

Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms

Simon Fischer and Ingo Steinwart

May 23, 2019

Institute for Stochastics and Applications

Faculty 8: Mathematics and Physics

University of Stuttgart

D-70569 Stuttgart Germany

{simon.fischer, ingo.steinwart}@mathematik.uni-stuttgart.de

Abstract

Learning rates for least-squares regression are typically expressed in terms of L_2 -norms. In this paper we extend these rates to norms stronger than the L_2 -norm without requiring the regression function to be contained in the hypothesis space. In the special case of Sobolev reproducing kernel Hilbert spaces used as hypotheses spaces, these stronger norms coincide with fractional Sobolev norms between the used Sobolev space and L_2 . As a consequence, not only the target function but also some of its derivatives can be estimated without changing the algorithm. From a technical point of view, we combine the well-known integral operator techniques with an embedding property, which so far has only been used in combination with empirical process arguments. This combination results in new finite sample bounds with respect to the stronger norms. From these finite sample bounds our rates easily follow. Finally, we prove the asymptotic optimality of our results in many cases.

Keywords Statistical Learning Theory, Regularized Kernel Methods, Least-Squares Regression, Interpolation Norms, Uniform Convergence, Learning Rates

1. Introduction

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ independently sampled from an unknown distribution P on $X \times \mathbb{R}$, the goal of non-parametric least-squares regression is to estimate the conditional mean function $f_P^* : X \rightarrow \mathbb{R}$ given by $f_P^*(x) := \mathbb{E}(Y|X = x)$. The function f_P^* is also known as regression function, we refer to [15] for basic information as well as various algorithms for this problem. In this work, we focus on regularized least-squares algorithms, which are also known as least-squares support vector machines (LS-SVM), see e.g. [31]. Recall that LS-SVMs construct a predictor $f_{D,\lambda}$ by solving the convex optimization problem

$$f_{D,\lambda} = \operatorname{argmin}_{f \in H} \left\{ \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \right\}, \quad (1)$$

where a reproducing kernel Hilbert space (RKHS) H over X is used as hypothesis spaces and $\lambda > 0$ is the so called regularization parameter. For a definition and basic properties of RKHSs see e.g. [31, Chapter 4]. Probably the most interesting theoretical challenge for this problem is to establish learning rates, either in expectation or in probability, for the generalization error

$$\|f_{D,\lambda} - f_P^*\| . \quad (2)$$

Here, we investigate (2) with respect to the norms of a continuous scale of suitable Hilbert spaces $H \subseteq [H]^\gamma \subseteq L_2$. In the following we assume that $[H]^0 = L_2$ and $[H]^1 = H$, see Section 2 for an exact definition. Moreover, in this paper we are mainly interested in the *hard learning* scenario $f_P^* \notin H$.

Let us briefly compare the two main techniques previously used in the literature to establish learning rates for (2): the *integral operator* technique, see e.g. [6, 7, 8, 2, 30, 4, 3, 9, 18, 16] and references therein, and the *empirical process* technique, see e.g. [19, 31, 33] and references therein. An advantage of the integral operator technique is, that it can provide learning rates for (2) with respect to a continuous scale of γ , including the L_2 -norm case $\gamma = 0$, see e.g. [3, 18]. In addition, it can be used to establish learning rates for *spectral regularization algorithms*, see e.g. [2, 3, 18] and further RKHS-based learning algorithms, see e.g. [21, 17, 25, 22, 23]. On the other hand, the empirical process techniques can so far only handle the L_2 -norm in (2), but in the hard learning scenario $f_P^* \notin H$, which is rarely investigated by the integral operator technique, it provides the fastest, and in many cases minimax optimal, L_2 -learning rates for (2), see [33]. In addition, it can be easily applied to learning algorithms (1) in which the least-squares loss function is replaced by other convex loss functions, see e.g. [13] for expectile regression and e.g. [11] for quantile regression.

In the present manuscript, which is an improvement of its first version [14], we extend and improve the results of [3, 18]. To be more precise, we extend the results of [3], which assume $f_P^* \in H$, to the hard learning case and the largest possible scale of γ , and compared to [18] we obtain faster rates of convergence for (2), if the RKHS enjoys a certain embedding property, which previously has only been used in [33, 9, 25]. In the hard learning scenario, we obtain, as a byproduct, the L_2 -learning rates of [33], as well as the very first L_∞ -norm learning rates. For a more detailed comparison with the literature see Section 5 and in particular Table 1 and Figure 1. Finally, we also prove the minimax optimality of our $[H]^\gamma$ -norm learning rates for all combinations of H and P for which the optimal L_2 -norm learning rates are known.

The rest of this work is organized as follows: We start in Section 2 with an introduction of notations and general assumptions. In Section 3 we present our learning rates and discuss their main assumptions. The consequences of our results for the special case of a Sobolev/Besov RKHS H and a marginal distribution P_X close to the uniform distribution can be found in Section 4. Note that in this case $[H]^\gamma$ coincide with the classical Besov spaces and the corresponding norms have a nice interpretation in terms of derivatives. Finally, we compare our result with other contributions in Section 5. All proofs can be found in Section 6.

Acknowledgment

We are especially grateful to Nicole Mücke for pointing us to the article of Lin, Rudi, Rosasco, and Cevher [18].

2. Preliminaries

Let (X, \mathcal{B}) be a measurable space used as *input space*, $Y = \mathbb{R}$ be the *output space* and P be an *unknown* probability distribution on $X \times Y$ with $|P|_2^2 := \int_{X \times Y} y^2 dP(x, y) < \infty$. Moreover, we denote the marginal distribution of P on X by $\nu := P_X$ and assume that (X, \mathcal{B}) is ν -complete. In the following, we fix a (regular) conditional probability $P(\cdot | x)$ of P given $x \in X$. Since the conditional mean function f_P^* is only ν -almost everywhere uniquely determined we use the symbol f_P^* for both, the ν -equivalence class and for the representative

$$f_P^*(x) = \int_Y y P(dy|x) .$$

If we use another representative we will explicitly point this out.

In the following, we fix a separable RKHS H on X with respect to a measurable (w.r.t. $\mathcal{B} \otimes \mathcal{B}$) and bounded kernel k . Let us recall some facts about the interplay between H and $L_2(\nu)$. Some of the following facts can already be found in Smale and Zhou [28, 29] and De Vito et al. [8, 7], but we follow the more recent contribution [32]. According to [32, Lemma 2.2, Lemma 2.3], and [31, Theorem 4.27] the—not necessarily injective—embedding $I_\nu : H \rightarrow L_2(\nu)$, mapping a function $f \in H$ to its ν -equivalence class $[f]_\nu$, is well-defined, Hilbert-Schmidt and the Hilbert-Schmidt norm satisfies

$$\|I_\nu\|_{\mathcal{L}_2(H, L_2(\nu))} = \|k\|_{L_2(\nu)} := \left(\int_X k(x, x) d\nu(x) \right)^{1/2} < \infty .$$

Moreover, the adjoint operator $S_\nu := I_\nu^* : L_2(\nu) \rightarrow H$ is an integral operator with respect to the kernel k , i.e.

$$(S_\nu f)(x) = \int_X k(x, x') f(x') d\nu(x')$$

for $x \in X$ and $f \in L_2(\nu)$. Next, we define the self-adjoint and positive semi-definite integral operators

$$T_\nu := I_\nu S_\nu : L_2(\nu) \rightarrow L_2(\nu) \quad \text{and} \quad C_\nu := S_\nu I_\nu : H \rightarrow H .$$

These operators are trace class and the trace norm satisfies

$$\|T_\nu\|_{\mathcal{L}_1(L_2(\nu))} = \|C_\nu\|_{\mathcal{L}_1(H)} = \|I_\nu\|_{\mathcal{L}_2(H, L_2(\nu))}^2 = \|S_\nu\|_{\mathcal{L}_2(L_2(\nu), H)}^2 . \quad (3)$$

If there is no danger of confusion we write $\|\cdot\|$ for the operator norm, $\|\cdot\|_2$ for the Hilbert-Schmidt norm, and $\|\cdot\|_1$ for the trace norm. The spectral theorem for self-adjoint compact operators yields an at most countable index set \mathcal{I} , a non-increasing summable sequence $(\mu_i)_{i \in \mathcal{I}} \subseteq (0, \infty)$, and a family $(e_i)_{i \in \mathcal{I}} \subseteq H$, such that $([e_i]_\nu)_{i \in \mathcal{I}}$ is an ONB of $\overline{\text{ran } T_\nu} \subseteq L_2(\nu)$ and $(\mu_i^{1/2} e_i)_{i \in \mathcal{I}}$ is an ONB of $(\ker I_\nu)^\perp \subseteq H$ with

$$T_\nu = \sum_{i \in \mathcal{I}} \mu_i \langle \cdot, [e_i]_\nu \rangle_{L_2(\nu)} [e_i]_\nu \quad \text{resp.} \quad C_\nu = \sum_{i \in \mathcal{I}} \mu_i \langle \cdot, \mu_i^{1/2} e_i \rangle_H \mu_i^{1/2} e_i , \quad (4)$$

see [32, Lemma 2.12] for details. Since we are mainly interested in the hard learning scenario $f_P^* \notin H$ we exclude finite index sets \mathcal{I} and assume $\mathcal{I} = \mathbb{N}$ in the following.

Let us recall some intermediate spaces introduced on [32, p. 384]. We call them *power spaces*. For

$\alpha \geq 0$ the α -power space is defined by

$$[H]_\nu^\alpha := \left\{ \sum_{i \geq 1} a_i \mu_i^{\alpha/2} [e_i]_\nu : (a_i)_{i \geq 1} \in \ell_2(\mathbb{N}) \right\} \subseteq L_2(\nu)$$

and equipped with the α -power norm

$$\left\| \sum_{i \geq 1} a_i \mu_i^{\alpha/2} [e_i]_\nu \right\|_{[H]_\nu^\alpha} := \|(a_i)_{i \geq 1}\|_{\ell_2(\mathbb{N})} = \left(\sum_{i \geq 1} a_i^2 \right)^{1/2}$$

for $(a_i)_{i \geq 1} \in \ell_2(\mathbb{N})$. If there is no danger of confusion we use the abbreviation $\|\cdot\|_\alpha := \|\cdot\|_{[H]_\nu^\alpha}$. Moreover, in the case of $\alpha = 1$ we introduce the notation $[H]_\nu := [H]_\nu^1$. The space $[H]_\nu^\alpha$ is a separable Hilbert space with ONB $(\mu_i^{\alpha/2} [e_i]_\nu)_{i \geq 1}$. Recall, that for $\alpha = 0$ we have $[H]_\nu^0 = \overline{\text{ran } I_\nu} \subseteq L_2(\nu)$ with $\|\cdot\|_0 = \|\cdot\|_{L_2(\nu)}$. Moreover, for $\alpha = 1$ we have $[H]_\nu^1 = \text{ran } I_\nu$ and $[H]_\nu^1$ is isometric isomorph to the closed subspace $(\ker I_\nu)^\perp$ of H via I_ν , i.e. $\|[f]_\nu\|_1 = \|f\|_H$ for $f \in (\ker I_\nu)^\perp$. For $0 < \beta < \alpha$ the embeddings

$$[H]_\nu^\alpha \hookrightarrow [H]_\nu^\beta \hookrightarrow [H]_\nu^0 = \overline{\text{ran } I_\nu} \subseteq L_2(\nu) \quad (5)$$

exist and are compact. For $\alpha > 0$ the α -power space is given by the image of the fractional integral operator, namely

$$[H]_\nu^\alpha = \text{ran } T_\nu^{\alpha/2} \quad \text{and} \quad \|T_\nu^{\alpha/2} f\|_\alpha = \|f\|_{L_2(\nu)}$$

for $f \in \overline{\text{ran } I_\nu}$. In addition, for $0 < \alpha < 1$ the α -power space is characterized in terms of interpolation spaces of the real method, see e.g. [34, Section 1.3.2] for a definition. To be more precise, [32, Theorem 4.6] yields

$$[H]_\nu^\alpha \cong [L_2(\nu), [H]_\nu]_{\alpha, 2}, \quad (6)$$

where the symbol \cong in (6) means that these spaces are isomorph, i.e. the sets coincide and the corresponding norms are equivalent. Note that for Sobolev/Besov RKHS and marginal distributions close to the uniform distribution, the interpolation space $[L_2(\nu), [H]_\nu]_{\alpha, 2}$ is well-known from the literature, see Section 4 for details.

