Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems

Abhishake Rastogi, Gilles Blanchard, Peter Mathé

Electronic Journal of Statistics
Vol. 14 (2020) 2798–2841
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1735

Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems

Abhishake Rastogi[*1] **and Gilles Blanchard**[2] **and Peter Mathé**[3]

[1] *Institute of Mathematics, University of Potsdam, Karl-Liebknecht-Strasse 24-25, 14476 Potsdam, Germany, e-mail:* abhishake@uni-potsdam.de

[2] *Universit Paris-Saclay, CNRS, Laboratoire de mathmatiques d'Orsay F-91405, Orsay, France, e-mail:* gilles.blanchard@universite-paris-saclay.fr

[3] *Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany, e-mail:* peter.mathe@wias-berlin.de

Abstract: We study a non-linear statistical inverse problem, where we observe the noisy image of a quantity through a non-linear operator at some random design points. We consider the widely used Tikhonov regularization (or method of regularization) approach to estimate the quantity for the non-linear ill-posed inverse problem. The estimator is defined as the minimizer of a Tikhonov functional, which is the sum of a data misfit term and a quadratic penalty term. We develop a theoretical analysis for the minimizer of the Tikhonov regularization scheme using the concept of reproducing kernel Hilbert spaces. We discuss optimal rates of convergence for the proposed scheme, uniformly over classes of admissible solutions, defined through appropriate source conditions.

MSC2020 subject classifications: Primary 65J20; Secondary 62G08, 62G20, 65J15, 65J22.
Keywords and phrases: Statistical inverse problem, Tikhonov regularization, reproducing kernel Hilbert space, general source condition, minimax convergence rates.

Received November 2019.

1. Introduction

Within the *classical setup* of *supervised learning* we are given random samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$, where the elements $y_i,\ i = 1, \dots, m$ are noisy observations of $g(x_i),\ i = 1, \dots, m$ at random points $x_i,\ i = 1, \dots, m$ of the form

$$y_i := g(x_i) + \varepsilon_i \quad \text{for } i = 1, \dots, m. \tag{1.1}$$

In this introduction, we will assume implicitly that $Y = \mathbb{R}$, though in the main body of the paper the more general setting where Y is a Hilbert space will be considered. We assume that the random observations of $\mathbf{z}$ are independently and identically distributed according to some unknown joint probability distri-

*Corresponding Author.

bution ρ on the sample space $Z = X \times Y$. Further, we assume that the joint probability measure ρ can be described as $\rho(x, y) = \rho(y|x)\nu(x)$, where $\rho(y|x)$ is the conditional distribution of y given x and ν is the marginal distribution on X. The noise terms $(\varepsilon_i)_{i=1}^m$ are independent centered random variables satisfying $\int_Y \varepsilon_i d\rho(y_i|x_i) = 0$, so that $g(x) = \int_Y y d\rho(y|x)$. The cardinality m of the samples $\mathbf{z}$ is called sample size. The goal is to learn the functional relationship g: this may be viewed as a *direct* problem. In the nonparametric regression literature, it is most often assumed either that the design is fixed on a grid, or, in the case of random design as above, that the X-marginal (denoted hereafter by ν) used for sampling is known, or at least closely comparable (e.g. having upper and lower bounded density) to a known reference measure such as Lebesgue. A crucial difference in the point of view of statistical learning is that no such assumptions are made; the marginal distribution ν is unknown to the user and can be quite arbitrary. This is the point of view adopted in the present work.

It is often assumed that the function g belongs to some reproducing kernel Hilbert space, say $\mathcal{H}_2$, so that pointwise function evaluations are well-defined and continuous. There is vast literature for this specific setup, originating in works on nonparametric regression using spline methods [39], and having known a resurgence in statistical learning [2, 5]. Since pointwise evaluations can be represented by scalar products in a reproducing kernel Hilbert space, a mathematically equivalent setup consists of assuming that the input x takes values in a Hilbert space and the function g is linear; this appears in functional data analysis [6, 12, 20].

We shall contrast this with the inverse problem, given in terms of some, in general, non-linear mapping $A\colon \mathcal{D}(A) \subseteq \mathcal{H}_1 \to \mathcal{H}_2$, acting between the real separable Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, and with $A(f) = g$. Thus, the goal is to learn the implicitly given element $f \in \mathcal{H}_1$, still from the finite samples $\mathbf{z}$. Formally we assume the model

$$y_i := g(x_i) + \varepsilon_i \quad \text{for } i = 1, \dots, m, \quad \text{where} \quad A(f) = g. \tag{1.2}$$

In the case of random observations, the literature is much scarcer than for the classical setup. For linear mappings $A\colon \mathcal{H}_1 \to \mathcal{H}_2$, this was analyzed in [3, 7], under the assumption that the norm in $\mathscr{L}^2(X, \nu; Y)$ is accessible to the user (at least up to a multiplicative constant); again, this is an unrealistic assumption if the only information on ν is available through the sample points $(x_1, \dots, x_m)$. In the case of general random design with unknown marginal distribution ν, this was analyzed in [4].

For non-linear mappings A, some structural assumptions on the nature of the non-linearity must be assumed. In the present study we shall use the most classical non-linearity assumption, first assumed in [33], and presented in detail in the monograph [11]. Roughly speaking, the mapping A is assumed to be Fréchet differentiable at the true solution, and the Fréchet derivative obeys some Lipschitz property, see Assumption 5 for the precise requirements. Such non-linear inverse problems occur in many situations, and examples are given in the seminal monograph [11]. Of special importance are problems of parameter identification in partial differential equations, and we mention the monograph [16,

Chapt. 1], and the more recent [34]. In such settings, the model we consider with random sampling and unknown sampling distribution ν ("learning" model) is particularly relevant if we have a good "physical" knowledge of a complex system (the known operator A), but with many unknown parameters, furthermore this system is only observed "in the wild", without the possibility for the user to choose or even know precisely the design distribution. This is typically of interest in fields such as economics, medicine, etc.

A widely used approach to stabilizing the estimation problem (1.1) is the Tikhonov regularization or regularized least-squares algorithm or method of regularization. The estimate of the solution of (1.1) is obtained by minimizing an objective function consisting of an error term measuring the fit to the data plus a smoothness term measuring the complexity of the quantity f. For the non-linear statistical inverse problem (1.2), and hence with $g = A(f)$, the corresponding regularization scheme over the hypothesis space $\mathcal{H}_1$ can be described as

$$f_{\mathbf{z},\lambda} = \underset{f\in\mathcal{D}(A)\subset\mathcal{H}_1}{\operatorname{argmin}} \left\{ \frac{1}{m}\sum_{i=1}^{m} \|A(f)(x_i) - y_i\|_Y^2 + \lambda \left\|f - \bar{f}\right\|_{\mathcal{H}_1}^2 \right\}. \tag{1.3}$$

Here $\bar{f} \in \mathcal{H}_1$ denotes some initial guess of the true solution, which offers the possibility to incorporate *a-priori* information. The regularization parameter λ is positive and controls the trade-off between the error term measuring the fitness of data and the complexity of the solution measured in the norm in $\mathcal{H}_1$.

The objective of this paper is to analyze the theoretical properties of the regularized least-squares estimator $f_{\mathbf{z},\lambda}$, in particular, the asymptotic performance of the regularization scheme is evaluated by the bounds and the rates of convergence of the regularized least-squares estimator $f_{\mathbf{z},\lambda}$. Precisely, we develop a non-asymptotic analysis of Tikhonov regularization (1.3) for the non-linear statistical inverse problems based on the tools that have been developed for the modern mathematical study of reproducing kernel methods. The challenges specific to the studied problem are that the considered model is an inverse problem (rather than a pure prediction problem) and non-linear. The upper rate of convergence for the regularized least-squares estimator $f_{\mathbf{z},\lambda}$ to the true solution is described in the probabilistic sense by exponential tail inequalities. For sample size m and the confidence level $0 < \eta < 1$, we establish the bounds of the form

$$\mathbb{P}_{\mathbf{z}\in Z^m} \left\{ \|f_{\mathbf{z},\lambda} - f\|_{\mathcal{H}_1} \le \varepsilon(m) \log\left(\frac{1}{\eta}\right) \right\} \ge 1 - \eta.$$

Here the function $m \mapsto \varepsilon(m)$ is a positive decreasing function and describes the rate of convergence as $m \to \infty$. The upper rate of convergence is complemented by a minimax lower bound for any learning algorithm for the considered non-linear statistical inverse problem. The lower rate result shows that the error rate attained by the Tikhonov regularization scheme for a suitable choice of the regularization parameter is optimal on a suitable class of probability measures.

Now we review previous results concerning regularization algorithms on different learning schemes which are directly comparable to our results: Caponnetto

TABLE 1
Convergence rates of the regularization schemes on different learning schemes: the parameter r refers to the smoothness of true solution in source conditions and $b > 1$ refers to the rate of eigenvalue decay of the covariance operators.

	Learning problem	Regularization	$\|f_{\mathbf{z},\lambda} - f\|_{\mathcal{H}_1}$	Smoothness	General source condition	Optimal rates
Caponnetto et al. [5]		Tikhonov				√
Rastogi et al. [29]	Direct	General	$m^{-\frac{br}{2br+b+1}}$	$0 \le r \le 1$	√	√
Lin et al. [18]		General			√	√
Blanchard et al. [4]	Linear inverse	General	$m^{-\frac{br}{2br+b+1}}$	$0 \le r \le 1$		√
Our Results	Non-linear inverse	Tikhonov	$m^{-\frac{br}{2br+b+1}}$	$\frac{1}{2} \le r \le 1$	√	√

et al. [5], Rastogi et al. [29], Lin et al. [18] and Blanchard et al. [4]. For convenience, we tried to present the most essential points in a unified way in Table 1.

In this table, the parameter r corresponds to a (Hölder type) smoothness assumption for the unknown true solution, and the parameter $b > 1$ corresponds to the decay rate of the eigenvalues of the covariance operator, both to be introduced below in Assumption 6, and Assumption 7, respectively.

The model (1.2) covers nonparametric regression under random design (which we also call the direct problem, i.e., $A = I$), and the linear statistical inverse problem. Thus, introducing a general non-linear operator A gives a unified approach to different learning problems. In the direct learning setting, Caponnetto et al. [5] established the minimax optimal rates of convergence for Tikhonov regularization under a Hölder source condition. Rastogi et al. [29] generalized these bounds for general regularization and under a general source condition. Lin et al. [18] obtained the error estimates in interpolation norms for general regularization in Hilbert space which particularly gives the optimal rates for general regularization in the reproducing kernel Hilbert space (RKHS). These results cover the case when the minimizer of the expected risk does not belong to the RKHS. Blanchard et al. [4] considered general regularization methods for the linear statistical inverse problem. They generalized the convergence analysis of the direct learning scheme to the inverse learning setting and achieved the minimax optimal rates of convergence for general regularization under a Hölder source condition. They considered that the image of the operator A is a reproducing kernel Hilbert space which can be seen as a special case of our general assumption that $Im(A)$ is contained in a reproducing kernel Hilbert space in the linear setting (when A is the linear operator). Here, we consider Tikhonov regularization for the non-linear statistical inverse problem. We obtain minimax optimal rates of convergence under a general source condition. The assumptions

on the non-linear operator A (see Assumption 5, and the condition (4.5), below) allow us to estimate the error bounds for the source condition under some additional constraint, which for Hölder source condition ($\phi(t) = t^r$) corresponds to the range $\frac{1}{2} \leq r \leq 1$.

In this nonlinear setup, other works include the milestone work [26] which considers asymptotic analysis for the generalized Tikhonov regularization for (1.2) using the linearization technique. The reference [3] considers a 2-step approach, in which, first an approximate g^δ of g in (1.2) is estimated using the observations $\{(x_i, y_i)\}_{i=1}^m$. Then, the regularization schemes are used to stably approximate the quantity f in (1.2). Again, the norm in $\mathscr{L}^2(X, \nu; Y)$ needs to be known.

The references [1] and [15, 40] consider respectively a Gauss-Newton algorithm and the method of regularization (Tikhonov regularization) for certain non-linear inverse problems, in the idealized setting where the noise is a (possibly weak) Gaussian element in a Hilbert space. In this, the noise ξ is a Hilbert space process on $\mathcal{H}_2$, such that the random variable $\xi(g) = \langle \xi, g \rangle$ satisfies $\mathbb{E}(\xi(g)) = 0$, $Var(\xi(g)) < \infty$ for any vector $g \in \mathcal{H}_2$, and $\mathbb{E}(\xi(g_1)\xi(g_2)) = \langle \mathbf{C}g_1, g_2 \rangle$ for all $g_1, g_2 \in \mathcal{H}_2$ and a bounded self-adjoint nonnegative operator $\mathbf{C}$. The white noise setting is when $\mathbf{C} = I$ (in that case $\xi \notin \mathcal{H}_2$ but $\xi(g) = \langle \xi, g \rangle$ holds in a weak sense), otherwise the setting is dubbed colored noise. This type of noise can cover random design sampling effects, but the output space of the operator is then taken to be $\mathscr{L}^2(X, \nu; Y)$ in such an approach; thus there again it is implicitly required that $\mathscr{L}^2(X, \nu; Y)$ is known for the construction of the considered methods. Loubes et al. [21] consider (1.2) under a fixed design and concentrate on the problem of model selection. Finally, the recent work [30] analyzes rates of convergence in a model where observations are of the form $h(Kf)(x)$ perturbed by noise, but only in a white noise model and for specific, uni-variate non-linear link functions h, and linear operator K.

Hence, the statistical inverse problems are considered with (most often Gaussian) white or colored noise in general. The 2-step approaches are well-studied to obtain the approximate solution of the inverse problems. In the present paper we make the following contributions:

- We consider the nonlinear statistical inverse problem with random design and random observation noise. The observation noise can be non-Gaussian (satisfying a Berstein-type moment assumption), and the random design distribution is unknown, which generally precludes approaches based on directly modeling the observation error as a Gaussian noise in a fixed, known Hilbert space.
- To directly approximate the quantity f in (1.2) from the observations, we consider the (1-step) non-linear Tikhonov regularization rather than 2-step approaches for inverse problems. Furthermore, we establish rates of convergence in terms of sample size, as the sample size tends to ∞.

The structure of the paper is as follows. In Section 2, we introduce the basic setup and notation for supervised learning problems in a reproducing kernel Hilbert space framework. In Sections 3 and 4, we discuss the main results of

this paper on consistency and error bounds of the regularized least-squares solution $f_{\mathbf{z},\lambda}$ under certain assumptions on the (unknown) joint probability measure ρ, and the (non-linear) mapping A. We establish minimax rates of convergence over the regularity classes defined through appropriate source conditions by using the concept of the effective dimension. In Section 5, we present a concluding discussion on some further aspects of the results. In the appendix, we establish the concentration inequalities, perturbation results and the proofs of consistency results, upper error bounds, and lower error bounds.

2. Setup and basic definitions

In this section, we discuss the mathematical concepts and definitions used in our analysis. We start with a brief description of the reproducing kernel Hilbert spaces since our approximation schemes will be built in such spaces. The vector-valued reproducing kernel Hilbert spaces are the extension of real-valued reproducing kernel Hilbert spaces, see e.g. [25].

Definition 2.1. *Let X be a non-empty set, $(Y, \langle\cdot,\cdot\rangle_Y)$ be a real separable Hilbert space and $\mathcal{H}$ be a Hilbert space of functions from X to Y. If the 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 $y \in Y$, then $\mathcal{H}$ is called vector-valued reproducing kernel Hilbert space.

For the Banach space $\mathcal{L}(Y)$ of bounded linear operators $Y \to Y$, a function $K : X \times X \to \mathcal{L}(Y)$ is said to be an operator-valued positive semi-definite kernel if for each pair $(x, z) \in X \times X$, $K(x, z)^* = K(z, x)$, and for every finite set of points $\{x_i\}_{i=1}^N \subset X$ and $\{y_i\}_{i=1}^N \subset Y$,

$$\sum_{i,j=1}^{N} \langle y_i, K(x_i, x_j)y_j\rangle_Y \geq 0.$$

For every operator-valued positive semi-definite kernel, $K : X \times X \to \mathcal{L}(Y)$, there exists a unique vector-valued reproducing kernel Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of functions from X to Y satisfying the following conditions:

(i) For all $x \in X$ and $y \in Y$, the function $K_x y = K(\cdot, x)y$, defined by

$$z \in X \mapsto (K_x y)(z) = K(z, x)y \in Y,$$

belongs to $\mathcal{H}$; this allows us to define the linear mapping $K_x : Y \to \mathcal{H} : y \mapsto K_x y$.

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

(iii) For all $f \in \mathcal{H}$, $x \in X$ and $y \in Y$, $\langle f(x), y\rangle_Y = \langle f, K_x y\rangle_{\mathcal{H}}$, in other words $f(x) = K_x^* f$ (reproducing property).

Moreover, there is a one-to-one correspondence between operator-valued positive semi-definite kernels and vector-valued reproducing kernel Hilbert spaces [25]. In special case, when Y is a bounded subset of $\mathbb{R}$, the reproducing kernel Hilbert space is said to be real-valued reproducing kernel Hilbert space. In this case, the operator-valued positive semi-definite kernel becomes the symmetric, positive semi-definite kernel $K : X \times X \to \mathbb{R}$ and each reproducing kernel Hilbert space $\mathcal{H}$ can be described as the completion of the span of the set $\{K_x \in \mathcal{H} : x \in X\}$ for $K_x : X \to \mathbb{R} : t \mapsto K_x(t) = K(x,t)$. Moreover, for every function f in the reproducing kernel Hilbert space $\mathcal{H}$, the reproducing property can be described as $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$.

We assume that the input space X be a Polish space, and the output space $(Y, \langle \cdot, \cdot \rangle_Y)$ be a real separable Hilbert space (both endowed with their Borel σ-algebras). Under these assumptions the conditional distribution $\rho(y|x)$ exists, see for example [37, Section A.3.2].

We specify the abstract framework for the present study. We consider that random observations $\{(x_i, y_i)\}_{i=1}^m$ follow the model $y = A(f)(x) + \varepsilon$ with the centered noise $\varepsilon = y - A(f_\rho)(x)$ (i.e., $\int_Y \varepsilon d\rho(y|x) = 0$). The operator A is assumed to be one-to-one.

Assumption 1 (True solution f_ρ)**.** *Given ρ, there exists $f_\rho \in \operatorname{int}(\mathcal{D}(A)) \subset \mathcal{H}_1$ such that*

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

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 $x \in X$,*

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

This Assumption is usually referred to as a *Bernstein-type assumption*. The assumption implies the bounds of the noise in the second and higher-order moments as follows:

$$\int_Y \|\varepsilon\|_Y^n \, d\rho(y|x) \leq \frac{n!}{2} \Sigma^2 M^{n-2}, \quad \forall n \geq 2.$$

Concerning the Hilbert space $\mathcal{H}_2$, we assume the following throughout the paper.

