Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

Balancing principle in supervised learning for a general regularization scheme

S. Lu, P. Mathe, S. Pereverzyev

RICAM-Report 2016-38

Balancing principle in supervised learning for a general regularization scheme

Shuai Lu

School of Mathematical Sciences, Fudan University, Shanghai, China 200433

Peter Mathé

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

Sergei V. Pereverzev

Johann Radon Institute for Computational and Applied Mathematics, Altenbergerstrasse 69, A-4040 Linz, Austria

Abstract

We discuss the problem of parameter choice in learning algorithms generated by a general regularization scheme. Such a scheme covers well-known algorithms as regularized least squares and gradient descent learning. It is known that in contrast to classical deterministic regularization methods, the performance of regularized learning algorithms is influenced not only by the smoothness of a target function, but also by the capacity of a space, where regularization is performed. In the infinite dimensional case the latter one is usually measured in terms of the effective dimension. In the context of supervised learning both the smoothness and effective dimension are intrinsically unknown *a priori*. Therefore we are interested in *a posteriori* regularization parameter choice, and we propose a new form of the balancing principle. An advantage of this strategy over the known rules such as cross-validation based adaptation is that it does not require any data splitting and allows the use of all available labeled data in the construction of regularized approximants. We provide the analysis of the proposed rule and demonstrate its advantage in simulations.

Keywords: supervised learning, general smoothness, balancing principle

2010 MSC: 65D15, 68T05, 68Q32

Email addresses: slu@fudan.edu.cn (Shuai Lu), peter.mathe@wias-berlin.de (Peter Mathé), sergei.pereverzyev@oeaw.ac.at (Sergei V. Pereverzev)

1. Introduction

The concept of a general regularization scheme was first proposed in Bakushinskii [1] to simultaneously treat different methods for solving linear ill-posed problems in a Hilbert space setting, such as regularized least-squares (Tikhonov scheme) and Landweber iteration. Such a general scheme has a long history in regularization theory. Starting from Evgeniou et al. [2] it is known that this can also profitably be used for learning from examples. Such use was also analyzed in Bauer et al. [3], Yao et al. [4], Gerfo et al. [5], Guo et al. [6], Zhou [7], where the corresponding regularization parameter was chosen *a priori*.

At the same time, it was observed in Caponnetto and De Vito [8], Caponnetto et al. [9] that, in contrast to deterministic regularization theory, the convergence or learning rates of regularized learning algorithms produced by a general regularization scheme is influenced not only by the smoothness of a target function, but also by the capacity of the hypothesis space given in terms of the effective dimension.

In the context of supervised learning, both the smoothness of the target function, as well as the effective dimension, depend on the unknown probability measure governing input-output relations. Therefore, any *a priori* choice of the regularization parameters, as this is discussed in the above-mentioned studies, cannot be effectively used in learning-from-examples.

For supervised learning algorithms originating from general regularization schemes this issue is discussed in several studies. De Vito et al. [10] deals only with the worst case behavior of the effective dimension. Caponnetto and Yao [11] presuppose that in addition to the training data one is provided with a sufficiently large number of input-output data for validation purposes.

We are interested in a choice of the regularization parameter that does not require any data splitting and allows us to use all available input-output data in the construction of regularized approximants. We shall employ the *balancing principle* that was used also in De Vito et al. [10], but this time we balance the unknown smoothness of a target function with the empirical effective dimension, and the balancing is performed in the empirical norm, only. In order to realize this approach we need to obtain novel error bounds in the empirical norm, as

well as in the norm of the hypothesis space. An interesting conclusion coming from our analysis is that previously considered parametrization of the effective dimension by power functions, see e.g. Caponnetto and De Vito [8], Caponnetto and Yao [11], Guo et al. [6] is too rough for hypothesis spaces generated by smooth kernels. Moreover, our numerical tests demonstrate an advantage of our approach compared with cross-validation based adaptation from Caponnetto and Yao [11], where part of the training data needs to be reserved only for validation. Note also that the present approach can be potentially combined with the idea Caponnetto et al. [9] to use unlabeled data for improving estimations of the effective dimension by the empirical one.

2. Setup, notation and behavior of the effective dimension

In this section we provide more details about the setup which is considered here, and we start with a discussion on typical behavior of the effective dimension.

2.1. Setup and notation

Within the context of learning from examples an input-output relation between $x \in X$ and $y \in Y$ is given by a joint probability distribution ρ on $X \times Y$. The distribution is only partially known through examples (training data) $z = \{z_i = (x_i, y_i),\ i = 1, 2, \ldots, n\}$. In the sequel we assume that the input space X is a compact domain or some manifold in $\mathbb{R}^d$, whereas, for simplicity, we let $Y \subset \mathbb{R}$ be a closed subset. Then the joint distribution ρ admits a disintegration $\rho(x, y) = \rho(y|x)\rho_X(x)$, where $\rho(y|x)$ is the conditional probability of y given x, and the distribution ρ_X is the marginal distribution for drawing $x \in X$. The goal is to establish, given training data z, an input-output relation $f_z \colon X \to Y$ which provides us with a good prediction $y = f_z(x)$ when a new input $x \in X$ is given.

The quality of a function $f \in L_2(X, \rho_X)$ to be a predictor for y from observation x is given by the functional (risk)

$$f \longrightarrow \mathcal{E}(f) = \int_{X \times Y} (y - f(x))^2 d\rho(x, y).$$

The unique minimizer (regression function) of this functional $\mathcal{E}$ is denoted by f_ρ, and it is called the target function. It is known that $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{\rho_X}$,

where we denote by $\|\cdot\|_{\rho_X}$ the standard norm in $L_2(X, \rho_X)$. Also, if the function, say f, does not depend on y we have that $\|f\|_{\rho_X} = \|f\|_\rho$.

To proceed further with practical learning, one needs a hypothesis space $\mathcal{H} \subset L_2(X, \rho_X)$, where a learning function f_z shall be taken from. Such a hypothesis space is usually chosen as a reproducing kernel Hilbert space $\mathcal{H} := \mathcal{H}_K$ in terms of a Mercer kernel $K : X \times X \to \mathbb{R}$ with $\kappa = \sqrt{\sup_{x \in X} K(x, x)}$.

Similar to Smale and Zhou [12], see also Lu and Pereverzev [13, Chapt. 4], we can define the continuous inclusion operator $I_K : \mathcal{H} \to L_2(X, \rho_X)$ and its adjoint $I_K^* : L_2(X, \rho_X) \to \mathcal{H}$. We furthermore introduce the covariance operator $T = I_K^* I_K : \mathcal{H} \to \mathcal{H}$ and an integral operator $L = I_K I_K^* : L_2(X, \rho_X) \to L_2(X, \rho_X)$ such that $\|T\|_{\mathcal{H} \to \mathcal{H}} = \|L\|_{L_2(X,\rho_X) \to L_2(X,\rho_X)} = \kappa$. Since the kernel is assumed to be bounded (by κ^2) the operators T and L are of trace class. For any function $f \in \mathcal{H}$ we can relate the RKHS $\mathcal{H}-$norm and $\rho-$norm by the following relation

$$\|f\|_\rho = \|\sqrt{T} f\|_{\mathcal{H}}, \tag{1}$$

which can be easily verified by a polar decomposition $I_K = U\sqrt{T}$ with a partial isometry U. Let $P : L_2(X, \rho_X) \to L_2(X, \rho_X)$ be the projection on the closure of the range of I_K in $L_2(X, \rho_X)$. Then, if $P f_\rho \in \text{Range}(I_k)$, the embedding equation

$$I_K f = f_\rho$$

is solvable and we can define its Moore-Penrose generalized solution $f^\dagger \in \mathcal{H}$, which is the best approximation of the target function $f_\rho \in L_2(X, \rho_X)$ by elements from $\mathcal{H}$ in $L_2(X, \rho)$. If $f_\rho \in \mathcal{H}$ then both functions coincide.

Along with both above operators T, L we shall also consider their empirical forms. Replacing ρ_X by the empirical measure $\rho_x = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$, we define the sampling operator $S_x : \mathcal{H} \to \mathbb{R}^n$ by $(S_x f)_i = f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}}$, $i = 1, \ldots, n$. Its adjoint operator $S_x^* : \mathbb{R}^n \to \mathcal{H}$ is similarly defined, and this further yields the empirical covariance operator $T_x = S_x^* S_x : \mathcal{H} \to \mathcal{H}$. Explicit forms of these operators can be found in Smale and Zhou [12] or Lu and Pereverzev [13, P.207] and we skip all these details. We just mention explicitly, that the empirical ρ_x-norm of a function $f \in L_2(X, \rho_X)$ is given by $\|f\|_{\rho_x} = \left(\frac{1}{n} \sum_{j=1}^n |f(x_j)|^2 \right)^{1/2} \left(= \left(\int |f(t)|^2 \, d\, \rho_x(t) \right)^{1/2} \right)$.

The error estimates which we will give next, depend on the *effective dimension* $\mathcal{N}(\lambda) = \mathcal{N}_T(\lambda),\ \lambda > 0$, of the covariance operator T, which increases as $\lambda \to 0$. Our analysis will use a *general regularization scheme* to obtain a family of reconstructions

$$f_z^\lambda = g_\lambda(T_x) S_x^* y, \tag{2}$$

where $\{g_\lambda\}$ is a family of operator functions with $0 < \lambda \leq \kappa$. Details for the regularization scheme as well as for properties of the effective dimension will be given, below. Then, under fairly general assumptions we shall derive the following error estimates. For $f^\dagger \in \mathcal{H}$ there will be an increasing continuous function $\varphi,\ \varphi(0) = 0$, such that with confidence at least $1 - \eta$ we have error estimates

$$\|f^\dagger - f_z^\lambda\|_\rho \leq C\sqrt{\lambda}\left(\varphi(\lambda) + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)\left(\log\frac{6}{\eta}\right)^3,$$
$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C\left(\varphi(\lambda) + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)\left(\log\frac{6}{\eta}\right)^2,$$

and in the empirical norm

$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C\sqrt{\lambda}\left(\varphi(\lambda) + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)\left(\log\frac{6}{\eta}\right)^3.$$

These bounds reveal the interesting feature that these are composed by some increasing function and a decreasing function in λ, respectively. This motivates an
80 *a posteriori* choice of the regularization parameter λ by the *balancing principle*, to be specified in detail, below.

2.2. Behavior of the effective dimension $\mathcal{N}(\lambda)$

As could be seen from the bounds given above, the effective dimension is an important ingredient to describe tight bounds for the error, see e.g. Caponnetto and De Vito [8]. For the covariance operator $T\colon \mathcal{H} \to \mathcal{H}$, given as $(Tf)(x) = \int k(x,y) f(y)\ \rho_X(dy)$, it is defined as

$$\mathcal{N}(\lambda) = \mathcal{N}_T(\lambda) := \mathrm{Tr}((\lambda I + T)^{-1} T), \quad \lambda > 0. \tag{3}$$

Since the operator T is trace class, the effective dimension is finite. In particular, this function is continuously decreasing from ∞ to 0. More properties are given

in Zhang [14], Lin et al. [15]. Since the operator T is not known often the following assumption on the polynomial decay of the effective dimension is made, and we refer to Caponnetto and De Vito [8, Def. 1], and Zhou [7, Theorem 2], for example.

Assumption 2.1 (polynomial decay)**.** *Assume that there exists an index* $\beta \in (0, 1]$ *such that*

$$\mathcal{N}(\lambda) \leq C_0^2 \lambda^{-\beta}, \quad \forall \lambda > 0.$$

Although this assumption provides us with polynomial rates we shall argue that this assumption is not reasonable when using RKHS based on smooth kernels, as e.g., radial basis functions (Gaussian kernels).

