Analysis and Applications (2015)

DOI: 10.1142/S0219530515500050

World Scientific
www.worldscientific.com

Learning rates for the risk of kernel-based quantile regression estimators in additive models

Andreas Christmann
Department of Mathematics
University of Bayreuth
Universitaetsstrasse 30, Bayreuth
95440 Bayreuth, Germany
andreas.christmann@uni-bayreuth.de

Ding-Xuan Zhou
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon Tong, Hong Kong
mazhou@cityu.edu.hk

Received 16 October 2014
Accepted 3 November 2014
Published 5 March 2015

Additive models play an important role in semiparametric statistics. This paper gives learning rates for regularized kernel-based methods for additive models. These learning rates compare favorably in particular in high dimensions to recent results on optimal learning rates for purely nonparametric regularized kernel-based quantile regression using the Gaussian radial basis function kernel, provided the assumption of an additive model is valid. Additionally, a concrete example is presented to show that a Gaussian function depending only on one variable lies in a reproducing kernel Hilbert space generated by an additive Gaussian kernel, but does not belong to the reproducing kernel Hilbert space generated by the multivariate Gaussian kernel of the same variance.

Keywords: Additive model; quantile regression; rate of convergence; support vector machine.

Mathematics Subject Classification 2010: 62G05, 62G08, 68Q32

1. Introduction

Additive models [30, 9, 10] provide an important family of models for semiparametric regression or classification. Some reasons for the success of additive models are their increased flexibility when compared to linear or generalized linear models and their increased interpretability when compared to fully nonparametric models. It is

well known that good estimators in additive models are in general less prone to the curse of high dimensionality than good estimators in fully nonparametric models. Many examples of such estimators belong to the large class of regularized kernel-based methods over a reproducing kernel Hilbert space H, see e.g., [21, 38]. In the last years many interesting results on learning rates of regularized kernel-based models for additive models have been published when the focus is on sparsity and when the classical least squares loss function is used, see e.g., [18, 1, 17, 19, 22, 33] and the references therein. Of course, the least squares loss function is differentiable and has many nice mathematical properties, but it is only locally Lipschitz continuous and therefore regularized kernel-based methods based on this loss function typically suffer on bad statistical robustness properties, even if the kernel is bounded. This is in sharp contrast to kernel methods based on a Lipschitz continuous loss function and on a bounded loss function, where results on upper bounds for the bias and on a bounded influence function are known, see e.g., [4] for the general case and [3] for additive models.

Therefore, we will here consider the case of regularized kernel-based methods based on a general convex and Lipschitz continuous loss function, on a general kernel, and on the classical regularizing term $\lambda\|\cdot\|_H^2$ for some $\lambda > 0$ which is a smoothness penalty but not a sparsity penalty, see e.g., [35, 36, 23, 32, 6, 26, 11, 7]. Such regularized kernel-based methods are now often called support vector machines (SVMs), although the notation was historically used for such methods based on the special hinge loss function and for special kernels only, we refer to [37, 2, 5].

In this paper we address the open question, whether an SVM with an additive kernel can provide a substantially better learning rate in high dimensions than an SVM with a general kernel, say a classical Gaussian RBF kernel, if the assumption of an additive model is satisfied. Our leading example covers learning rates for quantile regression based on the Lipschitz continuous but nondifferentiable pinball loss function, which is also called check function in the literature, see, e.g., [16, 15] for parametric quantile regression and [24, 34, 28] for kernel-based quantile regression. We will not address the question how to check whether the assumption of an additive model is satisfied because this would be a topic of a paper of its own. Of course, a practical approach might be to fit both models and compare their risks evaluated for test data. For the same reason we will also not cover sparsity.

Consistency of support vector machines generated by additive kernels for additive models was considered in [3]. In this paper we establish learning rates for these algorithms. Let us recall the framework with a complete separable metric space $\mathcal{X}$ as the input space and a closed subset $\mathcal{Y}$ of $\mathbb{R}$ as the output space. A Borel probability measure P on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is used to model the learning problem and an independent and identically distributed sample $D_n = \{(x_i, y_i)\}_{i=1}^n$ is drawn according to P for learning. A loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is used to measure the quality of a prediction function $f : \mathcal{X} \to \mathbb{R}$ by the local error $L(x, y, f(x))$. *Throughout the paper we assume that* L *is measurable,* $L(x, y, y) = 0$, *convex with respect to the*

third variable, and uniformly Lipschitz continuous satisfying

$$\sup_{(x,y)\in\mathcal{Z}} |L(x,y,t) - L(x,y,t')| \le |L|_1 |t-t'| \quad \forall\, t,t' \in \mathbb{R} \tag{1.1}$$

with a finite constant $|L|_1 \in (0,\infty)$.

Support vector machines (SVMs) considered here are kernel-based regularization schemes in a reproducing kernel Hilbert space (RKHS) H generated by a Mercer kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. With a shifted loss function $L^* : \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to \mathbb{R}$ introduced for dealing even with heavy-tailed distributions as $L^*(x,y,t) = L(x,y,t) - L(x,y,0)$, they take the form $f_{L,\mathbb{D}_n,\lambda}$ where for a general Borel measure ρ on $\mathcal{Z}$, the function $f_{L,\rho,\lambda}$ is defined by

$$f_{L,\rho,\lambda} = \arg\min_{f\in H}\{\mathcal{R}_{L^*,\rho}(f) + \lambda\|f\|_H^2\},$$

$$\mathcal{R}_{L^*,\rho}(f) = \int_{\mathcal{Z}} L^*(x,y,f(x))\, d\rho(x,y), \tag{1.2}$$

where $\lambda > 0$ is a regularization parameter. The idea to shift a loss function has a long history, see e.g., [14] in the context of M-estimators. It was shown in [4] that $f_{L,\rho,\lambda}$ is also a minimizer of the following optimization problem involving the original loss function L if a minimizer exists:

$$\min_{f\in H}\left\{\int_{\mathcal{Z}} L(x,y,f(x))\, d\rho(x,y) + \lambda\|f\|_H^2\right\}. \tag{1.3}$$

The additive model we consider consists of the *input space decomposition* $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_s$ with each $\mathcal{X}_j$ a complete separable metric space and a *hypothesis space*

$$\mathcal{F} = \{f_1 + \cdots + f_s : f_j \in \mathcal{F}_j, j = 1,\ldots,s\}, \tag{1.4}$$

where $\mathcal{F}_j$ is a set of functions $f_j : \mathcal{X}_j \to \mathbb{R}$ each of which is also identified as a map $(x_1,\ldots,x_s) \mapsto f_j(x_j)$ from $\mathcal{X}$ to $\mathbb{R}$. Hence the functions from $\mathcal{F}$ take the additive form $f(x_1,\ldots,x_s) = f_1(x_1) + \cdots + f_s(x_s)$. We mention, that there is strictly speaking a notational problem here, because in the previous formula each quantity x_j is an element of the set $\mathcal{X}_j$ which is a subset of the full input space $\mathcal{X}$, $j = 1,\ldots,s$, whereas in the definition of sample $D_n = \{(x_i,y_i)\}_{i=1}^n$ each quantity x_i is an element of the full input space $\mathcal{X}$, where $i = 1,\ldots,n$. Because these notations will only be used in different places and because we do not expect any misunderstandings, we think this notation is easier and more intuitive than specifying these quantities with different symbols.

The additive kernel $k = k_1 + \cdots + k_s$ is defined in terms of Mercer kernels k_j on $\mathcal{X}_j$ as

$$k((x_1,\ldots,x_s),(x_1',\ldots,x_s')) = k_1(x_1,x_1') + \cdots + k_s(x_s,x_s').$$

It generates an RKHS H which can be written in terms of the RKHS H_j generated by k_j on $\mathcal{X}_j$ corresponding to the form (1.4) as

$$H = \{f_1 + \cdots + f_s : f_j \in H_j, j = 1, \ldots, s\}$$

with norm given by

$$\|f\|_H^2 = \min_{\substack{f=f_1+\cdots+f_s \\ f_1\in H_1,\ldots,f_s\in H_s}} \|f_1\|_{H_1}^2 + \cdots + \|f_s\|_{H_s}^2.$$

The norm of $f := f_1 + \cdots + f_s$ satisfies

$$\|f_1 + \cdots + f_s\|_H^2 \le \|f_1\|_{H_1}^2 + \cdots + \|f_s\|_{H_s}^2, \quad f_1 \in H_1, \ldots, f_s \in H_s. \tag{1.5}$$

To illustrate advantages of additive models, we provide two examples of comparing additive with product kernels. The first example deals with Gaussian RBF kernels. All proofs will be given in Sec. 4.

