Journal of Machine Learning Research 19 (2018) 1-29 Submitted 11/16; Revised 8/18; Published 8/18

Parallelizing Spectrally Regularized Kernel Algorithms

Nicole Mücke NICOLE.MUECKE@MATHEMATIK.UNI-STUTTGART.DE
Institute of Stochastics and Applications, University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Gilles Blanchard* GILLES.BLANCHARD@MATH.UNI-POTSDAM.DE
Institute of Mathematics, University of Potsdam
Karl-Liebknecht-Strae 24-25
14476 Potsdam, Germany

Editor: Ingo Steinwart

Abstract

We consider a distributed learning approach in supervised learning for a large class of spectral regularization methods in an reproducing kernel Hilbert space (RKHS) framework. The data set of size n is partitioned into $m = O(n^\alpha)$, $\alpha < \frac{1}{2}$, disjoint subsamples. On each subsample, some spectral regularization method (belonging to a large class, including in particular Kernel Ridge Regression, L^2-boosting and spectral cut-off) is applied. The regression function f is then estimated via simple averaging, leading to a substantial reduction in computation time. We show that minimax optimal rates of convergence are preserved if m grows sufficiently slowly (corresponding to an upper bound for α) as $n \to \infty$, depending on the smoothness assumptions on f and the intrinsic dimensionality. In spirit, the analysis relies on a classical bias/stochastic error analysis.

Keywords: Distributed Learning, Spectral Regularization, Minimax Optimality

1. Introduction

Distributed learning (DL) algorithms are a standard tool for reducing computational burden in machine learning problems where massive datasets are involved. Assuming a complexity cost (for time and/or memory) of $O(n^\beta)$ ($\beta > 1$, $\beta \in [2, 3]$ being common) of the base learning algorithm without parallelization, dividing randomly data of cardinality n into m disjoint, equally-sized subsamples and processing them in parallel using the same base learning algorithm has therefore complexity cost of $O(m.(n/m)^\beta) = O(n^\beta/m^{\beta-1})$, roughly

*. Financial support by the DFG via Research Unit 1735 "Structural Inference in Statistics" as well as SFB 1294 "Data Assimilation" is gratefully acknowledged.

gaining a factor $m^{\beta-1}$ (for time and memory) compared to the single machine approach. The final output is obtained from averaging the individual outputs.

Recently, DL was studied in several machine learning contexts. In point estimation (Li et al., 2013), matrix factorization (Mackey et al., 2011), smoothing spline models and testing (Cheng and Shang, 2016), local average regression (Chang et al., 2017), in classification (Hsieh et al., 2014; Guo et al., 2015), and also in kernel ridge regression (Zhang et al., 2013; Lin et al., 2017; Xu et al., 2016).

In this paper, we study the DL approach for the statistical learning problem

$$Y_i := f(X_j) + \epsilon_i \, , j = 1 \, , \ldots , n \, , \tag{1}$$

at random i.i.d. data points $X_1, \ldots, X_n$ drawn according to a probability distribution ν on $\mathcal{X}$, where ϵ_j are independent centered noise variables. The unknown regression function f is real-valued and belongs to some reproducing kernel Hilbert space with bounded kernel K. We partition the given data set $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subset \mathcal{X} \times \mathbb{R}$ into m disjoint equal-size subsamples $D_1, \ldots, D_m$. On each subsample D_j, we compute a local estimator $\hat{f}^{\lambda}_{D_j}$, using a spectral regularization method. The final estimator for the target function f is obtained by simple averaging: $\bar{f}^{\lambda}_{D} := \frac{1}{m} \sum_{j=1}^{m} \hat{f}^{\lambda}_{D_j}$.
The non-distributed setting ($m = 1$) has been studied in the recent paper of Blanchard and Mücke (2017), building the root position of our results in the distributed setting, where weak and strong minimax optimal rates of convergence are established. Our aim is to extend these results to distributed learning and to derive minimax optimal rates. We again apply a fairly large class of spectral regularization methods, including the popular kernel ridge regression (KRR), L^2-boosting and spectral cut-off. Using the same notation as Blanchard and Mücke (2017), we let

$$T : f \in \mathcal{H}_K \mapsto \int f(x) K(x, \cdot) d\nu(x) \in \mathcal{H}_K$$

denote the kernel integral operator associated to K and the sampling measure ν. We denote $\bar{T} = \kappa^{-2} T$, with κ^2 the upper bound of K. Our rates of convergence are governed by a *source condition* assumption on f of the form

$$\Omega(r, R) := \{ f \in \mathcal{H}_K : f = \bar{T}^r h, \ \|h\|_{\mathcal{H}_K} \leq R \}$$

for some constants $r, R > 0$ as well as by the *ill-posedness* of the problem, as measured by an assumed power decay of the eigenvalues of T with exponent $b > 1$. We show that for $s \in [0, \frac{1}{2}]$ in the sense of p-th moment ($p \geq 1$) expectation

$$\left\| \bar{T}^s (f - \bar{f}^{\lambda_n}_{D}) \right\|_{\mathcal{H}_K} \lesssim R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{(r+s)}{2r+1+1/b}} \, , \tag{2}$$

for an appropriate choice of the regularization parameter λ_n , depending on the global sample size n as well as on R and the noise variance σ^2 (but not on the number m of subsample sets). Note that $s = 0$ corresponds to the reconstruction error (i.e. $\mathcal{H}_K$-norm), and $s = \frac{1}{2}$ to the

prediction error (i.e., $L^2(\nu)$-norm). The symbol $\lesssim$ means that the inequality holds up to a multiplicative constant that can depend on various parameters entering in the assumptions of the result, but not on n, m, σ, nor R. An important assumption is that the inequality $q \geq r + s$ should hold, where q is the *qualification* of the regularization method, a quantity defined in the classical theory of inverse problems (see Section 2.3 for a precise definition). Basic problems are the choice of the regularization parameter on the subsamples and, most importantly, the proper choice of m, since it is well known that choosing m too large gives a suboptimal convergence rate in the limit $n \to \infty$ (see, e.g., Xu et al., 2016).

Our approach to this problem is based on a relatively classical bias-variance decomposition principle. Choosing the global regularization parameter as the optimal choice for a *single* sample of size n results in a bias estimate which is identical for all subsamples, is unchanged by averaging, and is straightforward from the single-sample analysis. On the other hand, the reduced sample size of each of the m individual subsamples causes an inflation of variance. However, since the m subsamples are independent, so are the outputs of the learning algorithm applied to each one of them; as a consequence averaging reduces the inflated variance sufficiently to get minimax optimality. We can write the variance as a sum of independent random variables, allowing to successfully apply a Rosenthal's inequality in the Hilbert space setting due to Pinelis (1994). The technical "limiting factors" in this argument give rise to the limitation on the number of subsamples m; for m larger than the allowed range, some remainder terms are no longer negligible using our proof technique, and rate optimality is not guaranteed any longer.

The outline of the paper is as follows. Section 2 contains notation and the setting. Section 3 states our main result on distributed learning. Section 4 presents numerical studies. A concluding discussion in Section 5 contains a more detailed comparison of our results with related results available in the literature. Section 6 contains the proofs of the theorems.

2. Notation, statistical model and distributed learning algorithm

In this section, we specify the mathematical background and the statistical model for (distributed) regularized learning. We have included this section for self sufficiency and reader convenience. It essentially repeats the setting in Blanchard and Mücke (2017) in summarized form.

2.1 Kernel-induced operators

We assume that the input space $\mathcal{X}$ is a standard Borel space endowed with a probability measure ν, the output space is equal to $\mathbb{R}$. We let K be a real-valued positive semidefinite kernel on $\mathcal{X} \times \mathcal{X}$ which is bounded by κ^2. The associated reproducing kernel Hilbert space will be denoted by $\mathcal{H}_K$. It is assumed that all functions $f \in \mathcal{H}_K$ are measurable and bounded in supremum norm, i.e. $\|f\|_\infty \leq \kappa \|f\|_{\mathcal{H}_K}$ for all $f \in \mathcal{H}_K$. Therefore, $\mathcal{H}_K$ is a subset of $L^2(\mathcal{X}, \nu)$, with $S : \mathcal{H}_K \longrightarrow L^2(\mathcal{X}, \nu)$ being the inclusion operator, satisfying

$\|S\| \leq \kappa$. The adjoint operator $S^* : L^2(\mathcal{X}, \nu) \longrightarrow \mathcal{H}_K$ is identified as

$$S^* g = \int_{\mathcal{X}} g(x) K_x \, \nu(dx) \, ,$$

where K_x denotes the element of $\mathcal{H}_K$ equal to the function $t \mapsto K(x,t)$. The covariance operator $T : \mathcal{H}_K \longrightarrow \mathcal{H}_K$ is given by

$$T = \int_{\mathcal{X}} \langle \cdot, K_x \rangle_{\mathcal{H}_K} K_x \, \nu(dx) \, ,$$

which can be shown to be positive self-adjoint trace class (and hence is compact). The empirical versions of these operators, corresponding formally to taking the empirical distribution $\widehat{\nu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ in place of ν in the above formulas, are given by

$$S_{\mathbf{x}} : \mathcal{H}_K \longrightarrow \mathbb{R}^n \, , \qquad (S_{\mathbf{x}} f)_j = \langle f, K_{x_j} \rangle_{\mathcal{H}_K} \, ,$$

$$S_{\mathbf{x}}^* : \mathbb{R}^n \longrightarrow \mathcal{H}_K \, , \qquad S_{\mathbf{x}}^* \mathbf{y} = \frac{1}{n} \sum_{j=1}^n y_j K_{x_j} \, ,$$

$$T_{\mathbf{x}} := S_{\mathbf{x}}^* S_{\mathbf{x}} : \mathcal{H}_K \longrightarrow \mathcal{H}_K \, , \qquad T_{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^n \langle \cdot, K_{x_j} \rangle_{\mathcal{H}_K} K_{x_j} \, .$$

We introduce the shortcut notation $\bar{T} = \kappa^{-2} T$ and $\bar{T}_{\mathbf{x}} := \kappa^{-2} T_{\mathbf{x}}$, ensuring $\|\bar{T}\| \leq 1$ and $\|\bar{T}_x\| \leq 1$, for any $x \in \mathcal{X}$. Similarly, $\bar{S} = \kappa^{-1} S$ and $\bar{S}_{\mathbf{x}_j} := \kappa^{-1} S_{\mathbf{x}_j}$, ensuring $\|\bar{S}\| \leq 1$ and $\|\bar{S}_x\| \leq 1$, for any $x \in \mathcal{X}$. The numbers μ_j are the positive eigenvalues of $\bar{T}$ satisfying $0 < \mu_{j+1} \leq \mu_j$ for all $j > 0$ and $\mu_j \searrow 0$.