Indeed, the increase rate of the effective dimension (as $\lambda \to 0$) depends on the smoothness properties of the chosen kernel as well as on properties of the marginal density ρ_X. Specifically, for Gaussian kernels the RKHS spaces were described in Scovel et al. [16]. In Kühn [17] the author described the covering numbers of RKHS based on Gaussian kernels when the sampling density is uniform on the cube $[0, 1]^d$. These bounds easily extend to bounds on the decay of the singular numbers of the corresponding operator T, and it gives us an exponential decay for these. From this we deduce, see e.g. Blanchard and Mathé [18, Ex. 4], that the effective dimension will increase at a logarithmic rate. Therefore, we accompany Assumption 2.1 with

Assumption 2.2 (logarithmic decay)**.** *Assume that there exists a constant* $b > 0$ *such that*

$$\mathcal{N}(\lambda) \leq b \log(1/\lambda), \quad \forall \lambda > 0.$$

We shall exemplify this in numerical simulations for kernels of finite and infinite smoothness, based on the corresponding behavior for the empirical version, and for large sample sizes. Recall the representation of the empirical covariance operator for $u = \sum_{i=1}^n u_i K(x_i, \cdot)$ as

$$\begin{aligned} T_x u &= \frac{1}{n} \sum_{j=1}^n \left\langle \sum_{i=1}^n u_i K(x_i, \cdot), K(x_j, \cdot) \right\rangle K(x_j, \cdot) \\ &= \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^n u_i K(x_i, x_j) K(x_j, \cdot). \end{aligned}$$

We obtain the representation for the effective dimension $\mathcal{N}_{T_x}$ of the empirical operator T_x as

$$\mathcal{N}_{T_x}(\lambda) = \mathrm{Tr}(\lambda I + T_x)^{-1} T_x = \mathrm{Tr}\left((n\lambda I + \mathbb{K})^{-1}\mathbb{K}\right), \quad \lambda > 0, \tag{4}$$

where we denote by $\mathbb{K}$ the $n \times n$ matrix with components $K(x_i, x_j)$, $i, j = 1, \ldots, n$. For the subsequent discussion we choose different kernel functions on the unit interval $[0, 1]$ as

$$K_1(x,t) = xt + e^{-8(t-x)^2},$$
$$K_2(x,t) = \min\{x,t\} - xt.$$

First we highlight in Figure 1 that for large sample size, here for $n = 1000, 2000$ the empirical effective dimension $\mathcal{N}_{T_x}(\lambda)$ does not vary much, and hence is close to the effective dimension $\mathcal{N}(\lambda)$. The following result from Blanchard and Mücke

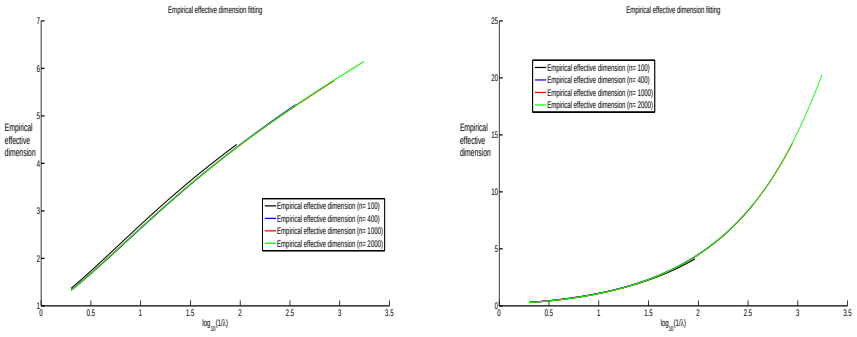

Figure 1: The empirical effective dimensions for kernel functions K_1 (left) and K_2 (right), and sample size $n = 100, 400, 1000, 2000$.

[19], communicated by the authors, is important; it improves previous estimate from Caponnetto et al. [9, Thm. 2], and it is given here with permission by the authors.

Theorem 2.3. *Blanchard and Mücke [19] Assume that $n\lambda \geq 4$. For any $0 < \eta < 1$, and letting $\delta := 2\log(4/\eta)/\sqrt{n\lambda}$ we have with probability $1 - \eta$ that*

$$(1-\delta)^{-1}\sqrt{\mathcal{N}(\lambda)} - \delta \leq \sqrt{\mathcal{N}_{T_x}(\lambda)} \leq (1+\delta)\sqrt{\mathcal{N}(\lambda)} + \delta.$$

Consequently, if $\delta^2 \leq \mathcal{N}_{T_x}(\lambda)$ then the bound

$$\mathcal{N}(\lambda) \leq 4(1+\delta)^2 \mathcal{N}_{T_x}(\lambda)$$

holds true.

Next we shall discuss whether a power-type behavior for the effective dimension seems reasonable. The definition of the effective dimension shows that the decay rate shall highly depend on the singular value of the empirical covariance operator T. To this end we consider the kernel functions K_1 and K_2 from above with t, x uniformly distributed in $[0, 1]$. We check the hypothesis that for these

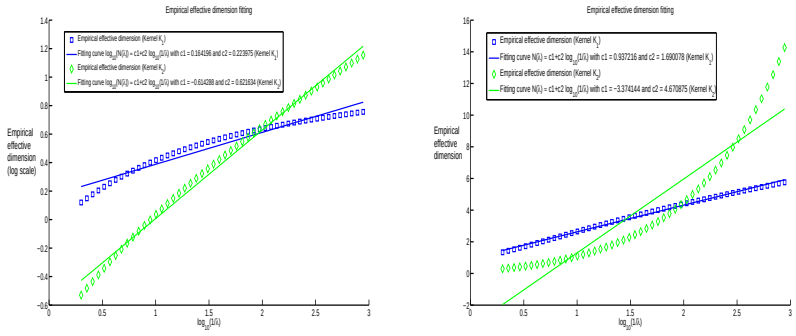

Figure 2: Fitting of the effective dimension for power-type and log-type behavior for the kernels K_1 (blue) and K_2 (green).

kernels the dependence on λ is of power-type or of log-type by drawing corresponding plots. In case of power-type the left panel in Figure 2 should exhibit straight lines. As can be seen, this seems to hold for the kernel K_2 (green), but this is less evident for the kernel K_1. In contrast, the right panel clearly shows that for the kernel K_2 the log-type behavior is violated, whereas for the kernel K_1 this seems to hold.

We draw the following conclusions. First, instead of estimating the exponent of the decay rate of the effective dimension it is recommended to directly use it. Moreover, for large enough sample size we may replace the effective dimension by the empirical counterpart. This is later used when proposing the adaptive choice of the regularization parameter in § 5.2. Finally, for smooth kernels, like K_1, a log-type behavior of the effective dimension is to be expected, and hence Assumption 2.2 is reasonable in such cases.

3. General regularization scheme and general source conditions

To have a stable recovery of the objective function $f^\dagger$, regularization is necessary to give some approximate reconstruction as presented in (2).

3.1. General regularization scheme

We first recall the definition of general regularization scheme in Mathé and Pereverzev [20] or Lu and Pereverzev [13, Def. 2.2].

Definition 3.1. Mathé and Pereverzev [20] A family g_λ, $0 < \lambda \leq \kappa$ is called regularization, if there are constants γ_0, $\gamma_{-1/2}$ and γ_{-1} for which

$$\sup_{0<\sigma\leq\kappa} \sqrt{\sigma}|g_\lambda(\sigma)| \leq \frac{\gamma_{-1/2}}{\sqrt{\lambda}}, \tag{5}$$

$$\sup_{0<\sigma\leq\kappa} |g_\lambda(\sigma)| \leq \frac{\gamma_{-1}}{\lambda}, \tag{6}$$

and

$$\sup_{0<\sigma\leq\kappa} |1 - \sigma g_\lambda(\sigma)| \leq \gamma_0, \tag{7}$$

such that

$$\sup_{0<\sigma\leq\kappa} |\sigma g_\lambda(\sigma)| \leq \gamma_0 + 1.$$

For the analysis it is convenient to introduce the residual function r_λ as

$$r_\lambda(\sigma) := 1 - \sigma g_\lambda(\sigma);$$

in particular we have that $|r_\lambda(\sigma)| \leq \gamma_0$, as seen from (7). The qualification of the regularization scheme, generated by $\{g_\lambda\}$, is the maximum value p for which

$$\sup_{0<\sigma\leq\kappa} |r_\lambda(\sigma)\sigma^p| \leq \gamma_p \lambda^p. \tag{8}$$

Direct computation shows that the qualification for standard (or iterative) Tikhonov regularization is $p = 1$ (or $p = m$ with m be the iteration number). The qualification of Landweber iteration, and truncated singular value decomposition (spectral cut-off) is $p = \infty$. For extended discussion, we refer to Lu and Pereverzev [13, Se.2.2]. The qualification has considerable impact on the error bounds.

3.2. General source conditions

To further establish error estimates we need to restrict our objective function $f^\dagger$ to a compact set with certain regularity properties, often called smoothness. The smoothness is then given in terms of a general source condition, generated by some index function.

Definition 3.2 (Index function)**.** A continuous non-decreasing function $\varphi\colon [0,\kappa] \to \mathbb{R}^+$ is called the *index function* if $\varphi(0) = 0$.

General source conditions are then given by assuming that

$$f^\dagger \in T_\varphi := \{f \in X : \quad f = \varphi(T)v, \quad \|v\|_{\mathcal{H}} \leq 1\}.$$

Moreover, the index function φ is covered (with some constant $0 < c \leq 1$) by the qualification p of the considered regularization scheme, if

$$c\frac{\lambda^p}{\varphi(\lambda)} \leq \inf_{\lambda \leq \sigma \leq \kappa} \frac{\sigma^p}{\varphi(\sigma)}, \quad 0 < \lambda \leq \kappa.$$

We shall focus on the following two classes $\mathcal{F}(d)$ and $\mathcal{F}_L(d)$ of index functions.

To this end we recall the concept of operator monotone functions, used in the context of learning in Bauer et al. [3] and De Vito et al. [10].

Definition 3.3 (Operator monotone, operator Lipschitz functions)**.** An index function $f\colon (0,\kappa) \to \mathbb{R}$ is called *operator monotone* if for any pair of non-negative self-adjoint operators A, B, $\|A\|_{\mathcal{H}\to\mathcal{H}}, \|B\|_{\mathcal{H}\to\mathcal{H}} \leq \kappa$ we have that $A \prec B$[1] implies that $f(A) \prec f(B)$. It is called *operator Lipschitz* if $\|f(A) - f(B)\|_{\mathcal{H}\to\mathcal{H}} \leq C\|A - B\|_{\mathcal{H}\to\mathcal{H}}$, for some constant C.

In these terms the smoothness classes are given as

$$\mathcal{F} := \{\varphi \text{ oper. monotone index function}\},$$

and

$$\mathcal{F}_L := \{\varphi = \vartheta\psi, \quad \psi \in \mathcal{F},\ \vartheta \ \text{ operator Lipschitz index function}\}.$$

We note that the decomposition $\varphi = \vartheta\psi$ is not unique and we can tune both functions ϑ, ψ allowing the Lipschitz constant for ϑ to equal 1. As one can

[1] For $A, B \in \mathcal{L}(\mathcal{H})$ we write $A \prec B$ if for all $x \in \mathcal{H}$ we have that $\langle Ax, x\rangle_{\mathcal{H}} \leq \langle Bx, x\rangle_{\mathcal{H}}$.

easily verify, if $r \in (0,1]$, the functions $t \to t^r$ belong to $\mathcal{F}$ covered by λ^p with $p = 1$ and they belong to $\mathcal{F}_L$ for $r \geq 1$.

We close the current section with the following result.

Lemma 3.4 (cf. Mathé and Pereverzev [21, Lem. 3])**.** *For each operator monotone index function φ there is a constant $1 \leq C < \infty$ such that $\varphi(t)/t \leq C\varphi(s)/s$, provided that $0 < s \leq t \leq \kappa$. Consequently, operator monotone index functions are covered by the qualification 1, hence $\sup_{0<\sigma\leq\kappa} |r_\lambda(\sigma)\varphi(\sigma)| \leq C\varphi(\lambda)$.*

4. Error estimates

In this section, we discuss error estimates for the $L_2(X, \rho_X)$, RKHS, and empirical norms between $f^\dagger$ and f_z^λ from (2), respectively. These norms will be important in the adaptive parameter choice presented in the Section 5.