Example 1.1. Let $s = 2$, $\mathcal{X}_1 = \mathcal{X}_2 = [0,1]$ and $\mathcal{X} = [0,1]^2$. Let $\sigma > 0$ and

$$k_1(u,v) = k_2(u,v) = \exp\left(-\frac{|u-v|^2}{\sigma^2}\right), \quad u, v \in [0,1].$$

The additive kernel $k((x_1,x_2),(x_1',x_2')) = k_1(x_1,x_1') + k_2(x_2,x_2')$ is given by

$$k((x_1,x_2),(x_1',x_2')) = \exp\left(-\frac{|x_1-x_1'|^2}{\sigma^2}\right) + \exp\left(-\frac{|x_2-x_2'|^2}{\sigma^2}\right). \tag{1.6}$$

Furthermore, the product kernel $k^{\Pi}((x_1,x_2),(x_1',x_2')) = k_1(x_1,x_1')\cdot k_2(x_2,x_2')$ is the standard Gaussian kernel given by

$$k^{\Pi}((x_1,x_2),(x_1',x_2')) = \exp\left(-\frac{|x_1-x_1'|^2 + |x_2-x_2'|^2}{\sigma^2}\right) \tag{1.7}$$

$$= \exp\left(-\frac{|(x_1,x_2)-(x_1',x_2')|^2}{\sigma^2}\right). \tag{1.8}$$

Define a Gaussian function f on $\mathcal{X} = [0,1]^2$ depending only on one variable by

$$f(x_1,x_2) = \exp\left(-\frac{|x_1|^2}{\sigma^2}\right). \tag{1.9}$$

Then $f \in H$ but

$$f \notin H_{k^{\Pi}}, \tag{1.10}$$

where $H_{k^{\Pi}}$ denotes the RKHS generated by the standard Gaussian RBF kernel k^{Π}.

The second example is about Sobolev kernels.

Example 1.2. Let $2 \le s \in \mathbb{N}$, $\mathcal{X}_1 = \cdots = \mathcal{X}_s = [0,1]$ and $\mathcal{X} = [0,1]^s$. Let

$$W^1[0,1] := \{u \in L_2([0,1]); D^{\alpha}u \in L_2([0,1]) \text{ for all } |\alpha| \le 1\}$$

be the Sobolev space consisting of all square integrable univariate functions whose derivative is also square integrable. It is an RKHS with a Mercer kernel k^* defined on $[0,1]^2$. If we take all the Mercer kernels $k_1, \ldots, k_s$ to be k^*, then $H_j = W^1[0,1]$ for each j. The additive kernel k is also a Mercer kernel and defines an RKHS

$$H = H_1 + \cdots + H_s = \{f_1(x_1) + \cdots + f_s(x_s) : f_1, \ldots, f_s \in W^1[0,1]\}.$$

However, the multivariate Sobolev space $W^1([0,1]^s)$, consisting of all square integrable functions whose partial derivatives are all square integrable, contains discontinuous functions and is not an RKHS.

Denote the marginal distribution of P on $\mathcal{X}_j$ as $P_{\mathcal{X}_j}$. Under the assumption that $H_j \subset \mathcal{F}_j \subset L_1(P_{\mathcal{X}_j})$ for each j and that H_j is dense in $\mathcal{F}_j$ in the $L_1(P_{\mathcal{X}_j})$-metric, it was proved in [3] that

$$\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) \to \mathcal{R}^*_{L^*,P,\mathcal{F}} := \inf_{f\in\mathcal{F}} \mathcal{R}_{L^*,P}(f) \quad (n \to \infty)$$

in probability as long as $\lambda = \lambda_n$ satisfies $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} \lambda_n^2 n = \infty$.

The rest of the paper has the following structure. Section 2 contains our main results on learning rates for SVMs based on additive kernels. Learning rates for quantile regression are treated as important special cases. Section 3 contains a comparison of our results with other learning rates published recently. Section 4 contains all the proofs and some results which can be interesting in their own.

2. Main Results on Learning Rates

In this paper we provide some learning rates for the support vector machines generated by additive kernels for additive models which helps improve the quantitative understanding presented in [3]. The rates are about asymptotic behaviors of the excess risk $\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}}$ and take the form $O(m^{-\alpha})$ with $\alpha > 0$. They will be stated under three kinds of conditions involving the hypothesis space H, the measure P, the loss L, and the choice of the regularization parameter λ.

2.1. *Approximation error in the additive model*

The first condition is about the approximation ability of the hypothesis space H. Since the output function $f_{L,\mathbb{D}_n,\lambda}$ is from the hypothesis space, the learning rates of the learning algorithm depend on the approximation ability of the hypothesis space H with respect to the optimal risk $\mathcal{R}^*_{L^*,P,\mathcal{F}}$ measured by the following approximation error.

Definition 2.1. The approximation error of the triple (H, P, λ) is defined as

$$\mathcal{D}(\lambda) = \inf_{f\in H} \{\mathcal{R}_{L^*,P}(f) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda\|f\|_H^2\}, \quad \lambda > 0. \tag{2.1}$$

To estimate the approximation error, we make an assumption about the minimizer of the risk

$$f^*_{\mathcal{F},P} = \arg\inf_{f\in\mathcal{F}} \mathcal{R}_{L^*,P}(f). \tag{2.2}$$

For each $j \in \{1,\ldots,s\}$, define the integral operator $L_{k_j} : L_2(P_{\mathcal{X}_j}) \to L_2(P_{\mathcal{X}_j})$ associated with the kernel k_j by

$$L_{k_j}(f)(x_j) = \int_{\mathcal{X}_j} k_j(x_j,u_j) f(u_j) dP_{\mathcal{X}_j}(u_j), \quad x_j \in \mathcal{X}_j, f \in L_2(P_{\mathcal{X}_j}).$$

We mention that L_{k_j} is a compact and positive operator on $L_2(P_{\mathcal{X}_j})$. Hence we can find its normalized eigenpairs $((\lambda_{j,\ell},\psi_{j,\ell}))_{\ell\in\mathbb{N}}$ such that $(\psi_{j,\ell})_{\ell\in\mathbb{N}}$ is an orthonormal basis of $L_2(P_{\mathcal{X}_j})$ and $\lambda_{j,\ell} \to 0$ as $\ell \to \infty$. Fix $r > 0$. Then we can define the rth power $L^r_{k_j}$ of L_{k_j} by

$$L^r_{k_j}\left(\sum_\ell c_{j,\ell}\,\psi_{j,\ell}\right) = \sum_\ell c_{j,\ell}\lambda^r_{j,\ell}\,\psi_{j,\ell}, \quad \forall\,(c_{j,\ell})_{\ell\in\mathbb{N}} \in \ell_2.$$

This is a positive and bounded operator and its range is well defined. The assumption $f^*_j = L^r_{k_j}(g^*_j)$ means f^*_j lies in this range.

Assumption 2.2. We assume $f^*_{\mathcal{F},P} \in L_\infty(P_{\mathcal{X}})$ and $f^*_{\mathcal{F},P} = f^*_1 + \cdots + f^*_s$ where for some $0 < r \le \frac{1}{2}$ and each $j \in \{1,\ldots,s\}$, $f^*_j : \mathcal{X}_j \to \mathbb{R}$ is a function of the form $f^*_j = L^r_{k_j}(g^*_j)$ with some $g^*_j \in L_2(P_{\mathcal{X}_j})$.

The case $r = \frac{1}{2}$ of Assumption 2.2 means each f^*_j lies in the RKHS H_j.

