Journal of Machine Learning Research 21 (2020) 1-38 Submitted 9/19; Revised 7/20; Published 10/20

Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms

Simon Fischer SIMON.FISCHER@MATHEMATIK.UNI-STUTTGART.DE
Ingo Steinwart INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE
Institute for Stochastics and Applications
Faculty 8: Mathematics and Physics
University of Stuttgart
70569 Stuttgart Germany

Editor: Lorenzo Rosasco

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 data set $D = \{(x_i, y_i)\}_{i=1}^n$ independently sampled from an unknown distribution P on $X \times Y$, the goal of non-parametric least-squares regression is to estimate the conditional mean function $f_P^* : X \to Y$ given by $f_P^*(x) \coloneqq \mathbb{E}(Y|X = x)$. The function f_P^* is also known as regression function, we refer to Györfi et al. (2002) for basic information as well as various algorithms for this problem. In this work, we focus on kernel-based regularized least-squares algorithms, which are also known as least-squares support vector machines (LS-SVMs), see e.g. Steinwart & Christmann (2008). Recall that LS-SVMs construct a predictor $f_{D,\lambda}$ by solving the convex optimization problem

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

where a reproducing kernel Hilbert space (RKHS) H over X is used as hypothesis space and $\lambda > 0$ is the so-called regularization parameter. For a definition and basic properties of RKHSs see e.g. Steinwart & Christmann (2008, 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

$$\|f_{D,\lambda} - f_P^*\| \ . \tag{2}$$

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

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., De Vito et al., 2005a,b, 2006; Bauer et al., 2007; Smale & Zhou, 2007; Caponnetto & De Vito, 2007; Blanchard & Mücke, 2017; Dicker et al., 2017; Lin et al., 2018; Lin & Cevher, 2018a, and references therein) and the *empirical process* technique (see e.g., Mendelson & Neeman, 2010; Steinwart & Christmann, 2008; Steinwart et al., 2009, 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., Blanchard & Mücke, 2017; Lin et al., 2018). In addition, it can be used to establish learning rates for *spectral regularization algorithms* (see e.g., Bauer et al., 2007; Blanchard & Mücke, 2017; Lin et al., 2018) and further kernel-based learning algorithms (see e.g., Mücke, 2019; Lin & Cevher, 2018b; Pillaud-Vivien et al., 2018; Mücke & Blanchard, 2018; Mücke et al., 2019). 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 Steinwart et al. (2009). This advantage of the empirical process technique in the hard learning scenario is based on the additional consideration of some *embedding property* of the RKHS, which has hardly been considered in combination with the integral operator technique so far. In a nutshell, this embedding property allows for an improved bound on the L_∞-norm of the regularized population predictor. In addition, the empirical process technique 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. Farooq & Steinwart (2018) for expectile regression and Eberts & Steinwart (2013) for quantile regression.

In the present manuscript, which is an improvement of its first version Fischer & Steinwart (2017), we apply the integral operator technique in combination with some embedding property, see (EMB) in Section 3 below for details, to learning scenarios including the case $f_P^* \notin H$. Recall that such embedding properties—as far as we know—have only been used by Steinwart et al. (2009), Dicker et al. (2017), and Pillaud-Vivien et al. (2018). By doing so, we extend and improve the results of Blanchard & Mücke (2017) and Lin et al. (2018). To be more precise, we extend the results of Blanchard & Mücke (2017), who only considered the case $f_P^* \in H$, to the hard learning case and the largest possible scale of γ. Moreover, compared to Lin et al. (2018) we obtain faster rates of convergence for (2), if the RKHS enjoys a certain embedding property. In the hard learning scenario, we obtain, as a byproduct, the L_2-learning rates of Steinwart et al. (2009), as well as the very first L_∞-norm learning rates in the hard learning scenario. For a more detailed comparison with the literature see Section 5 and in particular Table 1 and Figure 1. Finally, we 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. The consequences of our results for the special case of a Sobolev/Besov RKHS H 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.

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 $X \times \mathbb{R}$ with

$$|P|_2^2 := \int_{X\times\mathbb{R}} y^2 \, \mathrm{d}P(x,y) < \infty \ . \tag{3}$$

Moreover, we denote the marginal distribution of P on X by $\nu := P_X$. 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

$$f_P^*(x) = \int_{\mathbb{R}} y \ P(\mathrm{d}y|x) \ . \tag{4}$$

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 and bounded kernel k. Let us recall some facts about the interplay between H and $L_2(\nu)$. Some of the following results have already be shown by Smale & Zhou (2004, 2005) and De Vito et al. (2006, 2005b), but we follow the more recent contribution of Steinwart & Scovel (2012) because of its more general applicability. According to Steinwart & Scovel (2012, Lemma 2.2, Lemma 2.3) and Steinwart & Christmann (2008, Theorem 4.27) the—not necessarily injective—embedding $I_\nu : H \to L_2(\nu)$, mapping a function $f \in H$ to its ν-equivalence class $[f]_\nu$, is well-defined, Hilbert-Schmidt, and the Hilbert-Schmidt norm satisfies

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

Moreover, the adjoint operator $S_\nu := I_\nu^* : L_2(\nu) \to H$ is an integral operator with respect to the kernel k, i.e. for $f \in L_2(\nu)$ and $x \in X$ we have

$$(S_\nu f)(x) = \int_X k(x,x') f(x') \, \mathrm{d}\nu(x') \ . \tag{5}$$

Next, we define the self-adjoint and positive semi-definite integral operators

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

These operators are trace class and their trace norms satisfy

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

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 I, a non-increasing summable sequence $(\mu_i)_{i\in I} \subseteq (0,\infty)$, and a family $(e_i)_{i\in I} \subseteq H$, such that $([e_i]_\nu)_{i\in I}$ is an orthonormal basis (ONB) of $\overline{\operatorname{ran} I_\nu} \subseteq L_2(\nu)$ and $(\mu_i^{1/2} e_i)_{i\in I}$ is an ONB of $(\ker I_\nu)^\perp \subseteq H$ with

$$T_\nu = \sum_{i\in I} \mu_i \langle \cdot, [e_i]_\nu \rangle_{L_2(\nu)} [e_i]_\nu \qquad \text{and} \qquad C_\nu = \sum_{i\in I} \mu_i \langle \cdot, \mu_i^{1/2} e_i \rangle_H \, \mu_i^{1/2} e_i \ , \tag{6}$$

see Steinwart & Scovel (2012, Lemma 2.12) for details. Since we are mainly interested in the hard learning scenario $f_P^* \notin H$ we exclude finite I and assume $I = \mathbb{N}$ in the following.

Let us recall some intermediate spaces introduced by Steinwart & Scovel (2012, Equation 36). 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*

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

for $(a_i)_{i\geq 1} \in \ell_2(\mathbb{N})$, it becomes a Hilbert space. Moreover, $(\mu_i^{\alpha/2}[e_i]_\nu)_{i\geq 1}$ forms an ONB of $[H]_\nu^\alpha$ and consequently $[H]_\nu^\alpha$ is a separable Hilbert space. If there is no danger of confusion we use the abbreviation $\|\cdot\|_\alpha := \|\cdot\|_{[H]_\nu^\alpha}$. Furthermore, in the case of $\alpha = 1$ we introduce the notation $[H]_\nu := [H]_\nu^1$. Recall that for $\alpha = 0$ we have $[H]_\nu^0 = \overline{\operatorname{ran} I_\nu} \subseteq L_2(\nu)$ with $\|\cdot\|_0 = \|\cdot\|_{L_2(\nu)}$. Moreover, for $\alpha = 1$ we have $[H]_\nu^1 = \operatorname{ran} I_\nu$ and $[H]_\nu^1$ is isometrically isomorphic 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

$$[H]_\nu^\alpha \hookrightarrow [H]_\nu^\beta \hookrightarrow [H]_\nu^0 = \overline{\operatorname{ran} I_\nu} \subseteq L_2(\nu) \tag{7}$$

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

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

for $f \in \overline{\operatorname{ran} I_\nu}$. In addition, for $0 < \alpha < 1$, the α-power space is characterized in terms of interpolation spaces of the real method, see e.g. Triebel (1978, Section 1.3.2) for a definition. To be more precise, Steinwart & Scovel (2012, Theorem 4.6) proved

$$[H]_\nu^\alpha \cong \left[L_2(\nu), [H]_\nu \right]_{\alpha,2} \ , \tag{8}$$

where the symbol $\cong$ in (8) means that these spaces are isomorphic, i.e. the sets coincide and the corresponding norms are equivalent. Note that for Sobolev/Besov RKHSs and marginal distributions that are essentially the uniform distribution, the interpolation space $\left[L_2(\nu), [H]_\nu \right]_{\alpha,2}$ 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* satisfies 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

$$\mu_i \leq C\, i^{-1/p} \tag{EVD}$$

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

$$c\, i^{-1/p} \leq \mu_i \leq C\, i^{-1/p} \tag{EVD+}$$

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

$$\big\| [H]^\alpha_\nu \hookrightarrow L_\infty(\nu) \big\| \leq A \ . \tag{EMB}$$

This mean $[H]^\alpha_\nu$ is continuously embedded into $L_\infty(\nu)$ and the operator norm of the embedding is bounded by A. Because of (7) the larger α is, the weaker the embedding property is. Since our kernel k is bounded, (EMB) is always satisfied for $\alpha = 1$. Moreover, Part (iii) of Lemma 10 in Section 6 shows that (EMB) implies a polynomial eigenvalue decay of order $1/\alpha$ and hence we assume $p \leq \alpha$ in the following. Observe that the converse does not hold in general and consequently it is possible that we even have the strict inequality $p < \alpha$.

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: There is a constant $B > 0$ such that $f^*_P \in [H]^\beta_\nu$ and

$$\|f^*_P\|_\beta \leq B \ . \tag{SRC}$$

Note that $|P|_2 < \infty$, defined in (3), 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. Nonetheless, we included the case $1 \leq \beta \leq 2$ because it is no extra effort in the proof. Since we want to estimate $\|[f_{D,\lambda}]_\nu - f^*_P\|_\gamma$ and this expression is well-defined if and only if $f^*_P \in [H]^\gamma_\nu$, 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

$$\int_{\mathbb{R}} |y - f^*_P(x)|^m \ P(\mathrm{d}y|x) \leq \frac{1}{2} m!\, \sigma^2\, L^{m-2} \tag{MOM}$$

is satisfied for ν-almost all $x \in X$ and all $m \geq 2$. Note that (MOM) is satisfied for Gaussian noise with bounded variance, i.e. $P(\,\cdot\,|x) = \mathcal{N}(f^*_P(x), \sigma^2_x)$, where $x \mapsto \sigma_x \in (0, \infty)$ is a measurable and ν-almost surely 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-recognized in the statistical analysis of regularized least-squares algorithms (see e.g., Caponnetto & De Vito, 2007; Blanchard &

Mücke, 2017; Lin & Cevher, 2018a; Lin et al., 2018). 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 does *not* imply additional constraints, such as boundedness, on f_P^*. An embedding property slightly weaker than (EMB) was used by Steinwart et al. (2009) in combination with empirical process arguments. Dicker et al. (2017) used (EMB) to investigate benign scenarios with exponentially decreasing eigenvalues and $f_P^* \in H$, and Pillaud-Vivien et al. (2018) used (EMB) 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 enable us to prove the following result.

Theorem 1 (γ-Learning Rates) *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X with respect to a bounded and measurable kernel k, P be a probability distribution on $X \times \mathbb{R}$ with $|P|_2 < \infty$, and $\nu := P_X$ be the marginal distribution on X. Furthermore, let $B_\infty > 0$ be a constant with $\|f_P^*\|_{L_\infty(\nu)} \leq B_\infty$ and the Conditions* (EMB), (EVD) (SRC), *and* (MOM) *be 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 $\lambda_n \asymp (n/\log^r(n))^{-1/\alpha}$ for some $r > 1$ there is a constant $K > 0$ independent of $n \geq 1$ and $\tau \geq 1$ such that

$$\left\|[f_{D,\lambda_n}]_\nu - f_P^*\right\|_\gamma^2 \leq \tau^2 K \left(\frac{\log^r(n)}{n}\right)^{\frac{\beta-\gamma}{\alpha}} \tag{9}$$

is satisfied for sufficiently large $n \geq 1$ with P^n-probability not less than $1 - 4e^{-\tau}$.

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

$$\left\|[f_{D,\lambda_n}]_\nu - f_P^*\right\|_\gamma^2 \leq \tau^2 K \left(\frac{1}{n}\right)^{\frac{\beta-\gamma}{\beta+p}} \tag{10}$$

is satisfied for sufficiently large $n \geq 1$ with P^n-probability not less than $1 - 4e^{-\tau}$.

Theorem 1 is mainly based on a finite sample bound given in Section 6, see Theorem 16. We think that the statement of Theorem 1 can be proved for general regularization methods if one combines our technique, especially Lemma 17 and Lemma 18 from Section 6.2, with the results of Lin et al. (2018) and Lin & Cevher (2018a). However, we stick to the learning scheme (1) for simplicity. The proof of Theorem 1 reveals that the constants $K > 0$ just depend on the parameters and constants from (EMB), (EVD), (SRC), and (MOM), on the considered norm, i.e. on γ, on B_∞, and on the regularization parameter sequence $(\lambda_n)_{n\geq 1}$. Moreover, the index bound hidden in the phrase *for sufficient large $n \geq 1$* just depends on the parameters and constants from (EMB) and (EVD), on τ, on a lower bound $0 < c \leq 1$ for the operator norm $c \leq \|C_\nu\|$, and on the regularization parameter sequence $(\lambda_n)_{n\geq 1}$. The asymptotic behavior in n of the right hand side in (9) and (10), respectively, is called *learning rate* with respect to the γ-power norm or abbreviated γ-learning rate. Recall, for $\gamma = 0$, the norms on left hand sides of (9) and (10) coincide with the $L_2(\nu)$-norm.

Note that, for $\beta \geq \alpha$, the conditional mean function f_P^* is automatically ν-almost surely bounded, since we have $f_P^* \in [H]_\nu^\beta \hookrightarrow [H]_\nu^\alpha \hookrightarrow L_\infty(\nu)$, and in this case always Situation (10) applies. Moreover, in the case of $\alpha = p$, which was also considered by Steinwart et al. (2009, Corollary 6), we are always in Situation (10), too.

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

$$n^{-\frac{\beta-\gamma}{\max\{\beta+p,\alpha\}}} \quad .$$

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 have to assume (EVD+) to make sure that the eigenvalues do not decay faster than (EVD) guarantees.

