Universität Stuttgart

Sobolev Norm Learning Rates for Regularized Least-Squares Algorithm

Simon Fischer, Ingo Steinwart

Stuttgarter Mathematische Berichte

2017/003

mathematik

Fachbereich Mathematik
Fakultät Mathematik und Physik
Universität Stuttgart
Pfaffenwaldring 57
D-70 569 Stuttgart

E-Mail: preprints@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/preprints

ISSN **1613-8309**

Sobolev Norm Learning Rates for Regularized Least-Squares Algorithm

Simon Fischer and Ingo Steinwart

February 23, 2017

Institute for Stochastics and Applications
Faculty 8: Mathematics and Physics
University of Stuttgart
D-70569 Stuttgart Germany
{simon.fischer, ingo.steinwart}@mathematik.uni-stuttgart.de

Abstract

Learning rates for regularized least-squares algorithms are in most cases expressed with respect to the excess risk, or equivalently, the L_2-norm. For some applications, however, guarantees with respect to stronger norms such as the L_∞-norm, are desirable. We address this problem by establishing learning rates for a continuous scale of norms between the L_2- and the RKHS norm. As a byproduct we derive L_∞-norm learning rates, and in the case of Sobolev RKHSs we actually obtain Sobolev norm learning rates, which may also imply L_∞-norm rates for some derivatives. In all cases, we do not need to assume the target function to be contained in the used RKHS. Finally, we show that in many cases the derived rates are minimax optimal.

1. Introduction

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ independently sampled from an unknown distribution P on $X \times Y$ with $Y \subseteq \mathbb{R}$, the goal of non-parametric least-squares regression is to estimate the conditional mean function $f_P^* : X \to \mathbb{R}$ given by $f_P^*(x) := \mathbb{E}(Y|X = x)$. There are various different algorithms for this regression problem, see e.g. [8], but in this paper we focus on regularized

least-squares algorithms, which are also known as least-squares support vector machines (LS-SVM), see e.g. [11].

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 H is a reproducing kernel Hilbert space (RKHS) over X and $\lambda > 0$ is the so called regularization parameter. Probably the most interesting theoretical challenge for this and many other algorithms is to establish bounds, either in expectation or in probability, for

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

Here, the most frequently considered norm is the $L_2(\nu)$-norm, where $\nu := P_X$ denotes the marginal distribution of P on X, since a simple calculation shows that this norm equals the square root of the least squares excess risk. From a practical point of view another highly interesting norm is the supremum norm $\|\cdot\|_\infty$. However, this norm is only rarely investigated, probably because of the associated technical challenges. Yet another norm considered in (2) is the RKHS norm $\|\cdot\|_H$, since on the one hand side it bounds the $\|\cdot\|_\infty$ as long as the kernel is bounded, and on the other hand its consideration is less challenging than directly dealing with $\|\cdot\|_\infty$. However, the price for this convenience is that, instead of $f_P^* \in L_\infty(\nu)$, one even needs to assume $f_P^* \in H$. In this paper, we address these shortcomings by establishing learning rates for a continuous scale of norms $\|\cdot\|_\gamma$ between $\|\cdot\|_{L_2(\nu)}$ and $\|\cdot\|_H$, which in many interesting cases dominate both $\|\cdot\|_\infty$ and Sobolev type norms even if $f_P^* \notin H$. As a consequence of the latter, we also obtain $\|\cdot\|_\infty$-estimates that include derivatives. Last but not least we show that our resulting $\|\cdot\|_\gamma$-learning rates are in many cases minimax optimal.

Before we describe our results in a bit more detail, let us quickly introduce these intermediate spaces. To this end let us assume in the following that the kernel k of H is bounded. Then the integral operator $T_\nu : L_2(\nu) \to L_2(\nu)$ associated to the kernel k is well-defined, positive semi-definite, self-adjoint and nuclear. In particular, the powers $T_\nu^{\gamma/2}$ are defined for $\gamma > 0$ and it has been shown in [12, Equation (36)] that its image $[H]_\nu^\gamma := \operatorname{ran} T_\nu^{\gamma/2}$ can be equipped with some norm $\|\cdot\|_{[H]_\nu^\gamma}$, see also page 5 where these spaces are called *power spaces*. In fact, [12, Equation (36)] also introduces $[H]_\nu^0$, but in this case the equation above only holds if $H \subseteq L_2(\nu)$ is dense. Furthermore, [12, Theorem 4.6] shows that $\|\cdot\|_{[H]_\nu^\gamma}$ is equivalent to the interpolation

norm $\|\cdot\|_{[L_2(\nu),H]_{\gamma,2}}$ of the real method for all $\gamma \in (0,1)$. As a consequence $[H]^\gamma_\nu$ is equipped with a Besov type norm if H is a Sobolev space and ν is close to the uniform distribution on a suitable domain, see Section 4 for details. Last but not least, we have $[H]^1_\nu = H$ if the embedding $H \to L_2(\nu)$ is injective.

As in most papers investigating bounds on (2) we consider the following two types of assumptions:

(i) *eigenvalue decay*: $\mu_i \preccurlyeq i^{-1/p}$ for some $p \in (0,1)$, where $(\mu_i)_{i\geq 1}$ denotes the eigenvalues of T_ν.

(ii) *source condition*: $f^*_P \in [H]^\beta_\nu$ for some $0 < \beta \leq 2$.

To the best of our knowledge all papers considering stronger norms than the $L_2(\nu)$-norm restrict their investigations to the source condition case $\beta \geq 1$. However, this is a strong assumption since it implies the usually unrealistic $f^*_P \in H$. The novelty of our results is, that they even hold in the case $\beta < 1$, where $f^*_P \in H$ is no longer necessary. Moreover, for $1 \leq \beta \leq 2$ our $[H]^\gamma_\nu$-learning rates generalize the best (and optimal) already known learning rates. Furthermore, our rates for $\beta < 1$ are still optimal in many cases if we additionally use the assumption

(iii) *embedding property*: $[H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$ for some $0 < \alpha \leq 1$, i.e. the power space $[H]^\alpha_\nu$ is continuously embedded into $L_\infty(\nu)$,

taken from Steinwart et al. [13]. To be more precise, we obtain optimality in the case $\alpha < \beta$, in which we have $f^*_P \in [H]^\beta_\nu \hookrightarrow [H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$. In addition note that the embedding property always holds for $\alpha = 1$ if k is a bounded kernel. Let us now compare our results to some results from the literature, see Table 1 for an overview. To this end, we assume $Y = [-M, M]$ for some $M > 0$ and that k is a bounded measurable kernel whose (separable) RKHS H is dense in $L_2(\nu)$. Note that these assumptions form the largest common ground under which all papers considered in Table 1 achieve learning rates. In order to complete this comparison, let us briefly emphasize the specialty of each paper. Steinwart et al. [13] consider clipped LS-SVMs with a generalized regularization term $\lambda\|f\|^q_H$ for $q \geq 1$. Furthermore, instead of $f^*_P \in [H]^\beta_P$ and $[H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$ they used slightly weaker assumptions. This paper provides the fastest $L_2(\nu)$-learning rates in the case $\beta \in (0,1]$. Smale and Zhou [9] additionally provide faster rates in the noise-less case. Caponnetto and De Vito [3] prove their rates also for the case of multidimensional output and also consider rates for the best approximation if H is not dense in $L_2(\nu)$ and $f^*_P \in L_2(\nu)\backslash\overline{H}^{L_2(\nu)}$. Finally, Blanchard and Mücke [2] prove their results for an entire family of spectral regularization methods, which contains LS-SVMs as a special case.

The rest of this paper is organized as follows: In Section 2 we introduce the concepts we need to formulate our main results in Section 3. The subsequent section discusses the consequences for

publication	assumptions			learning rates (exponent)			
	$f_P^* \in [H]_\nu^\beta$	$[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)$	$\mu_i \preccurlyeq i^{-\frac{1}{p}}$	$L_2(\nu)$	$[L_2(\nu), H]_{\gamma,2}$	H	$L_\infty(\nu)$
our results	$0 < \beta \leq 2$	$0 < \alpha \leq 1$	$0 < p \leq \alpha$	$\frac{\beta}{\max\{\beta,\alpha\}+p}$	$\frac{(\beta-\gamma)_+}{\max\{\beta,\alpha\}+p}$	$\frac{(\beta-1)_+}{\beta+p}$	$\frac{(\beta-\alpha)_+}{\beta+p}$
Steinwart et al. [13]	$0 < \beta \leq 1$	$0 < \alpha \leq 1$	$p = \alpha$	$\frac{\beta}{\beta+p}$	x	x	x
Smale and Zhou [9]	$0 < \beta \leq 2$	$\alpha = 1$	$p = 1$	$\frac{\beta}{\max\{\beta,1\}+1}$	x	$\frac{(\beta-1)_+}{\beta+1}$	$\frac{(\beta-1)_+}{\beta+1}$
Caponnetto and De Vito [3]	$1 \leq \beta \leq 2$	$\alpha = 1$	$0 \leq p < 1$	$\frac{\beta}{\beta+p}$	x	x	x
Blanchard and Mücke [2]	$1 \leq \beta \leq 2$	$\alpha = 1$	$0 < p \leq 1$	$\frac{\beta}{\beta+p}$	$\frac{(\beta-\gamma)_+}{\beta+p}$	$\frac{(\beta-1)_+}{\beta+p}$	$\frac{(\beta-1)_+}{\beta+p}$

Table 1: Learning rates obtained by different authors. For simplicity, we ignore possible $\log(n)$-terms and just compare the exponent $r > 0$ of the polynomial part n^{-r}. The symbol „x“ means that the corresponding situation is not covered in this paper.

the special case of Besov RKHSs. For these spaces we will also see that the embedding property is often automatically satisfied. Last but not least we derive learning rates with respect to the $C^j(X)$-norms. The proofs of the main results can be found in Section 5.

2. Preliminaries

Setting Let $(X, \mathcal{B})$ be a measurable space (the *input space*), $Y = \mathbb{R}$ (the *output space*) and P an *unknown* distribution on $X \times Y$ with $|P|_2 := \int_{X\times Y} y^2 \, \mathrm{d}P(x,y) < \infty$. Moreover, we label the marginal distribution of P on X as $\nu := P_X$ and assume that $(X, \mathcal{B})$ is ν-complete. Furthermore, we fix a regular conditional probability $\big(P(\cdot|x)\big)_{x\in X}$ of P, which exists according to Dudley [6, Theorem 10.2.1 and Theorem 10.2.2]. Then the conditional mean function is given by $f_P^* = \big[x \mapsto \int_Y y \; P(\mathrm{d}y|x)\big]_\nu$, where $[f]_\nu$ denotes the ν-equivalence class of a measurable function $f : X \to \mathbb{R}$.

RKHS vs. L_2 We fix a separable RKHS H on X with respect to a ($\mathcal{B} \otimes \mathcal{B}$-)measurable and bounded kernel k. Let us recall some basic facts about the interplay between H and $L_2(\nu)$ from Steinwart and Scovel [12]. According to [12, Lemma 2.2, Lemma 2.3] and [11, Theorem 4.27] the (not necessarily injective) embedding $I_\nu : H \to L_2(\nu),\ f \mapsto [f]_\nu$ is well-defined, Hilbert-Schmidt

and the Hilbert-Schmidt norm fulfills

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

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

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

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

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

These operators are trace class and the trace norm is given by $\|T_\nu\|_{\mathcal{L}_1(L_2(\nu))} = \|C_\nu\|_{\mathcal{L}_1(H)} = \|I_\nu\|^2_{\mathcal{L}_2(H,L_2(\nu))} = \|S_\nu\|^2_{\mathcal{L}_2(L_2(\nu),H)}$. Please note that the operators I_ν, S_ν, T_ν and C_ν also depend on the RKHS H although this is not reflected in the notation. The spectral theorem for compact operators yields an at most countable index set $\mathcal{I} = \{1, 2, \ldots, N\}$ with $N \in \mathbb{N}_0$ resp. $\mathcal{I} = \mathbb{N}$, a positive, decreasing sequence $(\mu_i)_{i\in I} \in \ell_1(\mathcal{I})$ (i.e. it is summable) and a family $(e_i)_{i\in I} \subseteq H$, such that $(\mu_i^{1/2}\, e_i)_{i\in\mathcal{I}}$ is an ONS in H and $([e_i]_\nu)_{i\in\mathcal{I}}$ is an ONS in $L_2(\nu)$ with

$$C_\nu = \sum_{i\in\mathcal{I}} \mu_i \langle\,\cdot\,, \mu_i^{1/2}\, e_i\rangle_H\, \mu_i^{1/2}\, e_i, \qquad \text{resp.} \qquad T_\nu = \sum_{i\in\mathcal{I}} \mu_i \langle\,\cdot\,, [e_i]_\nu\rangle_{L_2(\nu)} [e_i]_\nu, \tag{3}$$

see Steinwart and Scovel [12, Lemma 2.12] for details.