A standard condition in the literature (e.g., [25]) for achieving decays of the form $\mathcal{D}(\lambda) = O(\lambda^r)$ for the approximation error (2.1) is $f^*_{\mathcal{F},P} = L^r_k(g^*)$ with some $g^* \in L_2(P_{\mathcal{X}})$. Here the operator L_k is defined by

$$L_k(f)(x_1,\ldots,x_s) = \int_{\mathcal{X}} \left(\sum_{j=1}^s k_j(x_j,x'_j)\right) f(x'_1,\ldots,x'_s) dP_{\mathcal{X}}(x'_1,\ldots,x'_s). \tag{2.3}$$

In general, this cannot be written in an additive form. However, the hypothesis space (1.4) takes an additive form $\mathcal{F} = \mathcal{F}_1 + \cdots + \mathcal{F}_s$. So it is natural for us to impose an additive expression $f^*_{\mathcal{F},P} = f^*_1 + \cdots + f^*_s$ for the target function $f^*_{\mathcal{F},P}$ with the component functions f^*_j satisfying the power condition $f^*_j = L^r_{k_j}(g^*_j)$.

The above natural assumption leads to a technical difficulty in estimating the approximation error: the function f^*_j has no direct connection to the marginal distribution $P_{\mathcal{X}_j}$ projected onto $\mathcal{X}_j$, hence existing methods in the literature (e.g., [25]) cannot be applied directly. Note that on the product space $\mathcal{X}_j \times \mathcal{Y}$, there is no natural probability measure projected from P, and the risk on $\mathcal{X}_j \times \mathcal{Y}$ is not defined.

Our idea to overcome the difficulty is to introduce an intermediate function $f_{j,\lambda}$. It may not minimize a risk (which is not even defined). However, it approximates the

component function f_j^* well. When we add up such functions $f_{1,\lambda} + \cdots + f_{s,\lambda} \in H$, we get a good approximation of the target function $f^*_{\mathcal{F},P}$, and thereby a good estimate of the approximation error. This is the first novelty of the paper.

Theorem 2.3. *Under Assumption* 2.2, *we have*

$$\mathcal{D}(\lambda) \le C_r \lambda^r \quad \forall\, 0 < \lambda \le 1, \tag{2.4}$$

where C_r is the constant given by

$$C_r = \sum_{j=1}^{s} (|L|_1 \|g_j^*\|_{L_2(P_{\mathcal{X}_j})} + \|g_j^*\|^2_{L_2(P_{\mathcal{X}_j})}).$$

2.2. *Special bounds for covering numbers in the additive model*

The second condition for our learning rates is about the capacity of the hypothesis space measured by ℓ_2-empirical covering numbers.

Definition 2.4. Let $\mathcal{G}$ be a set of functions on $\mathcal{Z}$ and $\mathbf{z} = \{z_1, \ldots, z_m\} \subset \mathcal{Z}$. For every $\epsilon > 0$, the **covering number of** $\mathcal{G}$ with respect to the empirical metric $d_{2,\mathbf{z}}$, given by $d_{2,\mathbf{z}}(f,g) = \{\frac{1}{m}\sum_{i=1}^m (f(z_i) - g(z_i))^2\}^{1/2}$ is defined as

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{G}, \epsilon) = \inf\left\{ \ell \in \mathbb{N} : \exists \{f_i\}_{i=1}^{\ell} \subset \mathcal{G} \text{ such that } \mathcal{G} = \bigcup_{i=1}^{\ell} \{f \in \mathcal{G} : d_{2,\mathbf{z}}(f, f_i) \le \epsilon\} \right\}$$

and the ℓ_2**-empirical covering number** of $\mathcal{G}$ is defined as

$$\mathcal{N}(\mathcal{G}, \epsilon) = \sup_{m \in \mathbb{N}} \sup_{\mathbf{z} \in \mathcal{Z}^m} \mathcal{N}_{2,\mathbf{z}}(\mathcal{G}, \epsilon).$$

Assumption 2.5. We assume $\kappa := \sum_{j=1}^s \sup_{x_j \in \mathcal{X}_j} \sqrt{k_j(x_j, x_j)} < \infty$ and that for some $\zeta \in (0, 2)$, $c_\zeta > 0$ and every $j \in \{1, \ldots, s\}$, the ℓ_2-empirical covering number of the unit ball of H_j satisfies

$$\log \mathcal{N}(\{f \in H_j : \|f\|_{H_j} \le 1\}, \epsilon) \le c_\zeta \left(\frac{1}{\epsilon}\right)^\zeta, \quad \forall\, \epsilon > 0. \tag{2.5}$$

The second novelty of this paper is to observe that the additive nature of the hypothesis space yields the following nice bound with a dimension-independent power exponent for the covering numbers of the balls of the hypothesis space H, to be proved in Sec. 4.4.

Theorem 2.6. *Under Assumption* 2.5, *for any $R \ge 1$ and $\epsilon > 0$, we have*

$$\log \mathcal{N}(\{f \in H : \|f\|_H \le R\}, \epsilon) \le s^{1+\zeta} c_\zeta \left(\frac{R}{\epsilon}\right)^\zeta, \quad \forall\, \epsilon > 0. \tag{2.6}$$

Remark 2.7. The bound for the covering numbers stated in Theorem 2.6 is special: the power ζ is independent of the number s of the components in the additive model. It is well known [8] in the literature of function spaces that the covering

numbers of balls of the Sobolev space W^h on the cube $[-1,1]^s$ of the Euclidean space $\mathbb{R}^s$ with regularity index $h > s/2$ has the following asymptotic behavior with $0 < c_{h,s} < C_{h,s} < \infty$:

$$c_{h,s}\left(\frac{R}{\epsilon}\right)^{s/h} \leq \log \mathcal{N}(\{f \in W^h : \|f\|_{W^h} \leq R\}, \epsilon) \leq C_{h,s}\left(\frac{R}{\epsilon}\right)^{s/h}.$$

Here the power $\frac{s}{h}$ depends linearly on the dimension s. Similar dimension-dependent bounds for the covering numbers of the RKHSs associated with Gaussian RBF-kernels can be found in [43, 44]. The special bound in Theorem 2.6 demonstrates an advantage of the additive model in terms of capacity of the additive hypothesis space.

2.3. *Learning rates for quantile regression*

The third condition for our learning rates is about the noise level in the measure P with respect to the hypothesis space. Before stating the general condition, we consider a special case for quantile regression, to illustrate our general results. Let $0 < \tau < 1$ be a quantile parameter. The quantile regression function $f_{P,\tau}$ is defined by its value $f_{P,\tau}(x)$ to be a τ-quantile of $P(\cdot|x)$, i.e. a value $u \in \mathcal{Y} = \mathbb{R}$ satisfying

$$\rho(\{y \in \mathcal{Y} : y \leq u\} \,|\, x) \geq \tau \quad \text{and} \quad \rho(\{y \in \mathcal{Y} : y \geq u\} \,|\, x) \geq 1 - \tau. \tag{2.7}$$

The regularization scheme for quantile regression considered here takes the form (1.2) with the loss function L given by the pinball loss as

$$L(x, y, t) = \begin{cases} (1-\tau)(t-y) & \text{if } t > y, \\ -\tau(t-y) & \text{if } t \leq y. \end{cases} \tag{2.8}$$

A noise condition on P for quantile regression is defined in [27, 28] as follows. To this end, let Q be a probability measure on $\mathbb{R}$ and $\tau \in (0,1)$. Then a real number q_τ is called τ-quantile of Q, if and only if q_τ belongs to the set

$$F_\tau^*(Q) := \{t \in \mathbb{R}, Q((-\infty, t]) \geq \tau \text{ and } Q([t, \infty)) \geq 1 - \tau\}.$$

It is well known that $F_\tau^*(Q)$ is a compact interval.

Definition 2.8. Let $\tau \in (0,1)$.

(1) A probability measure Q on $\mathbb{R}$ is said to have a **τ-quantile of type** 2, if there exist a τ-quantile $t^* \in \mathbb{R}$ and a constant $b_Q > 0$ such that, for all $s \in [0,2]$, we have

$$Q((t^* - s, t^*)) \geq b_Q s \quad \text{and} \quad Q((t^*, t^* + s)) \geq b_Q s. \tag{2.9}$$

(2) Let $p \in (0, \infty]$. We say that a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$ has a **τ-quantile of p-average type** 2 if the conditional probability measure $Q_x := \rho(\cdot \,|\, x)$ has

$\rho_{\mathcal{X}}$-almost surely a τ-quantile of type 2 and the function

$$\gamma : \mathcal{X} \to (0, \infty), \quad \gamma(x) := \gamma_{\rho(\cdot|x)} := b_{\rho(\cdot|x)},$$

where $b_{\rho(\cdot|x)} > 0$ is the constant defined in part (1), satisfies $\gamma^{-1} \in L^p_{\rho_{\mathcal{X}}}$.