Theorem 2 (γ-Lower Rates) *Let $(X, \mathcal{B})$ be a measurable space, 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 $K_0, K, r > 0$ such that for all learning methods $D \mapsto f_D$, all $\tau > 0$, and all sufficiently large $n \geq 1$ there is a distribution P on $X \times \mathbb{R}$ with $P_X = \nu$ satisfying $\|f_P^*\|_{L_\infty(\nu)} \leq B_\infty$,* (SRC) *with respect to β, B,* (MOM) *with respect to σ, L, and with P^n-probability not less than $1 - K_0\tau^{1/r}$*

$$\left\|[f_D]_\nu - f_P^*\right\|_\gamma^2 \geq \tau^2 K \left(\frac{1}{n}\right)^{\frac{\max\{\alpha,\beta\}-\gamma}{\max\{\alpha,\beta\}+p}} \quad . \tag{11}$$

In short, Theorem 2 states that there is no learning method satisfying a faster decaying γ-learning rate than

$$n^{-\frac{\max\{\alpha,\beta\}-\gamma}{\max\{\alpha,\beta\}+p}}$$

under the assumptions of Theorem 1 and (EVD+). The asymptotic behavior in n of the right hand side in (11) is called *(minimax) lower rate* with respect to the γ-power norm or abbreviated γ-lower rate. Theorem 2 extends the lower bounds previously obtained by Caponnetto & De Vito (2007), Steinwart et al. (2009), and Blanchard & Mücke (2017). To be more precise, Caponnetto & De Vito (2007, Theorem 2) considered only the case $f_P^* \in H$ and $\gamma = 0$, Steinwart et al. (2009, Theorem 9) considered only the case $\beta \geq \alpha$ and $\gamma = 0$, and Blanchard & Mücke (2017, Theorem 3.5) restricted their considerations to $f_P^* \in H$. In the case of $\alpha \leq \beta$, which implies the boundedness of f_P^*, the γ-learning rate of LS-SVMs stated in Theorem 1 coincides with the γ-lower rate from Theorem 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.

Remark 3 (Optimality and Boundedness) *Under the assumptions of Theorem 2,* but without *requiring the uniform boundedness of f_P^* by some constant B_∞, we can* improve *the*

γ-lower rate of Theorem 2. More precisely, a straightforward modification of Lemma 23 in Section 6 gives in the case of not uniformly bounded f_P^* the γ-lower rate

$$n^{-\frac{\beta-\gamma}{\beta+p}} \,.$$

Moreover, if we would be able to prove the γ-learning rates of Theorem 1 with a constant $K > 0$ independent of $\|f_P^*\|_{L_\infty(\nu)}$ then we would have optimality for our γ-learning rates in the case of $\beta > \alpha - p$ instead of $\beta \geq \alpha$.

Because of (EMB), the next remark is a direct consequence of Theorem 1 for $\gamma = \alpha$.

Remark 4 (L_∞-Learning Rates) *Under the assumptions of Theorem 1 in the case of $\beta > \alpha$ the following statement is true. For all regularization parameter sequences $(\lambda_n)_{n\geq 1}$ with $\lambda_n \asymp n^{1/(\beta+p)}$ 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*

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

for sufficiently large $n \geq 1$ with P^n-probability not less than $1 - 4e^{-\tau}$.

Note that all previous efforts to get L_∞-learning rates for the learning scheme (1) need to assume $f_P^* \in H$. Consequently, Remark 4 establishes 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 assume that X is a benign domain: Let $X \subseteq \mathbb{R}^d$ be a non-empty, open, connected, and bounded set with a

$$C_\infty\text{-boundary} \tag{DOM}$$

and be equipped with the Lebesgue-Borel σ-algebra $\mathcal{B}$. Furthermore, $L_2(X) := L_2(\mu)$ denotes the corresponding L_2-space.

Let us briefly introduce Sobolev and Besov Hilbert spaces. For a more detailed introduction see e.g. Adams & Fournier (2003). For $m \in \mathbb{N}$ we denote the *Sobolev space* of smoothness m by $W_m(X) := W_{m,2}(X)$, see e.g. Adams & Fournier (2003, 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 $B_{2,2}^r(X) := \big[L_2(X), W_m(X)\big]_{r/m,2}$, where $m := \min\{k \in \mathbb{N} : \ k > r\}$ see e.g. Adams & Fournier (2003, Section 7.30) for details. 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

$$B_{2,2}^r(X) \cong \big[L_2(X), B_{2,2}^t(X)\big]_{r/t,2} \tag{12}$$

for all $t > r > 0$, see e.g. Adams & Fournier (2003, Section 7.32) for details. Moreover, an extension of the Sobolev embedding theorem to Besov spaces guarantees that, for $r > d/2$,

each μ-equivalence class in $B^r_{2,2}(X)$ has a unique continuous and bounded representative, see e.g. Adams & Fournier (2003, Part c of Theorem 7.24). 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, for $r > j + d/2$,

$$B^r_{2,2}(X) \hookrightarrow C_j(X) \ . \tag{13}$$

Consequently, we define, for $r > d/2$, the *Besov RKHS* as the set of continuous representatives $H_r(X) \coloneqq \{f \in C_0(X) : \ [f]_\mu \in B^r_{2,2}(X)\}$ and equip this space with the norm $\|f\|_{H_r(X)} \coloneqq \|[f]_\mu\|_{B^r_{2,2}(X)}$. The Besov RKHS $H_r(X)$ is a separable RKHS with respect to a kernel k_r. Moreover, k_r is bounded and measurable, see e.g. Steinwart & Christmann (2008, Lemma 4.28 and Lemma 4.25).

In the following, we fix a Besov RKHS $H_r(X)$ for some $r > d/2$ and a probability measure P on $X \times \mathbb{R}$ such that the marginal distribution $\nu = P_X$ on X satisfies the following condition: The probability measure ν is equivalent to the Lebesgue measure μ on X, i.e. $\mu \ll \nu$, $\nu \ll \mu$, and there are constants $g, G > 0$ such that

$$g \leq \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \leq G \tag{LEB}$$

is μ-almost surely satisfied. For marginal distributions ν satisfying (LEB) we have $L_2(\nu) \cong L_2(X)$ and we can describe the power spaces of $H_r(X)$ according to (8), the interpolation property, and (12) by

$$[H_r(X)]^{u/r}_\nu \cong \big[L_2(\nu), [H_r(X)]_\nu\big]_{u/r,2} \cong \big[L_2(X), [H_r(X)]_\mu\big]_{u/r,2} \cong B^u_{2,2}(X) \tag{14}$$

for $0 < u < r$. As a consequence of (14), we have $f^*_P \in B^s_{2,2}(X)$ for some $0 < s < r$ if and only if (SRC) is satisfied for $\beta = s/r$. Next, if we combine (14) and (13) then we get (EMB) for all α with $\frac{d}{2r} < \alpha < 1$:

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

Finally, we consider the asymptotic behavior of the eigenvalues $(\mu_i)_{i \geq 1}$ of the integral operator T_ν. Carl & Stephani (1990, Equation 4.4.12) show that the eigenvalue μ_i of T_ν equals the squares of the approximation number $a_i^2(I_\nu)$ of the embedding $I_\nu : H_r(X) \to L_2(\nu)$. Since $L_2(\nu) \cong L_2(X)$ these approximation numbers are described by Edmunds & Triebel (1996, Equation 4 on p. 119), namely

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

To sum up, the eigenvalues satisfy (EVD+) for $p = \frac{d}{2r}$. The following corollaries are direct consequences of Part (ii) of Theorem 1 and Theorem 2 with $p = \frac{d}{2r}$, $\beta = s/r$, $\gamma = t/r$, and an $\alpha > p$ that is chosen sufficiently close to p.

Corollary 5 (Besov-Learning Rates) *Let $X \subseteq \mathbb{R}^d$ be a set satisfying* (DOM)*, $H_r(X)$ be a Besov RKHS on X with $r > d/2$, P be a probability distribution on $X \times \mathbb{R}$ with $|P|_2 < \infty$, and $\nu \coloneqq P_X$ be the marginal distribution on X such that* (LEB) *is satisfied. Furthermore,*

let $B, B_\infty > 0$ be constants with $\|f_P^*\|_{L_\infty(\mu)} \leq B_\infty$ and $\|f_P^*\|_{B^s_{2,2}(X)} \leq B$ for some $0 < s < r$, and the Condition (MOM) *be satisfied. Then, for $0 \leq t < s$ and a regularization parameter sequence $(\lambda_n)_{n\geq 1}$ with $\lambda_n \asymp n^{-r/(s+d/2)}$, 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*

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

for sufficiently large $n \geq 1$ with P^n-probability not less than $1 - 4e^{-\tau}$.

Note that the $B^t_{2,2}$-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 $r > s$, i.e. to pick a sufficiently small $H_r(X)$. Recall that the case $t = 0$ corresponds to L_2-norm learning rates.

Corollary 6 (Besov-Lower Rates) *Let $X \subseteq \mathbb{R}^d$ be a set satisfying* (DOM)*, $H_r(X)$ be a Besov RKHS on X with $r > d/2$, and ν be a probability distribution on X satisfying* (LEB)*. Then, for all parameters $0 \leq t < s < r$ with $s > d/2$ and all constants $\sigma, L, B, B_\infty > 0$, there exist $K_0, K, r > 0$ such that for all learning methods $D \mapsto f_D$, all $\tau > 0$, and all sufficiently large $n \geq 1$ there is a distribution P on $X \times \mathbb{R}$ with $P_X = \nu$ satisfying $\|f_P^*\|_{L_\infty(\nu)} \leq B_\infty$, $\|f_P^*\|_{B^t_{2,2}(X)} \leq B$,* (MOM) *with respect to σ, L, and with P^n-probability not less than $1 - K_0\tau^{1/r}$*

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

In short, Corollary 6 states that the rates from Corollary 5 are optimal for $s > d/2$.

Remark 7 *Under the assumptions of Corollary 6 in the case of $s \leq d/2$ for all sufficiently small $\varepsilon > 0$ the following lower bound is satisfied*

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

Finally, if we have $s > j + d/2$, for some integer $j \geq 0$, then the combination of Corollary 5 and (13) yields $C_j(X)$-norm learning rates. To this end, we denote by f_P^* the unique continuous representative of the ν-equivalence class f_P^* and apply Corollary 5 with a sufficiently small $t > j + d/2$.

Remark 8 ($C_j(X)$-Learning Rates) *Under the assumption of Corollary 5 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 $\lambda_n \asymp n^{-r/(s+d/2)}$ 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*

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

for sufficiently large $n \geq 1$ with P^n-probability not less than $1 - 4e^{-\tau}$.

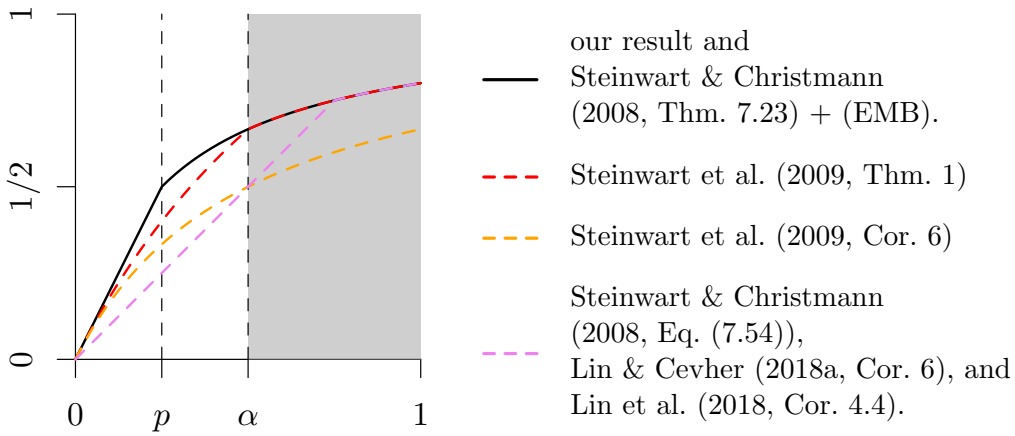

Figure 1: Plot of the exponent r of the L_2-learning rate n^{-r} over the smoothness β of f_P^* for a fixed RKHS H and a fixed marginal distribution $\nu = P_X$ which satisfy (EMB) and (EVD) with respect to $\alpha = 1/2$ and $p = 1/4$, respectively. Consequently, higher values correspond to faster learning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

Remark 8 suggests that $D \mapsto \partial^\alpha f_{D,\lambda}$, for some multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, is a reasonable estimator for the α-th derivative of the regression function $\partial^\alpha f_P^*$ if $f_P^* \in B_{2,2}^s(X)$ with some $s > |\alpha| + d/2 = \alpha_1 + \ldots + \alpha_d + d/2$. Note that the $\varepsilon > 0$ appears in the rates of Remark 7 and Remark 8 because we have to choose $\alpha > p$ and $t > j + d/2$, respectively.

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 obtained by many authors, see e.g. Caponnetto & De Vito (2007); Lin & Cevher (2018a) for L_2-rates and Blanchard & Mücke (2017); Lin et al. (2018) 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 Steinwart & Christmann (2008); Steinwart et al. (2009), which use empirical process techniques and Lin & Cevher (2018a); Lin et al. (2018), 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, the article of Lin et al. (2018) is the only one 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* (EMB) $[H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$	*Assumptions* (EVD) $\mu_i \preccurlyeq i^{-\frac{1}{p}}$	*Exponent r of the Learning Rate n^{-r}* $L_2(\nu)$	*Exponent r of the Learning Rate n^{-r}* $[H]^\gamma_\nu$ 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 & Christmann (2008, Thm. 7.23) + (EMB)				
Steinwart et al. (2009, Thm. 1)	$0 < \alpha \leq 1$	$0 < p \leq \alpha$	$\frac{\beta}{\max\{\beta+p,\beta+\alpha(1-\beta)\}}$	
Steinwart et al. (2009, Cor. 6)	$0 < \alpha \leq 1$	$p = \alpha$	$\frac{\beta}{\beta+\alpha}$	x
Steinwart & Christmann (2008, Eq. (7.54))	$\alpha = 1$	$0 < p \leq 1$	$\frac{\beta}{\max\{\beta+p,1\}}$	
Lin & Cevher (2018a, Cor. 6)				
Lin et al. (2018, Cor. 4.4)				$\frac{\beta-\gamma}{\max\{\beta+p,1\}}$

Table 1: Learning rates established by different authors for $f^*_P \in [H]^\beta_\nu$ 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. The article of Lin & Cevher (2018a) is an extended version of the conference paper Lin & Cevher (2018b). Lin & Cevher (2018a) investigate distributed gradient decent methods and spectral regularization algorithms. In Corollary 6 they provide the L_2-learning rate $n^{-\beta/\max\{\beta+p,1\}}$ in expectation for spectral regularization algorithms, containing the learning scheme (1) as special case. Lin et al. (2018) establish the γ-learning rate $n^{-(\beta-\gamma)/\max\{\beta+p,1\}}$ in probability for spectral regularization algorithms under more general source conditions, see Lin et al. (2018, Equation 18) for a definition. Both articles do not take any embedding property into account and hence we get at least the same rates and in case of (EMB) with $\alpha < 1$ we actually improve their rates iff $\beta + p < 1$. Let us illustrate this improvement in the case of a Besov RKHS $H_r(X)$ with smoothness r. To this end, we assume $f^*_P \in B^s_{2,2}(X)$ for some $s > 0$. Besides the condition $r > d/2$, which ensures that $H_r(X)$ is a RKHS, the only requirement for our Corollary 5 is $r > s$ in order to achieve the fastest known L_2-learning rate $n^{-s/(s+d/2)}$. Recall that this rate is independent of the smoothness r of the hypothesis space and is known to be optimal for $s > d/2$, see e.g. Corollary 6. In order to get the same L_2-learning rate by the results of Lin & Cevher (2018a) or Lin et al. (2018) the *additional* constraint $r \leq s + d/2$ has to be satisfied. Otherwise, they 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 Lin et al. (2018) in a similar way, i.e. to get our Besov-learning rates with the help of the results of Lin et al. (2018) the *additional* constraint $r \leq s + d/2$ has to be satisfied.

