lambda)}$, by using that $\lambda \mapsto \lambda\mathcal{N}_{L_v}(\lambda)$ is an increasing function, and $\lambda \leq \vartheta(\lambda)$, for λ small enough, we get

$$\lambda\mathcal{N}_{L_v}(\lambda) \leq \vartheta(\lambda)\mathcal{N}_{L_v}(\vartheta(\lambda)).$$

This together with Proposition 2 implies that

$$\varrho^2(\lambda)\mathcal{N}_{L_v}(\lambda) \leq \mathcal{N}_{L_v}\left(\frac{\lambda}{\varrho^2(\lambda)}\right) \leq 2\beta^{n+1}\widetilde{C}\mathcal{N}_{T_v}(\lambda). \tag{34}$$

Under the condition (12) from the estimates (13), (33), (34) we get with confidence $1-\eta/2$:

$$\left\|L^{-1}r_\lambda(T_{\mathbf{x}})L\right\|_{\mathcal{L}(\mathcal{H})} \leq 1 + (B+D)\beta^{n+1}\widetilde{C}C_{\kappa,\tilde{\kappa}}\log^4\left(\frac{4}{\eta}\right), \tag{35}$$

where $C_{\kappa,\tilde{\kappa}}$ depends on $\kappa, \tilde{\kappa}$.

I_2: By construction of f_ρ^R we have that $f_\rho^R = L^{-1}v$, $\|v\|_{\mathcal{H}} \leq R$. Using the fact that p covers ϱ we bound

$$\left\|\varrho(T_v)r_\lambda(T_{\mathbf{x}})Lf_\rho^R\right\|_{\mathcal{H}} \leq R\Xi^\varrho\left\|\varrho(T_{\mathbf{x}}+\lambda I)r_\lambda(T_{\mathbf{x}})\right\|_{\mathcal{L}(\mathcal{H})} \leq 2R\Xi^\varrho\varrho(\lambda). \tag{36}$$

I_3: For the last summand we argue

$$\begin{aligned}
&\left\|\varrho(T_v)g_\lambda(T_{\mathbf{x}})B_{\mathbf{x}}^*(S_{\mathbf{x}}Af_\rho - \mathbf{y})\right\|_{\mathcal{H}} \\
&\leq \Xi^{\frac{1}{2}}\Xi^\varrho\Psi\left\|g_\lambda(T_{\mathbf{x}})\varrho(T_{\mathbf{x}}+\lambda I)(T_{\mathbf{x}}+\lambda I)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H})} \\
&\leq \Xi^{\frac{1}{2}}\Xi^\varrho\Psi\sup_{t\in[0,\kappa^2]}\varrho(t+\lambda)(t+\lambda)^{\frac{1}{2}}\left|g_\lambda(t)\right| \\
&\leq \Xi^{\frac{1}{2}}\Xi^\varrho\Psi\left(\sup_{t\in[0,\kappa^2]}\varrho(t+\lambda)(t+\lambda)^{-\frac{1}{2}}\right)\left\{\lambda\sup_{t\in[0,\kappa^2]}\left|g_\lambda(t)\right| + \sup_{t\in[0,\kappa^2]}\left|tg_\lambda(t)\right|\right\} \\
&\leq \Xi^{\frac{1}{2}}\Xi^\varrho\Psi\{B+D\}\varrho(\lambda)\lambda^{-\frac{1}{2}},
\end{aligned} \tag{37}$$

where $\Xi^{1/2}$ and Ψ were as in (14) and (17).