One can show that a distribution Q having a τ-quantile of type 2 has a unique τ-quantile t^*. Moreover, if Q has a Lebesgue density h_Q then Q has a τ-quantile of type 2 if h_Q is bounded away from zero on $[t^* - a, t^* + a]$ since we can use $b_Q := \inf\{h_Q(t) : t \in [t^* - a, t^* + a]\}$ in (2.9). This assumption is general enough to cover many distributions used in parametric statistics such as Gaussian, Student's t, and logistic distributions (with $Y = \mathbb{R}$), Gamma and log-normal distributions (with $Y = [0, \infty)$), and uniform and Beta distributions (with $Y = [0, 1]$).

The following theorem, to be proved in Sec. 4, gives a learning rate for the regularization scheme (1.2) in the special case of quantile regression.

Theorem 2.9. *Suppose that $|y| \le |L|_0$ almost surely for some constant $|L|_0 > 0$, and that each kernel k_j is C^∞ with $\mathcal{X}_j \subset \mathbb{R}^{d_j}$ for some $d_j \in \mathbb{N}$. If Assumption 2.2 holds with $r = \frac{1}{2}$ and P has a τ-quantile of p-average type 2 for some $p \in (0, \infty]$, then by taking $\lambda = n^{-\frac{4(p+1)}{3(p+2)}}$, for any $\epsilon > 0$ and $0 < \delta < 1$, with confidence at least $1 - \delta$ we have*

$$\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} \le \widetilde{C} \left(\log \frac{2}{\delta} + \log \left(\log \frac{1}{\epsilon} + 2 \right) \right)^2 n^{\epsilon - \alpha(p)}, \qquad (2.10)$$

where $\widetilde{C}$ is a constant independent of n and δ and

$$\alpha(p) = \frac{2(p+1)}{3(p+2)}. \qquad (2.11)$$

Note that the exponent $\alpha(p)$ given by (2.11) for the learning rate in (2.10) is independent of the quantile level τ, of the number s of additive components in $f^*_{L^*,\mathcal{F},P} = f^*_1 + \cdots + f^*_s$, and of the dimensions $d_1, \ldots, d_s$ and

$$d = \sum_{j=1}^{s} d_j.$$

Further note that $\alpha(p) \in [\frac{1}{2}, \frac{2}{3})$, if $p \ge 2$, and $\alpha(p) \to \frac{2}{3}$ if $p \to \infty$. Because $\epsilon > 0$ can be arbitrarily close to 0, the learning rate, which is independent of the dimension d and given by Theorem 2.9, is close to $n^{-2/3}$ for large values of p and is close to $n^{-1/2}$ or better, if $p \ge 2$.

2.4. *General learning rates*

To state our general learning rates, we need an assumption on a *variance-expectation bound* which is similar to Definition 2.8 in the special case of quantile regression.

Assumption 2.10. We assume that there exist an exponent $\theta \in [0,1]$ and a positive constant c_θ such that

$$\int_{\mathcal{Z}} \{(L^*(x,y,f(x)) - L^*(x,y,f^*_{\mathcal{F},P}(x))^2\} dP(x,y)$$
$$\leq c_\theta (1+\|f\|_\infty)^{2-\theta} \{\mathcal{R}_{L^*,P}(f) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})\}^\theta, \quad \forall\, f \in \mathcal{F}. \tag{2.12}$$

Remark 2.11. Assumption 2.10 always holds true for $\theta = 0$. If the triple $(P, \mathcal{F}, L)$ satisfies some conditions, the exponent θ can be larger. For example, when L is the pinball loss (2.8) and P has a τ-quantile of p-average type q for some $p \in (0,\infty]$ and $q \in (1,\infty)$ as defined in [26], then $\theta = \min\{\frac{2}{q}, \frac{p}{p+1}\}$.

Theorem 2.12. *Suppose that $L(x,y,0)$ is bounded by a constant $|L|_0$ almost surely. Under Assumptions* 2.2, 2.5 *and* 2.10, *if we take $\epsilon > 0$ and $\lambda = n^{-\beta}$ for some $\beta > 0$, then for any $0 < \delta < 1$, with confidence at least $1-\delta$ we have*

$$\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} \leq \widetilde{C} \left(\log \frac{2}{\delta} + \log\left(\log \frac{1}{\epsilon} + 2\right)\right)^2 n^{\epsilon - \alpha(r,\beta,\theta,\zeta)}, \tag{2.13}$$

where $\alpha(r,\beta,\theta,\zeta)$ is given by

$$\min\left\{ r\beta, \frac{1}{2} + \beta\left(\frac{\theta(1+r)}{4} - \frac{1-r}{2}\right), \frac{4}{4-2\theta+\zeta\theta} - \beta, \right.$$
$$\left. \frac{2}{4-2\theta+\zeta\theta} - \frac{(1-r)\beta}{2}, \frac{2}{4-2\theta+\zeta\theta} - \frac{(1-r)\beta}{2} - \frac{\beta(1+r)\left(1-\frac{\theta}{2}\right)-1}{4} \right\} \tag{2.14}$$

and $\widetilde{C}$ is constant independent of n or δ (to be given explicitly in the proof).

3. Comparison of Learning Rates

We now add some theoretical and numerical comparisons on the goodness of our learning rates with those from the literature. As already mentioned in the introduction, some reasons for the popularity of additive models are flexibility, increased interpretability, and (often) a reduced proneness of the curse of high dimensions. Hence it is important to check, whether the learning rate given in Theorem 2.12 under the assumption of an additive model favorably compares to (essentially) optimal learning rates without this assumption. In other words, we need to demonstrate that the main goal of this paper is achieved by Theorems 2.9 and 2.12, i.e. that an SVM based on an additive kernel can provide a substantially better learning rate in high dimensions than an SVM with a general kernel, say a classical Gaussian RBF kernel, provided the assumption of an additive model is satisfied.

Remark 3.1. Our learning rate in Theorem 2.9 is new and optimal in the literature of SVM for quantile regression. Most learning rates in the literature of SVM for

quantile regression are given for projected output functions $\Pi_{|L|_0}(f_{L,\mathbb{D}_n,\lambda})$, while it is well known that projections improve learning rates [40]. Here the projection operator $\Pi_{|L|_0}$ is defined for any measurable function $f : \mathcal{X} \to \mathbb{R}$ by

$$\Pi_{|L|_0}(f)(x) = \begin{cases} f(x) & \text{if } |f(x)| \le |L|_0, \\ |L|_0 & \text{if } f(x) > |L|_0, \\ -|L|_0 & \text{if } f(x) < -|L|_0. \end{cases} \tag{3.1}$$

Sometimes this is called clipping. Such results are given in [28, 41]. For example, under the assumptions that P has a τ-quantile of p-average type 2, the approximation error condition (2.4) is satisfied for some $0 < r \le 1$, and that for some constants $a \ge 1, \xi \in (0,1)$, the sequence of eigenvalues (λ_i) of the integral operator L_k satisfies $\lambda_i \le a i^{-1/\xi}$ for every $i \in \mathbb{N}$, it was shown in [28] that with confidence at least $1-\delta$,

$$\mathcal{R}_{L^*,P}(\Pi_{|L|_0}(f_{L,\mathbb{D}_n,\lambda})) - \mathcal{R}^*_{L^*,P,\mathcal{F}} \le \widetilde{C} \log \frac{2}{\delta} n^{-\alpha},$$

where

$$\alpha = \min\left\{ \frac{(p+1)r}{(p+2)r + (p+1-r)\xi}, \frac{2r}{r+1} \right\}.$$

Here the parameter ξ measures the capacity of the RKHS H_k and it plays a similar role as half of the parameter ζ in Assumption 2.5. For a C^∞ kernel and $r = \frac{1}{2}$, one can choose ξ and ζ to be arbitrarily small and the above power index α can be taken as $\alpha = \min\{\frac{p+1}{p+2}, \frac{2}{3}\} - \epsilon$.