3. Main Results

Before we state the results we introduce the main assumptions. For $0 < p \leq 1$ we assume that the *eigenvalue decay* satisfy a polynomial upper bound of order $1/p$: There is a constant $C > 0$ such that the eigenvalues $(\mu_i)_{i \geq 1}$ of the integral operator satisfy

$$\mu_i \leq C i^{-1/p} \quad (\text{EVD})$$

for all $i \geq 1$. In order to establish the optimality of our results we obviously need to assume an exact polynomial asymptotic behavior of order $1/p$: There are constants $c, C > 0$ such that

$$c i^{-1/p} \leq \mu_i \leq C i^{-1/p} \quad (\text{EVD+})$$

is satisfied for all $i \geq 1$. Our next assumption is the *embedding property*, for $0 < \alpha \leq 1$: There is a constant $A > 0$ with

$$\| [H]_\nu^\alpha \hookrightarrow L_\infty(\nu) \| \leq A . \quad (\text{EMB})$$

This mean $[H]_\nu^\alpha$ is continuously embedded into $L_\infty(\nu)$ and the operator norm of the embedding is bounded by A . Because of (5) the larger α is the weaker the embedding property is. Since our kernel k is bounded (EMB) is always satisfied for $\alpha = 1$. Moreover, Lemma 6.2 (iii) in the proof section shows that (EMB) implies a polynomial eigenvalue decay of order $1/\alpha$. But the inverse does not hold in general and hence we assume $p \leq \alpha$ in the following.

Note that the Conditions (EMB) and (EVD)/(EVD+) just describe the interplay between the marginal distribution $\nu = P_X$ and the RKHS H . Consequently, they are independent of the conditional distribution $P(\cdot|x)$ and especially independent of the regression function f_P^* . In the following, we use a *source condition*, for $0 < \beta \leq 2$, to measure the smoothness of the regression function: $f_P^* \in [H]_\nu^\beta$ and there is a constant $B > 0$ with

$$\| f_P^* \|_\beta \leq B . \quad (\text{SRC})$$

Note that $|P|_2 < \infty$ already implies $f_P^* \in L_2(\nu)$. Moreover, (SRC) with $\beta \geq 1$ implies that f_P^* has a representative from H —in short $f_P^* \in H$ —and hence $\beta \geq 1$ excludes the hard learning scenario we are mainly interested in. We included the case $1 \leq \beta \leq 2$ because it is no extra effort in the proof. Since the generalization error $\| [f_{D,\lambda}]_\nu - f_P^* \|_\gamma$, with respect to the γ -power norm, for some $0 \leq \gamma \leq 1$, is well-defined if and only if $f_P^* \in [H]_\nu^\gamma$ we naturally have to assume $\beta \geq \gamma$ in the following. Finally, we introduce a *moment condition* to control the noise of the observations: There are constants $\sigma, L > 0$ such that

$$\int_Y |y - f_P^*(x)|^m P(dy|x) \leq \frac{1}{2} m! \sigma^2 L^{m-2} \quad (\text{MOM})$$

is satisfied for ν -a.a. $x \in X$ and all $m \geq 2$. Note that (MOM) is satisfied for Gaussian noise with bounded variance, i.e. $P(\cdot|x) = \mathcal{N}(f_P^*(x), \sigma_x^2)$, where $x \mapsto \sigma_x \in (0, \infty)$ is a measurable and ν -a.s. bounded function. Another sufficient condition is that P is concentrated on $X \times [-M, M]$ for some constant $M > 0$, i.e. $P(X \times [-M, M]) = 1$. The Conditions (EVD) and (SRC) are well-recognised in the statistical analysis of regularized least-squares algorithms, see e.g. [4, 3, 16, 18]. However, there is a whole zoo of moment conditions. We use (MOM) because (MOM) only constraints the discrepancy of the observation y to the *true* value $f_P^*(x)$ and hence do *not* imply additional constraints, such as boundedness, on f_P^* . An embedding property similar to (EMB) was used in [33] in combination with empirical process arguments, in [9] to investigate benign scenarios with exponentially decreasing eigenvalues and $f_P^* \in H$, and in [25] to investigate stochastic gradient methods. But embedding properties are new in combination with the integral operator technique in the hard learning scenario for the learning scheme (1) and enables us to prove the following result.

3.1 Theorem (γ -Learning Rates) *Let H be a separable RKHS on X with respect to a bounded and measurable kernel k and P be a probability distribution on $X \times Y$ such that $|P|_2 < \infty$ and (X, \mathcal{B}) is complete with respect to the marginal distribution $\nu := P_X$. Furthermore, we assume that there is a constant $B_\infty > 0$ with $\| f_P^* \|_{L_\infty(\nu)} \leq B_\infty$ and that the Conditions (EMB), (EVD) (SRC), and (MOM) are satisfied for some $0 < p \leq \alpha \leq 1$ and $0 < \beta \leq 2$. Then for $0 \leq \gamma \leq 1$ with $\gamma < \beta$ and a*

regularization parameter sequence $(\lambda_n)_{n \geq 1}$ the LS-SVM $D \mapsto f_{D, \lambda_n}$ with respect to H defined by (1) satisfies the following statements:

(i) In the case of $\beta + p \leq \alpha$ and $\lambda_n \asymp \left(\frac{\log^r(n)}{n}\right)^{1/\alpha}$ for some $r > 1$ there is a constant $K > 0$ independent of $n \geq 1$ and $\tau \geq 1$ such that

$$\| [f_{D, \lambda_n}]_\nu - f_P^* \|_\gamma^2 \leq \tau^2 K \left(\frac{\log^r(n)}{n} \right)^{\frac{\beta-\gamma}{\alpha}} \quad (7)$$

is satisfied for sufficient large $n \geq 1$ with P^n -probability $\geq 1 - 4e^{-\tau}$.

(ii) In the case of $\beta + p > \alpha$ and $\lambda_n \asymp \left(\frac{1}{n}\right)^{\frac{1}{\beta+p}}$ there is a constant $K > 0$ independent of $n \geq 1$ and $\tau \geq 1$ such that

$$\| [f_{D, \lambda_n}]_\nu - f_P^* \|_\gamma^2 \leq \tau^2 K \left(\frac{1}{n} \right)^{\frac{\beta-\gamma}{\beta+p}} \quad (8)$$

is satisfied for sufficient large $n \geq 1$ with P^n -probability $\geq 1 - 4e^{-\tau}$.

Theorem 3.1 is mainly based on a finite sample bound given in the proof section, see Theorem 6.7. The asymptotic behavior in n of the right hand side in (7) respectively (8) is called *learning rate* with respect to the γ -power norm or abbreviated γ -learning rate. Note that, for $\beta \geq \alpha$, the conditional mean function f_P^* is automatically ν -a.s. bounded, since we have $f_P^* \in [H]_\nu^\beta \hookrightarrow [H]_\nu^\alpha \hookrightarrow L_\infty(\nu)$, and in this case always situation (8) applies. Moreover, in the case of $\alpha = p$, which was also considered in [33, Corollary 6], we are always in situation (8), too. Recall, for $\gamma = 0$ the left hand side coincides with the $L_2(\nu)$ -norm. If we ignore the log-term in the obtained γ -learning rates then in both cases, $\beta + p \leq \alpha$ and $\beta + p > \alpha$, the γ -learning rate coincides with

$$\left(\frac{1}{n} \right)^{\frac{\beta-\gamma}{\max\{\beta+p, \alpha\}}}.$$

Finally, note that the asymptotic behavior of the regularization parameter sequence does not depend on the considered γ -power norm. Consequently, we get convergence with respect to *all* γ -power norms $0 \leq \gamma < \beta$ *simultaneously*. In order to investigate the optimality of our γ -learning rates the next theorem yields γ -lower rates. In doing so, we obviously have to assume (EVD+) to make sure that the eigenvalues do not decay faster than (EVD) guarantees.

3.2 Theorem (γ -Lower Rates) *Let H be a separable RKHS on X with respect to a bounded and measurable kernel k , and ν be a probability distribution on X such that (EMB) and (EVD+) are satisfied for some $0 < p \leq \alpha \leq 1$. Then, for all parameters $0 < \beta \leq 2$, $0 \leq \gamma \leq 1$ with $\gamma < \beta$ and all constants $\sigma, L, B, B_\infty > 0$ there exist $K, C, r > 0$ such that for all learning methods $D \mapsto f_D$, all $\tau > 0$, and all $n \geq 1$ sufficiently large there is a distribution P on $X \times Y$ with $P_X = \nu$ satisfying $\|f_P^*\|_{L_\infty(\nu)}^2 \leq B_\infty$, (SRC) with respect to B , (MOM) with respect to σ, L , and*

$$\| [f_D]_\nu - f_P^* \|_\gamma^2 \geq \tau^2 K \left(\frac{1}{n} \right)^{\frac{\max\{\alpha, \beta\} - \gamma}{\max\{\alpha, \beta\} + p}} \quad (9)$$

with P^n -probability $\geq 1 - C\tau^r$.

In short, Theorem 3.2 says that there is no learning method satisfying a faster decaying γ -learning

rate than

$$\left(\frac{1}{n}\right)^{\frac{\max\{\alpha,\beta\}-\gamma}{\max\{\alpha,\beta\}+p}}$$

under the assumptions of Theorem 3.1 and (EVD+). The asymptotic behavior in n of the right hand side in (9) is called (*minimax*) *lower rate* with respect to the γ -power norm or abbreviated γ -lower rate. Note that special cases of Theorem 3.2 can be found in [4, 33, 3]. In the case of $\alpha \leq \beta$, which implies the boundedness of f_P^* , the γ -learning rate of LS-SVMs stated in Theorem 3.1 coincides with the γ -lower rate from Theorem 3.2 and hence is optimal. The optimal rate in the case of $\alpha > \beta$, which does *not* imply the boundedness of f_P^* , is, *even for the L_2 -norm*, an outstanding problem for several decades which we cannot address, either.

3.3 Remark (Optimality and Boundedness) *Under the assumptions of Theorem 3.2, but without requiring the uniform boundedness of f_P^* by some constant B_∞ , the improved γ -lower rate*

$$\left(\frac{1}{n}\right)^{\frac{\beta-\gamma}{\beta+p}}$$

is satisfied. This requires only a straight forward modification of Lemma 6.13 in Section 6. If we would be able to prove the γ -learning rates of Theorem 3.1 with a constant $K > 0$ independent of B_∞ then we would have optimality for all $\beta > \alpha - p$ instead of $\beta \geq \alpha$.

We think that this would be a valuable next step on the way answering the question of optimality. However—to the best of our knowledge—[24] is the only article providing learning rates for unbounded target functions and is based on an empirical process arguments.

Because of (EMB), the next remark is a direct consequence of Theorem 3.1 for $\gamma = \alpha$.

3.4 Remark (L_∞ -Learning Rates) *Under the assumptions of Theorem 3.1 in the case of $\beta > \alpha$ the following statement is true. For all regularization parameter sequences $(\lambda_n)_{n \geq 1}$ with $\lambda_n \asymp \left(\frac{1}{n}\right)^{\frac{1}{\beta+p}}$ there is a constant $K > 0$ independent of $n \geq 1$ and $\tau \geq 1$ such that the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to H defined by (1) satisfies*

$$\| [f_{D,\lambda_n}]_\nu - f_P^* \|_{L_\infty(\nu)}^2 \leq \tau^2 K \left(\frac{1}{n}\right)^{\frac{\beta-\alpha}{\beta+p}}$$

for all $n \geq 1$ sufficiently large with P^n -probability $\geq 1 - 4e^{-\tau}$.

Note that all previous efforts to get L_∞ -learning rates for the learning scheme (1) need to assume $f_P^* \in H$. Consequently, we get the very first L_∞ -learning rates in the hard learning scenario.

4. Example: Besov RKHSs

In this section we illustrate our main results in the case of Besov RKHSs. To this end, we make the following general assumptions: Let $X \subseteq \mathbb{R}^d$ be a non-empty, open, connected and bounded set with a C_∞ -boundary and equipped with the Lebesgue σ -algebra \mathcal{B} such that (X, \mathcal{B}) is complete with respect to the Lebesgue measure μ . Furthermore, $L_2(X) := L_2(\mu)$ denotes the corresponding L_2 -space.

First, we briefly introduce the Sobolev and Besov spaces. For a more detailed introduction see e.g. Adams and Fournier [1]. Since we are only interested in Hilbert spaces, we restrict ourself to this

case. For $m \in \mathbb{N}$ we denote the *Sobolev space* of smoothness m by $W_m(X) := W_{m,2}(X)$, see e.g. [1, Definition 3.2] for a definition. For $r > 0$ the *Besov space* $B_{2,2}^r(X)$ is defined by means of the real interpolation method, namely $B_{2,2}^r(X) := [L_2(X), W_m(X)]_{r/m,2}$, where $m := \min\{k \in \mathbb{N} : k > r\}$, see [1, Section 7.30]. For $r = 0$ we define $B_{2,2}^0(X) := L_2(X)$. It is well-known that the Besov spaces $B_{2,2}^r(X)$ are separable Hilbert spaces and that they satisfy

$$B_{2,2}^r(X) \cong [L_2(X), B_{2,2}^t(X)]_{r/t,2} \quad (10)$$

for all $t > r > 0$, see e.g. [1, Section 7.32] for details. Moreover, for $r > d/2$ each μ -equivalence class in $B_{2,2}^r(X)$ has a unique continuous and bounded representative, see [1, Theorem 7.24 (c)]. In fact, for $r > j + d/2$, this representative is from the space $C_j(X)$ of j -times continuous differentiable and bounded functions with bounded derivatives. More precisely, the mapping of a μ -equivalence class to its (unique) continuous representative is linear and continuous, in short

$$B_{2,2}^r(X) \hookrightarrow C_j(X) . \quad (11)$$

Consequently, we define, for $r > d/2$, the *Besov RKHS* as the set of continuous representatives $H_r(X) := \{f \in C_0(X) : [f]_\mu \in B_{2,2}^r(X)\}$ and equip this space with the norm $\|f\|_{H_r(X)} := \|[f]_\mu\|_{B_{2,2}^r(X)}$. The Besov RKHS $H_r(X)$ is a separable RKHS with respect to a kernel k_r . Moreover, k_r is bounded and measurable according to [31, Lemma 4.28 and 4.25].