Summarizing, using the estimates of Propositions 4, 5, and (35)–(37), we get with confidence $1-\eta$ that

$$\left\|f_\rho - f_{\mathbf{z},\lambda}\right\|_{\mathcal{H}} \leq C\left[d(R) + \varrho(\lambda)\left\{R + \frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2\mathcal{N}_{T_v}(\lambda)}{m\lambda}}\right\}\right]\log^4\left(\frac{4}{\eta}\right). \tag{38}$$

For any parameter choice λ satisfying the condition (12) using the inequality (13) we get that

$$\frac{\kappa M}{m\lambda} \leq \frac{\kappa M}{\mathcal{N}_{T_v}(1)}$$

and

$$\sqrt{\frac{\Sigma^2 \mathcal{N}_{T_v}(\lambda)}{m\lambda}} \leq \Sigma.$$

This implies

$$R + \frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2 \mathcal{N}_{T_v}(\lambda)}{m\lambda}} \leq 2R, \tag{39}$$

provided that $R \geq \Sigma + \kappa M / \mathcal{N}_{T_v}(1)$. Inserting the bound from inequality (39) into the estimate (38) completes the proof. □

Proof (Proof of Theorem 2) For the minimizer f_ρ^R of the distance function defined in (20), the error can be expressed as follows:

$$L(f_\rho - f_{\mathbf{z},\lambda}) = r_\lambda(T_\mathbf{x}) L(f_\rho - f_\rho^R) + r_\lambda(T_\mathbf{x}) L f_\rho^R + g_\lambda(T_\mathbf{x}) B_\mathbf{x}^*(S_\mathbf{x} A f_\rho - \mathbf{y}).$$

First, we estimate the error in the interpolation norm for some index function ζ:

$$\begin{aligned} \left\| \zeta(T_v) L(f_\rho - f_{\mathbf{z},\lambda}) \right\|_\mathcal{H} \leq & d_q(R) \underbrace{\left\| \zeta(T_v) r_\lambda(T_\mathbf{x}) \right\|_{\mathcal{L}(\mathcal{H})}}_{I_1} + \underbrace{\left\| \zeta(T_v) r_\lambda(T_\mathbf{x}) L f_\rho^R \right\|_\mathcal{H}}_{I_2} \\ & + \underbrace{\left\| \zeta(T_v) g_\lambda(T_\mathbf{x}) B_\mathbf{x}^*(S_\mathbf{x} A f_\rho - \mathbf{y}) \right\|_\mathcal{H}}_{I_3}. \end{aligned} \tag{40}$$

I_1: We bound

$$\begin{aligned} \left\| \zeta(T_v) r_\lambda(T_\mathbf{x}) \right\|_{\mathcal{L}(\mathcal{H})} &\leq \left\| \zeta(T_v + \lambda I) r_\lambda(T_\mathbf{x}) \right\|_{\mathcal{L}(\mathcal{H})} \\ &\leq \Xi^\zeta \left\| \zeta(T_\mathbf{x} + \lambda I) r_\lambda(T_\mathbf{x}) \right\|_{\mathcal{L}(\mathcal{H})} \leq \Xi^\zeta c_p \zeta(\lambda). \end{aligned} \tag{41}$$

I_2: For the minimizer $f_\rho^R = L^{-q} g$ of the distance function (20), we observe from Proposition 1 that there is $v \in \mathcal{H}$ such that $L f_\rho^R = L^{-(q-1)} g = \varphi(T_v) v$, $\|v\|_\mathcal{H} \leq R$. Thus by assuming that the $\varphi = \varphi_1 \varphi_2$ (with φ_1 being sub-linear and φ_2 Lipschitz with constant one) we continue bounding

$$\begin{aligned} r_\lambda(T_\mathbf{x}) L f_\rho^R &= r_\lambda(T_\mathbf{x}) \varphi(T_v) v \\ &= r_\lambda(T_\mathbf{x}) \varphi_2(T_\mathbf{x}) \varphi_1(T_v) v + r_\lambda(T_\mathbf{x})(\varphi_2(T_v) - \varphi_2(T_\mathbf{x})) \varphi_1(T_v) v. \end{aligned}$$