Results, relating these different bounds, were first given in De Vito et al. [10, Prop. 1]. Here we provide a novel and enhanced form, with proof given in the appendix. We introduce the following (random) function

$$\Psi_{x,\lambda} := \|(\lambda I + T)^{-1/2}(T - T_x)\|_{\mathcal{H}\to\mathcal{H}}, \quad x \in X^n,\ \lambda > 0. \tag{9}$$

Proposition 4.1. *Assume that $\mathcal{H}$ is a RKHS with a bounded kernel. For any $f \in \mathcal{H}$ and a constant $\lambda > 0$ there holds*

$$\left| \|f\|_\rho^2 - \|f\|_{\rho_x}^2 \right| \leq \|(\lambda I + T)^{-1/2}(T - T_x)\|_{\mathcal{H}\to\mathcal{H}} \Big(\sqrt{\lambda}\|f\|_{\mathcal{H}} + \|f\|_\rho \Big) \|f\|_{\mathcal{H}}. \tag{10}$$

Consequently we have that

$$\|f\|_{\rho_x} \leq \|f\|_\rho + 2\max\left\{ \sqrt{\lambda}, \frac{1}{4}\Psi_{x,\lambda} \right\} \|f\|_{\mathcal{H}} \tag{11}$$

and

$$\frac{1}{\sqrt{2}}\|f\|_\rho \leq \|f\|_{\rho_x} + \left(\Psi_{x,\lambda} \left(\Psi_{x,\lambda} + \sqrt{\lambda} \right) \right)^{1/2} \|f\|_{\mathcal{H}}. \tag{12}$$

The latter bound (12) is more explicit (with high probability), cf. Corollary 4.7 below, and it will then provide order optimal (a priori) risk bounds.

We shall apply the above bounds to $f = f^\dagger - f_z^\lambda$. Additionally we introduce

$$\bar{f}_x^\lambda := g_\lambda(T_x) T_x f^\dagger \tag{13}$$

which has the following (defining) property.

Lemma 4.2. *Fox fixed $x_1, \ldots, x_n$ we denote by $\mathbf{E}_y$ be the expectation with respect to $(y_1, \ldots, y_n) \to \prod_{j=1}^n \rho(y_j|x_j)$. Then there holds $\mathbf{E}_y f_z^\lambda = \bar{f}_x^\lambda$.*

We thus have that

$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq \|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} + \|\bar{f}_x^\lambda - f_z^\lambda\|_{\mathcal{H}}, \tag{14}$$

and similar for the ρ- and ρ_x-norms, respectively. We call the first summand in the decomposition form (14) the approximation error, and the second summand the noise propagation error. Such a decomposition has been well analyzed both in the regularization theory Mathé and Pereverzev [22] and in learning theory De Vito et al. [10], Guo et al. [6].

4.1. Probabilistic estimates

Similar to the previous work in Smale and Zhou [12], Caponnetto and De Vito [8], Bauer et al. [3], Blanchard and Krämer [23], Guo et al. [6] we need some appropriate error estimates between the covariance operator T and its empirical form T_x. We collect the following results, cf. Caponnetto and De Vito [8], Bauer et al. [3], Blanchard and Krämer [23].

To this end it will be convenient to define the auxiliary functions $\mathcal{B}_{n,\lambda}$ and $\Upsilon(\lambda)$ (c.f. Guo et al. [6]) as

$$\mathcal{B}_{n,\lambda} := \frac{2\kappa}{\sqrt{n}} \left(\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{\mathcal{N}(\lambda)} \right), \quad \lambda > 0, \tag{15}$$

and

$$\Upsilon(\lambda) := \left(\frac{\mathcal{B}_{n,\lambda}}{\sqrt{\lambda}} \right)^2 + 1, \quad \lambda > 0, \tag{16}$$

which will prove useful in subsequent error estimates.

The probabilistic estimates are as follows. For any $\eta \in (0, 1)$, with confidence at least $1 - \eta$ we have that

$$\|T - T_x\|_{\mathcal{H}\to\mathcal{H}} \leq \frac{4\kappa^2}{\sqrt{n}} \log \frac{2}{\eta}, \tag{17}$$

$$\Psi_{x,\lambda} = \|(\lambda I + T)^{-1/2}(T - T_x)\|_{\mathcal{H}\to\mathcal{H}} \leq \mathcal{B}_{n,\lambda} \log \frac{2}{\eta}, \tag{18}$$

and

$$\|(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}} \leq 2\frac{M}{\kappa} \mathcal{B}_{n,\lambda} \log \frac{2}{\eta}. \tag{19}$$

We need an appropriate upper bound for $\|\lambda I+T)(\lambda I+T_x)^{-1}\|_{\mathcal{H}\to\mathcal{H}}$, cf. Guo et al. [6], and we have, with confidence at least $1-\eta$,

$$\Xi := \|(\lambda I + T)(\lambda I + T_x)^{-1}\|_{\mathcal{H}\to\mathcal{H}} \leq 2\left[\left(\frac{\mathcal{B}_{n,\lambda}\log\frac{2}{\eta}}{\sqrt{\lambda}}\right)^2 + 1\right]. \tag{20}$$

For the bound (19) to hold, we assume that there exists a constant $M > 0$, such that $|y| \leq M$ almost surely. As a consequence of the operator concavity of the function $t \mapsto t^{1/2}$, we obtain, see Blanchard and Krämer [23, Lem. A.7] that

$$\|(\lambda I + T)^{1/2}(\lambda I + T_x)^{-1/2}\|_{\mathcal{H}\to\mathcal{H}} \leq \Xi^{1/2}. \tag{21}$$

4.2. Error bounds

We start bounding the approximation errors and the propagation errors in the ρ- and $\mathcal{H}$-norms, in Propositions 4.3 and 4.4, respectively. The proofs will be given in the appendix.

Proposition 4.3 (Approximation error bound)**.** *Assume that $Pf_\rho \in \mathrm{Range}(I_K)$ and let $\bar{f}_x^\lambda$ be defined by (13).*

1. *If $f^\dagger \in T_\varphi$, $\varphi \in \mathcal{F}$ and if the regularization $g_\lambda(\sigma)$ has a qualification $p \geq 3/2$, then the approximation error is estimated as*

$$\|f^\dagger - \bar{f}_x^\lambda\|_\rho \leq C\sqrt{8(\gamma_0^2+\gamma_{3/2}^2)}\Xi^{3/2}\lambda^{1/2}\varphi(\lambda), \quad 0 < \lambda \leq \kappa,$$
$$\|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} \leq C(\gamma_0+\gamma_1)\Xi\varphi(\lambda), \quad 0 < \lambda \leq \kappa.$$

2. *If $f^\dagger \in T_\varphi$, $\varphi = \vartheta\psi \in \mathcal{F}_L$ and if the qualification of the regularization $g_\lambda(\sigma)$ covers $\vartheta(\lambda)\lambda^{3/2}$, then the approximation error is estimated as*

$$\|f^\dagger - \bar{f}_x^\lambda\|_\rho \leq C\sqrt{8(\gamma_\vartheta^2+\gamma_{\vartheta+3/2}^2)}\Xi^{3/2}\sqrt{\lambda}\varphi(\lambda) + \sqrt{\gamma_0^2+\gamma_{1/2}^2}\psi(\kappa)\Xi^{1/2}\sqrt{\lambda}\|T-T_x\|_{\mathcal{H}\to\mathcal{H}}$$

and

$$\|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} \leq C(\gamma_\vartheta+\gamma_{\vartheta+1})\Xi\varphi(\lambda) + \psi(\kappa)\gamma_0\|T-T_x\|_{\mathcal{H}\to\mathcal{H}}.$$

Concerning the propagation error, we adopt the corresponding result from Guo et al. [6].

Proposition 4.4 (Propagation error bound)**.** *Assume $Pf_\rho \in \mathrm{Range}(I_K)$ and let f_z^λ, $\bar{f}_x^\lambda$ defined by (2), (13). There holds*

$$\|\bar{f}_x^\lambda - f_z^\lambda\|_\rho \leq (\gamma_{-1} + \gamma_0 + 1)\Xi\|(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}},$$
$$\|\bar{f}_x^\lambda - f_z^\lambda\|_{\mathcal{H}} \leq (\gamma_{-1/2}^2 + \gamma_{-1}^2)^{1/2}\Xi^{1/2}\frac{1}{\sqrt{\lambda}}\|(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}}.$$

We summarize the error estimates in Propositions 4.3 and 4.4 in terms of the above functions $\mathcal{B}_{n,\lambda}$ from (15) and $\Upsilon(\lambda)$ from (16) as follows.

195 **Proposition 4.5.** *Assume that $Pf_\rho \in \mathrm{Range}(I_K)$, $\eta \in (0,1)$, and C a generic constant independent of n, λ and η.*

If $f^\dagger \in T_\varphi$, $\varphi \in \mathcal{F}$ and the regularization $g_\lambda(\sigma)$ has a qualification $p \geq 3/2$, then the total error, with confidence at least $1-\eta$, allows for error estimates

$$\|f^\dagger - f_z^\lambda\|_\rho \leq C\left(\Upsilon(\lambda)\mathcal{B}_{n,\lambda} + [\Upsilon(\lambda)]^{3/2}\sqrt{\lambda}\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^3,$$
$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C\left([\Upsilon(\lambda)]^{1/2}\frac{1}{\sqrt{\lambda}}\mathcal{B}_{n,\lambda} + \Upsilon(\lambda)\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^2,$$

and in the empirical norm

$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C\left(\Upsilon(\lambda)\mathcal{B}_{n,\lambda} + [\Upsilon(\lambda)]^{3/2}\sqrt{\lambda}\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^3.$$

If $f^\dagger \in T_\varphi$, $\varphi = \vartheta\psi \in \mathcal{F}_L$ and the qualification of the regularization $g_\lambda(\sigma)$ covers $\vartheta(\lambda)\lambda^{3/2}$, then the total error, with confidence at least $1-\eta$, allows for error estimates

$$\|f^\dagger - f_z^\lambda\|_\rho \leq C\left(\Upsilon(\lambda)\mathcal{B}_{n,\lambda} + [\Upsilon(\lambda)]^{1/2}\left(\frac{\lambda}{n}\right)^{1/2} + [\Upsilon(\lambda)]^{3/2}\sqrt{\lambda}\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^3,$$
$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C\left([\Upsilon(\lambda)]^{1/2}\frac{1}{\sqrt{\lambda}}\mathcal{B}_{n,\lambda} + \left(\frac{1}{n}\right)^{1/2} + \Upsilon(\lambda)\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^2,$$

and in the empirical norm

$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C\left(\Upsilon(\lambda)\mathcal{B}_{n,\lambda} + [\Upsilon(\lambda)]^{1/2}\left(\frac{\lambda}{n}\right)^{1/2} + [\Upsilon(\lambda)]^{3/2}\sqrt{\lambda}\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^3.$$

Proof. The bounds for the ρ- and $\mathcal{H}$-norms follow straightly after Propositions 4.3 and 4.4 by inserting the corresponding bounds (17)–(20), respectively. Notice that at present we may use the bound (11) in conjunction with the

bound (18), which gives

$$2\max\left\{\frac{1}{4}\Psi_{x,\lambda},\sqrt{\lambda}\right\} \le 2\sqrt{\lambda}\max\left\{1,\frac{1}{4}\frac{\mathcal{B}_{n,\lambda}}{\sqrt{\lambda}}\log(6/\eta)\right\}$$
$$\le \sqrt{\lambda\Upsilon(\lambda)}\max\left\{1,\frac{1}{4}\log(6/\eta)\right\}$$

and hence that

$$\frac{1}{2}\max\left\{\Psi_{x,\lambda},4\sqrt{\lambda}\right\}\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \le \sqrt{2\Upsilon(\lambda)\lambda}\|f^\dagger - f_z^\lambda\|_{\mathcal{H}}\log(6/\eta).$$

To obtain bounds in the empirical ρ_x-norm we use the above estimate with Proposition 4.1. Overall we obtain

$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \le C\left(\Upsilon(\lambda)\mathcal{B}_{n,\lambda} + [\Upsilon(\lambda)]^{3/2}\sqrt{\lambda}\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^3$$
$$+ \sqrt{2\Upsilon(\lambda)\lambda}\|f^\dagger - f_z^\lambda\|_{\mathcal{H}}\log(6/\eta)$$
$$\le C\left(\Upsilon(\lambda)\mathcal{B}_{n,\lambda} + [\Upsilon(\lambda)]^{3/2}\sqrt{\lambda}\varphi(\lambda)\right)\left(\log\frac{6}{\eta}\right)^3,$$

for a larger constant C, which gives the bound for the first case $\varphi \in \mathcal{F}$. The same reasoning also applies to obtain the bound for $\varphi \in \mathcal{F}_L$. □

The above general bounds can be refined once we shall assume a lower bound for the regularization parameter λ.