The learning rate in Theorem 2.9 may be improved by relaxing Assumption 2.2 to a Sobolev smoothness condition for $f^*_{\mathcal{F},P}$ and a regularity condition for the marginal distribution $P_{\mathcal{X}}$. For example, one may use a Gaussian kernel $k = k(n)$ depending on the sample size n and [29] achieve the approximation error condition (2.4) for some $0 < r < 1$. This is done for quantile regression in [42, 7]. Since we are mainly interested in additive models, we shall not discuss such an extension.

Example 3.2. Let $s = 2$, $\mathcal{X}_1 = \mathcal{X}_2 = [0,1]$ and $\mathcal{X} = [0,1]^2$. Let $\sigma > 0$ and the additive kernel k be given by (1.6) with k_1, k_2 in Example 1.1 as

$$k_1(u,v) = k_2(u,v) = \exp\left(-\frac{|u-v|^2}{\sigma^2}\right), \quad u, v \in [0,1].$$

If the function $f^*_{\mathcal{F},P}$ is given by (1.9), $|y| \le |L|_0$ almost surely for some constant $|L|_0 > 0$, and P has a τ-quantile of p-average type 2 for some $p \in (0,\infty]$, then by taking $\lambda = n^{-\frac{4(p+1)}{3(p+2)}}$, for any $\epsilon > 0$ and $0 < \delta < 1$, (2.10) holds with confidence at least $1-\delta$.

Remark 3.3. It is unknown whether the above learning rate can be derived by existing approaches in the literature (e.g., [28, 29, 41, 42, 7]) even after projection. Note that the kernel in the above example is independent of the sample size. It

would be interesting to see whether there exists some $r > 0$ such that the function f defined by (1.9) lies in the range of the operator $L^r_{k^\Pi}$. The existence of such a positive index would lead to the approximation error condition (2.4), see [25, 31].

Let us now add some numerical comparisons on the goodness of our learning rates given by Theorem 2.12 with those given by [7]. Their Corollary 4.12 gives (essentially) minimax optimal learning rates for (clipped) SVMs in the context of nonparametric quantile regression using one Gaussian RBF kernel on the whole input space under appropriate smoothness assumptions of the target function. Let us consider the case that the distribution P has a τ-quantile of p-average type 2, where $p = \infty$, and assume that both Corollary 4.12 in [7] and our Theorem 2.12 are applicable. That is, we assume in particular that P is a probability measure on $\mathcal{X} \times \mathcal{Y} := \mathbb{R}^d \times [-1, +1]$ and that the marginal distribution $P_{\mathcal{X}}$ has a Lebesgue density $g \in L_w(\mathbb{R}^d)$ for some $w \geq 1$. Furthermore, suppose that the optimal decision function $f^*_{L^*,\mathcal{F},P}$ has (to make Theorem 2.12 applicable with $r \in (0, \frac{1}{2}]$) the additive structure $f^*_{L^*,\mathcal{F},P} = f^*_1 + \cdots + f^*_s$ with each f^*_j as stated in Assumption 2.2, where $\mathcal{X}_j = \mathbb{R}^{d_j}$ and $d := \sum_{j=1}^s d_j$, with minimal risk $\mathcal{R}^*_{L^*,P,\mathcal{F}}$ and additionally fulfills (to make Corollary 4.12 in [7] applicable)

$$f^*_{L^*,P,\mathcal{F}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d) \cap B^\alpha_{2s,\infty}(\mathbb{R}^d),$$

where $s := \frac{w}{w-1} \in [1, \infty]$ and $B^\alpha_{2s,\infty}(\mathbb{R}^d)$ denotes a Besov space with smoothness parameter $\alpha \geq 1$. The intuitive meaning of α is, that increasing values of α correspond to increased smoothness. We refer to [8, pp. 25–27, 44] for details on Besov spaces. It is well known that the Besov space $B^\alpha_{p,q}(\mathbb{R}^d)$ contains the Sobolev space $W^\alpha_p(\mathbb{R}^d)$ for $\alpha \in \mathbb{N}$, $p \in (1, \infty)$, and $\max\{p, 2\} \leq q \leq \infty$, and that $W^\alpha_2(\mathbb{R}^d) = B^\alpha_{2,2}(\mathbb{R}^d)$. We mention that if all k_j are suitably chosen Wendland kernels, their reproducing kernel Hilbert spaces H_j are Sobolev spaces, see [39, Theorem 10.35, p. 160]. Furthermore, we use the same sequence of regularizing parameters as in [7, Corollaries 4.9 and 4.12], i.e.

$$\lambda_n = c_1 n^{-\beta_{ES}(d,\alpha,\theta)}, \quad \text{where } \beta_{ES}(d, \alpha, \theta) := \frac{2\alpha + d}{2\alpha(2 - \theta) + d}, \quad n \in \mathbb{N}, \tag{3.2}$$

where $d \in \mathbb{N}$, $\alpha \geq 1$, $\theta \in [0, 1]$, and c_1 is some user-defined positive constant independent of $n \in \mathbb{N}$. For reasons of simplicity, let us fix $c_1 = 1$. Then [7, Corollary 4.12] gives learning rates for the risk of SVMs for τ-quantile regression, if a single Gaussian RBF-kernel on $\mathcal{X} \subset \mathbb{R}^d$ is used for τ-quantile functions of p-average type 2 with $p = \infty$, which are of order

$$c_2 n^{\epsilon - \alpha_{ES}(d,\alpha)}, \quad \text{where } \alpha_{ES}(d, \alpha) = \frac{2\alpha}{2\alpha + d}.$$

Hence the learning rate in Theorem 2.9 is better than the one in [7, Corollary 4.12] in this situation, if

$$\alpha(r, \beta_{ES}(d, \alpha, \theta), \theta, \zeta) > \alpha_{ES}(d, \alpha),$$

Table 1. Comparison of exponents of learning rates.

$\theta \in [0,1]$	$\zeta \in (0,2)$	$\lim_{d\to\infty} \alpha_{ES}(d,\alpha)$ from [7, Corollary 4.12]	$\lim_{d\to\infty} \alpha(r,\beta_{ES}(d,\alpha,\theta),\theta,\zeta)$ from Theorem 2.12
>0	fixed	0	positive
1	1	0	$\min\{r,1/3\}$
1	3/2	0	$\min\{r,1/7\}$
1/2	1	0	$\min\{r,1/7\}$
0	fixed	0	0
$\in [0,1]$	$\to 2$	0	0

Note: The table lists the limits of the exponents $\lim_{d\to\infty} \alpha_{ES}(d,\alpha)$ from [7, Corollary 4.12] and $\lim_{d\to\infty} \alpha(r,\beta_{ES}(d,\alpha,\theta),\theta,\zeta)$ from Theorem 2.12, respectively, if the regularizing parameter $\lambda = \lambda_n$ is chosen in an optimal manner for the nonparametric setup, i.e. $\lambda_n = n^{-\beta_{ES}(d,\alpha,\theta)}$, with $\beta_{ES}(d,\alpha,\theta) \to 1$ for $d \to \infty$ and $\alpha \in [1,\infty)$. Recall that $r \in (0, \frac{1}{2}]$.

Table 2. Comparison of exponents of learning rates.

r	θ	ζ	$\lim_{d\to\infty} \alpha(r,\beta_{ES}(d,\alpha,\theta),\theta,\zeta)$
0.5	1	0.1	0.5
		1	0.333
		1.9	0.026
0.5	0.5	0.1	0.311
		1	0.143
		1.9	0.013
0.5	0.1	0.1	0.05
		1	0.026
		1.9	0.003
0.25	1	0.1	0.25
		1	0.25
		1.9	0.026
0.25	0.5	0.1	0.25
		1	0.143
		1.9	0.013
0.25	0.1	0.1	0.05
		1	0.026
		1.9	0.003
0.1	1	0.1	0.1
		1	0.1
		1.9	0.026
0.1	0.5	0.1	0.1
		1	0.1
		1.9	0.013
0.1	0.1	0.1	0.05
		1	0.026
		1.9	0.003

Note: The table lists the limits of the exponents $\lim_{d\to\infty} \alpha(r,\beta_{ES}(d,\alpha,\theta),\theta,\zeta)$ from Theorem 2.12, if the regularizing parameter λ is chosen in optimal manner for the nonparametric setup, i.e. $\lambda = n^{-(2\alpha+d)/(2\alpha(2-\theta)+d)}$ with $\alpha \in [1,\infty)$ and $\theta \in [0,1]$, see [7, Corollary 4.12].