Then we get

$$\begin{aligned}
\left\|\zeta(T_\nu)r_\lambda(T_{\mathbf{x}})Lf_\rho^R\right\|_{\mathcal{H}} &= \left\|\zeta(T_\nu)r_\lambda(T_{\mathbf{x}})\varphi(T_\nu)v\right\|_{\mathcal{H}} \\
&\le \Xi^\zeta\Big\{\left\|\zeta(T_{\mathbf{x}}+\lambda I)r_\lambda(T_{\mathbf{x}})\varphi_2(T_{\mathbf{x}})\varphi_1(T_\nu)v\right\|_{\mathcal{H}} \\
&\quad + \left\|\zeta(T_{\mathbf{x}}+\lambda I)r_\lambda(T_{\mathbf{x}})(\varphi_2(T_\nu)-\varphi_2(T_{\mathbf{x}}))\varphi_1(T_\nu)v\right\|_{\mathcal{H}}\Big\} \\
&\le R\Xi^\zeta\Big\{\left\|\zeta(T_{\mathbf{x}}+\lambda I)r_\lambda(T_{\mathbf{x}})\varphi_2(T_{\mathbf{x}})\varphi_1(T_{\mathbf{x}}+\lambda I)\right\|_{\mathcal{L}(\mathcal{H})} \\
&\quad \times\left\|\left(\frac{1}{\varphi_1}\right)(T_{\mathbf{x}}+\lambda I)\varphi_1(T_\nu+\lambda I)\right\|_{\mathcal{L}(\mathcal{H})} \\
&\quad +\varphi_1(\kappa^2)\left\|\zeta(T_{\mathbf{x}}+\lambda I)r_\lambda(T_{\mathbf{x}})\right\|_{\mathcal{L}(\mathcal{H})}\left\|T_\nu-T_{\mathbf{x}}\right\|_{\mathcal{L}(\mathcal{H})}\Big\} \\
&\le R\Xi^\zeta\Big\{\Xi^{\varphi_1}\sup_{t\in[0,\kappa^2]}\left\{\left|r_\lambda(t)\right|\varphi_2(t)\zeta(t+\lambda)\varphi_1(t+\lambda)\right\} \\
&\quad +\varphi_1(\kappa^2)\left\|T_\nu-T_{\mathbf{x}}\right\|_{\mathcal{L}(\mathcal{H})}\sup_{t\in[0,\kappa^2]}\left\{\left|r_\lambda(t)\right|\zeta(t+\lambda)\right\}\Big\} \\
&\le R2^q c_p\zeta(\lambda)\Xi^\zeta\Big\{\Xi^{\varphi_1}\varphi(\lambda)+\varphi_1(\kappa^2)\left\|T_\nu-T_{\mathbf{x}}\right\|_{\mathcal{L}(\mathcal{H})}\Big\},
\end{aligned} \tag{42}$$

because of the qualification of the regularization.

I_3: From the arguments used in (37), we get

$$\left\|\zeta(T_\nu)g_\lambda(T_{\mathbf{x}})B_{\mathbf{x}}^*(S_{\mathbf{x}}Af_\rho-\mathbf{y})\right\|_{\mathcal{H}} \le \Xi^{\frac{1}{2}}\Xi^\zeta\Psi\{B+D\}\zeta(\lambda)\lambda^{-\frac{1}{2}}. \tag{43}$$

Overall, using Propositions 4–5 and (41)–(43) in (40) we obtain with confidence $1-\eta$ that

$$\begin{aligned}
&\left\|\zeta(T_\nu)L\left(f_{\mathbf{z},\lambda}-f_\rho\right)\right\|_{\mathcal{H}} \\
&\quad\le C\zeta(\lambda)\left\{d_q(R)+R\left(\varphi(\lambda)+\frac{1}{\sqrt{m}}\right)+\frac{\kappa M}{m\lambda}+\sqrt{\frac{\Sigma^2\mathcal{N}_{T_\nu}(\lambda)}{m\lambda}}\right\}\log^4\left(\frac{4}{\eta}\right).
\end{aligned} \tag{44}$$

The fact that $\mathcal{N}_{T_\nu}(\lambda)$ is decreasing function of λ with the inequality (12) implies that

$$\frac{\kappa M}{m\lambda} \le \frac{\kappa M}{m\lambda}\frac{\mathcal{N}_{T_\nu}(\lambda)}{\mathcal{N}_{T_\nu}(1)} \le \frac{\kappa M}{\mathcal{N}_{T_\nu}(1)}\sqrt{\frac{\mathcal{N}_{T_\nu}(\lambda)}{m\lambda}}.$$

This, together with (44) yields the first result.