Assumption 3 (Vector valued reproducing kernel Hilbert space $\mathcal{H}_2$)**.** *We assume $\mathcal{H}_2$ to be a separable 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) For all $x \in X$, $K_x : Y \to \mathcal{H}_2$ is a Hilbert-Schmidt operator, and

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

implying in particular that $\mathcal{H}_2 \subset \mathscr{L}^2(X, \nu; Y)$.

(ii) The real-valued function $\varsigma : X \times X \to \mathbb{R}$*, defined by*

$$\varsigma(x,t) = \langle K_t v, K_x w \rangle_{\mathcal{H}_2},$$

is measurable $\forall v, w \in Y$*.*

Note that the separability assumption on $\mathcal{H}_2$ is in particular satisfied if the kernel K is continuous, see [37, Lemma 4.33], but this is not necessary.

Example 2.2 (Sobolev space)**.** *Certain Sobolev spaces satisfy the above assumptions, and the kernel K is completely explicit.*

Let $W^{k,2}(\mathbb{R}^d)$ be the Sobolev space of differential order k (based on the space $\mathscr{L}^2(\mathbb{R}^d, \nu; \mathbb{R})$), for the integer $k > \frac{d}{2}$, which is defined as the completion of $C_c^\infty(\mathbb{R}^d)$ with respect to the norm given by:

$$\|f\|_{\mathcal{H}}^2 = \|f\|_{W^{k,2}(\mathbb{R}^d)}^2 = \sum_{\nu=0}^{k} \sum_{\alpha \in \mathbb{Z}_+^d, |\alpha| \le \nu} \frac{\nu!}{\alpha!} \binom{k}{\nu} \int_{\mathbb{R}^d} \left| \frac{\partial^\nu f(x)}{\partial x^\nu} \right|^2 dx.$$

The Sobolev space $W^{k,2}(\mathbb{R}^d)$ is a reproducing kernel Hilbert space with the reproducing kernel K, given by (see [32, Sec. 1.3.5])

$$K(x,y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\exp(i\langle x-y, \xi \rangle)}{(1 + \|\xi\|^2)^k} d\xi, \qquad x, y \in \mathbb{R}^d,$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$.

It satisfies Assumption 3 with $\kappa^2 := (2\pi)^{-d} \int_{\mathbb{R}^d} (1 + \|\xi\|^2)^{-k} d\xi < \infty$.

Note that in the case of real-valued functions ($Y \subset \mathbb{R}$) we get $K_x \in \mathcal{H}_2$ and Assumption 3 simplifies to the condition that the kernel is measurable and $\kappa^2 := \sup_{x \in X} \|K_x\|_{\mathcal{H}_2}^2 = \sup_{x \in X} K(x,x) < \infty$.

The operator I_K denotes the canonical injection map $\mathcal{H}_2 \to \mathscr{L}^2(X, \nu; Y)$, that

$$\|I_K g\|_{\mathscr{L}^2(X,\nu;Y)}^2 = \int_X \|g(x)\|_Y^2 \, d\nu(x) = \int_X \|K_x^* g\|_Y^2 \, d\nu(x) \le \kappa^2 \|g\|_{\mathcal{H}_2}^2 .$$

We denote $L_K := I_K^* I_K \colon \mathcal{H}_2 \to \mathcal{H}_2$ the corresponding covariance operator.

3. Consistency

We establish consistency in expectation and almost surely of the Tikhonov regularization in the sense that $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \to 0$ as $|\mathbf{z}| = m \to \infty$. For this, we need weak assumptions on the operator.

Assumption 4 (Lipschitz continuity)**.** *We suppose that $\mathcal{D}(A)$ is weakly closed with the nonempty interior and $A : \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_2$ is Lipschitz continuous, one-to-one.*

The inequality $\|I_K g\|_{\mathscr{L}^2(X,\nu;Y)} \leq \kappa \|g\|_{\mathcal{H}_2}$ for $g \in \mathcal{H}_2$ and the continuity of the operator $A : \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_2$ implies that $I_K A : \mathcal{D}(A) \to \mathcal{H}_2 \hookrightarrow \mathscr{L}^2(X, \nu; Y)$ is also continuous. Since $\mathcal{D}(A)$ is weakly closed, the mapping $I_K A$ is weakly sequentially closed[1]. For the continuous and weakly sequentially closed operator A, there exists a global minimizer of the functional in (1.3). But it is not necessarily unique since A is non-linear (see [34, Section 4.1.1]).

The proofs of Theorems 3.1, 3.4 will be given in Appendix B.

Theorem 3.1. *Suppose that Assumptions 1, 3, 4 hold true and*

$$\sigma_\rho^2 := \int_Z \|y - A(f_\rho)(x)\|_Y^2 \, d\rho(x, y) < \infty.$$

Assume additionally that I_K is injective. Let $f_{\mathbf{z},\lambda}$ denote a (not necessarily unique) solution to the minimization problem (1.3) and assume that the regularization parameter $\lambda(m) > 0$ is chosen such that

$$\lambda \to 0, \quad \frac{1}{\lambda\sqrt{m}} \to 0 \text{ as } m \to \infty. \tag{3.1}$$

Then we have that

$$\mathbb{E}_{\mathbf{z}} \left(\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}^2 \right) \to 0 \text{ as } |\mathbf{z}| = m \to \infty. \tag{3.2}$$

Remark 3.2. *As can be seen from the proof, the existence of arbitrary moments, as required in Assumption 2 is not needed. Instead, only the existence of second moments is used, as seen from the introduction of σ_ρ.*

Remark 3.3. *A sufficient condition to ensure that I_K is injective is that the marginal ρ_X has full support on the Polish space X (this is also a necessary condition) and that the kernel K is continuous (see [37, Exercise 4.6]).*

The previous result can be strengthened as follows.

Theorem 3.4. *Suppose that Assumptions 1–4 hold true, and assume additionally that I_K is injective. Let $f_{\mathbf{z},\lambda}$ denote a (not necessarily unique) solution to the minimization problem (1.3) and assume that the regularization parameter $\lambda(m) > 0$ is chosen such that*

$$\lambda \to 0, \quad \frac{\log m}{\lambda\sqrt{m}} \to 0 \text{ as } m \to \infty. \tag{3.3}$$

Then we have that

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \to 0 \text{ almost surely as } m \to \infty. \tag{3.4}$$

[1] i.e., if a sequence $(f_m)_{m\in\mathbb{N}} \subset \mathcal{D}(A)$ converges weakly to some $f \in \mathcal{H}_1$ and if the sequence $(A(f_m))_{m\in\mathbb{N}} \subset \mathscr{L}^2(X, \nu; Y)$ converges weakly to some $g \in \mathscr{L}^2(X, \nu; Y)$, then $f \in \mathcal{D}(A)$ and $A(f) = g$.

4. Convergence rates

To derive rates of convergence additional assumptions are made on operator A. We need to introduce the corresponding notion of smoothness of the true solution f_ρ from Assumption 1. We discuss the class of probability measures defined through the appropriate source condition which describes the smoothness of the true solution.

Following the work of Engl et al. [11, Chapt. 10] on 'classical' non-linear inverse problems, we consider the following assumption:

Assumption 5 (Non-linearity of the operator)**.** *We assume that $\mathcal{D}(A)$ is convex with nonempty interior, $A : \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_2 \hookrightarrow \mathscr{L}^2(X, \nu; Y)$ is weakly sequentially closed and one-to-one. Furthermore, we assume that*

(i) A is Fréchet differentiable,

(ii) the Fréchet derivative $A'(f)$ of A at f is bounded in a ball $\mathcal{B}_d(f_\rho)$ of radius $d := 4\left\|f_\rho - \bar{f}\right\|_{\mathcal{H}_1}$, i.e., there exists $L < \infty$ such that

$$\|A'(f)\|_{\mathcal{H}_1 \to \mathcal{H}_2} \leq L \qquad \forall\ f \in \mathcal{B}_d(f_\rho) \cap \mathcal{D}(A) \subset \mathcal{H}_1,$$

and

(iii) there exists $\gamma \geq 0$ such that for all $f \in \mathcal{B}_d(f_\rho) \cap \mathcal{D}(A) \subset \mathcal{H}_1$ we have,

$$\|I_K \{A(f) - A(f_\rho) - A'(f_\rho)(f - f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)} \leq \frac{\gamma}{2} \|f - f_\rho\|^2_{\mathcal{H}_1}.$$

Remark 4.1. *The condition (iii) also holds true under the stronger assumption that A' is Lipschitz for the operator norm (see [11, Chapt. 10]), i.e.,*

$$\|I_K \{A'(f) - A'(f_\rho)\}\|_{\mathcal{H}_1 \to \mathscr{L}^2(X,\nu;Y)} \leq \gamma \|f - f_\rho\|_{\mathcal{H}_1}.$$

A sufficient condition for weak sequential closedness is that $\mathcal{D}(A)$ is weakly closed (e.g. closed and convex) and A is weakly continuous. Note that under the Fréchet differentiability of $A : \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_2$ (Assumption 5 (ii)), the operator A is Lipschitz continuous in a ball $\mathcal{B}_d(f_\rho)$ with Lipschitz constant L.

The nonlinearity assumption imposed on the operator A is standard, and its applicability for nonlinear illposed problems has been verified for several examples, see e.g. [17, Chapt. 4], [14] and also [31, Chapt. 7].

To illustrate the general setting, we also consider the following examples for nonlinear operators on Sobolev spaces as in Example 2.2. We shall assume that $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$ is the Sobolev space $W^{k,2}(\mathbb{R})$.

Example 4.2. *For $\mathcal{H} = W^{k,2}(\mathbb{R})$ described in Example 2.2, we consider the non-linear operator $A : \mathcal{H} \to \mathcal{H}$ given by:*

$$[A(f)](x) = \int_{\mathbb{R}^d} \vartheta(x, s)(f(s))^2 d\mu(s), \qquad x \in \mathbb{R}^d,\ f \in \mathcal{D}(A) \subset \mathcal{H},$$

where $\vartheta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is k-times differentiable. It can be checked that $A(f) \in \mathcal{H}$, with

$$\|A(f)\|_{\mathcal{H}} \leq C_k(\vartheta) \|f\|^2_\infty \leq \kappa^2 C_k(\vartheta) \|f\|^2_{\mathcal{H}},$$

where κ is as in Example 2.2, and

$$C_k(\vartheta) := \left\| \int_{\mathbb{R}^d} \vartheta(\cdot, s) d\mu(s) \right\|_{\mathcal{H}}$$

(assumed to be finite).

The Fréchet derivative of A at f is given by

$$[A'(f)g](x) = 2 \int_{\mathbb{R}^n} \vartheta(x, s) f(s) g(s) d\mu(s).$$

Then we have

$$\|A'(f_\rho)g\|_{\mathcal{H}} \leq 2C_k(\vartheta) \|f_\rho\|_\infty \|g\|_\infty \leq 2\kappa^2 C_k(\vartheta) \|f_\rho\|_{\mathcal{H}} \|g\|_{\mathcal{H}},$$

and

$$\begin{aligned} &\|I_K \{A'(f)g - A'(f_\rho)g\}\|_{\mathscr{L}^2(\mathbb{R}^d, \nu; \mathbb{R})} \\ \leq &\|A'(f)g - A'(f_\rho)g\|_\infty \leq \kappa \|A'(f)g - A'(f_\rho)g\|_{\mathcal{H}} \\ \leq &2\kappa^3 C_k(\vartheta) \|f - f_\rho\|_{\mathcal{H}} \|g\|_{\mathcal{H}}, \end{aligned}$$

so that Assumption 5 is satisfied.

Example 4.3. *Assume that Ω is of class $\mathcal{C}^1$, a nonempty, bounded, open subset of $\mathbb{R}^d$. For $\mathcal{H} = W^{k,2}(\mathbb{R})$, we consider the non-linear operator $A : \mathcal{H} \to \mathcal{H}$ of the form:*

$$A(f)(x) = G(x, f(x)), \tag{4.1}$$

where $G : \overline{\Omega} \times \mathbb{R} \to \mathbb{R}$, $G \in \mathcal{C}^{k+1}(\overline{\Omega} \times \mathbb{R})$. Such non-linear mappings are often called superposition operators.

From [38, Theorem 3.1], we observe that the operator A is continuous and from Theorem 4.1 ibid., the Fréchet derivative of A at f is given by

$$[A'(f)g](x) = g(x) D_y G(x, f(x)),$$

where D_y denotes the derivative of G with respect to the second coordinate. Further, the Fréchet derivative is Lipschitz continuous, so that Assumption 5 is satisfied.

Under the above non-linearity assumption on operator A we now introduce the related operators, which will turn out to be useful in the analysis of regularization schemes.

We recall that I_K denotes the canonical injection map $\mathcal{H}_2 \to \mathscr{L}^2(X, \nu; Y)$, and $A'(f_\rho)$ the Fréchet derivative of A at f_ρ. We define the operator $B : \mathcal{H}_1 \to \mathscr{L}^2(X, \nu; Y)$ given by

$$f \mapsto Bf := (I_K \circ (A'(f_\rho)))(f) = I_K(A'(f_\rho)f).$$

We denote $T := B^*B \colon \mathcal{H}_1 \to \mathcal{H}_1$ the corresponding covariance operator. The operators L_K from Section 2, and T are positive, self-adjoint and compact operators, even trace-class operators.

Observe that the operator B depends on I_K and f_ρ, thus on the joint probability measure ρ itself. It is bounded and satisfies $\|B\|_{\mathcal{H}_1 \to \mathscr{L}^2(X,\nu;Y)} \leq \kappa L$.

The consistency results as established in Section 3, yield convergence of the minimizers $f_{\mathbf{z},\lambda}$, as $|\mathbf{z}| = m$ tends to infinity, and the parameter λ is chosen appropriately. However, the rates of convergence may be arbitrarily slow. This phenomenon is known as the no free lunch theorem [10]. Therefore, we need some prior assumptions on the probability measure ρ to achieve uniform rates of convergence for learning algorithms.

Definition 4.4 (Index function)**.** *A function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ is said to be an index function if it is continuous and strictly increasing with $\phi(0) = 0$.*

Assumption 6 (General source condition)**.** *The true solution f_ρ belongs to the class $\Omega(\rho, \phi, R)$ with*

$$\Omega(\rho, \phi, R) := \left\{ f \in \mathcal{H}_1 : f - \bar{f} = \phi(T)v \text{ and } \|v\|_{\mathcal{H}_1} \leq R \right\},$$

where ϕ is an index function defined on the interval $[0, \kappa^2 L^2]$.

The general source condition $f_\rho \in \Omega(\rho, \phi, R)$, by allowing for the index functions ϕ, cover a wide range of source conditions, such as Hölder source condition $\phi(t) = t^r$ with $r \geq 0$, and logarithmic-type source condition $\phi(t) = t^p \log^{-\nu}\left(\frac{1}{t}\right)$ with $p \in \mathbb{N}$, $\nu \in [0, 1]$. The source sets $\Omega(\rho, \phi, R)$ are precompact sets in $\mathcal{H}_1$, since the operator T is compact. Observe that in contrast with the linear case, in the equation $f_\rho - \bar{f} = \phi(T)v$ from Assumption 6, the true solution f_ρ appears on both sides, since the operator T itself depends on it (through $A'(f_\rho)$). This condition is more easily interpreted as a condition on the "initial guess" $\bar{f}$, so that the initial error $(\bar{f} - f_\rho)$ should satisfy a source condition with respect to the operator linearized at the true solution. Assumption 6 is usually referred to as a general source condition, see e.g. [24], which is a measure of regularity of the true solution f_ρ. This is inspired, on the one hand, by the approach considered in previous works on statistical learning using kernels, and, on the other hand, by the "classical" literature on non-linear inverse problems. The true solution f_ρ is represented in terms of the marginal probability distribution ν over the input space X, and of the linearized operator at the true solution, respectively. Both aspects enter into Assumption 6.

Assumption 7 (Eigenvalue decay condition)**.** *The eigenvalues $(t_n)_{n\in\mathbb{N}}$ of the covariance operator L_K follow a polynomial decay, i.e., for fixed positive constants β and $b > 1$,*

$$t_n \leq \beta n^{-b} \quad \forall n \in \mathbb{N}.$$

Now under Assumption 5 (ii) using the relation for singular values $s_j(UV) \leq \|U\|\, s_j(V)$ for $j \in \mathbb{N}$ (see Chapter 11 [27]) we obtain,

$$s_j(T) \leq \|A'(f_\rho)\|^2_{\mathcal{H}_1 \to \mathcal{H}_2}\, s_j(L_K) \leq L^2 s_j(L_K).$$

Hence the polynomial decay condition on eigenvalues of the operator L_K implies that the eigenvalues of T also follows the polynomial decay.

Following the concept of Bauer et al. [2], and Blanchard et al. [4], we consider the classes of probability measures $\mathcal{P}_\phi$ and $\mathcal{P}_{\phi,b}$.

Definition 4.5 ($\mathcal{P}_\phi$ and $\mathcal{P}_{\phi,b}$)**.** *The class of probability measures $\mathcal{P}_\phi$ consists of ρ such that*

(i) The conditional distribution $\rho(y|x)$ satisfies the Assumptions 1, 2.
(ii) The true solution f_ρ, that corresponds to ρ, cf. Assumption 1, obeys the smoothness assumption 6.

Further, we define the set of probability measures $\mathcal{P}_{\phi,b} \subset \mathcal{P}_\phi$ consisting of those $\rho \in \mathcal{P}_\phi$ such that the sampling distribution ρ_X satisfies the eigenvalue decay assumption 7.

Both the classes $\mathcal{P}_\phi$, $\mathcal{P}_{\phi,b}$ depend on the observation noise distribution (reflected in the parameters $M > 0$, $\Sigma > 0$) and the smoothness properties of the true solution f_ρ (reflected in the parameters $R > 0$, $\phi > 0$). The class $\mathcal{P}_{\phi,b}$ also depends on the properties of the covariance operator L_K (reflected in terms of the eigenvalue decay parameter b).

We achieve optimal minimax rates of convergence using the concept of *effective dimension* of the operator L_K. For the trace class operator L_K, the effective dimension is defined as

$$\mathcal{N}(\lambda) = \mathcal{N}_{L_K}(\lambda) := \operatorname{tr}\left((L_K + \lambda I)^{-1} L_K\right), \quad \text{for } \lambda > 0.$$

For the infinite-dimensional operator L_K, the effective dimension is a continuously decreasing function of λ from ∞ to 0. For further discussion on the effective dimension, we refer to the literature [19, 22].

Under Assumptions 3, 5 (ii), the effective dimension $\mathcal{N}(\lambda)$ can trivially be estimated as follows,

$$\mathcal{N}(\lambda) \leq \left\|(L_K + \lambda I)^{-1}\right\|_{\mathcal{L}(\mathcal{H}_2)} \operatorname{tr}(L_K) \leq \frac{\kappa^2}{\lambda}, \quad \lambda > 0. \tag{4.2}$$

However, we know from [5, Prop. 3] that, under Assumption 7, we have the improved bound

$$\mathcal{N}(\lambda) \leq C_{\beta,b} \lambda^{-1/b}, \tag{4.3}$$

where $C_{\beta,b}$ is a positive constant depends on the parameters β and b.