In the following, we fix a Besov RKHS $H_r(X)$, with $r > d/2$, and a probability measure P on $X \times Y$ such that the marginal distribution $\nu = P_X$ on X satisfies $\nu \ll \mu$, $\mu \ll \nu$, and $g \leq \frac{d\nu}{d\mu} \leq G$ μ -a.s. for some constants $g, G > 0$. For such a marginal distribution we have $L_2(\nu) \cong L_2(X)$ and we can describe the power spaces of $H_r(X)$ according to (6), the interpolation property, and (10) by

$$[H_r(X)]_\nu^{u/r} \cong [L_2(\nu), [H_r(X)]_\nu]_{u/r,2} \cong [L_2(X), [H_r(X)]_\mu]_{u/r,2} \cong B_{2,2}^u(X) \quad (12)$$

for $0 < u < r$. As a consequence of (12), if $f_P^* \in B_{2,2}^s(X)$ for some $0 < s < r$, then (SRC) is satisfied with $\beta = s/r$. Next, if we combine (12) and (11) then we get (EMB) for all α with $\frac{d}{2r} < \alpha < 1$:

$$[H_r(X)]_\nu^\alpha \cong B_{2,2}^{\alpha r}(X) \hookrightarrow C_0(X) \hookrightarrow L_\infty(\nu) .$$

Finally, we consider the asymptotic behavior of the eigenvalues $(\mu_i)_{i \geq 1}$ of the integral operator T_ν . According to [5, Equation (4.4.12)] the eigenvalue μ_i of T_ν equals the squares of the approximation numbers $a_i^2(I_\nu)$ of the embedding $I_\nu : H_r(X) \rightarrow L_2(\nu)$. Since $L_2(\nu) \cong L_2(X)$ these approximation numbers are described in [12, Equation (4) on p. 119] by

$$\mu_i = a_i^2(I_\nu) \asymp i^{-2r/d} .$$

To sum up, the eigenvalues satisfy (EVD+) for $p = \frac{d}{2r}$. The following corollaries are direct consequences of Theorem 3.1 and Theorem 3.2 with $p = \frac{d}{2r}$, $\beta = s/r$, $\gamma = t/r$, and $\alpha > p$ sufficient close to p .

4.1 Corollary (Besov-Learning Rates) *Let $H_r(X)$ be a Besov RKHS with $r > d/2$ and P be a probability*

distribution on $X \times Y$ such that $|P|_2 < \infty$ and the marginal distribution $\nu := P_X$ satisfies $\nu \ll \mu$, $\mu \ll \nu$, and $g \leq \frac{d\nu}{d\mu} \leq G$ for some constants $g, G > 0$. Furthermore, we assume that there are constants $B, B_\infty > 0$, such that $\|f_P^*\|_{L_\infty(\mu)} \leq B_\infty$ and $\|f_P^*\|_{B_{2,2}^s(X)} \leq B$ for some $0 < s < r$, and that the Condition (MOM) is satisfied. Then for $0 \leq t < s$ and a regularization parameter sequence $(\lambda_n)_{n \geq 1}$ with $\lambda_n \asymp \left(\frac{1}{n}\right)^{\frac{r}{s+d/2}}$ there is a constant $K > 0$ independent of $n \geq 1$ and $\tau \geq 1$ such that the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to the Besov RKHS $H_r(X)$ defined by (1) satisfies

$$\|[f_{D,\lambda_n}]_\mu - f_P^*\|_{B_{2,2}^t(X)}^2 \leq \tau^2 K \left(\frac{1}{n}\right)^{\frac{s-t}{s+d/2}}$$

for sufficient large $n \geq 1$ with P^n -probability $\geq 1 - 4e^{-\tau}$.

Note that the $B_{2,2}^t$ -learning rate is independent of the chosen Besov RKHS $H_r(X)$. Besides $r > d/2$ the only requirement on the choice of $H_r(X)$, a user has to take care of, is $r > s$. Recall that the case $t = 0$ corresponds to L_2 -learning rates.

4.2 Corollary (Besov-Lower Rates) *Let $H_r(X)$ be a Besov RKHS with $r > d/2$ and ν be a probability distribution on X with $\nu \ll \mu$, $\mu \ll \nu$, and $g \leq \frac{d\nu}{d\mu} \leq G$ for some constants $g, G > 0$. Then for all parameters $0 \leq t < s < r$, $\varepsilon > 0$ sufficient small, and all constants $\sigma, L, B, B_\infty > 0$ there exist $K, C, r > 0$ such that for all learning methods $D \mapsto f_D$, all $\tau > 0$ and $n \geq 1$ sufficient large there is a distribution P on $X \times Y$ with $P_X = \nu$ satisfying $\|f_P^*\|_{L_\infty(\nu)} \leq B_\infty$, $\|f_P^*\|_{B_{2,2}^t(X)} \leq B$, (MOM) with respect to σ, L , and*

(i) in the case of $s \leq d/2$

$$\|[f_D]_\mu - f_P^*\|_{B_{2,2}^t(X)}^2 \geq \tau^2 K \left(\frac{1}{n}\right)^{1/2-t/d+\varepsilon}$$

(ii) in the case of $s > d/2$

$$\|[f_D]_\mu - f_P^*\|_{B_{2,2}^t(X)}^2 \geq \tau^2 K \left(\frac{1}{n}\right)^{\frac{s-t}{s+d/2}}$$

with P^n -probability $\geq 1 - C\tau^\tau$.

Note that in the case of $s \leq d/2$ the $\varepsilon > 0$ appears in the lower rate because we have to choose $\alpha > p$. In short, Corollary 4.2 says that the learning rates from Corollary 4.1 are optimal in the case of $s > d/2$. Finally, if we even have $s > j + d/2$, for some integer $j \geq 0$, then the combination of Corollary 4.1 and (11) yields $C_j(X)$ -learning rates. To this end, we denote by f_P^* the unique continuous representative of the μ -equivalence class f_P^* .

4.3 Remark ($C_j(X)$ -Learning Rates) *Under the assumption of Corollary 4.1 in the case of $s > j + d/2$ for some integer $j \geq 0$ the following statement is true. For all $0 < \varepsilon < \frac{s-(j+d/2)}{s+d/2}$ and each regularization parameter sequence $(\lambda_n)_{n \geq 1}$ with $\lambda_n \asymp \left(\frac{1}{n}\right)^{\frac{r}{s+d/2}}$ there is a constant $K > 0$ independent of $n \geq 1$ and $\tau \geq 1$ such that the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to the Besov RKHS $H_r(X)$ defined by (1) satisfies*

$$\|f_{D,\lambda_n} - f_P^*\|_{C_j(X)}^2 \leq \tau^2 K \left(\frac{1}{n}\right)^{\frac{s-(j+d/2)}{s+d/2}-\varepsilon}$$

for sufficient large $n \geq 1$ with P^n -probability $\geq 1 - 4e^{-\tau}$.

Note that learning rates for the Besov RKHS are already investigated e.g. by Steinwart et al. [33] but only for the $L_2(X)$ -norm. Our contributions are learning rates with respect to the $B_{2,2}^t(X)$ -norm and the $C_j(X)$ -norm.

5. Comparison

In this section we compare our results with learning rates previously obtained in the literature. Since in the case of $f_P^* \in [H]_\nu^\beta$ with $1 \leq \beta \leq 2$ we just recover the well-known optimal rates $n^{-\frac{\beta-\gamma}{\beta+p}}$ obtained by many authors, see e.g. [4, 16] for L_2 -rates and [3, 18] for general γ -rates, we focus on the hard learning scenario $0 < \beta < 1$. Furthermore, due to the large amount of results in the literature we limit our considerations to the best known results for the learning scheme (1), namely [31, 33], which use empirical process techniques and [16, 18], which use integral operator techniques. Moreover, we assume that P is concentrated on $X \times [-M, M]$ for some $M > 0$ and that k is a bounded measurable kernel with separable RKHS H . Note that these assumptions form the largest common ground under which all the considered contributions achieve L_2 -learning rates. In addition, [18] is the only result of the four articles listed above that considers general γ -learning rates. Finally, in order to keep the comparison clear we ignore log-terms in the learning rates. In Table 1 we give a short overview of the learning rates and in Figure 1 we plot the exponent r of the polynomial L_2 -learning rates n^{-r} over the smoothness $0 < \beta < 1$ of $f_P^* \in [H]_\nu^\beta$ for some fixed $0 < p \leq \alpha \leq 1$.

Articles	Assumptions		Learning Rates n^{-r} (exponent) in	
	(EMB) $[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)$	(EVD) $\mu_i \preceq i^{-\frac{1}{p}}$	$L_2(\nu)$	$[H]_\nu^\gamma$ for $\gamma < \beta$
our results	$0 < \alpha \leq 1$	$0 < p \leq \alpha$	$\frac{\beta}{\max\{\beta+p, \alpha\}}$	$\frac{\beta-\gamma}{\max\{\beta+p, \alpha\}}$
Steinwart and Christmann [31, Thm. 7.23] + (EMB)				
Steinwart et al. [33, Thm. 1]	$0 < \alpha \leq 1$	$0 < p \leq \alpha$	$\frac{\beta}{\max\{\beta+p, \beta+\alpha(1-\beta)\}}$	x
Steinwart et al. [33, Cor. 6]	$0 < \alpha \leq 1$	$p = \alpha$	$\frac{\beta}{\beta+\alpha}$	
Steinwart and Christmann [31, Eq. (7.54)]	$\alpha = 1$	$0 < p \leq 1$	$\frac{\beta}{\max\{\beta+p, 1\}}$	
Lin and Cevher [16, Cor. 6]				
Lin et al. [18, Cor. 4.4]				$\frac{\beta-\gamma}{\max\{\beta+p, 1\}}$

Table 1: Learning rates established by different authors for $f_P^* \in [H]_\nu^\beta$ with $0 < \beta < 1$. In order to keep the comparison clear we ignore log-terms in the learning rates. The *blue* results are based on integral operator techniques and the *green* ones are based on empirical process techniques. The *marked* parameter ranges are more restrictive than ours and the *marked* rates are never better than our rates and at least for some parameter ranges worse than our rates.