Power Spaces Let us recall some intermediate spaces introduced in Steinwart and Scovel [12, remark after Proposition 4.2]. We call them *power spaces*. For $\alpha \geq 0$ the *α-power space* is given by

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

and equipped with the *α-power space norm*

$$\left\|\sum_{i\in\mathcal{I}} a_i \mu_i^{\alpha/2} [e_i]_\nu\right\|_{[H]^\alpha_\nu} := \left\|(a_i)_{i\in\mathcal{I}}\right\|_{\ell_2(\mathcal{I})}$$

for $(a_i)_{i\in\mathcal{I}} \in \ell_2(\mathcal{I})$. In the special case $\alpha = 1$ we introduce the abbreviation $[H]_\nu := [H]^1_\nu$. Let us summarize some basic facts about these spaces: Since for every $f \in [H]^\alpha_\nu$ there exist a unique

sequence $(a_i)_{i\in\mathcal{I}} \in \ell_2(\mathcal{I})$ with $f = \sum_{i\in\mathcal{I}} a_i \mu_i^{\alpha/2}[e_i]_\nu$ such that this series converges unconditionally in $[H]^\alpha_\nu$, the α-power norm is well-defined. Furthermore, $[H]^\alpha_\nu$ is a separable Hilbert space with ONB $(\mu_i^{\alpha/2}[e_i]_\nu)_{i\in\mathcal{I}}$ and for $0 \leq \beta < \alpha$ the embedding $[H]^\alpha_\nu \hookrightarrow [H]^\beta_\nu$ exists and is compact. Recall, that $[H]^1_\nu = \operatorname{ran} I_\nu$ and $[H]^0_\nu = \overline{\operatorname{ran} I_\nu}^{L_2(\nu)}$ with $\|\cdot\|_{[H]^0_\nu} = \|\cdot\|_{L_2(\nu)}$ holds. In the case $0 < \alpha < 1$ the α-power spaces are characterized by

$$[H]^\alpha_\nu \cong [L_2(\nu), [H]_\nu]_{\alpha,2}. \tag{4}$$

This mean that these sets coincide and the corresponding norms are equivalent. Furthermore, we can choose constants in the norm equivalence that depend only on α. For details see Steinwart and Scovel [12, Theorem 4.6], whereby the dependency of the constants is not a part of this statement, but it is contained in the proof of that theorem. Finally remark, that the interpolation space $[L_2(\nu), [H]_\nu]_{\alpha,2}$ (of the real method) is for some measures ν and RKHSs H well-known from the literature.

3. Assumptions and Results

In this section we present the assumptions and results of this work and discuss their consequences.

Assumptions Here we define the set of probability measures P on $X \times Y$ which are considered in our main results. Let us start with the set of considered marginal distributions on X.

3.1 Assumption (Probability measures on X)

Let H be a separable RKHS on X with respect to a bounded and measurable kernel k. Furthermore, let $A, C > 0$ be some constants and $0 < p \leq \alpha \leq 1$ be some parameters. By $\mathfrak{N}_{H,\alpha,p} := \mathfrak{N}_{H,A,C,\alpha,p}$ we denote the set of all probability measures ν on X with the following properties:

(i) The measurable space $(X, \mathcal{B})$ is ν-complete.

(ii) The *embedding property* $[H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$ holds with $\|\operatorname{Id} : [H]^\alpha_\nu \to L_\infty(\nu)\| \leq A$.

(iii) The eigenvalues fulfill a *polynomial upper bound* of order $\frac{1}{p}$, i.e. $\mu_i \leq C\, i^{-1/p}$ for all $i \in \mathcal{I}$.

Furthermore, we introduce for a constant $c > 0$ and a parameter $0 < q \leq p$ the subset $\mathfrak{N}_{H,\alpha,p,q} := \mathfrak{N}_{H,A,C,c,\alpha,p,q} \subseteq \mathfrak{N}_{H,A,C,\alpha,p}$ of probability measures ν on X which additionally have the following property:

(iv) The eigenvalues fulfill a *polynomial lower bound* of order $\frac{1}{q}$, i.e. $c\, i^{-1/q} \leq \mu_i$ for all $i \in \mathcal{I}$.

The condition $p \leq \alpha$ is not restrictive because the existence of the embedding $[H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$ already implies a polynomial upper bound of order $\frac{1}{\alpha}$ for the eigenvalues (see the Paragraph *L_∞-Embedding* in Section 5). Thus we are just interested in tightenings of the eigenvalue decay. Although we omit the constants A, C, c in the notation and just write $\mathfrak{N}_{H,\alpha,p}$ resp. $\mathfrak{N}_{H,\alpha,p,q}$, this sets are provided with some fixed constants $A, C, c > 0$.

3.2 Assumption (Probability measures on $X \times Y$)

Let H be a separable RKHS on X with respect to a bounded and measurable kernel k and $\mathfrak{N}$ a set of probability measures on X. Furthermore, let $B, B_\infty, L, \sigma > 0$ be some constants and $0 < \beta \leq 2$ a parameter. Then we denote by $\mathfrak{P}_{H,\beta}(\mathfrak{N}) := \mathfrak{P}_{H,B,B_\infty,L,\sigma,\beta}(\mathfrak{N})$ the set of all probability measures P on $X \times Y$ with the following properties:

(i) $\nu := P_X \in \mathfrak{N}$,

(ii) $|P|_2^2 := \int_{X\times Y} y^2 \, \mathrm{d}P(x,y) < \infty$,

(iii) $f^*_P \in L_\infty(\nu) \cap [H]^\beta_\nu$ with $\|f^*_P\|^2_{L_\infty(\nu)} \leq B_\infty$ and $\|f^*_P\|^2_{[H]^\beta_\nu} \leq B$,

(iv) $\int_Y |y - f^*_P(x)|^m \, P(\mathrm{d}y|x) \leq \frac{1}{2} m! \, \sigma^2 \, L^{m-2}$ for ν-almost all $x \in X$ and all $m \geq 2$.

Condition (iv) holds for Gaussian noise with bounded variance, i.e. $P(\cdot|x) = \mathcal{N}(f^*_P(x), \sigma_x^2)$, where $x \mapsto \sigma_x \in (0,\infty)$ is a measurable and ν-a.s. bounded function. Another sufficient condition is that P is concentrated on $X \times [-M, M]$ for some constant $M > 0$, i.e. $P(X \times [-M, M]) = 1$. In most cases we use $\mathfrak{N} = \mathfrak{N}_{H,\alpha,p}$ or $\mathfrak{N} = \mathfrak{N}_{H,\alpha,p,q}$, so we introduce the abbreviations $\mathfrak{P}_{H,\beta,\alpha,p} := \mathfrak{P}_{H,\beta}(\mathfrak{N}_{H,\alpha,p})$ and $\mathfrak{P}_{H,\beta,\alpha,p,q} := \mathfrak{P}_{H,\beta}(\mathfrak{N}_{H,\alpha,p,q})$.

Upper Rates The next theorem is the main result and contains our $[H]^\gamma_\nu$-learning rates.

3.3 Theorem ($[H]^\gamma_\nu$-Learning Rates)

Let H be a separable RKHS on X with respect to a bounded and measurable kernel k, $0 < p \leq \alpha \leq 1$, $0 \leq \gamma \leq 1$, $0 \leq \gamma < \beta \leq 2$ and $\lambda = (\lambda_n)_{n\geq 1}$ a sequence of regularization parameters. Then the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to H fulfills

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathfrak{P}_{H,\beta,\alpha,p}} P^n\Big(D : \ \big\|[f_{D,\lambda_n}]_\nu - f^*_P\big\|^2_{[H]^\gamma_\nu} > \tau a_n\Big) = 0$$

if one of the following two conditions hold:

(i) $\beta \leq \alpha$, $\lambda_n \asymp \big(\frac{\log(n)}{n}\big)^{\frac{1}{\alpha+p}}$ and $a_n = \big(\frac{\log(n)}{n}\big)^{\frac{\beta-\gamma}{\alpha+p}}$,

(ii) $\beta > \alpha$, $\lambda_n \asymp \left(\frac{1}{n}\right)^{\frac{1}{\beta+p}}$ *and* $a_n = \left(\frac{1}{n}\right)^{\frac{\beta-\gamma}{\beta+p}}$.

In the following we call such sequences $(a_n)_{n\geq 1}$ *upper rates* or *learning rates*. Obviously, every sequence $(\tilde{a}_n)_{n\geq 1}$ which decreases at most with the speed of $(a_n)_{n\geq 1}$ (i.e. $a_n = \mathcal{O}(\tilde{a}_n)$) is also an upper rate for and on every smaller set of probability measures $\mathfrak{P} \subseteq \mathfrak{P}_{H,\beta,\alpha,p}$ at least the same learning rate is achieved. Recall, since $\|\cdot\|_{[H]^\gamma_\nu} = \|\cdot\|_{L_2(\nu)}$ holds for $\gamma = 0$, our upper rates contain the special case of the $L_2(\nu)$-norm, witch coincides with the LS-excess-risk (see the Paragraph *LS-SVM* in Section 5). Because of the assumed embedding $[H]^\alpha_\nu \hookrightarrow L_\infty(\nu)$, the following corollary is a direct consequence of the $[H]^\gamma_\nu$-learning rates in the case $\gamma = \alpha$.

3.4 Corollary ($L_\infty(\nu)$-Learning Rates)

Let H be a separable RKHS on X with respect to a bounded and measurable kernel k, $0 < p \leq \alpha \leq 1$, $0 < \alpha < \beta \leq 2$ and $\lambda = (\lambda_n)_{n\geq 1}$ a sequence of regularization parameters.Then the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to H fulfills

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathfrak{P}_{H,\beta,\alpha,p}} P^n\Big(D:\ \big\|[f_{D,\lambda_n}]_\nu - f_P^*\big\|^2_{L_\infty(\nu)} > \tau a_n\Big) = 0$$

if $\lambda_n \asymp \left(\frac{1}{n}\right)^{\frac{1}{\beta+p}}$ and $a_n = \left(\frac{1}{n}\right)^{\frac{\beta-\alpha}{\beta+p}}$ holds.

Remark, that in the case $0 < \alpha < \beta < 1$ we get $L_\infty(\nu)$-learning rates even though the conditional mean function f_P^* do not have to lie in the RKHS.

Lower Rates In order to investigate the optimality of our learning rates the next theorem yields lower rates for the minimax probabilities.

3.5 Theorem ($[H]^\gamma_\nu$-Minimax Lower Rates)

Let H be a separable RKHS on X with respect to a bounded and measurable kernel k, $0 < q \leq p \leq \alpha \leq 1$, $0 \leq \gamma \leq 1$, $0 \leq \gamma < \beta \leq 2$, such that $\mathfrak{N}_{H,\alpha,p,q}$ is not empty. Then it holds

$$\lim_{\tau\to 0^+} \liminf_{n\to\infty} \inf_{D\mapsto f_D} \sup_{P\in\mathfrak{P}_{H,\beta,\alpha,p,q}} P^n\Big(D:\ \big\|[f_D]_\nu - f_P^*\big\|^2_{[H]^\gamma_\nu} > \tau b_n\Big) = 1$$

for $b_n = \left(\frac{1}{n}\right)^{\frac{\max\{\alpha,\beta\}-\gamma}{\max\{\alpha,\beta\}+q-\gamma(1-\frac{q}{p})}}$. The infimum is taken over all measurable learning methods with respect to $\mathfrak{P}_{H,\beta,\alpha,p,q}$ and γ, i.e. maps $(X\times Y)^n \to \{f : X \to Y \text{ measurable}\}$, $D \mapsto f_D$ such that $(X\times Y)^n \to [0,\infty]$, $D \mapsto \|[f_D]_\nu - f_P^*\|_{[H]^\gamma_\nu}$ is for all $P \in \mathfrak{P}_{H,\beta,\alpha,p,q}$ measurable with respect to the universal completion of the product-σ-algebra.

In the following we call such sequences $(b_n)_{n\geq 1}$ *(minimax) lower rates*. Obviously, every sequence $(\tilde{b}_n)_{n\geq 1}$ which decreases at least with the same speed as $(b_n)_{n\geq 1}$ (i.e. $\tilde{b}_n = \mathcal{O}(b_n)$) is also a lower rate for this set of probability measures and on every larger set of probability measures $\mathfrak{P} \supseteq \mathfrak{P}_{H,\beta,\alpha,p,q}$ at least the same lower rate holds. The meaning of a $[H]^\gamma_\nu$-lower rate $(b_n)_{n\geq 1}$ is, that no measurable learning method can fulfill a $[H]^\gamma_\nu$-learning rate $(a_n)_{n\geq 1}$ in the sense of Theorem 3.3 that decreases faster than $(b_n)_{n\geq 1}$ (i.e. $a_n = o(b_n)$). In the case $q = p$ and $\alpha < \beta$ the $[H]^\gamma_\nu$-learning rates of LS-SVMs stated in Theorem 3.3 coincide with the $[H]^\gamma_\nu$-minimax lower rates from Theorem 3.5 and therefore are optimal in the $[H]^\gamma_\nu$-minimax sense.