4.1. Upper rates of convergence

In Theorems 4.6–4.7, we present the upper error bounds for the regularized least-squares solution $f_{\mathbf{z},\lambda}$ over the class of probability measures $\mathcal{P}_\phi$. We establish the error bounds for both the direct learning setting in the sense of the $\mathscr{L}^2(X,\nu;Y)$-norm reconstruction error $\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)}$ and the inverse problem setting in the sense of the $\mathcal{H}_1$-norm reconstruction error $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}$. Since the explicit expression of $f_{\mathbf{z},\lambda}$ is not known, we use the definition (1.3) of the regularized least-squares solution $f_{\mathbf{z},\lambda}$ to derive the error bounds. We use the

linearization techniques for the operator A in the neighborhood of the true solution f_ρ under the (Fréchet) differentiability of A. We estimate the error bounds for the regularized least-squares estimator by measuring the complexity of the true solution f_ρ and the effect of random sampling. The rates of convergence are governed by the noise condition (Assumption 2), the general source condition (Assumption 6) and the ill-posedness of the problem, as measured by an assumed power decay (Assumption 7) of the eigenvalues of T with exponent $b > 1$. The effect of random sampling and the complexity of f_ρ are measured through Assumption 2 and Assumption 6 in Proposition A.3 and Proposition C.1, respectively. In addition to this, we briefly discuss two additional assumptions considered in the analysis.

The error bound discussed in the following theorem holds non-asymptotically, but this holds with sufficiently small regularization parameter λ and sufficiently large sample size m. Given the parameters κ, L, M, Σ, d (from Assumptions 2–3, 5), for fixed η and λ, we can choose sufficiently large sample size m such that

$$8\kappa^2 \max\left(1, \frac{L(M+\Sigma)}{\kappa d}\right) \log\left(\frac{4}{\eta}\right) \leq \sqrt{m}\lambda. \tag{4.4}$$

The condition (4.4) says that as the regularization parameter λ decreases, the sample size must increase. This condition will be automatically satisfied under the parameter choice considered later in Theorem 4.8.

Under the source condition $f_\rho - \bar{f} = \phi(T)v$ for $\phi(t) = \sqrt{t}\psi(t)$, we have that $f_\rho - \bar{f} = T^{1/2}\psi(T)v = T^{1/2}w$ for $\psi(T)v = w$. We assume that

$$2\gamma \|w\|_{\mathcal{H}_1} < 1. \tag{4.5}$$

The additional assumption (4.5) is a "smallness" condition that imposes a constraint between $\|w\|_{\mathcal{H}_1}$ and the non-linearity as measured by the parameter γ in Assumption 5 (iii). This condition ensures that the initial guess is close enough to the true solution and the residual error on linearizing the nonlinear operator A at the true solution is small enough in order to achieve the rates of convergence. This fact will be clear from the proofs of Theorem 4.6 and Theorem 4.7. For the latter norm to be finite for any function satisfying the source condition $f_\rho \in \Omega(\rho, \phi, R)$, it requires that $\phi(t)/\sqrt{t}$ remains bounded near 0, in particular, if $\phi(t) = t^r$, that $r \geq \frac{1}{2}$.

The proofs of Theorems 4.6–4.8 will be given in Appendix C.

Theorem 4.6. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_\phi$ *where* $\phi(t) = \sqrt{t}$. *Suppose Assumptions 1–3, 5–6 and the conditions* (4.4), (4.5) *hold true. Then, for all* $0 < \eta < 1$, *for the regularized least-squares estimator* $f_{\mathbf{z},\lambda}$ *(not necessarily unique) in (1.3) with the confidence* $1-\eta$ *the following upper bound holds:*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C_1 \left\{ R\sqrt{\lambda} + \frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2 \mathcal{N}(\lambda)}{m\lambda}} \right\} \log\left(\frac{4}{\eta}\right)$$

and

$$\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)} \leq C_2\sqrt{\lambda}\left\{R\sqrt{\lambda} + \frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m\lambda}}\right\}\log\left(\frac{4}{\eta}\right),$$

where C_1 and C_2 depend on the parameters γ, L, R.

In the above theorem we discussed the error bounds for the Hölder source condition (Assumption 6) with $\phi(t) = \sqrt{t}$. In the following theorem, we discuss the error bound for the general source condition with the suitable assumptions on the function ϕ.

Theorem 4.7. *Let $\mathbf{z}$ be i.i.d. samples drawn according to the probability measure $\rho \in \mathcal{P}_\phi$ where $\phi(t) = \sqrt{t}\psi(t)$ is an index function satisfying the conditions that $\psi(t)$ and $\sqrt{t}/\psi(t)$ are nondecreasing functions. Suppose Assumptions 1–3, 5–6 and the conditions (4.4), (4.5) hold true. Then, for all $0 < \eta < 1$, for the regularized least-squares estimator $f_{\mathbf{z},\lambda}$ (not necessarily unique) in (1.3) with the confidence $1 - \eta$ the following upper bound holds:*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C\left\{R\phi(\lambda) + \frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m\lambda}}\right\}\log\left(\frac{4}{\eta}\right),$$

where C depends on the parameters γ, L, $\|w\|_{\mathcal{H}_1}$.

Note that error bounds for $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}$ in both Theorem 4.6 and Theorem 4.7 are the same up to a constant factor which depends on the parameters γ, L, $\|w\|_{\mathcal{H}_1}$.

In Theorems 4.6–4.7, the error estimates reveal the interesting fact that the error terms consist of increasing and decreasing functions of λ which led to propose a choice of regularization parameter by balancing the error terms. We derive the rates of convergence for the regularized least-squares estimator based on a data-independent (a-priori) parameter choice of λ for the classes of probability measures $\mathcal{P}_\phi$ and $\mathcal{P}_{\phi,b}$. The effective dimension plays a crucial role in the error analysis of the regularized least-squares learning algorithm. In Theorem 4.8, we derive the rate of convergence for the regularized least-squares solution $f_{\mathbf{z},\lambda}$ under the general source condition $f_\rho \in \Omega(\rho, \phi, R)$ for the parameter choice rule for λ based on the index function ϕ and the sample size m. For the class of probability measures $\mathcal{P}_{\phi,b}$, the polynomial decay condition (Assumption 7) on the spectrum of the operator T also enters into the picture and the parameter b enters in the parameter choice by the estimate (4.3) of the effective dimension. For this class, we derive the minimax optimal rate of convergence in terms of the index function ϕ, the sample size m, and the parameter b.

Theorem 4.8. *Under the same assumptions of Theorem 4.7, the convergence of the regularized least-squares estimator $f_{\mathbf{z},\lambda}$ in (1.3) to the true solution f_ρ can be described as:*

(i) For the class of probability measures $\mathcal{P}_\phi$ with the parameter choice $\lambda =$

$\Theta^{-1}\left(m^{-1/2}\right)$ *where* $\Theta(t) = t\phi(t)$, *we have*

$$\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C'\phi\left(\Theta^{-1}\left(m^{-1/2}\right)\right)\log\left(\frac{4}{\eta}\right)\right\} \geq 1-\eta,$$

where C' *depends on the parameters* γ, L, $\|w\|_{\mathcal{H}_1}$, R, κ, M, Σ *and*

$$\lim_{\tau\to\infty}\limsup_{m\to\infty}\sup_{\rho\in\mathcal{P}_\phi}\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} > \tau\phi\left(\Theta^{-1}\left(m^{-1/2}\right)\right)\right\} = 0.$$

(ii) For the class of probability measures $\mathcal{P}_{\phi,b}$ *under Assumption 7 and the parameter choice* $\lambda = \Psi^{-1}\left(m^{-1/2}\right)$ *where* $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, *we have*

$$\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C''\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)\log\left(\frac{4}{\eta}\right)\right\} \geq 1-\eta,$$

where C'' *depends on the parameters* γ, L, $\|w\|_{\mathcal{H}_1}$, R, κ, M, Σ, b, β *and*

$$\lim_{\tau\to\infty}\limsup_{m\to\infty}\sup_{\rho\in\mathcal{P}_{\phi,b}}\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} > \tau\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)\right\} = 0.$$

Notice that the rates given for the class P_ϕ are worse than the one for the (smaller) class $P_{\phi,b}$, which is easily seen from the fact that $t^{1/2+1/(2b)} \geq t$ for $b > 1$, and hence $\Psi(t) \geq \Theta(t)$ for $t \in [0,1]$.

We obtain the following corollary as a consequence of Theorem 4.8.

Corollary 4.9. *Under the same assumptions of Theorem 4.7 with the Hölder's source condition* $f_\rho \in \Omega(\rho,\phi,R)$, $\phi(t) = t^r$, *the convergence of the regularized least-squares estimator* $f_{\mathbf{z},\lambda}$ *in (1.3) to the true solution* f_ρ *can be described as:*

(i) For the class of probability measures $\mathcal{P}_\phi$ *with the parameter choice* $\lambda = m^{-\frac{1}{2r+2}}$, *for all* $0 < \eta < 1$, *we have with the confidence* $1-\eta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C' m^{-\frac{r}{2r+2}}\log\left(\frac{4}{\eta}\right) \quad \text{for } \frac{1}{2} \leq r \leq 1.$$

(ii) For the class of probability measures $\mathcal{P}_{\phi,b}$ *with the parameter choice* $\lambda = m^{-\frac{b}{2br+b+1}}$, *for all* $0 < \eta < 1$, *we have with the confidence* $1-\eta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C'' m^{-\frac{br}{2br+b+1}}\log\left(\frac{4}{\eta}\right) \quad \text{for } \frac{1}{2} \leq r \leq 1.$$

We obtain the following corollary as a consequence of Theorem 4.6.

Corollary 4.10. *Under the same assumptions of Theorem 4.6 with the Hölder's source condition* $f_\rho \in \Omega(\rho,\phi,R)$, $\phi(t) = t^{1/2}$, *the convergence of the regularized least-squares estimator* $f_{\mathbf{z},\lambda}$ *in (1.3) to the true solution* f_ρ *can be described as:*

(i) For the class of probability measures $\mathcal{P}_\phi$ *with the parameter choice* $\lambda = m^{-\frac{1}{3}}$*, for all* $0 < \eta < 1$*, we have with the confidence* $1-\eta$*,*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C_1' m^{-\frac{1}{6}} \log\left(\frac{4}{\eta}\right)$$

and

$$\|I_K\left\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\right\}\|_{\mathscr{L}^2(X,\nu;Y)} \leq C_2' m^{-\frac{1}{3}} \log\left(\frac{4}{\eta}\right),$$

where C_1' *and* C_2' *depends on the parameters* γ*,* L*,* $\|w\|_{\mathcal{H}_1}$*,* κ*,* M*,* Σ*.*

(ii) For the class of probability measures $\mathcal{P}_{\phi,b}$ *with the parameter choice* $\lambda = m^{-\frac{b}{2b+1}}$*, for all* $0 < \eta < 1$*, we have with the confidence* $1-\eta$*,*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C_1'' m^{-\frac{b}{4b+2}} \log\left(\frac{4}{\eta}\right)$$

and

$$\|I_K\left\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\right\}\|_{\mathscr{L}^2(X,\nu;Y)} \leq C_2'' m^{-\frac{b}{2b+1}} \log\left(\frac{4}{\eta}\right),$$

where C_1'' *and* C_2'' *depends on the parameters* γ*,* L*,* $\|w\|_{\mathcal{H}_1}$*,* κ*,* M*,* Σ*,* b*,* β*.*

Now we compare the error bounds established for direct learning setting in the sense of $\mathscr{L}^2(X,\nu;Y)$-norm reconstruction error $\|I_K\left\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\right\}\|_{\mathscr{L}^2(X,\nu;Y)}$ and the inverse problem setting in the sense of the $\mathcal{H}_1$-norm reconstruction error $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}$. Since under the condition (4.4) from (B.13) we have that $f_{\mathbf{z},\lambda} \in \mathcal{B}_d(f_\rho) \cap \mathcal{D}(A) \subset \mathcal{H}_1$ with confidence $1-\eta/2$, therefore with Assumption 5 linearizing the operator A at f_ρ (i.e., $A(f_{\mathbf{z},\lambda}) = A(f_\rho) + A'(f_\rho)(f_{\mathbf{z},\lambda} - f_\rho) + r(f_{\mathbf{z},\lambda})$) we conclude that

$$\begin{aligned}
&\|I_K\left\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\right\}\|_{\mathscr{L}^2(X,\nu;Y)} \\
&= \|I_K\left\{A'(f_\rho)(f_{\mathbf{z},\lambda} - f_\rho) + r(f_{\mathbf{z},\lambda})\right\}\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\leq \|B(f_{\mathbf{z},\lambda} - f_\rho)\|_{\mathscr{L}^2(X,\nu;Y)} + \|I_K r(f_{\mathbf{z},\lambda})\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\leq \left\|T^{1/2}(f_{\mathbf{z},\lambda} - f_\rho)\right\|_{\mathcal{H}_1} + \frac{\gamma}{2}\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}^2.
\end{aligned} \tag{4.6}$$

Thus bounding the prediction norm $\|I_K\left\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\right\}\|_{\mathscr{L}^2(X,\nu;Y)}$ corresponds to learning bound in which the first norm consists of some target function $T^{1/2}f_\rho$ which has additional smoothness $1/2$, on the other hand, the second term is square of the reconstruction error in $\mathcal{H}_1$-norm, therefore this might result in a higher rate. Indeed, this heuristics is validated from Theorem 4.6 and Corollary 4.10, where we observe that the prediction norm has a faster convergence rate than the reconstruction error in $\mathcal{H}_1$-norm.

The assumptions on the non-linear operator A (see Assumption 5, and the condition (4.5)) allow us to estimate the reconstruction error bounds in $\mathcal{H}_1$-norm for Hölder source condition ($\phi(t) = t^r$) corresponds to the range $\frac{1}{2} \leq r$. It is well-known that Tikhonov regularization has the saturation effect at $r = 1$ (since

it has qualification 1), therefore we cannot improve the rates of convergence beyond $r = 1$. From (4.6) we observe that for the prediction error we have additional smoothness 1/2 in the bound on the right-hand side, therefore we only estimate the prediction error for $r = \frac{1}{2}$. For the higher smoothness ($\frac{1}{2} \leq r$), the rates for the prediction error cannot be improved.

4.2. Lower rates of convergence

In this section, we discuss the lower rates of convergence for non-linear statistical inverse problems over a subclass of the probability measures $\mathcal{P}_{\phi,b}$. The Kullback-Leibler information and Fano inequalities are the main ingredients in the analysis of the estimates for the minimum possible error. The Kullback-Leibler divergence between two probability measures ρ_1 and ρ_2 is defined as

$$\mathcal{K}(\rho_1, \rho_2) := \int_Z \log(g(z)) d\rho_1(z),$$

where g is the density of ρ_1 with respect to ρ_2, that is, $\rho_1(E) = \int_E g(z) d\rho_2(z)$ for all measurable sets E.

To obtain the lower bound, we define a family of probability measures ρ_f parameterized by suitable vectors $f \in \mathcal{D}(A) \subset \mathcal{H}_1$. We assume that Y is finite-dimensional space with a basis $\{v_j\}_{j=1}^d$. Then for each $f \in \mathcal{D}(A) \subset \mathcal{H}_1$, we associate the probability measure on the sample space Z:

$$\rho_f(x, y) := \frac{1}{2dJ} \sum_{j=1}^{d} \left(a_j(x) \delta_{y + dJv_j} + b_j(x) \delta_{y - dJv_j} \right) \nu(x), \tag{4.7}$$

where $a_j(x) = J - \langle A(f), K_x v_j \rangle_{\mathcal{H}_2}$, $b_j(x) = J + \langle A(f), K_x v_j \rangle_{\mathcal{H}_2}$, $J = 4\kappa \|A(f)\|_{\mathcal{H}_2}$ and $\delta_{y-\xi}$ denotes the Dirac measure on Y with unit mass at $y = \xi$.

Following the analysis of Caponnetto et al. [5] and DeVore et al. [9] we establish the lower rates of convergence for the non-linear statistical inverse problems that can be attained by any learning algorithm. The main steps are the following. To obtain the lower rates of convergence for learning algorithms, we generate N_ε-vectors $(f_1, \ldots, f_{N_\varepsilon})$ 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 vectors are separated by constant times ε with respect to the norm in Hilbert space $\mathcal{H}_1$ (Proposition D.2 (i)). Then we construct the probability measures $\rho_i = \rho_{f_i}$ from (4.7), parameterized by f_i's ($1 \leq i \leq N_\varepsilon$) with small Kullback-Leibler divergence to each other (Proposition D.2 (ii)) and are therefore statistically close. Finally, we obtain the lower rates of convergence on applying [9, Lemma 3.3] using the Kullback-Leibler information.

Assumption 8. *For the lower rates of convergence, we assume the following conditions on the non-linear operator A:*

(i) A is Fréchet differentiable.

(ii) At the initial guess $\bar{f}$, we denote

$$\|A'(\bar{f})\|_{\mathcal{H}_1 \to \mathcal{H}_2} =: L.$$

(iii) There exists $\gamma \geq 0$ such that for all $f, \tilde{f} \in \mathcal{D}(A) \subset \mathcal{H}_1$ in a sufficiently large ball around $\bar{f}$ we have,

$$\left\| I_K \left\{ A'(\tilde{f}) - A'(f) \right\} \right\|_{HS} \leq \gamma \|\tilde{f} - f\|_{\mathcal{H}_1}.$$

(iv) The function ϕ is a continuous increasing function with $\phi(0) = 0$ and $\theta(t) = \phi(t^2)$ is Lipschitz continuous with the constant L_θ. For the operators $T = A'(f)^* I_K^* I_K A'(f)$ and $\overline{T} = A'(\bar{f})^* I_K^* I_K A'(\bar{f})$:

$$\phi(T) = R_f \phi(\overline{T}) \text{ and } \|R_f - I\|_{\mathcal{L}(\mathcal{H}_1)} \leq \zeta \left\| f - \bar{f} \right\|_{\mathcal{H}_1},$$

where f belongs to the sufficiently large ball, $R_f : \mathcal{H}_1 \to \mathcal{H}_1$ is a family of bounded linear operators and ζ is a positive constant.

(v) The eigenvalues $(t_n)_{n \in \mathbb{N}}$ of the operator $\overline{T} = A'(\bar{f})^* I_K^* I_K A'(\bar{f})$ follow the polynomial decay: For fixed positive constants α, β, and $b > 1$,

$$\alpha n^{-b} \leq t_n \leq \beta n^{-b} \quad \forall n \in \mathbb{N}.$$

In contrast to upper rates of convergence for Tikhonov regularization, we require the additional assumption (iv) on A for the lower rates. This condition is the generalization of the following condition used in [13] for the Landweber iteration:

$$A'(f) = R_f A'(\bar{f}) \text{ and } \|R_f - I\|_{\mathcal{L}(\mathcal{H}_1)} \leq \zeta \left\| f - \bar{f} \right\|_{\mathcal{H}_1}, \ f \in B_d(\bar{f}),$$

which implies that the Fréchet derivative of A is Lipschitz continuous in $B_d(\bar{f})$. Note that in the linear case $R_f \equiv I$; therefore, Assumption 8 (iv) may be interpreted as a further restriction on the "non-linearity" of A.