Integral operator techniques: Lin and Cevher [16], which is an extended version of the conference paper [17], investigates distributed gradient decent methods and spectral regularization algorithms. In [16, Corollary 6] they provide the L_2 -learning rate $n^{-\frac{\beta}{\max\{\beta+\alpha, 1\}}}$ (in expectation) for spectral regularization algorithms, containing the learning scheme (1) as special case. Lin et al. [18] establish the γ -learning rate $n^{-\frac{\beta-\gamma}{\max\{\beta+p, 1\}}}$ (in probability) for spectral regularization algorithms under more general source conditions, see [18, Equation (18)]. Both articles, [16] and [18], do not take into account any embedding property and hence in case of (EMB) with $\alpha < 1$ we improve their rates iff $\beta + p < 1$. Let

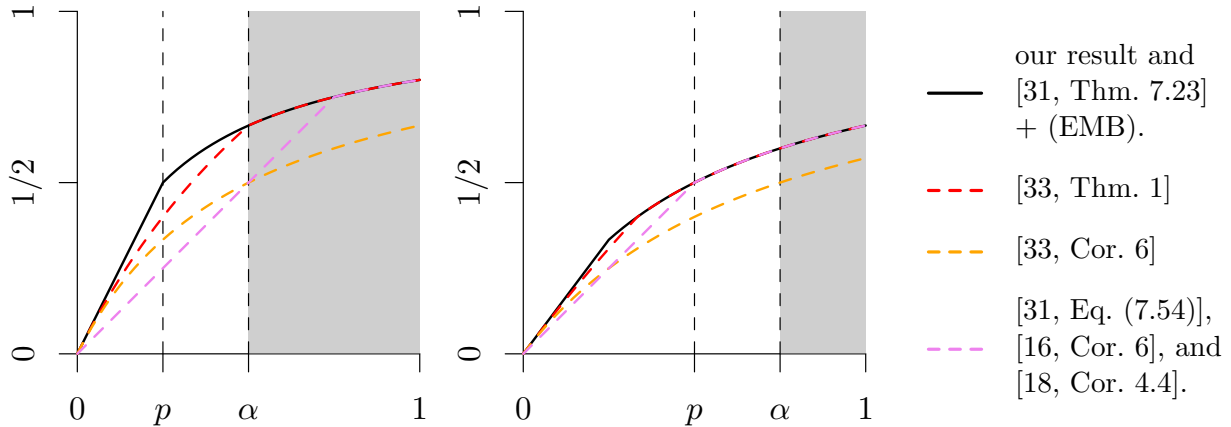


Figure 1: Plot of the exponent r of the L_2 -learning rate n^{-r} over the smoothness β of f_P^* for a fixed RKHS H and a fixed marginal distribution $\nu = P_X$ which satisfy (EMB) and (EVD) with respect to α resp. p . Consequently, higher values correspond to faster learning rates. In the gray shaded range the best rates are known to be optimal. The left plot corresponds to $\alpha = 1/2$ and $p = 1/4$ whereby the right plot corresponds to $\alpha = 3/4$ and $p = 1/2$.

us illustrate this improvement in the case of a Besov RKHS $H_r(X)$ with smoothness r . To this end, we assume $f_P^* \in B_{2,2}^s(X)$ for some $s > 0$. Besides the condition $r > d/2$, which ensures that $H_r(X)$ is a RKHS, the only requirement is $r > s$ in order to achieve the fastest known L_2 -learning rate $n^{-\frac{2s}{2s+d}}$. Recall that this rate is independent of the smoothness r and is known to be optimal for $s > d/2$. In order to get the same L_2 -learning rate by the results of [16, 18] the *additional* constraint $r < s + d/2$ has to be satisfied. Otherwise, [16, 18] only yield the L_2 -rate $n^{-s/r}$, which gets worse with increasing smoothness r . Consequently, taking (EMB) into account facilitates the choice of r . Moreover, for learning rates with respect to Besov norms our results improve those of [18] in a similar way.

Empirical process techniques: [31] provide an oracle inequality in [31, Theorem 7.23] under a slightly weaker assumptions than (EVD) and (SRC). As already mentioned in [31, Equation (7.54)], this oracle inequality together with [31, Example 7.3] leads to the L_2 -rate $n^{-\frac{\beta}{\max\{\beta+p, 1\}}}$. Consequently, the rate in [31, Equation (7.54)] coincides with the results in [16, 18] and is even better by a logarithmic factor. Inspired by Mendelson and Neeman [19, Lemma 5.1] Steinwart et al. used an embedding property, slightly weaker than (EMB), for the first time in [33, Theorem 1]. Moreover, [33, Theorem 1] was used in [33, Corollary 6] to establish, in the case of $p = \alpha$, the L_2 -rate $n^{-\frac{\beta}{\beta+\alpha}}$. But the proof remains valid in the general case $p \leq \alpha$ and hence [33, Theorem 1] yields the rate $n^{-\frac{\beta}{\max\{\beta+p, \beta+\alpha(1-\beta)\}}}$. This rate is worse than ours iff $\alpha < 1$ and $\beta < 1 - p/\alpha$. If we combine [31, Theorem 7.23 and Example 7.3] with (EMB) then we recover our L_2 -rate from Theorem 3.1 even without logarithmic factor. Finally, it is to mention that [31, 33] consider the *clipped* predictor. The influence of this clipping is not clear, but it maybe the reason for avoiding the logarithmic factors appearing in some results obtained by the integral operator techniques.

To sum up, we use the integral operator technique to recover the best known, and in many cases optimal, L_2 -learning rates previously only obtained by the empirical process technique. In addition, we improve the best known γ -learning rates from [18] for the learning scheme (1) whenever (EMB) is satisfied for $0 < \alpha < 1$ as well as (SRC) and (EVD) are satisfied for $\beta + p < 1$. Recall that the

empirical process technique is not able to provide general γ -learning rates yet. Finally, we show that our γ -learning rates are optimal in all cases where the optimal L_2 -norm learning rate is known.

6. Proofs

First, we summarize some well-known facts that we need for the proofs of our main results. To this end, we use the notation and general assumptions from Section 2.

Since we assume that H is separable according to [32, Corollary 3.2] there exists a ν -zero set $N \subseteq X$, such that k is given by

$$k(x, x') = \sum_{i \geq 1} \mu_i e_i(x) e_i(x')$$

for all $x, x' \in X \setminus N$. Furthermore, the boundedness of k implies $\sum_{i \geq 1} \mu_i e_i^2(x) \leq A^2$ for ν -a.a. $x \in X$ and a constant $A \geq 0$. Motivated by this statement we say, for $\alpha > 0$, that the α -power of k is ν -a.s. bounded if there exists a constant $A \geq 0$ with

$$\sum_{i \geq 1} \mu_i^\alpha e_i^2(x) \leq A^2 \tag{13}$$

for ν -a.a. $x \in X$. Furthermore, we write $\|k_\nu^\alpha\|_\infty$ for the smallest constant with this property and set $\|k_\nu^\alpha\|_\infty := \infty$ if there is no such constant. Consequently, $\|k_\nu^\alpha\|_\infty < \infty$ is an abbreviation of the phrase *the α -power of k is ν -a.s. bounded*. We refer to [32, Proposition 4.2] for the logic behind this notation. Because of the above introduction $\|k_\nu^1\|_\infty < \infty$ is always satisfied. Since the measurable spaces (X, \mathcal{B}) is ν -complete the following theorem from [32, Theorem 5.3] gives an equivalent characterization.

6.1 Theorem (L_∞ -Embeddings) *For $0 < \alpha \leq 1$ the following statements are equivalent:*

- (i) *The α -power of k is ν -a.s. bounded.*
- (ii) $[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)$.

If one (and thus both) of the statements above is true, we have

$$\|[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)\| = \|k_\nu^\alpha\|_\infty .$$

Note that the claimed equality is not a part of [32, Theorem 5.3] but it is contained in the proof of that theorem. The following lemma summarizes further implications of the embedding property.

6.2 Lemma *For $0 < p, \alpha \leq 1$ the following statements are true:*

- (i) (EMB) *implies* $\|[e_i]_\nu\|_{L_\infty(\nu)} \leq \|k_\nu^\alpha\|_\infty \mu_i^{-\alpha/2}$ *for all* $i \geq 1$.
- (ii) (EMB) *implies* $(\mu_i)_{i \geq 1} \in \ell_\alpha(\mathbb{N})$. *If, in addition, the eigenfunctions are uniformly bounded, i.e. $\sup_{i \geq 1} \|[e_i]_\nu\|_{L_\infty(\nu)} < \infty$, then the inverse implication is true.*
- (iii) (EMB) *implies* (EVD) *for* $p = \alpha$. *If, in addition, the eigenfunctions are uniformly bounded, then (EVD) w.r.t. p implies (EMB) for all $\alpha > p$.*

Note that uniformly bounded eigenfunction have been considered e.g. in [19, Assumption 4.1] and [33, Theorem 2], see also the discussion after [32, Theorem 5.3].

Proof. (i) is clear since $(\mu_i^{\alpha/2}[e_i]_\nu)_{i \geq 1}$ is an ONB of $[H]_\nu^\alpha$. (ii) From [32, Proposition 4.4] we know $\sum_{i \geq 1} \mu_i^\alpha \leq \|k_\nu^\alpha\|_\infty^2 < \infty$. The inverse is a consequence of (13). (iii) follows from (ii) together with the monotonicity of the eigenvalues $(\mu_i)_{i \geq 1}$. The inverse is a consequence of: (EVD) w.r.t. p implies $(\mu_i)_{i \geq 1} \in \ell_\alpha(\mathbb{N})$ for $\alpha > p$. \square

The *effective dimension* $\mathcal{N}_\nu : (0, \infty) \rightarrow [0, \infty)$ is defined by

$$\mathcal{N}_\nu(\lambda) := \text{tr}((C_\nu + \lambda)^{-1}C_\nu) = \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda}$$

for $\lambda > 0$, where tr denotes the trace operator. This quantity is widely used in the statistical analysis of LS-SVMs, see e.g. [4, 3, 16, 18], and depends on the decay of the eigenvalues $(\mu_i)_{i \geq 1}$. More precisely, (EVD) for $0 < p \leq 1$ is equivalent to the existence of a constant $C_p > 0$ with

$$\mathcal{N}_\nu(\lambda) \leq C_p \lambda^{-p} \tag{14}$$

for all $\lambda > 0$. In the case $p < 1$ we can choose $C_p = C^p/(1-p)$ according to [4, Proposition 3]. For $p(=\alpha) = 1$ we have $\mathcal{N}_\nu(\lambda) \leq \|C_\nu\|_1 \|(C_\nu + \lambda)^{-1}\| \leq \|C_\nu\|_1 \lambda^{-1}$ and $\|C_\nu\|_1 = \sum_{i \geq 1} \mu_i \leq \|k_\nu^\alpha\|_\infty^2 =: C_p$. For the inverse implication we combine (14) with $i \frac{\mu_i}{\mu_i + \lambda} \leq \mathcal{N}_\nu(\lambda)$ and $\lambda = \mu_i$ to get (EVD) with $C = (2C_p)^{1/p}$.

The *LS-risk* of a measurable function $f : X \rightarrow \mathbb{R}$ is defined by

$$\mathcal{R}_P(f) := \int_{X \times Y} (y - f(x))^2 \, dP(x, y)$$

and the *Bayes-LS-risk* $\mathcal{R}_P^* := \inf_{f: X \rightarrow \mathbb{R}} \mathcal{R}_P(f)$ is achieved by the conditional mean function f_P^* . Moreover, the *LS-excess-risk* is given by $\mathcal{R}_P(f) - \mathcal{R}_P^* = \|[f]_\nu - f_P^*\|_{L_2(\nu)}^2$ and minimizing the LS-risk is equivalent to approximating the conditional mean function in the $L_2(\nu)$ -norm. For $\lambda > 0$ the unique minimizer of

$$\inf_{f \in H} \left\{ \lambda \|f\|_H^2 + \mathcal{R}_P(f) \right\} \tag{15}$$

is given by

$$f_{P,\lambda} := (C_\nu + \lambda)^{-1} g_P \in H, \tag{16}$$

with $g_P := S_\nu f_P^*$ and some times called *infinite sample solution*. As already mentioned in the proof of [29, Theorem 4] the spectral decomposition in (4) yields

$$f_{P,\lambda} = \sum_{i \geq 1} \frac{\mu_i^{1/2}}{\mu_i + \lambda} a_i \mu_i^{1/2} e_i \in (\ker I_\nu)^\perp, \quad \text{and} \quad f_P^* - [f_{P,\lambda}]_\nu = \sum_{i \geq 1} \frac{\lambda}{\mu_i + \lambda} a_i [e_i]_\nu, \tag{17}$$

with $a_i := \langle f_P^*, [e_i]_\nu \rangle_{L_2(\nu)}$ for $i \geq 1$. For the second identity in (17) we have to assume $f_P^* \in [H]_\nu^0$. Note that the predictor $f_{D,\lambda}$, for a data set $D = \{(x_i, y_i)\}_{i=1}^n$, given in (1) is the unique minimizer of (15) w.r.t. the *empirical* measure $D := \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ and hence $f_{D,\lambda}$ is given by (16) w.r.t. the corresponding empirical quantities. For the later proof we will need the integral operator $C_{D_X} : H \rightarrow H$ w.r.t. the empirical marginal distribution $D_X = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ on X . In order to avoid subsubscripts we denote this operator by C_D .

6.1. Some Bounds

In this section we further exploit the spectral representations in (4). The first lemma describes the connection of the γ -power norm and the H -norm.