Discussion Recall that for $\gamma = 0$, the same $[H]^\gamma_\nu$-upper and lower rates and thereby optimal rates are established in the publication [3], but only for the case $\alpha = 1 < \beta$, and [2] extend these optimal rates to all $\gamma \in [0, 1]$. In other words, we further generalize these results to the case $\alpha < \beta \leq 1$, in which the conditional mean function f^*_P does not have to be in the RKHS. Unfortunately, for $\beta \leq \alpha$ our lower and upper rates do no longer match, nonetheless they improve the results from [9]. To be more precise, for $\gamma = 0$ and $\beta \leq \alpha$, [9] only obtained the upper rates of Theorem 3.3 for automatically satisfied case $p = \alpha = 1$, and therefore our rates are faster whenever $p < \alpha \leq 1$ or $p \leq \alpha < 1$ holds. Similarly, we improve the rates of [9] for $\gamma = 1$ and $\beta > 1 = \alpha = p$ whenever we actually have $p < 1$. Finally, the only case, in which our rates are worse than the best known rates is for $\beta \leq \alpha = p$ and $\gamma = 0$. In this case, the best known upper rates $a^*_n := \left(\frac{1}{n}\right)^{\frac{\beta}{\beta+p}}$, which were proven in [13], do not match our lower rates either. Namely, for $p = q$ we have

$$\text{our lower rate} = \left(\frac{1}{n}\right)^{\frac{\alpha}{\alpha+p}} \leq \left(\frac{1}{n}\right)^{\frac{\beta}{\beta+p}} \leq \left(\frac{\log(n)}{n}\right)^{\frac{\beta}{\alpha+p}} = \text{our upper rate}\,.$$

Consequently, the following questions remain open: Is the exponent $\frac{\beta}{\beta+p}$ optimal in the case $\beta \leq \alpha = p$, $\gamma = 0$? Can the techniques from [13] be adapted to improve the $[H]^\gamma_\nu$-upper rates for $\gamma > 0$ in the case $\beta \leq \alpha = p$? And last but not least, are our $L_\infty(\nu)$-rates optimal?

4. Example: Besov RKHSs

In this section we consider the specific case of Besov RKHSs. To this end we make the following assumptions: Let $X \subseteq \mathbb{R}^d$ be a non-empty open, connected and bounded set with a C_∞-boundary and equipped with the Lebesgue σ-algebra. Furthermore, we denote by μ the d-dimensional Lebesgue measure on X and by $L_2(X) := L_2(\mu)$ the corresponding L_2-space.

Introduction We briefly introduce the Sobolev and Besov spaces. To this end we follow the lines of Adams and Fournier [1, Chapter 7 Besov Spaces]. Because we are only interested in Hilbert space, we restrict ourself to this special cases. For $m \in \mathbb{N}$ the *Sobolev space* $W_m(X)$ is given by the linear space

$$W_m(X) := \{f \in L_2(X) : \ \partial^\alpha f \in L_2(X) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq m\}$$

equipped with the norm $\|f\|^2_{W_m(X)} := \sum_{|\alpha|\leq m} \|\partial^\alpha f\|^2_{L_2(X)}$. For $r > 0$ we define the *Besov space* $B_r(X)$ by means of the real interpolation method, namely $B_r(X) := [L_2(X), W_m(X)]_{\frac{r}{m},2}$, where $m := \min\{k \in \mathbb{N} : \ k > r\}$. A consequence of the reiteration property of the real interpolation method is

$$B_r(X) \cong [L_2(X), B_t(X)]_{\frac{r}{t},2} \tag{5}$$

for all $t > r > 0$. It is well-known that the Besov spaces $B_r(X)$ are separable Hilbert spaces with

$$B_r(X) \hookrightarrow C_j(X) \tag{6}$$

for $r > j + \frac{d}{2}$. Here $C_j(X)$ denotes the space of j-times continuous differentiable bounded functions with bounded derivatives. Therefore we can define the *Besov RKHS*

$$H_r(X) := \{f \in C_0(X) : \ [f]_\mu \in B_r(X)\}$$

for $r > \frac{d}{2}$ and equip this space with the norm $\|f\|_{H_r(X)} := \|[f]_\mu\|_{B_r(X)}$. The Besov RKHS is a separable RKHS with respect a kernel k_r since this space is isometric isomorph to $B_r(X)$ and this space is embedded into $C_0(X)$. Moreover, k_r is bounded and measurable, for details see Steinwart and Christmann [11, Lemma 4.28 and Lemma 4.25]. To describe the power spaces of $H_r(X)$ with respect to a probability measure ν on X we restrict ourself to the following set of measures.

4.1 Assumption (Probability measures on X for Besov RKHSs)
Let $G > 0$ be a constant with $G^{-1} \leq \mu(X) \leq G$. Then we denote by $\mathfrak{N}_{X,\mu} := \mathfrak{N}_{X,G,\mu}$ the set of all probability measures ν on X with $\nu \ll \mu$, $\mu \ll \nu$ such that $G^{-1} \leq \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \leq G$ holds ν-a.s.

Using Equation (4), the interpolation property and Equation (5) yield

$$[H_r(X)]^{u/r}_\nu \cong [L_2(\nu), [H_r(X)]_\nu]_{\frac{u}{r},2} \cong [L_2(X), [H_r(X)]_\mu]_{\frac{u}{r},2} \cong B_u(X) \tag{7}$$

for $\nu \in \mathfrak{N}_{X,\mu}$ and $r > u > 0$, where the constants of the overall norm equivalence can be chosen just depending on G, u, r and the underlying geometry of X.

4.2 Lemma (Comparison - Probability measures on X)
For all constants $G^{-1} \leq \mu(X) \leq G$ for the set $\mathfrak{N}_{X,\mu}$ and all parameters $r > \frac{d}{2}$, $1 > \alpha > \frac{d}{2r}$ and $p = q = \frac{d}{2r}$ there are constants $A, C, c > 0$ for the set $\mathfrak{N}_{H_r(X),\alpha,p,q}$ such that $\mathfrak{N}_{X,\mu} \subseteq \mathfrak{N}_{H_r(X),\alpha,p,q}$ holds.

Proof. Recall that, $H_r(X)$ is a separable RKHS with respect to a measurable and bounded kernel k_r. Let $\nu \in \mathfrak{N}_{X,\mu}$. Due to $1 > \alpha > \frac{d}{2r}$ it holds $r > \alpha r > \frac{d}{2}$, and therefore Equation (7) and (6) yield $[H_r(X)]^\alpha_\nu \cong B_{\alpha r}(X) \hookrightarrow C_0(X) \hookrightarrow L_\infty(\nu)$. Now the eigenvalues of T_ν (with respect to k_r) equal the squares of the approximation numbers of $I_\nu : H_r(X) \to L_2(\nu)$ (see Carl and Stephani [4, Equation (4.4.12)] and Steinwart [10, Section 2 and 3]). Because of Edmunds and Triebel [7, p. 119] (see also the discussion around Steinwart [10, Equation (37)]) these eigenvalues $(\mu_i)_{i\in\mathcal{I}}$ behave asymptoticly like $\mu_i \asymp i^{-\frac{2r}{d}}$. In both cases the constants can be chosen just dependent on G, α, r and the underlying geometry of X. Thus we can choose $A, C, c > 0$ such that the assertion holds. □

4.3 Assumption (Probability measures on $X \times Y$ for Besov RKHSs)
Let $\mathfrak{N}$ be a set of probability measures on X. Furthermore, let $E, B_\infty, L, \sigma > 0$ be some constants and $s > 0$ a parameter. Then we denote by $\mathfrak{P}_{X,s}(\mathfrak{N}) := \mathfrak{P}_{X,E,B_\infty,L,\sigma,s}(\mathfrak{N})$ the set of all probability measures P on $X \times Y$ with the following properties:

(i) $\nu := P_X \in \mathfrak{N}$,

(ii) $|P|_2 < \infty$,

(iii) $f^*_P \in L_\infty(\mu) \cap B_s(X)$ with $\|f^*_P\|_{L_\infty(\mu)} \leq B_\infty$ and $\|f^*_P\|_{B_s(X)} \leq E$

(iv) $\int_Y |y - f^*_P(x)|^m \, P(\mathrm{d}y|x) \leq \frac{1}{2} m! \, \sigma^2 \, L^{m-2}$ for ν-almost all $x \in X$ and all $m \geq 2$.

In most cases we use $\mathfrak{N} = \mathfrak{N}_{X,\mu}$, so we introduce the abbreviation $\mathfrak{P}_{X,s} := \mathfrak{P}_{X,s}(\mathfrak{N}_{X,\mu})$.

4.4 Lemma (Comparison - Probability measures on $X \times Y$)
For all parameters $r > \frac{d}{2}$, $r > s > 0$, $\beta = \frac{s}{r}$ and all constants $E, B_\infty, L, \sigma > 0$ for the set $\mathfrak{P}_{X,s}$ there is a constant $B > 0$ such that $\mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{X,\mu}) \subseteq \mathfrak{P}_{X,s}$ holds with respect to the constants B, B_∞, L, σ. Furthermore, there is another constant $B > 0$ such that the inverse inclusion $\mathfrak{P}_{X,s} \subseteq \mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{X,\mu})$ holds.

Proof. We just have to compare Assumption (iii). But this is a direct consequence of Equation (7) and $L_\infty(\mu) = L_\infty(\nu)$ for $\nu \in \mathfrak{N}_{X,\mu}$. □

Upper Rates In order to obtain learning rates in the Besov setting we exploit Theorem 3.3 with the help of Lemma 4.2 and Lemma 4.4.

4.5 Theorem ($B_t(X)$-Learning Rates)
Let $r > \frac{d}{2}$, $r > s > t$ and $\lambda = (\lambda_n)_{n\geq 1}$ a sequence of regularization parameters. Then the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to $H_r(X)$ fulfills

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathfrak{P}_{X,s}} P^n\Big(D:\ \|[f_{D,\lambda_n}]_\nu - f_P^*\|^2_{B_t(X)} > \tau a_n\Big) = 0$$

if one of the following two conditions hold:

(i) $s \leq \frac{d}{2}$, $\lambda_n \asymp \big(\frac{\log(n)}{n}\big)^{\frac{2r}{2d+\varepsilon}}$ *and* $a_n = \big(\frac{\log(n)}{n}\big)^{\frac{2s-2t}{2d+\varepsilon}}$ *for some* $0 < \varepsilon < 2r - d$,

(ii) $s > \frac{d}{2}$, $\lambda_n \asymp \big(\frac{1}{n}\big)^{\frac{2r}{2s+d}}$ *and* $a_n = \big(\frac{1}{n}\big)^{\frac{2s-2t}{2s+d}}$.

Proof. We set $p := q := \frac{d}{2r}$, $\beta := \frac{s}{r}$ and $\gamma := \frac{t}{r}$. For $s \leq \frac{d}{2}$ we choose $\alpha := \frac{d+\varepsilon}{2r}$ and for $s > \frac{d}{2}$ we choose $\frac{s}{r} > \alpha > \frac{d}{2r}$. According to Lemma 4.4 and Lemma 4.2 there are constants for the set $\mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{X,\mu})$ resp. for the set $\mathfrak{N}_{H_r(X),\alpha,p,q}$ such that $\mathfrak{P}_{X,s} \subseteq \mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{X,\mu}) \subseteq \mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{H_r(X),\alpha,p})$ holds. Hence the assertion is a consequence of Theorem 3.3 combined with Equation (7). □

Because of Equation (6), the following corollary is a consequence of Theorem 4.5 with $j+\frac{d}{2} < t < s$.

4.6 Corollary ($C_j(X)$-Learning Rates)
Let $j \geq 0$ be a non-negative integer, $r > s > j + \frac{d}{2}$ and $\lambda = (\lambda_n)_{n\geq 1}$ a sequence of regularization parameters. Then the LS-SVM $D \mapsto f_{D,\lambda_n}$ with respect to $H_r(X)$ fulfills

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathfrak{P}_{X,s}} P^n\Big(D:\ \|f_{D,\lambda_n} - f_P^*\|^2_{C_j(X)} > \tau a_n\Big) = 0$$

if $\lambda_n \asymp \big(\frac{1}{n}\big)^{\frac{2r}{2s+d}}$ and $a_n = \big(\frac{1}{n}\big)^{\frac{2s-2j-d}{2s+d}-\varepsilon}$ for some $0 < \varepsilon < \frac{2s-2j-d}{2s+d}$. Here f_P^* also denotes the unique continuous representative of the ν-equivalence class f_P^*.

Lower Rates In order to obtain minimax lower rates in the Besov setting we adapt the proof of Theorem 3.5 with the help of Lemma 4.4.

4.7 Theorem ($B_t(X)$-Minimax Lower Rates)
Let $r > \frac{d}{2}$, $r > s > t$. Then it holds

$$\lim_{\tau\to 0^+} \liminf_{n\to\infty} \inf_{D\mapsto f_D} \sup_{P\in\mathfrak{P}_{X,s}} P^n\Big(D:\ \|[f_D]_\nu - f_P^*\|^2_{B_t(X)} > \tau b_n\Big) = 1$$

if one of the following two conditions hold:

(i) $s \le \frac{d}{2}$ *and* $b_n = \big(\frac{1}{n}\big)^{\frac{d-2t}{2d}+\varepsilon}$ *for some sufficient small* $\varepsilon > 0$,

(ii) $s > \frac{d}{2}$ *and* $b_n = \big(\frac{1}{n}\big)^{\frac{2s-2t}{2s+d}}$.

The infimum in the above expression is taken over all measurable learning methods with respect to $\mathfrak{P}_{X,s}$ and t, i.e. maps $(X\times Y)^n \to \{f : X \to Y \text{ measurable}\}$, $D \mapsto f_D$ such that $(X\times Y)^n \to [0,\infty]$, $D \mapsto \|[f_D]_\nu - f_P^*\|_{B_t(X)}$ is for all $P \in \mathfrak{P}_{X,s}$ measurable with respect to the universal completion of the product-σ-algebra.