Empirical process techniques. Steinwart & Christmann (2008) provide an oracle inequality in Theorem 7.23 under a slightly weaker assumption than (EVD). As already mentioned there (Steinwart & Christmann, 2008, Equation 7.54), this oracle inequality leads, under a slightly weaker assumption than (SRC), to the L_2-rate $n^{-\beta/\max\{\beta+p,1\}}$. This rate coincides with the results of Lin & Cevher (2018a) and Lin et al. (2018), and is even better by a logarithmic factor. Inspired by Mendelson & Neeman (2010, Lemma 5.1), Steinwart et al. (2009) were the first using an embedding property, slightly weaker than (EMB), to derive finite sample bounds, see Steinwart et al. (2009, Theorem 1). Moreover, Theorem 1 of Steinwart et al. (2009) was used in Corollary 6 of that article to establish, in the case of $p = \alpha$, the L_2-rate $n^{-\beta/(\beta+\alpha)}$. But the proof remains valid in the general case $p \leq \alpha$ and hence Steinwart et al. (2009, Theorem 1) get the L_2-rate $n^{-\beta/\max\{\beta+p,\beta+\alpha(1-\beta)\}}$. This rate is never better than ours and is worse than ours iff $\alpha < 1$ and $\beta < 1 - p/\alpha$. If we combine the oracle inequality of Steinwart & Christmann (2008, Theorem 7.23) with (EMB) then we recover our L_2-rate from Theorem 1 even without logarithmic factor. However, recall that the empirical process technique is not able to provide general γ-learning rates yet. Finally, it is to mention that both contributions, Steinwart & Christmann (2008) and Steinwart et al. (2009), consider the *clipped* predictor. The influence of this clipping is not clear, but it could be the reason for avoiding the logarithmic factors appearing in some learning rates obtained by 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 Lin et al. (2018) for the learning scheme (1) whenever (EMB) is satisfied for some $0 < \alpha < 1$ as well as (SRC) and (EVD) are satisfied for $\beta + p < 1$. Finally, we show that our γ-learning rates are optimal in all cases in which 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, Steinwart & Scovel (2012, Corollary 3.2) show that there exists a ν-zero set $N \subseteq X$, such that k is given by

$$k(x, x') = \sum_{i \geq 1} \mu_i \, e_i(x) e_i(x') \tag{15}$$

for all $x, x' \in X \backslash N$. Furthermore, the boundedness of k implies $\sum_{i \geq 1} \mu_i e_i^2(x) \leq A^2$ for ν-almost all $x \in X$ 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

$$\sum_{i \geq 1} \mu_i^\alpha e_i^2(x) \leq A^2 \tag{16}$$

for ν-almost all $x \in X$. 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 Steinwart & Scovel (2012, Proposition 4.2) for the logic behind this notation. Because of the representation in

(15) and the boundedness of k we always have $\|k_\nu^1\|_\infty < \infty$. The following theorem allows an alternative characterization of (EMB).

Theorem 9 (L_∞-Embedding) *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, and ν be a probability distribution on X. Then the following equality is satisfied, for $\alpha > 0$,*

$$\big\|[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)\big\| = \|k_\nu^\alpha\|_\infty \ . \tag{17}$$

Note that with the help of (17) the Condition (EMB) can be written as $\|k_\nu^\alpha\|_\infty \leq A$. The statement of Theorem 9 is part of Steinwart & Scovel (2012, Theorem 5.3), but we give an alternative proof below, which is more basic and does *not* require the ν-completeness of the measurable space $(X, \mathcal{B})$. Moreover, our proof of Theorem 9 remains true in the situation considered by Steinwart & Scovel (2012, Theorem 5.3), i.e. for σ-finite measures ν and (possibly unbounded) kernels k whose RKHS H is compactly embedded into $L_2(\nu)$. In this respect, we generalize Theorem 5.3 of Steinwart & Scovel (2012). We restricted our consideration to bounded kernels and probability measures only for convenience since we do not need this generality in the rest of this work.

Proof First we prove '$\geq$'. To this end, we assume that $\mathrm{Id} : [H]_\nu^\alpha \to L_\infty(\nu)$ exists and is bounded and hence $\|\operatorname{Id}\| < \infty$. Since $(\mu_i^{\alpha/2}[e_i]_\nu)_{i\in I}$ is an ONB of $[H]_\nu^\alpha$ for every sequence $a = (a_i)_{i\geq 1} \in \ell_2(\mathbb{N})$ the series $\sum_{i\geq 1} a_i \mu_i^{\alpha/2}[e_i]_\nu$ converges in $[H]_\nu^\alpha$ and hence it also converges in $L_\infty(\nu)$. As a result, there is a representative $f_a : X \to \mathbb{R}$ with $[f_a]_\nu = \sum_{i\geq 1} a_i \mu_i^{\alpha/2}[e_i]_\nu \in [H]_\nu^\alpha$ and a set $N_a \subseteq X$ with $\nu(N_a) = 0$ such that

$$f_a(x) = \sum_{i\geq 1} a_i \mu_i^{\alpha/2} e_i(x) \qquad \text{and} \qquad |f_a(x)| \leq \Big\|\sum_{i\geq 1} a_i \mu_i^{\alpha/2}[e_i]_\nu\Big\|_{L_\infty(\nu)}$$

for all $x \in X\backslash N_a$. Consequently, for all $x \in X\backslash N_a$, we find

$$|f_a(x)| \leq \Big\|\sum_{i\geq 1} a_i \mu_i^{\alpha/2}[e_i]_\nu\Big\|_{L_\infty(\nu)} \leq \|\operatorname{Id}\| \cdot \Big\|\sum_{i\geq 1} a_i \mu_i^{\alpha/2}[e_i]_\nu\Big\|_{[H]_\nu^\alpha} = \|\operatorname{Id}\| \cdot \|a\|_{\ell_2(\mathbb{N})} \ .$$

Since the closed unit ball $\overline{B}_{\ell_2(\mathbb{N})}$ of $\ell_2(\mathbb{N})$ is separable there is a countable dense subset $B \subseteq \overline{B}_{\ell_2(\mathbb{N})}$. If we define the set $N := \bigcup_{a\in B} N_a \subseteq X$ then we have $\nu(N) = 0$ since B is countable. Moreover, the denseness of B in $\overline{B}_{\ell_2(\mathbb{N})}$ implies

$$\sum_{i\geq 1} \mu_i^\alpha e_i^2(x) = \|(\mu_i^{\alpha/2} e_i(x))\|_{\ell_2(\mathbb{N})}^2 = \sup_{a\in B} |\langle a, (\mu_i^{\alpha/2} e_i(x))_{i\geq 1}\rangle_{\ell_2(\mathbb{N})}|^2 = \sup_{a\in B} |f_a(x)|^2 \leq \|\operatorname{Id}\|^2$$

for all $x \in X\backslash N$ and hence $\|k_\nu^\alpha\|_\infty \leq \|\operatorname{Id}\|$.

Now we prove '$\leq$'. To this end, we assume $\|k_\nu^\alpha\|_\infty < \infty$ and choose some $[f]_\nu \in [H]_\nu^\alpha$ with $\|[f]_\nu\|_\alpha \leq 1$. Then there is a (unique) sequence $a = (a_i)_{i\geq 1} \in \ell_2(\mathbb{N})$ with $\|a\|_{\ell_2(\mathbb{N})} \leq 1$ and $[f]_\nu = \sum_{i\geq 1} a_i \mu_i^{\alpha/2}[e_i]_\nu$. Using Hölder's inequality we get

$$|f(x)| \leq \|a\|_{\ell_2(\mathbb{N})} \left(\sum_{i\geq 1} \mu_i^\alpha e_i^2(x)\right)^{1/2} \leq \|k_\nu^\alpha\|_\infty$$

for ν-almost all $x \in X$. Consequently, we have $\|[f]_\nu\|_{L_\infty(\nu)} \leq \|k_\nu^\alpha\|_\infty$ for all $[f]_\nu \in [H]_\nu^\alpha$ with $\|[f]_\nu\|_\alpha \leq 1$ and this proves $\|\operatorname{Id}\| \leq \|k_\nu^\alpha\|_\infty$. ∎

The following lemma summarizes further implications of (EMB).

Lemma 10 *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, and ν be a probability distribution on X. Then the following statements are true, for $0 < p, \alpha \leq 1$:*

(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 converse implication is true.*

(iii) (EMB) *implies* (EVD) *for* $p = \alpha$. *If, in addition, the eigenfunctions are uniformly bounded, then* (EVD) *w.r.t.* $0 < p < 1$ *implies* (EMB) *for all* $\alpha > p$.

Note that uniformly bounded eigenfunction have been considered e.g. by Mendelson & Neeman (2010, Assumption 4.1) and Steinwart et al. (2009, Theorem 2), see also the discussion after Theorem 5.3 of Steinwart & Scovel (2012).

Proof For the proof we silently use the Identity (17) in Theorem 9.

(i) Using (EMB) and the fact that $(\mu_i^{\alpha/2}[e_i]_\nu)_{i\geq 1}$ is an ONB of $[H]_\nu^\alpha$ yields the assertion

$$\|\mu_i^{\alpha/2}[e_i]_\nu\|_{L_\infty(\nu)} \leq \|k_\nu^\alpha\|_\infty \big\|\mu_i^{\alpha/2}[e_i]_\nu\big\|_\alpha = \|k_\nu^\alpha\|_\infty \quad .$$

(ii) The first statement in (ii) is from Steinwart & Scovel (2012, Theorem 5.3). The converse under the additional assumption of uniformly bounded eigenfunctions is a direct consequence of (16).

(iii) If (EMB) is satisfied for α, then the monotonicity of the eigenvalues $(\mu_i)_{i\geq 1}$ and Statement (ii) imply, for $i \geq 1$,

$$i\mu_i^\alpha \leq \sum_{j=1}^{i} \mu_j^\alpha \leq \sum_{j\geq 1} \mu_j^\alpha =: D < \infty \quad .$$

Consequently, (EVD) is satisfied for $p = \alpha$ and $C := D^{1/\alpha}$. For the converse we assume (EVD) w.r.t. $0 < p < 1$. As a consequence, we have $\sum_{i\geq 1} \mu_i^\alpha \leq C^\alpha \sum_{i\geq 1} i^{-\alpha/p} < \infty$ for all $\alpha > p$ and together with Part (ii) this gives the assertion. ∎

Recall that the *effective dimension* $\mathcal{N}_\nu : (0, \infty) \to [0, \infty)$ is defined by

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

where tr denotes the trace operator. The effective dimension is widely used in the statistical analysis of LS-SVMs (see e.g., Caponnetto & De Vito, 2007; Blanchard & Mücke, 2017; Lin & Cevher, 2018a; Lin et al., 2018). The following lemma establishes a connection between (EVD) and the asymptotic behavior of $\mathcal{N}_\nu(\lambda)$ for $\lambda \to 0^+$.

Lemma 11 *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, and ν be a probability distribution on X. Then the following statements are equivalent, for $0 < p \leq 1$:*

(i) There is a constant $D > 0$ such that the following inequality is satisfied, for $\lambda > 0$,

$$\mathcal{N}_\nu(\lambda) \leq D\lambda^{-p} \quad .$$

(ii) (EVD) *is satisfied for p, i.e. there is a constant $C > 0$ with $\mu_i \leq Ci^{-1/p}$ for all $i \geq 1$.*

Note that (i)⇒(ii) for $p < 1$ is from Caponnetto & De Vito (2007, Proposition 3).