provided the assumption of the additive model is valid. Table 1 lists the values of $\alpha(r, \beta_{ES}(d, \alpha, \theta), \theta, \zeta)$ from (2.14) for some finite values of the dimension d, where $\alpha \in [1, \infty)$. All of these values of $\alpha(r, \beta_{ES}(d, \alpha, \theta), \theta, \zeta)$ are positive with the exceptions if $\theta = 0$ or $\zeta \to 2$. This is in contrast to the corresponding exponent in the learning rate by [7, Corollary 4.12], because

$$\lim_{d \to \infty} \alpha_{ES}(d, \alpha) = \lim_{d \to \infty} \frac{2\alpha}{2\alpha + d} = 0, \quad \forall\, \alpha \in [1, \infty).$$

Table 2 and Figs. 1 and 2 give additional information on the limit $\lim_{d\to\infty} \times \alpha(r, \beta_{ES}(d, \alpha, \theta), \theta, \zeta)$. Of course, higher values of the exponent indicates faster rates of convergence. It is obvious, that an SVM based on an additive kernel has a significantly faster rate of convergence in higher dimensions d compared to SVM based on a single Gaussian RBF kernel defined on the whole input space, of course under the assumption that the additive model is valid. The figures seem to indicate that our learning rate from Theorem 2.12 is probably not optimal for small dimensions. However, the main focus of the present paper is on high dimensions.

We now briefly comment on the goodness of the learning rate provided by Theorem 2.9. Let us assume that the distribution P on $\mathcal{X} \times \mathcal{Y} := \mathbb{R}^d \times [-1, +1]$ has

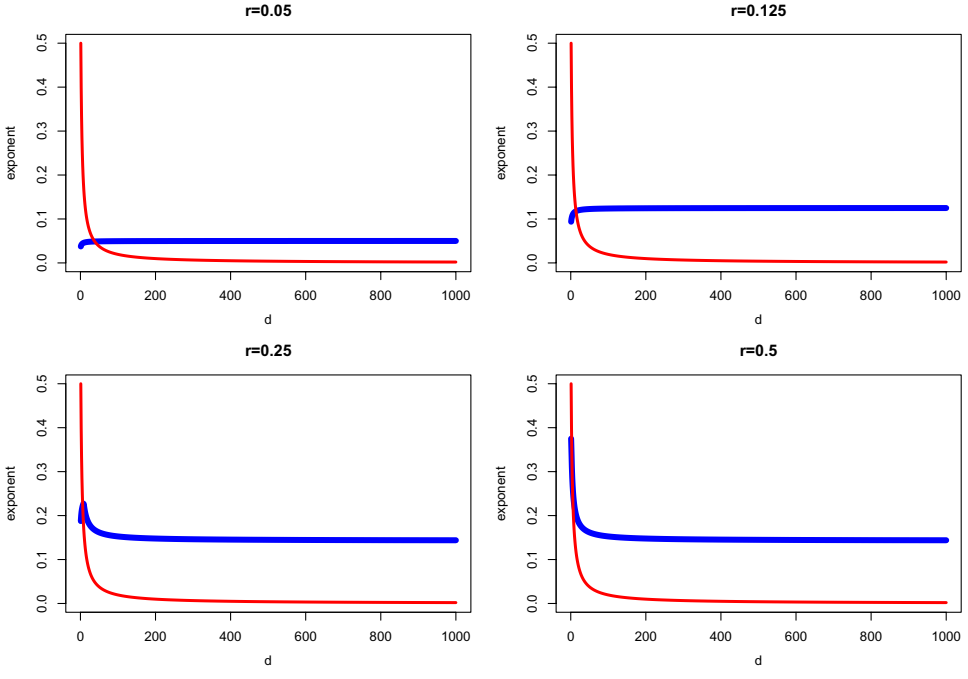

Fig. 1. Plots of exponents $\lim_{d\to\infty} \alpha(r, \beta_{ES}(d, \alpha, \theta), \theta, \zeta)$ from Theorem 2.12 (thick curve) and [7, Corollary 4.12] (thin curve) versus the dimension d, if the regularizing parameter $\lambda = \lambda_n$ is chosen in an optimal manner for the nonparametric setup, i.e. $\lambda_n = n^{-(2\alpha+d)/(2\alpha(2-\theta)+d)}$ with $\alpha = 1$. We set $\theta = 0.5$ and $\zeta = 1$.

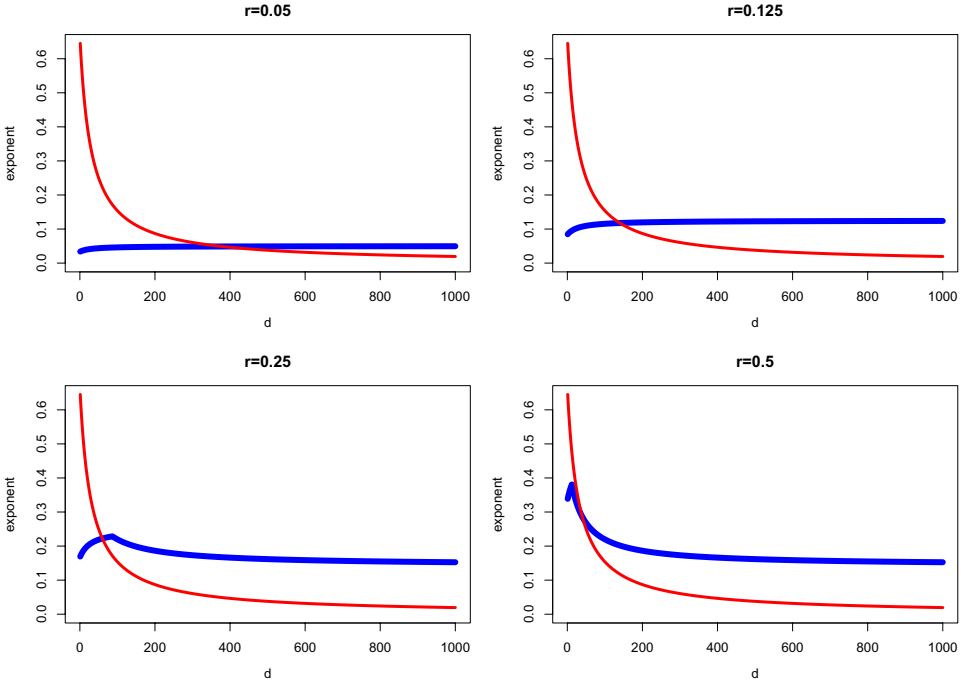

Fig. 2. Similar to Fig. 1, but for $\alpha = 10$.

a τ-quantile of p-average type $q = 2$ for some $p \in (1, \infty]$. Furthermore, consider the sequence of regularizing parameters

$$\lambda := c_1 n^{-\beta_{ES}(d,\alpha,\theta)}, \quad \text{with } \beta_{ES}(d, \alpha, \theta) := \frac{2\alpha + d}{2\alpha(2-\theta) + d},$$

where $c_1 > 0$, $\alpha \geq 1$, and $\theta \in [0, 1]$. For reasons of simplicity, we set $c_1 = 1$. Under the assumptions of Corollary 4.9 in [7], the learning rate for the risk of SVMs for τ-quantile regression, when a single Gaussian RBF-kernel on $\mathcal{X} = \mathbb{R}^d$ is used, is then of order

$$c_2 n^{\epsilon - \alpha_{ES}(d,\alpha,\theta)}, \quad \text{where } \alpha_{ES}(d, \alpha, \theta) = \frac{2\alpha}{2\alpha(2-\theta) + d},$$

where $c_2 > 0$ is a constant independent of n. If α, θ, and p are chosen such that $\frac{2\alpha+d}{2\alpha(2-\theta)+d} = \frac{4(p+1)}{3(p+2)}$ is fulfilled with $d \in \mathbb{N}$, we can make a fair comparison between the learning rates given by [7, Corollary 4.9] and by Theorem 2.9, respectively. Obviously, the learning rate given in Theorem 2.9 favorably compares to the one given by [7, Corollary 4.9] for high dimensions d, if the assumption of an additive model is satisfied, because the exponent $\alpha(p) = \frac{2(p+1)}{3(p+2)}$ in Theorem 2.9 is positive and independent of $d \in \mathbb{N}$, whereas $\alpha_{ES}(d, \alpha, \theta) \to 0$, if $d \to \infty$.