Proof. Because of a measurability issue we can not apply Theorem 3.5 and have to repeat the proof in the Besov setting. We set $p := q := \frac{d}{2r}$, $\beta := \frac{s}{r}$ and $\gamma := \frac{t}{r}$. For $s \le \frac{d}{2}$ we choose $1 > \alpha > \frac{d}{2r}$ sufficient small and for $s > \frac{d}{2}$ we choose $\frac{s}{r} > \alpha > \frac{d}{2r}$ arbitrary. As the constant G in the density bound of Assumption 4.1 is restricted to $G^{-1} \le \mu(X) \le G$, it holds $\nu := \frac{1}{\mu(X)}\mu \in \mathfrak{N}_{X,\mu}$ and according to Lemma 4.4 there are constants for the set $\mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{X,\mu})$ such that $\mathfrak{P}_{H_r(X),\beta}(\{\nu\}) \subseteq \mathfrak{P}_{H_r(X),\beta}(\mathfrak{N}_{X,\mu}) \subseteq \mathfrak{P}_{X,s}$ holds. Together with Equation (7) the proof remains a literally repetition of the proof of Theorem 3.5, so we omit the details. □

5. Proofs

First we summarize some well-known facts we need for the proof of our main results. To this end we use the notation from Section 2.

L_∞-Embedding Recall, that according to Steinwart and Scovel [12, Corollary 3.2] there exists a ν-zero set $N \subseteq X$, such that

$$k(x,x') = \sum_{i\in\mathcal{I}} \mu_i e_i(x) e_i(x') \tag{8}$$

holds for all $x, x' \in X\backslash N$ (because H is separable). Furthermore, the boundedness of k implies $\sum_{i\in\mathcal{I}} \mu_i e_i^2(x) \le A^2$ for ν-almost all $x \in X$ and a constant $A \ge 0$. 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

$$\sum_{i \in \mathcal{I}} \mu_i^\alpha e_i^2(x) \leq A^2 \tag{9}$$

for ν-almost all $x \in X$. Furthermore, by abuse of notation we write $\|k_\nu^\alpha\|_{L_\infty(\nu)}$ for the smallest constant with this property. Is there no such constant we set $\|k_\nu^\alpha\|_{L_\infty(\nu)} := \infty$. Thus we can just write $\|k_\nu^\alpha\|_{L_\infty(\nu)} < \infty$ as abbreviation of the phrase *the α-power of k is ν-a.s. bounded*. Because of the above introduction it always holds $\|k_\nu^1\|_{L_\infty(\nu)} < \infty$ for bounded kernels with separable RKHS. We recall the following theorem from Steinwart and Scovel [12, Theorem 5.3].

5.1 Theorem (L_∞-Embeddings)
Let $0 < \alpha \leq 1$ be a parameter. Then the following statements are equivalent:

(i) It holds $\|k_\nu^\alpha\|_{L_\infty(\nu)} < \infty$.

(ii) The embedding $\mathrm{Id} : [H]_\nu^\alpha \to L_\infty(\nu)$ is well-defined and continuous.

In this case it holds

$$\|\mathrm{Id} : [H]_\nu^\alpha \to L_\infty(\nu)\|_{\mathcal{L}([H]_\nu^\alpha, L_\infty(\nu))} = \|k_\nu^\alpha\|_{L_\infty(\nu)}. \tag{10}$$

Note that the claimed equality is not a part of Steinwart and Scovel [12, Theorem 5.3] but it is contained in the proof of that theorem. A further consequence of the ν-a.s. boundedness of the α-power is $\sum_{i \in \mathcal{I}} \mu_i^\alpha \leq \|k_\nu^\alpha\|_{L_\infty(\nu)}^2 < \infty$ (see [12, Proposition 4.4]) and the monotony of $(\mu_i)_{i \in \mathcal{I}}$ implies a polynomial decay of order $\frac{1}{\alpha}$. More precise, $\mu_i \leq \|k_\nu^\alpha\|_{L_\infty(\nu)}^{2/\alpha} i^{-1/\alpha}$ holds for all $i \in \mathcal{I}$.

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

$$\mathcal{N}_\nu(\lambda) := \mathrm{tr}\big((C_\nu + \lambda)^{-1} C_\nu\big) = \sum_{i \in \mathcal{I}} \frac{\mu_i}{\mu_i + \lambda}$$

for $\lambda > 0$, where tr denotes the trace operator. This quantity is widely used in the analysis of LS-SVMs and depends on the decay of the eigenvalues $(\mu_i)_{i \in \mathcal{I}}$. More precise, if there is a constant $C > 0$ and a parameter $0 < p \leq 1$, such that $\mu_i \leq C i^{-1/p}$ holds for all $i \in \mathcal{I}$ we get

$$\mathcal{N}_\nu(\lambda) \leq C_p \lambda^{-p} \tag{11}$$

for all $\lambda > 0$, where $C_p := \frac{C^p}{1-p}$ if $p < 1$ resp. $C_p = \|k\|_{L_2(\nu)}^2$ if $p = 1$. In the case $p < 1$ see Caponnetto and De Vito [3, Proposition 3] for details and for $p = 1$ this is a consequence of $\mathcal{N}_\nu(\lambda) = \mathrm{tr}((C_\nu + \lambda)^{-1} C_\nu) \leq \|(C_\nu + \lambda)^{-1}\|_{\mathcal{L}(H)} \|C_\nu\|_{\mathcal{L}_1(H)}$ together with $\|(C_\nu + \lambda)^{-1}\|_{\mathcal{L}(H)} \leq \lambda^{-1}$.

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

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

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^*. More precise, the *LS-excess-risk* is given by $\mathcal{R}_P(f) - \mathcal{R}_P^* = \|f - f_P^*\|_{L_2(\nu)}^2$ and minimizing the LS-risk is equivalent to approximating the conditional mean function in the $L_2(\nu)$-norm. For $\lambda > 0$ the minimization problem

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

is called *LS-SVM-problem* with respect to H, P and the regularization parameter $\lambda > 0$. Since

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

with $g_P := S_\nu f_P^*$ is the unique minimizer of the LS-SVM-problem, $f_{P,\lambda}$ is called *LS-SVM-solution* with respect to H, P and λ. Using the spectral decomposition from Equation (3) we get

$$f_{P,\lambda} = \sum_{i\in\mathcal{I}} \frac{\mu_i^{1/2}}{\mu_i + \lambda} \, a_i \, \mu_i^{1/2} e_i \in H, \qquad \text{and} \qquad f_P^* - [f_{P,\lambda}]_\nu = \sum_{i\in\mathcal{I}} \frac{\lambda}{\mu_i + \lambda} \, a_i \, [e_i]_\nu \tag{12}$$

for $a_i := \langle f_P^*, [e_i]_\nu \rangle_{L_2(\nu)}$ $(i \in \mathcal{I})$. Obviously, for the second identity we have to assume $f_P^* \in [H]_\nu^0$. Recall, that the predictor $f_{D,\lambda}$ for the dataset D defined in Equation (1) is the LS-SVM-solution with respect to the corresponding *empirical* measure, which is given by $D := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. As a consequence of Steinwart and Christmann [11, Theorem 6.23] the map LS-SVM $(X \times Y)^n \to H$, $D \mapsto f_{D,\lambda}$ is measurable with respect to the universal completion of product-σ-algebra on $(X \times Y)^n$. Hence we can measure the probability

$$P^n\Big(D \in (X\times Y)^n : \; \|[f_{D,\lambda}]_\nu - f_P^*\|_{[H]_\nu^\gamma} < \varepsilon \Big)$$

for $\lambda > 0$, $\varepsilon > 0$ and $0 \le \gamma \le 1$ if we extend P^n to the universal completion.

Upper Rates

Using the standard technique, we split the estimation of $\|[f_{D,\lambda}]_\nu - f_P^*\|^2_{[H]^\gamma_\nu}$ into two parts:

$$\|[f_{D,\lambda}]_\nu - f_P^*\|^2_{[H]^\gamma_\nu} \leq 2\|[f_{D,\lambda} - f_{P,\lambda}]_\nu\|^2_{[H]^\gamma_\nu} + 2\|[f_{P,\lambda}]_\nu - f_P^*\|^2_{[H]^\gamma_\nu},$$

the *estimation error* $\|[f_{D,\lambda} - f_{P,\lambda}]_\nu\|^2_{[H]^\gamma_\nu}$ and the *approximation error* $\|[f_{P,\lambda}]_\nu - f_P^*\|^2_{[H]^\gamma_\nu}$. First we consider the approximation error, which depends on the source condition.

5.2 Lemma (Approximation Error)

Let $0 \leq \beta \leq 2$, P a probability measure on $X \times Y$ and H a separable RKHS on X with respect to a bounded and measurable kernel k. If $f_P^* \in [H]^\beta_\nu$, then

$$\|[f_{P,\lambda}]_\nu - f_P^*\|^2_{[H]^\gamma_\nu} \leq \|f_P^*\|^2_{[H]^\beta_\nu} \lambda^{\beta-\gamma}$$

holds for all $\lambda > 0$ and all $0 \leq \gamma \leq \beta$.

Proof. The spectral representation from Equation (12) holds because of $f_P^* \in [H]^\beta_\nu \subseteq [H]^0_\nu = \overline{\operatorname{ran} I_\nu}$. Since $(\mu_i^{\gamma/2}[e_i]_\nu)_{i\in\mathcal{I}}$ is an ONB of $[H]^\gamma_\nu$ Parseval yields

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

If we estimate the fraction on the right hand side with Lemma A.1 and use the fact, that $(\mu_i^{\beta/2}[e_i]_\nu)_{i\in\mathcal{I}}$ is an ONB of $[H]^\beta_\nu$, we get

$$\left\|f_P^* - [f_{P,\lambda}]_\nu\right\|^2_{[H]^\gamma_\nu} \leq \lambda^{\beta-\gamma} \sum_{i\in\mathcal{I}} \mu_i^{-\beta} a_i^2 = \lambda^{\beta-\gamma} \|f_P^*\|^2_{[H]^\beta_\nu}. \qquad \square$$

The following oracle inequality controls the estimation error.

5.3 Theorem (Estimation Error - Oracle Inequality)

Let $0 \leq \alpha, \gamma \leq 1$ be some parameters, H a separable RKHS on X with respect to a bounded and measurable kernel k and P a probability measure on $X \times Y$ with $|P|_2 < \infty$. Furthermore, we assume

(i) the source condition*: $f_P^* \in L_\infty(\nu) \cap [H]^\gamma_\nu$,*

(ii) the embedding property*: $\|k_\nu^\alpha\|_{L_\infty(\nu)} < \infty$ and*

(iii) the moment condition*: there are some constants $\sigma, L > 0$ with*

$$\int_Y |y - f_P^*(x)|^m \, P(\mathrm{d}y|x) \leq \frac{1}{2} m! \, \sigma^2 \, L^{m-2}$$

for ν-almost all $x \in X$ and all $m \geq 2$.

Then for $\tau \geq 1$, $\lambda > 0$ and $n \geq A_{\lambda,\tau}$

$$\|[f_{D,\lambda} - f_{P,\lambda}]_\nu\|_{[H]^\gamma_\nu}^2 \leq 128 \frac{\tau^2}{n\lambda^\gamma} \left(5\mathcal{N}_{P_X}(\lambda)\sigma_\lambda^2 + \|k_{P_X}^\alpha\|_{L_\infty(P_X)}^2 \frac{L_\lambda^2}{n\lambda^\alpha} \right)$$

holds with P^n-probability $\geq 1 - 4e^{-\tau}$, where we set

(i) $A_{\lambda,\tau} := \max\{256\tau^2 \|k_\nu^\alpha\|_{L_\infty(\nu)}^2 \lambda^{-\alpha} \mathcal{N}_\nu(\lambda), \; 16\tau \|k_\nu^\alpha\|_{L_\infty(\nu)}^2 \lambda^{-\alpha}, \; \tau\}$,

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

We split the proof of this theorem into several lemmas.

5.4 Lemma (Power Norm on $\operatorname{ran} I_\nu$)

Under the assumptions of Theorem 5.3 it holds $\|[f]_\nu\|_{[H]^\gamma_\nu} \leq \big\|C_\nu^{\frac{1-\gamma}{2}} f\big\|_H$ for all $f \in H$.