Proof (i)⇐(ii) For $p < 1$ this implication is a consequence of Caponnetto & De Vito (2007, Proposition 3) for $D := C^p/(1-p)$. For $p = 1$ the properties of the trace operator yields $\mathcal{N}_\nu(\lambda) \leq \|C_\nu\|_1 \, \|(C_\nu + \lambda)^{-1}\|$. Since C_ν is a positive semi-definite operator we have $\|(C_\nu + \lambda)^{-1}\| \leq \lambda^{-1}$. Moreover, using the ONS $([e_i]_\nu)_{i\geq 1}$ in $L_2(\nu)$ and the monotone convergence theorem, the trace norm can be bounded by

$$\|C_\nu\|_1 = \sum_{i\geq 1} \mu_i = \sum_{i\geq 1} \mu_i \int_X e_i^2(x) \, \mathrm{d}\nu(x) = \int_X \sum_{i\geq 1} \mu_i e_i^2(x) \, \mathrm{d}\nu(x) \leq \|k_\nu^1\|_\infty^2 =: D$$

(i)⇒(ii) Since $(\mu_i)_{i\geq 1}$ is non-increasing also $(\mu_i/(\mu_i + \lambda))_{i\geq 1}$ is non-increasing for all $\lambda > 0$. Consequently, we have, for $i \geq 1$ and $\lambda > 0$,

$$i\frac{\mu_i}{\mu_i + \lambda} \leq \sum_{j=1}^{i} \frac{\mu_j}{\mu_j + \lambda} \leq \mathcal{N}_\nu(\lambda) \leq D\lambda^{-p} \quad .$$

Using this inequality for $\lambda = \mu_i$ we get $i \leq 2D\mu_i^{-p}$ for all $i \geq 1$ and this yields (EVD) w.r.t. p and $C = (2D)^{1/p}$. ∎

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

$$\mathcal{R}_P(f) := \int_{X\times\mathbb{R}} \bigl(y - f(x)\bigr)^2 \, \mathrm{d}P(x, y)$$

and the *Bayes-LS-risk* $\mathcal{R}_P^* := \inf_{f:X\to\mathbb{R}} \mathcal{R}_P(f)$ is achieved by the conditional mean function f_P^*, see e.g. Steinwart & Christmann (2008, Example 2.6). Moreover, the *LS-excess-risk* is given by $\mathcal{R}_P(f) - \mathcal{R}_P^* = \|[f]_\nu - f_P^*\|_{L_2(\nu)}^2$, see e.g. Steinwart & Christmann (2008, Example 2.6), and minimizing the LS-risk is therefore equivalent to approximating the conditional mean function in the $L_2(\nu)$-norm. For $\lambda > 0$ the unique minimizer of

$$\inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_P(f) \tag{18}$$

can be easily calculated by means of derivatives and is given by

$$f_{P,\lambda} := (C_\nu + \lambda)^{-1} g_P \in H \qquad \text{with} \quad g_P := S_\nu f_P^* \quad , \tag{19}$$

see e.g. Smale & Zhou (2005, Equations 7.4 and Equation 7.5). Note that (5) and (4) together with the properties of the conditional distribution $P(\cdot\,|x)$ yield

$$g_P = \int_X k(x,\,\cdot\,) \int_{\mathbb{R}} y\; P(\mathrm{d}y|x)\; \mathrm{d}P_X(x) = \int_{X\times\mathbb{R}} yk(x,\,\cdot\,)\; \mathrm{d}P(x,y)\;\;. \tag{20}$$

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 (18) w.r.t. the *empirical* measure $D \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ where $\delta_{(x,y)}$ denotes the Dirac measure at (x,y). Consequently, $f_{D,\lambda}$ is given by (19) w.r.t. the corresponding empirical quantities, namely

$$f_{D,\lambda} = (C_\delta + \lambda)^{-1} g_D \in H\;\;, \tag{21}$$

where δ denotes the marginal distribution of D on X, i.e. $\delta = D_X = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$.

6.1 Some Bounds

In this subsection we further exploit the spectral representations in (6) in order to establish some bounds which we use several times in the proofs of our main results.

Recall from Steinwart & Scovel (2012, Theorem 2.11 and Lemma 2.12) that $(\mu_i^{1/2} e_i)_{i\geq 1}$ is an ONB of $(\ker I_\nu)^\perp$, $([e_i]_\nu)_{i\geq 1}$ is an ONB of $\overline{\operatorname{ran} I_\nu} = [H]_\nu^0$, and

$$S_\nu = \sum_{i\geq 1} \mu_i^{1/2} \langle [e_i]_\nu,\,\cdot\,\rangle_{L_2(\nu)} \mu_i^{1/2} e_i\;\;. \tag{22}$$

As the representation in (19) indicates, the operator $(C_\nu + \lambda)^{-a}$, for $a > 0$, plays a crucial role in the following. To this end, we fix an arbitrary ONB $(\tilde{e}_j)_{j\in J}$ of $\ker I_\nu$, with $J \cap \mathbb{N} = \emptyset$, and bring up the following spectral representation

$$(C_\nu + \lambda)^{-a} = \sum_{i\geq 1} (\mu_i + \lambda)^{-a} \langle \mu_i^{1/2} e_i,\,\cdot\,\rangle_H\; \mu_i^{1/2} e_i + \lambda^{-a} \sum_{j\in J} \langle \tilde{e}_j,\,\cdot\,\rangle_H\; \tilde{e}_j\;\;. \tag{23}$$

Note that $(\tilde{e}_j)_{j\in J} \subseteq H$ are normalized in contrast to $(e_i)_{i\geq 1} \subseteq H$, which are not normalized to be aligned with the literature. Moreover, by normalizing $(e_i)_{i\geq 1}$ we get the ONB $(\mu_i^{1/2} e_i)_{i\geq 1} \cup (\tilde{e}_j)_{j\in J}$ of H, where J is at most countably infinite since H is separable.

Next, we present a spectral representation for $f_{P,\lambda}$ which is well-known from Smale & Zhou (2005, proof of Theorem 4). To this end, we use the abbreviation $a_i \coloneqq \langle f_P^*, [e_i]_\nu\rangle_{L_2(\nu)}$, for $i \geq 1$. A combination of (19) with the representations in (22) and (23), for $a = 1$, yields

$$f_{P,\lambda} = \sum_{i\geq 1} \frac{\mu_i^{1/2}}{\mu_i + \lambda}\, a_i\, \mu_i^{1/2} e_i \in (\ker I_\nu)^\perp\;. \tag{24}$$

If we additionally assume $f_P^* \in \overline{\operatorname{ran} I_\nu} = [H]_\nu^0$, then $f_P^* = \sum_{i\geq 1} a_i [e_i]_\nu$ holds and together with (24) we have

$$f_P^* - [f_{P,\lambda}]_\nu = \sum_{i\geq 1} \frac{\lambda}{\mu_i + \lambda}\, a_i\, [e_i]_\nu\;\;. \tag{25}$$

The first lemma describes the connection of the γ-power norm and the H-norm.

Lemma 12 *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, and ν be a probability distribution on X. Then, for $0 \leq \gamma \leq 1$ and $f \in H$, the inequality*

$$\|[f]_\nu\|_\gamma \leq \left\|C_\nu^{\frac{1-\gamma}{2}} f\right\|_H$$

is satisfied. If, in addition, $\gamma < 1$ or $f \perp \ker I_\nu$ is satisfied, then equality holds.

Proof Let us fix a $f \in H$. Since $(\mu_i^{1/2} e_i)_{i \geq 1}$ is an ONB of $(\ker I_\nu)^\perp$, there exits a $g \in \ker I_\nu$ with $f = \sum_{i \geq 1} b_i \, \mu_i^{1/2} e_i + g$, where $b_i = \langle f, \mu_i^{1/2} e_i \rangle_H$ for all $i \geq 1$. Since $[g]_\nu = 0$ we have $[f]_\nu = \sum_{i \geq 1} b_i \mu_i^{1/2} [e_i]_\nu$ and together with Parseval's identity w.r.t. the ONB $(\mu_i^{\gamma/2}[e_i]_\nu)_{i \geq 1}$ of $[H]_\nu^\gamma$ this yields

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

For $\gamma < 1$ the spectral decomposition in (6) together with Parseval's identity w.r.t. the ONS $(\mu_i^{1/2} e_i)_{i \geq 1}$ in H yields

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

This proves the claimed equality in the case of $\gamma < 1$. For $\gamma = 1$ we have $C_\nu^{\frac{1-\gamma}{2}} = \mathrm{Id}_H$ and the Pythagorean theorem together with Parseval's identity yields

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

This gives the claimed equality if $f \perp \ker I_\nu$, i.e. $g = 0$, as well as the claimed inequality for general $f \in H$. ∎

The next lemma describes how the effective dimension comes into play. Note that parts of the next lemma are already mentioned by Rudi et al. (2015) in the discussion after Assumption 3.

Lemma 13 *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, and ν be a probability distribution on X. Then the following equality is satisfied, for $\lambda > 0$,*

$$\int_X \left\| (C_\nu + \lambda)^{-1/2} k(x, \cdot) \right\|_H^2 \, \mathrm{d}\nu(x) = \mathcal{N}_\nu(\lambda) \ . \tag{26}$$

If, in addition, $\|k_\nu^\alpha\|_\infty < \infty$ is satisfied, then the following inequality is satisfied, for $\lambda > 0$ and ν-almost all $x \in X$,

$$\left\| (C_\nu + \lambda)^{-1/2} k(x, \cdot) \right\|_H^2 \leq \|k_\nu^\alpha\|_\infty^2 \lambda^{-\alpha} \ . \tag{27}$$

Note that the inequality in (27) is the place where we benefit from (EMB).

Proof Let us fix a $\lambda > 0$. Since H is separable and k is measurable the map $X \to H$ given by $x \mapsto k(x, \cdot)$ is measurable, see e.g. Steinwart & Christmann (2008, Lemma 4.25) and hence $x \mapsto \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2$ is measurable, too. Using the ONB $(\tilde{e}_j)_{j \in J}$ of $\ker I_\nu$ introduced before Equation (23) and the reproducibility property of the kernel k we get the following series representation which converges in H

$$\begin{aligned} k(x, \cdot) &= \sum_{i \geq 1} \langle \mu_i^{1/2} e_i, k(x, \cdot) \rangle_H \, \mu_i^{1/2} e_i + \sum_{j \in J} \langle \tilde{e}_j, k(x, \cdot) \rangle_H \, \tilde{e}_j \\ &= \sum_{i \geq 1} \mu_i^{1/2} e_i(x) \, \mu_i^{1/2} e_i + \sum_{j \in J} \tilde{e}_j(x) \, \tilde{e}_j \end{aligned}$$

for all $x \in X$. Together with (23), for $a = 1/2$, and Parseval's identity we get

$$\begin{aligned} \|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 &= \left\| \sum_{i \geq 1} \frac{\mu_i^{1/2} e_i(x)}{(\mu_i + \lambda)^{1/2}} \mu_i^{1/2} e_i + \lambda^{-1/2} \sum_{j \in J} \tilde{e}_j(x) \, \tilde{e}_j \right\|_H^2 \\ &= \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda} e_i^2(x) + \frac{1}{\lambda} \sum_{j \in J} \tilde{e}_j^2(x) \end{aligned}$$

for all $x \in X$. Recall that the index set J is at most countable since H is separable. Moreover, $\tilde{e}_j \in \ker I_\nu$ for all $j \in J$ implies that the second summand on the right hand side vanishes for ν-almost all $x \in X$. Consequently, we have

$$\|(C_\nu + \lambda)^{-1/2} k(x, \cdot)\|_H^2 = \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda} e_i^2(x) \tag{28}$$

for ν-almost all $x \in X$. Now, (26) is a consequence of (28), the monotone convergence theorem, and the fact that $([e_i])_{i \geq 1}$ is an ONS in $L_2(\nu)$, namely

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

Finally, (27) is a consequence of (28) and Lemma 25, namely

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

is satisfied for ν-almost all $x \in X$. ∎

The next lemma uses the representations in (24) and (25) to provide bounds on the γ-power norm of $[f_{P,\lambda}]_\nu - f_P^*$ and $[f_{P,\lambda}]_\nu$.

Lemma 14 *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, P be a probability distribution on $X \times \mathbb{R}$ with $|P|_2 < \infty$, and $\nu := P_X$ be the marginal distribution on X. If $f_P^* \in [H]_\nu^\beta$ is satisfied for some $0 \leq \beta \leq 2$, then the following bounds are satisfied, for all $\lambda > 0$:*

$$\|[f_{P,\lambda}]_\nu - f_P^*\|_\gamma^2 \leq \|f_P^*\|_\beta^2 \, \lambda^{\beta - \gamma} \qquad \text{for all } 0 \leq \gamma \leq \beta, \tag{29}$$

$$\|[f_{P,\lambda}]_\nu\|_\gamma^2 \leq \|f_P^*\|_{\min\{\gamma,\beta\}}^2 \lambda^{-(\gamma - \beta)_+} \qquad \text{for all } \gamma \geq 0. \tag{30}$$

Here we used the abbreviation $t_+ := \max\{0,t\}$ for $t \in \mathbb{R}$. Note that (29) in the case of $\gamma \in \{0,1\}$ is covered by Smale & Zhou (2005, 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 (24), we can use the equality from Lemma 12 and exchange the left hand sides of (29) by $\|f_{P,\lambda} - f_P^*\|_H^2$ in the case of $\beta \geq \gamma = 1$. Analogously, we can exchange the left hand side in (30) by $\|f_{P,\lambda}\|_H^2$ for $\gamma = 1$.

Proof Let us first show (29). Since $f_P^* \in [H]_\nu^\beta \subseteq [H]_\nu^0$ we can use the spectral representation in (25). Then, Parseval's identity w.r.t. the ONB $(\mu_i^{\gamma/2}[e_i]_\nu)_{i\geq 1}$ of $[H]_\nu^\gamma$ yields

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

If we estimate the fraction on the right hand side with Lemma 25 and apply Parseval's identity w.r.t. the ONB $(\mu_i^{\beta/2}[e_i]_\nu)_{i\geq 1}$ of $[H]_\nu^\beta$, then we get

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

In order to show (30) we use the spectral representation in (24) and Parseval's identity

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

In the case of $\gamma \leq \beta$ we estimate the fraction by 1 and then Parseval's identity 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 $\gamma > \beta$ we additionally use Lemma 25 and get

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

Whereby, in the last equality we used Parseval's identity again. ∎

If we combine the bounds from Lemma 14 with (EMB) we directly obtain the following $L_\infty(\nu)$ bounds. Note that some parts of the following lemma are already stated by Steinwart & Scovel (2012, Corollary 5.5).

Corollary 15 *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, P be a probability distribution on $X \times \mathbb{R}$ with $|P|_2 < \infty$, and $\nu := P_X$ be the marginal distribution on X. If $f_P^* \in [H]_\nu^\beta$ and* (EMB) *are satisfied for some $0 \leq \beta \leq 2$ and $0 < \alpha \leq 1$, respectively, then the following bounds are satisfied, for all $0 < \lambda \leq 1$:*

$$\|[f_{P,\lambda}]_\nu - f_P^*\|_{L_\infty(\nu)}^2 \leq \left(\|f_P^*\|_{L_\infty(\nu)} + \|k_\nu^\alpha\|_\infty \|f_P^*\|_\beta\right)^2 \lambda^{\beta-\alpha} \tag{31}$$

$$\|[f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}^2 \leq \|k_\nu^\alpha\|_\infty^2 \|f_P^*\|_{\min\{\alpha,\beta\}}^2 \lambda^{-(\alpha-\beta)_+} \ . \tag{32}$$

Proof The bound in (32) is a direct consequence of the Identity (17) in Theorem 9 and (30) with $\gamma = \alpha$.