6.3 Lemma *For $0 \leq \gamma \leq 1$ and $f \in H$ we have $\|[f]_\nu\|_\gamma \leq \|C_\nu^{\frac{1-\gamma}{2}} f\|_H$. If, in addition, $\gamma < 1$ or $f \perp \ker I_\nu$ is satisfied then we have equality.*

Proof. Let us fix a function $f \in H$. Because $(\mu_i^{1/2} e_i)_{i \geq 1}$ is an ONB of $(\ker I_\nu)^\perp$, there exists a $g \in \ker I_\nu$ with $f = \sum_{i \geq 1} a_i \mu_i^{1/2} e_i + g$, where $a_i = \langle f, \mu_i^{1/2} e_i \rangle_H$ for all $i \geq 1$. Since $(\mu_i^{\gamma/2} [e_i]_\nu)_{i \geq 1}$ is an ONB of $[H]_\nu^\gamma$ Parseval yields

$$\|[f]_\nu\|_\gamma^2 = \left\| \sum_{i \geq 1} a_i \mu_i^{\frac{1-\gamma}{2}} \mu_i^{\gamma/2} [e_i]_\nu \right\|_\gamma^2 = \sum_{i \geq 1} \mu_i^{1-\gamma} a_i^2 .$$

For $\gamma < 1$ the spectral decomposition in (4) together with Parseval w.r.t. the ONS $(\mu_i^{1/2} e_i)_{i \geq 1}$ in H yields

$$\|C_\nu^{\frac{1-\gamma}{2}} f\|_H^2 = \left\| \sum_{i \geq 1} \mu_i^{\frac{1-\gamma}{2}} a_i \mu_i^{1/2} e_i \right\|_H^2 = \sum_{i \geq 1} \mu_i^{1-\gamma} a_i^2 .$$

For $\gamma = 1$ we have $C_\nu^{\frac{1-\gamma}{2}} = \text{Id}_H$. Pythagoras and Parseval w.r.t. the ONS $(\mu_i^{1/2} e_i)_{i \geq 1}$ in H yields

$$\|C_\nu^{\frac{1-\gamma}{2}} f\|_H^2 = \left\| \sum_{i \geq 1} a_i \mu_i^{1/2} e_i + g \right\|_H^2 = \left\| \sum_{i \geq 1} a_i \mu_i^{1/2} e_i \right\|_H^2 + \|g\|_H^2 = \sum_{i \geq 1} a_i^2 + \|g\|_H^2 . \quad \square$$

The next lemma describes how the effective dimension comes into play. Parts of this lemma are already contained in [27, Assumption 3] and the following discussion.

6.4 Lemma *The following statements are satisfied for all $\lambda > 0$:*

- (i) *If $\|k_\nu^\alpha\|_\infty < \infty$ then $\|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 \leq \|k_\nu^\alpha\|_\infty^2 \lambda^{-\alpha}$ for ν -a.a. $x \in X$.*
- (ii) $\int_X \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 d\nu(x) = \mathcal{N}_\nu(\lambda)$.

Proof. Let us fix $\lambda > 0$. Since H is separable and k measurable the map $X \rightarrow H$, $x \mapsto k(x, \cdot)$ is measurable, see [31, Lemma 4.25]. Consequently, $\|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2$ depends measurable on x . Let us fix an arbitrary ONB $(e_j)_{j \in \mathcal{J}}$ of $\ker I_\nu$. Thus, $(\mu_i^{1/2} e_i)_{i \geq 1} \cup (e_j)_{j \in \mathcal{J}}$ is an ONB of H and k satisfies

$$k(x, \cdot) = \sum_{i \geq 1} \mu_i^{1/2} e_i(x) \mu_i^{1/2} e_i + \sum_{j \in \mathcal{J}} e_j(x) e_j .$$

for all $x \in X$. Together with the spectral decomposition in (4) and Parseval we get

$$\|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 = \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda} e_i^2(x) + \frac{1}{\lambda} \sum_{j \in \mathcal{J}} e_j^2(x)$$

for all $x \in X$. Since H is separable the index set \mathcal{J} is at most countable. Moreover, $e_j \in \ker I_\nu$ for all $j \in \mathcal{J}$ implies that the second summand on the r.h.s. vanishes for ν -a.a. $x \in X$. Consequently, we have

$$\|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 = \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda} e_i^2(x)$$

for ν -a.a. $x \in X$. Now, Statement (i) is a consequence of Lemma A.1

$$\sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda} e_i^2(x) = \sum_{i \geq 1} \frac{\mu_i^{1-\alpha}}{\mu_i + \lambda} \mu_i^\alpha e_i^2(x) \leq \left(\sum_{i \geq 1} \mu_i^\alpha e_i^2(x) \right) \sup_{i \geq 1} \frac{\mu_i^{1-\alpha}}{\mu_i + \lambda} \leq \|k_\nu^\alpha\|_\infty^2 \lambda^{-\alpha}$$

for ν -a.a. $x \in X$. In order to prove Statement (ii) we use the fact that $([e_i])_{i \geq 1}$ is an ONS in $L_2(\nu)$ and the monotone convergence theorem

$$\int_X \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 d\nu(x) = \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda} \int_X e_i^2(x) d\nu(x) = \text{tr}((C_\nu + \lambda)^{-1} C_\nu) . \quad \square$$

The next lemma uses the representations in (17) to provide bounds on the γ -power norm of $f_{P,\lambda}$ and $f_P^* - f_{P,\lambda}$.

6.5 Lemma *Let $f_P^* \in [H]_\nu^\beta$ for some $0 \leq \beta \leq 2$. Then the following bounds are satisfied for $\lambda > 0$:*

$$\|[f_{P,\lambda}]_\nu - f_P^*\|_\gamma^2 \leq \|f_P^*\|_\beta^2 \lambda^{\beta-\gamma} \quad \text{for } 0 \leq \gamma \leq \beta, \quad (18)$$

$$\|[f_{P,\lambda}]_\nu\|_\gamma^2 \leq \|f_P^*\|_{\gamma \wedge \beta}^2 \lambda^{-(\gamma-\beta)_+} \quad \text{for } \gamma \geq 0. \quad (19)$$

Here we used the abbreviation $(\gamma - \beta)_+ = \max\{0, \gamma - \beta\}$. Note that (18) in the case of $\gamma \in \{0, 1\}$ is contained in [29, Theorem 4]. Since, in the case $\beta \geq \gamma = 1$, the ν -equivalence class f_P^* has a (unique) representative $f_P^* \in H$ with $f_P^* \perp \ker I_\nu$ and $f_{P,\lambda} \perp \ker I_\nu$ holds according to (17) we can use the equality from Lemma 6.4 and exchange the left hand side of (18) and (19) by $\|f_{P,\lambda} - f_P^*\|_H$ resp. $\|f_{P,\lambda}\|_H$.

Proof. (18) Since $f_P^* \in [H]_\nu^\beta \subseteq [H]_\nu^0$ we can use the spectral representations in (17). Parseval w.r.t. the ONB $(\mu_i^{\gamma/2} [e_i]_\nu)_{i \geq 1}$ of $[H]_\nu^\gamma$ yields

$$\|f_P^* - [f_{P,\lambda}]_\nu\|_\gamma^2 = \lambda^2 \sum_{i \geq 1} \left(\frac{\mu_i^{-\gamma/2}}{\mu_i + \lambda} \right)^2 a_i^2 = \lambda^2 \sum_{i \geq 1} \left(\frac{\mu_i^{\frac{\beta-\gamma}{2}}}{\mu_i + \lambda} \right)^2 \mu_i^{-\beta} a_i^2 .$$

If we estimate the fraction on the r.h.s. with Lemma A.1 and use the fact, that $(\mu_i^{\beta/2} [e_i]_\nu)_{i \geq 1}$ is an ONB of $[H]_\nu^\beta$, we get

$$\|f_P^* - [f_{P,\lambda}]_\nu\|_\gamma^2 \leq \lambda^{\beta-\gamma} \sum_{i \geq 1} \mu_i^{-\beta} a_i^2 = \lambda^{\beta-\gamma} \|f_P^*\|_\beta^2 .$$

(19) Again the spectral representation in (17) together with Parseval yields

$$\|[f_{P,\lambda}]_\nu\|_\gamma^2 = \sum_{i \geq 1} \left(\frac{\mu_i}{\mu_i + \lambda} \right)^2 \mu_i^{-\gamma} a_i^2 .$$

In the case of $\gamma \leq \beta$ we estimate the fraction by 1 and then Parseval gives us

$$\|[f_{P,\lambda}]_\nu\|_\gamma^2 \leq \sum_{i \geq 1} \mu_i^{-\gamma} a_i^2 = \|f_P^*\|_\gamma^2 .$$

In the case of $\gamma > \beta$ additionally use Lemma A.1 and get

$$\| [f_{P,\lambda}]_\nu \|_\gamma^2 = \sum_{i \geq 1} \left(\frac{\mu_i^{1-\frac{\gamma-\beta}{2}}}{\mu_i + \lambda} \right)^2 \mu_i^{-\beta} a_i^2 \leq \lambda^{-(\gamma-\beta)} \sum_{i \geq 1} \mu_i^{-\beta} a_i^2 = \lambda^{-(\gamma-\beta)} \| f_P^* \|_\beta^2 . \quad \square$$

From the bounds obtained in the previous lemma we directly get the following $L_\infty(\nu)$ bounds. Note that some parts of the following lemma are already known from [32, Corollary 5.5].

6.6 Corollary *Let $f_P^* \in [H]_\nu^\beta$ and $\|k_\nu^\alpha\|_\infty < \infty$ for some $0 \leq \beta \leq 2$ and $0 < \alpha \leq 1$. Then the following bounds are satisfied for $\lambda > 0$:*

$$\| [f_{P,\lambda}]_\nu - f_P^* \|_{L_\infty(\nu)}^2 \leq 2(\|f_P^*\|_{L_\infty(\nu)}^2 + \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_\beta^2) \lambda^{\beta-\alpha} \quad \text{if } f_P^* \in L_\infty(\nu) \text{ and } \lambda \leq 1, \quad (20)$$

$$\| [f_{P,\lambda}]_\nu \|_{L_\infty(\nu)}^2 \leq \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_{\alpha \wedge \beta}^2 \lambda^{-(\alpha-\beta)_+} . \quad (21)$$

Proof. (21) is a direct consequence of Theorem 6.1 and (19) with $\gamma = \alpha$. (20) In the case of $\beta \leq \alpha$ we use the triangle inequality, Inequality (21), and $\lambda \leq 1$

$$\begin{aligned} \| f_P^* - [f_{P,\lambda}]_\nu \|_{L_\infty(\nu)}^2 &\leq 2\|f_P^*\|_{L_\infty(\nu)}^2 + 2\| [f_{P,\lambda}]_\nu \|_{L_\infty(\nu)}^2 \\ &\leq 2(\|f_P^*\|_{L_\infty(\nu)}^2 + \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_\beta^2) \lambda^{-(\alpha-\beta)} . \end{aligned}$$

In the case of $\beta > \alpha$ we use Theorem 6.1 and (18)

$$\| f_P^* - [f_{P,\lambda}]_\nu \|_{L_\infty(\nu)}^2 \leq \|k_\nu^\alpha\|_\infty^2 \|f_P^* - [f_{P,\lambda}]_\nu\|_\alpha^2 \leq \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_\beta^2 \lambda^{\beta-\alpha} . \quad \square$$

6.2. Upper Rates

Using the standard technique, we split $\| [f_{D,\lambda}]_\nu - f_P^* \|_\gamma$ into two parts:

$$\| [f_{D,\lambda}]_\nu - f_P^* \|_\gamma \leq \| [f_{D,\lambda} - f_{P,\lambda}]_\nu \|_\gamma + \| [f_{P,\lambda}]_\nu - f_P^* \|_\gamma , \quad (22)$$

the *estimation error* $\| [f_{D,\lambda} - f_{P,\lambda}]_\nu \|_\gamma$ and the *approximation error* $\| [f_{P,\lambda}]_\nu - f_P^* \|_\gamma$. A bound on the approximation error is given in Lemma 6.5 and the following oracle inequality controls the estimation error.

6.7 Theorem (Estimation Error - Oracle Inequality) *Let H be a separable RKHS on X w.r.t. a bounded and measurable kernel k , P be a probability measure on $X \times Y$ with $|P|_2 < \infty$ and marginal distribution $\nu = P_X$, and $0 \leq \gamma \leq 1$. Furthermore, we assume $\|f_P^*\|_{L_\infty(\nu)} < \infty$, (SRC) with $\beta = \gamma$, (EMB) for $0 < \alpha \leq 1$, and (MOM). If we define the abbreviations*

- (i) $g_\lambda := \log\left(2e\mathcal{N}_\nu(\lambda) \frac{\|C_\nu\| + \lambda}{\|C_\nu\|}\right)$,
- (ii) $A_{\lambda,\tau} := 8\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda \lambda^{-\alpha}$, and
- (iii) $L_\lambda := \max\{L, \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}\}$

then for $\tau \geq 1$, $\lambda > 0$, and $n \geq A_{\lambda, \tau}$ we have with P^n -probability $\geq 1 - 4e^{-\tau}$

$$\left\| C_{\nu}^{\frac{1-\gamma}{2}} (f_{D, \lambda} - f_{P, \lambda}) \right\|_H^2 \leq \frac{576\tau^2}{n\lambda^\gamma} \left(\sigma^2 \mathcal{N}_\nu(\lambda) + \|k_\nu^\alpha\|_\infty^2 \frac{\|f_P^* - [f_{P, \lambda}]_\nu\|_{L_2(\nu)}^2}{\lambda^\alpha} + 2\|k_\nu^\alpha\|_\infty^2 \frac{L_\lambda^2}{n\lambda^\alpha} \right).$$

According to Lemma 6.3 the same result is true for $\|[f_{D, \lambda} - f_{P, \lambda}]_\nu\|_\gamma^2$. Moreover, in the case of $\gamma = 1$ the left hand side coincides with $\|f_{D, \lambda} - f_{P, \lambda}\|_H$. Our proof is base on an argument tracing back to [30]. We refine the analysis with some ideas from [4] and [16] under the embedding property. We split the proof into several lemmas: the fist one improves [16, Lemma 18] under the additional Assumption (EMB).