Summarizing, the following conclusion seems to be fair. If an additive model is valid, its structure is known, and the dimension d of $X = \mathbb{R}^d$ is high, then it

makes sense to use an additive kernel, because (i) from a theoretical point of view: faster rate of convergence, (ii) from the big data point of view: the same accuracy of estimating the risk can in principle be achieved already with much smaller data sets, (iii) from an applied point of view: increased interpretability and flexibility.

4. Proofs

This section contains all the proofs of this paper. As some of the results may be interesting in their own, we treat the topics of estimating the approximation error, the proof of the somewhat surprising assertion in Example 1.1, sample error estimates, and the proofs of our learning rates from Sec. 2 in different subsections.

4.1. *Estimating the approximation error*

To carry out our analysis, we need an error decomposition framework.

Lemma 4.1. *There holds*

$$\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2 \le \mathcal{S} + \mathcal{D}(\lambda), \tag{4.1}$$

where the terms are defined as

$$\begin{aligned}\mathcal{S} &= \{\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda})\} \\ &\quad + \{\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,P,\lambda}) - \mathcal{R}_{L^*,P}(f_{L,P,\lambda})\},\end{aligned} \tag{4.2}$$

$$\mathcal{D}(\lambda) = \mathcal{R}_{L^*,P}(f_{L,P,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda\|f_{L,P,\lambda}\|_H^2. \tag{4.3}$$

Proof. We compare the risk with the empirical risk and write $\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda})$ as $\{\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda})\} + \mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda})$. Then we add and subtract a term involving the function $f_{L,P,\lambda}$ to find

$$\begin{aligned}&\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2 \\ &\quad= \{\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda})\} \\ &\qquad+ \{(\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda}) + \lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2) - (\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,P,\lambda}) + \lambda\|f_{L,P,\lambda}\|_H^2)\} \\ &\qquad+ \{\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,P,\lambda}) - \mathcal{R}_{L^*,P}(f_{L,P,\lambda})\} \\ &\qquad+ \{\mathcal{R}_{L^*,P}(f_{L,P,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda\|f_{L,P,\lambda}\|_H^2\}.\end{aligned}$$

But $(\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda}) + \lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2) - (\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,P,\lambda}) + \lambda\|f_{L,P,\lambda}\|_H^2) \le 0$ by the definition of $f_{L,\mathbb{D}_n,\lambda}$. Then the desired statement is proved. □

In the error decomposition (4.1), the first term $\mathcal{S}$ is called *sample error* and will be dealt with later on. The second term $\mathcal{D}(\lambda)$ is the *approximation error* which can be stated equivalently by Definition 2.1.

In this section we estimate the approximation error based on Assumption 2.2. Our estimation is based on the following lemma which is proved by the same method as that in [25]. Recall that the integral operator L_{k_j} is a positive operator on $L_2(P_{\mathcal{X}_j})$, hence $L_{k_j}+\lambda I$ is invertible.

Lemma 4.2. *Let $j \in \{1,\dots,s\}$ and $0 < r \le \frac{1}{2}$. Assume $f_j^* = L_{k_j}^r(g_j^*)$ for some $g_j^* \in L_2(P_{\mathcal{X}_j})$. Define an intermediate function $f_{j,\lambda}$ on $\mathcal{X}_j$ by*

$$f_{j,\lambda} = (L_{k_j}+\lambda I)^{-1} L_{k_j}(f_j^*). \tag{4.4}$$

Then we have

$$\|f_{j,\lambda}-f_j^*\|_{L_2(P_{\mathcal{X}_j})}^2 + \lambda \|f_{j,\lambda}\|_{k_j}^2 \le \lambda^{2r}\|g_j^*\|_{L_2(P_{\mathcal{X}_j})}^2. \tag{4.5}$$

Proof. If $\{(\lambda_i,\psi_i)\}_{i\ge1}$ are the normalized eigenpairs of the integral operator L_{k_j}, then the system $\{\sqrt{\lambda_i}\psi_i : \lambda_i > 0\}$ is orthogonal in H_j.

Write $g_j^* = \sum_{i\ge1} d_i\psi_i$ with $\|\{d_i\}\|_{\ell^2} = \|g_j^*\|_{L_2(P_{\mathcal{X}_j})} < \infty$. Then $f_j^* = \sum_{i\ge1} \times \lambda_i^r d_i\psi_i$ and

$$f_{j,\lambda} - f_j^* = (L_{k_j}+\lambda I)^{-1}L_{k_j}(f_j^*) - f_j^* = -\sum_{i\ge1}\frac{\lambda}{\lambda_i+\lambda}\lambda_i^r d_i\psi_i.$$

Hence

$$\begin{aligned}\|f_{j,\lambda}-f_j^*\|_{L_2(P_{\mathcal{X}_j})}^2 &= \sum_{i\ge1}\left(\frac{\lambda}{\lambda_i+\lambda}\lambda_i^r d_i\right)^2 \\ &= \lambda^{2r}\sum_{i\ge1}\left(\frac{\lambda}{\lambda_i+\lambda}\right)^{2(1-r)}\left(\frac{\lambda_i}{\lambda_i+\lambda}\right)^{2r} d_i^2.\end{aligned}$$

Also,

$$\|f_{j,\lambda}\|_{k_j}^2 = \left\|\sum_{i\ge1}\frac{\lambda_i}{\lambda_i+\lambda}\lambda_i^r d_i\psi_i\right\|_{k_j}^2 = \left\|\sum_{i\ge1}\frac{\lambda_i^{\frac{1}{2}+r}}{\lambda_i+\lambda}d_i\sqrt{\lambda_i}\psi_i\right\|_{k_j}^2 = \sum_{i\ge1}\frac{\lambda_i^{1+2r}}{(\lambda_i+\lambda)^2}d_i^2.$$

Therefore, we have

$$\begin{aligned}&\|f_{j,\lambda}-f_j^*\|_{L_2(P_{\mathcal{X}_j})}^2 + \lambda\|f_{j,\lambda}\|_{k_j}^2 \\ &\quad= \lambda^{2r}\sum_{i\ge1}\left\{\left(\frac{\lambda}{\lambda_i+\lambda}\right)^{2(1-r)}\left(\frac{\lambda_i}{\lambda_i+\lambda}\right)^{2r} + \left(\frac{\lambda}{\lambda_i+\lambda}\right)^{1-2r}\left(\frac{\lambda_i}{\lambda_i+\lambda}\right)^{1+2r}\right\} d_i^2 \\ &\quad\le \lambda^{2r}\sum_{i\ge1}\left\{\frac{\lambda}{\lambda_i+\lambda}+\frac{\lambda_i}{\lambda_i+\lambda}\right\} d_i^2 = \lambda^{2r}\|\{d_i\}\|_{\ell^2}^2 = \lambda^{2r}\|g_j^*\|_{L_2(P_{\mathcal{X}_j})}^2.\end{aligned}$$

This proves the desired bound. □

4.2. *Proof of Theorem* 2.3

Proof of Theorem 2.3. Observe that $f_{j,\lambda} \in H_j$. So $f_{1,\lambda} + \cdots + f_{s,\lambda} \in H$ and by the definition of the approximation error, we have

$$\mathcal{D}(\lambda) \leq \mathcal{R}_{L^*,P}(f_{1,\lambda} + \cdots + f_{s,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda \|f_{1,\lambda} + \cdots + f_{s,\lambda}\|^2_H.$$

But

$$\mathcal{R}^*_{L^*,P,\mathcal{F}} = \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P}) = \mathcal{R}_{L^*,P}(f^*_1 + \cdots + f^*_s)$$

according to Assumption 2.2. Using the inequality in (1.5), we obtain

$$\mathcal{D}(\lambda) \leq \mathcal{R}_{L^*,P}(f_{1,\lambda} + \cdots + f_{s,\lambda}) - \mathcal{R}_{L^*,P}(f^*_1 + \cdots + f^*_s) + \lambda \sum_{j=1}^{s} \|f_{j,\lambda}\|^2_{H_j}.$$