We recall the definition of the functions $\mathcal{B}_{n,\lambda}$, $\Upsilon(\lambda)$, and $\Psi_{x,\lambda}$ in (15), (16), and (9), respectively. The following is important.

Lemma 4.6. *There exists a λ_* satisfying $\mathcal{N}(\lambda_*)/\lambda_* = n$. For $\lambda_* \le \lambda \le \kappa$, there holds*

$$\mathcal{B}_{n,\lambda} \le \frac{2\kappa}{\sqrt{n}}\left(\sqrt{2}\kappa + \sqrt{\mathcal{N}(\lambda)}\right). \tag{22}$$

This yields

$$\Upsilon(\lambda) \le 1 + \left(4\kappa^2 + 2\kappa\right)^2 \tag{23}$$

and (for $n \ge \kappa$) also

$$\mathcal{B}_{n,\lambda}\left(\mathcal{B}_{n,\lambda} + \sqrt{\lambda}\right) \le (1+2\kappa)^4\min\left\{\lambda,\sqrt{\frac{\kappa}{n}}\right\}. \tag{24}$$

We postpone the proof to the appendix.

We stress that within the above range $\lambda_* \leq \lambda \leq \kappa$ the function $\Upsilon(\lambda)$ is bounded. This observation extends to $\lambda \geq c\lambda_*$, whenever $0 < c \leq 1$. Indeed, albeit the function $\mathcal{N}(\lambda)$ is decreasing, the companion $\lambda \mapsto \lambda\mathcal{N}(\lambda)$ is non-decreasing, see e.g. Lin et al. [15, Lem. 2.2]. Therefore we find for $0 < c \leq 1$ that

$$\mathcal{N}(c\lambda_*) = \frac{c\lambda_*\mathcal{N}(c\lambda_*)}{c\lambda_*} \leq \frac{\lambda_*\mathcal{N}(\lambda_*)}{c\lambda_*} = \frac{1}{c}\mathcal{N}(\lambda_*).$$

Before establishing the bounds for the overall error, we shall use the estimate from (24) for enhancing the bound (12).

Corollary 4.7. *With probability at least $1-\eta$ (for $\eta \leq 2/e$) we have*

$$\frac{1}{\sqrt{2}}\|f\|_\rho \leq \|f\|_{\rho_x} + \left[(1+2\kappa)^2 \log(2/\eta)\right] \min\left\{\sqrt{\lambda}, \left(\frac{\kappa}{n}\right)^{1/4}\right\} \|f\|_{\mathcal{H}}.$$

Remark 4.8. The above bound has close relation to the previous bound, as given in De Vito et al. [10, Prop. 1]: With probability at least $1-\eta$ and by assuming that $\lambda \geq n^{-1/2}$, we have for any $f \in \mathcal{H}$ that [2]

$$\|f\|_\rho \lesssim \|f\|_{\rho_x} + \frac{C(\eta)}{n^{1/4}}\|f\|_{\mathcal{H}} \tag{25}$$

with $C(\eta) > \max\{\log(2/\eta)^{1/4}, 1\}$. As one can see, the second regime in the bound from Corollary (25) is used. The corollary is based on (24). From its proof we saw that this part results from the case that the parameter λ is larger than $\sqrt{\kappa/n}$, which ignores the role of the effective dimension.

Thus it is interesting to discuss when the optimal parameter λ_{apriori} defined in (26) below obeys $\lambda_{\text{apriori}} < n^{-1/2}$, in which case the new bound from Corollary 4.7 is superior. Looking at the balancing condition (26) this is the case when

$$\lambda_{\text{apriori}} \prec \frac{1}{\sqrt{n}} = \frac{\varphi(\lambda_{\text{apriori}})\sqrt{\lambda}_{\text{apriori}}}{\mathcal{N}(\lambda_{\text{apriori}})},$$

where the symbol $\prec$ means that $\lambda_{\text{apriori}} = \lambda_{\text{apriori}}(n)$ decays to zero faster than the right hand side. Under power type smoothness $\varphi(\lambda) = \lambda^{r-1/2}$, and under Assumption 2.1 (with exponent β) this means $2r+\beta < 2$. Thus for low smoothness and fast decay of the singular numbers of the operator T the previous

[2] Actually, in that study the roles of the ρ- and ρ_x-norms are interchanged. But the proof uses the bound (17), and the roles of ρ_x- and ρ-norms are converted.

bounds yield sub-optimal reconstruction rates. Under the bound (25) optimal rates can only be guaranteed for $r \in [1/2, 1-\beta/2]$. The latter is in accordance with the results obtained in De Vito et al. [10, Thm. 3 and examples], where the exponent $r - 1/2$, here, corresponds to r in that study.

We recall the definition of λ_* from Lemma 4.6. For $\lambda \geq \lambda_*$ we can refine Proposition 4.5 by using Lemma 4.6, i.e., replacing $\mathcal{B}_{n,\lambda}$ and $\Upsilon(\lambda)$ by the corresponding bounds.

Theorem 4.9. *Assume that $Pf_\rho \in \mathrm{Range}(I_K)$, and that $\lambda \geq \lambda_*$. There is a generic constant C, independent of n, λ, such that the following estimates hold for any $0 < \eta < 1$.*

If $f^\dagger \in T_\varphi$ with $\varphi \in \mathcal{F}$ or $\varphi = \vartheta\psi \in \mathcal{F}_L$, and the regularization $g_\lambda(\sigma)$ has a qualification $p \geq 3/2$, or if its qualification covers $\vartheta(\lambda)\lambda^{3/2}$, correspondingly, then with confidence at least $1-\eta$ the total error allows for estimates

$$\|f^\dagger - f_z^\lambda\|_\rho \leq C\sqrt{\lambda}\left(\varphi(\lambda) + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)\left(\log\frac{6}{\eta}\right)^3,$$

$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C\left(\varphi(\lambda) + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)\left(\log\frac{6}{\eta}\right)^2,$$

and in the empirical norm

$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C\sqrt{\lambda}\left(\varphi(\lambda) + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)\left(\log\frac{6}{\eta}\right)^3.$$

Proof. For $\lambda \geq \lambda_*$ we use Lemma 4.6. Inserting these bounds into the estimates from Proposition 4.5 we find the upper bounds as given. □

Remark 4.10. We highlight the role of both summands in $\varphi(\lambda)+\sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}$ from the view point of regularization theory for statistical inverse problems $Y^\sigma = Tx+\sigma\xi$ under Gaussian white noise. In [18, Prop. 1] the error bound

$$\left(\mathbb{E}\|f^\dagger - f_\lambda^\sigma\|_{\mathcal{H}}^2\right)^{1/2} \leq \|f^\dagger - f_\lambda\|_{\mathcal{H}} + \sqrt{2}\gamma_{-1}\sigma\sqrt{\frac{\mathcal{N}(\lambda)}{\lambda}}$$

was given (There is no discretization considered in [18] and we freely adopt that formalism to the present context). Of course, the first summand corresponds to the bias, and it depends on the underlying smoothness assumption, and it is bounded by $\varphi(\lambda)$. The second one corresponds to the standard deviation in the bias-variance decomposition in statistical inverse problems, if the noise level

is $\sigma = 1/\sqrt{n}$. To get this standard deviation small as $n \to \infty$ the restriction $\lambda \geq \sigma^2 = \frac{1}{n}$ is a 'natural' (minimal) consequence.

5. Choice of the regularization parameter

We shall discuss several choices of the regularization parameter λ. First, as this is standard, we derive the rates inferred from Theorem 4.9. Then we propose a novel *a posteriori* choice, based on the balancing principle, see e.g. De Vito et al. [10], Lu and Pereverzev [13]. Finally, in the numerical simulations we compare the newly proposed parameter choice with the cross-validation type approach from Caponnetto and Yao [11].

5.1. A Priori choice of the regularization parameter

The above general bounds result in the following *a priori* error bounds, i.e., when the effective dimension and the smoothness are known. Indeed, the upper bounds in Theorem 4.9 reveal that these are sums of the increasing function $\varphi(\lambda)$ and the decreasing function $\lambda \mapsto \sqrt{\mathcal{N}(\lambda)/(\lambda n)}$. The *a priori* choice of the parameter λ thus aims at minimizing, precisely 'balancing', both components with a λ_{apriori} such that

$$\varphi(\lambda_{\text{apriori}}) = \sqrt{\mathcal{N}(\lambda_{\text{apriori}})/(n\lambda_{\text{apriori}})}. \tag{26}$$

It is important to stress that the parameter λ_{apriori} is within the range required for the application of Theorem 4.9, for instance, $\lambda_{\text{apriori}} \geq c\lambda_*$ with a specific constant c defined below. Indeed, if $\varphi(\kappa) \leq 1$ we consider

$$\frac{\lambda_*}{\mathcal{N}(\lambda_*)} = \frac{1}{n} = \frac{\lambda_{\text{apriori}}}{\mathcal{N}(\lambda_{\text{apriori}})}\varphi^2(\lambda_{\text{apriori}}) \leq \frac{\lambda_{\text{apriori}}}{\mathcal{N}(\lambda_{\text{apriori}})}\varphi^2(\kappa) \leq \frac{\lambda_{\text{apriori}}}{\mathcal{N}(\lambda_{\text{apriori}})},$$

and the monotonicity of the function $\lambda \mapsto \lambda/\mathcal{N}(\lambda)$ yields that $\lambda_* \leq \lambda_{\text{apriori}}$. Otherwise, if $c := \left(\varphi^2(\kappa)\right)^{-1} < 1$, then we argue

$$\frac{\lambda_{\text{apriori}}}{\mathcal{N}(\lambda_{\text{apriori}})} \geq c\frac{1}{n} = c\frac{\lambda_*}{\mathcal{N}(\lambda_*)} \geq \frac{c\lambda_*}{\mathcal{N}(c\lambda_*)}.$$

Again, the monotonicity of the function $\lambda \mapsto \lambda/\mathcal{N}(\lambda)$ yields that $\lambda_{\text{apriori}} \geq c\lambda_*$.

Under Assumption 2.1 the balancing of error estimates in Theorem 4.9, precisely the corresponding upper bounds, allows for the following learning rates, see also Caponnetto and De Vito [8, Thm. 1] for similar results for penalized

least squares, and Guo et al. [6, Cor. 1] for general regularization but restricted to the ρ-norm.