6.8 Lemma *Let the assumptions of Theorem 6.7 be satisfied. For $\tau \geq 1$, $\lambda > 0$, and $n \geq 1$ the operator norm satisfies with ν^n -probability $\geq 1 - 2e^{-\tau}$*

$$\|(C_\nu + \lambda)^{-1/2} (C_\nu - C_D) (C_\nu + \lambda)^{-1/2}\| \leq \frac{4\|k_\nu^\alpha\|_\infty^2 \tau g \lambda}{3n\lambda^\alpha} + \sqrt{\frac{2\|k_\nu^\alpha\|_\infty^2 \tau g \lambda}{n\lambda^\alpha}}.$$

Proof. First, we define $C_x : H \rightarrow H$ the *integral* operator w.r.t. the point measure at $x \in X$, i.e.

$$C_x f = f(x)k(x, \cdot) = \langle f, k(x, \cdot) \rangle_H k(x, \cdot).$$

Since the operator C_x has rank one C_x is a Hilbert-Schmidt operator. Now, we consider the random variable $\xi_1 : X \rightarrow \mathcal{L}_2(H)$,

$$\xi_1(x) := (C_\nu + \lambda)^{-1/2} C_x (C_\nu + \lambda)^{-1/2}$$

with values in the space of Hilbert-Schmidt operators on H . Since H is a separable RKHS w.r.t. a measurable and bounded kernel, the map $X \rightarrow \mathcal{L}_2(H)$, $x \mapsto C_x$ is bounded and measurable. Moreover, the maps $x \mapsto C_x$ and $x \mapsto \xi_1(x)$ are Bochner integrable w.r.t. a arbitrary probability measure μ on X and [10, Chapter II.2 Theorem 6] yields

$$\mathbb{E}_\mu \xi_1 = (C_\nu + \lambda)^{-1/2} (\mathbb{E}_{x \sim \mu} C_x) (C_\nu + \lambda)^{-1/2} = (C_\nu + \lambda)^{-1/2} C_\mu (C_\nu + \lambda)^{-1/2}$$

If we exploit this identity in the case of $\mu = \nu = P_X$ and $\mu = D_X$, then we get

$$\frac{1}{n} \sum_{i=1}^n (\xi_1(x_i) - \mathbb{E}_\nu \xi_1) = \mathbb{E}_{D_X} \xi_1 - \mathbb{E}_\nu \xi_1 = (C_\nu + \lambda)^{-1/2} (C_D - C_\nu) (C_\nu + \lambda)^{-1/2}$$

for all $D = ((x_i, y_i))_{i=1}^n \in (X \times Y)^n$. Using the self-adjointness of $(C_\nu + \lambda)^{-1/2}$ we get $\xi_1(x) = \langle \cdot, h_x \rangle_H h_x$ with $h_x := (C_\nu + \lambda)^{-1/2} k(x, \cdot) \in H$ for all $x \in X$. An application of Lemma 6.4 yields the supremum bound

$$\|\xi_1(x)\| = \|h_x\|_H^2 \leq \|k_\nu^\alpha\|_\infty^2 \lambda^{-\alpha} =: B$$

for ν -a.a. $x \in X$. For two self-adjoint operators A, B on a Hilbert space we write $A \preceq B$ iff $B - A$ is a positive operator. Since $\xi_1(x)^2 = \xi_1(x) \|h_x\|_H^2 \preceq B \xi_1(x)$ also the *variance* bound

$$\mathbb{E}_\nu (\xi_1^2) \preceq B \mathbb{E}_\nu \xi_1 = B (C_\nu + \lambda)^{-1} C_\nu =: V.$$

Moreover, we have $\|V\| = B \frac{\|C_\nu\|}{\|C_\nu\| + \lambda} \leq B$ and $\text{tr}(V) = B\mathcal{N}_\nu(\lambda)$. Consequently, Theorem A.3 is applicable and yields the assertion because of $g(V) = g_\lambda$. \square

6.9 Lemma *Let the assumptions of Theorem 6.7 be satisfied. For $\tau \geq 1$, $\lambda > 0$ and $n \geq 1$, we have with ν^n -probability $\geq 1 - 2e^{-\tau}$*

$$\begin{aligned} & \left\| (C_\nu + \lambda)^{-1/2} \left((g_D - C_D f_{P,\lambda}) - (g_P - C_\nu f_{P,\lambda}) \right) \right\|_H^2 \\ & \leq \frac{64\tau^2}{n} \left(\sigma^2 \mathcal{N}_\nu(\lambda) + \|k_\nu^\alpha\|_\infty^2 \frac{\|f_P^* - [f_{P,\lambda}]_\nu\|_0^2}{\lambda^\alpha} + 2\|k_\nu^\alpha\|_\infty^2 \frac{\max\{L^2, \|f_P^* - [f_{P,\lambda}]_\nu\|_{L^\infty(\nu)}^2\}}{n\lambda^\alpha} \right). \end{aligned}$$

Proof. We consider the random variable $\xi_2 : X \times Y \rightarrow H$,

$$\xi_2(x, y) := (y - f_{P,\lambda}(x))(C_\nu + \lambda)^{-1/2} k(x, \cdot) .$$

Since H is a separable RKHS w.r.t. a bounded and measurable kernel and (MOM) is satisfied, the maps ξ_2 and $(x, y) \mapsto (y - f_{P,\lambda}(x))k(x, \cdot)$ are measurable and Bochner integrable w.r.t. P and the empirical measure D . [10, Chapter II.2 Theorem 6] yields for $Q \in \{P, D\}$

$$\mathbb{E}_Q \xi_2 = (C_\nu + \lambda)^{-1/2} \left(\mathbb{E}_{(x,y) \sim Q} y k(x, \cdot) - \mathbb{E}_{x \sim Q} f_{P,\lambda}(x) k(x, \cdot) \right) = (C_\nu + \lambda)^{-1/2} (g_Q - C_{Q_X} f_{P,\lambda}) .$$

Consequently, we get

$$\frac{1}{n} \sum_{i=1}^n \left(\xi_2(x_i, y_i) - \mathbb{E}_P \xi_2 \right) = \mathbb{E}_D \xi_2 - \mathbb{E}_P \xi_2 = (C_\nu + \lambda)^{-1/2} \left((g_D - C_D f_{P,\lambda}) - (g_P - C_\nu f_{P,\lambda}) \right) .$$

In order to apply Bernstein's inequality we need to bound the m -th moment for $m \geq 2$:

$$\mathbb{E}_P \|\xi_2\|_H^m = \int_X \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^m \int_Y |y - f_{P,\lambda}(x)|^m P(dy|x) d\nu(x) . \quad (23)$$

First, we consider the inner integral: Using the triangle inequality and (MOM) yields

$$\begin{aligned} \int_Y |y - f_{P,\lambda}(x)|^m P(dy|x) & \leq 2^{m-1} \left(\|\text{id}_Y - f_P^*(x)\|_{L_m(P(\cdot|x))}^m + |f_P^*(x) - f_{P,\lambda}(x)|^m \right) \\ & \leq \frac{1}{2} m! (2L)^{m-2} 2\sigma^2 + 2^{m-1} |f_P^*(x) - f_{P,\lambda}(x)|^m . \end{aligned}$$

for ν -a.a. $x \in X$. If we plug this bound into the outer integral we get two terms. For both terms we use Lemma 6.4. The first term is estimated by

$$\frac{1}{2} m! (2L)^{m-2} 2\sigma^2 \int_X \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^m d\nu(x) \leq \frac{1}{2} m! \left(\frac{2L \|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}} \right)^{m-2} 2\sigma^2 \mathcal{N}_\nu(\lambda)$$

and the term second by

$$\begin{aligned}
& 2^{m-1} \int_X \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^m |f_P^*(x) - f_{P,\lambda}(x)|^m d\nu(x) \\
& \leq \frac{1}{2} \left(\frac{2\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}} \right)^m \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}^{m-2} \int_X |f_P^*(x) - f_{P,\lambda}(x)|^2 d\nu(x) \\
& \leq \frac{1}{2} m! \left(\frac{2\|k_\nu^\alpha\|_\infty \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}}{\lambda^{\alpha/2}} \right)^{m-2} 2 \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_2(\nu)}^2 \frac{\|k_\nu^\alpha\|_\infty^2}{\lambda^\alpha} .
\end{aligned}$$

Continuing estimate (23) we get that $\mathbb{E}_P \|\xi_2\|_H^m$ is bounded by

$$\frac{1}{2} m! \left(\frac{2\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}} \max\{L, \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}\} \right)^{m-2} 2 \left(\sigma^2 \mathcal{N}_\nu(\lambda) + \|f_P^* - [f_{P,\lambda}]_\nu\|_0^2 \frac{\|k_\nu^\alpha\|_\infty^2}{\lambda^\alpha} \right)$$

and an application of Bernstein's inequality from Theorem A.2 yield the assertion. \square

Proof of Theorem 6.7. Let us fix $\tau \geq 1$, $\lambda > 0$ and $n \geq A_{\lambda,\tau}$. For $D \in (X \times Y)^n$ the representation $f_{D,\lambda} = (C_D + \lambda)^{-1} g_D$ from (16), w.r.t. the empirical measure D , yields

$$C_\nu^{\frac{1-\gamma}{2}} (f_{D,\lambda} - f_{P,\lambda}) = C_\nu^{\frac{1-\gamma}{2}} (C_D + \lambda)^{-1} (g_D - (C_D + \lambda) f_{P,\lambda}) .$$

If we combine this with the identity $\text{Id}_H = (C_\nu + \lambda)^{-1/2} (C_\nu + \lambda)^{1/2}$ then we get

$$\left\| C_\nu^{\frac{1-\gamma}{2}} (f_{D,\lambda} - f_{P,\lambda}) \right\|_H \leq \left\| C_\nu^{\frac{1-\gamma}{2}} (C_\nu + \lambda)^{-1/2} \right\| \tag{24a}$$

$$\cdot \left\| (C_\nu + \lambda)^{1/2} (C_D + \lambda)^{-1} (C_\nu + \lambda)^{1/2} \right\| \tag{24b}$$

$$\cdot \left\| (C_\nu + \lambda)^{-1/2} (g_D - (C_D + \lambda) f_{P,\lambda}) \right\|_H \tag{24c}$$

for all $D \in (X \times Y)^n$. Now, we consider the three factors on the r.h.s. separately. Let us start with Factor (24a). An application of Lemma A.1 yields

$$(24a) = \left\| C_\nu^{\frac{1-\gamma}{2}} (C_\nu + \lambda)^{-1/2} \right\| = \sup_{i \geq 1} \left(\frac{\mu_i^{1-\gamma}}{\mu_i + \lambda} \right)^{1/2} \leq \lambda^{-\gamma/2} . \tag{25}$$

Factor (24c) can be rearranged using $f_{P,\lambda} = (C_\nu + \lambda)^{-1} g_P$ from (16)

$$\begin{aligned}
(C_\nu + \lambda)^{-1/2} (g_D - (C_D + \lambda) f_{P,\lambda}) &= (C_\nu + \lambda)^{-1/2} (g_D - (C_D - C_\nu + C_\nu + \lambda) f_{P,\lambda}) \\
&= (C_\nu + \lambda)^{-1/2} ((g_D - C_D f_{P,\lambda}) - (g_P - C_\nu f_{P,\lambda})) .
\end{aligned}$$

Consequently, the square of Factor (24c) coincides with the r.h.s. in Lemma 6.9 and this lemma provides a suitable bound. Finally, in order to estimate (24b) we start with the following identity

$$\begin{aligned}
(C_D + \lambda) &= (C_D - C_\nu + C_\nu + \lambda) \\
&= (C_\nu + \lambda)^{1/2} (\text{Id} - (C_\nu + \lambda)^{-1/2} (C_\nu - C_D) (C_\nu + \lambda)^{-1/2}) (C_\nu + \lambda)^{1/2} .
\end{aligned}$$

If we take the inverse and multiply it by the factor $(C_\nu + \lambda)^{1/2}$ from the left and from the right, then

we get

$$(24b) = \left\| \left(\text{Id} - (C_\nu + \lambda)^{-1/2} (C_\nu - C_D) (C_\nu + \lambda)^{-1/2} \right)^{-1} \right\|.$$

Lemma 6.8 gives us an estimate for the operator norm of $(C_\nu + \lambda)^{-1/2} (C_\nu - C_D) (C_\nu + \lambda)^{-1/2}$. Continuing the estimate from Lemma 6.8 with $n \geq A_{\lambda, \tau}$ yields

$$\left\| (C_\nu + \lambda)^{-1/2} (C_\nu - C_D) (C_\nu + \lambda)^{-1/2} \right\| \leq \frac{2}{3}$$

with ν^n -probability $\geq 1 - 2e^{-\tau}$. Consequently, the Neumann series is applicable and gives

$$(24b) \leq \sum_{k=0}^{\infty} \left(\frac{2}{3} \right)^k = 3 \quad (26)$$

with ν^n -probability $\geq 1 - 2e^{-\tau}$. Now, we get the claimed bound, with P^n -probability $\geq 1 - 4e^{-\tau}$, if we continue the estimate in (24) with (25), Lemma 6.9, and (26). \square