Proof. We fix some $f \in H$. Because $(\mu_i^{1/2} e_i)_{i \in \mathcal{I}}$ is an ONB of $(\ker I_\nu)^\perp$, there is a $g \in \ker I_\nu$ with $f = \sum_{i \in \mathcal{I}} a_i \mu_i^{1/2} e_i + g$, where $a_i = \langle f, \mu_i^{1/2} e_i \rangle_H$ for all $i \in \mathcal{I}$. Parseval yields

$$\|[f]_\nu\|_{[H]^\gamma_\nu}^2 = \Big\| \sum_{i \in \mathcal{I}} a_i \mu_i^{\frac{1-\gamma}{2}} \mu_i^{\gamma/2} [e_i]_\nu \Big\|_{[H]^\gamma_\nu}^2 = \sum_{i \in \mathcal{I}} \mu_i^{1-\gamma} a_i^2,$$

because $(\mu_i^{\gamma/2}[e_i]_\nu)_{i \in \mathcal{I}}$ is an ONB of $[H]^\gamma_\nu$. If $\gamma < 1$ holds, then the spectral decomposition from Equation (3) yields

$$\|C_\nu^{\frac{1-\gamma}{2}} f\|_H^2 = \Big\| \sum_{i \in \mathcal{I}} \mu_i^{\frac{1-\gamma}{2}} a_i \mu_i^{1/2} e_i \Big\|_H^2 = \sum_{i \in \mathcal{I}} \mu_i^{1-\gamma} a_i^2,$$

where we used in the second equality again Parseval with respect to the ONS $(\mu_i^{1/2} e_i)_{i \in \mathcal{I}}$ of H. For $\gamma = 1$ we get $C_\nu^{\frac{1-\gamma}{2}} = \mathrm{Id}_H$ and

$$\|C_\nu^{\frac{1-\gamma}{2}} f\|_H^2 = \Big\| \sum_{i \in \mathcal{I}} a_i \mu_i^{1/2} e_i + g \Big\|_H^2 = \Big\| \sum_{i \in \mathcal{I}} a_i \mu_i^{1/2} e_i \Big\|_H^2 + \|g\|_H^2 \geq \sum_{i \in \mathcal{I}} a_i^2,$$

where we used $\sum_{i \in \mathcal{I}} a_i \mu_i^{1/2} e_i \perp g$ and again Parseval. □

Next we bring some $L_\infty(\nu)$ and $L_2(\nu)$ bounds forward, because we use them later several times.

5.5 Lemma (L_2 and L_∞ Bound)
Under the assumptions of Theorem 5.3 the following statements are true for all $\lambda > 0$:

(i) $\|(C_\nu + \lambda)^{-1/2} k(x,\cdot)\|_H^2 \le \|k_\nu^\alpha\|_{L_\infty(\nu)}^2 \lambda^{-\alpha}$ *for ν-almost all $x \in X$.*

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

Proof. Let $\lambda > 0$. Because we assume a separable RKHS H with respect to a measurable kernel the map $X \to H,\ x \mapsto k(x,\cdot)$ is measurable and thus also $X \to \mathbb{R},\ x \mapsto \|(C_\nu + \lambda)^{-1/2} k(x,\cdot)\|_H^2$ is measurable. Let us fix an arbitrary ONB $(e_j)_{j\in\mathcal{J}}$ of $\ker I_\nu$. Thus $(\mu_i^{1/2} e_i)_{i\in\mathcal{I}} \cup (e_j)_{j\in\mathcal{J}}$ is an ONB of H and it holds

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

for all $x \in X$. Together with the spectral decomposition from Equation (3) and Parseval we get

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

for all $x \in X$. Because H is separable the index set $\mathcal{J}$ is at most countable. Thus $e_j \in \ker I_\nu$ for all $j \in \mathcal{J}$ imply that the second summand on the right hand side vanishes for ν-almost all $x \in X$. Hence

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

holds for ν-almost all $x \in X$. In order to prove Statement (i) we use Lemma A.1.

$$\sum_{i\in\mathcal{I}} \frac{\mu_i}{\mu_i + \lambda} e_i^2(x) = \sum_{i\in\mathcal{I}} \frac{\mu_i^{1-\alpha}}{\mu_i + \lambda} \mu_i^\alpha e_i^2(x) \le \Big(\sum_{i\in\mathcal{I}} \mu_i^\alpha e_i^2(x)\Big) \sup_{i\in\mathcal{I}} \frac{\mu_i^{1-\alpha}}{\mu_i + \lambda} \le \|k_\nu^\alpha\|_{L_\infty(\nu)}^2 \lambda^{-\alpha}$$

for ν-almost all $x \in X$. To prove Statement (ii) we use the fact that $([e_i])_{i\in\mathcal{I}}$ is an ONS in $L_2(\nu)$ and the monotone convergence theorem

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

In the next two lemmas we prefer the more detailed notation P_X instead of ν for the marginal distribution of P on X to avoid misunderstandings.

5.6 Lemma (Oracle Inequality (Part I))
Let the assumption of Theorem 5.3 hold. For $\tau \geq 1$, $\lambda > 0$ and $n \geq A_{\lambda,\tau}$ the estimate

$$\|[f_{D,\lambda} - f_{P,\lambda}]\|^2_{[H]^\gamma_{P_X}} \leq \frac{4}{\lambda^\gamma} \Big\| (C_{P_X} + \lambda)^{-1/2} \big((g_D - C_{D_X} f_{P,\lambda}) - (g_P - C_{P_X} f_{P,\lambda}) \big) \Big\|^2_H$$

holds with P^n-probability $\geq 1 - 2e^{-\tau}$.

The following proof is a modification of Caponnetto and De Vito [3, proof of Theorem 4 (Step 2.1 and Step 3.1)]. We extend the applicability of this proof from the parameter rage $\gamma = 0$ and $\alpha = 1$ to $0 \leq \alpha, \gamma \leq 1$.

Proof. Let us fix a $\tau \geq 1$ $\lambda > 0$ and $n \geq A_{\lambda,\tau}$. For $D \in (X \times Y)^n$ Lemma 5.4 yields

$$\|[f_{D,\lambda} - f_{P,\lambda}]\|_{[H]^\gamma_{P_X}} \leq \|C^{\frac{1-\gamma}{2}}_{P_X} (f_{D,\lambda} - f_{P,\lambda})\|_H.$$

Using the representation $f_{D,\lambda} = (C_{D_X} + \lambda)^{-1} g_D$ we get

$$C^{\frac{1-\gamma}{2}}_{P_X} (f_{D,\lambda} - f_{P,\lambda}) = C^{\frac{1-\gamma}{2}}_{P_X} (C_{D_X} + \lambda)^{-1} (g_D - (C_{D_X} + \lambda) f_{P,\lambda}),$$

Together with the identitie $\mathrm{Id}_H = (C_{P_X} + \lambda)^{-1/2} (C_{P_X} + \lambda)^{1/2}$ it follows

$$\begin{aligned} \|[f_{D,\lambda} - f_{P,\lambda}]\|_{[H]^\gamma_P} \leq{}& \big\| C^{\frac{1-\gamma}{2}}_{P_X} (C_{P_X} + \lambda)^{-1/2} \big\|_{\mathcal{L}(H)} && \text{(13a)} \\ & \cdot \big\| (C_{P_X} + \lambda)^{1/2} (C_{D_X} + \lambda)^{-1} (C_{P_X} + \lambda)^{1/2} \big\|_{\mathcal{L}(H)} && \text{(13b)} \\ & \cdot \|(C_{P_X} + \lambda)^{-1/2} (g_D - (C_{D_X} + \lambda) f_{P,\lambda})\|_H && \text{(13c)} \end{aligned}$$

for all $D \in (X \times Y)^n$. Now we successive estimate the three factors on the right hand side. (13a) Because C_{P_X} is self-adjoint and positive semi-definite the spectrum of the operator is contained in $\sigma(C_{P_X}) \subseteq [0, \infty)$ and therefore it holds

$$(13a) = \big\| C^{\frac{1-\gamma}{2}}_{P_X} (C_{P_X} + \lambda)^{-1/2} \big\|_{\mathcal{L}(H)} = \sup_{t \in \sigma(C_{P_X})} \Big(\frac{t^{1-\gamma}}{t + \lambda} \Big)^{1/2} \leq \lambda^{-\gamma/2},$$

where we used Lemma A.1 in the last step. (13c) This term can be rearranged using the identity

$f_{P,\lambda} = (C_{P_X} + \lambda)^{-1} g_P$

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

Consequently we get

$$\begin{aligned}\|[f_{D,\lambda} - f_{P,\lambda}]\|_{[H]_P^\gamma} \leq\ & \frac{1}{\lambda^{\gamma/2}} \big\|(C_{P_X} + \lambda)^{1/2} (C_{D_X} + \lambda)^{-1} (C_{P_X} + \lambda)^{1/2}\big\|_{\mathcal{L}(H)} \\ & \cdot \big\|(C_{P_X} + \lambda)^{-1/2}\big((g_D - C_{D_X} f_{P,\lambda}) - (g_P - C_{P_X} f_{P,\lambda})\big)\big\|_H.\end{aligned} \tag{14}$$

for all $D \in (X \times Y)^n$ and it remains to estimate the Factor (13b) which is the main part of the proof. In order to estimate (13b) we start with the following identity

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

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

$$(13b) = \big\|\big(\mathrm{Id} - (C_{P_X} + \lambda)^{-1/2} (C_{P_X} - C_{D_X}) (C_{P_X} + \lambda)^{-1/2}\big)^{-1}\big\|_{\mathcal{L}(H)}.$$

Now we apply the Berstein inequality to estimate the norm of the operator $(C_{P_X} + \lambda)^{-1/2}(C_{P_X} - C_{D_X})(C_{P_X} + \lambda)^{-1/2}$ and afterwards we use the Neumann series to get an estimate for (13b). To this end we consider the random variable $\xi_1 : X \to \mathcal{L}_2(H)$,

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

with values in the space of Hilbert-Schmidt operators on H. Where $C_{\{x\}} : H \to H$ denotes the *integral* operator with respect to the empirical measure of the point $\{x\}$, i.e.

$$C_{\{x\}} f = f(x) k(x, \cdot) = \langle f, k(x, \cdot)\rangle_H \, k(x, \cdot).$$

Because the range of the operator $C_{\{x\}}$ is one dimensional it is especially a Hilbert-Schmidt operator. Since H is a separable RKHS with respect to a measurable and bounded kernel, the map $X \to \mathcal{L}_2(H),\ x \mapsto C_{\{x\}}$ is bounded and measurable. Moreover, the map $x \mapsto C_{\{x\}}$ and $x \mapsto \xi_1(x)$ are Bochner integrable with respect to a arbitrary probability measure μ on X and

Diestel and Uhl [5, Chapter II.2 Theorem 6] yields

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

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

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

for all $D = ((x_i, y_i))_{i=1}^n \in (X \times Y)^n$. Using the self-adjointness of $(C_{P_X} + \lambda)^{-1/2}$ we get $\xi_1(x) = \langle \cdot, g_x\rangle_H \, g_x$ for all $x \in X$ with $g_x := (C_{P_X} + \lambda)^{-1/2} k(x, \cdot)$. Applying Lemma 5.5 yields the supremum bound

$$\|\xi_1(x)\|_{\mathcal{L}_2(H)} = \|g_x\|_H^2 \leq \|k_{P_X}^\alpha\|_{L_\infty(P_X)}^2 \lambda^{-\alpha} =: L_1$$

for P_X-almost all $x \in X$ and the variance bound

$$\int_X \|\xi_1(x)\|_{\mathcal{L}_2(H)}^2 \, \mathrm{d}P_X(x) \leq L_1 \int_X \|(C_{P_X} + \lambda)^{-1/2} k(x, \cdot)\|_H^2 \, \mathrm{d}P_X(x) = L_1 \mathcal{N}_{P_X}(\lambda) =: \sigma_1^2.$$

The separability of H implies the separability of $\mathcal{L}_2(H)$ and hence the Bernstein inequality (Corollary A.3) is applicable. Therefore

$$\big\|(C_{P_X} + \lambda)^{-1/2}(C_{P_X} - C_{D_X})(C_{P_X} + \lambda)^{-1/2}\big\|_{\mathcal{L}_2(H)} \leq 4\tau\left(\sqrt{\frac{\sigma_1^2}{n}} + \frac{L_1}{n}\right)$$

holds with ${P_X}^n$-probability $\geq 1 - 2e^{-\tau}$. Because we have chosen $n \geq A_{\lambda,\tau} \geq \max\{(16\tau\sigma_1)^2, 16\tau L_1\}$ we get

$$4\tau\sqrt{\frac{\sigma_1^2}{n}} = \frac{4\tau\sigma_1}{\sqrt{n}} \leq \frac{4\tau\sigma_1}{16\tau\sigma_1} = \frac{1}{4} \qquad \text{and} \qquad 4\tau\frac{L_1}{n} \leq \frac{4\tau L_1}{16\tau L_1} = \frac{1}{4}.$$

Combining these estimates we get

$$\big\|(C_{P_X} + \lambda)^{-1/2}(C_{P_X} - C_{D_X})(C_{P_X} + \lambda)^{-1/2}\big\|_{\mathcal{L}_2(H)} \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

with ${P_X}^n$-probability $\geq 1 - 2e^{-\tau}$. Because the Hilbert-Schmidt norm dominates the operator norm the Neumann series is applicable and yields

$$(13\text{b}) = \big\|\big(\mathrm{Id} - (C_{P_X} + \lambda)^{-1/2}(C_{P_X} - C_{D_X})(C_{P_X} + \lambda)^{-1/2}\big)^{-1}\big\|_{\mathcal{L}(H)} \leq \sum_{k=0}^{\infty}\Big(\frac{1}{2}\Big)^k = 2$$

with ${P_X}^n$-probability $\geq 1 - 2e^{-\tau}$. Together with the inequality (14) the statement follows. □

5.7 Lemma (Oracle Inequality (Part 2))
Let the assumption of Theorem 5.3 hold. For $\tau \geq 1$, $\lambda > 0$ and $n \geq 1$ the estimate

$$\begin{aligned} &\left\|(C_{P_X} + \lambda)^{-1/2}\big((g_D - C_{D_X} f_{P,\lambda}) - (g_P - C_{P_X} f_{P,\lambda})\big)\right\|_H^2 \\ &\leq \frac{32\tau^2}{n}\left(5\mathcal{N}_{P_X}(\lambda)\sigma_\lambda^2 + \|k_{P_X}^\alpha\|_{L_\infty(P_X)}^2 \frac{L_\lambda^2}{n\lambda^\alpha}\right) \end{aligned}$$

holds with P^n-probability $\geq 1 - 2e^{-\tau}$.