For the last two estimates in Theorem 2, by using Proposition 1 we get

$$\left\|f_\rho-f_{\mathbf{z},\lambda}\right\|_{\mathcal{H}} = \left\|L^{-1}\left\{L(f_\rho-f_{\mathbf{z},\lambda})\right\}\right\|_{\mathcal{H}} \le \left\|\varrho(T_\nu)L(f_\rho-f_{\mathbf{z},\lambda})\right\|_{\mathcal{H}},$$

and

$$\left\|I_\nu A(f_{\mathbf{z},\lambda}-f_\rho)\right\|_{\mathscr{L}^2(X,\nu;Y)} = \left\|T_\nu^{1/2}L(f_{\mathbf{z},\lambda}-f_\rho)\right\|_{\mathcal{H}}.$$

These two upper bounds can now be estimated from the general bound by letting $\zeta := \varrho$ and $\zeta(t) := t^{\frac{1}{2}}$, respectively. We also use that ϱ^2 is sub-linear, and this completes the proof. □

C Probabilistic bounds

In the following proposition, we present the standard perturbation inequalities in learning theory which measure the effect of random sampling in the probabilistic sense. The following two propositions can be proved using the arguments given in Step 2.1. of (Caponnetto and De Vito, 2007, Thm. 4).

Proposition 4 *Suppose Assumptions 1–3 hold true, then for $m \in \mathbb{N}$ and $0 < \eta < 1$, each of the following estimate holds with the confidence $1 - \eta$,*

$$\Psi = \Psi(\lambda) := \left\| (T_\nu + \lambda I)^{-1/2} B^*_{\mathbf{x}}(\mathbf{y} - S_{\mathbf{x}} A(f_\rho)) \right\|_{\mathcal{H}} \leq 2\left(\frac{\kappa M}{m\sqrt{\lambda}} + \sqrt{\frac{\Sigma^2 \mathcal{N}_{T_\nu}(\lambda)}{m}} \right) \log\left(\frac{2}{\eta}\right),$$

$$\Upsilon = \Upsilon(\lambda) := \left\| (T_\nu + \lambda I)^{-1/2} (T_\nu - T_{\mathbf{x}}) \right\|_{HS} \leq 2\left(\frac{\kappa^2}{m\sqrt{\lambda}} + \sqrt{\frac{\kappa^2 \mathcal{N}_{T_\nu}(\lambda)}{m}} \right) \log\left(\frac{2}{\eta}\right),$$

$$\|T_\nu - T_{\mathbf{x}}\|_{HS} \leq 2\left(\frac{\kappa^2}{m} + \frac{\kappa^2}{\sqrt{m}} \right) \log\left(\frac{2}{\eta}\right)$$

and

$$\Lambda = \Lambda(\lambda) := \left\| (L_\nu + \lambda I)^{-1/2} (L_{\mathbf{x}} - L_\nu) \right\|_{HS} \leq 2\left(\frac{\tilde{\kappa}^2}{m\sqrt{\lambda}} + \sqrt{\frac{\tilde{\kappa}^2 \mathcal{N}_{L_\nu}(\lambda)}{m}} \right) \log\left(\frac{2}{\eta}\right).$$

In the following proposition, the probabilistic estimate of the first term can be established under the condition (12) on the regularization parameter λ, and the sample size m. The last two estimates are obtained by using (Blanchard, 2019, Prop. A.2).