2.2 Noise assumption and prior classes

In our setting of kernel learning, the sampling is assumed to be random i.i.d., where each observation point (X_i, Y_i) follows the model $Y = f_\rho(X) + \varepsilon$. For (X, Y) having distribution ρ, we assume that the conditional expectation wrt. ρ of Y given X exists and belongs to $\mathcal{H}_K$, that is, it holds for ν-almost all $x \in X$:

$$\mathbb{E}_\rho[Y | X = x] = \bar{S}_x f_\rho \, , \text{ for some } f_\rho \in \mathcal{H}_K \, . \tag{3}$$

Furthermore, we will make the following assumption on the observation noise distribution: There exists $\sigma > 0$ and $M > 0$ such that for any $l \geq 2$

$$\mathbb{E}[\left| Y - \bar{S}_X f_\rho \right|^l \mid X] \leq \frac{1}{2} l! \sigma^2 M^{l-2} \, , \quad \nu - \text{a.s.} \tag{4}$$

To derive nontrivial rates of convergence, we concentrate our attention on specific subsets (also called *models*) of the class of probability measures. If $\mathcal{P}$ denotes the set of all probability distributions on $\mathcal{X}$, we define classes of sampling distributions by introducing a decay

condition on the *effective dimension* $\mathcal{N}(\lambda)$, being a measure for the complexity of $\mathcal{H}_K$ with respect to the marginal distribution ν: For $\lambda \in (0,1]$ we set

$$\mathcal{N}(\lambda) = \text{Trace}[\ (\bar{T}+\lambda)^{-1}\bar{T}\ ]\ . \tag{5}$$

Note that $\mathcal{N}(\lambda) \leq 1$. For any $b > 1$ we introduce

$$\mathcal{P}^{<}(b) := \{\nu \in \mathcal{P} :\ \mathcal{N}(\lambda) \leq C_b(\kappa^2\lambda)^{-\frac{1}{b}}\ \}\ . \tag{6}$$

In De Vito and Caponnetto, 2006, Proposition 3, it is shown that such a condition is implied by polynomially decreasing eigenvalues of $\bar{T}$. More precisely, if the eigenvalues μ_i satisfy $\mu_j \leq \beta/j^b\ \ \forall j \geq 1$ or $b > 1$ and $\beta > 0$, then

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

For a subset $\Omega \subseteq \mathcal{H}_K$, we let $\mathcal{K}(\Omega)$ be the set of regular conditional probability distributions $\rho(\cdot|\cdot)$ on $\mathcal{B}(\mathbb{R})\times\mathcal{X}$ such that (3) and (4) hold for some $f_\rho \in \Omega$. We will focus on a *Hölder-type source condition*, i.e. given $r > 0, R > 0$ and $\nu \in \mathcal{P}$, we define

$$\Omega(r,R) := \{f \in \mathcal{H}_K : f = \bar{T}^r h,\ \|h\|_{\mathcal{H}_K} \leq R\}. \tag{7}$$

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

$$\mathcal{M}(r,R,\mathcal{P}')\ :=\ \{\ \rho(dx,dy) = \rho(dy|x)\nu(dx)\ :\ \rho(\cdot|\cdot) \in \mathcal{K}(\Omega(r,R)),\ \nu \in \mathcal{P}'\ \}\ , \tag{8}$$

with $\mathcal{P}' = \mathcal{P}^{<}(b)$. As a consequence, the class of models depends not only on the smoothness properties of the solution (reflected in the parameters $R > 0,\ r > 0$), but also essentially on spectral properties of $\bar{T}$, reflected in $\mathcal{N}(\lambda)$.

2.3 Spectral regularization

In this subsection, we introduce the class of linear regularization methods based on spectral theory for self-adjoint linear operators. These are standard methods for finding stable solutions for ill-posed inverse problems. Originally, these methods were developed in the deterministic context (see Engl et al., 2000). Later on, they have been applied to probabilistic problems in machine learning (see, e.g., Bauer et al., 2007; De Vito and Caponnetto, 2006; Dicker et al., 2017 or Blanchard and Mücke, 2017).

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

(i) There exists a constant $D' < \infty$ such that for any $0 < \lambda \leq 1$

$$\sup_{0\leq t\leq 1} |tg_\lambda(t)| \leq D'\ . \tag{9}$$

(ii) There exists a constant $E < \infty$ such that for any $0 < \lambda \leq 1$

$$\sup_{0 \leq t \leq 1} |g_\lambda(t)| \leq \frac{E}{\lambda} \,. \tag{10}$$

(iii) Defining the residual $r_\lambda(t) := 1 - g_\lambda(t)t$ *, there exists a constant $\gamma_0 < \infty$ such that for any $0 < \lambda \leq 1$*

$$\sup_{0 \leq t \leq 1} |r_\lambda(t)| \leq \gamma_0 \,.$$

It has been shown in e.g. Gerfo et al. (2008), Dicker et al. (2017), Blanchard and Mücke (2017) that attainable learning rates are essentially linked with the qualification of the regularization $\{g_\lambda\}_\lambda$, being the maximal q such that for any $0 < \lambda \leq 1$

$$\sup_{0 \leq t \leq 1} |r_\lambda(t)| t^q \leq \gamma_q \lambda^q. \tag{11}$$

for some constant $\gamma_q > 0$. Note that by (iii) , using interpolation, we have validity of (11) also for any $q' \in [0, q]$ with constant $\gamma_{q'} = \gamma_0^{1-\frac{q'}{q}} \gamma_q^{\frac{q'}{q}}$.

The most popular examples include:

Example 1 *(Tikhonov Regularization, Kernel Ridge Regression) The choice $g_\lambda(t) = \frac{1}{\lambda + t}$ corresponds to Tikhonov regularization. In this case we have $D' = E = \gamma_0 = 1$. The qualification of this method is $q = 1$ with $\gamma_q = 1$.*

Example 2 *(Landweber Iteration, gradient descent) The Landweber Iteration (gradient descent algorithm with constant stepsize) is defined by*

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

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

Example 3 (ν- method) *The $\nu-$ method belongs to the class of so called semi-iterative regularization methods. This method has finite qualification $q = \nu$ with γ_q a positive constant. Moreover, $D = 1$ and $E = 2$. The filter is given by $g_k(t) = p_k(t)$, a polynomial of degree $k - 1$, with regularization parameter $\lambda \sim k^{-2}$, which makes this method much faster as e.g. gradient descent.*

2.4 Distributed learning algorithm

We let $D = \{(x_j, y_j)\}_{j=1}^n \subset \mathcal{X} \times \mathcal{Y}$ be the dataset, which we partition into m disjoint subsamples[1] $D_1, \ldots, D_m$, each having size $\frac{n}{m}$. Denote the jth data subsample by $(\mathbf{x}_j, \mathbf{y}_j) \in (\mathcal{X} \times \mathbb{R})^{\frac{n}{m}}$. On each subsample we compute a local estimator for a suitable a-priori parameter choice $\lambda = \lambda_n$ according to

$$f_{D_j}^{\lambda_n} := g_{\lambda_n}(\bar{T}_{\mathbf{x}_j}) \bar{S}_{\mathbf{x}_j}^* \mathbf{y}_j \ . \tag{12}$$

By f_D^λ we will denote the estimator using the whole sample $m = 1$. The final estimator is given by simple averaging of the local ones:

$$\bar{f}_D^\lambda := \frac{1}{m} \sum_{j=1}^m f_{D_j}^\lambda \ . \tag{13}$$

3. Main results

This section presents our main results. Theorem 3 and Theorem 4 contain separate estimates on the approximation error and the sample error and lead to Corollary 5 which gives an upper bound for the error $\left\| \bar{T}^s (f_\rho - \bar{f}_D^\lambda) \right\|_{\mathcal{H}_K}$ and presents an upper rate of convergence for the sequence of distributed learning algorithms.

For the sake of the reader we recall Theorem 6, which was already shown in Blanchard and Mücke (2017), presenting the minimax optimal rate for the single machine problem. This yields an estimate on the difference between the single machine and the distributed learning algorithm in Corollary 7.

We want to track the precise behavior of these rates not only for what concerns the exponent in the number of examples n, but also in terms of their scaling (multiplicative constant) as a function of some important parameters (namely the noise variance σ^2 and the complexity radius R in the source condition, see Remark 9 below). For this reason, we introduce a notion of a family of rates over a family of models. More precisely, we consider an indexed family $(\mathcal{M}_\theta)_{\theta \in \Theta}$, where for all $\theta \in \Theta$, $\mathcal{M}_\theta$ is a class of Borel probability distributions on $\mathcal{X} \times \mathbb{R}$ satisfying the basic general assumptions (3) and (4). We consider rates of convergence in the sense of the p-th moments of the estimation error, where $1 \leq p < \infty$ is a fixed real number.

1. For the sake of simplicity, throughout this paper we assume that n is divisible by m. This could always be achieved by disregarding some data; alternatively, it is straightforward to show that admitting one smaller block in the partition does not affect the asymptotic results of this paper. We shall not try to discuss this point in greater detail. In particular, we shall not analyze in which general framework our simple averages could be replaced by weighted averages.

As already mentioned in the introduction, our proofs are based on a classical bias-variance decomposition as follows: Introducing

$$\tilde{f}_D^{\lambda} = \frac{1}{m}\sum_{j=1}^{m} g_\lambda(\bar{T}_{\mathbf{x}_j})\bar{T}_{\mathbf{x}_j} f_\rho \,, \tag{14}$$

we write

$$\begin{aligned}\bar{T}^s(f_\rho - \bar{f}_D^{\lambda}) &= \bar{T}^s(f_\rho - \tilde{f}_D^{\lambda}) + \bar{T}^s(\tilde{f}_D^{\lambda} - \bar{f}_D^{\lambda}) \\ &= \frac{1}{m}\sum_{j=1}^{m} \bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) f_\rho + \frac{1}{m}\sum_{j=1}^{m} \bar{T}^s g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j) \,. \end{aligned} \tag{15}$$

In all the forthcoming results in this section, we assume:

Assumption 2 *Let $s \in [0, \frac{1}{2}]$, $p \geq 1$ and consider the model $\mathcal{M}_{\sigma,M,R} := \mathcal{M}(r, R, \mathcal{P}^<(b))$ where $r > 0$ and $b > 1$ are fixed, and $\theta = (R, M, \sigma)$ varies in $\Theta = \mathbb{R}^3_+$. Given a sample $D \subset (\mathcal{X} \times \mathbb{R})$ of size n, define $\bar{f}_D^{\lambda_n}$, $f_D^{\lambda_n}$ as in Section 2.4 and $\tilde{f}_D^{\lambda_n}$ as in (14), using a regularization function of qualification $q \geq r + s$, with parameter sequence*

$$\lambda_n := \lambda_{n,(\sigma,R)} := \min\left(\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b}{2br+b+1}}, 1\right) \,, \tag{16}$$

independent on M. Define the sequence

$$a_n := a_{n,(\sigma,R)} := R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b(r+s)}{2br+b+1}} \,. \tag{17}$$

We recall that we shall always assume that n is a multiple of m. With these preparations, our main results are:

Theorem 3 (Approximation error) *Under Assumption 2, we have: If the number m_n of subsample sets satisfies*

$$m_n \leq n^\alpha \,, \quad \alpha < \frac{2b\min\{r,1\}}{2br+b+1} \,, \tag{18}$$

then

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s(f_\rho - \tilde{f}_D^{\lambda_n})\right\|^p_{\mathcal{H}_K}\right]^{\frac{1}{p}}}{a_n} < \infty \,.$$