Proof. Let us fix a $\tau \geq 1$ $\lambda > 0$ and $n \geq 1$. To prove this statement we define the random variable $\xi_2 : X \times Y \to H$,

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

Since H is a separable RKHS with respect to a bounded and measurable kernel and the moment condition holds, we get that ξ_2 and $(x, y) \mapsto (y - f_{P,\lambda}(x))k(x, \cdot)$ are measurable and Bochner integrable with respect to an arbitrary probability measure Q on $X \times Y$. Diestel and Uhl [5, Chapeter II.2 Theorem 6] yields

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

If we use this identity for the case $Q = P$ and $Q = D$ we get

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

To apply the Bernstein inequality (Theorem A.2) we need to bound the m-th moment for $m \geq 2$. Let us fix some $m \geq 2$. By the definition we get

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

First we consider the inner integral: Using the triangle inequality yields

$$\int_Y |y - f_{P,\lambda}(x)|^m \, P(\mathrm{d}y|x) \leq 2^{m-1}(\|\operatorname{id}_Y - f_P^*(x)\|_{L_m(P(\cdot|x))}^m + |f_P^*(x) - f_{P,\lambda}(x)|^m).$$

for P_X-almost all $x \in X$. Furthermore, it follows by the moment condition

$$\int_Y |y - f_{P,\lambda}(x)|^m \, P(\mathrm{d}y|x) \leq 2^{m-1}\Big(\frac{1}{2} m! \sigma^2 L^{m-2} + \|f_P^* - [f_{P,\lambda}]_{P_X}\|_{L_\infty(P_X)}^m\Big) \leq \frac{1}{2} \, m! \, (2\sigma_\lambda)^2 \, (2L_\lambda)^{m-2}$$

for P_X-almost all $x \in X$. If we plug this into the initial equation and use Lemma 5.5 we get

$$\begin{aligned}
\mathbb{E}_P \|\xi_2\|^m &\leq \frac{1}{2} \, m! \, (2\sigma_\lambda)^2 \, (2L_\lambda)^{m-2} \int_X \|(C_{P_X} + \lambda)^{-1/2} k(x, \cdot)\|_H^m P_X(x) \\
&\leq \frac{1}{2} \, m! \, \big(4\mathcal{N}_{P_X}(\lambda)\sigma_\lambda^2\big) \, \big(2\|k_{P_X}^\alpha\|_{L_\infty(P_X)} \lambda^{-\alpha/2} L_\lambda\big)^{m-2}.
\end{aligned}$$

Because H is separable the Bernstein inequality (Theorem A.2) is applicable and yields

$$\begin{aligned}
&\Big\|(C_{P_X} + \lambda)^{-1/2}\Big((g_D - C_{D_X} f_{P,\lambda}) - (g_P - C_{P_X} f_{P,\lambda})\Big)\Big\|_H^2 \\
\leq &4\tau^2 \left(\sqrt{\frac{20\mathcal{N}_{P_X}(\lambda)\sigma_\lambda^2}{n}} + \frac{2\|k_{P_X}^\alpha\|_{L_\infty(P_X)} \lambda^{-\alpha/2} L_\lambda}{n}\right)^2 \\
\leq &\frac{8\tau^2}{n}\left(20\mathcal{N}_{P_X}(\lambda)\sigma_\lambda^2 + 4\|k_{P_X}^\alpha\|_{L_\infty(P_X)}^2 \frac{L_\lambda^2}{n\lambda^\alpha}\right)
\end{aligned}$$

with P^n-probability $\geq 1 - 2e^{-\tau}$. □

Now the proof of Theorem 5.3 is just an application of Lemma 5.6 and Lemma 5.7. In order to simplify the statement in Theorem 5.3 under the assumption $P \in \mathfrak{P}_{H,\beta,\alpha,p}$ we need the next lemma.

5.8 Lemma

Let $P \in \mathfrak{P}_{H,\beta,\alpha,p}$ be a probability measure. Then there are constants $N, V > 0$ depending only on $\mathfrak{P}_{H,\beta,\alpha,p}$ such that $A_{\lambda,\tau} \leq N\frac{\tau^2}{\lambda^{p+\alpha}}$ and $L_\lambda^2, \sigma_\lambda^2 \leq \frac{V}{\lambda^{\max\{0,\alpha-\beta\}}}$ for all $0 < \lambda \leq 1$ and $\tau \geq 1$.

Proof. Let $P \in \mathfrak{P}_{H,\beta,\alpha,p}$, $0 < \lambda \leq 1$ and $\tau \geq 1$. From Equation (8) we get $\|k\|_{L_2(\nu)}^2 \leq \|k_\nu^1\|_{L_\infty(\nu)}^2 \leq \|k_\nu^\alpha\|_{L_\infty(\nu)}^2$ and Theorem 5.1 yields $\|k\|_{L_2(\nu)}^2 \leq \|k_\nu^\alpha\|_{L_\infty(\nu)}^2 \leq A$, since we have the embedding property. If we redefine $C_p := A$ in the case $p = 1$ then Equation (11) still holds and $N := \max\{256AC_p, 16A, 1\}$ is a possible constant. In order to prove the second inequality we distinguish

two cases: For $\beta \leq \alpha$ we proceed by

$$\|f_P^* - [f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}^2 \leq 2\|f_P^*\|_{L_\infty}^2 + 2\|[f_{P,\lambda}]_\nu\|_{L_\infty(\nu)}^2 \leq 2B_\infty + 2A\|[f_{P,\lambda}]_\nu\|_{[H]_\nu^\alpha}^2.$$

Using the spectral representation from Equation (12) and Parseval yields

$$\|[f_{P,\lambda}]_\nu\|_{[H]_\nu^\alpha}^2 = \Big\|\sum_{i\in\mathcal{I}} \frac{\mu_i^{1-\alpha/2}}{\mu_i+\lambda}\, a_i\, \mu_i^{\alpha/2}[e_i]_\nu\Big\|_{[H]_\nu^\alpha}^2 = \sum_{i\in\mathcal{I}}\Big(\frac{\mu_i^{1-\alpha/2}}{\mu_i+\lambda}\Big)^2 a_i^2 = \sum_{i\in\mathcal{I}}\Big(\frac{\mu_i^{1+\frac{\beta-\alpha}{2}}}{\mu_i+\lambda}\Big)^2 a_i^2\mu_i^{-\beta}.$$

If we estimate the fraction on the right hand side with Lemma A.1, then we get

$$\|[f_{P,\lambda}]_\nu\|_{[H]_\nu^\alpha}^2 \leq \lambda^{\beta-\alpha}\sum_{i\in\mathcal{I}} a_i^2\mu_i^{-\beta} = \lambda^{\beta-\alpha}\|f_P^*\|_{[H]_\nu^\beta}^2 \leq B\lambda^{-(\alpha-\beta)}.$$

Therefore it holds $L_\lambda^2, \sigma_\lambda^2 \leq V\lambda^{-(\alpha-\beta)}$ with $V := \max\{L^2, \sigma^2, 2B_\infty + 2AB\}$. In the case $\beta > \alpha$ we apply Lemma 5.2:

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

Hence it holds $L_\lambda^2, \sigma_\lambda^2 \leq \max\{L^2, \sigma^2, AB\} \leq V$. □

With this preparations we simplify the estimation for the measures in the set $\mathfrak{P}_{H,\beta,\alpha,p}$.

5.9 Corollary (Estimation on $\mathfrak{P}_{H,\beta,\alpha,p}$)
Let H be a separable RKHS on X with respect to a bounded and measurable kernel k, $\sigma > 0$, $p, \alpha, \gamma \in [0,1]$ and $\beta \in [0,2]$ with $0 < p \leq \alpha \leq 1$ resp. $0 \leq \gamma < \beta \leq 2$. Then there exists constants $K, N > 0$, depending only on $\mathfrak{P}_{H,\beta,\alpha,p}$ such that for all $0 < \lambda \leq 1$, $\tau \geq 1$ and $n \geq N\frac{\tau^2}{\lambda^{\alpha+p}}$

$$\|[f_{D,\lambda}]_\nu - f_P^*\|_{[H]_\nu^\gamma}^2 \leq K\Big(\lambda^{\beta-\gamma} + \frac{\tau^2}{n\lambda^{\gamma+p+\max\{0,\alpha-\beta\}}}\big(1 + \frac{1}{n\lambda^{\alpha-p}}\big)\Big)$$

holds with P^n-probability $\geq 1 - 4e^{-\tau}$ for all $P \in \mathfrak{P}_{H,\beta,\alpha,p}$.

Proof. Let $P \in \mathfrak{P}_{H,\beta,\alpha,p}$, $0 < \lambda \leq 1$, $\tau \geq 1$ and $n \geq N\frac{\tau^2}{\lambda^{\alpha+p}}$, whereby N is the constant from Lemma 5.8. Then Theorem 5.3 is applicable and yields

$$\|[f_{D,\lambda} - f_{P,\lambda}]_\nu\|_{[H]_\nu^\gamma}^2 \leq 128V\max\{5C_p, A\}\frac{\tau^2}{n\lambda^{\gamma+p+\max\{0,\alpha-\beta\}}}\left(1 + \frac{1}{n\lambda^{\alpha-p}}\right)$$

with P^n-probability $\geq 1 - 4e^{-\tau}$. Together with Lemma 5.2 we get the assertion for $K := 2\max\{128V\max\{5C_p, A\}, B\}$. □

If we exploit the previous results, then we can prove the claimed learning rates.

Proof of Theorem 3.3. Because for the given asymptotic of the regularization parameter sequence $(\lambda_n)_{n\geq 1}$ there is an index bound n_τ such that $n \geq N\frac{\tau^2}{\lambda_n^{\alpha+p}}$ holds for all $n \geq n_\tau$, we can apply Corollary 5.9. Since the term $\frac{1}{n\lambda_n^{\alpha-p}}$ is bounded we get the claimed rates. □

Lower Rates

In order to prove $[H]^\gamma_\nu$-minimax lower rates we establish the following lower bound.

5.10 Lemma (Lower Bound)
Let H be a separable RKHS on X with respect to a bounded and measurable kernel k, $0 < q \leq p \leq \alpha \leq 1$, $0 \leq \gamma \leq 1$ and $0 \leq \gamma < \beta \leq 2$ such that there exists a $\nu \in \mathfrak{N}_{H,\alpha,p,q}$. In addition, we set $v := \frac{\gamma}{p}$ and $u := \frac{q}{\max\{\alpha,\beta\}-\gamma}$. Then there are constants $0 < \varepsilon_0 \leq 1$ and $C_1, C_2 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$ there are $P_0, P_1, \ldots, P_{M_\varepsilon} \in \mathfrak{P}_{H,\beta}(\{\nu\})$ with the following properties:

(i) It holds $2^{C_2\varepsilon^{-u}} \leq M_\varepsilon \leq 2^{3C_2\varepsilon^{-u}}$.

(ii) It holds $\|f^*_{P_i} - f^*_{P_j}\|^2_{[H]^\gamma_\nu} \geq 4\varepsilon$ for all $i, j \in \{0, 1, \ldots, M_\varepsilon\}$ with $i \neq j$.

(iii) It holds

$$\inf_{\Psi} \max_{j=0,1,\ldots,M_\varepsilon} P_j^n\big(D :\ \Psi(D) \neq j\big) \geq \frac{\sqrt{M_\varepsilon}}{\sqrt{M_\varepsilon}+1}\left(1 - C_1 n\varepsilon^{1+u(v+1)} - \frac{1}{2\log(M_\varepsilon)}\right)$$

for all $n \geq 1$, where the infimum is taken over all measurable functions $\Psi : (X \times Y)^n \to \{0, 1, \ldots, M_\varepsilon\}$ with respect to the universal completion of the product-σ-algebra.

Please note, that the probability measures P_j also depend on ε although we omit this in the notation. Remember that we need just one probability measures ν on X with the required properties to construct distributions on $X \times Y$ that are *difficult* to learn. The proof is an application of the following theorem from Tsybakov [14].

5.11 Theorem (Lower Bound)
Let $P_0, P_1, \ldots, P_M$ with $M \geq 2$ be a family of probability measures on a measurable space $(\Omega, \mathcal{A})$.