The proof of the following theorem will be given in Appendix D.

Theorem 4.11. *Let $\mathbf{z}$ be i.i.d. samples drawn according to the probability measure $\rho \in \mathcal{P}_{\phi,b}$ under the hypothesis $\dim(Y) = d < \infty$. Then under Assumptions 3, 8 for $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, the estimator $f_{\mathbf{z}}^l$ corresponding to any learning algorithm l ($\mathbf{z} \to f_{\mathbf{z}}^l \in \mathcal{H}_1$) converges with the following lower rate:*

$$\lim_{\tau \to 0} \liminf_{m \to \infty} \inf_{l \in \mathcal{A}} \sup_{\rho \in \mathcal{P}_{\phi,b}} \mathbb{P}_{\mathbf{z} \in Z^m} \left\{ \|f_{\mathbf{z}}^l - f_\rho\|_{\mathcal{H}_1} > \tau \phi \left(\Psi^{-1} \left(m^{-1/2} \right) \right) \right\} = 1,$$

where $\mathcal{A}$ denotes the set of all learning algorithms $l : \mathbf{z} \to f_{\mathbf{z}}^l$.

We obtain the following corollary as a consequence of Theorem 4.11.

Corollary 4.12. *Under the same assumptions of Theorem 4.11, for any learning algorithm with Hölder's source condition $f_\rho \in \Omega(\rho, \phi, R)$, $\phi(t) = t^r$, the lower rates of convergence can be described as*

$$\lim_{\tau \to 0} \liminf_{m \to \infty} \inf_{l \in \mathcal{A}} \sup_{\rho \in \mathcal{P}_{\phi,b}} \mathbb{P}_{\mathbf{z} \in Z^m} \left\{ \|f_{\mathbf{z}}^l - f_\rho\|_{\mathcal{H}_1} > \tau m^{-\frac{br}{2br+b+1}} \right\} = 1.$$

The choice of parameter $\lambda(m)$ is said to be optimal if, for this choice of the parameter, the upper rate of convergence coincides with the minimax lower rate. For the class of probability measures $\mathcal{P}_{\phi,b}$ with the parameter choice $\lambda = \Psi^{-1}(m^{-1/2})$, Theorem 4.8 shares the upper rate of convergence with the lower rate of convergence in Theorem 4.11. Therefore the choice of the parameter is optimal.

5. Discussion

Our analysis guarantees the consistency of the Tikhonov regularization algorithm and provides a finite sample bound for non-linear statistical inverse problems in vector-valued setting, therefore the results can be applied to the multi-task learning problem. We also discussed the asymptotic worst-case analysis for any learning algorithm in this setting, showing optimality in the minimax sense on a suitable class of priors. The rates of convergence presented in Section 4 are asymptotic in nature, i.e., all parameters are fixed as $m \to \infty$. This provides a mathematical foundation for nonlinear inverse problems in the statistical learning framework. The considered framework generalizes previously proposed settings for different learning schemes: direct, linear inverse learning problem.

Impact of the effective dimension

The upper rates were represented in terms of the index function ϕ from Assumption 6, and the effective dimension $\mathcal{N}(\lambda)$ of the governing operator L_K. This is seen from the basic probabilistic bound, given in Proposition A.3, and this holds regardless of the fact that $\lambda \to \mathcal{N}(\lambda)$ decays at a polynomial rate. However, the construction for the lower bounds makes use of this constraint. Also, the Corollaries 4.9 and 4.10 can be given a handy representation of the upper bounds under power type decay.

Saturation effect

In Theorem 4.6 we highlighted the upper rates, both for the errors $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}$, and $\|I_K \{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)}$ in the limiting case when smoothness is given through the index function $\phi(t) = \sqrt{t}$; and these differ by a factor $\sqrt{\lambda}$. We emphasize that for higher smoothness $\phi(t) = \sqrt{\psi}(t)$ with an additional index function ψ this cannot be expected to remain valid. This is known from the linear case and is due to the saturation effect of Tikhonov regularization.

Relation to classical regularization theory

Within the present study, the smoothness assumption 6 is based on the composed operator $B = I_K \circ [A'(f_\rho)]$ through $T = B^*B$. This is in contrast to classical regularization theory, when the corresponding operator is $[A'(f_\rho)]^* A'(f_\rho)$.

By assuming an appropriate link condition between the operators $[A'(f_\rho)]^* A'(f_\rho)$ and T, one can transfer the obtained rates results from the present context to the standard ones and we refer to the corresponding calculus established in [23].

Parameter choice

The a-priori parameter choice considered in our analysis depends on the smoothness parameters b, ϕ. In practice, a posteriori parameter choice rule (data-dependent) for the regularization parameter λ such as the discrepancy principle, balancing principle, quasi-optimality principle with theoretical justification is required, so that we can turn our results to data-dependent minimax adaptivity even in the absence of a-priori knowledge of the regularity parameters.

Approximate solution of Tikhonov regularization

For non-linear mappings, the solution of the Tikhonov regularization scheme (1.3) is not explicitly given. Therefore, in practice, we need to find an approximate solution of the considered scheme. Suppose $\mathcal{H}_1$ is a reproducing kernel Hilbert space corresponding the kernel K_1, then the general representer theorem (see, e.g. [35, Theorem 16.1]) holds for the scheme (1.3). By this theorem, the solution of the scheme is given by

$$f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} K(\cdot, x_i)c_i, \quad x_i \in X,\ c_i \in Y.$$

Thus, in this case, by using this representation, the infinite dimensional minimization problem (1.3) reduces to a finite dimensional minimization problem. Hence, we only need to approximate the coefficients $\{c_i\}_{i=1}^m$. To this end, we can apply the gradient descent or stochastic gradient approach to approximate the solution of (1.3), being understood that the theoretical convergence of such schemes remains a challenging point due to the potential existence of local minima.

Appendix A: Notation and probabilistic estimates

Here we introduce some relevant operators.

Definition A.1 (Sampling operator)**.** *For a discrete ordered set* $\mathbf{x} = (x_i)_{i=1}^m$*, the sampling operator* $S_{\mathbf{x}} : \mathcal{H}_2 \to Y^m$ *is defined as*

$$S_{\mathbf{x}}(g) := (g(x_1), \ldots, g(x_m)).$$

We equip the product Hilbert space Y^m with the scalar product$\langle (y_i)_{i=1}^m, (y_i')_{i=1}^m \rangle = \frac{1}{m}\sum_{i=1}^m \langle y_i, y_i' \rangle$, and denote the associated Hilbert norm $\|\mathbf{y}\|_m^2 =$

$\frac{1}{m}\sum_{i=1}^{m}\|y_i\|_Y^2$ for $\mathbf{y}=(y_1,\dots,y_m)$. Then the adjoint $S_{\mathbf{x}}^*: Y^m \to \mathcal{H}_2$ is given by

$$S_{\mathbf{x}}^*\mathbf{c}=\frac{1}{m}\sum_{i=1}^{m}K_{x_i}c_i, \qquad \forall \mathbf{c}=(c_1,\dots,c_m)\in Y^m.$$

Under Assumption 3, the sampling operator is bounded by κ, since

$$\|S_{\mathbf{x}}g\|_m^2=\frac{1}{m}\sum_{i=1}^{m}\|g(x_i)\|_Y^2=\frac{1}{m}\sum_{i=1}^{m}\left\|K_{x_i}^*g\right\|_Y^2\le\kappa^2\|g\|_{\mathcal{H}_2}^2.$$

The sampling versions are the operators $B_{\mathbf{x}}:=S_{\mathbf{x}}\circ(A'(f_\rho))$ and $T_{\mathbf{x}}:=B_{\mathbf{x}}^*B_{\mathbf{x}}$. The operator $T_{\mathbf{x}}$ is positive and self-adjoint. Under Assumptions 3, 5 (ii), the operator $B_{\mathbf{x}}$ is bounded and satisfies $\|B_{\mathbf{x}}\|_{\mathcal{H}_1\to Y^m}\le\kappa L$. We also recall that $L_K=I_K^*I_K$ for the canonical injection map $I_K:\mathcal{H}_2\to\mathscr{L}^2(X,\nu;Y)$ and $T=B^*B$ for $B=I_K\circ[A'(f_\rho)]$. These operators will be used in our analysis.

The following inequality is based on the results of Pinelis and Sakhanenko [28].

Proposition A.2. *Let $\mathcal{H}$ be a real separable Hilbert space and ξ be a random variable on (Ω,ρ) with values in $\mathcal{H}$. If there exist two constant Q and S satisfying*

$$\mathbb{E}_\omega\left\{\|\xi(\omega)-\mathbb{E}_\omega(\xi)\|_{\mathcal{H}}^n\right\}\le\frac{1}{2}n!S^2Q^{n-2}\quad\forall n\ge 2,$$

then for any $0<\eta<1$ and for all $m\in\mathbb{N}$,

$$\mathbb{P}\left\{(\omega_1,\dots,\omega_m)\in\Omega^m:\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(\omega_i)-\mathbb{E}_\omega(\xi)\right\|_{\mathcal{H}}\le 2\left(\frac{Q}{m}+\frac{S}{\sqrt{m}}\right)\log\left(\frac{2}{\eta}\right)\right\}\ge 1-\eta.$$

In the following proposition, we measure the effect of random sampling using Assumption 2. The quantities describe the probabilistic estimates of the perturbation measure due to random sampling. These bounds are standard in learning theory, and can be found in [4, Proposition 5.2, 5.5].

Proposition A.3. *Let $\mathbf{z}$ be i.i.d. random samples with Assumptions 1–3, then for $m\in\mathbb{N}$ and $0<\eta<1$, each of the following estimates holds with the confidence $1-\eta$,*

$$\left\|\frac{1}{m}\sum_{i=1}^{m}(L_K+\lambda I)^{-1/2}K_{x_i}(y_i-A(f_\rho)(x_i))\right\|_{\mathcal{H}_2}\le 2\left(\frac{\kappa M}{m\sqrt{\lambda}}+\sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m}}\right)\log\left(\frac{2}{\eta}\right),$$

$$\left\|\frac{1}{m}\sum_{i=1}^{m}K_{x_i}(y_i-A(f_\rho)(x_i))\right\|_{\mathcal{H}_2}\le 2\left(\frac{\kappa M}{m}+\frac{\kappa\Sigma}{\sqrt{m}}\right)\log\left(\frac{2}{\eta}\right)$$

and

$$\|S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K\|_{\mathcal{L}_2(\mathcal{H}_2)} \leq 2\left(\frac{\kappa^2}{m} + \frac{\kappa^2}{\sqrt{m}}\right)\log\left(\frac{2}{\eta}\right).$$

Proposition A.4. *For $m \in \mathbb{N}$ and $0 < \eta < 1$, under Assumptions 3, the following estimates hold with the confidence $1 - \eta/2$,*

$$\|S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K\|_{\mathcal{L}_2(\mathcal{H}_2)} \leq \frac{\lambda}{2},$$
$$\|(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1}(L_K + \lambda I)\|_{\mathcal{L}_2(\mathcal{H}_2)} \leq 2$$

and

$$\|(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|_{\mathcal{L}_2(\mathcal{H}_2)} \leq \sqrt{2}$$

provided that

$$8\kappa^2 \log(4/\eta) \leq \sqrt{m}\lambda. \tag{A.1}$$

The proofs of the first and second expressions are the content of [29, Theorem 3.1]. The third expression is obtained from the second using the Cordes inequality [4, Prop. 5.7].

Appendix B: Proof of the consistency results

Throughout the analysis we use the following identity in the real Hilbert space $\mathcal{H}$:

$$\|f - h\|_{\mathcal{H}}^2 - \|f - g\|_{\mathcal{H}}^2 = \|g - h\|_{\mathcal{H}}^2 - 2\langle f - g, h - g\rangle_{\mathcal{H}} \qquad f, g, h \in \mathcal{H}.$$

Proof of Theorem 3.1. By the definition of $f_{\mathbf{z},\lambda}$ as a solution to the minimization problem (1.3), we get the inequality

$$\|S_{\mathbf{x}} A(f_{\mathbf{z},\lambda}) - \mathbf{y}\|_m^2 + \lambda\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1}^2 \leq \|S_{\mathbf{x}} A(f_\rho) - \mathbf{y}\|_m^2 + \lambda\|f_\rho - \bar{f}\|_{\mathcal{H}_1}^2.$$

It follows that

$$\begin{aligned} &\|S_{\mathbf{x}}\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_m^2 + \lambda\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1}^2 \\ \leq &2\langle S_{\mathbf{x}}\{A(f_\rho) - A(f_{\mathbf{z},\lambda})\}, S_{\mathbf{x}} A(f_\rho) - \mathbf{y}\rangle_m + \lambda\|f_\rho - \bar{f}\|_{\mathcal{H}_1}^2. \end{aligned} \tag{B.1}$$

Consequently, we get

$$\begin{aligned} &\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)}^2 + \lambda\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1}^2 \\ \leq &2\langle A(f_\rho) - A(f_{\mathbf{z},\lambda}), S_{\mathbf{x}}^*\{S_{\mathbf{x}} A(f_\rho) - \mathbf{y}\}\rangle_{\mathcal{H}_2} + \lambda\|f_\rho - \bar{f}\|_{\mathcal{H}_1}^2 \\ &+ \langle (L_K - S_{\mathbf{x}}^* S_{\mathbf{x}})\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}, A(f_{\mathbf{z},\lambda}) - A(f_\rho)\rangle_{\mathcal{H}_2} \end{aligned} \tag{B.2}$$

and

$$\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1}^2 \leq \frac{2}{\lambda}\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|_{\mathcal{H}_2}\|S_{\mathbf{x}}^*\{S_{\mathbf{x}} A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2} + \|f_\rho - \bar{f}\|_{\mathcal{H}_1}^2.$$

Under the Lipschitz continuity of the operator A (Assumption 4) (i.e., $\|A(f) - A(f_\rho)\|_{\mathcal{H}_2} \leq L \|f - f_\rho\|_{\mathcal{H}_1}$ for $f \in \mathcal{H}_1$) and triangle inequality we get,

$$\begin{aligned}&\|f_{\mathbf{z},\lambda} - \bar{f}\|^2_{\mathcal{H}_1}\\ \leq&\frac{2L}{\lambda}(\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1} + \|\bar{f} - f_\rho\|_{\mathcal{H}_1})\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2} + \|f_\rho - \bar{f}\|^2_{\mathcal{H}_1}\end{aligned}$$

which implies

$$\begin{aligned}&\left(\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1} - \frac{L}{\lambda}\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2}\right)^2\\ \leq&\left(\|f_\rho - \bar{f}\|_{\mathcal{H}_1} + \frac{L}{\lambda}\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2}\right)^2.\end{aligned}$$

Then it gives

$$\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1} \leq \frac{2L}{\lambda}\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2} + \left\|f_\rho - \bar{f}\right\|_{\mathcal{H}_1}. \tag{B.3}$$

Using the triangle inequality $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} - \left\|\bar{f} - f_\rho\right\|_{\mathcal{H}_1} \leq \left\|f_{\mathbf{z},\lambda} - \bar{f}\right\|_{\mathcal{H}_1}$ we get

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq \frac{2L}{\lambda}\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2} + 2\left\|f_\rho - \bar{f}\right\|_{\mathcal{H}_1}.$$

Now squaring both sides and taking expectation with respect to $\mathbf{z}$ we obtain,

$$\mathbb{E}_{\mathbf{z}}\left(\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{\mathcal{H}_1}\right) \leq \frac{8L^2}{\lambda^2}\mathbb{E}_{\mathbf{z}}\left(\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|^2_{\mathcal{H}_2}\right) + 8\left\|f_\rho - \bar{f}\right\|^2_{\mathcal{H}_1}. \tag{B.4}$$

Under Assumptions 1, 3 and $\sigma^2_\rho = \mathbb{E}_z\left(\|y - A(f_\rho)(x)\|^2_Y\right) < \infty$ we have that

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}}\left(\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|^2_{\mathcal{H}_2}\right) \qquad\qquad (B.5)\\ =&\frac{1}{m^2}\mathbb{E}_{\mathbf{z}}\left(\sum_{i,j=1}^m \left\langle K_{x_i}\left\{y_i - A(f_\rho)(x_i)\right\}, K_{x_j}\left\{y_j - A(f_\rho)(x_j)\right\}\right\rangle_{\mathcal{H}_2}\right)\\ =&\frac{1}{m^2}\mathbb{E}_{\mathbf{z}}\left(\sum_{i=1}^m \|K_{x_i}\left\{y_i - A(f_\rho)(x_i)\right\}\|^2_{\mathcal{H}_2}\right)\\ \leq&\frac{\kappa^2}{m^2}\mathbb{E}_{\mathbf{z}}\left(\sum_{i=1}^m \|y_i - A(f_\rho)(x_i)\|^2_Y\right) = \frac{\kappa^2\sigma^2_\rho}{m},\end{aligned}$$

and from [36, Lemma 1] we have,

$$\mathbb{E}_{\mathbf{x}}\left(\|S^*_{\mathbf{x}}S_{\mathbf{x}} - L_K\|^2_{\mathcal{L}(\mathcal{H}_2)}\right) \leq \frac{\kappa^2}{m}. \tag{B.6}$$

Using (B.5) in (B.4) we get,

$$\mathbb{E}_{\mathbf{z}}\left(\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{\mathcal{H}_1}\right) \leq \frac{8\kappa^2\sigma^2_\rho L^2}{\lambda^2 m} + 8\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1} \tag{B.7}$$

from which with the parameter choice rule (3.1) we deduce that

$$\limsup_{m\to\infty} \mathbb{E}_{\mathbf{z}}\left(\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{\mathcal{H}_1}\right) \leq 8\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1}. \tag{B.8}$$

Hence, we observe that $a^2 := \sup_{m\in\mathbb{N}} \mathbb{E}_{\mathbf{z}}\left(\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{\mathcal{H}_1}\right) < \infty$. Now, we show that there exists a subsequence of $(f_{\mathbf{z},\lambda})_{m\in\mathbb{N}}$, denoted by $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$, such that

$$\mathbb{E}_{\mathbf{z}(k)}\left(\langle f_{\mathbf{z}(k),\lambda} - f_\rho, f\rangle_{\mathcal{H}_1}\right) \to \left\langle \tilde{f}, f\right\rangle_{\mathcal{H}_1} \text{ as } k\to\infty \tag{B.9}$$

for some $\tilde{f} \in \mathcal{H}_1$ and for all $f \in \mathcal{H}_1$.