Corollary 5.1. *Let the assumptions in Theorem 4.9 be satisfied. Suppose that smoothness is given by an index function $\varphi(\lambda) = \lambda^{r-1/2}$ with $r \geq 1/2$, and that the effective dimension obeys Assumption 2.1. Choose $\lambda = n^{-\frac{1}{2r+\beta}}$, then with confidence at least $1-\eta$, there holds*

$$\|f^\dagger - f_z^\lambda\|_\rho \leq C n^{-\frac{r}{2r+\beta}} \left(\log \frac{6}{\eta}\right)^3,$$
$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C n^{-\frac{r-1/2}{2r+\beta}} \left(\log \frac{6}{\eta}\right)^2,$$
$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C n^{-\frac{r}{2r+\beta}} \left(\log \frac{6}{\eta}\right)^3.$$

For logarithmic decay of the effective dimension, i.e., under Assumption 2.2 the corresponding results are as follows.

Corollary 5.2. *Let the assumptions in Theorem 4.9 be satisfied. Suppose that smoothness is given by an index function $\varphi(\lambda) = \lambda^{r-1/2}$ with $r \geq 1/2$, and that the effective dimension obeys Assumption 2.2. If $\lambda \asymp \left(\frac{\log n}{n}\right)^{1/(2r)}$ then with confidence at least $1-\eta$, there holds*

$$\|f^\dagger - f_z^\lambda\|_\rho \leq C \sqrt{\frac{\log n}{n}} \left(\log \frac{6}{\eta}\right)^3,$$
$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C \left(\frac{\log n}{n}\right)^{(r-1/2)/(2r)} \left(\log \frac{6}{\eta}\right)^2,$$
$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C \sqrt{\frac{\log n}{n}} \left(\log \frac{6}{\eta}\right)^3.$$

These bounds show that the learning rates are almost parametric in the ρ and ρ_x-norms, respectively. Actually, the bound in the ρ-norm is order optimal, and we briefly sketch the corresponding concept.

As in DeVore et al. [24, Sect. 3] we consider the *minimax rate* of learning over a set $\Theta \subset L_2(X, \rho_X)$ as

$$e_m(\Theta, \rho_X) := \inf_{\hat{f}} \sup_{f_\rho \in \Theta} \mathbb{E}_{\rho^m} \|\hat{f} - f_\rho\|_\rho, \tag{27}$$

where the infimum is taken over arbitrary learning algorithms. Then we have the following result. Notice that the validity of Assumption 2.2 holds for covariance operators T with exponential decay of the singular numbers, say $t_n \asymp e^{-\gamma n}$, $n = 1, 2, \ldots$.

Theorem 5.3. *Suppose that the smoothness index function φ increases with at most polynomial rate, and that ρ_X is such that the covariance operator T has exponentially decaying singular numbers. Then*

$$e_m(T_\varphi, \rho_X) \geq c\sqrt{\frac{\log m}{m}}.$$

The proof of this result relies on DeVore et al. [24, Thm. 3.1], and the construction given in [8]. We briefly sketch the arguments in the appendix.

5.2. Adaptive choices of the regularization parameter

In this section, we propose a modification of the parameter choice based on the balancing principle, as outlined in De Vito et al. [10] or Lu and Pereverzev [13, Ch.4.3]. Theorem 4.9 reveals that we need to balance the terms $\varphi(\lambda)$ and $\sqrt{\mathcal{N}(\lambda)/(n\lambda)}$ in the RKHS $\mathcal{H}$-norm, while in the ρ_x-norm this is to be done for the terms $\sqrt{\lambda}\varphi(\lambda)$ and $\sqrt{\mathcal{N}(\lambda)/n}$. We observe the following. First, in both cases the first term is increasing while the second one is decreasing as a function of the parameter λ. Furthermore, in both cases we can compute norm differences $\|f_z^\lambda - f_z^{\lambda'}\|$ for different parameters λ, λ'. Indeed, for a function $f = \sum_{i=1}^n \alpha_i K(x_i, \cdot) \in \mathcal{H}$ we have

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 &= \left\langle \sum_{i=1}^n \alpha_i K(x_i, \cdot), \sum_{i=1}^n \alpha_i K(x_i, \cdot) \right\rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j)\end{aligned}$$

by the reproducing property of $K(x, s)$. In particularly, by choosing $f_z^\beta = \sum_{i=1}^n \alpha_i^\beta K(x_i, \cdot)$ and $f_z^\lambda = \sum_{i=1}^n \alpha_i^\lambda K(x_i, \cdot)$ we obtain the RKHS norm

$$\|f_z^\beta - f_z^\lambda\|_{\mathcal{H}}^2 = (\alpha^\beta - \alpha^\lambda)\mathbb{K}(\alpha^\beta - \alpha^\lambda)^T$$

and in the empirical ρ_x-norm

$$\|f_z^\beta - f_z^\lambda\|_{\rho_x}^2 = \frac{1}{n}(\alpha^\beta - \alpha^\lambda)\mathbb{K}^2(\alpha^\beta - \alpha^\lambda)^T,$$

where $\mathbb{K}$ is the $n \times n$ matrix composed from $K(x_i, x_j)$, $i, j = 1, \ldots, n$. As one can observe, both, the $\mathcal{H}$- and the ρ_x-norms are computable if the kernel function $K(x, s)$ is known explicitly.

We turn to the implementation of the adaptive choice rule. To this end we choose a parameter $\lambda_{\text{start}} \leq c\lambda_*$ together with a natural number N and a spacing $\mu > 1$.

Then we consider the grid

$$\Delta_N := \left\{\lambda_i, \quad \lambda_i = \lambda_{\text{start}}\mu^i,\ i = 1, \dots, N\right\}. \tag{28}$$

If $N \geq \lceil \frac{\log(\kappa/\lambda_{\text{start}})}{\log(\mu)} \rceil$ then $\lambda_N \geq \kappa$, such that the whole range of potential values λ is covered.

We choose the regularization parameter either according to

$$\lambda_+ = \max\left\{\lambda_i : \|f_z^{\lambda_i} - f_z^{\lambda_j}\|_\circ \leq 4\mathcal{S}(n, \eta, \lambda_j),\ j = 1, \dots, i\right\} \tag{29}$$

or

$$\bar{\lambda} = \max\left\{\lambda_i : \|f_z^{\lambda_j} - f_z^{\lambda_{j-1}}\|_\circ \leq 4\mathcal{S}(n, \eta, \lambda_{j-1}),\ j = 1, \dots, i-1\right\}, \tag{30}$$

where $\circ \in \{\mathcal{H}, \rho_x\}$, and with $\mathcal{S}(n, \eta, \lambda) = C\sqrt{\mathcal{N}(\lambda)/(n\lambda)}$ in the RKHS $\mathcal{H}$-norm, and $\mathcal{S}(n, \eta, \lambda) = C\sqrt{\mathcal{N}(\lambda)/n}$ in the ρ_x-norm, respectively. Then De Vito et al. [10, Thm.1-2], or Lu and Pereverzev [13, Prop.4.5-4.6] show that both above parameter choice rules provide order optimal error estimates. We formulate the result for this adaptive choice of the regularization parameter. To this end we introduce the following function $\theta = \theta_{\mathcal{N},\varphi}$, resulting from the required balancing, see (26), as

$$\theta_{\mathcal{N},\varphi}(\lambda) := \frac{\sqrt{\lambda}\varphi(\lambda)}{\sqrt{\mathcal{N}(\lambda)}}, \quad \lambda > 0. \tag{31}$$

This is an increasing continuous function with $\lim_{\lambda\to 0}\theta_{\mathcal{N},\varphi}(\lambda) = 0$ (index function), and that this function is identically given both in case that $\circ = \mathcal{H}$ or $\circ = \rho_x$.

Theorem 5.4. *Let the assumptions in Theorem 4.9 be satisfied. If the regularization parameter λ is chosen within the grid (28) as $\lambda = \lambda_+$ by (29), or as $\lambda = \bar{\lambda}$ by (30), respectively, then with confidence at least $1 - \eta$, there holds*

$$\|f^\dagger - f_z^\lambda\|_{\mathcal{H}} \leq C\varphi(\theta^{-1}_{\mathcal{N},\varphi}(n^{-1/2}))\left(\log\frac{6}{\eta}\right)^2,$$

$$\|f^\dagger - f_z^\lambda\|_{\rho_x} \leq C\sqrt{\theta^{-1}_{\mathcal{N},\varphi}(n^{-1/2})}\varphi(\theta^{-1}_{\mathcal{N},\varphi}(n^{-1/2}))\left(\log\frac{6}{\eta}\right)^3,$$

with a generic constant C.

Actually, adaptive risk bounds (in the ρ-norm) are also possible, and we sketch the reasoning. We let $\lambda_{+,\mathcal{H}}$ and λ_{+,ρ_x} (and likewise $\bar{\lambda}_{\mathcal{H}}$, $\bar{\lambda}_{\rho_x}$) the parameter choices according to both cases in (29) or (30). Then we define $\hat{\lambda}_+ = \min\{\lambda_{+,\rho_x}, \lambda_{+,\mathcal{H}}\}$, and similarly $\hat{\bar{\lambda}} := \min\{\bar{\lambda}_{\rho_x}, \bar{\lambda}_{\mathcal{H}}\}$. It is straight forward to check that the Assumptions 4.2 of Proposition 4.7 in Lu and Pereverzev [13] are fulfilled, and hence based on the bound in Corollary 4.7 both choices (with high probability) result in error bounds (with $\hat{\lambda}$ one of the above choices)

$$\|f^\dagger - f_z^{\hat{\lambda}}\|_\rho \leq C\left(\sqrt{\lambda_0(n)}\varphi(\lambda_0(n)) + n^{-1/4}\varphi(\lambda_0(n))\right)\left(\log\frac{6}{\eta}\right)^3,$$

where $\lambda_0(n) = \theta_{\mathcal{N},\varphi}^{-1}(n^{-1/2})$. As it was exploited in [13], if the true parameter obeys $\lambda_0(n) \geq cn^{-1/2}$, which is, for example, the case under the conditions of Corollary 5.2 with $r \geq 1$, then this yields order optimality. In the low smoothness case, see the discussion in Remark 4.8, the above choices yield convergence, however at a suboptimal rate $n^{-1/4}\varphi(\lambda_0(n))\left(\log\frac{6}{\eta}\right)^3$. It is not clear to the authors whether an oracle type parameter choice can be derived by the bound from Corollary 4.7 in the low smoothness regime.

6. Numerical simulation

In this section we consider some numerical simulation verifying the advantage of the novel error bound based on the effective dimension $\mathcal{N}(\lambda)$, in particular its alternative empirical form $\mathcal{N}_{T_x}(\lambda)$. Especially, for a small sample size it allows for a better approximation than the cross-validation based rule from Caponnetto and Yao [11], which is also described below.

6.1. Numerical parameter choice rule by using the empirical effective dimension

In all tests, the target function is fixed as in Micchelli and Pontil [25], and hence given by

$$f_\rho(x) = \frac{1}{10}\left(x + 2\left(e^{-8(\frac{4}{3}\pi - x)^2} - e^{-8(\frac{\pi}{2} - x)^2} - e^{-8(\frac{3}{2}\pi - x)^2}\right)\right),\ x \in [0, 2\pi]. \quad (32)$$

The training set $z = \{x_i, y_i\}_{i=1}^n$ consists of independent and uniformly distributed x_i in the interval $[0, 2\pi]$, the responses $y_i = f_\rho(x_i) + \zeta_i$ are given with independent random noise ζ_i uniformly distributed in the interval $[-0.05, 0.05]$. The RKHS $\mathcal{H}$ is generated by the kernel function $K(x, s) = xt + e^{-8(t-x)^2}$ with

$t, x \in [0, 2\pi]$, such that the target function f_ρ from (32) belongs to the chosen RKHS $\mathcal{H}$.