To prove (31) we can assume without loss of generality $f_P^* \in L_\infty(\nu)$. In the case of $\beta \leq \alpha$ we use the triangle inequality, Inequality (32), and $\lambda \leq 1$ to find

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

In the case $\beta > \alpha$, Bound (31) is a consequence of the Identity (17) in Theorem 9 and (29) with $\gamma = \alpha$. ■

6.2 Upper Rates

In order to establish upper bounds, we split $\|[f_{D,\lambda}]_\nu - f_P^*\|_\gamma$ into two parts:

$$\left\|[f_{D,\lambda}]_\nu - f_P^*\right\|_\gamma \leq \left\|[f_{D,\lambda} - f_{P,\lambda}]_\nu\right\|_\gamma + \left\|[f_{P,\lambda}]_\nu - f_P^*\right\|_\gamma \ , \tag{33}$$

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 has already been given in Lemma 14 and the following inequality controls the estimation error.

Theorem 16 (Error Control Inequality) *Let $(X, \mathcal{B})$ be a measurable space, H be a separable RKHS on X w.r.t. a bounded and measurable kernel k, P be a probability distribution on $X \times \mathbb{R}$ with $|P|_2 < \infty$, and $\nu := P_X$ be the marginal distribution on X. Furthermore, let $\|f_P^*\|_{L_\infty(\nu)} < \infty$, $\|k_\nu^\alpha\|_\infty < \infty$, and* (MOM) *be satisfied. Then for the abbreviations*

$$g_\lambda := \log\left(2e\mathcal{N}_\nu(\lambda)\frac{\|C_\nu\| + \lambda}{\|C_\nu\|}\right), \tag{34}$$

$$A_{\lambda,\tau} := 8\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda \lambda^{-\alpha}, \ \textit{and} \tag{35}$$

$$L_\lambda := \max\{L, \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}\} \tag{36}$$

and $0 \leq \gamma \leq 1$, $\tau \geq 1$, $\lambda > 0$, and $n \geq A_{\lambda,\tau}$, the following bound is satisfied with P^n-probability not less than $1 - 4e^{-\tau}$

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

According to Lemma 12 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 based on an argument tracing back to Smale & Zhou (2007). We refine the analysis with some ideas of Caponnetto & De Vito (2007) and Lin & Cevher (2018a) under the embedding property. We split the proof into several lemmas: the first one improves Lemma 18 of Lin & Cevher (2018a) under the additional Assumption (EMB).

Lemma 17 *Let the assumptions of Theorem 16 be satisfied and g_λ as defined in (34). Then, for $\tau \geq 1$, $\lambda > 0$, and $n \geq 1$, the following operator norm bound is satisfied with ν^n-probability not less than $1 - 2e^{-\tau}$*

$$\left\|(C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}\right\| \leq \frac{4\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda}{3n\lambda^\alpha} + \sqrt{\frac{2\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda}{n\lambda^\alpha}} \quad . \tag{37}$$

Proof This is a consequence of the concentration inequality in Theorem 27, but before we start with the main part of the proof we recall some well-known facts about the mapping $\otimes : H \times H \to \mathcal{L}_2(H)$ into the space of Hilbert-Schmidt operators defined by $f \otimes g \coloneqq \langle f, \cdot\,\rangle_H\, g$. Since $f \otimes g$ has rank one, $f \otimes g$ is a Hilbert-Schmidt operator. Furthermore, $\otimes$ is bilinear, satisfies the following Hilbert-Schmidt norm and operator norm identity

$$\|f \otimes g\|_2 = \|f \otimes g\| = \|f\|_H \|g\|_H \quad , \tag{38}$$

and hence $\otimes$ is continuous. Moreover, the adjoint operator is given by $(f \otimes g)^* = g \otimes f$ and $\langle (f \otimes g)h, p\rangle_H = \langle f, h\rangle_H \langle g, p\rangle_H$ for $h, p \in H$. As a result, $f \otimes f$ is, for all $f \in H$, a self-adjoint positive semi-definite Hilbert-Schmidt operator.

Now, we consider $C_x : H \to H$ the *integral* operator w.r.t. the point measure at $x \in X$,

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

and define the random variables $\xi_0, \xi_1 : X \to \mathcal{L}_2(H)$ by

$$\xi_0(x) \coloneqq C_x \qquad \text{and} \qquad \xi_1(x) \coloneqq (C_\nu + \lambda)^{-1/2} C_x (C_\nu + \lambda)^{-1/2} \quad .$$

Using the definition of the bilinear operator $\otimes$, the self-adjointness of $(C_\nu + \lambda)^{-1/2}$, and the abbreviation $h_x \coloneqq (C_\nu + \lambda)^{-1/2} k(x, \cdot\,)$ we can represent ξ_0 and ξ_1 as follows

$$\begin{aligned} \xi_0(x) f &= \big(k(x, \cdot\,) \otimes k(x, \cdot\,)\big) f \\ \xi_1(x) f &= \big\langle k(x, \cdot\,), (C_\nu + \lambda)^{-1/2} f \big\rangle_H\, (C_\nu + \lambda)^{-1/2} k(x, \cdot\,) \\ &= \big\langle (C_\nu + \lambda)^{-1/2} k(x, \cdot\,), f \big\rangle_H\, (C_\nu + \lambda)^{-1/2} k(x, \cdot\,) \\ &= (h_x \otimes h_x) f \quad . \end{aligned} \tag{39}$$

Since H is a separable RKHS w.r.t. a measurable kernel, the map $X \to H$, $x \mapsto k(x, \cdot\,)$ is measurable, see e.g. Steinwart & Christmann (2008, Lemma 4.25). Consequently, ξ_0 and ξ_1 are measurable, as compositions of measurable functions. Combining (38) with the representations in (39) and Lemma 13 we get the supremum bounds, w.r.t. the Hilbert-Schmidt norm and the operator norm,

$$\begin{aligned} &\|\xi_0(x)\|_2 = \|\xi_0(x)\| = \|k(x, \cdot\,)\|_H^2 = k(x, x) \leq \|k_\nu^1\|_\infty^2 \qquad \text{and} \\ &\|\xi_1(x)\|_2 = \|\xi_1(x)\| = \|(C_\nu + \lambda)^{-1/2} k(x, \cdot\,)\|_H^2 \leq \|k_\nu^\alpha\|_\infty^2 \lambda^{-\alpha} =: B \end{aligned} \tag{40}$$

for ν-almost all $x \in X$. As a consequence of the boundedness w.r.t. the Hilbert-Schmidt norm, the mappings ξ_0 and ξ_1 are Bochner-integrable w.r.t. every probability measure μ on X. Combining Diestel & Uhl (1977, Theorem 6 in Chapter II.2) and $\mathbb{E}_{x\sim\mu} C_x = C_\mu$ yields

$$\mathbb{E}_\mu \xi_1 = (C_\nu + \lambda)^{-1/2} \big(\mathbb{E}_{x\sim\mu} C_x\big) (C_\nu + \lambda)^{-1/2} = (C_\nu + \lambda)^{-1/2} C_\mu (C_\nu + \lambda)^{-1/2} \quad . \tag{41}$$

If we exploit (41) in the case of $\mu = \nu = P_X$ and $\mu = \delta = D_X$, then we get

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

for all $D = \bigl((x_i,y_i)\bigr)_{i=1}^n \in (X\times\mathbb{R})^n$. Consequently, the left hand side of our claimed inequality (37) coincides with the left hand side in Theorem 27 w.r.t. the random variable ξ_1. A supremum bound for ξ_1 is already established in (40) and $\xi_1(x)$ is a positive semi-definite self-adjoint Hilbert-Schmidt operator because of the representation in (39) and the properties of $\otimes$. Finally, we have to provide a *variance* bound for ξ_1. To this end, recall that for two self-adjoint operators R and S on a Hilbert space we write $R \preccurlyeq S$ iff $S - R$ is a positive semi-definite operator. The representation $\xi_1(x) = h_x \otimes h_x$ from (39) together with the supremum bound in (40) yields

$$\xi_1(x)^2 = \xi_1(x)\xi_1(x) = \|h_x\|_H^2\langle h_x, \cdot\,\rangle_H h_x = \|h_x\|_H^2\xi_1(x) \preccurlyeq B\xi_1(x)$$

for all ν-almost all $x \in X$. Since the relation $\preccurlyeq$ remains true if we integrate both sides we get from the identity in (41) with $\mu = \nu$ the variance bound

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

Note that V is a self-adjoint positive semi-definite operator as an integral over self-adjoint positive semi-definite operators. Moreover, using the spectral representation of C_ν in (6) and the spectral representation of $(C_\nu+\lambda)^{-1/2}$ in (23) with $a = 1/2$ we get

$$V = B\sum_{i\geq 1}\frac{\mu_i}{\mu_i+\lambda}\langle \mu_i^{1/2}e_i, \cdot\,\rangle_H\,\mu_i^{1/2}e_i \quad .$$

Since the operator norm coincides with the largest eigenvalue we get

$$\|V\| = B\frac{\mu_1}{\mu_1+\lambda} = B\frac{\|C_\nu\|}{\|C_\nu\|+\lambda} \quad .$$

Moreover, the trace coincides with the sum of the eigenvalues and hence

$$\operatorname{tr}(V) = B\sum_{i\geq 1}\frac{\mu_i}{\mu_i+\lambda} = B\mathcal{N}_\nu(\lambda) \quad .$$

Consequently, Theorem 27 is applicable and together with $\|V\| \leq B$ and $g(V) = g_\lambda$ Theorem 27 yields the assertion. ∎

Lemma 18 *Let the assumptions of Theorem 16 be satisfied and L_λ as in* (36)*. Then, for $\tau \geq 1$, $\lambda > 0$, and $n \geq 1$, the following bound is satisfied with P^n-probability not less than $1 - 2e^{-\tau}$*

$$\begin{aligned}&\bigl\|(C_\nu+\lambda)^{-1/2}\bigl((g_D - C_\delta f_{P,\lambda}) - (g_P - C_\nu f_{P,\lambda})\bigr)\bigr\|_H^2 \\ &\leq \frac{64\tau^2}{n}\left(\sigma^2\mathcal{N}_\nu(\lambda) + \|k_\nu^\alpha\|_\infty^2\frac{\|f_P^* - [f_{P,\lambda}]_\nu\|_0^2}{\lambda^\alpha} + 2\|k_\nu^\alpha\|_\infty^2\frac{L_\lambda^2}{n\lambda^\alpha}\right) \quad .\end{aligned} \tag{42}$$

Proof We consider the random variables $\xi_0, \xi_2 : X \times \mathbb{R} \to H$ defined by

$$\begin{aligned}\xi_0(x,y) &:= (y - f_{P,\lambda}(x))k(x,\,\cdot\,)\\ \xi_2(x,y) &:= (C_\nu + \lambda)^{-1/2}\xi_0(x,y) \ .\end{aligned}$$

Since H is a separable RKHS w.r.t. the measurable kernel k the mappings $x \mapsto k(x,\,\cdot\,)$ and $f_{P,\lambda}$ are measurable, see Steinwart & Christmann (2008, Lemma 4.24 and Lemma 4.25). Consequently, ξ_0 and ξ_2 are measurable, as compositions of measurable functions. Moreover, since our kernel k is bounded also $f_{P,\lambda}$ is bounded and

$$\|\xi_0(x,y)\|_H = |y - f_{P,\lambda}(x)|\,\|k(x,\,\cdot\,)\|_H \leq (|y| + \|f_{P,\lambda}\|_{L_\infty(\nu)})\|k_\nu^1\|_\infty$$

is satisfied for ν-almost all $x \in X$. As a result ξ_0 is Bochner-integrable w.r.t. all probability measures Q on $X \times \mathbb{R}$ with

$$|Q|_1 := \int_{X\times\mathbb{R}} |y| \,\mathrm{d}Q(x,y) < \infty \ .$$

An analogous bound shows that ξ_2 is Bochner-integrable w.r.t. such measures Q. Combining Diestel & Uhl (1977, Theorem 6 in Chapter II.2) and (20) yields

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

If we use this identity for $Q = D$ and $Q = P$, then we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \Big(\xi_2(x_i,y_i) - \mathbb{E}_P\xi_2\Big) &= \mathbb{E}_D\xi_2 - \mathbb{E}_P\xi_2\\ &= (C_\nu + \lambda)^{-1/2}\Big((g_D - C_\delta f_{P,\lambda}) - (g_P - C_\nu f_{P,\lambda})\Big)\end{aligned}$$

and therefore the left hand side of our claimed Inequality (42) coincides with the left hand side of Bernstein's inequality for H-valued random variables from Theorem 26. Consequently, it remains to bound the m-th moment of ξ_2, for $m \geq 2$,

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

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

$$\begin{aligned}\int_{\mathbb{R}} |y - f_{P,\lambda}(x)|^m \ P(\mathrm{d}y|x) &\leq 2^{m-1}\Big(\|\operatorname{Id}_{\mathbb{R}} - f_P^*(x)\|_{L_m(P(\,\cdot\,|x))}^m + |f_P^*(x) - f_{P,\lambda}(x)|^m\Big)\\ &\leq \frac{1}{2}m!(2L)^{m-2}2\sigma^2 + 2^{m-1}|f_P^*(x) - f_{P,\lambda}(x)|^m \ .\end{aligned}$$

for ν-almost all $x \in X$. If we plug this bound into the outer integral and use the abbreviation $h_x := (C_\nu + \lambda)^{-1/2}k(x,\,\cdot\,)$ we get

$$\begin{aligned}\mathbb{E}_P\|\xi_2\|_H^m \leq \frac{1}{2}m!(2L)^{m-2}2\sigma^2 &\int_X \|h_x\|_H^m \,\mathrm{d}\nu(x)\\ &+ 2^{m-1}\int_X \|h_x\|_H^m\,|f_P^*(x) - f_{P,\lambda}(x)|^m \,\mathrm{d}\nu(x) \ .\end{aligned} \tag{43}$$