We define the sequence (ξ_m) of random variables $\xi_m = f_{\mathbf{z},\lambda} - f_\rho \in \mathcal{H}_1$. Since $\mathbb{E}_{\mathbf{z}}(\|\xi_m\|^2_{\mathcal{H}_1}) \leq a^2$, Banach-Alaoglu-Bourbaki theorem and the fact that $\mathcal{H}_1$ is a Hilbert space imply that, possible passing to a subsequence $\xi_k = f_{\mathbf{z}(k),\lambda} - f_\rho \in \mathcal{H}_1$, there exists $\xi^* \in \mathcal{H}_1$ such that

$$\lim_{k\to\infty} \mathbb{E}_{\mathbf{z}(k)}\left(\langle \xi_k, \xi\rangle_{\mathcal{H}_1}\right) = \mathbb{E}_{\mathbf{z}(k)}\left(\langle \xi^*, \xi\rangle_{\mathcal{H}_1}\right), \quad \forall \xi \in \mathcal{H}_1.$$

Define $\tilde{f} = \mathbb{E}_{\mathbf{z}(k)}(\xi^*) \in \mathcal{H}_1$ and, given $f \in \mathcal{H}_1$, choose $\xi = f$ be the constant random variable, then, since the scalar product commutes with expectation,

$$\lim_{k\to\infty} \mathbb{E}_{\mathbf{z}(k)}\left(\langle f_{\mathbf{z}(k),\lambda} - f_\rho, f\rangle_{\mathcal{H}_1}\right) = \left\langle \tilde{f}, f\right\rangle_{\mathcal{H}_1}$$

which is (B.9).

From inequality (B.2) we get:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left(\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|^2_{\mathscr{L}^2(X,\nu;Y)}\right) + \lambda\mathbb{E}_{\mathbf{z}}\left(\|f_{\mathbf{z},\lambda} - \bar{f}\|^2_{\mathcal{H}_1}\right)\\
\leq& 2\mathbb{E}_{\mathbf{z}}\left(\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|_{\mathcal{H}_2}\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|_{\mathcal{H}_2}\right)\\
&+ \mathbb{E}_{\mathbf{z}}\left(\|S^*_{\mathbf{x}}S_{\mathbf{x}} - L_K\|_{\mathcal{L}(\mathcal{H}_2)}\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|^2_{\mathcal{H}_2}\right) + \lambda\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1}\\
\leq& 2\left[\mathbb{E}_{\mathbf{z}}\left(\|S^*_{\mathbf{x}}\{S_{\mathbf{x}}A(f_\rho) - \mathbf{y}\}\|^2_{\mathcal{H}_2}\right)\right]^{1/2}\left[\mathbb{E}_{\mathbf{z}}\left(\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|^2_{\mathcal{H}_2}\right)\right]^{1/2}\\
&+ \left[\mathbb{E}_{\mathbf{z}}\left(\|S^*_{\mathbf{x}}S_{\mathbf{x}} - L_K\|^2_{\mathcal{L}(\mathcal{H}_2)}\right)\right]^{1/2}\left[\mathbb{E}_{\mathbf{z}}\left(\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|^4_{\mathcal{H}_2}\right)\right]^{1/2} + \lambda\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1}.
\end{aligned}$$

Under the Lipschitz continuity of A, from (B.5), (B.6), (B.8) with the parameter choice rule (3.1), we obtain

$$\mathbb{E}_{\mathbf{z}}\left(\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|^2_{\mathscr{L}^2(X,\nu;Y)}\right) \to 0 \text{ as } m\to\infty. \tag{B.10}$$

We have $\mathcal{D}(A)$ is weakly closed and $A : \mathcal{H}_1 \to \mathcal{H}_2$ is Lipschitz continuous, this implies that $I_K A : \mathcal{H}_1 \to \mathscr{L}^2(X,\nu;Y)$ is weakly sequentially closed.

Now from (B.9), (B.10) we obtain a subsequence again denoted by $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$ such that $\langle f_{\mathbf{z}(k),\lambda} - f_\rho, f\rangle_{\mathcal{H}_1} \to \left\langle \tilde{f}, f\right\rangle_{\mathcal{H}_1}$ for some $\tilde{f} \in \mathcal{H}_1$, for all $f \in \mathcal{H}_1$ and $\|I_K\{A(f_{\mathbf{z}(k),\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)} \to 0$ as $k\to\infty$ almost surely. The

assumed injectivity of the mappings I_K and A yields the injectivity of the composition $I_K \circ A$. Therefore the weak closedness of $I_K \circ A$ implies that $\tilde{f} = 0$.

Our next aim is to prove the convergence (3.2). By contradiction, assume that there exists an $\varepsilon > 0$ and a subsequence $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$ such that

$$\mathbb{E}_{\mathbf{z}(k)}\left(\|f_{\mathbf{z}(k),\lambda} - f_\rho\|^2_{\mathcal{H}_1}\right) \geq \varepsilon \text{ for all } k \in \mathbb{N}. \tag{B.11}$$

We have the identity

$$\|f_{\mathbf{z}(k),\lambda} - f_\rho\|^2_{\mathcal{H}_1} = \|f_{\mathbf{z}(k),\lambda} - \bar{f}\|^2_{\mathcal{H}_1} + \|f_\rho - \bar{f}\|^2_{\mathcal{H}_1} + 2\langle \bar{f} - f_{\mathbf{z}(k),\lambda}, f_\rho - \bar{f}\rangle_{\mathcal{H}_1}. \tag{B.12}$$

Using the same arguments as above, we can again find a further subsequence $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$ such that $\mathbb{E}_{\mathbf{z}}\left\langle f_{\mathbf{z}(k),\lambda} - f_\rho, f\right\rangle \to 0$ for all $f \in \mathcal{H}_1$ as $k \to \infty$. Hence from the inequalities (B.8) and (B.12), we obtain

$$\begin{aligned}
&\limsup_{k\to\infty} \mathbb{E}_{\mathbf{z}(k)}\left(\|f_{\mathbf{z}(k),\lambda} - f_\rho\|^2_{\mathcal{H}_1}\right) \\
\leq& 2\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1} + 2\limsup_{k\to\infty} \mathbb{E}_{\mathbf{z}(k)}\left(\langle \bar{f} - f_{\mathbf{z}(k),\lambda}, f_\rho - \bar{f}\rangle_{\mathcal{H}_1}\right) \\
=& 2\limsup_{k\to\infty} \mathbb{E}_{\mathbf{z}(k)}\left(\langle f_\rho - f_{\mathbf{z}(k),\lambda}, f_\rho - \bar{f}\rangle_{\mathcal{H}_1}\right) = 0,
\end{aligned}$$

which contradicts (B.11). This completes the proof of the desired result (3.2). □

Proof of Theorem 3.4. From the inequality (B.3) and Proposition A.3 under Assumptions 1–4, the following inequality holds with the confidence $1 - \eta/2$:

$$\|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1} \leq \frac{4\kappa(M+\Sigma)L}{\lambda\sqrt{m}} \log\left(\frac{4}{\eta}\right) + \|f_\rho - \bar{f}\|_{\mathcal{H}_1}. \tag{B.13}$$

Choosing the parameter $\eta(m) = 4/m^2$, we obtain

$$\mathbb{P}_{\mathbf{z}\in Z^m}\left\{E_m : \left\|f_{\mathbf{z},\lambda} - \bar{f}\right\|_{\mathcal{H}_1} > 8\kappa L(M+\Sigma)\frac{\log m}{\lambda\sqrt{m}} + \left\|f_\rho - \bar{f}\right\|_{\mathcal{H}_1}\right\} \leq \frac{2}{m^2}.$$

Therefore, the sum of the probabilities of the events E_m is finite:

$$\sum_{m=1}^{\infty} \mathbb{P}_{\mathbf{z}\in Z^m}(E_m) \leq \sum_{m=1}^{\infty} \frac{2}{m^2} < \infty.$$

Hence applying the Borel-Cantelli lemma we get,

$$\mathbb{P}_{\mathbf{z}}\left(\limsup_{m\to\infty} E_m\right) = 0$$

from which with the parameter choice rule (3.3) we deduce that

$$\limsup_{m\to\infty} \|f_{\mathbf{z},\lambda} - \bar{f}\|_{\mathcal{H}_1} \leq \|f_\rho - \bar{f}\|_{\mathcal{H}_1} \tag{B.14}$$

almost surely. Note that $f_{\mathbf{z},\lambda}$ is finite almost surely due to (B.14). Hence, there exists a subsequence of $(f_{\mathbf{z},\lambda})_{m\in\mathbb{N}}$ which weakly converges to some $\tilde{f}$. We denote the subsequence by $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$, i.e., $f_{\mathbf{z}(k),\lambda} \rightharpoonup \tilde{f}$. The next step of the proof is to show that $\tilde{f} = f_\rho$.

From inequality (B.2) and Proposition A.3 under Assumptions 1–3, the following inequality holds with confidence $1-\eta$,

$$\begin{aligned}
&\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|^2_{\mathscr{L}^2(X,\nu;Y)} \\
\leq &\frac{4\kappa(M+\Sigma)}{\sqrt{m}}\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|_{\mathcal{H}_2}\log\left(\frac{4}{\eta}\right) \\
&+ \frac{4\kappa^2}{\sqrt{m}}\|A(f_{\mathbf{z},\lambda}) - A(f_\rho)\|^2_{\mathcal{H}_2}\log\left(\frac{4}{\eta}\right) + \lambda\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1}.
\end{aligned}$$

Using the arguments similar to above, with the parameter choice rule (3.3) we obtain $\|I_K\{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)} \to 0$ almost surely. Now $f_{\mathbf{z}(k),\lambda} \rightharpoonup \tilde{f}$ in $\mathcal{H}_1$ and $\|I_K\{A(f_{\mathbf{z}(k),\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)} \to 0$ as $k \to \infty$ almost surely, hence the weak closedness and one-to-one assumption on assumption on $I_K A$ imply that $\tilde{f} = f_\rho$.

Our next aim is to prove the convergence (3.4). By contradiction, assume that there exists an $\varepsilon > 0$ and a subsequence $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$ such that

$$\|f_{\mathbf{z}(k),\lambda} - f_\rho\|^2_{\mathcal{H}_1} \geq \varepsilon. \tag{B.15}$$

We have the identity

$$\|f_{\mathbf{z}(k),\lambda} - f_\rho\|^2_{\mathcal{H}_1} = \|f_{\mathbf{z}(k),\lambda} - \bar{f}\|^2_{\mathcal{H}_1} + \|f_\rho - \bar{f}\|^2_{\mathcal{H}_1} + 2\langle \bar{f} - f_{\mathbf{z}(k),\lambda}, f_\rho - \bar{f}\rangle_{\mathcal{H}_1}. \tag{B.16}$$

Using the same arguments as above, we can again find a further subsequence $(f_{\mathbf{z}(k),\lambda})_{k\in\mathbb{N}}$ which weakly converges to f_ρ. Hence from the inequalities (B.14) and (B.16), we obtain almost surely,

$$\begin{aligned}
\limsup_{m\to\infty}\|f_{\mathbf{z}(k),\lambda} - f_\rho\|^2_{\mathcal{H}_1} &\leq 2\|f_\rho - \bar{f}\|^2_{\mathcal{H}_1} + 2\limsup_{m\to\infty}\langle \bar{f} - f_{\mathbf{z}(k),\lambda}, f_\rho - \bar{f}\rangle_{\mathcal{H}_1} \\
&= 2\limsup_{m\to\infty}\langle f_\rho - f_{\mathbf{z}(k),\lambda}, f_\rho - \bar{f}\rangle_{\mathcal{H}_1} = 0,
\end{aligned}$$

which contradicts (B.15). This completes the proof of the desired result (3.4). □

Appendix C: Proof of upper rates

Regularization schemes given by an explicit regularization formula for direct and linear inverse learning problems are well-studied in the reproducing kernel Hilbert space setting [29, 4]. Optimal convergence rates were established for these kernel methods. In contrast, the main difficulty for the problem considered here arises from the nonlinearity of the operator A. Because of this, the minimizer in the Tikhonov regularization scheme is not explicit, and in the proofs we

have to rely only on optimality properties of the variational formulation (1.3). For classical nonlinear inverse problems [11] the convergence analysis is developed for 2-step approaches. The approach in this study finds its difficulty due to the random sampling and then the empirical error based on the samples. To achieve the optimal convergence rates, the error bound should be bounded in terms of $\mathcal{S}_e = \left\|(L_K + \lambda I)^{-1/2} S_{\mathbf{x}}^*(S_{\mathbf{x}} A(f_\rho) - \mathbf{y})\right\|_{\mathcal{H}_2}$ which then yields error bounds in terms of effective dimension, allowing us to achieve optimal convergence rates.

We introduce the operator $\Xi := S_{\mathbf{x}}(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1} S_{\mathbf{x}}^*$ and the element $\Delta := S_{\mathbf{x}} A(f_\rho) - \mathbf{y}$.

Proof of Theorem 4.6. By the definition of $f_{\mathbf{z},\lambda}$ as a solution to the minimization problem (1.3), the inequality holds true:

$$\|S_{\mathbf{x}} A(f_{\mathbf{z},\lambda}) - \mathbf{y}\|_m^2 + \lambda \left\|f_{\mathbf{z},\lambda} - \bar{f}\right\|_{\mathcal{H}_1}^2 \leq \|S_{\mathbf{x}} A(f_\rho) - \mathbf{y}\|_m^2 + \lambda \left\|f_\rho - \bar{f}\right\|_{\mathcal{H}_1}^2$$

which implies

$$\begin{aligned} &\|S_{\mathbf{x}} \{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_m^2 + \lambda \|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}^2 \\ \leq &2\lambda \left\langle f_\rho - \bar{f}, f_\rho - f_{\mathbf{z},\lambda}\right\rangle_{\mathcal{H}_1} + 2 \left\langle A(f_\rho) - A(f_{\mathbf{z},\lambda}), S_{\mathbf{x}}^* \Delta\right\rangle_{\mathcal{H}_2}. \end{aligned}$$

Under the conditions (i) and (iii) of Assumption 5, for $f \in \mathcal{B}_d(f_\rho)$ we get

$$A(f) = A(f_\rho) + A'(f_\rho)(f - f_\rho) + r(f) \tag{C.1}$$

holds with

$$\|I_K r(f)\|_{\mathscr{L}^2(X,\nu;Y)} \leq \frac{\gamma}{2} \|f - f_\rho\|_{\mathcal{H}_1}^2 \tag{C.2}$$

and

$$\begin{aligned} &\|r(f)\|_{\mathcal{H}_2} = \|A(f) - A(f_\rho) - A'(f_\rho)(f - f_\rho)\|_{\mathcal{H}_2} \qquad (C.3) \\ = &\left\| \int_0^1 \{A'(f_\rho + t(f - f_\rho)) - A'(f_\rho)\}(f - f_\rho)\, dt \right\|_{\mathcal{H}_2} \\ \leq &\int_0^1 \|\{A'(f_\rho + t(f - f_\rho)) - A'(f_\rho)\}\|_{\mathcal{H}_1 \to \mathcal{H}_2} \|f - f_\rho\|_{\mathcal{H}_1}\, dt \leq 2L \|f - f_\rho\|_{\mathcal{H}_1}. \end{aligned}$$

Note that under condition (4.4), from inequality (B.13) we get $f_{\mathbf{z},\lambda} \in \mathcal{B}_d(f_\rho)$ with confidence $1 - \eta/2$ for $d = 4\left\|f_\rho - \bar{f}\right\|_{\mathcal{H}_1}$, therefore using the linearization of the non-linear operator A in (C.1) at $f_{\mathbf{z},\lambda}$ and under Assumption 6 we obtain,

$$\begin{aligned} &\|I_K \{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}\|_{\mathscr{L}^2(X,\nu;Y)}^2 + \lambda \|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}^2 \\ \leq &2\lambda \left\langle T^{1/2} v, f_\rho - f_{\mathbf{z},\lambda}\right\rangle_{\mathcal{H}_1} + 2 \left\langle A(f_\rho) - A(f_{\mathbf{z},\lambda}), L_K (L_K + \lambda I)^{-1} S_{\mathbf{x}}^* \Delta\right\rangle_{\mathcal{H}_2} \\ &+ 2\lambda \left\langle A(f_\rho) - A(f_{\mathbf{z},\lambda}), (L_K + \lambda I)^{-1} S_{\mathbf{x}}^* \Delta\right\rangle_{\mathcal{H}_2} \\ &+ \left\langle (L_K - S_{\mathbf{x}}^* S_{\mathbf{x}}) \{A(f_{\mathbf{z},\lambda}) - A(f_\rho)\}, A(f_{\mathbf{z},\lambda}) - A(f_\rho)\right\rangle_{\mathcal{H}_2} \end{aligned}$$

$$
\begin{aligned}
&\leq 2\lambda \left\langle v, T^{1/2}(f_\rho - f_{\mathbf{z},\lambda}) \right\rangle_{\mathcal{H}_1} + 2\mathcal{S}_e \left\| I_K \left\{ A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\quad + 2\sqrt{\lambda} L \mathcal{S}_e \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1} + I_1 \left\| A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\|_{\mathcal{H}_2}^2 \\
&\leq 2\lambda R \left\| I_K \left\{ A(f_\rho) - A(f_{\mathbf{z},\lambda}) + r(f_{\mathbf{z},\lambda}) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\quad + 2\mathcal{S}_e \left\| I_K \left\{ A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\quad + 2\sqrt{\lambda} L \mathcal{S}_e \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1} + L^2 I_1 \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1}^2 \\
&\leq 2\left(\lambda R + \mathcal{S}_e\right) \left\| I_K \left\{ A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} + 2\sqrt{\lambda} L \mathcal{S}_e \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1} \\
&\quad + L^2 I_1 \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1}^2 + \lambda\gamma R \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1}^2,
\end{aligned}
$$

where $I_1 = \left\| S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K \right\|_{L(\mathcal{H}_2)}$ and $\mathcal{S}_e = \left\| (L_K + \lambda I)^{-1/2} S_{\mathbf{x}}^* (S_{\mathbf{x}} A(f_\rho) - \mathbf{y}) \right\|_{\mathcal{H}_2}$.