Following the discussion in Section 2.1, we numerically verify the adaptive parameter choice rule by directly using the empirical effective dimension $\mathcal{N}_{T_x}(\lambda_i)$ for $i = 1, \dots N$. Specifically, we replace the effective dimension $\mathcal{N}(\lambda)$ by the empirical form $\mathcal{N}_{T_x}(\lambda)$, and hence we consider the following form of the regularization parameter $\tilde{\lambda}$ according to

$$\tilde{\lambda} = \max \left\{ \lambda_i : \|f_z^{\lambda_j} - f_z^{\lambda_{j-1}}\|_{\rho_x} \leq M \sqrt{\frac{\mathcal{N}_{T_x}(\lambda_j)}{n}}, j = 1, \dots, i-1 \right\}.$$

Remark 6.1. We need to tune the constant M in front of the above right hand side. Actually, when deriving the asymptotic behaviors in Theorem 4.9 we have considered upper bounds, especially for the constant. This is enough for asymptotical considerations, i.e., for large sample size. In practical computations such bounds may be too rough. As it appears in the practical implementation, the factor $M = \kappa^{-2}$ decreases the influence of the upper bound constants. Notice, that we can easily approximate κ from the sample by taking $\kappa \approx \kappa_x = \mathrm{Tr}(T_x)$.

The results of the simulation are shown in Figure 3.

6.2. Comparison with cross-validation

We shall compare the current approach with the cross-validation type parameter choice rule in Caponnetto and Yao [11]. More precisely, we equally split the data set into a learning set z_{train} and a validation set z_{vali} with $z_{\text{train}} = \{z_{i,\text{train}}\} = \{(x_{i,\text{train}}, y_{i,\text{train}})\}$ and $z_{\text{vali}} = \{z_{i,\text{vali}}\} = \{(x_{i,\text{vali}}, y_{i,\text{vali}})\}$. The set z_{train} is used to provide approximating functions $f^\lambda_{z_{\text{train}}}$ where λ is taken from the regularization parameter set (28). The regularization parameter is chosen as below

$$\hat{\lambda} = \arg\min_{\Delta_N} \frac{1}{|z_{\text{vali}}|} \sum_{i=1}^{|z_{\text{vali}}|} \left(f^\lambda_{z_{\text{train}}}(x_{i,\text{vali}}) - y_{i,\text{vali}} \right)^2$$

where $|z_{\text{vali}}|$ is the sample size of the validation set z_{vali}. The comparison with the sampling sizes $n = 30, 100, 1000, 2000$ is displayed in Figure 3. In almost all examples the adaptive parameter choice rule which uses the empirical effective dimension outperforms adaptive cross-validation.

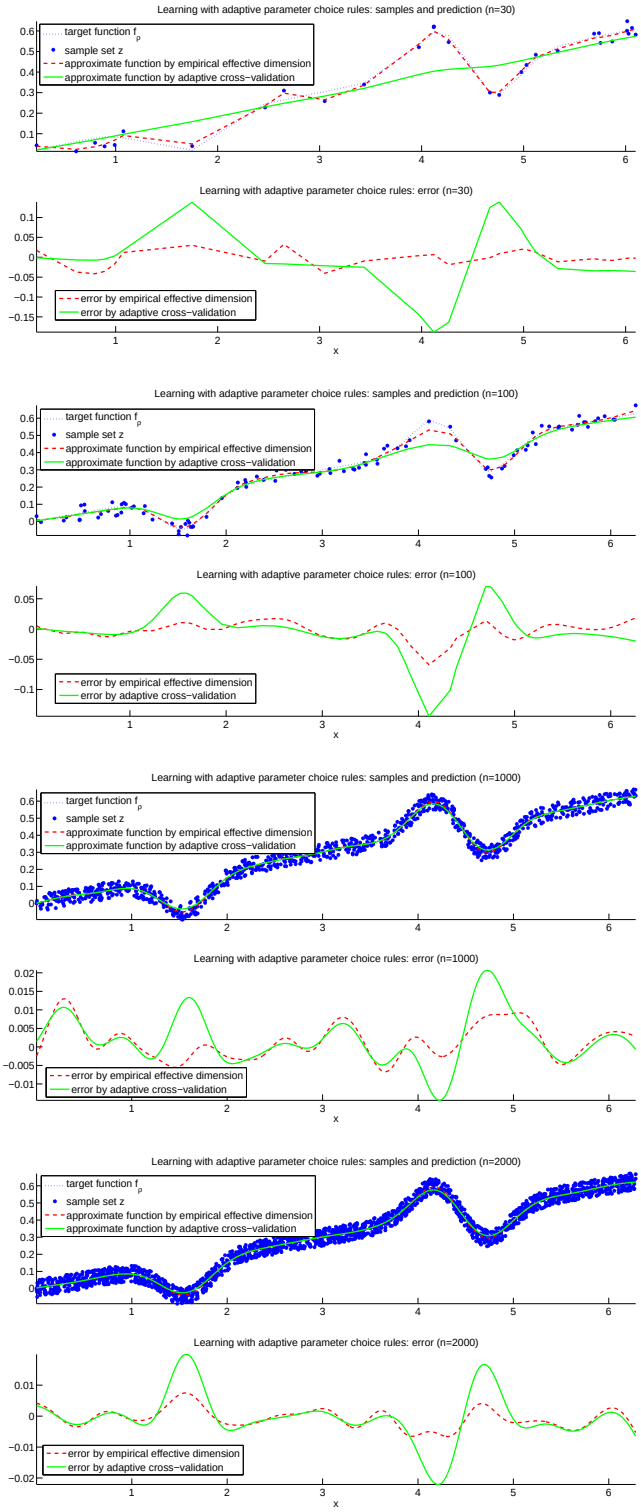

Figure 3: Comparison between the adaptive parameter choice rule with discrete empirical effective dimension and adaptive cross-validation for different sampling sizes $n = 30, 100, 1000, 2000$.

Acknowledgments

This work started when the first two authors visited S. V. Pereverzev at the Johann Radon Institute for Computational and Applied Mathematics (RICAM) in 2016. Both authors would like to thank him for the invitation and kind hospitality. S. Lu is supported by NSFC (no.11522108,91630309), Special Funds for Major State Basic Research Projects of China (2015CB856003) and Shanghai Municipal Education Commission (no.16SG01). S. V. Pereverzev is partially supported by the Austrian Science Foundation (FWF), project I1669, and by the EU-Horizin 2020 MSC-RISE project AMMODIT.

7. Appendix

We first collect a few general bounds which can be obtained from the spectral theory of non-negative self-adjoint operators A, $\|A\|_{\mathcal{H}\to\mathcal{H}} \le \kappa$. First, for any continuous (measurable) function $m\colon [0,\kappa] \to \mathbb{R}$ we find that

$$\|m(A)\,(\lambda I + A)^{1/2}\,v\|_{\mathcal{H}} = \left(\lambda\|m(A)v\|_{\mathcal{H}}^2 + \|m(A)A^{1/2}v\|_{\mathcal{H}}^2\right)^{1/2}. \tag{A.33}$$

This can be seen from

$$\begin{aligned}\|m(A)\,(\lambda I + A)^{1/2}\,v\|_{\mathcal{H}}^2 &= \langle m(A)\,(\lambda I + A)^{1/2}\,v, m(A)\,(\lambda I + A)^{1/2}\,v\rangle_{\mathcal{H}}\\ &= \langle(\lambda I + A)\,m(A)v, m(A)v\rangle_{\mathcal{H}}\\ &= \lambda\|m(A)v\|_{\mathcal{H}}^2 + \|m(A)A^{1/2}v\|_{\mathcal{H}}^2,\end{aligned}$$

where we have implemented the commuting properties of $m(A)$ and $\lambda I + A$ (or A). Later, we shall apply this to operators T and T_x, respectively.

We start using (A.33) to enhance the bounds from Guo et al. [6].

Lemma 7.1.

1. *Let g_λ be any regularization with qualification one. Then for any compact self-adjoint operator A, $\|A\|_{\mathcal{H}\to\mathcal{H}} \le \kappa$, there holds*

$$\|r_\lambda(A)(\lambda I + A)\|_{\mathcal{H}\to\mathcal{H}} \le (\gamma_0 + \gamma_1)\lambda, \tag{A.34}$$

and

$$\|r_\lambda(A)(\lambda I + A)^{1/2}\|_{\mathcal{H}\to\mathcal{H}} \le \sqrt{\gamma_0^2 + \gamma_{1/2}^2}\sqrt{\lambda}. \tag{A.35}$$

2. *If g_λ has qualification at least $3/2$ then*

$$\|r_\lambda(A)(\lambda I + A)^{3/2}\|_{\mathcal{H}\to\mathcal{H}} \le \sqrt{8(\gamma_0^2 + \gamma_{3/2}^2)}\lambda^{3/2}.$$

The estimate (A.34) relates arbitrary regularization with qualification at least equal to 1 to Tikhonov regularization. Item 2 will be used, when bounding estimates in the ρ-norm.

Proof of Lemma 7.1. The first assertion is seen from the triangle inequality directly. For (A.35) we use (A.33) with $m(A) := r_\lambda(A)$ to derive

$$\|r_\lambda(A)(\lambda I + A)^{1/2}v\|_{\mathcal{H}\to\mathcal{H}} = \left(\lambda\|r_\lambda(A)v\|_{\mathcal{H}}^2 + \|r_\lambda(A)A^{1/2}v\|_{\mathcal{H}}^2\right)^{1/2},$$

and the result follows from Definition 3.1 and (8).

To prove Item 2, we observe, by the Hölder's inequality, that $(a+b)^t \le 2^t\left(a^t + b^t\right),\ a, b \ge 0,\ t > 0$. Let $(t_j, u_j),\ j = 1, 2\ldots$ be the singular value decomposition for the operator A, i.e., $Au_j = s_j u_j,\ j = 1, 2, \ldots$. We obtain that

$$\begin{aligned}\langle(\lambda I + A)^3 u_j, u_j\rangle_{\mathcal{H}} = (\lambda + s_j)^3 &\le 8\left(\lambda^3 + s_j^3\right)\\ &= 8\langle\left(\lambda^3 I + A^3\right)u_j, u_j\rangle_{\mathcal{H}}, \quad j = 1, 2, \ldots.\end{aligned}$$

By linearity of A this extends to arbitrary $u \in \mathcal{H}$, hence we have that

$$\langle(\lambda I + A)^3 u, u\rangle_{\mathcal{H}} \le 8\langle\left(\lambda^3 I + A^3\right)u, u\rangle_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

We then let $u\colon = r_\lambda(A)^2 v,\ \|v\|_{\mathcal{H}} \le 1$ to see that

$$\begin{aligned}\|r_\lambda(A)(\lambda I + A)^{3/2}v\|_{\mathcal{H}\to\mathcal{H}}^2 &= \langle(\lambda I + A)^3 r_\lambda(A)v, r_\lambda(A)v\rangle_{\mathcal{H}}\\ &\le 8\langle\left(\lambda^3 I + A^3\right)r_\lambda(A)v, r_\lambda(A)v\rangle_{\mathcal{H}}\\ &= 8\left(\lambda^3\|r_\lambda(A)v\|_{\mathcal{H}}^2 + \|r_\lambda(A)A^{3/2}v\|_{\mathcal{H}}^2\right)\\ &\le 8\left(\lambda^3\gamma_0^2 + \gamma_{3/2}^2\lambda^3\right)\|v\|_{\mathcal{H}}^2\\ &= 8\left(\gamma_0^2 + \gamma_{3/2}^2\right)\lambda^3\|v\|_{\mathcal{H}}^2,\end{aligned}$$

which proves the second assertion. □

Proof of Proposition 4.1. Notice that for any $f \in \mathcal{H}$ we have

$$\begin{aligned}\left|\|f\|_\rho^2 - \|f\|_{\rho_x}^2\right| &= \langle(T - T_x)f, f\rangle\\ &\le \|(T - T_x)f\|_{\mathcal{H}}\,\|f\|_{\mathcal{H}}.\end{aligned}$$