Using Lemma 13, the first term in (43) can be bounded by

$$\begin{aligned}\frac{1}{2}m!(2L)^{m-2}2\sigma^2\int_X \|h_x\|_H^m \,\mathrm{d}\nu(x) &\leq \frac{1}{2}m!(2L)^{m-2}2\sigma^2\left(\frac{\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}}\right)^{m-2}\int_X \|h_x\|_H^2 \,\mathrm{d}\nu(x)\\ &= \frac{1}{2}m!\left(\frac{2L\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}}\right)^{m-2}2\sigma^2\mathcal{N}_\nu(\lambda)\\ &\leq \frac{1}{2}m!\left(\frac{2L_\lambda\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}}\right)^{m-2}2\sigma^2\mathcal{N}_\nu(\lambda)\ ,\end{aligned}$$

where we only used $L \leq L_\lambda$ in the last step. Again, using Lemma 13, the second term in (43) can be bounded by

$$\begin{aligned}&2^{m-1}\int_X \|h_x\|_H^m\,|f_P^*(x) - f_{P,\lambda}(x)|^m \,\mathrm{d}\nu(x)\\ &\leq \frac{1}{2}\left(\frac{2\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}}\right)^m \|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}^{m-2}\int_X |f_P^*(x) - f_{P,\lambda}(x)|^2 \,\mathrm{d}\nu(x)\\ &= \frac{1}{2}\left(\frac{2\|k_\nu^\alpha\|_\infty\|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}}{\lambda^{\alpha/2}}\right)^{m-2}\|f_P^* - [f_{P,\lambda}]_\nu\|_{L_2(\nu)}^2\frac{4\|k_\nu^\alpha\|_\infty^2}{\lambda^\alpha}\\ &\leq \frac{1}{2}m!\left(\frac{2L_\lambda\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}}\right)^{m-2}\|f_P^* - [f_{P,\lambda}]_\nu\|_{L_2(\nu)}^2\frac{2\|k_\nu^\alpha\|_\infty^2}{\lambda^\alpha}\ ,\end{aligned}$$

where we only used $\|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)} \leq L_\lambda$ and $2 \leq m!$ in the last step. Continuing Estimate (43) we get

$$\mathbb{E}_P\|\xi_2\|_H^m \leq \frac{1}{2}m!\left(\frac{2L_\lambda\|k_\nu^\alpha\|_\infty}{\lambda^{\alpha/2}}\right)^{m-2}2\Big(\sigma^2\mathcal{N}_\nu(\lambda) + \|f_P^* - [f_{P,\lambda}]_\nu\|_0^2\frac{\|k_\nu^\alpha\|_\infty^2}{\lambda^\alpha}\Big)$$

and an application of Bernstein's inequality from Theorem 26 with $L = 2L_\lambda\|k_\nu^\alpha\|_\infty\lambda^{-\alpha/2}$ and $\sigma^2 = 2\big(\sigma^2\mathcal{N}_\nu(\lambda) + \|f_P^* - [f_{P,\lambda}]_\nu\|_0^2\,\|k_\nu^\alpha\|_\infty^2\lambda^{-\alpha}\big)$ yield the assertion. ∎

Proof of Theorem 16. Let us fix some $\tau \geq 1$, $\lambda > 0$, and $n \geq A_{\lambda,\tau}$. For $D \in (X \times \mathbb{R})^n$ the representation $f_{D,\lambda} = (C_\delta + \lambda)^{-1}g_D$ from (21) yields

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

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

$$\begin{aligned}\Big\|C_\nu^{\frac{1-\gamma}{2}}\big(f_{D,\lambda} - f_{P,\lambda}\big)\Big\|_H^2 \leq\ & \big\|C_\nu^{\frac{1-\gamma}{2}}(C_\nu + \lambda)^{-1/2}\big\|^2 && \text{(44a)}\\ &\cdot \big\|(C_\nu + \lambda)^{1/2}(C_\delta + \lambda)^{-1}(C_\nu + \lambda)^{1/2}\big\|^2 && \text{(44b)}\\ &\cdot \|(C_\nu + \lambda)^{-1/2}(g_D - (C_\delta + \lambda)f_{P,\lambda})\|_H^2 && \text{(44c)}\end{aligned}$$

for all $D \in (X \times \mathbb{R})^n$. Now, we consider the three factors on the right hand side separately. Let us start with Term (44a). An application of Lemma 25 yields

$$\big\|C_\nu^{\frac{1-\gamma}{2}}(C_\nu + \lambda)^{-1/2}\big\|^2 = \sup_{i\geq 1}\frac{\mu_i^{1-\gamma}}{\mu_i + \lambda} \leq \lambda^{-\gamma}\ . \tag{45}$$

Next, Factor (44c) can be rearranged using $f_{P,\lambda} = (C_\nu + \lambda)^{-1} g_P$ from (19):

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

Consequently, the Factor (44c) coincides with the right hand side in Lemma 18 and this lemma yields

$$\begin{aligned}&\|(C_\nu + \lambda)^{-1/2}(g_D - (C_\delta + \lambda) f_{P,\lambda})\|_H^2 \\ &\leq \frac{64\tau^2}{n}\left(\sigma^2 \mathcal{N}_\nu(\lambda) + \|k_\nu^\alpha\|_\infty^2 \frac{\|f_P^* - [f_{P,\lambda}]_\nu\|_0^2}{\lambda^\alpha} + 2\|k_\nu^\alpha\|_\infty^2 \frac{L_\lambda^2}{n\lambda^\alpha}\right)\end{aligned} \tag{46}$$

with P^n-probability not less than $1 - 2e^{-\tau}$. Finally, in order to estimate (44b) we start with the following identity

$$\begin{aligned}C_\delta + \lambda &= C_\delta - C_\nu + C_\nu + \lambda \\ &= -(C_\nu - C_\delta) + (C_\nu + \lambda)^{1/2}(C_\nu + \lambda)^{1/2} \\ &= (C_\nu + \lambda)^{1/2}\left(\mathrm{Id} - (C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}\right)(C_\nu + \lambda)^{1/2} \ .\end{aligned}$$

Plugging this into (44b), we get

$$\begin{aligned}&\left\|(C_\nu + \lambda)^{1/2}(C_\delta + \lambda)^{-1}(C_\nu + \lambda)^{1/2}\right\|^2 \\ &= \left\|\left(\mathrm{Id} - (C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}\right)^{-1}\right\|^2 \ .\end{aligned}$$

Lemma 17 gives us an estimate for the operator norm of $(C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}$. Continuing the estimate from Lemma 17 with $n \geq A_{\lambda,\tau}$ and $A_{\lambda,\tau} = 8\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda \lambda^{-\alpha}$ from (35) yields

$$\begin{aligned}\left\|(C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}\right\| &\leq \frac{4}{3} \cdot \frac{\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda}{n\lambda^\alpha} + \sqrt{2 \cdot \frac{\|k_\nu^\alpha\|_\infty^2 \tau g_\lambda}{n\lambda^\alpha}} \\ &\leq \frac{4}{3} \cdot \frac{1}{8} + \sqrt{2 \cdot \frac{1}{8}} = \frac{2}{3}\end{aligned}$$

with ν^n-probability not less than $1 - 2e^{-\tau}$. Consequently, the inverse of

$$\mathrm{Id} - (C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}$$

can be represented by the Neumann series. In particular, the Neumann series gives us the following bound on (44b)

$$\begin{aligned}&\left\|(C_\nu + \lambda)^{1/2}(C_\delta + \lambda)^{-1}(C_\nu + \lambda)^{1/2}\right\|^2 \\ &= \left\|\left(\mathrm{Id} - (C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}\right)^{-1}\right\|^2 \\ &\leq \left(\sum_{k=0}^\infty \left\|(C_\nu + \lambda)^{-1/2}(C_\nu - C_\delta)(C_\nu + \lambda)^{-1/2}\right\|^k\right)^2 \\ &\leq \left(\sum_{k=0}^\infty \left(\frac{2}{3}\right)^k\right)^2 = 9\end{aligned} \tag{47}$$

with ν^n-probability not less than $1 - 2e^{-\tau}$. Now, if we combine the estimate in (44) with (45), (46), and (47), then we get the claimed bound, with P^n-probability not less than $1 - 4e^{-\tau}$. ∎

Proof of Theorem 1. Let us fix some $\tau \geq 1$ and some lower bound $0 < c \leq 1$ with $c \leq \|C_\nu\|$. First, we show that Theorem 16 is applicable. To this end, we prove in both cases, $\beta + p \leq \alpha$ and $\beta + p > \alpha$, that there is an index bound $n_0 \geq 1$ such that $n \geq A_{\lambda_n,\tau}$ is satisfied for all $n \geq n_0$. Since $\lambda_n \to 0$ we choose $n_0' \geq 1$ such that $\lambda_n \leq c \leq \min\{1, \|C_\nu\|\}$ for all $n \geq n_0'$. Using the definitions of $A_{\lambda_n,\tau}$ and g_λ in (35) and (34), respectively, $\lambda_n \leq c \leq \|C_\nu\|$, $\mathcal{N}_\nu(\lambda_n) \leq D\lambda_n^{-p}$ from Lemma 11, and $\|k_\nu^\alpha\|_\infty \leq A$ from (EMB) and (17) we get, for $n \geq n_0'$,

$$\begin{aligned}
\frac{A_{\lambda_n,\tau}}{n} &= 8\|k_\nu^\alpha\|_\infty^2 \tau \frac{g_{\lambda_n}}{n\lambda_n^\alpha} \\
&= 8\|k_\nu^\alpha\|_\infty^2 \tau \frac{\log\bigl(2e\mathcal{N}_\nu(\lambda_n)(1 + \lambda_n/\|C_\nu\|)\bigr)}{n\lambda_n^\alpha} \\
&\leq 8A^2\tau \frac{\log\bigl(4eD\lambda_n^{-p}\bigr)}{n\lambda_n^\alpha} \\
&= 8A^2\tau \biggl(\frac{\log\bigl(4eD\bigr)}{n\lambda_n^\alpha} + p\frac{\log\bigl(\lambda_n^{-1}\bigr)}{n\lambda_n^\alpha} \biggr) \; .
\end{aligned}$$

Consequently, it is enough to show $\frac{\log(\lambda_n^{-1})}{n\lambda_n^\alpha} \to 0$. To this end, we consider the cases $\beta + p \leq \alpha$ and $\beta + p > \alpha$ separately.

(i) In the case of $\beta + p \leq \alpha$ we have $\lambda_n \asymp \bigl(n/\log^r(n)\bigr)^{-1/\alpha}$ for some $r > 1$ and hence

$$\frac{\log\bigl(\lambda_n^{-1}\bigr)}{n\lambda_n^\alpha} \asymp \frac{\log(n)}{n(n/\log^r(n))^{-1}} = \frac{1}{\log^{r-1}(n)} \to 0 \; .$$

(ii) In the case of $\beta + p > \alpha$ we have $1 - \frac{\alpha}{\beta+p} > 0$, $\lambda_n \asymp n^{-1/(\beta+p)}$, and hence

$$\frac{\log\bigl(\lambda_n^{-1}\bigr)}{n\lambda_n^\alpha} \asymp \frac{\log(n)}{n^{1-\frac{\alpha}{\beta+p}}} \to 0 \; .$$

Consequently, there is a $n_0 \geq n_0'$ with $n \geq A_{\lambda_n,\tau}$ for all $n \geq n_0$. Moreover, n_0 just depends on $(\lambda_n)_{n\geq 1}$, c, τ, A, D, and on the parameters α, p.

Let $n \geq n_0$ be fixed. From Lemma 12 and Theorem 16 we get the bound

$$\|[f_{D,\lambda_n} - f_{P,\lambda_n}]_\nu\|_\gamma^2 \leq \frac{576\tau^2}{n\lambda_n^\gamma} \biggl(\sigma^2\mathcal{N}_\nu(\lambda_n) + \|k_\nu^\alpha\|_\infty^2 \frac{\|f_P^* - [f_{P,\lambda_n}]_\nu\|_{L_2(\nu)}^2}{\lambda_n^\alpha} + 2\|k_\nu^\alpha\|_\infty^2 \frac{L_{\lambda_n}^2}{n\lambda_n^\alpha} \biggr) \; .$$

Continuing this estimate by using $\mathcal{N}_\nu(\lambda_n) \leq D\lambda_n^{-p}$ from Lemma 11, $\|k_\nu^\alpha\|_\infty \leq A$ from (EMB) and (17), and $\|f_P^* - [f_{P,\lambda_n}]_\nu\|_{L_2(\nu)}^2 \leq B^2\lambda_n^\beta$ from Lemma 14 and (SRC) we get

$$\|[f_{D,\lambda_n} - f_{P,\lambda_n}]_\nu\|_\gamma^2 \leq 576\frac{\tau^2}{n\lambda_n^\gamma} \biggl(\sigma^2 D\lambda_n^{-p} + A^2B^2\lambda_n^{\beta-\alpha} + 2A^2\frac{L_{\lambda_n}^2}{n\lambda_n^\alpha} \biggr) \; . \tag{48}$$

Combining the definition of L_λ in (36) with Corollary 15 and $\lambda_n \leq 1$ we get

$$\begin{aligned} L_{\lambda_n}^2 &= \max\bigl\{L^2, \|f_P^* - [f_{P,\lambda_n}]_\nu\|_{L_\infty(\nu)}^2\bigr\} \\ &\leq \max\bigl\{L^2, \bigl(\|f_P^*\|_{L_\infty(\nu)} + \|k_\nu^\alpha\|_\infty \|f_P^*\|_\beta\bigr)^2 \lambda_n^{-(\alpha-\beta)}\bigr\} \\ &\leq K_0\, \lambda_n^{-(\alpha-\beta)_+} \end{aligned}$$

with $K_0 := \max\bigl\{L^2, \bigl(B_\infty + AB\bigr)^2\bigr\}$. For the first and second addend in (48) we use again $\lambda_n \leq 1$ and get

$$\sigma^2 D \lambda_n^{-p} + A^2 B^2 \lambda_n^{\beta-\alpha} \leq \bigl(\sigma^2 D + A^2 B^2\bigr) \max\{\lambda_n^{-p}, \lambda_n^{-(\alpha-\beta)}\} = K_1 \lambda_n^{-\max\{p,\alpha-\beta\}}$$

with $K_1 := \sigma^2 D + A^2 B^2$. Plugging both bounds into (48) gives us

$$\begin{aligned} \|[f_{D,\lambda_n} - f_{P,\lambda_n}]_\nu\|_\gamma^2 &\leq 576 \frac{\tau^2}{n\lambda_n^\gamma} \left(K_1 \lambda_n^{-\max\{p,\alpha-\beta\}} + 2A^2 K_0 \frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+}} \right) \\ &= 576 \frac{\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}} \left(K_1 + 2A^2 K_0 \frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}} \right) . \end{aligned}$$

Next, we show that the second term in the brackets is bounded. To this end, we consider the cases $\beta + p \leq \alpha$ and $\beta + p > \alpha$ separately.