Proposition 5 *Suppose Assumption 3 and the condition (12) hold true. Let $\zeta : \mathbb{R}^+ \to \mathbb{R}^+$ be a nondecreasing and sub-linear function, then for $m \in \mathbb{N}$ and $0 < \eta < 1$, each of the following estimates hold with the confidence $1 - \eta$,*

$$\Upsilon = \left\| (T_\nu + \lambda I)^{-\frac{1}{2}} (T_\nu - T_{\mathbf{x}}) \right\|_{HS} \leq \sqrt{\lambda} 2\kappa(2\kappa + 1) \log\left(\frac{2}{\eta}\right),$$

$$\Xi^s = \Xi^s(\lambda) := \left\| (T_{\mathbf{x}} + \lambda I)^{-s} (T_\nu + \lambda I)^s \right\|_{\mathcal{L}(\mathcal{H})} \leq \left(\frac{\Upsilon}{\sqrt{\lambda}} + 1 \right)^{2s} \leq \left((2\kappa + 1)^2 \log\left(\frac{2}{\eta}\right) \right)^{2s}$$

for $0 \leq s \leq 1$ and

$$\Xi^{\zeta} = \Xi^{\zeta}(\lambda) := \left\| \left(\frac{1}{\zeta}\right)(T_{\mathbf{x}} + \lambda I)\zeta(T_{\nu} + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})} \leq \left(\frac{\Upsilon}{\sqrt{\lambda}} + 1\right)^2 \leq \left((2\kappa + 1)^2 \log\left(\frac{2}{\eta}\right)\right)^2.$$

Lemma 2 *Suppose Assumption 4 holds true. Let g_λ be any regularization with residual function r_λ. Then for $\upsilon(t) = t/\varrho(t)$, we have that*

$$\left\| L^{-1} r_\lambda(T_{\mathbf{x}}) L \right\|_{\mathcal{L}(\mathcal{H})} \leq 1 + (B + D)\left(\Xi^{\varrho}\Xi^{\upsilon} + \Xi\varrho(\lambda)(\varrho(\lambda) + 1)\frac{\Lambda}{\sqrt{\lambda}} \right). \tag{45}$$

Proof For $L_{\mathbf{x}} = A^* S_{\mathbf{x}}^* S_{\mathbf{x}} A$ and $L_{\nu} = A^* I_{\nu}^* I_{\nu} A$ with the fact that $T_{\mathbf{x}} - T_{\nu} = L^{-1}\left(L_{\mathbf{x}} - L_{\nu}\right)L^{-1}$, the proof will be based on the following decomposition

$$\begin{aligned} L^{-1} r_\lambda(T_{\mathbf{x}}) L =& I - L^{-1} g_\lambda(T_{\mathbf{x}}) T_{\nu} L + L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1}(L_{\nu} - L_{\mathbf{x}}) \\ =& I - L^{-1} g_\lambda(T_{\mathbf{x}}) T_{\nu} L + \lambda L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1}(L_{\nu} + \lambda I)^{-1}\left(L_{\nu} - L_{\mathbf{x}}\right) \\ &+ L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1} L_{\nu}(L_{\nu} + \lambda I)^{-1}\left(L_{\nu} - L_{\mathbf{x}}\right), \end{aligned}$$

and this yields the estimate

$$\begin{aligned} &\left\| L^{-1} r_\lambda(T_{\mathbf{x}}) L \right\|_{\mathcal{L}(\mathcal{H})} \\ &\quad \leq 1 + \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) T_{\nu} L \right\|_{\mathcal{L}(\mathcal{H})} + \lambda \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1}(L_{\nu} + \lambda I)^{-1}\left(L_{\nu} - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\ &\qquad + \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1} L_{\nu}(L_{\nu} + \lambda I)^{-1}\left(L_{\nu} - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\ &\quad = 1 + I_1 + \lambda I_2 + I_3. \end{aligned} \tag{46}$$