Theorem 4 (Sample Error) *Under Assumption 2, we have: If the number m_n of subsample sets satisfies*

$$m_n \leq n^{\alpha}\,, \quad \alpha < \frac{2br}{2br+b+1}\,, \tag{19}$$

Then

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}}}{a_n} < \infty\,.$$

And, as consequence (by (15) and applying the triangle inequality for the L^p-norm):

Corollary 5 *Under Assumption 2, we have: If the number m_n of subsample sets satisfies*

$$m_n \leq n^{\alpha}\,, \quad \alpha < \frac{2b\min\{r,1\}}{2br+b+1}\,, \tag{20}$$

then the sequence (17) *is an upper rate of convergence in L^p for all $p>0$, for the interpolation norm of parameter s, for the sequence of estimated solutions $(\bar{f}_D^{\lambda_{n,(\sigma,R)}})$ over the family of models $(\mathcal{M}_{\sigma,M,R})_{(\sigma,M,R)\in\mathbb{R}^3_+}$, i.e.*

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(f_\rho - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}}}{a_n} < \infty\,.$$

Theorem 6 (Blanchard and Mücke, 2017) *The sequence* (17) *is an upper rate of convergence in L^p for all $p>0$, for the interpolation norm of parameter s, for the sequence of estimated solutions $(f_D^{\lambda_{n,(\sigma,R)}})$ over the family of models $(\mathcal{M}_{\sigma,M,R})_{(\sigma,M,R)\in\mathbb{R}^3_+}$, i.e.*

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(f_\rho - f_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}}}{a_n} < \infty\,.$$

Combining Corollary 5 with Theorem 6 by applying the triangle inequality immediately yields:

Corollary 7 *If the number m_n of subsample sets satisfies*

$$m_n \leq n^{\alpha}\,, \quad \alpha < \frac{2b\min\{r,1\}}{2br+b+1}\,, \tag{21}$$

then

$$\sup_{(\sigma,M,R)\in\mathbb{R}^3_+} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(f_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}}}{a_n} < \infty\,.$$

Remark 8 *Our results in the distributed setting slightly differ from those obtained in Theorem 6 from Blanchard and Mücke (2017) in two several respects:*

- *While in the single machine approach, rates of convergence are obtained for any $p > 0$, the proofs in Section 6 only hold for $p \geq 1$ due to loss of subadditivity of p-th moments for $0 < p < 1$.*
- *While the upper upper rates of convergence in Blanchard and Mücke (2017) are derived over classes of marginals ν induced by assuming a decay condition for the eigenvalues of $\bar{T}$, we somewhat enlarge this class by assuming a decay condition for $\mathcal{N}(\lambda)$ in (6). Theorem 6 also holds under this weaker condition. Note that it is an open problem if lower rates of convergence can also be obtained by weakening the condition for eigenvalue decay.*

Remark 9 (Signal-to-noise-ratio) *Our results show that the choice of the regularization parameter λ_n in (16) and thus the rate of convergence a_n in (17) highly depend on the signal-to-noise-ratio $\frac{\sigma^2}{R^2}$, a quantity which naturally appears in the theory of regularization of ill-posed inverse problems. As a general rule, the degree of regularization should increase with the level of noise in the data, i.e., the importance of the priors should increase as the model fit decreases. Our theoretical results precisely show this behavior.*

4. Numerical studies

In this section we numerically study the error in $\mathcal{H}_K$-norm, corresponding to $s = 0$ in Corollary 5 (in expectation with $p = 2$) both in the single machine and distributed learning setting. Our main interest is to study the upper bound for our theoretical exponent α, parametrizing the size of subsamples in terms of the total sample size, $m = n^\alpha$, in different smoothness regimes. In addition we shall demonstrate in which way parallelization serves as a form of regularization.

More specifically, we let $\mathcal{H}_K = H_0^1[0,1]$ be the Sobolev space consisting of absolutely continuous functions f on $[0,1]$ with weak derivative of order 1 in $L^2[0,1]$, with boundary condition $f(0) = f(1) = 0$. The reproducing kernel is given by $K(x,t) = x \wedge t - xt$. For all experiments in this section, we simulate data from the regression model

$$Y_i = f_\rho(X_i) + \epsilon_i \,, \qquad i = 1, \ldots, n \,,$$

where the input variables $X_i \sim Unif[0,1]$ are uniformly distributed and the noise variables $\varepsilon_i \sim N(0, \sigma^2)$ are normally distributed with standard deviation $\sigma = 0.005$. We choose the target function f_ρ according to two different cases, namely $r < 1$ (*low smoothness regime*) and $r = \infty$ (*high smoothness regime*). To accurately determine the degree of smoothness $r > 0$, we apply Proposition 10 below by explicitly calculating the Fourier coefficients $(\langle f_\rho, e_j \rangle_{\mathcal{H}_K})_{j \in \mathbb{N}}$, where $e_j(x) = \frac{\sqrt{2}}{\pi j} \cos(\pi j x)$, for $j \in \mathbb{N}^*$, forms an ONB of $\mathcal{H}_K$. Recall that the rate of eigenvalue decay is explicitly given by $b = 2$, meaning that we have full control over all parameters in (21). We need the following characterization:

Proposition 10 (Engl et al., 2000, Prop. 3.13) *Let $\mathcal{H}_K, \mathcal{H}_2$ be separable Hilbert spaces and $S : \mathcal{H}_K \longrightarrow \mathcal{H}_2$ be a compact linear operator with singular system* [2] *$\{\sigma_j, \varphi_j, \psi_j\}$. Denoting by $S^\dagger$ the generalized inverse*[3] *of S, one has for any $r > 0$ and $g \in \mathcal{H}_2$:*

g is in the domain of $S^\dagger$ and $S^\dagger g \in \mathrm{Im}((S^*S)^r)$ if and only if

$$\sum_{j=0}^{\infty} \frac{|\langle g, \psi_j \rangle_{\mathcal{H}_2}|^2}{\sigma_j^{2+4r}} < \infty .$$

In our case, $\mathcal{H}_K$ is as above, $\mathcal{H}_2$ is $L^2([0,1])$ with Lebesgue measure and $S : H_0^1[0,1] \to L^2([0,1])$ is the inclusion. Since $H_0^1[0,1]$ is dense in $L^2([0,1])$, we know that $(\mathrm{Im}(S))^\perp$ is trivial, giving $SS^\dagger = id$ on $\mathrm{Im}(S)$. Furthermore, $\varphi_j = e_j$ is a normalized eigenbasis of $T = S^*S$ with eigenvalues $\sigma_j^2 = (\pi j)^{-2}$. With $\psi_j = \frac{S\varphi_j}{\|S\varphi_j\|_{L^2}}$ we obtain for $f \in H_0^1[0,1]$

$$\langle Sf, \psi_j \rangle_{L^2} = \left\langle Sf, \frac{Se_j}{\|Se_j\|} \right\rangle_{L^2} = \left\langle f, \frac{S^*Se_j}{\|Se_j\|} \right\rangle_{H_0^1} = \sigma_j \langle f, e_j \rangle_{H_0^1}.$$

Thus, applying Proposition 10 gives

Corollary 11 *For S and $T = S^*S$ defined in Section 2, we have for any $r > 0$: $f \in \mathrm{Im}(T^r)$ if and only if*

$$\sum_{j=1}^{\infty} j^{4r} |\langle f, e_j \rangle_{L^2}|^2 < \infty .$$

Thus, as expected, abstract smoothness measured by the parameter r in the source condition corresponds in this special case to decay of the classical Fourier coefficients which by the classical theory of Fourier series measures smoothness of the periodic continuation of $f \in L^2([0,1])$ to the real line.

4.1 Low smoothness regime

We choose $f_\rho(x) = \frac{1}{2}x(1-x)$ which clearly belongs to $\mathcal{H}_K$. A straightforward calculation gives the Fourier coefficient $\langle f_\rho, e_j \rangle = -2(\pi j)^{-2}$ for j odd (vanishing for j even). Thus, by the above criterion, f_ρ satisfies the source condition $f_\rho \in Ran(\bar{T}^r)$ precisely for $0 < r < 0.75$. (Observe that although f_ρ is smooth on $[0,1]$, its periodic continuation on the real line is not, hence the low smoothness regime.) According to Theorem 6, the worst case rate in the single machine problem is given by $n^{-\gamma}$, with $\gamma = 0.25$. Regularization is done using the $\nu-$ method (see Example 3), with qualification $q = \nu = 1$. Recall that the stopping index

2. i.e., the φ_j are the normalized eigenfunctions of S^*S with eigenvalues σ_j^2 and $\psi_j = S\varphi_j/\|S\varphi_j\|$; thus $S = \sum \sigma_j \langle \varphi_j, \cdot \rangle \psi_j$.
3. the unique unbounded linear operator with domain $\mathrm{Im}(S) \oplus (\mathrm{Im}(S))^\perp$ in $\mathcal{H}_2$ vanishing on $(\mathrm{Im}(S))^\perp$ and satisfying $SS^\dagger = 1$ on $\mathrm{Im}(S)$, with range orthogonal to the null space $N(S)$.

k_{stop} serves as the regularization parameter λ, where $k_{stop} \sim \lambda^{-2}$. We consider sample sizes from $500, \ldots, 9000$. In the model selection step, we estimate the performance of different models and choose the *oracle stopping time* $\hat{k}_{oracle}$ by minimizing the reconstruction error:

$$\hat{k}_{oracle} = \arg\min_k \left(\frac{1}{M} \sum_{j=1}^{M} \left\| f_\rho - \hat{f}_j^k \right\|_{\mathcal{H}_K}^2 \right)^{\frac{1}{2}}$$

over $M = 30$ runs.

In the model assessment step, we partition the dataset into $m \sim n^\alpha$ subsamples, for any $\alpha \in \{0, 0.05, 0.1, \ldots, 0.85\}$. On each subsample we regularize using the oracle stopping time $\hat{k}_{oracle}$ (determined by using the whole sample). Corresponding to Corollary 5, the accuracy should be comparable to the one using the whole sample as long as $\alpha < 0.5$. In Figure 1 (left panel) we plot the reconstruction error $\|\bar{f}^{\hat{k}} - f_\rho\|_{\mathcal{H}_K}$ versus the ratio $\alpha = \log(m)/\log(n)$ for different sample sizes. We execute each simulation $M = 30$ times. The plot supports our theoretical finding. The right panel shows the reconstruction error versus the total number of samples using different partitions of the data. The black curve ($\alpha = 0$) corresponds to the baseline error ($m = 0$, no partition of data). Error curves below a threshold $\alpha < 0.6$ are roughly comparable, whereas curves above this threshold show a gap in performances.

In another experiment we study the performances in case of (very) different regularization: Only partitioning the data (no regularization), underregularization (higher stopping index) and overregularization (lower stopping index). The outcome of this experiment amplifies the regularization effect of parallelizing. Figure 2 shows the main point: Overregularization is always hopeless, underregularization is better. In the extreme case of (almost) no regularization, there is a sharp minimum in the reconstruction error which is only slightly larger than the minimax optimal value for the oracle regularization parameter and which is achieved at an attractively large degree of parallelization. Qualitatively, this agrees very well with the intuitive notion that parallelizing serves as regularization.