Now we apply (A.33) for $m(T) = I$ to obtain that

$$\begin{aligned}\|(T - T_x)f\|_{\mathcal{H}} &= \|(T - T_x)(\lambda I + T)^{-1/2}(\lambda I + T)^{1/2} f\|_{\mathcal{H}} \\ &\leq \|(T - T_x)(\lambda I + T)^{-1/2}\|_{\mathcal{H}\to\mathcal{H}} \|(\lambda I + T)^{1/2} f\|_{\mathcal{H}} \\ &\leq \|(T - T_x)(\lambda I + T)^{-1/2}\|_{\mathcal{H}\to\mathcal{H}} \left(\lambda \|f\|_{\mathcal{H}}^2 + \|\sqrt{T} f\|_{\mathcal{H}}^2\right)^{1/2} \\ &\leq \|(\lambda I + T)^{-1/2}(T - T_x)\|_{\mathcal{H}\to\mathcal{H}} \Big(\sqrt{\lambda}\|f\|_{\mathcal{H}} + \|f\|_{\rho}\Big),\end{aligned}$$

by the Cauchy-Schwarz inequality, which implies the first estimate. To prove the first consequence, we recall the function $\Psi_{x,\lambda}$ from (9), and we assign

$$a := \max\left\{4\sqrt{\lambda}, \Psi_{x,\lambda}\right\}.$$

Thus we may and do assume that (10) holds in the form

$$\left|\|f\|_{\rho}^2 - \|f\|_{\rho_x}^2\right| \leq a\left(\sqrt{\lambda}\|f\|_{\mathcal{H}} + \|f\|_{\rho}\right)\|f\|_{\mathcal{H}}.$$

This in turn yields, by noticing that $\sqrt{\lambda} \leq a/4$, the bound

$$\begin{aligned}\|f\|_{\rho_x}^2 &\leq \|f\|_{\rho}^2 + a\left(\sqrt{\lambda}\|f\|_{\mathcal{H}} + \|f\|_{\rho}\right)\|f\|_{\mathcal{H}} \leq \|f\|_{\rho}^2 + a\left(\frac{a}{4}\|f\|_{\mathcal{H}} + \|f\|_{\rho}\right)\|f\|_{\mathcal{H}} \\ &= \left(\|f\|_{\rho} + \frac{a}{2}\|f\|_{\mathcal{H}}\right)^2,\end{aligned}$$

from which the first estimate is an easy consequence.

To prove the second consequence we use that $ab \leq a^2/4 + b^2,\ a, b \geq 0$, and we thus start from (10) to find

$$\begin{aligned}\left|\|f\|_{\rho}^2 - \|f\|_{\rho_x}^2\right| &\leq \Psi_{x,\lambda}\sqrt{\lambda}\|f\|_{\mathcal{H}}^2 + \Psi_{x,\lambda}\|f\|_{\rho}\|f\|_{\mathcal{H}} \\ &\leq \Psi_{x,\lambda}\sqrt{\lambda}\|f\|_{\mathcal{H}}^2 + \frac{1}{4}\|f\|_{\rho}^2 + \Psi_{x,\lambda}^2\|f\|_{\mathcal{H}}^2.\end{aligned}$$

This yields

$$\frac{1}{2}\|f\|_{\rho}^2 \leq \frac{3}{4}\|f\|_{\rho}^2 \leq \|f\|_{\rho_x}^2 + \left(\Psi_{x,\lambda}^2 + \Psi_{x,\lambda}\sqrt{\lambda}\right)\|f\|_{\mathcal{H}}^2,$$

from which the desired bound follows. □

Proof of Proposition 4.3. We first prove the results for $f^\dagger \in T_\varphi$, $\varphi \in \mathcal{F}$, with an operator monotone index function φ. Notice, for any $\lambda > 0$

$$\begin{aligned}\|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} &= \|(I - g_\lambda(T_x)T_x)\varphi(T)v\|_{\mathcal{H}} \\ &\leq \underbrace{\|r_\lambda(T_x)T_x)(\lambda I + T)\|_{\mathcal{H}\to\mathcal{H}}}_{:=\mathcal{I}} \underbrace{\|(\lambda I + T)^{-1}\varphi(T)v\|_{\mathcal{H}}}_{:=\mathcal{II}}.\end{aligned}$$

We now estimate each term $\mathcal{I}$, $\mathcal{II}$ in the right-hand side of above inequality. In fact, by using Lemma 7.1 and the definition of the function Ξ we bound

$$\begin{aligned}\mathcal{I} &= \|r_\lambda(T_x)T_x)(\lambda I + T_x)(\lambda I + T_x)^{-1}(\lambda I + T)\|_{\mathcal{H}\to\mathcal{H}} \\ &\leq \|r_\lambda(T_x)(\lambda I + T_x)\|_{\mathcal{H}\to\mathcal{H}}\|(\lambda I + T_x)^{-1}(\lambda I + T)\|_{\mathcal{H}\to\mathcal{H}} \\ &\leq (\gamma_0 + \gamma_1)\lambda\Xi.\end{aligned}$$

The estimate for the second term $\|(\lambda I+T)^{-1}\varphi(T)v\|_{\mathcal{H}}$ is proven noticing $(\lambda I + T)^{-1} = \frac{1}{\lambda}(I - T(\lambda I+T)^{-1})$ which is in the weighted form of the residual function $r_\lambda(T)$ of Tikhonov regularization. Since it has qualification one and the index function φ is operator monotone, we find from Lemma 3.4 that

$$\mathcal{II} \leq C\frac{\varphi(\lambda)}{\lambda}\|v\|_{\mathcal{H}}.$$

Combing both estimates for $\mathcal{I}$, $\mathcal{II}$ above, we obtain

$$\begin{aligned}\|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} &\leq \|(I - g_\lambda(T_x)T_x)(\lambda I + T)\|_{\mathcal{H}\to\mathcal{H}}\|(\lambda I + T)^{-1}\varphi(T)v\|_{\mathcal{H}} \\ &\leq C(\gamma_0 + \gamma_1)\Xi\varphi(\lambda).\end{aligned}$$

The proof of $\|f^\dagger - \bar{f}_x^\lambda\|_\rho$ can be carried out in a similar manner. Using the definition of the norm, we derive

$$\begin{aligned}\|f^\dagger - \bar{f}_x^\lambda\|_\rho = \|\sqrt{T}(f^\dagger - \bar{f}_x^\lambda)\|_{\mathcal{H}} &\leq \|\sqrt{T}(I - g_\lambda(T_x)T_x)\varphi(T)v\|_{\mathcal{H}} \\ &\leq \underbrace{\|(\lambda I + T)^{1/2}(I - g_\lambda(T_x)T_x)(\lambda I + T)\|_{\mathcal{H}\to\mathcal{H}}}_{:=\mathcal{III}}\underbrace{\|(\lambda I + T)^{-1}\varphi(T)v\|_{\mathcal{H}}}_{:=\mathcal{IV}}.\end{aligned}$$

Factor $\mathcal{IV}$ equals the factor $\mathcal{II}$ from above, and we use that bound. For factor $\mathcal{III}$ we use that the operators $r_\lambda(T) = (I - g_\lambda(T_x)T_x)$ and $(\lambda I + T_x)$ commute, and we use the bound (21) to obtain

$$\begin{aligned}\mathcal{III} &\leq \|(\lambda I + T)^{1/2}(\lambda I + T_x)^{-1/2}(\lambda I + T_x)^{1/2}r_\lambda(T_x)(\lambda I + T)\|_{\mathcal{H}\to\mathcal{H}} \\ &\leq \Xi^{1/2}\|r_\lambda(T_x)(\lambda I + T_x)^{3/2}\|_{\mathcal{H}\to\mathcal{H}}\Xi.\end{aligned}$$

The middle factor was bounded in Lemma 7.1(2) which results in

$$\mathcal{III} \leq \sqrt{8(\gamma_0^2 + \gamma_{3/2}^2)}\Xi^{3/2}\lambda^{3/2}$$

and consequently

$$\|f^\dagger - \bar{f}_x^\lambda\|_\rho \leq C\sqrt{8(\gamma_0^2 + \gamma_{3/2}^2)}\Xi^{3/2}\lambda^{1/2}\varphi(\lambda),$$

proving Item 1.

We turn to proving Item 2. For $f^\dagger \in T_\varphi$, $\varphi \in \mathcal{F}_L$, $\varphi = \vartheta\psi$, we take the following decomposition

$$f^\dagger - \bar{f}_x^\lambda = (I - g_\lambda(T_x)T_x)\Big(\vartheta(T_x)\psi(T) + (\vartheta(T) - \vartheta(T_x))\psi(T)\Big)v.$$

This gives

$$\|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} \le \underbrace{\|r_\lambda(T_x)\vartheta(T_x)\psi(T)v\|_{\mathcal{H}}}_{:=\mathcal{V}} + \underbrace{\|r_\lambda(T_x)(\vartheta(T) - \vartheta(T_x))\psi(T)v\|_{\mathcal{H}}}_{:=\mathcal{VI}},$$

and we bound each term $\mathcal{V}$, $\mathcal{VI}$, separately. As above we obtain that

$$\begin{aligned}
\mathcal{V} &\le \|r_\lambda(T_x)\vartheta(T_x)(\lambda + T)(\lambda + T)^{-1}\psi(T)v\|_{\mathcal{H}} \\
&\le \|r_\lambda(T_x)\vartheta(T_x)(\lambda + T_x)\|_{\mathcal{H}\to\mathcal{H}}\Xi\|(\lambda + T)^{-1}\psi(T)v\|_{\mathcal{H}} \\
&\le C(\gamma_\vartheta + \gamma_{\vartheta+1})\vartheta(\lambda)\lambda\Xi\frac{\psi(\lambda)}{\lambda}\|v\|_{\mathcal{H}} \\
&= C(\gamma_\vartheta + \gamma_{\vartheta+1})\|v\|_{\mathcal{H}}\Xi\varphi(\lambda).
\end{aligned}$$

The bound for the second term is given by

$$\begin{aligned}
\mathcal{VI} &\le \|r_\lambda(T_x)(\vartheta(T) - \vartheta(T_x))\|_{\mathcal{H}\to\mathcal{H}}\|\psi(T)v\|_{\mathcal{H}} \\
&\le \psi(\kappa)\gamma_0\|v\|_{\mathcal{H}}\|T - T_x\|_{\mathcal{H}\to\mathcal{H}}.
\end{aligned}$$

Thus

$$\|f^\dagger - \bar{f}_x^\lambda\|_{\mathcal{H}} \le C\Big((\gamma_\vartheta + \gamma_{\vartheta+1})\Xi\varphi(\lambda) + \psi(\kappa)\gamma_0\|T - T_x\|_{\mathcal{H}\to\mathcal{H}}\Big).$$