We observe that $\left\| g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})} \leq B + D$. For the function $\upsilon(t) = t/\varrho(t)$, we can bound I_1 as

$$\begin{aligned} &\left\| L^{-1} g_\lambda(T_{\mathbf{x}}) T_{\nu} L \right\|_{\mathcal{L}(\mathcal{H})} \\ &\quad \leq \left\| L^{-1}\frac{1}{\varrho}\left(T_{\nu} + \lambda I\right) \right\|_{\mathcal{L}(\mathcal{H})} \left\| \varrho\left(T_{\nu} + \lambda I\right)\frac{1}{\varrho}\left(T_{\mathbf{x}} + \lambda I\right) \right\|_{\mathcal{L}(\mathcal{H})} \left\| g_\lambda(T_{\mathbf{x}})\left(T_{\mathbf{x}} + \lambda I\right) \right\|_{\mathcal{L}(\mathcal{H})} \\ &\quad \times \left\| \left(T_{\mathbf{x}} + \lambda I\right)^{-1} \varrho\left(T_{\mathbf{x}} + \lambda I\right)\left(T_{\nu} + \lambda I\right)\frac{1}{\varrho}\left(T_{\nu} + \lambda I\right) \right\|_{\mathcal{L}(\mathcal{H})} \\ &\quad \times \left\| \varrho\left(T_{\nu} + \lambda I\right)\left(T_{\nu} + \lambda I\right)^{-1} T_{\nu} L \right\|_{\mathcal{L}(\mathcal{H})} \\ &\quad \leq \Xi^{\upsilon}\Xi^{\varrho}(B + D)\left\| L^{-1}\frac{1}{\varrho}\left(T_{\nu} + \lambda I\right) \right\|_{\mathcal{L}(\mathcal{H})} \left\| \varrho\left(T_{\nu} + \lambda I\right)\left(T_{\nu} + \lambda I\right)^{-1} T_{\nu} L \right\|_{\mathcal{L}(\mathcal{H})}. \end{aligned}$$

It remains to bound the second and third factors. From Proposition 1 we find that

$$\left\| L^{-1}\frac{1}{\varrho}(T_{\nu} + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})} \leq \left\| \varrho(T_{\nu})\frac{1}{\varrho}(T_{\nu} + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})} \leq \left\| \varrho(T_{\nu} + \lambda I)\frac{1}{\varrho}(T_{\nu} + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})} = 1.$$

Again, under Assumption 4 we find that

$$\left\| \varrho\left(T_{\nu} + \lambda I\right)\left(T_{\nu} + \lambda I\right)^{-1} T_{\nu} L \right\|_{\mathcal{L}(\mathcal{H})} \leq \left\| \varrho\left(T_{\nu} + \lambda I\right)\left(T_{\nu} + \lambda I\right)^{-1} T_{\nu}\frac{1}{\varrho}(T_{\nu}) \right\|_{\mathcal{L}(\mathcal{H})} \leq 1,$$

which finally yields that $I_1 \leq \Xi^{\upsilon}\Xi^{\varrho}(B+D)$.

The terms I_2, I_3 can be bounded as

$$\begin{aligned}
I_2 =& \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1} (L_\nu + \lambda I)^{-1} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \frac{1}{\sqrt{\lambda}} \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1} \right\|_{\mathcal{L}(\mathcal{H})} \left\| (L_\nu + \lambda I)^{-1/2} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \frac{1}{\sqrt{\lambda}} \left\| \varrho(T_\nu) g_\lambda(T_{\mathbf{x}}) \varrho(T_\nu) \right\|_{\mathcal{L}(\mathcal{H})} \left\| (L_\nu + \lambda I)^{-1/2} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \frac{1}{\sqrt{\lambda}} \left\| \varrho(T_\nu) (T_\nu + \lambda I)^{-1/2} \right\|^2_{\mathcal{L}(\mathcal{H})} \left\| (T_\nu + \lambda I)(T_{\mathbf{x}} + \lambda I)^{-1} \right\|_{\mathcal{L}(\mathcal{H})} \left\| g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda I) \right\|_{\mathcal{L}(\mathcal{H})} \\
& \times \left\| (L_\nu + \lambda I)^{-1/2} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \frac{\varrho^2(\lambda)}{\lambda^{3/2}} (B+D) \Xi \left\| (L_\nu + \lambda I)^{-1/2} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})},
\end{aligned} \tag{47}$$