We emphasize that numerical results seem to indicate that parallelization is possible to a slightly larger degree than indicated by our theoretical estimate. A similar result was reported in the paper Zhang et al. (2013), which also treats the low smoothness case.

4.2 High smoothness regime

We choose $f_\rho(x) = \frac{1}{2\pi} \sin(2\pi x)$, which corresponds to just one non-vanishing Fourier coefficient and by our criterion Corollary 11 has $r = \infty$. In view of our main Corollary 5 this requires a regularization method with higher qualification; we take the *Gradient Descent* method (see Example 2).

The appearance of the term $2b \min\{1, r\}$ in our theoretical result 5 gives a predicted value $\alpha = 0$ (and would imply that parallelization is strictly forbidden for infinite smoothness). More specifically, the left panel in Figure 3 shows the absence of any plateau for the reconstruction error as a function of α. This corresponds to the right panel showing that

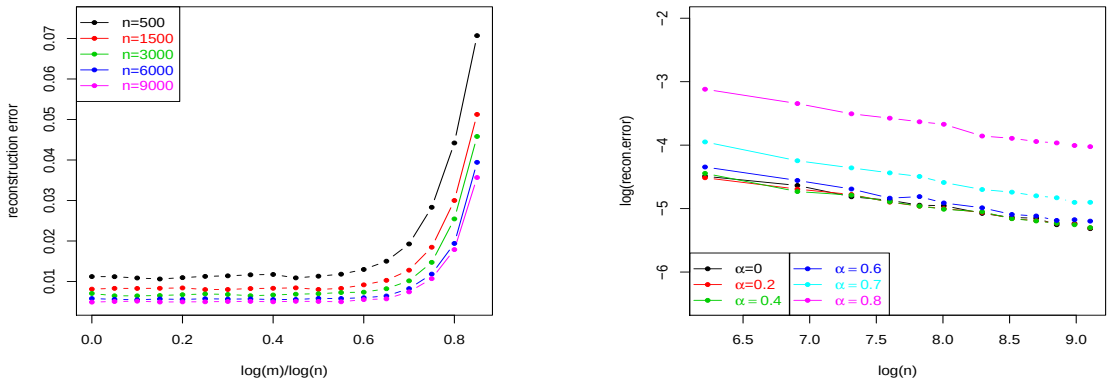

Figure 1: The reconstruction error $\|\bar{f}_D^{k_{oracle}} - f_\rho\|_{\mathcal{H}_K}$ in the low smoothness case. Left plot: Reconstruction error curves for various (but fixed) total sample sizes, as a function of the number m of subsamples. Right plot: Reconstruction error curves for various subsample number scalings $m = n^\alpha$, as a function of the sample size (on log-scale).

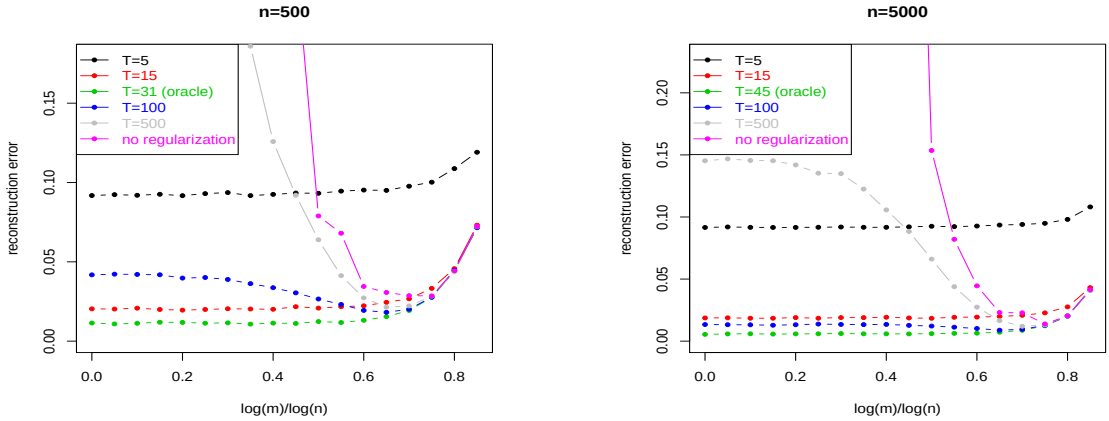

Figure 2: The reconstruction error $\|\bar{f}_D^{\lambda} - f_\rho\|_{\mathcal{H}_K}$ in the low smoothness case. Left plot: Error curves for different stopping times for $n = 500$, as a function of the number m of subsamples. Right plot: Error curves for different stopping times for $n = 5000$, as a function of the number m of subsamples.

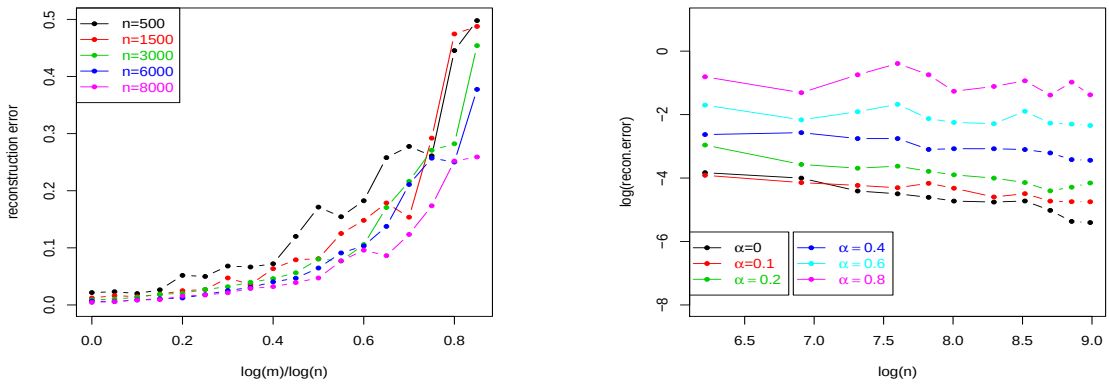

Figure 3: The reconstruction error $\|\bar{f}_D^{\lambda_{oracle}} - f_\rho\|_{\mathcal{H}_K}$ in the high smoothness case. Left plot: Reconstruction error curves for various (but fixed) sample sizes as a function of the number m of subsamples. Right plot: Reconstruction error curves for various subsample number scalings $m = n^\alpha$, as a function of the sample size (on log-scale).

no group of values of α performs roughly equivalently, meaning that we do not have any optimality guarantees.

Plotting different values of regularization in Figure 4 we again identify overregularization as hopeless, while severe underregularization exhibits a sharp minimum in the reconstruction error. But its value at roughly 0.25 is much less attractive compared to the case of low smoothness where the error is an order of magnitude less.

5. Discussion

Minimax Optimality: We have shown that for a large class of spectral regularization methods the error of the distributed algorithm $\|\bar{T}^s(\bar{f}_D^{\lambda_n} - f_\rho)\|_{\mathcal{H}_K}$ satisfies the same upper bound as the error $\left\|\bar{T}^s(f_D^{\lambda_n} - f_\rho)\right\|_{\mathcal{H}_K}$ for the single machine problem, if the regularization parameter λ_n is chosen according to (16), provided the number of subsamples grows sufficiently slowly with the sample size n. Since, the rates for the latter are minimax optimal (Blanchard and Mücke, 2017), our rates in Corollary 5 are minimax optimal also.

Comparison with other results: Zhang et al. (2013) derive minimax-optimal rates in three settings: finite rank kernels, sub-Gaussian decay of eigenvalues of the kernel and polynomial decay, provided m satisfies a certain upper bound, depending on the rate of decay of the eigenvalues under two crucial assumptions on the eigenfunctions of the integral operator associated to the kernel: For any $j \in \mathbb{N}$

$$\mathbb{E}[\phi_j(X)^{2k}] \leq \rho^{2k} \,, \tag{22}$$

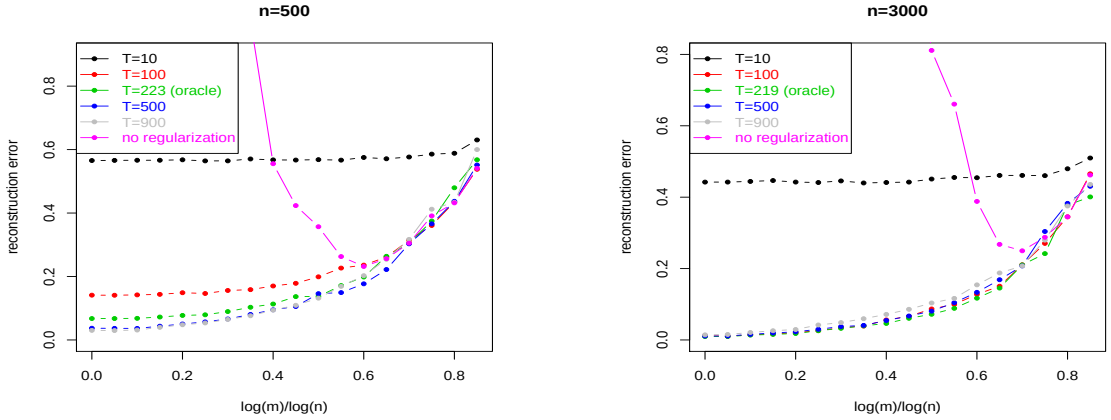

Figure 4: The reconstruction error $\|\bar{f}_D^{\lambda} - f_\rho\|$ in the high smoothness case. Left plot: Error curves for different stopping times for $n = 500$ samples, as a function of the number of subsamples. Right plot: Error curves for different stopping times for $n = 5000$ samples, as a function of the number of subsamples.

for some $k \geq 2$ and $\rho < \infty$ or even stronger, it is assumed that the eigenfunctions are uniformly bounded, i.e.

$$\sup_{x \in \mathcal{X}} |\phi_j(x)| \leq \rho \,, \tag{23}$$

or any $j \in \mathbb{N}$ and some $\rho < \infty$. We shall describe in more detail the case of polynomially decaying eigenvalues, which corresponds to our setting. Assuming eigenvalue decay $\mu_j \lesssim j^{-b}$ with $b > 1$, the authors choose a regularization parameter $\lambda_n = n^{-\frac{b}{b+1}}$ and

$$m \lesssim \left(\frac{n^{\frac{b(k-4)-k}{b+1}}}{\rho^{4k} \log^k(n)} \right)^{\frac{1}{k-2}} \,.$$

leading to an error in L^2-norm

$$\mathbb{E}[\|\bar{f}_D^{\lambda_n} - f_\rho\|_{L^2}^2] \lesssim n^{-\frac{b}{b+1}} \,,$$

being minimax optimal. Note that this choice of λ_n and the resulting rate correspond to our case $r = 0$, i.e., no smoothness of f_ρ is assumed (just that f_ρ belongs to the RKHS).