Proof of Theorem 3.1. In both cases, $\beta + p \leq \alpha$ and $\beta + p > \alpha$, for the given asymptotic of the regularization parameter sequence $(\lambda_n)_{n \geq 1}$ there is an index bound $n_0 \geq 1$ such that $\lambda_n \leq 1 \wedge \|C_\nu\|$ and $n \geq A_{\lambda_n, \tau}$ is satisfied for all $n \geq n_0$. Consequently, for $n \geq n_0$, we can apply Theorem 6.7. Together with Lemma 6.3 and (14) we get

$$\| [f_{D, \lambda_n} - f_{P, \lambda_n}]_\nu \|_\gamma^2 \leq K_0 \frac{\tau^2}{n \lambda_n^{\gamma + \max\{p, \alpha - \beta\}}} \left(1 + \frac{1}{n \lambda_n^{\max\{\alpha, 2\alpha - \beta\} - \max\{p, \alpha - \beta\}}} \right).$$

with $K_0 = 576 \max\{\sigma^2 C_p + \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_\beta^2, 2\|k_\nu^\alpha\|_\infty^2 \max\{L, 2(\|f_P^*\|_{L_\infty(\nu)}^2 + \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_\beta^2)\}$. Since the second term inside the brackets is, in both cases, bounded there is a constant $K_1 > 0$ with

$$\| [f_{D, \lambda_n} - f_{P, \lambda_n}]_\nu \|_\gamma^2 \leq K_0 K_1 \frac{\tau^2}{n \lambda_n^{\gamma + \max\{p, \alpha - \beta\}}}.$$

Together with Equation (22) and Lemma 6.5 we get the assertion, in both cases,

$$\| [f_{D, \lambda_n}]_\nu - f_P^* \|_\gamma^2 \leq \tau^2 2 (\|f_P^*\|_\beta^2 + K_0 K_1) \lambda_n^{\beta - \gamma}. \quad \square$$

6.3. Lower Rates

In order to prove γ -lower rates we establish the following lower bound.

6.10 Lemma (Lower Bound) *Let H be a separable RKHS on X w.r.t. a bounded and measurable kernel k , and ν be a probability distribution on X such that (EMB) and (EVD+) are satisfied for some $0 < p \leq \alpha \leq 1$. Then, for all parameters $0 < \beta \leq 2$, $0 \leq \gamma \leq 1$ with $\gamma < \beta$ and all constants $\sigma, L, B, B_\infty > 0$ there exist constants $0 < \varepsilon_0 \leq 1$ and $C_1, C_2 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$ there are probability measures $P_0, P_1, \dots, P_{M_\varepsilon}$ with marginal distribution ν on X satisfying $\|f_{P_j}^*\|_{L_\infty(\nu)}^2 \leq B_\infty$, (SRC) w.r.t. B , and (MOM) w.r.t. σ, L . Moreover, these measures satisfy:*

$$(i) \quad 2^{C_2 \varepsilon^{-u}} \leq M_\varepsilon \leq 2^{3C_2 \varepsilon^{-u}},$$

$$(ii) \quad \|f_{P_i}^* - f_{P_j}^*\|_\gamma^2 \geq 4\varepsilon \text{ for all } i, j \in \{0, 1, \dots, M_\varepsilon\} \text{ with } i \neq j, \text{ and}$$

(iii) $\inf_{\Psi} \max_{j=0,1,\dots,M_\varepsilon} P_j^n(D : \Psi(D) \neq j) \geq \frac{\sqrt{M_\varepsilon}}{\sqrt{M_\varepsilon}+1} \left(1 - C_1 n \varepsilon^{\frac{\max\{\alpha,\beta\}+p}{\max\{\alpha,\beta\}-\gamma}} - \frac{1}{2 \log(M_\varepsilon)}\right)$ for all $n \geq 1$, where the infimum is taken over all measurable functions $\Psi : (X \times Y)^n \rightarrow \{0, 1, \dots, M_\varepsilon\}$.

Note that the probability measures P_j also depend on ε although we omit this in the notation. We recall that just one probability measure ν on X with the required properties is needed to construct distributions on $X \times Y$ that are *difficult* to learn. The proof is an application of the following theorem from Tsybakov [36].

6.11 Theorem (Lower Bound) *Let P_0, P_1, \dots, P_M be a family of probability measures on a measurable space (Ω, \mathcal{A}) with $M \geq 2$. Moreover, we assume $P_j \ll P_0$ for all $j = 1, \dots, M$ and $\alpha_* := \frac{1}{M} \sum_{j=1}^M K(P_j, P_0) \in (0, \infty)$, where $K(P_j, P_0)$ denotes the Kullback-Leibler divergence from P_0 to P_j . Then,*

$$\inf_{\Psi} \max_{j=0,1,\dots,M} P_j(\omega \in \Omega : \Psi(\omega) \neq j) \geq \frac{\sqrt{M}}{\sqrt{M}+1} \left(1 - \frac{3\alpha_*}{\log(M)} - \frac{1}{2 \log(M)}\right)$$

is satisfied, where the infimum is taken over all measurable functions $\Psi : \Omega \rightarrow \{0, 1, \dots, M\}$.

Proof. From Tsybakov [36, Proposition 2.3] we know, that

$$\sup_{0 < \tau < 1} \frac{\tau M}{1 + \tau M} \left(1 + \frac{\alpha_* + \sqrt{\frac{\alpha_*}{2}}}{\log(\tau)}\right)$$

is a lower bound for the l.h.s. If we choose $\tau = M^{-1/2}$ and use the estimate $\sqrt{2\alpha_*} \leq \frac{1}{2} + \alpha_*$ afterwards, then we get the assertion. \square

We use this theorem for the measurable space $\Omega = (X \times Y)^n$ and follow the suggestion of [4, 3] in order to construct a family of probability measures P_0, P_1, \dots, P_M . To this end, let the assumptions of Lemma 6.10 be satisfied and set $\bar{\sigma} := \min\{\sigma, L\}$. Moreover, we define for a measurable function $f : X \rightarrow Y$ and $x \in X$ the conditional distribution $P_f(\cdot | x) := \mathcal{N}(f(x), \bar{\sigma}^2)$ as the normal distribution on $Y = \mathbb{R}$ with mean $f(x)$ and variance $\bar{\sigma}^2$. Consequently, $P_f(A) := \int_X \int_Y \mathbb{1}_A(x, y) P_f(dy|x) d\nu(x)$ for $A \in \mathcal{B} \otimes \mathcal{B}(Y)$ defines a probability measure on $X \times Y$ with marginal distribution ν on X . For this reason the corresponding power spaces $[H]_\nu^\alpha$ are independent of f . Since $P_f = P_{f'}$ is satisfied for $f' = f$ ν -a.s. we define $P_{[f]_\nu}$ for ν -equivalence classes. Moreover, for $f \in L_2(\nu)$ we get $\|P_f\|_2^2 = \bar{\sigma}^2 + \|f\|_{L_2(\nu)}^2 < \infty$ and the conditional mean function $f_{P_f}^*$ of P_f coincides with f . The properties of the normal distribution implies (MOM) w.r.t. $\sigma = L = \bar{\sigma}$, namely

$$\int_Y |y - f(x)|^m P_f(dy|x) = \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{m+1}{2}\right) (\bar{\sigma}\sqrt{2})^m \leq \frac{1}{2} m! \bar{\sigma}^m$$

for all $x \in X$, where Γ denotes the gamma function. To sum up, we reduced the construction of probability measures to the construction of functions $f_0, f_1, \dots, f_M \in L_\infty(\nu) \cap [H]_\nu^\beta$ with $\|f\|_{L_\infty(\nu)}^2 \leq B_\infty$ and $\|f\|_\beta^2 \leq B$. Before we start with the construction the following lemma from [3, Proposition 6.2] describes the Kullback-Leibler divergence between these measures.

6.12 Lemma (Kullback-Leibler Divergence) *For $f, f' \in L_2(\nu)$ and $n \geq 1$ the Kullback-Leibler divergence satisfies*

$$K(P_f^n, P_{f'}^n) := \int_{(X \times Y)^n} \log\left(\frac{dP_f^n}{dP_{f'}^n}\right) dP_{f'}^n = \frac{n}{2\bar{\sigma}^2} \|f - f'\|_{L_2(\nu)}^2 .$$

For the construction of suitable functions we use binary strings $\omega = (\omega_1, \dots, \omega_m) \in \{0, 1\}^m$ and define

$$f_\omega := 2 \left(\frac{8\varepsilon}{m} \right)^{1/2} \sum_{i=1}^m \omega_i \mu_{i+m}^{\gamma/2} [e_{i+m}]_\nu$$

for $0 < \varepsilon \leq 1$. Since the sum is finite we have $f_\omega \in [H]_\nu \subseteq L_\infty(\nu) \cap [H]_\nu^\beta$. Next, we establish sufficient conditions on ε and m such that the bounds $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ and $\|f_\omega\|_\beta^2 \leq B$ are satisfied.

6.13 Lemma *Under the assumptions of Lemma 6.10 there are constants $U > 0$ and $0 < \varepsilon_1 \leq 1$ such that the bounds $\|f_\omega\|_\beta^2 \leq B$ and $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ are satisfied for all $0 < \varepsilon \leq \varepsilon_1$ and all $m \leq U\varepsilon^{-u}$ with $u := \frac{p}{\max\{\alpha, \beta\} - \gamma}$.*

Proof. Let us fix $m \in \mathbb{N}$ and $0 < \varepsilon \leq 1$. (EVD+) and $\gamma < \beta$ implies

$$\|f_\omega\|_\beta^2 = \frac{32\varepsilon}{m} \sum_{i=1}^m \omega_i^2 \mu_{i+m}^{-(\beta-\gamma)} \leq 32\varepsilon \mu_{2m}^{-(\beta-\gamma)} \leq 32c^{\gamma-\beta} 2^{\frac{\beta-\gamma}{p}} \varepsilon m^{-\frac{\beta-\gamma}{p}} .$$

Consequently, for $m \leq U_1 \varepsilon^{-\frac{p}{\beta-\gamma}}$ with $U_1 := \frac{1}{2} c^p (B/32)^{\frac{p}{\beta-\gamma}}$ we have $\|f_\omega\|_\beta^2 \leq B$. In the case of $\gamma < \alpha$ the embedding property (EMB) $\|[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)\| =: A$ together with an analogous argument yields $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ for $m \leq U_2 \varepsilon^{-\frac{p}{\alpha-\gamma}}$ with $U_2 := \frac{1}{2} c^p \left(\frac{B_\infty}{32A^2} \right)^{\frac{p}{\alpha-\gamma}}$. So for $U := \min\{U_1, U_2\}$ and $\varepsilon_1 := \min\{1, U^{1/u}\}$ we get the assertion. In the case of $\gamma \geq \alpha$ (EVD) implies

$$\|f_\omega\|_{L_\infty(\nu)}^2 \leq A^2 \|f_\omega\|_\alpha^2 \leq \frac{32\varepsilon}{m} A^2 \sum_{i=1}^m \mu_{i+m}^{\gamma-\alpha} \leq 32A^2 \varepsilon \mu_m^{\gamma-\alpha} \leq 32A^2 C^{\gamma-\alpha} \varepsilon m^{-\frac{\gamma-\alpha}{p}} \quad (27)$$

and we get $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ for $0 < \varepsilon \leq \frac{B_\infty}{32A^2 C^{\gamma-\alpha}}$. Since $\gamma \geq \alpha$ implies $\beta > \alpha$ the assertion follows for $U := U_1$ and $\varepsilon_1 := \min\{\frac{B_\infty}{32A^2 C^{\gamma-\alpha}}, U_1^{1/u}\}$. \square

If $\omega' = (\omega'_1, \dots, \omega'_m) \in \{0, 1\}^m$ is another binary string, we investigate the norm of the difference $f_\omega - f_{\omega'}$. To this end, we set $v := \frac{2}{p}$ and use an analogue estimate as in (27), for $\alpha = 0$, which yields

$$\|f_\omega - f_{\omega'}\|_{L_2(\nu)}^2 \leq 32C^\gamma \varepsilon m^{-v} . \quad (28)$$

In order to obtain a lower bound on the γ -power norm, we assume $\sum_{i=1}^m (\omega_i - \omega'_i)^2 \geq \frac{m}{8}$, i.e. the distance between ω and ω' is *large*:

$$\|f_\omega - f_{\omega'}\|_\gamma^2 = \frac{32\varepsilon}{m} \sum_{i=1}^m (\omega_i - \omega'_i)^2 \geq 4\varepsilon . \quad (29)$$

The following lemma is from Tsybakov [36, Lemma 2.9] and suggests that there are many binary strings with large distances.