It gives

$$
\begin{aligned}
&\left(\left\| I_K \left\{ A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} - \lambda R - \mathcal{S}_e \right)^2 \\
&+ \left(\sqrt{\lambda\gamma_1} \left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1} - \frac{L}{\sqrt{\gamma_1}} \mathcal{S}_e \right)^2 \\
&\leq \left(\lambda R + \mathcal{S}_e\right)^2 + \frac{L^2}{\gamma_1} \mathcal{S}_e^2
\end{aligned}
$$

where $\gamma_1 = 1 - \gamma R - L^2 I_1/\lambda$. This implies

$$
\left\| I_K \left\{ A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} \leq 2R\lambda + \left(2 + \frac{L}{\sqrt{\gamma_1}} \right) \mathcal{S}_e
$$

and

$$
\left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1} \leq \frac{1}{\sqrt{\gamma_1}} R\sqrt{\lambda} + \left(\frac{1}{\sqrt{\gamma_1}} + \frac{2L}{\gamma_1} \right) \frac{\mathcal{S}_e}{\sqrt{\lambda}}.
$$

Now under Assumptions 1–3 using the estimates of Proposition A.3, the inequality (4.4) and (4.5), we obtain that $\gamma_1 = 1/2 - \gamma R > 0$ and with the probability $1 - \eta$,

$$
\begin{aligned}
&\left\| I_K \left\{ A(f_{\mathbf{z},\lambda}) - A(f_\rho) \right\} \right\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\leq 2R\lambda + \frac{2(L + 2\sqrt{\gamma_1})}{\sqrt{\gamma_1}} \left(\frac{\kappa M}{m\sqrt{\lambda}} + \sqrt{\frac{\Sigma^2 \mathcal{N}(\lambda)}{m}} \right) \log\left(\frac{4}{\eta} \right)
\end{aligned}
$$

and

$$
\left\| f_{\mathbf{z},\lambda} - f_\rho \right\|_{\mathcal{H}_1} \leq \frac{1}{\sqrt{\gamma_1}} R\sqrt{\lambda} + \frac{2(2L + \sqrt{\gamma_1})}{\gamma_1} \left(\frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2 \mathcal{N}(\lambda)}{m\lambda}} \right) \log\left(\frac{4}{\eta} \right).
$$

which implies the desired result. □

For the analysis of Tikhonov regularization under general source condition, we consider the linearized and population version (i.e. using theoretical expectation under ρ) of the regularization scheme (1.3):

$$f_\lambda^l := \arg\min_{f\in\mathcal{H}_1}\left\{\int_Z \|A(f_\rho)(x)+A'(f_\rho)(f-f_\rho)(x)-y\|_Y^2 d\rho(x,y)+\lambda\|f-\bar{f}\|_{\mathcal{H}_1}^2\right\}.$$

Under Assumption 1, using the fact $\mathcal{E}(f) := \int_Z \|A(f_\rho)(x)+A'(f_\rho)(f-f_\rho)(x)-y\|_Y^2 d\rho(x,y) = \|T^{1/2}(f-f_\rho)\|_{\mathcal{H}_1}^2 + \mathcal{E}(f_\rho)$, we get

$$f_\lambda^l = (T+\lambda I)^{-1}(Tf_\rho+\lambda\bar{f}). \tag{C.4}$$

In the following proposition, we estimate the error bound of approximation error $f_\lambda^l - f_\rho$ which describes the complexity of the true solution f_ρ. The approximation error is independent of the samples $\mathbf{z}$.

Proposition C.1. *Suppose Assumptions 1, 6 holds true. Then under the assumption that $\phi(t)$ and $t/\phi(t)$ are non-decreasing functions, we have*

$$\left\|f_\lambda^l - f_\rho\right\|_{\mathcal{H}_1} \le R\phi(\lambda).$$

Proof. From the definition of f_λ^l in (C.4) and Assumption 6 we get,

$$f_\rho - f_\lambda^l = \lambda(T+\lambda I)^{-1}\phi(T)v.$$

Under the assumption that $\phi(t)$ and $t/\phi(t)$ are non-decreasing functions, we obtain,

$$\left\|f_\lambda^l - f_\rho\right\|_{\mathcal{H}_1} \le R\phi(\lambda). \qquad \square$$

Under Assumption 6 from Proposition C.1, we observe that $f_\lambda^l \in \mathcal{D}(A)\cap\mathcal{B}_d(f_\rho)$, provided λ is sufficiently small.

In the following theorem, we estimate the quantity $f_\lambda^l - f_{\mathbf{z},\lambda}$ and use the bound of the approximation error from the above proposition to find the error bound for $f_\rho - f_{\mathbf{z},\lambda}$.

Proof of Theorem 4.7. The main idea of the proof is to compare $f_{\mathbf{z},\lambda}$ and f_λ^l. From the definition of $f_{\mathbf{z},\lambda}$ in (1.3), we have

$$\|S_{\mathbf{x}}A(f_{\mathbf{z},\lambda})-\mathbf{y}\|_m^2+\lambda\|f_{\mathbf{z},\lambda}-\bar{f}\|_{\mathcal{H}_1}^2 \le \left\|S_{\mathbf{x}}A(f_\lambda^l)-\mathbf{y}\right\|_m^2+\lambda\|f_\lambda^l-\bar{f}\|_{\mathcal{H}_1}^2. \tag{C.5}$$

Using the linearization of operator A in (C.1) we reexpress the inequality (C.5) as follows,

$$\begin{aligned}\|f_{\mathbf{z},\lambda}-f_\lambda^l\|_{\mathcal{H}_1}^2 \le 2\langle f_\lambda^l-f_{\mathbf{z},\lambda}, f_\lambda^l-\bar{f}\rangle_{\mathcal{H}_1} + \frac{1}{\lambda}\big\{&\|B_{\mathbf{x}}(f_\lambda^l-f_\rho)+\Delta+S_{\mathbf{x}}(r(f_\lambda^l))\|_m^2\\ &-\|B_{\mathbf{x}}(f_{\mathbf{z},\lambda}-f_\rho)+\Delta+S_{\mathbf{x}}(r(f_{\mathbf{z},\lambda}))\|_m^2\big\}\end{aligned}$$

Now we decompose the second and third term in the right-hand side as follows:

$$\begin{aligned}
&\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1}^2 \\
\leq& 2\langle f_\lambda^l - f_{\mathbf{z},\lambda}, f_\lambda^l - \bar{f}\rangle_{\mathcal{H}_1} + \frac{1}{\lambda}\left\{\|\Xi\Delta + S_{\mathbf{x}}(r(f_\lambda^l))\|_m^2\right. \\
&+2\langle \Xi\Delta + S_{\mathbf{x}}(r(f_\lambda^l)), B_{\mathbf{x}}(f_\lambda^l - f_\rho) + (I-\Xi)\Delta\rangle_m \\
&-\|B_{\mathbf{x}}(f_{\mathbf{z},\lambda} - f_\lambda^l) + \Xi\Delta + S_{\mathbf{x}}(r(f_{\mathbf{z},\lambda}))\|_m^2 \\
&\left.-2\langle B_{\mathbf{x}}(f_{\mathbf{z},\lambda} - f_\lambda^l) + \Xi\Delta + S_{\mathbf{x}}(r(f_{\mathbf{z},\lambda})), B_{\mathbf{x}}(f_\lambda^l - f_\rho) + (I-\Xi)\Delta\rangle_m\right\}.
\end{aligned}$$

The fourth term in the right-hand side is negative, therefore it can be ignored, leading to:

$$\begin{aligned}
&\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1}^2 \qquad\qquad (C.6)\\
\leq&\frac{1}{\lambda}\left\{2\langle f_\lambda^l - f_{\mathbf{z},\lambda}, \lambda(f_\lambda^l - \bar{f}) + T_{\mathbf{x}}(f_\lambda^l - f_\rho) + B_{\mathbf{x}}^*(I-\Xi)\Delta\rangle_{\mathcal{H}_1}\right. \\
&+ 2\langle r(f_\lambda^l) - r(f_{\mathbf{z},\lambda}), S_{\mathbf{x}}^* B_{\mathbf{x}}(f_\lambda^l - f_\rho) + S_{\mathbf{x}}^*(I-\Xi)\Delta\rangle_{\mathcal{H}_2} \\
&\left.+ \|\Xi\Delta + S_{\mathbf{x}}(r(f_\lambda^l))\|_m^2\right\}.
\end{aligned}$$

The definition of f_λ^l in (C.4) implies that

$$\lambda(f_\lambda^l - \bar{f}) = T(f_\rho - f_\lambda^l). \qquad (C.7)$$

Therefore, from inequality (C.6), using Assumption 5 (ii) and (C.7) we get:

$$\begin{aligned}
\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1}^2 \leq&\frac{1}{\lambda}\left\{2\langle f_\lambda^l - f_{\mathbf{z},\lambda}, (T_{\mathbf{x}} - T)(f_\lambda^l - f_\rho) + B_{\mathbf{x}}^*(I-\Xi)\Delta\rangle_{\mathcal{H}_1}\right. \qquad (C.8)\\
&+ 2\langle r(f_\lambda^l) - r(f_{\mathbf{z},\lambda}), S_{\mathbf{x}}^* B_{\mathbf{x}}(f_\lambda^l - f_\rho) + S_{\mathbf{x}}^*(I-\Xi)\Delta\rangle_{\mathcal{H}_2} \\
&\left.+ 2\|\Xi\Delta\|_m^2 + 2\|S_{\mathbf{x}}(r(f_\lambda^l))\|_m^2\right\} \\
\leq&\frac{2}{\lambda}\left\{\langle f_\lambda^l - f_{\mathbf{z},\lambda}, (T_{\mathbf{x}} - T)(f_\lambda^l - f_\rho)\right. \\
&+ \lambda A'(f_\rho)^*(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1} S_{\mathbf{x}}^*\Delta\rangle_{\mathcal{H}_1} \\
&+ \langle r(f_\lambda^l) - r(f_{\mathbf{z},\lambda}), (S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K)A'(f_\rho)(f_\lambda^l - f_\rho) \\
&+ \lambda(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1} S_{\mathbf{x}}^*\Delta\rangle_{\mathcal{H}_2} \\
&+ \langle I_K\{r(f_\lambda^l) - r(f_{\mathbf{z},\lambda})\}, B(f_\lambda^l - f_\rho)\rangle_{\mathscr{L}^2(X,\nu;Y)} + \|\Xi\Delta\|_m^2 \\
&\left.+ \|I_K r(f_\lambda^l)\|_{\mathscr{L}^2(X,\nu;Y)}^2 + \langle (S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K) r(f_\lambda^l), r(f_\lambda^l)\rangle_{\mathcal{H}_2}\right\} \\
\leq&\frac{2}{\lambda}\left\{\left\|f_{\mathbf{z},\lambda} - f_\lambda^l\right\|_{\mathcal{H}_1}\left(L^2 I_1\mathcal{A}_e + \sqrt{\lambda}LI_2\mathcal{S}_e\right)\right. \\
&+ \left\|r(f_\lambda^l) - r(f_{\mathbf{z},\lambda})\right\|_{\mathcal{H}_2}\left(LI_1\mathcal{A}_e + \sqrt{\lambda}I_2\mathcal{S}_e\right) \\
&+ \|I_K\{r(f_\lambda^l) - r(f_{\mathbf{z},\lambda})\}\|_{\mathscr{L}^2(X,\nu;Y)}\|B(f_\lambda^l - f_\rho)\|_{\mathscr{L}^2(X,\nu;Y)} \\
&\left.+ I_2^2\mathcal{S}_e^2 + \|I_K r(f_\lambda^l)\|_{\mathscr{L}^2(X,\nu;Y)}^2 + I_1\|r(f_\lambda^l)\|_{\mathcal{H}_2}^2\right\},
\end{aligned}$$

where $\mathcal{A}_e = \left\|f_\lambda^l - f_\rho\right\|_{\mathcal{H}_1}$, $\mathcal{S}_e = \left\|(L_K+\lambda I)^{-1/2}S_{\mathbf{x}}^*(S_{\mathbf{x}}A(f_\rho)-\mathbf{y})\right\|_{\mathcal{H}_2}$, $I_1 = \|S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K\|_{\mathcal{L}(\mathcal{H}_2)}$ and $I_2 = \|(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\|_{\mathcal{L}(\mathcal{H}_2)}$.

We have $\|Bf\|_{\mathscr{L}^2(X,\nu;Y)} \le \left\|(T+\lambda I)^{1/2}f\right\|_{\mathcal{H}_1}$, $f \in \mathcal{D}(A) \subset \mathcal{H}_1$, therefore we obtain,

$$\left\|B(f_\rho - f_\lambda^l)\right\|_{\mathscr{L}^2(X,\nu;Y)} = \lambda\left\|B(T+\lambda I)^{-1}T^{1/2}w\right\|_{\mathscr{L}^2(X,\nu;Y)} \le \lambda\|w\|_{\mathcal{H}_1}. \tag{C.9}$$

Using the inequalities (C.2), (C.3), (C.9) in (C.8) we obtain,

$$\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1}^2 \le \frac{2}{\lambda}\left\{\lambda\gamma\|w\|_{\mathcal{H}_1}\left\|f_{\mathbf{z},\lambda} - f_\lambda^l\right\|_{\mathcal{H}_1}^2 + \delta_1\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1} + \delta_2\right\},$$

where $\delta_1 = 3L^2I_1\mathcal{A}_e + 3\sqrt{\lambda}LI_2\mathcal{S}_e$ and $\delta_2 = I_2^2\mathcal{S}_e^2 + 4\sqrt{\lambda}LI_2\mathcal{A}_e\mathcal{S}_e + \gamma^2\mathcal{A}_e^4/4 + 3\gamma\lambda\|w\|_{\mathcal{H}_1}\mathcal{A}_e^2/2 + 8L^2I_1\mathcal{A}_e^2$.

Under the condition (4.5) we have,

$$\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1}^2 \le \frac{2}{\gamma_2\lambda}\left\{\delta_1\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1} + \delta_2\right\},$$

where $\gamma_2 = 1 - 2\gamma\|w\|_{\mathcal{H}_1}$.

We have,

$$\left(\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1} - \frac{\delta_1}{\lambda\gamma_2}\right)^2 \le \frac{\delta_1^2}{\lambda^2\gamma_2^2} + \frac{2\delta_2}{\lambda\gamma_2},$$

which implies

$$\|f_{\mathbf{z},\lambda} - f_\lambda^l\|_{\mathcal{H}_1} \le \frac{2\delta_1}{\lambda\gamma_2} + \sqrt{\frac{2\delta_2}{\lambda\gamma_2}}.$$

Using the triangle inequality $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \le \left\|f_{\mathbf{z},\lambda} - f_\lambda^l\right\|_{\mathcal{H}_1} + \left\|f_\lambda^l - f_\rho\right\|_{\mathcal{H}_1}$ we obtain,

$$\begin{aligned}&\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1}\\ &\le \left\{c_1 + c_2\frac{2I_1}{\lambda} + c_3\sqrt{\frac{2I_1}{\lambda}} + c_4\sqrt{\frac{I_2}{2}}\right\}\mathcal{A}_e + \left\{c_5\frac{I_2}{\sqrt{2}} + c_6\sqrt{\frac{I_2}{2}}\right\}\left(\frac{\mathcal{S}_e}{2\sqrt{\lambda}}\right),\end{aligned}$$

where $c_1 = 1 + \gamma\|w\|_{\mathcal{H}_1}/\sqrt{2\gamma_2} + \sqrt{3\gamma\|w\|_{\mathcal{H}_1}/\gamma_2}$, $c_2 = 3L^2/\gamma_2$, $c_3 = \sqrt{8L^2/\gamma_2}$, $c_4 = \sqrt{4L/\gamma_2}$, $c_5 = 2\sqrt{2/\gamma_2} + 12\sqrt{2}L/\gamma_2$ and $c_6 = 2\sqrt{4L/\gamma_2}$.

Now using the estimate of Proposition A.4 with the inequality (4.4), we get with the probability $1-\eta/2$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \le (c_1+c_2+c_3+c_4)\mathcal{A}_e + (c_5+c_6)\left(\frac{\mathcal{S}_e}{2\sqrt{\lambda}}\right).$$

Under Assumptions 1–3, 6 from Proposition A.3, C.1, we obtain with the confidence $1-\eta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \le (c_1+c_2+c_3+c_4)R\phi(\lambda) + (c_5+c_6)\left(\frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m\lambda}}\right)\log\left(\frac{4}{\eta}\right).$$

which implies the desired result. □

Proof of Theorem 4.8. (i) Under the parameter choice $\lambda = \Theta^{-1}\left(m^{-1/2}\right)$ we have

$$\frac{1}{m\lambda} \leq \frac{\phi(\lambda)}{\sqrt{m}}.$$

From Theorem 4.7 and the bound (4.2), it follows that with the confidence $1-\eta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C'\phi\left(\Theta^{-1}\left(m^{-1/2}\right)\right)\log\left(\frac{4}{\eta}\right). \tag{C.10}$$

where $C' := (c_1+c_2+c_3+c_4+c_5+c_6)(R+\kappa M+\kappa L\Sigma)$.
Now defining $\tau := C'\log\left(\frac{4}{\eta}\right)$ gives

$$\eta = \eta_\tau = 4e^{-\tau/C'}.$$

The estimate (C.10) can be reexpressed as

$$\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} > \tau R\phi\left(\Theta^{-1}\left(m^{-1/2}\right)\right)\right\} \leq \eta_\tau. \tag{C.11}$$

(ii) From the condition (4.4) we have $8\kappa^2 \leq \sqrt{m}\lambda$. This together with the parameter choice $\lambda = \Psi^{-1}\left(m^{-1/2}\right)$ implies that

$$\frac{1}{m\lambda} = \frac{\lambda^{-\frac{1}{2}+\frac{1}{2b}}\phi(\lambda)}{\sqrt{m}} \leq \frac{\lambda^{\frac{1}{2}+\frac{1}{2b}}\phi(\lambda)}{8\kappa^2}.$$

Now for $\lambda \geq 1$ and $b > 1$ we have $\lambda^{-\frac{1}{2}+\frac{1}{2b}} \leq 1$, therefore $\frac{1}{m\lambda} \leq \phi(\lambda)$. On the other hand, for $\lambda \leq 1$ we have $\lambda^{\frac{1}{2}+\frac{1}{2b}} \leq 1$, therefore $\frac{1}{m\lambda} \leq \frac{\phi(\lambda)}{8\kappa^2}$. Hence, from Theorem 4.7 and the inequality (4.3), it follows that with the confidence $1-\eta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} \leq C''\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)\log\left(\frac{4}{\eta}\right), \tag{C.12}$$

where $C'' := (c_1+c_2+c_3+c_4+c_5+c_6)(R+\kappa M\max(1,\frac{1}{8\kappa^2})+\Sigma\sqrt{C_{\beta,b}})$.
Now defining $\tau := C''\log\left(\frac{4}{\eta}\right)$ gives

$$\eta = \eta_\tau = 4e^{-\tau/C''}.$$

The estimate (C.12) can be reexpressed as

$$\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|_{\mathcal{H}_1} > \tau R\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)\right\} \leq \eta_\tau. \tag{C.13}$$

Then from (C.11) and (C.13), our conclusions follow. □

Appendix D: Proof of lower rates

The following proposition is a variant of Proposition 4 [5] for the non-linear statistical inverse problem.

Proposition D.1. *For the probability measure ρ_f, defined in (4.7), parameterized by $f \in \mathcal{D}(A) \subset \mathcal{H}_1$:*

(i) The solution f_ρ for the probability measure $\rho = \rho_f$ is f.
(ii) The probability measure ρ_f satisfies Assumption 2 provided that

$$dJ + J/4 \leq M \text{ and } 2dJ \leq \Sigma. \tag{D.1}$$

Proof. The first point can be easily observed. Now we check the condition on the probability measure ρ_f for the second point.