For $k < 4$ the bound becomes less meaningful (compared to the case where $k \geq 4$) since $m \to 0$ as $n \to \infty$ in this case (for any sort of eigenvalue decay). On the other hand, if k and b might be taken arbitrarily large, corresponding to almost bounded eigenfunctions and arbitrarily large polynomial decay of eigenvalues, m might be chosen proportional to $n^{1-\epsilon}$, for any $\epsilon > 0$. As might be expected, replacing the L^{2k} bound on the eigenfunctions by a bound in L^∞, gives an upper bound on m which simply is the limit for $k \to \infty$ in the

bound given above, namely

$$m \lesssim \frac{n^{\frac{b-1}{b+1}}}{\rho^4 \log n} ,$$

which for large b behaves as above. Granted bounds on the eigenfunctions in L^{2k} for (very) large k, this is a strong result. While the decay rate of the eigenvalues can be determined by the smoothness of K (see, e.g., Ferreira and Menegatto, 2009 and references therein), it is a widely open question which general properties of the kernel imply estimates as in (22) and (23) on the eigenfunctions.

Zhou (2002) even gives a counterexample and presents a C^∞ Mercer kernel on $[0, 1]$ where the eigenfunctions of the corresponding integral operator are *not* uniformly bounded. Thus, smoothness of the kernel is not a sufficient condition for (23) to hold.

Moreover, we point out that the upper bound (22) on the eigenfunctions (and thus the upper bound for m in Zhang et al., 2013) depends on the unknown marginal distribution ν. Only the strongest assumption, a bound in sup-norm (23), does not depend on ν. Concerning this point, our approach is "agnostic".

As already mentioned in the Introduction, these bounds on the eigenfunctions have been eliminated by Lin et al. (2017), for KRR, imposing polynomial decay of eigenvalues as above. This is very similar to our approach. As a general rule, our bounds on m and the bounds obtained by Lin et al. (2017) are worse than the bounds of Zhang et al. (2013) for eigenfunctions in (or close to) L^∞, but in the complementary case where nothing is known on the eigenfunctions m still can be chosen as an increasing function of n, namely $m = n^\alpha$. More precisely, choosing λ_n as in (16), Lin et al. (2017) derive as an upper bound

$$m \lesssim n^\alpha , \qquad \alpha = \frac{2br}{2br + b + 1} ,$$

with r being the smoothness parameter arising in the source condition. We recall here that due to our assumption $q \geq r + s$, the smoothness parameter r is restricted to the interval $(0, \frac{1}{2}]$ for KRR ($q = 1$) and L^2-risk ($s = \frac{1}{2}$).

Our results (which hold for a general class of spectral regularization methods) are in some ways comparable to those of Lin et al. (2017). Specialized to KRR, our estimates for the exponent α in $m = O(n^\alpha)$ coincide with the result of Lin et al. (2017). Furthermore, we emphasize that Zhang et al. (2013) and Lin et al. (2017) estimate the DL-error only for $s = 1/2$ in our notation (corresponding to $L^2(\nu)-$norm), while our result holds for all values of $s \in [0, 1/2]$ which smoothly interpolates between $L^2(\nu)$-norm and RKHS-norm and, in addition, for all values of $p \in [1, \infty)$. Thus, our results also apply to the case of non-parametric inverse regression, where one is particularly interested in the reconstruction error, i.e. $\mathcal{H}_K$-norm (see, e.g., Blanchard and Mücke, 2017). Additionally, we precisely analyze the dependence of the noise variance σ^2 and the complexity radius R in the source condition.

Concerning general strategy, while Lin et al. (2017) use a novel second order decomposition in an essential way, our approach is more classical. We clearly distinguish between estimat-

ing the approximation error and the sample error. The bias using a subsample should be of the same order as when using the whole sample, whereas the estimation error is higher on each subsample, but gets reduced by averaging by writing the variance as a sum of i.i.d. random variables (which allows to use Rosenthal's inequality).

Finally, we want to mention the recent works of Lin and Zhou (2018) and Guo et al. (2017), which were worked out indepently from our work. Guo et al. (2017) also treat general spectral regularization methods (going beyond kernel ridge) and obtain essentially the same results, but with error bounds only in L^2-norm, excluding inverse learning problems. Lin and Zhou (2018) investigate distributed learning on the example of gradient descent algorithms, which have infinite qualification and allow larger smoothness of the regression function. They are able to improve the upper bound for the number of local machines to

$$m \lesssim \frac{n^\alpha}{\log^5(n)+1}\,, \qquad \alpha < \frac{br}{2br+b+1},$$

which is larger in the case $r > 2$. In the intermediate case $1 < r < 2$, our bound in (20) is still better. An interesting feature is the fact that it is possible to allow more local machines by using additional unlabeled data. This indicates that finding the upper bound for the number of machines in the high smoothness regime is still an open problem.

Adaptivity: It is clear from the theoretical results that both the regularization parameter λ and the allowed cardinality of subsamples m depend on the parameters r and b, which in general are unknown. Thus, an adaptive approach to both parameters b and r for choosing λ and m is of interest. To the best of our knowledge, there are yet no rigorous results on adaptivity in this more general sense. Progress in this field may well be crucial in finally assessing the relative merits of the distributed learning approach as compared with alternative strategies to effectively deal with large data sets.

We sketch an alternative naive approach to adaptivity, based on hold-out in the direct case, where we consider each $f \in \mathcal{H}_K$ also as a function in $L^2(\mathcal{X},\nu)$. We split the data $\mathbf{z} \in (\mathcal{X}\times\mathcal{Y})^n$ into a training and validation part $\mathbf{z} = (\mathbf{z}^t, \mathbf{z}^v)$ of cardinality m_t, m_v. We further subdivide $\mathbf{z}^t$ into m_k subsamples, roughly of size m_t/m_k, where $m_k \le m_t$, $k = 1,2,\dots$ is some strictly decreasing sequence. For each k and each subsample $\mathbf{z}_j$, $1 \le j \le m_k$, we define the estimators $\hat{f}^\lambda_{\mathbf{z}_j}$ as in (12) and their average

$$\bar{f}^\lambda_{k,\mathbf{z}^t} := \frac{1}{m_k}\sum_{j=1}^{m_k} \hat{f}^\lambda_{\mathbf{z}_j}\,. \tag{24}$$

Here, λ varies in some sufficiently fine lattice Λ. Then evaluation on $\mathbf{z}^v$ gives the associated empirical L^2- error

$$\mathrm{Err}^\lambda_k(\mathbf{z}^v) := \frac{1}{m_v}\sum_{i=1}^{m_v}(y^v_i - \bar{f}^\lambda_{k,\mathbf{z}^t}(x^v_i))^2\,, \quad \mathbf{z}^v = (\mathbf{y}^v, \mathbf{x}^v)\,, \quad \mathbf{y}^v = (y^v_1,\dots,y^v_{m_v})\,, \tag{25}$$

leading us to define

$$\hat{\lambda}_k := \mathrm{argmin}_{\lambda\in\Lambda}\mathrm{Err}^\lambda_k(\mathbf{z}^v)\,, \qquad \mathrm{Err}(k) := \mathrm{Err}^{\hat{\lambda}_k}_k(\mathbf{z}^v). \tag{26}$$

Then, an appropriate stopping criterion for k might be to stop at

$$k^* := \min\{k \geq 3 \ : \ \Delta(k) \leq \delta \inf_{2\leq j<k} \Delta(j)\} \ , \qquad \Delta(j) := |\mathrm{Err}(j) - \mathrm{Err}(j-1)| \ , \tag{27}$$

for some $\delta < 1$ (which might require tuning). The corresponding regularization parameter is $\hat{\lambda} = \hat{\lambda}_{k^*}$, given by (26). At least intuitively, it is then reasonable to define a purely data driven estimator as

$$\widehat{f}_n := \bar{f}^{\hat{\lambda}}_{k^*,\mathbf{z}^t} \ . \tag{28}$$

Note that the training data $\mathbf{z}^t$ enter the definition of $\widehat{f}_n$ via the explicit formula (24) encoding our kernel based approach, while $\mathbf{z}^v$ serves to determine $(k^*, \hat{\lambda}^*)$ via minimization of the empirical L^2-error and a criterion, which tells one to stop where $\mathrm{Err}(j)$ does not appreciably improve anymore. It is open if such a procedure achieves optimal rates, and we leave this for future research.

6. Proofs

For ease of reading we make use of the following conventions:

- for a (bounded) linear operator A, $\|A\|$ denotes the operator norm;
- we are interested in a precise dependence of multiplicative constants on the parameters σ, M, R, m, n and η. (To be clear about the role of the latter quantity: the proofs rely on high-probability statements on deviations, typically holding with high probability $1-\eta$.)
- the dependence of multiplicative constants on various other parameters, including the kernel parameter κ, the interpolating parameter $s \in [0, \frac{1}{2}]$, the parameters arising from the regularization method, $b > 1$, $\beta > 0$, $r > 0$, etc. will (generally) be omitted and simply indicated by the symbol $\blacktriangle$.
- the dependence of the norm parameter p will also be indicated, but will not be given explicitly.
- the values of $C_{\blacktriangle}$ and $C_{\blacktriangle,p}$ might change from line to line.
- the expression "for n sufficiently large" means that the statement holds for $n \geq n_0$, with n_0 potentially depending on all model parameters (including σ, M and R), but not on η.

6.1 Preliminaries

Proposition 12 (Guo et al., 2017, Proposition 1) *Define*

$$\mathcal{B}_n(\lambda) := \left[1 + \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}}\right)^2\right] . \tag{29}$$

For any $\lambda > 0$, $\eta \in (0,1]$, with probability at least $1-\eta$ one has

$$\left\|(\bar{T}_{\mathbf{x}} + \lambda)^{-1}(\bar{T} + \lambda)\right\| \leq 8 \log^2(2\eta^{-1}) \mathcal{B}_n(\lambda) \; . \tag{30}$$

Corollary 13 *Let $\eta \in (0,1)$. For $n \in \mathbb{N}$ let $\tilde{\lambda}_n$ be implicitly defined as the unique solution of $\mathcal{N}(\tilde{\lambda}_n) = n\tilde{\lambda}_n$. Then for any $\lambda \in [\max(\tilde{\lambda}_n, n^{-1}), 1]$, one has*

$$\mathcal{B}_n(\lambda) \leq 10 \; .$$

In particular,

$$\left\|(\bar{T}_{\mathbf{x}} + \lambda)^{-1}(\bar{T} + \lambda)\right\| \leq 80 \log^2(2\eta^{-1})$$

holds with probability at least $1-\eta$.

We remark that the trace of $\bar{T}$ is bounded by 1. This ensures that the interval $[\tilde{\lambda}_n, 1]$ is non-empty.

Proof [of Corollary 13] Let $\tilde{\lambda}_n$ be defined via $\mathcal{N}(\tilde{\lambda}_n) = n\tilde{\lambda}_n$. Since $\mathcal{N}(\lambda)/\lambda$ is decreasing, we have for any $\lambda \geq \tilde{\lambda}_n$

$$\sqrt{\frac{\mathcal{N}(\lambda)}{n\lambda}} \leq \sqrt{\frac{\mathcal{N}(\tilde{\lambda}_n)}{n\tilde{\lambda}_n}} = 1 \; .$$

Inserting this bound as well as $n\lambda \geq 1$ into (29) and (30) leads to the conclusion. ∎

Corollary 14 *Assume the marginal distribution ν of $\mathcal{X}$ belongs to $\mathcal{P}^<(b,\beta)$ with $b > 1$ and $\beta > 0$. If λ_n is defined by (16) and if*

$$m_n \leq n^\alpha \; , \qquad \alpha < \frac{2br}{2br + b + 1} \; ,$$

one has

$$\mathcal{B}_{\frac{n}{m_n}}(\lambda_n) \leq 2 \; ,$$

provided n is sufficiently large.