Moreover, we assume that $P_j \ll P_0$ holds for all $j = 1, \dots, M$ and that $\alpha_* := \frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \in (0, \infty)$, where $K(P_j, P_0)$ denotes the Kullback-Leibler divergence from P_0 to P_j. Then it holds

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

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

Proof. From Tsybakov [14, Proposition 2.3] we know, that

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

is a lower bound for the left hand side. If we choose $\tau = \frac{1}{\sqrt{M}}$ and use the estimation $\sqrt{2\alpha_*} \leq \frac{1}{2} + \alpha_*$ afterwards, then we get the assertion. □

We use this theorem for the measurable space $\Omega = (X \times Y)^n$ with a fixed $n \geq 1$ and equip this space with the universal completion of the product-σ-algebra. Furthermore, we follow the suggestion of Caponnetto and De Vito [3] as well as Blanchard and Mücke [2] in order to construct a family of probability measures $P_0, P_1, \dots, P_M \in \mathfrak{P}_{H,\beta}(\{\nu\})$. In the following, let the assumptions of Lemma 5.10 hold and set $\bar{\sigma} := \min\{\sigma, L\}$. Then we define for a measurable function $f : X \to Y$ and $x \in X$ the distribution $P_f(\,\cdot\,|x) := \mathcal{N}(f(x), \bar{\sigma}^2)$ as the normal distribution on $Y = \mathbb{R}$ with mean $f(x)$ and variance $\bar{\sigma}^2$. Hence $P_f(A) := \int_X \int_Y \mathbb{1}_A(x, y)\ P_f(\mathrm{d}y|x)\ \mathrm{d}\nu(x)$ for $A \in \mathcal{B} \otimes \mathcal{B}(Y)$ defines a probability measure on $X \times Y$ with marginal distribution ν on X, i.e. $(P_f)_X = \nu$. For that reason the corresponding power spaces are independent of f. The following lemma describes the Kullback-Leibler divergence for this measures.

5.12 Lemma (Kullback-Leibler Divergence)
For $f, f' \in \mathcal{L}_2(\nu)$ and $n \geq 1$ it holds $P_f^n \ll P_{f'}^n$ and $P_f^n \gg P_{f'}^n$. Furthermore, the Kullback-Leibler divergence fulfills

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

Proof. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be the density of $\mathcal{N}(0, \bar{\sigma}^2)$ with respect to the Lebesgue measure. Than

$$(x, y) \mapsto \frac{\varphi(y - f(x))}{\varphi(y - f'(x))}$$

is the density of P_f with respect to $P_{f'}$ and therefore $P_f \ll P_{f'}$. Hence it also holds for the product measures $P_f^n \ll P_{f'}^n$. It is well-known fact that $K(P_f^n, P_{f'}^n) = nK(P_f, P_{f'})$ holds and the determination of $K(P_f, P_{f'})$ is a standard procedure, so we omit it. □

Because $P_f = P_{f'}$ for $f' = f$ ν-a.s. we can define $P_{[f]_\nu}$ for ν-equivalence classes. For $f \in L_2(\nu)$ we get $|P_f|_2^2 = \bar{\sigma}^2 + \|f\|_{L_2(\nu)}^2 < \infty$ and $f_{P_f}^* = f$. Moreover, the properties of the normal distribution implies

$$\int_Y |y - f(x)|^m \, P_f(\mathrm{d}y|x) = \frac{1}{\sqrt{\pi}} \Gamma\Big(\frac{m+1}{2}\Big) \big(\bar{\sigma}\sqrt{2}\big)^m \leq \frac{1}{2} m! \, \bar{\sigma}^m$$

for all $x \in X$, where Γ labels the gamma function. Hence for $f \in L_\infty(\nu) \cap [H]_\nu^\beta$ with $\|f\|_{L_\infty(\nu)}^2 \leq B_\infty$ and $\|f\|_{[H]_\nu^\beta}^2 \leq B$ the requirements of Assumption 3.2 are fulfilled, i.e. we have $P_f \in \mathfrak{P}_{H,\beta}(\{\nu\})$. So we reduced the construction of probability measures to the construction of appropriate functions $f_0, f_1, \ldots, f_M \in L_\infty(\nu) \cap [H]_\nu^\beta$. To this end we use binary strings $\omega = (\omega_1, \ldots, \omega_m) \in \{0,1\}^m$ and define

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

for $0 < \varepsilon \leq 1$. Because f_ω is a finite linear combination of the eigenvectors $[e_i]_\nu$ of T_ν it holds $f_\omega \in [H]_\nu \subseteq L_\infty(\nu) \cap [H]_\nu^\beta$. First we want to establish sufficient conditions on ε and m, that $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ and $\|f_\omega\|_{[H]_\nu^\beta}^2 \leq B$ holds.

5.13 Lemma

Under the assumptions of Lemma 5.10 there is are constants $U > 0$ and $0 < \varepsilon_1 \leq 1$ depending only on $\mathfrak{P}_{H,\beta}(\{\nu\})$, such that $\|f_\omega\|_{[H]_\nu^\beta}^2 \leq B$ and $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ holds for all $0 < \varepsilon \leq \varepsilon_1$ and all $m \leq U\varepsilon^{-u}$.

Proof. Let $m \in \mathbb{N}$ and $0 < \varepsilon \leq 1$. The polynomial lower bound and $\gamma < \beta$ implies

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

Hence for $m \leq U_1 \varepsilon^{-\frac{q}{\beta-\gamma}}$ with $U_1 := \frac{1}{2} c^q \big(\frac{B}{32}\big)^{\frac{q}{\beta-\gamma}}$ it holds $\|f_\omega\|_{[H]_\nu^\beta}^2 \leq B$. In the case $\gamma < \alpha$ the embedding property together with an analogues argument yields $\|f_\omega\|_{L_\infty(\nu)}^2 \leq B_\infty$ for $m \leq U_2 \varepsilon^{-\frac{q}{\alpha-\gamma}}$ with $U_2 := \frac{1}{2} c^q \big(\frac{B_\infty}{32A}\big)^{\frac{q}{\alpha-\gamma}}$. So for $U := \min\{U_1, U_2\}$ and $\varepsilon_0 := \min\{1, U^{1/u}\}$ we get the

assertion. In the case $\gamma \geq \alpha$ the polynomial upper bound implies

$$\|f_\omega\|^2_{L_\infty(\nu)} \leq A\|f_\omega\|^2_{[H]^\alpha_\nu} \leq \frac{32\varepsilon}{m} A \sum_{i=1}^{m} \mu_{i+m}^{\gamma-\alpha} \leq 32A\,\varepsilon \mu_m^{\gamma-\alpha} \leq 32AC^{\gamma-\alpha}\varepsilon m^{-\frac{\gamma-\alpha}{p}} \tag{15}$$

and we get $\|f_\omega\|^2_{L_\infty(\nu)} \leq B_\infty$ for $0 < \varepsilon \leq C^{\gamma-\alpha}\frac{B_\infty}{32A}$. So for $U := U_1$ and $\varepsilon_0 := \min\{C^{\gamma-\alpha}\frac{B_\infty}{32A}, U_1^{1/u}\}$ we get the assertion in the case $\gamma \geq \alpha$. □

If $\omega' = (\omega'_1, \ldots, \omega'_m) \in \{0,1\}^m$ is an other binary string, we investigate the norm of the difference $f_\omega - f_{\omega'}$. Analogue estimates as in Equation (15) yields

$$\|f_\omega - f_{\omega'}\|^2_{L_2(\nu)} \leq 32C^\gamma\,\varepsilon m^{-\frac{\gamma}{p}}. \tag{16}$$

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

$$\|f_\omega - f_{\omega'}\|^2_{[H]^\gamma_\nu} = \frac{32\varepsilon}{m}\sum_{i=1}^{m}(\omega_i - \omega'_i)^2 \geq 4\varepsilon. \tag{17}$$

The following lemma is from Tsybakov [14, Lemma 2.9] and suggests that there are many binary strings with large distances.

5.14 Lemma (Gilbert-Varshamov Bound)
For $m \geq 8$ there are $\{\omega^{(0)}, \ldots, \omega^{(M)}\} \subseteq \{0,1\}^m$ with $M \geq 2^{m/8}$ such that $\omega^{(0)} = (0, \ldots, 0)$ and

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

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

Now we are ready to prove Lemma 5.10.

Proof of Lemma 5.10. Let us define $\varepsilon_0 := \min\{\varepsilon_1, (U/8)^{1/u}\}$ and $m_\varepsilon := \lfloor U\varepsilon^{-u}\rfloor$, where we used the notation from Lemma 5.13. Now fix a $n \geq 1$ and a $0 < \varepsilon \leq \varepsilon_0$. Since $m_\varepsilon \geq 8$, Lemma 5.14 yields $M_\varepsilon := \lceil 2^{m_\varepsilon/8}\rceil \geq 2$ binary strings $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M_\varepsilon)} \in \{0,1\}^{m_\varepsilon}$ with *large distances*. If we define $f_j := f_{\omega^{(j)}}$ and $P_j := P_{f_j}$ for $j = 0, 1, \ldots, M_\varepsilon$, then from Lemma 5.13 we get $P_j \in \mathfrak{P}_{H,\beta}(\{\nu\})$ for all $j = 0, 1, \ldots, M_\varepsilon$. Due to the definition of M_ε, m_ε and $m_\varepsilon \geq 8$ we get $\frac{8}{9}U\varepsilon^{-u} \leq m_\varepsilon \leq U\varepsilon^{-u}$ and $2^{\frac{U}{9}\varepsilon^{-u}} \leq 2^{m_\varepsilon/8} \leq M_\varepsilon \leq 2^{m_\varepsilon/4} \leq 2^{\frac{U}{4}\varepsilon^{-u}}$ and Statement (i) holds for $C_2 := \frac{U}{9}$. Assertion (ii)

is a consequence of the large distance between the binary strings and the discussion around Equation (17). Lemma 5.12, Equation (16) and $m_\varepsilon \geq \frac{8}{9} U \varepsilon^{-u}$ yield

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

where $C_3 := \frac{16 C^\gamma 9^v}{\bar{\sigma}^2 (8U)^v}$. Combining Theorem 5.11 and Statement (i) we get Assertion (iii) for $C_1 := \frac{3C_3}{C_2 \log(2)}$. □

Now the proof of Theorem 3.5 remains an application of Lemma 5.10 and the general reduction scheme from Tsybakov [14, Section 2.2].

Proof of Theorem 3.5. Using the notation of Lemma 5.10 and $r := \frac{1}{1+u(v+1)}$. We fix a $\tau > 0$ and choose an index bound n_τ with $\varepsilon_n := \tau \big(\frac{1}{n}\big)^r \leq \varepsilon_0$ for $n \geq n_\tau$. Let us fix a $n \geq n_\tau$. The application of Lemma 5.10 with $\varepsilon = \varepsilon_n$ yields some probability measures $P_0, P_1, \ldots, P_{M_\varepsilon} \in \mathfrak{P}_{H,\beta}(\{\nu\}) \subseteq \mathfrak{P}_{H,\beta,\alpha,p,q}$. First we estimate the left hand side of the inequality in statement (iii) of Lemma 5.10. Therefore we choose an arbitrary measurable learning method $D \mapsto f_D$ and define the measurable function $\Psi : (X \times Y)^n \to \{0, 1, \ldots, M_\varepsilon\}$,

$$\Psi(D) := \operatorname*{argmin}_{j=0,1,\ldots,M_\varepsilon} \|[f_D]_\nu - f_j\|_{[H]_\nu^\gamma}.$$

Then for $j \in \{0, 1, \ldots, M_\varepsilon\}$ and $D \in (X \times Y)^n$ with $\Psi(D) \neq j$ it holds

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

and therefore $P_j^n(D : \|[f_D]_\nu - f_{P_j}^*\|_{[H]_\nu^\gamma}^2 \geq \varepsilon) \geq P_j^n(D : \Psi(D) \neq j)$. Because of $P_j \in \mathfrak{P}_{H,\beta,\alpha,p,q}$ for all $j = 0, 1, \ldots, M_\varepsilon$ we get

$$\begin{aligned}
\inf_\Psi \max_{j=0,1,\ldots,M_\varepsilon} P_j^n\big(D :\ \Psi(D) \neq j\big) &\leq \max_{j=0,1,\ldots,M_\varepsilon} P_j^n\big(D :\ \|[f_D]_\nu - f_{P_j}^*\|_{[H]_\nu^\gamma}^2 \geq \varepsilon\big) \\
&\leq \sup_{P \in \mathfrak{P}_{H,\beta,\alpha,p,q}} P^n\big(D :\ \|[f_D]_\nu - f_P^*\|_{[H]_\nu^\gamma}^2 \geq \varepsilon\big).
\end{aligned}$$

Since we considered an arbitrary measurable learning method this holds also for the infimum over all measurable learning methods. Since $\varepsilon = \varepsilon_n$ with an arbitrary $n \geq n_\tau$ and $M_{\varepsilon_n} \to \infty$, we get

$$\liminf_{n\to\infty} \inf_{D \mapsto f_D} \sup_{P \in \mathfrak{P}_{H,\beta,\alpha,p,q}} P^n\big(D :\ \|[f_D]_\nu - f_P^*\|_{[H]_\nu^\gamma}^2 \geq \varepsilon_n\big) \geq 1 - 3C_1 \tau^{1+u(v+1)}.$$

Another limit $\tau \to 0^+$ yields the assertion because $r = \frac{\max\{\alpha,\beta\}-\gamma}{\max\{\alpha,\beta\}+q-\gamma(1-\frac{q}{p})}$. □

A. Appendix

A.1 Lemma

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

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

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

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

A.2 Theorem (Bernstein Inequality without Supremum Bound)

Let $(\Omega, \mathcal{B}, P)$ be a probability space, H a separable Hilbert space and $\xi : \Omega \to H$ 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 it holds

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

for all $\tau \geq 1$, $n \geq 1$ and $\sigma_*^2 := 5\sigma^2$ and $L_* := L$.