6.14 Lemma (Gilbert-Varshamov Bound) *For $m \geq 8$ and $M \geq 2^{m/8}$ there exist binary strings $\omega^{(0)}, \dots, \omega^{(M)} \in \{0, 1\}^m$ with $\omega^{(0)} = (0, \dots, 0)$ and*

$$\sum_{i=1}^m (\omega_i^{(j)} - \omega_i^{(k)})^2 \geq \frac{m}{8}$$

for all $j \neq k$, where $\omega^{(k)} = (\omega_1^{(k)}, \dots, \omega_m^{(k)})$.

Now, we are ready to prove Lemma 6.10.

Proof of Lemma 6.10. Using the notation from Lemma 6.13 we define $\varepsilon_0 := \min\{\varepsilon_1, (U/9)^{1/u}\}$ and $m_\varepsilon := \lfloor U\varepsilon^{-u} \rfloor$. Now, we fix a $n \geq 1$ and a $0 < \varepsilon \leq \varepsilon_0$. Since $m_\varepsilon \geq 9$, Lemma 6.14 yields for $M_\varepsilon := \lceil 2^{m_\varepsilon/8} \rceil \geq 3$ binary strings $\omega^{(0)}, \omega^{(1)}, \dots, \omega^{(M_\varepsilon)} \in \{0, 1\}^{m_\varepsilon}$ with *large distances*. If we define $f_j := f_{\omega^{(j)}}$ then $P_j := P_{f_j}$ for $j = 0, 1, \dots, M_\varepsilon$ satisfy the assumptions of Lemma 6.10 according to Lemma 6.13. It remains to prove the Statements (i)–(iii). Due to the definition of M_ε , m_ε and $m_\varepsilon \geq 9$ we get $\frac{8}{9}U\varepsilon^{-u} \leq m_\varepsilon \leq U\varepsilon^{-u}$ and $2^{\frac{U}{9}\varepsilon^{-u}} \leq 2^{m_\varepsilon/8} \leq M_\varepsilon \leq 2^{m_\varepsilon/4} \leq 2^{\frac{U}{3}\varepsilon^{-u}}$ and (i) is satisfied for $C_2 := \frac{U}{9}$. Statement (ii) is a consequence of the large distance between the binary strings and (29). Lemma 6.12, (28) and $m_\varepsilon \geq \frac{8}{9}U\varepsilon^{-u}$ yield

$$\frac{1}{M_\varepsilon} \sum_{j=1}^{M_\varepsilon} K(P_{f_j}^n, P_{f_0}^n) = \frac{n}{2\bar{\sigma}^2 M_\varepsilon} \sum_{j=1}^{M_\varepsilon} \|f_j - f_0\|_{L_2(\nu)}^2 \leq \frac{16C^\gamma}{\bar{\sigma}^2} n\varepsilon m_\varepsilon^{-v} = C_3 n\varepsilon^{1+uv}$$

where $C_3 := \frac{16C^\gamma 9^v}{\bar{\sigma}^2 (8U)^v}$. Combining Theorem 6.11 and (i) yields (iii) for $C_1 := \frac{3C_3}{C_2 \log(2)}$. \square

Now, the proof of Theorem 3.2 remains an application of Lemma 6.10 and the general reduction scheme from Tsybakov [36, Section 2.2].

Proof of Theorem 3.2. Let $D \mapsto f_{D,\lambda}$ be a (measurable) learning method. Furthermore, we use the notation of Lemma 6.10, set $r := \frac{\max\{\alpha, \beta\} - \gamma}{\max\{\alpha, \beta\} + p}$, and fix $\tau > 0$ and $n \geq 1$ with $\varepsilon_n := \tau \left(\frac{1}{n}\right)^r \leq \varepsilon_0$. It remains to show that there is a distribution P which is difficult to learn for the considered learning method. Lemma 6.10, for $\varepsilon = \varepsilon_n$, provides possible candidates P_0, P_1, \dots, P_{M_n} each satisfying the requirements of Theorem 3.2. Next, we estimate the left hand side of the inequality (iii) of Lemma 6.10. To this end, we define the measurable function $\Psi : (X \times Y)^n \rightarrow \{0, 1, \dots, M_n\}$,

$$\Psi(D) := \operatorname{argmin}_{j=0,1,\dots,M_n} \|[f_D]_\nu - f_j\|_\gamma .$$

For $j \in \{0, 1, \dots, M_n\}$ and $D \in (X \times Y)^n$ with $\Psi(D) \neq j$ we have

$$2\sqrt{\varepsilon_n} \leq \|f_{P_{\Psi(D)}}^* - f_{P_j}^*\|_\gamma \leq \|f_{P_{\Psi(D)}}^* - [f_D]_\nu\|_\gamma + \|[f_D]_\nu - f_{P_j}^*\|_\gamma \leq 2\|[f_D]_\nu - f_{P_j}^*\|_\gamma$$

and hence $P_j^n(D : \Psi(D) \neq j) \leq P_j^n(D : \|[f_D]_\nu - f_{P_j}^*\|_\gamma^2 \geq \varepsilon_n)$. According to (iii) of Lemma 6.10 there is $P \in \{P_0, \dots, P_{M_n}\}$ with

$$P^n(D : \|[f_D]_\nu - f_{P_j}^*\|_\gamma^2 \geq \tau n^{-r}) \geq \frac{\sqrt{M_n}}{\sqrt{M_n} + 1} \left(1 - C_1 \tau^{1/r} - \frac{1}{2 \log(M_n)}\right) .$$

Since $M_n \rightarrow \infty$ for $n \rightarrow \infty$ we can choose n sufficient large such that the right hand side is bounded from below by $1 - 2C_1 \tau^{1/r}$. \square

A. Auxiliary Results and Concentration Inequalities

A.1 Lemma For $\lambda > 0$ and $0 \leq \alpha \leq 1$ we consider the function $f_{\lambda,\alpha} : [0, \infty) \rightarrow \mathbb{R}$, $f_{\lambda,\alpha}(t) := \frac{t^\alpha}{\lambda+t}$. In the case $\alpha = 0$ this function is strict monotonically decreasing and in the case of $\alpha = 1$ strict monotonically increasing. Furthermore, the supremum of $f_{\lambda,\alpha}$ satisfied

$$\frac{1}{2}\lambda^{\alpha-1} \leq \sup_{t \geq 0} f_{\lambda,\alpha}(t) \leq \lambda^{\alpha-1},$$

where we use $0^0 := 1$. In the case of $\alpha < 1$ the function $f_{\lambda,\alpha}$ attain its supremum at $t^* := \frac{\lambda\alpha}{1-\alpha}$.

Proof. This could be easily proved, using the derivative of $f_{\lambda,\alpha}$. □

The following Bernstein type inequality for Hilbert space valued random variables is due to [26]. However we use a version from [4, Proposition 2].

A.2 Theorem (Bernstein's Inequality) Let (Ω, \mathcal{B}, P) be a probability space, H be a separable Hilbert space and $\xi : \Omega \rightarrow H$ be a random variable with

$$\mathbb{E}_P \|\xi\|_H^m \leq \frac{1}{2} m! \sigma^2 L^{m-2}$$

for all $m \geq 2$. Then for all $\tau \geq 1$ and $n \geq 1$ we have

$$P^n \left((\omega_i)_{i=1}^n \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P \xi \right\|_H^2 \geq 32\tau^2 \left(\frac{\sigma^2}{n} + \frac{L^2}{n^2} \right) \right) \leq 2e^{-\tau}.$$

Proof. This is a direct consequence of [4, Proposition 2] with $\eta = 2e^{-\tau}$ and

$$\mathbb{E}_P \|\xi - \mathbb{E}_P \xi\|_H^m \leq 2^{m-1} (\mathbb{E}_P \|\xi\|_H^m + \|\mathbb{E}_P \xi\|_H^m) \leq 2^m \mathbb{E}_P \|\xi\|_H^m.$$

Note that we consider the squared norm and hence additionally apply $(a+b)^2 \leq 2(a^2 + b^2)$ for $a, b \geq 0$. □

The following Bernstein type inequality for Hilbert-Schmidt operator valued random variables is due to [20]. However we use a version from [16, Lemma 26], see also [35] for an introduction to this topic.

A.3 Theorem Let (Ω, \mathcal{B}, P) be a probability space, H be a separable Hilbert space and $\xi : \Omega \rightarrow \mathcal{L}_2(H)$ be a random variable with values in the set of self-adjoint Hilbert-Schmidt operators. Furthermore, we assume that the operator norm is P -a.s. bounded, i.e. $\|\xi\| \leq B$ P -a.s. and that there is a self-adjoint positive semi-definite trace class operator V with $\mathbb{E}_P(\xi^2) \preceq V$, i.e. $V - \mathbb{E}_P(\xi^2)$ is positive semi-definite. Then for all $\tau \geq 1$ and $n \geq 1$ we have for $g(V) := \log\left(\frac{2e \operatorname{tr}(V)}{\|V\|}\right)$

$$P^n \left((\omega_i)_{i=1}^n \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P \xi \right\| \geq \left(\frac{4\tau B g(V)}{3n} + \sqrt{\frac{2\tau \|V\| g(V)}{n}} \right) \right) \leq 2e^{-\tau}.$$

Proof. This is a direct consequence of [16, Lemma 26] with $\delta = 2e^{-\tau}$. Furthermore, we used $\|\xi - \mathbb{E}_P \xi\| \leq 2\|\xi\|$, $\mathbb{E}_P(\xi - \mathbb{E}_P \xi)^2 \preceq \mathbb{E}_P(\xi^2)$, and $\log(c/\delta) \leq \tau \log(ec/2)$ for $c \geq 2$. □

References

- [1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, Elsevier/Academic Press, Amsterdam, second edition, 2003.
- [2] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23:52–72, 2007.
- [3] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18:971–1013, 2017.
- [4] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7:331–368, 2007.
- [5] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*, Cambridge University Press, Cambridge, 1990.
- [6] E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5:59–85, 2005.
- [7] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.
- [8] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for Tikhonov regularization. *Anal. Appl. (Singap.)*, 4:81–99, 2006.
- [9] 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. Stat.*, 11:1022–1047, 2017.
- [10] J. Diestel and J. J. Uhl, Jr. *Vector Measures*, American Mathematical Society, Providence, R.I., 1977. With a foreword by B. J. Pettis, Mathematical Surveys, No. 15.
- [11] M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.*, 7:1–42, 2013.
- [12] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, Cambridge University Press, Cambridge, 1996.
- [13] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108:203–227, 2019.
- [14] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv e-prints*, 1702.07254v1, 2017.
- [15] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*, Springer, New York, 2002.
- [16] J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *arXiv e-prints*, 1801.07226v2, 2018.
- [17] J. Lin and V. Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning*, page 27, 2018.
- [18] J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Appl. Comput. Harmon. Anal.*, 2018.
- [19] S. Mendelson and J. Neeman. Regularization in kernel learning. *Ann. Statist.*, 38:526–565, 2010.

- [20] S. Minsker. On some extensions of Bernstein’s inequality for self-adjoint operators. *Statist. Probab. Lett.*, 127:111–119, 2017.
- [21] N. Mücke. Reducing training time by efficient localized kernel regression. *arXiv e-prints*, 1707.03220v3, 2017.
- [22] N. Mücke and G. Blanchard. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.*, 19:1–29, 2018.
- [23] N. Mücke, G. Neu, and L. Rosasco. Beating SGD saturation with tail-averaging and minibatching. *arXiv e-prints*, 1902.08668v1, 2019.
- [24] Page and Grünewälder. Ivanov-regularised least-squares estimators over large RKHSs and their interpolation spaces. *arXiv e-prints*, 1706.03678v3, 2017.
- [25] L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems 31*, pages 8114–8124. Curran Associates, Inc., 2018.
- [26] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory Probab. Appl.*, 30:143–148, 1986.
- [27] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems 28*, pages 1657–1665. Curran Associates, Inc., 2015.
- [28] S. Smale and D.-X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41:279–306, 2004.
- [29] S. Smale and D.-X. Zhou. Shannon sampling II: Connections to learning theory. *Appl. Comput. Harmon. Anal.*, 19:285–302, 2005.
- [30] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26:153–172, 2007.
- [31] I. Steinwart and A. Christmann. *Support Vector Machines*, Springer, New York, 2008.
- [32] I. Steinwart and C. Scovel. Mercer’s theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35:363–417, 2012.
- [33] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.
- [34] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland Publishing Co., Amsterdam-New York, 1978.
- [35] J. A. Tropp. An introduction to matrix concentration inequalities. *arXiv e-prints*, 1501.01571v1, 2015.
- [36] A. B. Tsybakov. *Introduction to Nonparametric Estimation*, Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.