(i) In the case of $\beta + p \leq \alpha$ we have $0 < p \leq \alpha - \beta$ and

$$\alpha + (\alpha - \beta)_+ - \max\{p, \alpha - \beta\} = \alpha \ .$$

Since $\lambda_n \asymp \bigl(n/\log^r(n)\bigr)^{-1/\alpha}$ for some $r > 1$ we get

$$\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}} = \frac{1}{n\lambda_n^\alpha} \asymp \frac{1}{\log^r(n)} \ .$$

(ii) In the case of $\beta + p > \alpha$ we have $p > \alpha - \beta$, $\lambda_n \asymp n^{-1/(\beta+p)}$, and hence

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

Using $p > \alpha - \beta$ again gives us

$$1 - \frac{\alpha + (\alpha-\beta)_+ - p}{\beta + p} = \frac{2p - (\alpha-\beta) - (\alpha-\beta)_+}{\beta + p} > 0$$

Consequently, there is a constant $K_2 > 0$ with

$$\|[f_{D,\lambda_n} - f_{P,\lambda_n}]_\nu\|_\gamma^2 = 576 \frac{\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}} \bigl(K_1 + 2A^2 K_0 K_2\bigr)$$

for all $n \geq n_0$. Combining this with the splitting in (33) and with Lemma 14 yields, for $K_3 := 576(K_1 + 2A^2 K_0 K_2)$,

$$\begin{aligned} \bigl\|[f_{D,\lambda}]_\nu - f_P^*\bigr\|_\gamma^2 &\leq 2B^2 \lambda_n^{\beta-\gamma} + 2K_3 \frac{\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}} \\ &\leq \tau^2 \lambda_n^{\beta-\gamma} \left(2B^2 + 2K_3 \frac{1}{n\lambda_n^{\max\{\alpha,\beta+p\}}} \right) . \end{aligned}$$

Since in both cases, $\beta + p \leq \alpha$ and $\beta + p > \alpha$, we have $\lambda_n \succcurlyeq n^{-1/\max\{\alpha,\beta+p\}}$ the term inside the brackets is bounded by some constant $K > 0$ and hence we have

$$\left\|[f_{D,\lambda}]_\nu - f_P^*\right\|_\gamma^2 \leq \tau^2 K \lambda_n^{\beta-\gamma}$$

for all $n \geq n_0$. This is the assertion, in both cases. ■

6.3 Lower Rates

We establish the following lower bound in order to prove γ-lower rates .

Lemma 19 (Lower Bound) *Let $(X, \mathcal{B})$ be a measurable space, 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 the following statement is satisfied. For all $0 < \varepsilon \leq \varepsilon_0$ there is a $M_\varepsilon \geq 1$ with*

$$2^{C_2 \varepsilon^{-u}} \leq M_\varepsilon \leq 2^{3C_2 \varepsilon^{-u}} \tag{49}$$

for $u := \frac{p}{\max\{\alpha,\beta\}-\gamma}$ and there are probability measures $P_0, P_1, \ldots, P_{M_\varepsilon}$ with marginal distribution $(P_j)_X = \nu$ on X, $\|f_{P_j}^*\|_{L_\infty(\nu)} \leq B_\infty$, (SRC) *w.r.t. β, B, and* (MOM) *w.r.t. σ, L. Moreover, $P_0, P_1, \ldots, P_{M_\varepsilon}$ satisfy*

$$\|f_{P_i}^* - f_{P_j}^*\|_\gamma^2 \geq 4\varepsilon \tag{50}$$

for all $i, j \in \{0, 1, \ldots, M_\varepsilon\}$ with $i \neq j$ and

$$\max_{j=0,1,\ldots,M_\varepsilon} P_j^n\big(D : \ \Psi(D) \neq j\big) \geq \frac{\sqrt{M_\varepsilon}}{1+\sqrt{M_\varepsilon}}\Big(1 - C_1 n \varepsilon^{\frac{\max\{\alpha,\beta\}+p}{\max\{\alpha,\beta\}-\gamma}} - \frac{1}{2\log(M_\varepsilon)}\Big) \tag{51}$$

for all $n \geq 1$ and all measurable functions $\Psi : (X \times \mathbb{R})^n \to \{0, 1, \ldots, M_\varepsilon\}$.

Note that the probability measures P_j also depend on ε although we omit this in the notation. Moreover, just one probability measure ν on X with the required properties is needed to construct distributions on $X \times \mathbb{R}$ that are *difficult* to learn. The proof of Lemma 19 is an application of Tsybakov (2009, Proposition 2.3) stated in the following theorem. To this end, recall that the *Kullback-Leibler divergence* of two probability measures P_1, P_2 on some measurable space $(\Omega, \mathcal{A})$ is given by

$$K(P_1, P_2) := \int_\Omega \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_1$$

if $P_1 \ll P_2$ and otherwise $K(P_1, P_2) := \infty$.

Theorem 20 (Lower Bound) *Let $M \geq 2$, $(\Omega, \mathcal{A})$ be a measurable space, $P_0, P_1, \ldots, P_M$ be probability measures on $(\Omega, \mathcal{A})$ with $P_j \ll P_0$ for all $j = 1, \ldots, M$, and $0 < \alpha_* < \infty$ with*

$$\frac{1}{M}\sum_{j=1}^M K(P_j, P_0) \leq \alpha_* \ .$$

Then, for all measurable functions $\Psi : \Omega \to \{0, 1, \ldots, M\}$, the following bound is satisfied

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

Proof From Tsybakov (2009, Proposition 2.3) we know, that

$$\max_{j=0,1,\ldots,M} P_j\big(\omega \in \Omega : \ \Psi(\omega) \neq j\big) \geq \sup_{0<\tau<1} \frac{\tau M}{1+\tau M}\left(1 + \frac{\alpha_* + \sqrt{\alpha_*/2}}{\log(\tau)}\right)$$

is satisfied. If we choose $\tau = M^{-1/2}$, then we get

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

where we used the estimate $\sqrt{2\alpha_*} \leq 1/2 + \alpha_*$ in the last inequality. ∎

We use this theorem for the measurable space $\Omega = (X \times \mathbb{R})^n$ and follow the suggestion of Caponnetto & De Vito (2007) and Blanchard & Mücke (2017) in order to construct a family of probability measures $P_0, P_1, \ldots, P_M$. To this end, let the assumptions of Lemma 19 be satisfied and set $\bar{\sigma} := \min\{\sigma, L\}$. Moreover, we define for a measurable function $f : X \to \mathbb{R}$ and $x \in X$ the conditional distribution $P_f(\,\cdot\,|x) := \mathcal{N}(f(x), \bar{\sigma}^2)$ as the normal distribution on $\mathbb{R}$ with mean $f(x)$ and variance $\bar{\sigma}^2$. Consequently,

$$P_f(A) := \int_X \int_{\mathbb{R}} \mathbb{1}_A(x, y)\ P_f(\mathrm{d}y|x)\ \mathrm{d}\nu(x) , \tag{52}$$

for $A \in \mathcal{B} \otimes \mathcal{B}(\mathbb{R})$, defines a probability measure on $X \times \mathbb{R}$ with marginal distribution $(P_f)_X = \nu$ on X. For this reason, the corresponding power spaces $[H]^\alpha_\nu$ are independent of f. Since $P_f = P_{f'}$ is satisfied if $f' = f$ ν-a.s. we define $P_{[f]_\nu}$ for ν-equivalence classes. Moreover, for $f \in L_2(\nu)$, we get $|P_f|_2^2 = \bar{\sigma}^2 + \|f\|^2_{L_2(\nu)} < \infty$ and the conditional mean function $f^*_{P_f}$ of P_f coincides with f.

Lemma 21 (Moment Condition) *For a measurable function $f : X \to \mathbb{R}$ the probability measure P_f defined in* (52) *satisfies* (MOM) *for $\sigma = L = \bar{\sigma}$.*

Proof Let us fix an $x \in X$ and an $m \geq 2$. Since $P_f(\,\cdot\,|x) = \mathcal{N}(f(x), \bar{\sigma}^2)$, the mapping $y \mapsto (y - f(x))/\bar{\sigma}$ is standard normally distributed under the measure $P_f(\,\cdot\,|x)$ and

$$\int_{\mathbb{R}} |y - f(x)|^m\ P_f(\mathrm{d}y|x) = \bar{\sigma}^m \mathbb{E}|Z|^m$$

with some standard normally distributed random variable Z. Consequently, it remains to show $\mathbb{E}|Z|^m \leq m!/2$. For $m = 2k$ with some $k \geq 1$ the moments of Z are well-known, see e.g. Bauer (1996, Equation 4.20),

$$\mathbb{E}|Z|^m = (m-1)(m-3) \cdot \ldots \cdot 3 \cdot 1 \leq m!/m \leq m!/2 \ . \tag{53}$$

For $m = 2k-1$ with some $k \geq 2$ we use Hölder's inequality to get $(\mathbb{E}|Z|^m)^{1/m} \leq (\mathbb{E}|Z|^{m+1})^{1/(m+1)}$. Using (53) with $m+1 = 2k$ and $m \geq 3$ we get

$$\mathbb{E}|Z|^m \leq \big(m(m-2)\cdot\ldots\cdot 3\cdot 1\big)^{\frac{m}{m+1}} \leq \big(m!/2\big)^{\frac{m}{m+1}} \leq m!/2$$

This gives the assertion for all $m \geq 2$. ∎

To sum up, we reduced the construction of probability measures to the construction of functions $f_0, f_1, \ldots, f_M \in L_\infty(\nu) \cap [H]^\beta_\nu$ with $\|f_j\|^2_{L_\infty(\nu)} \leq B_\infty$ and $\|f_j\|^2_\beta \leq B$ for $j = 0, 1, \ldots, M$. Before we start with the construction we recall the following lemma from Blanchard & Mücke (2017, Proposition 6.2).

Lemma 22 (Kullback-Leibler Divergence) *For $f, f' \in L_2(\nu)$ and the corresponding probability measures P_f, $P_{f'}$ defined in* (52) *the Kullback-Leibler divergence satisfies, for $n \geq 1$,*

$$K(P_f^n, P_{f'}^n) = \frac{n}{2\bar{\sigma}^2}\|f - f'\|^2_{L_2(\nu)} \ .$$

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

$$f_\omega \coloneqq 2\Big(\frac{8\varepsilon}{m}\Big)^{1/2} \sum_{i=1}^m \omega_i \, \mu_{i+m}^{\gamma/2} \, [e_{i+m}]_\nu \tag{54}$$

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

Lemma 23 *Under the assumptions of Lemma 19 and $u = \frac{p}{\max\{\alpha,\beta\}-\gamma}$, for all $0 \leq \beta \leq 2$ and $0 \leq \gamma \leq 1$ with $\gamma < \beta$, there are constants $U > 0$ and $0 < \varepsilon_1 \leq 1$ such that for all $0 < \varepsilon \leq \varepsilon_1$ and all $m \leq U\varepsilon^{-u}$ the function f_ω defined in* (54) *satisfies the bounds $\|f_\omega\|_\beta \leq B$ and $\|f_\omega\|_{L_\infty(\nu)} \leq B_\infty$ for all $\omega \in \{0,1\}^m$.*

Note that, if we do *not* require the functions f_ω to be uniformly bounded, i.e. we omit the condition $\|f_\omega\|_{L_\infty(\nu)} \leq B_\infty$, then the same result is satisfied for $u = \frac{p}{\beta-\gamma}$.

Proof Let us fix $m \in \mathbb{N}$ and $0 < \varepsilon \leq 1$. First, we consider the condition $\|f_\omega\|_\beta \leq B$. Using (EVD+) and $\gamma < \beta$ we get

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

for $m \leq U_1 \varepsilon^{-\frac{p}{\beta-\gamma}}$ with $U_1 \coloneqq \big(B^2/32\big)^{\frac{p}{\beta-\gamma}} c^p/2$. Next, we consider the condition $\|f_\omega\|_{L_\infty(\nu)} \leq B_\infty$ for the cases $\gamma < \alpha$ and $\gamma \geq \alpha$ separately. In the case of $\gamma < \alpha$, (EMB) together with an analogues argument with α instead of β yields

$$\|f_\omega\|^2_{L_\infty(\nu)} \leq A^2 \|f_\omega\|^2_\alpha \leq B_\infty^2$$

for $m \leq U_2 \varepsilon^{-\frac{p}{\alpha-\gamma}}$ with $U_2 := \big((B_\infty/A)^2/32\big)^{\frac{p}{\alpha-\gamma}} c^p/2$. Consequently, for $U := \min\{U_1, U_2\}$, both conditions, $\|f_\omega\|_\beta \leq B$ and $\|f_\omega\|_{L_\infty(\nu)} \leq B_\infty$, are satisfied if

$$m \leq U \min\Big\{\varepsilon^{-\frac{p}{\beta-\gamma}}, \varepsilon^{-\frac{p}{\alpha-\gamma}}\Big\} = U\varepsilon^{-\min\{\frac{p}{\beta-\gamma}, \frac{p}{\alpha-\gamma}\}} = U\varepsilon^{-u} \quad .$$

Note that there is some $m \in \mathbb{N}$ satisfying this bound since we ensure $U\varepsilon^{-u} \geq 1$ by choosing $0 < \varepsilon \leq \varepsilon_1 := \min\{1, U^{1/u}\}$. In the case of $\gamma \geq \alpha$, (EMB) and (EVD) implies

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

for all $m \geq 1$ and $0 < \varepsilon \leq B_\infty^2/(32A^2C^{\gamma-\alpha})$. Since $\gamma \geq \alpha$ and $\beta > \gamma$ implies $\beta > \alpha$ and $u = \frac{p}{\beta-\gamma}$, both conditions, $\|f_\omega\|_\beta \leq B$ and $\|f_\omega\|_{L_\infty(\nu)} \leq B_\infty$, are satisfied for $m \leq U\varepsilon^{-u}$ and $0 < \varepsilon \leq \varepsilon_1$ with $U := U_1$ and $\varepsilon_1 := \min\{B_\infty^2/(32A^2C^{\gamma-\alpha}), U_1^{1/u}\}$. ∎

If $\omega' = (\omega'_1, \ldots, \omega'_m) \in \{0,1\}^m$ is an other binary string, we investigate the norm of the difference $f_\omega - f_{\omega'}$. To this end, we set $v := \gamma/p$ and use (EVD)

$$\|f_\omega - f_{\omega'}\|^2_{L_2(\nu)} = \frac{32\varepsilon}{m} \sum_{i=1}^m (\omega_i - \omega'_i)^2 \mu_i^\gamma \leq 32\varepsilon\mu_m^\gamma \leq 32C^\gamma\, \varepsilon m^{-v} \quad . \tag{55}$$

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

$$\|f_\omega - f_{\omega'}\|^2_\gamma = \frac{32\varepsilon}{m} \sum_{i=1}^m (\omega_i - \omega'_i)^2 \geq 4\varepsilon \quad . \tag{56}$$

The following lemma is from Tsybakov (2009, Lemma 2.9) and claims that there are many binary strings with large distances.