Proof [of Corollary 14] We can for starters assume that n is sufficiently large so that $\lambda_n < 1$, i.e. $\lambda_n = \left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b}{2br+b+1}}$ from (16). Recall that $\nu \in \mathcal{P}^<(b,\beta)$ implies $\mathcal{N}(\lambda_n) \leq C_\blacktriangle \lambda_n^{-\frac{1}{b}}$. Looking at the terms entering in $\mathcal{B}_{\frac{n}{m}}(\lambda_n)$, see (29), we have first, using the definition of λ_n in (16):

$$\frac{\mathcal{N}(\lambda_n)}{\frac{n}{m}\lambda_n} \leq C_\blacktriangle m \frac{\lambda_n^{-\frac{b+1}{b}}}{n} = C_\blacktriangle \frac{m}{n} \left(\frac{nR^2}{\sigma^2}\right)^{\frac{b+1}{2br+b+1}} \; ,$$

which (for fixed R, σ and other parameters entering in $C_\blacktriangle$) is $O(m_n n^{-\frac{2br}{2br+r+1}})$, and hence $o(1)$ provided

$$m_n \leq n^\alpha \,, \qquad \alpha < \frac{2br}{2br+b+1} \,.$$

For the second term entering in $\mathcal{B}_{\frac{n}{m}}(\lambda_n)$, we have

$$\frac{1}{\frac{n}{m}\lambda_n} = \frac{m}{n}\left(\frac{nR^2}{\sigma^2}\right)^{\frac{b}{2br+b+1}} \,,$$

which is $O(m_n n^{-\frac{2br+1}{2br+b+1}}) = o(1)$, provided

$$m_n \leq n^\alpha \,, \qquad \alpha < \frac{2br+1}{2br+b+1} \,,$$

which is implied by the previous stronger condition. ■

We shortly illustrate how Corollary 13 and Proposition 12 will be used. Let $u \in [0,1]$, $\tilde{\lambda}_n \leq \lambda$ as above and $f \in \mathcal{H}_K$. We have for any bounded operator A

$$\begin{aligned} \left\|\bar{T}^u A\right\| &= \left\|\bar{T}^u(\bar{T}+\lambda)^{-u}(\bar{T}+\lambda)^u(\bar{T}_{\mathbf{x}}+\lambda)^{-u}(\bar{T}_{\mathbf{x}}+\lambda)^u A\right\| \\ &\leq \left\|\bar{T}^u(\bar{T}+\lambda)^{-u}\right\| \left\|(\bar{T}+\lambda)^u(\bar{T}_{\mathbf{x}}+\lambda)^{-u}\right\| \left\|(\bar{T}_{\mathbf{x}}+\lambda)^u A\right\| \\ &\leq 8\log^{2u}(2\eta^{-1})\mathcal{B}_n(\lambda)^u \left\|(\bar{T}_{\mathbf{x}}+\lambda)^u A\right\| \,, \end{aligned} \tag{31}$$

with probability at least $1-\eta$, for any $\eta \in (0,1)$; for the last inequality we have used that the first factor is less than 1, and for the second factor Proposition 12 in combination with the Cordes inequality (see Proposition 22 in the Appendix). In particular, for any $\max(\tilde{\lambda}_n, n^{-1}) \leq \lambda$ (with $\tilde{\lambda}_n$ as in Corollary 13)

$$\left\|\bar{T}^u A\right\| \leq 80^u \log^{2u}(2\eta^{-1}) \left\|(\bar{T}_{\mathbf{x}}+\lambda)^u A\right\| \,, \tag{32}$$

with probability at least $1-\eta$.

In the following, we constantly use (31). Furthermore, to bound terms involving residuals we will frequently use the following estimate: for $v \geq 0, u \in [0,2]$, and provided $u+v \leq q$ (q being the qualification):

$$\begin{aligned} \sup_{t\in[0,1]} |r_\lambda(t)t^v(t+\lambda)^u| &\leq 2\Big(\sup_{t\in[0,1]} \left|r_\lambda(t)t^{v+u}\right| + \lambda^u \sup_{t\in[0,1]} |r_\lambda(t)t^v|\Big) \\ &\leq C_\blacktriangle \lambda^{v+u}, \end{aligned} \tag{33}$$

using twice (11) since $q \geq u+v$.

6.2 Approximation error bound

Recall that ν denotes the X-marginal of the sampling distribution ρ and $\mathcal{P}$ the set of all probability distributions on the input space $\mathcal{X}$.

Lemma 15 *Let $\nu \in \mathcal{P}$, $v \in \mathbb{R}$ and let $\mathbf{x} \in \mathcal{X}^{\frac{n}{m}}$ be an i.i.d. sample of size n/m, drawn according to ν. Assume the regularization $(g_\lambda)_\lambda$ has qualification $q \geq v+1+s$. Then with probability at least $1-\eta$:*

$$\left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}})\bar{T}_{\mathbf{x}}^v(\bar{T}-\bar{T}_{\mathbf{x}})\right\| \leq C_\blacktriangle \log^4(4\eta^{-1})\lambda^{s+v+1}\mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda)\left(\frac{m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right).$$

Proof [of Lemma 15] From (30),(31) and from Proposition 20 recalled in the Appendix, one has

$$\begin{aligned}\left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}})\bar{T}_{\mathbf{x}}^v(\bar{T}-\bar{T}_{\mathbf{x}})\right\| &\leq C_\blacktriangle \log^{2(s+1)}(4\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda) \\ &\qquad \left\|(\bar{T}_{\mathbf{x}}+\lambda)^s r_\lambda(\bar{T}_{\mathbf{x}})\bar{T}_{\mathbf{x}}^v(\bar{T}_{\mathbf{x}}+\lambda)\right\| \left\|(\bar{T}+\lambda)^{-1}(\bar{T}-\bar{T}_{\mathbf{x}})\right\| \\ &\leq C_\blacktriangle \log^4(4\eta^{-1})\lambda^{s+v+1}\mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda)\left(\frac{m}{n\lambda}+\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right),\end{aligned}$$

for any $\lambda \in (0,1]$, $\eta \in (0,1]$, with probability at least $1-\eta$. We also used that $s \leq \frac{1}{2}$, and the estimate (33). ■

Lemma 16 *Let $\nu \in \mathcal{P}$, $v \in \mathbb{R}$ and let $\mathbf{x} \in \mathcal{X}^{\frac{n}{m}}$ be an i.i.d. sample of size n/m drawn according to ν. Assume the regularization $(g_\lambda)_\lambda$ has qualification $q \geq v+s$. Then for any $\lambda \in (0,1]$, $\eta \in (0,1]$, with probability at least $1-\eta$*

$$\left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}})\bar{T}_{\mathbf{x}}^v\right\| \leq C_\blacktriangle \log^{2s}(2\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s}(\lambda)\lambda^{s+v},$$

for some $C_\blacktriangle < \infty$.

Proof [of Lemma 16] Using (31), (33), since $q \geq v+s$, it holds

$$\begin{aligned}\left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}})\bar{T}_{\mathbf{x}}^v\right\| &\leq C_\blacktriangle \log^{2s}(2\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s}(\lambda)\left\|(\bar{T}_{\mathbf{x}}+\lambda)^s r_\lambda(\bar{T}_{\mathbf{x}})\bar{T}_{\mathbf{x}}^v\right\| \\ &\leq C_\blacktriangle \log^{2s}(2\eta^{-1})\mathcal{B}_{\frac{n}{m}}^{s}(\lambda)\lambda^{s+v},\end{aligned}$$

with probability at least $1-\eta$. ■

Proposition 17 (Expectation of approximation error) *Let $f_\rho \in \Omega(r,R)$, $\lambda \in (0,1]$ and let $\mathcal{B}_{\frac{n}{m}}(\lambda)$ be defined in (29). Assume the regularization has qualification $q \geq r+s$. For any $p \geq 1$ one has:*

1. *If* $r \leq 1$, *then*

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^\lambda) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} R \, \lambda^{s+r} \, \mathcal{B}_{\frac{n}{m}}^{s+r}(\lambda) \, .$$

2. *If* $r > 1$, *then*

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^\lambda) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} R \lambda^s \mathcal{B}_{\frac{n}{m}}^{s+1}(\lambda) \left(\lambda^r + \lambda \left(\frac{m}{n\lambda} + \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}} \right) \right) \, .$$

In 1. and 2. the constant $C_{\blacktriangle,p}$ *does not depend on* $(\sigma, M, R) \in \mathbb{R}_+^3$.

Proof [of Proposition 17] Since $f_\rho \in \Omega(r, R)$,

$$\begin{aligned}
\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^\lambda) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} &= \left[\mathbb{E}_{\rho^{\otimes n}} \left\| \frac{1}{m} \sum_{j=1}^m \bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) f_\rho \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} \\
&\leq \frac{1}{m} \sum_{j=1}^m \left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) f_\rho \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} \\
&\leq \frac{R}{m} \sum_{j=1}^m \left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^r \right\|^p \right]^{\frac{1}{p}} \, . \qquad (34)
\end{aligned}$$

The first inequality is just the triangle inequality for the p-norm $\|f\|_p = \mathbb{E}[\|f\|_{\mathcal{H}_K}^p]^{\frac{1}{p}}$. We bound the expectation for each separate subsample of size $\frac{n}{m}$ by first deriving a probabilistic estimate and then we integrate.