and

$$\begin{aligned}
I_3 =& \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) L^{-1} L_\nu (L_\nu + \lambda I)^{-1} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) B_\nu^* \right\|_{\mathscr{L}^2(X,\nu;Y)\to\mathcal{H}} \left\| I_\nu A (L_\nu + \lambda I)^{-1} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{H}\to\mathscr{L}^2(X,\nu;Y)} \\
= & \left\| L^{-1} g_\lambda(T_{\mathbf{x}}) T_\nu^{1/2} \right\|_{\mathcal{L}(\mathcal{H})} \left\| L_\nu^{1/2} (L_\nu + \lambda I)^{-1} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \left\| \varrho(T_\nu) g_\lambda(T_{\mathbf{x}}) T_\nu^{1/2} \right\|_{\mathcal{L}(\mathcal{H})} \left\| (L_\nu + \lambda I)^{-1/2} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})} \\
\leq & \frac{\varrho(\lambda)}{\lambda^{1/2}} (B+D) \Xi \left\| (L_\nu + \lambda I)^{-1/2} \left(L_\nu - L_{\mathbf{x}}\right) \right\|_{\mathcal{L}(\mathcal{H})}.
\end{aligned} \tag{48}$$

This complete the proof.

Authors' contributions All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

Funding Open Access funding enabled and organized by Projekt DEAL. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) under Collaborative Research Centre SFB1294 (SFB-1294/1 - 318763901) and The Berlin Mathematics Research Center MATH+ (EXC-2046/1 - 390685689).

Availability of data and material Not applicable.Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent to participate Not applicable.

Consent for publication Not applicable.

Ethics approval Not applicable.

References

Agapiou, S., & Mathé, P. (2022). Designing truncated priors for direct and inverse Bayesian problems. *Electronic Journal of Statistics, 16*(1), 158–200.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68,* 337–404.

Bauer, F., Pereverzev, S., & Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of Complexity, 23*(1), 52–72.

Baumeister, J. (1987). *Stable solution of inverse problems. Advanced lectures in mathematics*. Friedrich Vieweg & Sohn.

Bhatia, R. (1997). Matrix analysis. In Grad. texts Math. (Vol. 169). Springer-Verlag, New York.

Blanchard, G., & Mathé, P. (2012). Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse Problems, 28*(11), 115011.

Blanchard, G., Mathé, P., & Mücke, N. (2019). Lepskii principle in supervised learning. arXiv:1905.10764.

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

Blanchard, G., & Mücke, N. (2020). Kernel regression, minimax rates and effective dimensionality: Beyond the regular case. *Analysis and Applications, 18*(04), 683–696.

Böttcher, A., Hofmann, B., Tautenhahn, U., & Yamamoto, M. (2006). Convergence rates for Tikhonov regularization from different kinds of smoothness conditions. *Applicable Analysis, 85*(5), 555–578.

Caponnetto, A., & De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics, 7*(3), 331–368.

Engl, H. W., Hanke, M., & Neubauer, A. (1996). *Regularization of inverse problems, volume 375. Math. Appl.* Kluwer Academic Publishers Group.

Gugushvili, S., van der Vaart, A., & Yan, D. (2020). Bayesian linear inverse problems in regularity scales. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques, 56*(3), 2081–2107.

Guo, Z.-C., Lin, S.-B., & Zhou, D.-X. (2017). Learning theory of distributed spectral algorithms. *Inverse Problems, 33,* 74009.

Hofmann, B. (2006). Approximate source conditions in Tikhonov–Phillips regularization and consequences for inverse problems with multiplication operators. *Mathematical Methods in the Applied Sciences, 29*(3), 351–371.

Hofmann, B., & Mathé, P. (2007). Analysis of profile functions for general linear regularization methods. *SIAM Journal on Numerical Analysis, 45*(3), 1122–1141.