Lemma 24 (Gilbert-Varshamov Bound) *For $m \geq 8$ there exists some $M \geq 2^{m/8}$ and some binary strings $\omega^{(0)}, \ldots, \omega^{(M)} \in \{0,1\}^m$ with $\omega^{(0)} = (0, \ldots, 0)$ and*

$$\sum_{i=1}^m \big(\omega_i^{(j)} - \omega_i^{(k)}\big)^2 \geq m/8 \tag{57}$$

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

Proof of Lemma 19. Using the constants $U > 0$ and $0 < \varepsilon_1 \leq 1$ from Lemma 23 we define $\varepsilon_0 := \min\{\varepsilon_1, (U/9)^{1/u}\}$ and $m_\varepsilon := \lfloor U\varepsilon^{-u} \rfloor$. Now, we fix an $n \geq 1$ and an $0 < \varepsilon \leq \varepsilon_0$. Since $m_\varepsilon \geq 9$, Lemma 24 yields at least $M_\varepsilon := \lceil 2^{m_\varepsilon/8} \rceil \geq 3$ binary strings $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M_\varepsilon)} \in \{0,1\}^{m_\varepsilon}$ satisfying (57). According to Lemma 23, for $j = 0, 1, \ldots, M_\varepsilon$, the corresponding functions $f_j := f_{\omega^{(j)}}$ defined in (54) satisfy the bounds $\|f_j\|_{L_\infty(\nu)} \leq B_\infty$ and

$\|f_j\|_\beta \le B$. Consequently, for $j = 0, 1, \dots, M_\varepsilon$, the corresponding probability distribution $P_j := P_{f_j}$ defined in (52) satisfies $\|f^*_{P_j}\|_{L_\infty(\nu)} \le B_\infty$ and (SRC) w.r.t. β, B. Recall that P_j additionally satisfies $(P_j)_X = \nu$ and (MOM) w.r.t. σ, L according to Lemma 21. It remains to prove the Statements (49),(50), and (51). Due to the definitions of M_ε, m_ε and $m_\varepsilon \ge 9$ we get $8U/9\ \varepsilon^{-u} \le m_\varepsilon \le U\varepsilon^{-u}$ and

$$2^{U/9\ \varepsilon^{-u}} \le 2^{m_\varepsilon/8} \le M_\varepsilon \le 2^{m_\varepsilon/4} \le 2^{U/3\ \varepsilon^{-u}} \quad .$$

Consequently, (49) is satisfied for $C_2 := U/9$. The inequality in (50) is a consequence of our choice of the binary strings with (57) and the inequality in (56). Lemma 22 and (55) yield

$$\frac{1}{M_\varepsilon}\sum_{j=1}^{M_\varepsilon} K(P_j^n, P_0^n) = \frac{n}{2\bar\sigma^2 M_\varepsilon}\sum_{j=1}^{M_\varepsilon} \|f_j - f_0\|^2_{L_2(\nu)} \le \frac{16C^\gamma}{\bar\sigma^2}\, n\varepsilon m_\varepsilon^{-v} \quad .$$

Furthermore, using $m_\varepsilon \ge 8U/9\ \varepsilon^{-u}$ we find

$$\frac{1}{M_\varepsilon}\sum_{j=1}^{M_\varepsilon} K(P_j^n, P_0^n) \le C_3 n\varepsilon^{1+uv} =: \alpha_*$$

with $C_3 := \frac{16C^\gamma 9^v}{\bar\sigma^2(8U)^v}$. For a measurable function $\Psi : (X \times \mathbb{R})^n \to \{0, 1, \dots, M_\varepsilon\}$, Theorem 20 and $M_\varepsilon \ge 2^{C_2\varepsilon^{-u}}$ from (49) yields

$$\begin{aligned}\max_{j=0,1,\dots,M_\varepsilon} P_j^n\big(D :\ \Psi(D) \ne j\big) &\ge \frac{\sqrt{M_\varepsilon}}{1+\sqrt{M_\varepsilon}}\left(1 - \frac{3C_3 n\varepsilon^{1+uv}}{\log(M_\varepsilon)} - \frac{1}{2\log(M_\varepsilon)}\right) \\ &\ge \frac{\sqrt{M_\varepsilon}}{1+\sqrt{M_\varepsilon}}\left(1 - \frac{3C_3}{C_2\log(2)} n\varepsilon^{1+uv+u} - \frac{1}{2\log(M_\varepsilon)}\right) \quad .\end{aligned}$$

Since $1 + uv + u = \frac{\max\{\alpha,\beta\}+p}{\max\{\alpha,\beta\}-\gamma}$, this gives us (51) for $C_1 := \frac{3C_3}{C_2\log(2)}$. ∎

Now, the proof of Theorem 2 remains an application of Lemma 19 and the general reduction scheme from Tsybakov (2009, Section 2.2).

Proof of Theorem 2. Let $D \mapsto f_{D,\lambda}$ be a (measurable) learning method. Furthermore, we use the notation of Lemma 19, set $r := \frac{\max\{\alpha,\beta\}-\gamma}{\max\{\alpha,\beta\}+p}$, and fix $\tau > 0$ and $n \ge 1$ with $\varepsilon_n := \tau n^{-r} \le \varepsilon_0$. It remains to show that there is a distribution P which is difficult to learn for the considered learning method. Lemma 19, for $\varepsilon = \varepsilon_n$, provides possible candidates $P_0, P_1, \dots, P_{M_n}$, with $M_n := M_{\varepsilon_n}$, each satisfying the requirements of Theorem 2. Next, we estimate the left hand side of the inequality in (51). To this end, we consider the measurable function $\Psi : (X \times \mathbb{R})^n \to \{0, 1, \dots, M_n\}$ defined by

$$\Psi(D) := \operatorname*{argmin}_{j=0,1,\dots,M_n} \|[f_D]_\nu - f_j\|_\gamma \quad . \tag{58}$$

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

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

Consequently, for all $j = 0, 1, \ldots, M_n$ we find

$$P_j^n\big(D : \Psi(D) \neq j\big) \leq P_j^n\big(D : \|[f_D]_\nu - f^*_{P_j}\|_\gamma^2 \geq \varepsilon_n\big) \ .$$

According to (51), for Ψ defined in (58), we have

$$\begin{aligned}\max_{j=0,1,\ldots,M_n} P^n\big(D : \|[f_D]_\nu - f^*_P\|_\gamma^2 \geq \varepsilon_n\big) &\geq \max_{j=0,1,\ldots,M_n} P^n\big(D : \Psi(D) \neq j\big) \\ &\geq \frac{\sqrt{M_n}}{\sqrt{M_n}+1}\Big(1 - C_1\tau^{1/r} - \frac{1}{2\log(M_n)}\Big) \ .\end{aligned}$$

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

Acknowledgments

The authors are especially grateful to Nicole Mücke for pointing them to the article of Lin, Rudi, Rosasco, and Cevher (2018). Moreover, the authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Simon Fischer.

Appendix A. Auxiliary Results and Concentration Inequalities

Lemma 25 *Let, for $\lambda > 0$ and $0 \leq \alpha \leq 1$, the function $f_{\lambda,\alpha} : [0,\infty) \to \mathbb{R}$ be defined by $f_{\lambda,\alpha}(t) := t^\alpha/(\lambda + t)$. In the case $\alpha = 0$ the function $f_{\lambda,\alpha}$ is decreasing and in the case of $\alpha = 1$ the function $f_{\lambda,\alpha}$ is increasing. Furthermore, the supremum of $f_{\lambda,\alpha}$ satisfies the following bound*

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

In the case of $0 < \alpha < 1$ the function $f_{\lambda,\alpha}$ attain its supremum at $t^* := \lambda\alpha/(1-\alpha)$.

Proof In order to prove this statement we use the derivative of $f_{\lambda,\alpha}$, which is given by

$$f'_{\lambda,\alpha}(t) = \frac{\alpha t^{\alpha-1}(\lambda + t) - t^\alpha}{(\lambda+t)^2} \ .$$

For $\alpha = 0$ we have $f'_{\lambda,\alpha}(t) = -(\lambda+t)^{-2} < 0$ and hence $\sup_{t\geq 0} f_{\lambda,\alpha}(t) = f_{\lambda,\alpha}(0) = \lambda^{\alpha-1}$. For $\alpha = 1$ we have $f'_{\lambda,\alpha}(t) = \lambda(\lambda+t)^{-2} > 0$ and hence $\sup_{t\geq 0} f_{\lambda,\alpha}(t) = \lim_{t\to\infty} f_{\lambda,\alpha}(t) = 1 = \lambda^{\alpha-1}$. For $0 < \alpha < 1$ the derivative $f'_{\lambda,\alpha}$ has a unique root at $t^* = \alpha\lambda/(1-\alpha)$. Since $f_{\lambda,\alpha}(0) = 0$ and $\lim_{t\to\infty} f_{\lambda,\alpha}(t) = 0$ holds, $f_{\lambda,\alpha}$ attains its global maximum at t^* and

$$\sup_{t\geq 0} f_{\lambda,\alpha}(t) = f_{\lambda,\alpha}(t^*) = \lambda^{\alpha-1}\alpha^\alpha(1-\alpha)^{1-\alpha} \ .$$

Since $g(\alpha) := \alpha^\alpha(1-\alpha)^{1-\alpha}$ is bounded by 1 the upper bound follows. The derivative

$$g'(\alpha) = g(\alpha)\log\left(\frac{\alpha}{1-\alpha}\right)$$

of g has a unique root at $\alpha = 1/2$ and hence the lower bound follows from $g(\alpha) \geq g(1/2) = 1/2$ for all $0 < \alpha < 1$. ■

The following Bernstein type inequality for Hilbert space valued random variables is due to Pinelis & Sakhanenko (1986). However we use a version from Caponnetto & De Vito (2007, Proposition 2).

Theorem 26 (Bernstein's Inequality) *Let $(\Omega, \mathcal{B}, P)$ be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ 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 $\tau \geq 1$ and $n \geq 1$, the following concentration inequality is satisfied

$$P^n \left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P \xi \right\|_H^2 \geq 32 \frac{\tau^2}{n} \left(\sigma^2 + \frac{L^2}{n} \right) \right) \leq 2e^{-\tau} \ .$$

Proof The m-th moment of the centered random variable $\xi - \mathbb{E}_P \xi$ is bounded by

$$\mathbb{E}_P \|\xi - \mathbb{E}_P \xi\|_H^m \leq 2^{m-1} \big(\mathbb{E}_P \|\xi\|_H^m + \|\mathbb{E}_P \xi\|_H^m \big) \leq 2^m \mathbb{E}_P \|\xi\|_H^m \leq \frac{1}{2} m! (2L)^{m-2} 4\sigma^2 \ .$$

Since we consider the squared norm the assertion is a direct consequence of Caponnetto & De Vito (2007, Proposition 2) with $\eta = 2e^{-\tau}$, $L = 2L$, and $\sigma^2 = 4\sigma^2$. ■

The following Bernstein type inequality for Hilbert-Schmidt operator valued random variables is due to Minsker (2017). However we use a version from Lin & Cevher (2018a, Lemma 26), see also Tropp (2015) for an introduction to this topic.

Theorem 27 *Let $(\Omega, \mathcal{B}, P)$ be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to \mathcal{L}_2(H)$ be a random variable with values in the set of self-adjoint Hilbert-Schmidt operators. Furthermore, let the operator norm be P-a.s. bounded, i.e. $\|\xi\| \leq B$ P-a.s. and V be a self-adjoint positive semi-definite trace class operator with $\mathbb{E}_P(\xi^2) \preccurlyeq V$, i.e. $V - \mathbb{E}_P(\xi^2)$ is positive semi-definite. Then, for $g(V) := \log\big(2e \operatorname{tr}(V)/\|V\|\big)$, $\tau \geq 1$, and $n \geq 1$, the following concentration inequality is satisfied*

$$P^n \left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P \xi \right\| \geq \frac{4\tau B g(V)}{3n} + \sqrt{\frac{2\tau \|V\| \, g(V)}{n}} \right) \leq 2e^{-\tau} \ .$$

Recall that $\|V\|$ denotes the operator norm and tr the trace operator.

Proof This is a direct consequence of Lemma 26 from Lin & Cevher (2018a) with $\delta = 2e^{-\tau}$ applied to the centered random variable $\xi - \mathbb{E}_P \xi$. Furthermore, we used $\|\xi - \mathbb{E}_P \xi\| \leq 2B$ and $\mathbb{E}_P(\xi - \mathbb{E}_P \xi)^2 \preccurlyeq \mathbb{E}_P(\xi^2) \preccurlyeq V$. Finally, β defined by Lin & Cevher (2018a, Lemma 26) can be bounded by

$$\beta := \log\left(\frac{4 \operatorname{tr}(V)}{\|V\| \delta} \right) = \log\left(\frac{2 \operatorname{tr}(V)}{\|V\|} \right) + \tau \leq \tau g(V)$$

because of $\tau \geq 1$ and $\log\big(2 \operatorname{tr}(V)/\|V\|\big) > 0$. ■

References

R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

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

H. Bauer. *Probability Theory*. De Gruyter, Berlin, 1996.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18:971–1013, 2017.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7:331–368, 2007.

B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5:59–85, 2005a.

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, 2005b.

E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for Tikhonov regularization. *Anal. Appl. (Singap.)*, 4:81–99, 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. Stat.*, 11: 1022–1047, 2017.

J. Diestel and J. J. Uhl, Jr. *Vector Measures*. American Mathematical Society, Providence, 1977.

M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.*, 7:1–42, 2013.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108:203–227, 2018.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv e-prints*, 1702.07254v1, 2017.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, New York, 2002.

J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *arXiv e-prints*, 1801.07226v2, 2018a.

J. Lin and V. Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning*, 2018b.

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.

S. Mendelson and J. Neeman. Regularization in kernel learning. *Ann. Statist.*, 38:526–565, 2010.

S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statist. Probab. Lett.*, 127:111–119, 2017.

N. Mücke. Reducing training time by efficient localized kernel regression. In *Proceedings of Machine Learning Research*, pages 2603–2610. PMLR, 2019.

N. Mücke and G. Blanchard. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.*, 19:1–29, 2018.

N. Mücke, G. Neu, and L. Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems 32*, pages 12568–12577. Curran Associates, 2019.

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, 2018.

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory Probab. Appl.*, 30:143–148, 1986.

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, 2015.

S. Smale and D.-X. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc.*, 41:279–306, 2004.

S. Smale and D.-X. Zhou. Shannon sampling II: Connections to learning theory. *Appl. Comput. Harmon. Anal.*, 19:285–302, 2005.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26:153–172, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

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.

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.

H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland Publishing Co., Amsterdam, 1978.

J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8:1–230, 2015.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.