Under the condition (D.1) for the conditional probability measure $\rho_f(y|x)$ we have,

$$\begin{aligned}
&\int_Y \left(e^{\|y-A(f)(x)\|_Y/M} - \frac{\|y-A(f)(x)\|_Y}{M} - 1 \right) d\rho_f(y|x) \\
\leq &\int_Y \|y-A(f)(x)\|_Y^2 d\rho_f(y|x) \sum_{i=2}^{\infty} \frac{(dJ + \|A(f)(x)\|_Y)^{i-2}}{M^i i!} \\
\leq &2d^2J^2 \sum_{i=2}^{\infty} \frac{(dJ + \|A(f)(x)\|_Y)^{i-2}}{M^i i!} \leq \frac{\Sigma^2}{2M^2}
\end{aligned}$$

which implies that for the solution $f_\rho = f$ the probability measure ρ_f satisfies Assumption 2. □

Proposition D.2. *Under Assumptions 3, 8, there is an $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$, there exists $N_\varepsilon \in \mathbb{N}$ and each $f_1, \ldots, f_{N_\varepsilon} \in \mathcal{H}_1$ (depending on ε) satisfying:*

(i) For $i = 1, \ldots, N_\varepsilon$, $f_i \in \Omega(\rho_{f_i}, \phi, R)$ and for any $i, j = 1, \ldots, N_\varepsilon$ with $i \neq j$,

$$\varepsilon\upsilon \leq \|f_i - f_j\|_{\mathcal{H}_1},$$

where $\upsilon = 1 - \|I - R_{f_i}\|_{\mathcal{L}(\mathcal{H}_1)} - \|I - R_{f_j}\|_{\mathcal{L}(\mathcal{H}_1)}$ is positive for sufficiently small ε and R_{f_i} are defined in Assumption 8 (iv).

(ii) Let $\rho_i := \rho_{f_i}$, $\rho_j := \rho_{f_j}$ be given by (4.7) for $f_i \in \Omega(\rho_i, \phi, R)$ and $f_j \in \Omega(\rho_j, \phi, R)$, $i, j = 1, \ldots, N_\varepsilon$, then Kullback-Leibler information $K(\rho_{f_i}, \rho_{f_j})$ fulfills the inequality:

$$\mathcal{K}(\rho_{f_i}, \rho_{f_j}) \leq \frac{16}{15dJ^2} \|I_K\{A(f_i) - A(f_j)\}\|^2_{\mathcal{L}^2(X,\nu;Y)}. \tag{D.2}$$

Further, it holds

$$\mathcal{K}(\rho_i, \rho_j) \leq \widetilde{C} \left(\frac{\varepsilon^2}{\ell_\varepsilon^b} + \varepsilon^4 \right), \tag{D.3}$$

where $N_\varepsilon \geq e^{\ell_\varepsilon/24}$ for $\ell_\varepsilon = \left\lfloor \frac{1}{2} \left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)} \right)^{1/b} \right\rfloor$ and $\widetilde{C} = \frac{16c''}{15dJ^2}$.

(iii) The eigenvalues $(t_n^i)_{n\in\mathbb{N}}$ *of the operators* $T_i = A'(f_i)^* I_K^* I_K A'(f_i)$ *follow the polynomial decay for the each* f_i *(*$1 \le i \le N_\varepsilon$*): For fixed positive constants* α_i, β_i *and* $b > 1$,

$$\alpha_i n^{-b} \le t_n^i \le \beta_i n^{-b} \quad \forall n \in \mathbb{N}.$$

Proof. For the initial guess $\bar{f}$ of the solution of the functional (1.3), let $(e_n)_{n\in\mathbb{N}}$ be an orthonormal basis of the Hilbert space $\mathcal{H}_1$ of eigenvectors of the operator $\overline{T} = A'(\bar{f})^* I_K^* I_K A'(\bar{f})$ corresponding to the eigenvalues $(t_n)_{n\in\mathbb{N}}$. For given $\varepsilon > 0$, we define

$$v = \sum_{n=\ell+1}^{2\ell} \frac{\varepsilon \pi^{n-\ell} e_n}{\sqrt{\ell}\phi(t_n)},$$

where $\pi = (\pi^1, \ldots, \pi^\ell) \in \{-1, +1\}^\ell$.

Under the polynomial decay condition $\alpha \le n^b t_n$ on the eigenvalues of the operator $\overline{T}$, we get

$$\|v\|_{\mathcal{H}_1}^2 = \sum_{n=\ell+1}^{2\ell} \frac{\varepsilon^2}{\ell\phi^2(t_n)} \le \sum_{n=\ell+1}^{2\ell} \frac{\varepsilon^2}{\ell\phi^2\left(\frac{\alpha}{n^b}\right)} \le \frac{\varepsilon^2}{\phi^2\left(\frac{\alpha}{2^b\ell^b}\right)} \le R^2,$$

for

$$\ell = \ell_\varepsilon = \left\lfloor \frac{1}{2} \left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)} \right)^{1/b} \right\rfloor, \tag{D.4}$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

We choose $\varepsilon_\circ$ such that $\ell_{\varepsilon_\circ} > 16$. Then from Proposition 6 [5], for every positive $\varepsilon < \varepsilon_\circ$ ($\ell_\varepsilon > \ell_{\varepsilon_\circ}$) there exists an integer $N_\varepsilon \in \mathbb{N}$ and $\pi_1, \ldots, \pi_{N_\varepsilon} \in \{-1, +1\}^{\ell_\varepsilon}$ such that for all $1 \le i, j \le N_\varepsilon,\ i \ne j$ it holds

$$\sum_{n=1}^{\ell_\varepsilon} (\pi_i^n - \pi_j^n)^2 \ge \ell_\varepsilon \tag{D.5}$$

and

$$\log(N_\varepsilon) \ge \ell_\varepsilon/24. \tag{D.6}$$

Now we construct N_ε-vectors satisfying the source condition (Assumption 6). For ε such that $0 < \varepsilon < \varepsilon_\circ$, we define

$$v_i = \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\varepsilon \pi_i^{n-\ell_\varepsilon} e_n}{\sqrt{\ell_\varepsilon}\phi(t_n)}, \tag{D.7}$$

where $\pi_i = (\pi_i^1, \ldots, \pi_i^{\ell_\varepsilon}) \in \{-1, +1\}^{\ell_\varepsilon}$ for $i = 1, \ldots, N_\varepsilon$. Hence from (D.4), we observe that $\|v_i\|_{\mathcal{H}_1} \le R$.

Suppose $F(f) = \phi(T)v + \bar{f}$ for $B = I_K A'(f)$, $T = B^*B$ and some $v \in \mathcal{H}_1$, then from Assumptions 3, 8 (iv) for the Lipschitz continuous function $\theta(t) =$

$\phi(t^2)$ from Propositions D.4, D.5 under the Lipschitz continuity of the Fréchet derivative of the operator A we obtain,

$$\begin{aligned}
&\left\|F(\tilde{f}) - F(f)\right\|_{\mathcal{H}_1} \\
\leq &\left\|\{\phi(\widetilde{T}) - \phi(T)\}v\right\|_{\mathcal{H}_1} \leq \left\|\{\theta(\widetilde{T}^{1/2}) - \theta(T^{1/2})\}v\right\|_{\mathcal{H}_1} \\
\leq &\|v\|_{\mathcal{H}_1} \left\|\theta(\widetilde{T}^{1/2}) - \theta(T^{1/2})\right\|_{\mathcal{L}(\mathcal{H}_1)} \leq \|v\|_{\mathcal{H}_1} \left\|\theta(\widetilde{T}^{1/2}) - \theta(T^{1/2})\right\|_{HS} \\
\leq &L_\theta \|v\|_{\mathcal{H}_1} \left\|\widetilde{T}^{1/2} - T^{1/2}\right\|_{HS} \leq \sqrt{2}L_\theta \|v\|_{\mathcal{H}_1} \left\|\widetilde{B} - B\right\|_{HS} \\
\leq &\sqrt{2}L_\theta \|v\|_{\mathcal{H}_1} \left\|I_K \left\{A'(\tilde{f}) - A'(f)\right\}\right\|_{HS} \leq \gamma\sqrt{2}L_\theta \|v\|_{\mathcal{H}_1} \left\|\tilde{f} - f\right\|_{\mathcal{H}_1},
\end{aligned}$$

where $\widetilde{B} = I_K A'(\tilde{f})$ and $\widetilde{T} = \widetilde{B}^* \widetilde{B}$.

If $\gamma\sqrt{2}L_\theta \|v\|_{\mathcal{H}_1} < 1$, then F is a contraction map. Hence, there exists a fixed point $f_* \in \mathcal{H}_1$ such that

$$f_* = F(f_*) = \phi(T_*)v + \bar{f}, \tag{D.8}$$

where $T_* = (I_K A'(f_*))^* I_K A'(f_*)$.

Hence for each v_i defined in (D.7) from (D.8) there exist f_i $(1 \leq i \leq N_\varepsilon)$ such that

$$f_i - \bar{f} = \phi(T_i)v_i,$$

where $B_i = I_K A'(f_i)$ and $T_i = B_i^* B_i$, i.e., $f_i \in \Omega(\rho_{f_i}, \phi, R)$ provided that $\gamma\sqrt{2}L_\theta \|v_i\|_{\mathcal{H}_1} < 1$ for $1 \leq i \leq N_\varepsilon$ which can be satisfied by making the quantity $\|v_i\|_{\mathcal{H}_1}$ arbitrarily small as $\varepsilon \to 0$.

Under Assumption 8 (iv) from eqn. (D.7) for all $1 \leq i, j \leq N_\varepsilon$, we get,

$$f_i - \bar{f} = \phi(T_i)v_i = R_{f_i}\phi(\overline{T})v_i = \{I - (I - R_{f_i})\} \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\varepsilon\pi_i^{n-\ell_\varepsilon} e_n}{\sqrt{\ell_\varepsilon}}$$

and

$$\begin{aligned}
f_i - f_j = &\sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\varepsilon(\pi_i^{n-\ell_\varepsilon} - \pi_j^{n-\ell_\varepsilon})e_n}{\sqrt{\ell_\varepsilon}} - (I - R_{f_i}) \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\varepsilon\pi_i^{n-\ell_\varepsilon} e_n}{\sqrt{\ell_\varepsilon}} \\
&+ (I - R_{f_j}) \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\varepsilon\pi_j^{n-\ell_\varepsilon} e_n}{\sqrt{\ell_\varepsilon}}
\end{aligned} \tag{D.9}$$

which implies from (D.5) that

$$\|f_i - \bar{f}\|_{\mathcal{H}_1} \leq \varepsilon(1 + \|I - R_{f_i}\|_{\mathcal{L}(\mathcal{H}_1)}) \tag{D.10}$$

and

$$\varepsilon\upsilon \leq \|f_i - f_j\|_{\mathcal{H}_1} \tag{D.11}$$

where $\upsilon = 1 - \|I - R_{f_i}\|_{\mathcal{L}(\mathcal{H}_1)} - \|I - R_{f_j}\|_{\mathcal{L}(\mathcal{H}_1)}$.

Then under Assumption 8 (iv) and (D.10) we have,

$$\|I - R_f\|_{\mathcal{L}(\mathcal{H}_1)} \leq \zeta \left\|f - \bar{f}\right\|_{\mathcal{H}_1} \leq \zeta\varepsilon(1 + \|I - R_f\|_{\mathcal{L}(\mathcal{H}_1)})$$

which implies that

$$\|I - R_f\|_{\mathcal{L}(\mathcal{H}_1)} \leq \frac{\zeta\varepsilon}{1 - \zeta\varepsilon}.$$

From Assumptions 3, 8 (ii) and (D.9) we get,

$$\begin{aligned}
&\|\overline{B}(f_i - f_j)\|_{\mathscr{L}^2(X,\nu;Y)} \\
\leq &\|\overline{B}\phi(\overline{T})(v_i - v_j)\|_{\mathscr{L}^2(X,\nu;Y)} + \varepsilon\kappa L(\|I - R_{f_i}\|_{\mathcal{L}(\mathcal{H}_1)} + \|I - R_{f_j}\|_{\mathcal{L}(\mathcal{H}_1)}),
\end{aligned}$$

where $\overline{B} = I_K \circ (A'(\bar{f}))$.

Now from Assumptions 8 (iv), (v) and (D.10) we get,

$$\begin{aligned}
&\|\overline{B}(f_i - f_j)\|_{\mathscr{L}^2(X,\nu;Y)} \qquad (D.12)\\
\leq &\left(\sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{t_n\varepsilon^2\left(\pi_i^{n-\ell_\varepsilon} - \pi_j^{n-\ell_\varepsilon}\right)^2}{\ell_\varepsilon}\right)^{1/2} + \varepsilon\kappa L\zeta(\|f_i - \bar{f}\|_{\mathcal{H}_1} + \|f_j - \bar{f}\|_{\mathcal{H}_1}) \\
\leq &\left(\sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\beta\varepsilon^2\left(\pi_i^{n-\ell_\varepsilon} - \pi_j^{n-\ell_\varepsilon}\right)^2}{\ell_\varepsilon n^b}\right)^{1/2} + c\varepsilon^2 \\
\leq &\left(\sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{4\beta\varepsilon^2}{\ell_\varepsilon n^b}\right)^{1/2} + c\varepsilon^2 \leq \left(\frac{4\beta\varepsilon^2}{\ell_\varepsilon}\int_{\ell_\varepsilon}^{2\ell_\varepsilon} \frac{1}{x^b}dx\right)^{1/2} + c\varepsilon^2 \leq c'\frac{\varepsilon}{\ell_\varepsilon^{b/2}} + c\varepsilon^2,
\end{aligned}$$

where $c = \kappa L\zeta(2 + \|I - R_{f_i}\|_{\mathcal{L}(\mathcal{H}_1)} + \|I - R_{f_j}\|_{\mathcal{L}(\mathcal{H}_1)})$ and $c' = \left(\frac{4\beta}{(b-1)}\left(1 - \frac{1}{2^{b-1}}\right)\right)^{1/2}$.

Note that the Lipschitz continuity of the Fréchet derivative of the operator A (Assumption 8 (iii)) imply that

$$A(f_i) = A(\bar{f}) + A'(\bar{f})(f_i - \bar{f}) + r(f_i)$$

holds with

$$\|I_K r(f_i)\|_{\mathscr{L}^2(X,\nu;Y)} \leq \frac{\gamma}{2}\|f_i - \bar{f}\|_{\mathcal{H}_1}^2.$$

Hence, for $1 \leq i, j \leq N_\varepsilon$, from the inequality (D.10), (D.12) we have,

$$\begin{aligned}
&\|I_K\{A(f_i) - A(f_j)\}\|_{\mathscr{L}^2(X,\nu;Y)}^2 \qquad (D.13)\\
= &\left\{\|\overline{B}(f_i - f_j)\|_{\mathscr{L}^2(X,\nu;Y)} + \|I_K\{r(f_i) - r(f_j)\}\|_{\mathscr{L}^2(X,\nu;Y)}\right\}^2 \\
\leq &2\|\overline{B}(f_i - f_j)\|_{\mathscr{L}^2(X,\nu;Y)}^2 + 2\|I_K\left\{r(f_i) - r(f_j)\right\}\|_{\mathscr{L}^2(X,\nu;Y)}^2 \\
\leq &2\|\overline{B}(f_i - f_j)\|_{\mathscr{L}^2(X,\nu;Y)}^2 + \gamma^2\|f_i - \bar{f}\|_{\mathcal{H}_1}^4 + \gamma^2\|f_j - \bar{f}\|_{\mathcal{H}_1}^4
\end{aligned}$$

$$\leq c'' \left(\frac{\varepsilon^2}{\ell_\varepsilon^b} + \varepsilon^4 \right),$$

where $c'' = 4c^2 + 4c'^2 + \gamma^2\{(1 + \|I - R_{f_i}\|_{\mathcal{L}(\mathcal{H}_1)})^4 + (1 + \|I - R_{f_j}\|_{\mathcal{L}(\mathcal{H}_1)})^4\}$.

Under Assumption 8 (iv), if $\left\|f - \bar{f}\right\|_{\mathcal{H}_1} < 1/\zeta$, then from Neumann series, we have that $\left\|R_f^{-1}\right\|_{\mathcal{L}(\mathcal{H}_1)} < \infty$. Therefore,

$$\phi(T) = R_f \phi(T_*) \quad \text{and} \quad \phi(T_*) = R_f^{-1} \phi(T).$$

Now using the relation for singular values $s_j(AB) \leq \|A\| s_j(B)$ for $j \in \mathbb{N}$ (see Chapter 11 [27]) we obtain,

$$\phi(s_j(T)) = s_j(\phi(T)) \leq \|R_f\|_{\mathcal{L}(\mathcal{H}_1)} s_j(\phi(T_*)) = \|R_f\|_{\mathcal{L}(\mathcal{H}_1)} \phi(s_j(T_*)) \quad \text{(D.14)}$$

and

$$\phi(s_j(T_*)) = s_j(\phi(T_*)) \leq \left\|R_f^{-1}\right\|_{\mathcal{L}(\mathcal{H}_1)} s_j(\phi(T)) = \left\|R_f^{-1}\right\|_{\mathcal{L}(\mathcal{H}_1)} s_j(\phi(T)) \quad \text{(D.15)}$$

Consequently, for small enough $\left\|f_i - \bar{f}\right\|_{\mathcal{H}_1}$ corresponding to small ε, the eigenvalues of T_i and T_* decay in the same order, hence in the polynomial order.

The inequality (D.2) can be proved similar to Proposition 4 [5]. We obtain the desired results from the inequalities (D.6), (D.11), (D.13), (D.14), (D.15) with (D.2). □

The following theorem is a restatement of Theorem 3.1 of [9] in the non-linear statistical inverse problem setting.

Proposition D.3. *For any learning algorithm ($\mathbf{z} \to f_{\mathbf{z}} \in \mathcal{H}_1$) under the hypothesis* $\dim(Y) = d < \infty$, *Assumption 8 and the condition* (D.1), *there exists a probability measure $\rho_* \in \mathcal{P}_{\phi,b}$ and $f_{\rho_*} \in \mathcal{H}_1$ such that for all $0 < \varepsilon < \varepsilon_\circ$, $f_{\mathbf{z}}$ can be approximated as*

$$\mathbb{P}_{\mathbf{z} \in Z^m} \{\|f_{\mathbf{z}} - f_{\rho_*}\|_{\mathcal{H}_1} > \varepsilon\upsilon/2\} \geq \min \left\{ \frac{1}{1 + e^{-\ell_\varepsilon/24}}, \vartheta e^{\left(\frac{\ell_\varepsilon}{48} - \frac{\tilde{C} m \varepsilon^2}{\ell_\varepsilon^b} - \tilde{C} m \varepsilon^4 \right)} \right\}$$

where $\vartheta = e^{-3/e}$ *and* $\ell_\varepsilon = \left\lfloor \frac{1}{2} \left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)} \right)^{1/b} \right\rfloor$.