Hofmann, B., & Mathé, P. (2018). Tikhonov regularization with oversmoothing penalty for non-linear ill-posed problems in Hilbert scales. *Inverse Problems, 34*(1), 15007.

Hofmann, B., & Mathé, P. (2020). A priori parameter choice in Tikhonov regularization with oversmoothing penalty for non-linear ill-posed problems. In J. Cheng, S. Lu, & M. Yamamoto (Eds.), *Inverse problems related top* (pp. 169–176). Springer.

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

Lin, K., Shuai, L., & Mathé, P. (2015). Oracle-type posterior contraction rates in Bayesian inverse problems. *Inverse Problems Imaging, 9*(3), 895–915.

Mair, B. A. (1994). Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM Journal on Mathematical Analysis, 25*(1), 135–147.

Mathé, P. (2019). Bayesian inverse problems with non-commuting operators. *Mathematics of Computation, 88*(320), 2897–2912.

Mathé, P., & Pereverzev, S. V. (2003). Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems, 19*(3), 789–803.

Mathé, P., & Tautenhahn, U. (2006). Interpolation in variable Hilbert scales with application to inverse problems. *Inverse Problems, 22*(6), 2271–2297.

Mathé, P., & Tautenhahn, U. (2007). Error bounds for regularization methods in Hilbert scales by using operator monotonicity. *Far East Journal of Mathematical Sciences, 24*(1), 1.

Micchelli, C. A., & Pontil, M. (2005). On learning vector-valued functions. *Neural Computation, 17*(1), 177–204.

Mücke, N., & Reiss, E. (2020). Stochastic gradient descent in Hilbert scales: Smoothness, preconditioning and earlier stopping. arXiv:2006.10840.

Nair, M. T. (1999). On Morozov's method for Tikhonov regularization as an optimal order yielding algorithm. *Journal of Analytical and Applied, 18,* 37–46.

Nair, M. T. (2002). Optimal order results for a class of regularization methods using unbounded operators. *Integral Equations and Operator Theory, 44*(1), 79–92.

Nair, M. T., Pereverzev, S. V., & Tautenhahn, U. (2005). Regularization in Hilbert scales under general smoothing conditions. *Inverse Problem, 21*(6), 1851–1869.

Natterer, F. (1984). Error bounds for Tikhonov regularization in Hilbert scales. *Applicable Analysis, 18*(1–2), 29–37.

Neubauer, A. (1988). An a posteriori parameter choice for tikhonov regularization in Hilbert scales leading to optimal convergence rates. *SIAM Journal on Numerical Analysis, 25*(6), 1313–1326.

Peller, V. V. (2016). Multiple operator integrals in perturbation theory. *Bulletin of Mathematical Sciences, 6*(1), 15–88.

Rastogi, A., Blanchard, G. & Mathé, P. (2020). Convergence analysis of Tikhonov regularization for non-linear statistical inverse learning problems. *Electronic Journal of Statistics, 14*(2), 2798–2841.

Rastogi, A., & Sampath, S. (2017). Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics, 3,* 3.

Shuai, L., Mathé, P., & Pereverzev, S. V. (2020). Balancing principle in supervised learning for a general regularization scheme. *Applied and Computational Harmonic Analysis, 48*(1), 123–148.

Shuai, L., & Pereverzev, S. (2013). *Regularization theory for ill-posed problems: Selected topics* (Vol. 58). Walter de Gruyter.

Smale, S., & Zhou, D.-X. (2003). Estimating the approximation error in learning theory. *Analysis and Applications, 01*(01), 17–41.

Tautenhahn, U. (1996). Error estimates for regularization methods in Hilbert scales. *SIAM Journal on Numerical Analysis, 33*(6), 2120–2130.

Zhang, T. (2002). Effective dimension and generalization of kernel learning. In *Proceedings of 15th International Conference Neural Information Processing System*, (pp. 454–461), MIT Press, Cambridge, MA.