Consider first the case where $r \leq 1$. Using (31), the Cordes inequality (Proposition 22 in the Appendix), and (33) one has for any $j = 1, \ldots, m$,

$$\begin{aligned}
\left\| \bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^r \right\| &\leq C_\blacktriangle \log^{2(s+r)}(4\eta^{-1}) \mathcal{B}_{\frac{n}{m}}^{s+r}(\lambda) \left\| (\bar{T}_{\mathbf{x}_j} + \lambda)^s r_\lambda(\bar{T}_{\mathbf{x}_j}) (\bar{T}_{\mathbf{x}_j} + \lambda)^r \right\| \\
&\leq C_\blacktriangle \log^3(4\eta^{-1}) \lambda^{s+r} \mathcal{B}_{\frac{n}{m}}^{s+r}(\lambda) \, ,
\end{aligned}$$

with probability at least $1 - \eta$ and where $\mathcal{B}_{\frac{n}{m}}(\lambda)$ is defined in (29). Recall that the regularization has qualification $q \geq r + s$. By integration one has

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^r \right\|^p \right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} \, \lambda^{s+r} \, \mathcal{B}_{\frac{n}{m}}^{s+r}(\lambda) \, ,$$

for some $C_{\blacktriangle,p} < \infty$, not depending on σ, M, R. Finally, from (34)

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^\lambda) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} R \, \lambda^{s+r} \, \mathcal{B}_{\frac{n}{m}}^{s+r}(\lambda) \, .$$

In the case where $r \geq 1$, we write $r = k + u$, with $k = \lfloor r \rfloor$ and $u = r - k < 1$. We shall use the decomposition

$$\bar{T}^k = \sum_{l=0}^{k-1} \bar{T}_{\mathbf{x}}^l (\bar{T} - \bar{T}_{\mathbf{x}}) \bar{T}^{k-(l+1)} + \bar{T}_{\mathbf{x}}^k \, . \qquad (35)$$

We proceed by bounding (34) according to decomposition (35) . For any $j = 1, \dots, m$, one has

$$\begin{aligned}\left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^{k+u}\right\|^p\right]^{\frac{1}{p}} &\leq \sum_{l=0}^{k-1} \left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^l_{\mathbf{x}_j} (\bar{T} - \bar{T}_{\mathbf{x}_j}) \bar{T}^{k-(l+1)+u}\right\|^p\right]^{\frac{1}{p}} \\ &\quad + \left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^k_{\mathbf{x}_j} \bar{T}^u\right\|^p\right]^{\frac{1}{p}} \\ &\leq \sum_{l=0}^{k-1} \left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^l_{\mathbf{x}_j} (\bar{T} - \bar{T}_{\mathbf{x}_j})\right\|^p\right]^{\frac{1}{p}} \\ &\quad + \left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^k_{\mathbf{x}_j} \bar{T}^u\right\|^p\right]^{\frac{1}{p}} . \end{aligned} \tag{36}$$

Here we use that $\left\|\bar{T}^{k-(l+1)+u}\right\|$ is bounded by 1. By Lemma 16 and by (31), (33), with probability at least $1 - \eta$

$$\begin{aligned}\left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^k_{\mathbf{x}_j} \bar{T}^u\right\| &\leq C_{\blacktriangle} \log^{2(s+u)}(2\eta^{-1}) \mathcal{B}^{s+u}_{\frac{n}{m}}(\lambda) \left\|(\bar{T}_{\mathbf{x}_j} + \lambda)^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^k_{\mathbf{x}_j} (\bar{T}_{\mathbf{x}_j} + \lambda)^u\right\| \\ &\leq C_{\blacktriangle} \log^{2(s+u)}(2\eta^{-1}) \mathcal{B}^{s+u}_{\frac{n}{m}}(\lambda) \lambda^{s+r},\end{aligned}$$

and thus integration yields

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^r_{\mathbf{x}_j} \bar{T}^u\right\|^p\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} \mathcal{B}^{s+u}_{\frac{n}{m}}(\lambda) \lambda^{s+r} . \tag{37}$$

For estimating the first term in (36) we may use Lemma 15. For any $l = 0, \dots, k-1$, we have $l + s + 1 \leq k + s \leq r + s \leq q$, hence for any $j = 1, \dots, m$ with probability at least $1 - \eta$

$$\left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^l_{\mathbf{x}_j} (\bar{T} - \bar{T}_{\mathbf{x}_j})\right\| \leq C_{\blacktriangle} \log^4(8\eta^{-1}) \lambda^{s+l+1} \mathcal{B}^{s+1}_{\frac{n}{m}}(\lambda) \left(\frac{m}{n\lambda} + \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) .$$

Again by integration, since $\lambda^l \leq 1$ for any $l = 0, \dots, k-1$, one has

$$\sum_{l=0}^{k-1} \left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s r_\lambda(\bar{T}_{\mathbf{x}_j}) \bar{T}^l_{\mathbf{x}_j} (\bar{T} - \bar{T}_{\mathbf{x}_j})\right\|^p\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} \lfloor r \rfloor \lambda^{s+1} \mathcal{B}^{s+1}_{\frac{n}{m}}(\lambda) \left(\frac{m}{n\lambda} + \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) . \tag{38}$$

Finally, combining (37) and (38) into (36), then (34), gives in the case where $r > 1$

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\|\bar{T}^s (f_\rho - \tilde{f}^\lambda_D)\right\|^p_{\mathcal{H}_K}\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} \lambda^s \mathcal{B}^{s+1}_{\frac{n}{m}}(\lambda) \left(\lambda^r + \lambda \left(\frac{m}{n\lambda} + \sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right)\right) .$$

The rest of the proof follows from (36).

■

Proof [of Theorem 3] Let λ_n be as defined by (16). According to Corollary 14, we have $\mathcal{B}_{\frac{n}{m_n}}(\lambda_n) \leq 2$ provided $\alpha < \frac{2br}{2br+b+1}$, for n sufficiently large. We can also assume n sufficiently large so that $\lambda_n < 1$, i.e., $R\lambda_n^{r+s} = a_n$ (from (16), (17)). Under these conditions, we immediately obtain from the first part of Proposition 17 in the case where $r \leq 1$

$$\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^{\lambda_n}) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} \leq C_{\blacktriangle,p} R \, \lambda_n^{s+r} = C_{\blacktriangle,p} \, a_n \, .$$

We turn to the case where $r > 1$. We apply the second part of Proposition 17. By Corollary 14 we have

$$\begin{aligned}\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^{\lambda_n}) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}} &\leq C_{\blacktriangle,p} R \lambda_n^s \mathcal{B}_{\frac{n}{m_n}}^{s+1}(\lambda_n) \left(\lambda_n^r + \lambda_n \left(\frac{m_n}{n\lambda_n} + \sqrt{\frac{m_n \mathcal{N}(\lambda_n)}{n\lambda_n}} \right) \right) \\ &\leq C_{\blacktriangle,p} \, R\lambda_n^s \, \left(\lambda_n^r + \lambda_n \left(\frac{m_n}{n\lambda_n} + \sqrt{m_n} \frac{R}{\sigma} \lambda_n^r \right) \right),\end{aligned}$$

where we used that $\mathcal{N}(\lambda_n) \leq C_{\blacktriangle} \lambda_n^{-1/b}$ and $\sigma \sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{n\lambda_n}} = R\lambda_n^r$ coming from the definition of λ_n, and $\lambda_n < 1$. Furthermore,

$$\frac{m_n}{n\lambda_n} = o\left(\sqrt{m}_n \lambda_n^r\right) \, ,$$

provided

$$m_n \leq n^\alpha \, , \qquad \alpha < \frac{2(br+1)}{2br+b+1} \, .$$

Finally, for n sufficiently large, $\frac{R}{\sigma}\sqrt{m_n}\lambda_n \leq 1$, provided that

$$\alpha < \frac{2b}{2br+b+1} \, .$$

As a result, for any $p \geq 1$:

$$\limsup_{n\to\infty} \sup_{\rho \in \mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}} \left\| \bar{T}^s (f_\rho - \tilde{f}_D^{\lambda_n}) \right\|_{\mathcal{H}_K}^p \right]^{\frac{1}{p}}}{a_n} \leq C_{\blacktriangle,p} \, ,$$

for some $C_{\blacktriangle,p} < \infty$, not depending on σ, M, R. ■

6.3 Sample error bound

The main idea for deriving an upper bound for the sample error is to identify it as a sum of unbiased Hilbert space-valued i.i.d. variables and then to apply a suitable version of Rosenthal's inequality.

Given $\lambda \in (0,1]$, we define the random variable $\xi_\lambda : (\mathcal{X} \times \mathbb{R})^{\frac{n}{m}} \longrightarrow \mathcal{H}_K$ by

$$\xi_\lambda(\mathbf{x}, \mathbf{y}) := \bar{T}^s g_\lambda(\bar{T}_{\mathbf{x}})(\bar{T}_{\mathbf{x}} f_\rho - \bar{S}^*_{\mathbf{x}}\mathbf{y}) \, .$$

Recall that according to Assumption (3), the conditional expectation w.r.t. ρ of Y given X satisfies

$$\mathbb{E}_\rho[Y|X = x] = \bar{S}_x f_\rho \, ,$$

implying that ξ_λ has zero expectation (since $\bar{T}_{\mathbf{x}} = \bar{S}^*_{\mathbf{x}}\bar{S}_{\mathbf{x}}$). Thus,

$$\bar{T}^s(\tilde{f}^\lambda_D - \bar{f}^\lambda_D) \;=\; \frac{1}{m}\sum_{j=1}^m \xi_\lambda(\mathbf{x}_j, \mathbf{y}_j) \tag{39}$$

is a sum of centered i.i.d. random variables.

Furthermore, we need the following result (Pinelis, 1994, Theorem 5.2) , which generalizes Rosenthal's (1970) inequalities (originally only formulated for real valued random variables) to random variables with values in a Banach space. For Hilbert spaces this looks particularly nice.

Proposition 18 *Let $\mathcal{H}$ be a Hilbert space and $\xi_1, \ldots, \xi_m$ be a finite sequence of independent, mean zero $\mathcal{H}$- valued random variables. If $2 \le p < \infty$, then there exists a constant $C_p > 0$, only depending on p, such that*

$$\left(\mathbb{E}\Big\| \frac{1}{m}\sum_{j=1}^m \xi_j \Big\|^p_{\mathcal{H}}\right)^{\frac{1}{p}} \;\le\; \frac{C_p}{m} \;\max\left\{\left(\sum_{j=1}^m \mathbb{E}\|\xi_j\|^p_{\mathcal{H}}\right)^{\frac{1}{p}}, \left(\sum_{j=1}^m \mathbb{E}\|\xi_j\|^2_{\mathcal{H}}\right)^{\frac{1}{2}}\right\} . \tag{40}$$

We remark in passing that Dirksen (2011), Corollary 1.22, establishes the interesting result that in addition to the upper bound in (40) there is also a corresponding lower bound where the constant C_p is replaced by another constant $C'_p > 0$, only depending on p.

Proposition 19 (Expectation of sample error) *Let ρ be a source distribution belonging to $\mathcal{M}_{\sigma,M,R}$, $s \in [0, \frac{1}{2}]$ and let $\lambda \in (0,1]$. Define $\mathcal{B}_{\frac{n}{m}}(\lambda)$ as in (29). Assume the regularization has qualification $q \ge r + s$. For any $p \ge 1$ one has:*

$$\left[\mathbb{E}_{\rho^{\otimes n}}\big\|\bar{T}^s(\tilde{f}^\lambda_D - \bar{f}^\lambda_D)\big\|^p_{\mathcal{H}_K}\right]^{\frac{1}{p}} \le C_{\blacktriangle,p}\, m^{-\frac{1}{2}} \mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}+s}\lambda^s \;\left(\frac{mM}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) ,$$

where C_p does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$.