Proof of Theorem A.2. We want to apply Caponnetto and De Vito [3, Proposition 2]. To this end we first prove

$$\mathbb{E}_P \|\xi - \mathbb{E}_P\xi\|_H^m \leq \frac{1}{2} m! 4\sigma^2 (L+\sigma)^{m-2} \tag{18}$$

for all $m \geq 2$. Let us fix $m \geq 2$. Because of $\|\mathbb{E}_P\xi\|_H \leq \mathbb{E}_P\|\xi\|_H \leq \sigma$ it holds

$$\mathbb{E}_P \|\xi - \mathbb{E}_P\xi\|_H^m \leq \mathbb{E}_P(\|\xi\|_H + \sigma)^m = \sum_{k=0}^m \binom{m}{k} \mathbb{E}_P(\|\xi\|_H^k)\, \sigma^{m-k}.$$

If we omit the first two terms of the sum, then we can apply our assumptions:

$$\sum_{k=2}^{m} \binom{m}{k} \mathbb{E}_P(\|\xi\|_H^k)\, \sigma^{m-k} \leq \frac{1}{2} m! \sigma^2 \sum_{k=2}^{m} \frac{1}{(m-k)!} L^{k-2} \sigma^{m-k}.$$

Shifting the index and using $\frac{1}{((m-2)-k)!} \leq \binom{m-2}{k}$ yields

$$\sum_{k=2}^{m} \binom{m}{k} \mathbb{E}_P(\|\xi\|_H^k)\, \sigma^{m-k} \leq \frac{1}{2} m! \sigma^2 \sum_{k=0}^{m-2} \binom{m-2}{k} L^k \sigma^{(m-2)-k} = \frac{1}{2} m! \sigma^2 (L+\sigma)^{m-2}.$$

Now let us estimate the first two terms: The first term ($k = 0$) is bounded by $\sigma^m \leq \frac{1}{2} m! \sigma^2 (L + \sigma)^{m-2}$ and the second therm ($k = 1$) by $m\sigma^m \leq \frac{1}{2} m!\, 2\sigma^2 (L+\sigma)^{m-2}$. Together we get Equation (18). Using [3, Proposition 2] and $\sqrt{\frac{4\sigma^2}{n}} + \frac{L+\sigma}{n} \leq \sqrt{\frac{5\sigma^2}{n}} + \frac{L}{n}$ yields the statement. □

A.3 Corollary (Bernstein Inequality with Supremum Bound)

Let $(\Omega, \mathcal{B}, P)$ be a probability space, H a separable Hilbert space and $\xi : \Omega \to H$ a random variable with supremum bound $L := \|\xi\|_{L_\infty(P)} < \infty$ and variance bound $\sigma^2 := \mathbb{E}_P \|\xi\|_H^2 < \infty$. Then Theorem A.2 holds with $\sigma_*^2 = 4\sigma^2$ and $L_* = 2L$.

Proof. Because of the assumptions the requirements of Theorem A.2 are fulfilled for L and σ^2. Since $\sigma \leq L$ we get by the penultimate step in the proof of Theorem A.2. □

References

[1] R.A. Adams and J.J.F. Fournier. *Sobolev Spaces.* Elsevier Science, Amsterdam, 2nd edition, 2003.

[2] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *ArXiv e-prints*, 1604.04054, 2016.

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

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

[5] J. Diestel and J.J. Uhl. *Vector measures.* American Mathematical Society, 1977.

[6] R.M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2nd edition, 2004.

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

[8] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[9] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

[10] I. Steinwart. A short note on the comparison of interpolation widths, wntropy numbers, and kolmogorov widths. *J. Approx. Theory*, 215:13–27, 2017.

[11] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer-Verlag, New York, 2008.

[12] 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.

[13] 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.

[14] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, New York, 2008.

Simon Fischer
Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
E-Mail: Simon.Fischer@mathematik.uni-stuttgart.de

Ingo Steinwart
Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
E-Mail: steinwart@mathematik.uni-stuttgart.de
WWW: http://www.isa.uni-stuttgart.de/Steinwart/

Erschienene Preprints ab Nummer 2012-001

Komplette Liste: `http://www.mathematik.uni-stuttgart.de/preprints`

2017-003 *Fischer, S.; Steinwart, I.:* Sobolev Norm Learning Rates for Regularized Least-Squares Algorithm

2017-002 *Farooq, M.; Steinwart, I.:* Learning Rates for Kernel-Based Expectile Regression

2017-001 *Bauer, B.; Devroye, L; Kohler, M.; Krzyzak, A.; Walk, H.:* Nonparametric Estimation of a Function From Noiseless Observations at Random Points

2016-006 *Devroye, L.; Györfi, L.; Lugosi, G.; Walk, H.:* On the measure of Voronoi cells

2016-005 *Kohls, C.; Kreuzer, C.; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Optimal Control Problems with Control Constraints

2016-004 *Blaschzyk, I.; Steinwart, I.:* Improved Classification Rates under Refined Margin Conditions

2016-003 *Feistauer, M.; Roskovec, F.; Sändig, AM.:* Discontinuous Galerkin Method for an Elliptic Problem with Nonlinear Newton Boundary Conditions in a Polygon

2016-002 *Steinwart, I.:* A Short Note on the Comparison of Interpolation Widths, Entropy Numbers, and Kolmogorov Widths

2016-001 *Köster, I.:* Sylow Numbers in Spectral Tables

2015-016 *Hang, H.; Steinwart, I.:* A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

2015-015 *Steinwart, I.:* Representation of Quasi-Monotone Functionals by Families of Separating Hyperplanes

2015-014 *Muhammad, F.; Steinwart, I.:* An SVM-like Approach for Expectile Regression

2015-013 *Nava-Yazdani, E.:* Splines and geometric mean for data in geodesic spaces

2015-012 *Kimmerle, W.; Köster, I.:* Sylow Numbers from Character Tables and Group Rings

2015-011 *Györfi, L.; Walk, H.:* On the asymptotic normality of an estimate of a regression functional

2015-010 *Gorodski, C, Kollross, A.:* Some remarks on polar actions

2015-009 *Apprich, C.; Höllig, K.; Hörner, J.; Reif, U.:* Collocation with WEB–Splines

2015-008 *Kabil, B.; Rodrigues, M.:* Spectral Validation of the Whitham Equations for Periodic Waves of Lattice Dynamical Systems

2015-007 *Kollross, A.:* Hyperpolar actions on reducible symmetric spaces

2015-006 *Schmid, J.; Griesemer, M.:* Well-posedness of Non-autonomous Linear Evolution Equations in Uniformly Convex Spaces

2015-005 *Hinrichs, A.; Markhasin, L.; Oettershagen, J.; Ullrich, T.:* Optimal quasi-Monte Carlo rules on higher order digital nets for the numerical integration of multivariate periodic functions

2015-004 *Kutter, M.; Rohde, C.; Sändig, A.-M.:* Well-Posedness of a Two Scale Model for Liquid Phase Epitaxy with Elasticity

2015-003 *Rossi, E.; Schleper, V.:* Convergence of a numerical scheme for a mixed hyperbolic-parabolic system in two space dimensions

2015-002 *Döring, M.; Györfi, L.; Walk, H.:* Exact rate of convergence of kernel-based classification rule

2015-001 *Kohler, M.; Müller, F.; Walk, H.:* Estimation of a regression function corresponding to latent variables

2014-021 *Neusser, J.; Rohde, C.; Schleper, V.:* Relaxed Navier-Stokes-Korteweg Equations for Compressible Two-Phase Flow with Phase Transition

2014-020 *Kabil, B.; Rohde, C.:* Persistence of undercompressive phase boundaries for isothermal Euler equations including configurational forces and surface tension

2014-019 *Bilyk, D.; Markhasin, L.:* BMO and exponential Orlicz space estimates of the discrepancy function in arbitrary dimension

2014-018 *Schmid, J.:* Well-posedness of non-autonomous linear evolution equations for generators whose commutators are scalar

2014-017 *Margolis, L.:* A Sylow theorem for the integral group ring of $PSL(2, q)$

2014-016 *Rybak, I.; Magiera, J.; Helmig, R.; Rohde, C.:* Multirate time integration for coupled saturated/unsaturated porous medium and free flow systems

2014-015 *Gaspoz, F.D.; Heine, C.-J.; Siebert, K.G.:* Optimal Grading of the Newest Vertex Bisection and H^1-Stability of the L_2-Projection

2014-014 *Kohler, M.; Krzyżak, A.; Walk, H.:* Nonparametric recursive quantile estimation

2014-013 *Kohler, M.; Krzyżak, A.; Tent, R.; Walk, H.:* Nonparametric quantile estimation using importance sampling

2014-012 *Györfi, L.; Ottucsák, G.; Walk, H.:* The growth optimal investment strategy is secure, too.

2014-011 *Györfi, L.; Walk, H.:* Strongly consistent detection for nonparametric hypotheses

2014-010 *Köster, I.:* Finite Groups with Sylow numbers $\{q^x, a, b\}$

2014-009 *Kahnert, D.:* Hausdorff Dimension of Rings

2014-008 *Steinwart, I.:* Measuring the Capacity of Sets of Functions in the Analysis of ERM

2014-007 *Steinwart, I.:* Convergence Types and Rates in Generic Karhunen-Loève Expansions with Applications to Sample Path Properties

2014-006 *Steinwart, I.; Pasin, C.; Williamson, R.; Zhang, S.:* Elicitation and Identification of Properties

2014-005 *Schmid, J.; Griesemer, M.:* Integration of Non-Autonomous Linear Evolution Equations

2014-004 *Markhasin, L.:* L_2- and $S^r_{p,q}B$-discrepancy of (order 2) digital nets

2014-003 *Markhasin, L.:* Discrepancy and integration in function spaces with dominating mixed smoothness

2014-002 *Eberts, M.; Steinwart, I.:* Optimal Learning Rates for Localized SVMs

2014-001 *Giesselmann, J.:* A relative entropy approach to convergence of a low order approximation to a nonlinear elasticity model with viscosity and capillarity

2013-016 *Steinwart, I.:* Fully Adaptive Density-Based Clustering

2013-015 *Steinwart, I.:* Some Remarks on the Statistical Analysis of SVMs and Related Methods

2013-014 *Rohde, C.; Zeiler, C.:* A Relaxation Riemann Solver for Compressible Two-Phase Flow with Phase Transition and Surface Tension

2013-013 *Moroianu, A.; Semmelmann, U.:* Generalized Killling spinors on Einstein manifolds

2013-012 *Moroianu, A.; Semmelmann, U.:* Generalized Killing Spinors on Spheres

2013-011 *Kohls, K; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Control Constrained Optimal Control Problems

2013-010 *Corli, A.; Rohde, C.; Schleper, V.:* Parabolic Approximations of Diffusive-Dispersive Equations

2013-009 *Nava-Yazdani, E.; Polthier, K.:* De Casteljau's Algorithm on Manifolds

2013-008 *Bächle, A.; Margolis, L.:* Rational conjugacy of torsion units in integral group rings of non-solvable groups

2013-007 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups over composition algebras

2013-006 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups, semifields, and translation planes

2013-005 *Eck, C.; Kutter, M.; Sändig, A.-M.; Rohde, C.:* A Two Scale Model for Liquid Phase Epitaxy with Elasticity: An Iterative Procedure

2013-004 *Griesemer, M.; Wellig, D.:* The Strong-Coupling Polaron in Electromagnetic Fields

2013-003 *Kabil, B.; Rohde, C.:* The Influence of Surface Tension and Configurational Forces on the Stability of Liquid-Vapor Interfaces

2013-002 *Devroye, L.; Ferrario, P.G.; Györfi, L.; Walk, H.:* Strong universal consistent estimate of the minimum mean squared error

2013-001 *Kohls, K.; Rösch, A.; Siebert, K.G.:* A Posteriori Error Analysis of Optimal Control Problems with Control Constraints

2012-013 *Diaz Ramos, J.C.; Dominguez Vázquez, M.; Kollross, A.:* Polar actions on complex hyperbolic spaces

2012-012 *Moroianu; A.; Semmelmann, U.:* Weakly complex homogeneous spaces

2012-011 *Moroianu; A.; Semmelmann, U.:* Invariant four-forms and symmetric pairs

2012-010 *Hamilton, M.J.D.:* The closure of the symplectic cone of elliptic surfaces

2012-009 *Hamilton, M.J.D.:* Iterated fibre sums of algebraic Lefschetz fibrations

2012-008 *Hamilton, M.J.D.:* The minimal genus problem for elliptic surfaces

2012-007 *Ferrario, P.:* Partitioning estimation of local variance based on nearest neighbors under censoring

2012-006 *Stroppel, M.:* Buttons, Holes and Loops of String: Lacing the Doily

2012-005 *Hantsch, F.:* Existence of Minimizers in Restricted Hartree-Fock Theory

2012-004 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Unitals admitting all translations

2012-003 *Hamilton, M.J.D.:* Representing homology classes by symplectic surfaces

2012-002 *Hamilton, M.J.D.:* On certain exotic 4-manifolds of Akhmedov and Park

2012-001 *Jentsch, T.:* Parallel submanifolds of the real 2-Grassmannian