Proof. Let $\varepsilon \leq \varepsilon_0$ and $f_1, \ldots, f_{N_\varepsilon}$ be as in Proposition D.2. Then we define the sets,

$$A_i = \left\{ \mathbf{z} \in Z^m : \|f_{\mathbf{z}} - f_i\|_{\mathcal{H}_1} < \frac{\varepsilon\upsilon}{2} \right\}, \text{ for } 1 \leq i \leq N_\varepsilon.$$

It is clear from (D.11) that $A_i \cap A_j = \emptyset$ if $i \neq j$. On applying Lemma 3.3 [9] with the probability measures $\rho_{f_i}^m$, $1 \leq i \leq N_\varepsilon$, we obtain that either

$$p := \max_{1 \leq i \leq N_\varepsilon} \rho_{f_i}^m(A_i^c) \geq \frac{N_\varepsilon}{N_\varepsilon + 1} \quad \text{(D.16)}$$

or

$$\min_{1\le j\le N_\varepsilon}\frac{1}{N_\varepsilon}\sum_{i=1,i\neq j}^{N_\varepsilon}\mathcal{K}(\rho^m_{f_i},\rho^m_{f_j})\ge\Psi_{N_\varepsilon}(p), \tag{D.17}$$

where $\Psi_{N_\varepsilon}(p)=\log(N_\varepsilon)+(1-p)\log\left(\frac{1-p}{p}\right)-p\log\left(\frac{N_\varepsilon-p}{p}\right)$. Further,

$$\begin{aligned}\Psi_{N_\varepsilon}(p)\ge&(1-p)\log(N_\varepsilon)+(1-p)\log(1-p)-\log(p)+2p\log(p)\\ \ge&-\log(p)+\log(\sqrt{N_\varepsilon})-3/e.\end{aligned} \tag{D.18}$$

Since minimum value of $x\log(x)$ is $-1/e$ on $[0,1]$.

For the joint probability measures $\rho^m_{f_i}$, $\rho^m_{f_j}$ $(\rho_{f_i},\rho_{f_j}\in\mathcal{P}_{\phi,b},\ 1\le i,j\le N_\varepsilon)$ from the inequality (D.3) we get,

$$\mathcal{K}(\rho^m_{f_i},\rho^m_{f_j})=m\mathcal{K}(\rho_{f_i},\rho_{f_j})\le\widetilde{C}m\left(\frac{\varepsilon^2}{\ell^b_\varepsilon}+\varepsilon^4\right). \tag{D.19}$$

Therefore the inequalities (D.16), (D.17), together with (D.18) and (D.19) implies

$$\begin{aligned}p:=&\max_{1\le i\le N_\varepsilon}\left(\mathbb{P}\left\{\mathbf{z}\in Z^m:\|f_{\mathbf{z}}-f_i\|_{\mathcal{H}_1}>\frac{\upsilon\varepsilon}{2}\right\}\right)\\ \ge&\min\left\{\frac{N_\varepsilon}{N_\varepsilon+1},\sqrt{N_\varepsilon}e^{-\frac{3}{e}-\widetilde{C}m\left(\frac{\varepsilon^2}{\ell^b_\varepsilon}+\varepsilon^4\right)}\right\}.\end{aligned}$$

From the estimate (D.6) for the probability measure ρ_* such that $p=\rho^m_*(A^c_i)$ the desired result follows. □

Proof of Theorem 4.11. From Proposition D.3 for some probability measure $\rho^*\in\mathcal{P}_{\phi,b}$ with $0<\varepsilon<\varepsilon_0$ we get,

$$\begin{aligned}&\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z}}-f_{\rho_*}\|_{\mathcal{H}_1}>\frac{\upsilon\varepsilon}{2}\right\}\\ \ge&\min\left\{\frac{1}{1+e^{-\ell_\varepsilon/24}},\vartheta e^{-\frac{1}{48}}e^{\left\{\frac{1}{96}\left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)}\right)^{1/b}-\widetilde{C}m\varepsilon^2\left(\frac{2^{2b-1}\phi^{-1}(\varepsilon/R)}{\alpha-2^{2b-1}\phi^{-1}(\varepsilon/R)}\right)-\widetilde{C}m\varepsilon^4\right\}}\right\},\end{aligned}$$

where $\vartheta=e^{-3/e}$ and $\ell_\varepsilon=\left\lfloor\frac{1}{2}\left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)}\right)^{1/b}\right\rfloor$.

Given $\tau>0$ for all $m\in\mathbb{N}$, we choose $\varepsilon_m=\tau R\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)$. Since ε_m tends to 0 when m tends to $+\infty$, therefore for m large enough $\varepsilon_m\le\varepsilon_0$. So Proposition D.3 applies to ensure,

$$\begin{aligned}&\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z}}-f_{\rho_*}\|_{\mathcal{H}_1}>\frac{\tau\upsilon R}{2}\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)\right\}\\ \ge&\min\left\{\frac{1}{1+e^{-\ell_\varepsilon/24}},\vartheta e^{-\frac{1}{48}}e^{c(m)}\right\},\end{aligned}$$

where

$$c(m) = \left(\Psi^{-1}\left(m^{-1/2}\right)\right)^{-1/b}\left\{\frac{\alpha^{1/b}}{96} - \frac{\widetilde{C}\tau^2R^2 2^{2b-1}}{\alpha - 2^{2b-1}\Psi^{-1}\left(m^{-1/2}\right)} - \widetilde{C}\tau^4R^4\left(\frac{\phi^2\left(\Psi^{-1}\left(m^{-1/2}\right)\right)}{\Psi^{-1}\left(m^{-1/2}\right)}\right)\right\}.$$

Now as m tends to ∞, $\varepsilon \to 0$ and $\ell_\varepsilon \to \infty$. Therefore, we conclude that

$$\lim_{\tau\to 0}\liminf_{m\to\infty}\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}_{\phi,b}}\mathbb{P}_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z}}^l - f_\rho\|_{\mathcal{H}_1} > \frac{\tau\upsilon R}{2}\phi\left(\Psi^{-1}\left(m^{-1/2}\right)\right)\right\} = 1. \quad \square$$

Proposition D.4. *[8, Lemma 7] Let $F,\ \widetilde{F} : \mathcal{H} \to \mathcal{H}$ be the self-adjoint, Hilbert-Schmidt operators over the separable Hilbert space $\mathcal{H}$. If $\theta(t)$ is Lipschitz continuous with Lipschitz constant $L_\theta \geq 0$, then θ is also operator Lipschitz in Hilbert-Schmidt norm:*

$$\left\|\theta(F) - \theta(\widetilde{F})\right\|_{HS} \leq L_\theta\left\|F - \widetilde{F}\right\|_{HS}.$$

Proposition D.5. *Let $B,\ \widetilde{B} : \mathcal{H}_1 \to \mathcal{H}_2$ be the Hilbert-Schmidt operators over the arbitrary separable Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$ and $T = B^*B$, $\widetilde{T} = \widetilde{B}^*\widetilde{B}$. Then*

$$\left\|T^{1/2} - \widetilde{T}^{1/2}\right\|_{HS} \leq \sqrt{2}\left\|B - \widetilde{B}\right\|_{HS}.$$

Proof. Let $(e_i, f_i, \mu_i)_{i\in\mathbb{N}}$ and $(\tilde{e}_i, \tilde{f}_i, \tilde{\mu}_i)_{i\in\mathbb{N}}$ be the singular value decompositions of the operators B and $\widetilde{B}$, i.e., $B = \sum_{i=1}^{\infty}\mu_i\langle\cdot, e_i\rangle_{\mathcal{H}_1}f_i$ and $\widetilde{B} = \sum_{i=1}^{\infty}\tilde{\mu}_i\langle\cdot, \tilde{e}_i\rangle_{\mathcal{H}_1}\tilde{f}_i$. The values $(\mu_i)_{i\in\mathbb{N}}$ and $(\tilde{\mu}_i)_{i\in\mathbb{N}}$ are the singular values of the operators B and $\widetilde{B}$, respectively. The vectors $(e_i)_{i\in\mathbb{N}}$ and $(\tilde{e}_i)_{i\in\mathbb{N}}$ are the orthonormal basis of the Hilbert space $\mathcal{H}_1$ and the eigenvectors of the operators $T = B^*B$ and $\widetilde{T} = \widetilde{B}^*\widetilde{B}$, respectively. The vectors $(f_i)_{i\in\mathbb{N}}$ and $(\tilde{f}_i)_{i\in\mathbb{N}}$ are the orthonormal basis of the Hilbert space $\mathcal{H}_2$ and the eigenvectors of the operators BB^* and $\widetilde{B}\widetilde{B}^*$, respectively.

We have

$$\begin{aligned}\left\|\{T^{1/2} - \widetilde{T}^{1/2}\}e_j\right\|_{\mathcal{H}_1}^2 &= \langle e_j, Te_j\rangle_{\mathcal{H}_1} + \langle e_j, \widetilde{T}e_j\rangle_{\mathcal{H}_1} - 2\langle T^{1/2}e_j, \widetilde{T}^{1/2}e_j\rangle_{\mathcal{H}_1} \\ &= \|Be_j\|_{\mathcal{H}_2}^2 + \left\|\widetilde{B}e_j\right\|_{\mathcal{H}_2}^2 - 2\mu_j\sum_{i=1}^{\infty}\tilde{\mu}_i\langle e_j, \tilde{e}_i\rangle_{\mathcal{H}_1}^2\end{aligned} \tag{D.20}$$

and

$$\left\|\{B - \widetilde{B}\}e_j\right\|_{\mathcal{H}_2}^2 = \|Be_j\|_{\mathcal{H}_2}^2 + \left\|\widetilde{B}e_j\right\|_{\mathcal{H}_2}^2 - 2\langle Be_j, \widetilde{B}e_j\rangle_{\mathcal{H}_2} \tag{D.21}$$

$$= \|Be_j\|^2_{\mathcal{H}_2} + \left\|\widetilde{B}e_j\right\|^2_{\mathcal{H}_2} - 2\mu_j \sum_{i=1}^{\infty} \tilde{\mu}_i \langle e_j, \tilde{e}_i\rangle_{\mathcal{H}_1} \langle f_j, \tilde{f}_i\rangle_{\mathcal{H}_2}$$

$$\geq \|Be_j\|^2_{\mathcal{H}_2} + \left\|\widetilde{B}e_j\right\|^2_{\mathcal{H}_2} - \mu_j \sum_{i=1}^{\infty} \tilde{\mu}_i \left(\langle e_j, \tilde{e}_i\rangle^2_{\mathcal{H}_1} + \langle f_j, \tilde{f}_i\rangle^2_{\mathcal{H}_2}\right).$$

Similarly,

$$\left\|\{B^* - \widetilde{B}^*\}f_j\right\|^2_{\mathcal{H}_1} \geq \|B^* f_j\|^2_{\mathcal{H}_1} + \left\|\widetilde{B}^* f_j\right\|^2_{\mathcal{H}_1} - \mu_j \sum_{i=1}^{\infty} \tilde{\mu}_i \left(\langle e_j, \tilde{e}_i\rangle^2_{\mathcal{H}_1} + \langle f_j, \tilde{f}_i\rangle^2_{\mathcal{H}_2}\right) \tag{D.22}$$

and

$$\left\|\{(BB^*)^{1/2} - (\widetilde{B}\widetilde{B}^*)^{1/2}\}f_j\right\|^2_{\mathcal{H}_1} = \|B^* f_j\|^2_{\mathcal{H}_1} + \left\|\widetilde{B}^* f_j\right\|^2_{\mathcal{H}_1} - 2\mu_j \sum_{i=1}^{\infty} \tilde{\mu}_i \langle f_j, \tilde{f}_i\rangle^2_{\mathcal{H}_2}. \tag{D.23}$$

From (D.20), (D.21), (D.22), (D.23) we obtain,

$$\left\|\{T^{1/2} - \widetilde{T}^{1/2}\}e_j\right\|^2_{\mathcal{H}_1} + \left\|\{(BB^*)^{1/2} - (\widetilde{B}\widetilde{B}^*)^{1/2}\}f_j\right\|^2_{\mathcal{H}_1}$$
$$\leq \left\|\{B - \widetilde{B}\}e_j\right\|^2_{\mathcal{H}_2} + \left\|\{B^* - \widetilde{B}^*\}f_j\right\|^2_{\mathcal{H}_1}$$

which implies

$$\left\|T^{1/2} - \widetilde{T}^{1/2}\right\|^2_{HS} + \left\|(BB^*)^{1/2} - (\widetilde{B}\widetilde{B}^*)^{1/2}\right\|^2_{HS} \leq \left\|B - \widetilde{B}\right\|^2_{HS} + \left\|B^* - \widetilde{B}^*\right\|^2_{HS}.$$

Hence,

$$\left\|T^{1/2} - \widetilde{T}^{1/2}\right\|^2_{HS} \leq 2\left\|B - \widetilde{B}\right\|^2_{HS}. \qquad \square$$

Acknowledgements

The authors are grateful to the reviewers for their helpful comments that led to improve the quality of the paper. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901.

References

[1] Bauer, F., Hohage, T. and Munk, A. (2009). Iteratively regularized Gauss-Newton method for nonlinear inverse problems with random noise. *SIAM J. Numer. Anal.* **47** 1827–1846. MR2505875

[2] Bauer, F., Pereverzev, S. and Rosasco, L. (2007). On regularization algorithms in learning theory. *J. Complex.* **23** 52–72. MR2297015

[3] Bissantz, N., Hohage, T. and Munk, A. (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Probl.* **20** 1773–1789. MR2107236

[4] Blanchard, G. and Mücke, N. (2018). Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.* **18** 971–1013. MR3833647

[5] Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249

[6] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Stat. & Probab. Lett.* **45** 11–22. MR1718346

[7] Cavalier, L. (2011). Inverse problems in statistics. In *Inverse Probl. high-dimensional Estim.*, *Lect. Notes Stat. Proc.* **203** 3–96. Springer, Heidelberg. MR2868199

[8] De Vito, E., Rosasco, L. and Toigo, A. (2014). Learning sets with separating kernels. *Appl. Comput. Harmon. Anal.* **37** 185–217. MR3223462

[9] DeVore, R., Kerkyacharian, G., Picard, D. and Temlyakov, V. (2006). Approximation methods for supervised learning. *Found. Comput. Math.* **6** 3–58. MR2198214

[10] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilistic theory of pattern recognition* **31**. Applications of Mathematics, Springer, New York. MR1383093

[11] Engl, H. W., Hanke, M. and Neubauer, A. (1996). *Regularization of inverse problems* **375**. Math. Appl., Kluwer Academic Publishers Group, Dordrecht, The Netherlands. MR1408680

[12] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Stat.* **35** 70–91. MR2332269

[13] Hanke, M., Neubauer, A. and Scherzer, O. (1995). A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.* **72** 21–37. MR1359706

[14] Hein, T. and Hofmann, B. (2003). On the nature of ill-posedness of an inverse problem arising in option pricing. *Inverse Probl.* **19** 1319–1338. MR2036533

[15] Hohage, T. and Pricop, M. (2008). Nonlinear Tikhonov regularization in Hilbert scales for inverse boundary value problems with random noise. *Inverse Probl. Imaging* **2** 271–290. MR2395144

[16] Isakov, V. (2017). *Inverse problems for partial differential equations*, third ed. *Applied Mathematical Sciences* **127**. Springer, Cham. MR3616276

[17] Ito, K. and Jin, B. (2015). *Inverse problemss: Tikhonov Theory And Algorithms. Series on Applied Mathematics* **22**. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ. MR3244283

[18] Lin, J., Rudi, A., Rosasco, L. and Cevher, V. (2020). Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Appl. Comput. Harmon. Anal.* **48** 868–890. MR4068943

[19] Lin, K., Lu, S. and Mathé, P. (2015). Oracle-type posterior contraction rates in Bayesian inverse problems. *Inverse Probl. Imaging* **9**

895–915. MR3406424
[20] Lin, S.-B. and Zhou, D.-X. (2018). Optimal Learning Rates for Kernel Partial Least Squares. *J. Fourier Anal. Appl.* **24** 908–933. MR3802296
[21] Loubes, J.-M. and Ludena, C. (2010). Penalized estimators for non linear inverse problems. *ESAIM Probab. Stat.* **14** 173–191. MR2741964
[22] Lu, S., Mathé, P. and Pereverzev, S. V. (2020). Balancing principle in supervised learning for a general regularization scheme. *Appl. Comput. Harmon. Anal.* **48** 123–148. MR4016987
[23] Mathé, P. (2019). Bayesian inverse problems with non-commuting operators. *Math. Comput.* **88** 2897–2912. MR3985479
[24] Mathé, P. and Pereverzev, S. V. (2003). Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Probl.* **19** 789–803. MR1984890
[25] Micchelli, C. A. and Pontil, M. (2005). On learning vector-valued functions. *Neural Comput.* **17** 177–204. MR2175914
[26] O'Sullivan, F. (1990). Convergence characteristics of methods of regularization estimators for nonlinear operator equations. *SIAM J. Numer. Anal.* **27** 1635–1649. MR1080343
[27] Pietsch, A. (1980). *Operator ideals*. North-Holland Mathematical Library, 20. North-Holland Publishing Co., Amsterdam-New York. MR0582655
[28] Pinelis, I. and Sakhanenko, A. I. (1986). Remarks on inequalities for large deviation probabilities. *Theory Probab. Its Appl.* **30** 143–148. MR0779438
[29] Rastogi, A. and Sampath, S. (2017). Optimal rates for the regularized learning algorithms under general source condition. *Front. Appl. Math. Stat.* **3** 3.
[30] Ray, K. and Schmidt-Hieber, J. (2016). Minimax theory for a class of nonlinear statistical inverse problems. *Inverse Probl.* **32** 65003. MR3493583
[31] Rieder, A. (2003). *Keine Probleme mit inversen Problemen*. Friedr. Vieweg & Sohn, Braunschweig. MR2030046
[32] Saitoh, S. and Sawano, Y. (2016). *Theory of reproducing kernels and applications*. Springer, Singapore. MR3560890
[33] Scherzer, O., Engl, H. W. and Kunisch, K. (1993). Optimal a posteriori parameter choice for Tikhonov regularization for solving nonlinear ill-posed problems. *SIAM J. Numer. Anal.* **30** 1796–1838. MR1249043
[34] Schuster, T., Kaltenbacher, B., Hofmann, B. and Kazimierski, K. S. (2012). *Regularization methods in Banach spaces. Radon Series on Computational and Applied Mathematics* **10**. Walter de Gruyter GmbH & Co. KG, Berlin. MR2963507
[35] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press. MR3277164
[36] Smale, S. and Zhou, D.-X. (2005). Shannon sampling II: Connections to learning theory. *Appl. Comput. Harmon. Anal.* **19** 285–302. MR2186447
[37] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Information Science and Statistics, Springer. MR2450103

[38] Valent, T. (1987). *Boundary value problems of finite elasticity – local theorems on existence, uniqueness, and analytic dependence on data* **31**. Springer Tracts in Natural Philosophy. MR0917733

[39] Wahba, G. (1990). *Spline models for observational data*. Philadelphia, PA: SIAM. MR1045442

[40] Werner, F. and Hofmann, B. (2020). Convergence analysis of (statistical) inverse problems under conditional stability estimates. *Inverse Probl.* **36** 015004. MR4047904