Proof [of Proposition 19] Let $\lambda \in (0,1]$ and $p \ge 2$. From Proposition 18

$$\begin{aligned}\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s\tilde{f}^\lambda_D - \bar{f}^\lambda_D\right\|^p_{\mathcal{H}_K}\right]^{\frac{1}{p}} &= \left[\mathbb{E}_{\rho^{\otimes n}}\left\|\frac{1}{m}\sum_{j=1}^m \xi_\lambda(\mathbf{x}_j, \mathbf{y}_j)\right\|^p_{\mathcal{H}_K}\right]^{\frac{1}{p}} \\ &\le \frac{C_p}{m}\;\max\left\{\left(\sum_{j=1}^m \left[\mathbb{E}_{\rho^{\otimes n}}\|\xi_\lambda(\mathbf{x}_j, \mathbf{y}_j)\|^p_{\mathcal{H}_K}\right]\right)^{\frac{1}{p}}, \left(\sum_{j=1}^m \left[\mathbb{E}_{\rho^{\otimes n}}\|\xi_\lambda(\mathbf{x}_j, \mathbf{y}_j)\|^2_{\mathcal{H}_K}\right]\right)^{\frac{1}{2}}\right\} . \end{aligned} \tag{41}$$

Again, the estimates in expectation will follow from integrating a bound holding with high probability. By (31), one has for any $j = 1, \dots, m$,

$$\begin{aligned}\|\xi_\lambda(\mathbf{x}_j, \mathbf{y}_j)\|_{\mathcal{H}_K} &= \|\bar{T}^s g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)\|_{\mathcal{H}_K} \\ &\leq 8\log^{2s}(4\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\lambda)^s \|(\bar{T}_{\mathbf{x}_j} + \lambda)^s g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} f_\rho - \bar{S}^*_{\mathbf{x}_j}\mathbf{y}_j)\|_{\mathcal{H}_K} , \end{aligned} \tag{42}$$

holding with probability at least $1 - \frac{\eta}{2}$, where $\mathcal{B}_{\frac{n}{m}}(\lambda)$ is defined in (29). We proceed by splitting:

$$(\bar{T}_{\mathbf{x}_j} + \lambda)^s g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} f_\rho - \bar{S}^\star_{\mathbf{x}_j}\mathbf{y}_j) = H^{(1)}_{\mathbf{x}_j} \cdot H^{(2)}_{\mathbf{x}_j} \cdot h^\lambda_{\mathbf{z}_j} ,$$

with

$$\begin{aligned} H^{(1)}_{\mathbf{x}_j} &:= ||(\bar{T}_{\mathbf{x}_j} + \lambda)^s g_\lambda(\bar{T}_{\mathbf{x}_j})(\bar{T}_{\mathbf{x}_j} + \lambda)^{\frac{1}{2}}||, \\ H^{(2)}_{\mathbf{x}_j} &:= ||(\bar{T}_{\mathbf{x}_j} + \lambda)^{-\frac{1}{2}}(\bar{T} + \lambda)^{\frac{1}{2}}||, \\ h^\lambda_{\mathbf{z}_j} &:= ||(\bar{T} + \lambda)^{-\frac{1}{2}}(\bar{T}_{\mathbf{x}_j} f_\rho - \bar{S}^\star_{\mathbf{x}_j}\mathbf{y}_j)||_{\mathcal{H}_K} . \end{aligned}$$

The first term is estimated using (9),(10) and gives for $s \in [0, \frac{1}{2}]$

$$\begin{aligned} H^{(1)}_{\mathbf{x}_j} &\leq \sup_{t\in[0,1]} \left(g_\lambda(t)(t+\lambda)^{s+\frac{1}{2}} \right) \\ &\leq 2\left(\sup_{t\in[0,1]} g_\lambda(t) t^{s+\frac{1}{2}} + \lambda^{s+\frac{1}{2}} \sup_{t\in[0,1]} g_\lambda(t) \right) \\ &\leq 2\left(\Big(\sup_{t\in[0,1]} g_\lambda(t) \Big)^{\frac{1}{2}-s} \Big(\sup_{t\in[0,1]} g_\lambda(t) t \Big)^{s+\frac{1}{2}} + \lambda^{s+\frac{1}{2}} \sup_{t\in[0,1]} g_\lambda(t) \right) \\ &\leq C_{\blacktriangle} \lambda^{s-\frac{1}{2}} . \end{aligned} \tag{43}$$

The second term is now bounded using (31) once more. One has with probability at least $1 - \frac{\eta}{4}$

$$H^{(2)}_{\mathbf{x}_j} \leq 8\log(8\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}} . \tag{44}$$

Finally, $h^\lambda_{\mathbf{z}_j}$ is estimated using Proposition 21:

$$h^\lambda_{\mathbf{z}_j} \leq 2\log(8\eta^{-1}) \left(\frac{mM}{n\sqrt{\lambda}} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n}} \right) , \tag{45}$$

holding with probability at least $1 - \frac{\eta}{4}$. Thus, combining (43), (44) and (45) with (42) gives for any $j = 1, \dots, m$,

$$\|\xi_\lambda(\mathbf{x}_j, \mathbf{y}_j)\|_{\mathcal{H}_K} \leq C_{\blacktriangle} \log^{2(s+1)}(8\eta^{-1})\mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}+s}\lambda^s \left(\frac{mM}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}} \right) ,$$

with probability at least $1 - \eta$. Integration gives for any $p \geq 2$:

$$\sum_{j=1}^m \left[\mathbb{E}_{\rho^{\otimes n}} \left\| \xi_\lambda(\mathbf{x}_j, \mathbf{y}_j) \right\|^p_{\mathcal{H}_K} \right] \leq C_{\blacktriangle,p} m\mathcal{A}^p ,$$

with

$$\mathcal{A} := \mathcal{A}_{\frac{n}{m}}(\lambda) := \mathcal{B}_{\frac{n}{m}}(\lambda)^{\frac{1}{2}+s}\lambda^s \left(\frac{mM}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda)}{n\lambda}}\right) .$$

Combining this with (41) implies, since $p \geq 2$:

$$\begin{aligned}\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(\tilde{f}_D^\lambda - \bar{f}_D^\lambda)\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}} &\leq \frac{C_{\blacktriangle,p}}{m}\max\left((m\mathcal{A}^p)^{\frac{1}{p}}, \left(m\mathcal{A}^2\right)^{\frac{1}{2}}\right) \\ &= \frac{C_{\blacktriangle,p}}{m}\,\mathcal{A}\max\left(m^{\frac{1}{p}}, m^{\frac{1}{2}}\right) \\ &= \frac{C_{\blacktriangle,p}}{\sqrt{m}}\,\mathcal{A}\,,\end{aligned}$$

where $C_{\blacktriangle,p}$ does not depend on $(\sigma, M, R) \in \mathbb{R}_+^3$. The result for the case $1 \leq p \leq 2$ immediately follows from Hölder's inequality. ■

Proof [of Theorem 4] Let λ_n be as defined by (16); as earlier we assume n is big enough so that $\lambda_n < 1$. According to Corollary 14, we have $\mathcal{B}_{\frac{n}{m}}(\lambda_n) \leq 2$ provided $\alpha < \frac{2br}{2br+b+1}$ and n is sufficiently large. Under this condition we immediately obtain from Proposition 19:

$$\begin{aligned}\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}} &\leq \frac{C_{\blacktriangle,p}}{\sqrt{m}}\,\lambda_n^s \left(\frac{mM}{n\lambda_n} + \sigma\sqrt{\frac{m\mathcal{N}(\lambda_n)}{n\lambda_n}}\right) \\ &\leq C_{\blacktriangle,p}\lambda_n^s \left(\frac{\sqrt{m}M}{n\lambda_n} + \sigma\sqrt{\frac{\lambda_n^{-\frac{1}{b}}}{n\lambda_n}}\right) ,\end{aligned}$$

where we used again that $\mathcal{N}(\lambda_n) \leq C_{\blacktriangle}\lambda_n^{-1/b}$; now

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

provided

$$m_n \leq n^\alpha , \qquad \alpha < \frac{2(br+1)}{2br+b+1} .$$

Recalling that $\sigma\sqrt{\frac{\lambda_n^{-1/b}}{n\lambda_n}} = R\lambda_n^r = \lambda_n^{-s}a_n$, we arrive at

$$\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}} \leq C_{\blacktriangle,p}\, a_n .$$

As a result, for any $p \geq 1$:

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}_{\sigma,M,R}} \frac{\left[\mathbb{E}_{\rho^{\otimes n}}\left\|\bar{T}^s(\tilde{f}_D^{\lambda_n} - \bar{f}_D^{\lambda_n})\right\|_{\mathcal{H}_K}^p\right]^{\frac{1}{p}}}{a_n} \leq C_{\blacktriangle,p} ,$$

for some $C_{\blacktriangle,p}$, not depending on the model parameters $(\sigma, M, R) \in \mathbb{R}_+^3$, thus leading to the conclusion. ■

Appendix A

Proposition 20 (see e.g. Blanchard and Mücke, 2017, Proposition 5.3) *For any $n \in \mathbb{N}$, $\lambda \in (0,1]$ and $\eta \in (0,1)$, one has with probability at least $1-\eta$:*

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

where $\|.\|_{HS}$ denotes the Hilbert-Schmidt norm. (Since the operator norm is bounded by the Hilbert-Schmidt norm, the above statement also holds for the operator norm.)

Proposition 21 (see e.g. Blanchard and Mücke, 2017, Proposition 5.2) *For $n \in \mathbb{N}$, $\lambda \in (0,1]$ and $\eta \in (0,1]$, it holds with probability at least $1-\eta$:*

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

Proposition 22 (Cordes Inequality, see e.g. Bhatia, 1997, Theorem IX.2.1-2) *Let A, B be to self-adjoint, positive operators on a Hilbert space. Then for any $s \in [0,1]$:*

$$\|A^s B^s\| \leq \|AB\|^s . \tag{46}$$

References

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

R. Bhatia. *Matrix Analysis*. Springer, 1997.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 2017.

X. Chang, S.-B. Lin, and Y. Wang. Divide and conquer local average regression. *Electron. J. Statist.*, 11(1):1326–1350, 2017. doi: 10.1214/17-EJS1265. URL `https://doi.org/10.1214/17-EJS1265`.

G. Cheng and Z. Shang. Computational limits of divide-and-conquer method. Technical report, arXiv:1512.09226, 2016.

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

L. H. Dicker, D. P. Foster, and D. Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electron. J. Statist.*, 11(1):1022–1047, 2017. doi: 10.1214/17-EJS1258.

S. Dirksen. *Noncommutative and vector-valued Rosenthal inequalities.* PhD thesis, Delft Univ. Technology, 2011.

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

J. C. Ferreira and V. A. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, May 2009. ISSN 1420-8989.

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

Q. Guo, B.-W. Chen, S. Rho, W. Ji, F. Jiang, X. Ji, and S.-Y. Kung. Efficient divide-and-conquer classification based on parallel feature-space decomposition for distributed systems. *IEEE Systems Journal*, 2015.

Z.-C. Guo, S.-B. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

C. J. Hsieh, S. Si, and I. Dhillon. A divide-and-conquer solver for kernel support vector machine. *Proceedings of the 31. International Conference on Machine Learning*, 32(1): 575–583, 2014.

R. Li, D. K. J. Lin, and B. Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29 (5):399–409, 2013.

S. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18:1–31, 2017.

S.-B. Lin and D.-X. Zhou. Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, 47(2):249–276, 2018.

L. Mackey, A. Talwalkar, and M. I. Jordan. Divide-and-conquer matrix factorization. *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2011.

I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

H. P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

C. Xu, Y. Zhang, R. Li, and X. Wu. On the feasibility of distributed kernel regression for big data. *IEEE Transactions on Knowledge and Data Engineering*, 28:3041–3052, 2016.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. *JMLR: Workshop and Conference Proceedings*, 30:1–26, 2013.

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18 (3):739–767, 2002.