For the ρ-norm we similarly have

$$\|f^\dagger - \bar{f}_x^\lambda\|_\rho \le \underbrace{\|\sqrt{T}r_\lambda(T_x)\vartheta(T_x)\psi(T)v\|_{\mathcal{H}}}_{:=\mathcal{VII}} + \underbrace{\|\sqrt{T}r_\lambda(T_x)(\vartheta(T) - \vartheta(T_x))\psi(T)v\|_{\mathcal{H}}}_{:=\mathcal{VIII}}$$

where we find, by using the definition of Ξ and Lemma 7.1(2) that

$$\begin{aligned}
\mathcal{VII} &\le \|(\lambda I + T)^{1/2}r_\lambda(T_x)\vartheta(T_x)\psi(T)v\|_{\mathcal{H}} \\
&\le \Xi^{3/2}\|r_\lambda(T_x)\vartheta(T_x)(\lambda I + T_x)^{3/2}\|_{\mathcal{H}\to\mathcal{H}}\|(\lambda I + T)^{-1}\psi(T)v\|_{\mathcal{H}} \\
&\le C\Xi^{3/2}\sqrt{8(\gamma_\vartheta^2 + \gamma_{\vartheta+3/2}^2)}\vartheta(\lambda)\lambda^{3/2}\frac{\psi(\lambda)}{\lambda}\|v\|_{\mathcal{H}} \\
&\le C\sqrt{8(\gamma_\vartheta^2 + \gamma_{\vartheta+3/2}^2)}\Xi^{3/2}\sqrt{\lambda}\varphi(\lambda)
\end{aligned}$$

and, by using estimate (A.35) from Lemma 7.1, we get

$$\begin{aligned}\mathcal{VIII} &\leq \|(\lambda I + T)^{1/2} r_\lambda(T_x)(\vartheta(T) - \vartheta(T_x))\psi(T)v\|_{\mathcal{H}} \\ &\leq \|(\lambda I + T)^{1/2} r_\lambda(T_x)\|_{\mathcal{H}\to\mathcal{H}} \|T - T_x\|_{\mathcal{H}\to\mathcal{H}} \psi(\kappa)\|v\|_{\mathcal{H}} \\ &\leq \Xi^{1/2} \|(\lambda I + T_x)^{1/2} r_\lambda(T_x)\|_{\mathcal{H}\to\mathcal{H}} \|T - T_x\|_{\mathcal{H}\to\mathcal{H}} \psi(\kappa)\|v\|_{\mathcal{H}} \\ &\leq \sqrt{(\gamma_0^2 + \gamma_{1/2}^2)}\psi(\kappa)\|v\|_{\mathcal{H}} \Xi^{1/2}\sqrt{\lambda}\|T - T_x\|_{\mathcal{H}\to\mathcal{H}},\end{aligned}$$

where we applied the bound (A.33) for $A := T_x$. We thus conclude

$$\begin{aligned}\|f^\dagger - \bar{f}_x^\lambda\|_\rho \leq C\Big(&\sqrt{8(\gamma_\vartheta^2 + \gamma_{\vartheta+3/2}^2)}\Xi^{3/2}\sqrt{\lambda}\varphi(\lambda) \\ &+ \sqrt{(\gamma_0^2 + \gamma_{1/2}^2)}\psi(\kappa)\Xi^{1/2}\sqrt{\lambda}\|T - T_x\|_{\mathcal{H}\to\mathcal{H}}\Big),\end{aligned}$$

which proves the second item, and completes the proof of the proposition. □

Proof of Proposition 4.4. By direct calculation, we derive

$$\begin{aligned}\|\bar{f}_x^\lambda - f_z^\lambda\|_{\mathcal{H}} &\leq \|g_\lambda(T_x)(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}} \\ &= \|g_\lambda(T_x)(\lambda I + T_x)^{1/2}(\lambda I + T_x)^{-1/2}(\lambda I + T)^{1/2}(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}} \\ &\leq (\gamma_{-1/2}^2 + \gamma_{-1}^2)^{1/2}\Xi^{1/2}\frac{1}{\sqrt{\lambda}}\|(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}},\end{aligned}$$

where we used (A.33) with $m(T_x) = g_\lambda(T_x)$, and we similarly find that

$$\begin{aligned}\|\bar{f}_x^\lambda - f_z^\lambda\|_\rho &\leq \|\sqrt{T}g_\lambda(T_x)(\lambda + T)^{1/2}(\lambda + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}} \\ &\leq \|(\lambda + T)^{1/2} g_\lambda(T_x)(\lambda + T)^{1/2}(\lambda + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}} \\ &\leq \Xi\|g_\lambda(T_x)(\lambda I + T_x)\|_{\mathcal{H}}\|(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}} \\ &\leq (\gamma_{-1} + \gamma_0 + 1)\Xi\|(\lambda I + T)^{-1/2}(S_x^* y - T_x f^\dagger)\|_{\mathcal{H}},\end{aligned}$$

which completes the proof. □

Proof of Lemma 4.6. Notice that the function $\lambda \to \mathcal{N}(\lambda)/\lambda$ is decreasing from ∞ to zero, such that the parameter $\lambda_* = \lambda_*(n)$ exists and is well-defined.

From the definition of λ_* we see that $n = \frac{\mathcal{N}(\lambda_*)}{\lambda_*} \geq \frac{\mathcal{N}(\kappa)}{\lambda_*}$, such that $n\lambda_* \geq 1/2$, which yields (22). Moreover, by monotonicity we find that

$$\frac{\mathcal{B}_{n,\lambda}}{\sqrt{\lambda}} \leq \frac{\mathcal{B}_{n,\lambda_*}}{\sqrt{\lambda_*}} = 2\kappa\left(\frac{\kappa}{n\lambda_*} + \sqrt{\frac{\mathcal{N}(\lambda_*)}{n\lambda_*}}\right) \leq 2\kappa\,(2\kappa + 1)\,, \tag{A.36}$$

which in turn gives (23). The bound (A.36) also implies that

$$\mathcal{B}_{n,\lambda}(\mathcal{B}_{n,\lambda}+\sqrt{\lambda}) \le (1+2\kappa)^2\sqrt{\lambda}\mathcal{B}_{n,\lambda}.$$

Now we distinguish two cases. For $\lambda \ge \sqrt{\kappa/n}$ we use that $\lambda\mathcal{N}(\lambda) \le \kappa$, which in turn implies

$$\begin{aligned}\sqrt{\lambda}\mathcal{B}_{n,\lambda} &\le \frac{2\kappa}{\sqrt{n}}\left(\frac{\kappa}{\sqrt{n}}+\sqrt{\lambda\mathcal{N}(\lambda)}\right)\\ &\le \frac{4\kappa\sqrt{\kappa}}{\sqrt{n}} \le (1+2\kappa)^2\sqrt{\frac{\kappa}{n}}. \end{aligned} \tag{A.37}$$

In the remaining case that $\lambda \le \sqrt{\kappa/n}$ we use that $\lambda\mathcal{N}(\lambda) \le \lambda^2 n$ since we assumed that $\lambda \ge \lambda_*$. This gives

$$\sqrt{\lambda}\mathcal{B}_{n,\lambda} \le \frac{2\kappa}{\sqrt{n}}\left(\frac{\kappa}{\sqrt{n}}+\lambda\sqrt{n}\right) \le 2\kappa\left(\frac{\kappa}{n}+\lambda\right) \le 2\kappa(1+\kappa)\lambda \le (1+2\kappa)^2\lambda. \tag{A.38}$$

Both bounds from (A.37) and (A.38) can be summarized as this is done in (24), and the proof is complete. □

Proof of Theorem 5.3. First, given $0 < \delta < 1$ we assign the cardinality $m := \lfloor 2\log(1/\delta)\rfloor$.

Since the smoothness function φ is at most polynomial, so is its companion function $\theta(t) := \sqrt{t}\varphi(t)$. Let (t_n, u_n) denote the singular value decomposition of the operator T. Under exponential decay we find a constant C_θ such that

$$\frac{1}{m}\sum_{n=1}^{m}\frac{1}{\theta^2(t_n)} \le C_\theta^2 e^m \le C_\theta^2\delta^{-2}.$$

Let $\tau = \frac{1}{C_\theta}\min\left\{1, \frac{1}{4\kappa\varphi(\kappa)}\right\}$. For the above m we apply [8, Prop. 6] to find $N = N_\delta$ and signs $\sigma_1, \dots, \sigma_N \in \{-1,+1\}^m$ such that

$$\frac{1}{m}\sum_{n=1}^{m}\left(\sigma_i^n - \sigma_j^n\right)^2 \ge 1,$$

where in addition $N_\delta \ge e^{m/24} \ge \left(\frac{1}{e\delta^2}\right)^{1/24}$.

With these signs we construct the functions

$$g_i := \tau\delta\frac{1}{\sqrt{m}}\sum_{n=1}^{m}\sigma_i^n\frac{1}{\theta(t_n)}e_n, \quad i = 1,\dots,N.$$

By construction this gives the norm bounds

$$\|g_i\|_{\mathcal{H}}^2 = \tau^2\delta^2\frac{1}{m}\sum_{n=1}^{m}|\sigma_i^n|^2\frac{1}{\theta^2(t_n)} \le \tau^2\delta^2 C_\theta^2\delta^{-2} \le 1.$$

To each function g_i we assign the function $f_i = \varphi(T)g_i,\ i = 1, \dots, N$, and we find that

$$\|f_i\|_{C(X)} \leq \kappa \|f_i\|_{\mathcal{H}} \leq \kappa\varphi(\kappa)\|g_i\|_{\mathcal{H}} \leq 1/4.$$

These functions also form a *tight packing* of T_φ, because we have that

$$\|f_i - f_j\|_\rho^2 = \|\theta(T)(g_i - g_j)\|_{\mathcal{H}}^2 = \tau^2\delta^2 \frac{1}{m} \sum_{n=1}^{m} \left|\sigma_i^n - \sigma_j^n\right|^2,$$

which gives that $\tau\delta \leq \|f_i - f_j\|_\rho \leq 4\tau\delta,\ i \neq j$. Thus the assumptions of DeVore et al. [24, Thm. 3.1] are fulfilled, and there is some measure ρ such that

$$\mathbb{E}_{\rho^m} \|\hat{f} - f_\rho\|_\rho \geq \tau\delta^*/4$$

whenever the sample size n satisfies $\log(N_{\delta^*}) \geq 16\tau^2 n\,(\delta^*)^2$. The left hand side is a decreasing function (in δ) whereas the right hand side is increasing. Calibration of both sides, and taking into account the lower bound for N_δ, this shows that there is some $\alpha > 0$ such that this holds true for $\delta^2 = \alpha\frac{\log(n)}{n}$, which completes the proof. □

References

[1] A. B. Bakushinskii, A general method of constructing regularizing algorithms for a linear ill-posed equation in Hilbert space, U.S.S.R. Comput. Math. and Math. Phys. 7 (1967) 279–287.

[2] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.

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

[4] Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, Constr. Approx. 26 (2007) 289–315.

[5] L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, A. Verri, Spectral algorithms for supervised learning, Neural Comput. 20 (2008) 1873–1897.

[6] Z.-C. Guo, S. Lin, D.-X. Zhou, Learning theory of distributed spectral algorithm, 2016. Submitted.

[7] D.-X. Zhou, Distributed learning algorithms (2016). MFO Report No. 33/2016.

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

[9] A. Caponnetto, L. Rosasco, E. De Vito, A. Verri, Empirical effective dimension and optimal rates for regularized least squares algorithm, Technical Report, Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, 2005.

[10] E. De Vito, S. Pereverzyev, L. Rosasco, Adaptive kernel methods using the balancing principle, Found. Comput. Math. 10 (2010) 455–479.

[11] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Anal. Appl. (Singap.) 8 (2010) 161–183.

[12] S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.

[13] S. Lu, S. V. Pereverzev, Regularization theory for ill-posed problems. Selected topics., volume 58 of *Inverse and Ill-posed Problems Series*, Walter De Gruyter, Berlin, 2013.

[14] T. Zhang, Effective dimension and generalization of kernel learning, NIPs (2002) 454–461.

[15] K. Lin, S. Lu, P. Mathé, Oracle-type posterior contraction rates in Bayesian inverse problems, Inverse Probl. Imaging 9 (2015) 895–915.

[16] C. Scovel, D. Hush, I. Steinwart, J. Theiler, Radial kernels and their reproducing kernel Hilbert spaces, J. Complexity 26 (2010) 641–660.

[17] T. Kühn, Covering numbers of Gaussian reproducing kernel Hilbert spaces, J. Complexity 27 (2011) 489–499.

[18] G. Blanchard, P. Mathé, Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration, Inverse Problems 28 (2012) 115011, 23pp.

[19] G. Blanchard, N. Mücke, Empirical effective dimension, 2016. Private communication.

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

[21] P. Mathé, S. V. Pereverzev, Regularization of some linear ill-posed problems with discretized random noisy data, Math. Comp. 75 (2006) 1913–1929.

[22] P. Mathé, S. V. Pereverzev, Discretization strategy for linear ill-posed problems in variable Hilbert scales, Inverse Problems 19 (2003) 1263–1277.

[23] G. Blanchard, N. Krämer, Optimal learning rates for kernel conjugate gradient regression, NIPs (2010) 226–234.

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

[25] C. A. Micchelli, M. Pontil, Learning the kernel function via regularization, J. Mach. Learn. Res. 6 (2005) 1099–1125.