Applying the Lipschitz property (1.1), the excess risk term can be estimated as

$$\begin{aligned}
&\mathcal{R}_{L^*,P}(f_{1,\lambda} + \cdots + f_{s,\lambda}) - \mathcal{R}_{L^*,P}(f^*_1 + \cdots + f^*_s) \\
&\quad = \int_{\mathcal{Z}} L^*(x, y, f_{1,\lambda}(x_1) + \cdots + f_{s,\lambda}(x_s)) dP(x, y) \\
&\qquad - \int_{\mathcal{Z}} L^*(x, y, f^*_1(x_1) + \cdots + f^*_s(x_s)) dP(x, y) \\
&\quad \leq \int_{\mathcal{Z}} |L|_1 \left| \sum_{j=1}^{s} f_{j,\lambda}(x_j) - \sum_{j=1}^{s} f^*_j(x_j) \right| dP(x, y) \\
&\quad \leq |L|_1 \sum_{j=1}^{s} \int_{\mathcal{X}_j} |f_{j,\lambda}(x_j) - f^*_j(x_j)| dP_{\mathcal{X}_j}(x_j).
\end{aligned}$$

But

$$\int_{\mathcal{X}_j} |f_{j,\lambda}(x_j) - f^*_j(x_j)| dP_{\mathcal{X}_j}(x_j) = \|f_{j,\lambda} - f^*_j\|_{L_1(P_{\mathcal{X}_j})} \leq \|f_{j,\lambda} - f^*_j\|_{L_2(P_{\mathcal{X}_j})}.$$

The bound (4.5) implies the following two inequalities

$$\|f_{j,\lambda} - f^*_j\|^2_{L_2(P_{\mathcal{X}_j})} \leq \lambda^{2r} \|g^*_j\|^2_{L_2(P_{\mathcal{X}_j})} \tag{4.6}$$

and

$$\lambda \|f_{j,\lambda}\|^2_{k_j} \leq \lambda^{2r} \|g^*_j\|^2_{L_2(P_{\mathcal{X}_j})}. \tag{4.7}$$

Taking square roots on both sides in (4.6) yields

$$\mathcal{D}(\lambda) \leq \sum_{j=1}^{s} (|L|_1 \|f_{j,\lambda} - f^*_j\|_{L_2(P_{\mathcal{X}_j})} + \lambda \|f_{j,\lambda}\|^2_{H_j}).$$

This together with (4.7) and Lemma 4.2 gives

$$\mathcal{D}(\lambda) \leq \sum_{j=1}^{s} \{|L|_1 \lambda^r \|g^*_j\|_{L_2(P_{\mathcal{X}_j})} + \lambda^{2r} \|g^*_j\|^2_{L_2(P_{\mathcal{X}_j})}\}$$

and completes the proof of the statement. □

4.3. *Proof of the assertion in Example* 1.1

Proof of Example 1.1. The function f can be written as $f = f_1 + 0$ where f_1 is a function on $\mathcal{X}_1$ given by $f_1(x_1, x_1') = k_1(x_1, 0) \in H_1$. So $f \in H$.

Now we prove (1.10). Assume to the contrary that $f \in H_{k^{\Pi}}$. We apply a characterization of the RKHS $H_{k^{\Pi}}$ given in [20, Theorem 1] as

$$H_{k^{\Pi}} = \left\{ f = e^{-\frac{\|x\|^2}{\sigma^2}} \sum_{|\alpha|=0}^{\infty} w_\alpha x^\alpha : \|f\|_K^2 = \sum_{\ell=0}^{\infty} \frac{\ell!}{(2/\sigma^2)^\ell} \sum_{|\alpha|=\ell} \frac{w_\alpha^2}{C_\alpha^\ell} < \infty \right\}, \tag{4.8}$$

where $\|x\|^2 = |x_1|^2 + |x_2|^2$ and $C_\alpha^\ell = \frac{\ell!}{\alpha_1!\alpha_2!}$ for $\alpha = (\alpha_1, \alpha_2) \in \mathbb{Z}_+^2$. Since $f \in H_{k^{\Pi}}$, we have

$$f(x_1, x_2) = \exp\left\{ -\frac{|x_1|^2}{\sigma^2} \right\} = e^{-\frac{|x_1|^2+|x_2|^2}{\sigma^2}} \sum_{|\alpha|=0}^{\infty} w_\alpha x^\alpha,$$

where the coefficient sequence $\{w_\alpha : \alpha \in \mathbb{Z}_+^2\}$ satisfies

$$\|f\|_K^2 = \sum_{\ell=0}^{\infty} \frac{\ell!}{(2/\sigma^2)^\ell} \sum_{|\alpha|=\ell} \frac{w_\alpha^2}{C_\alpha^\ell} < \infty.$$

It follows that

$$\exp\left\{ \frac{|x_2|^2}{\sigma^2} \right\} = \sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{|x_2|^2}{\sigma^2} \right)^m = \sum_{|\alpha|=0}^{\infty} w_\alpha x^\alpha.$$

Hence

$$w_\alpha = \begin{cases} \frac{1}{m!\sigma^{2m}}, & \text{if } \alpha = (0, 2m) \text{ with } m \in \mathbb{Z}_+, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\|f\|_K^2 = \sum_{m=0}^{\infty} \frac{(2m)!}{(2/\sigma^2)^{2m}} \frac{w_{(0,2m)}^2}{C_{(0,2m)}^{2m}} = \sum_{m=0}^{\infty} \frac{(2m)!}{(2/\sigma^2)^{2m}} \left(\frac{1}{m!\sigma^{2m}} \right)^2 = \sum_{m=0}^{\infty} \frac{(2m)!}{2^{2m}(m!)^2}.$$

Finally we apply the Stirling's approximation:

$$\sqrt{2\pi m} \left(\frac{m}{\pi} \right)^m \le m! \le \frac{e}{\sqrt{2\pi}} \sqrt{2\pi m} \left(\frac{m}{\pi} \right)^m,$$

and find

$$\|f\|_K^2 = \sum_{m=0}^{\infty} \frac{(2m)!}{2^{2m}(m!)^2} \ge \sum_{m=0}^{\infty} \frac{\sqrt{2\pi(2m)} \left(\frac{2m}{\pi} \right)^{2m}}{2^{2m} \left(\frac{e}{\sqrt{2\pi}} \sqrt{2\pi m} \left(\frac{m}{\pi} \right)^m \right)^2} = \sum_{m=0}^{\infty} \frac{2\sqrt{\pi}}{e^2 \sqrt{m}} = \infty.$$

This is a contradiction. Therefore, $f \notin H_{k^{\Pi}}$. This proves the conclusion in Example 1.1. □

4.4. *Sample error estimates*

In this subsection we bound the sample error $\mathcal{S}$ defined by (4.2) by Assumption 2.10. It can first be decomposed in two terms:

$$\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2, \tag{4.9}$$

where

$$\begin{aligned}\mathcal{S}_1 &= \{\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})\} \\ &\quad - \{\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}_{L^*,\mathbb{D}_n}(f^*_{\mathcal{F},P})\}, \qquad (4.10)\\ \mathcal{S}_2 &= \{\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,P,\lambda}) - \mathcal{R}_{L^*,\mathbb{D}_n}(f^*_{\mathcal{F},P})\} \\ &\quad - \{\mathcal{R}_{L^*,P}(f_{L,P,\lambda}) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})\}. \qquad (4.11)\end{aligned}$$

The second term $\mathcal{S}_2$ can be bounded easily by the Bernstein inequality.

Lemma 4.3. *Under Assumptions* 2.2 *and* 2.10, *for any* $0 < \lambda \le 1$ *and* $0 < \delta < 1$, *with confidence* $1 - \frac{\delta}{2}$, *we have*

$$\mathcal{S}_2 \le C_1' \log\frac{2}{\delta} \max\left\{\frac{\lambda^{\frac{r-1}{2}}}{n}, \frac{\lambda^{\frac{r-1}{2}+\frac{\theta(r+1)}{4}}}{\sqrt{n}}\right\}, \tag{4.12}$$

where C_1' *is a constant independent of* δ, n *or* λ *and given explicitly in the proof, see* (4.13).

Proof. Consider the random variable ξ on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ defined by

$$\xi(x,y) = L^*(x,y,f_{L,P,\lambda}(x)) - L^*(x,y,f^*_{\mathcal{F},P}(x)), \quad z = (x,y) \in \mathcal{Z}.$$

Here $\mathcal{B}(\mathcal{Z})$ denotes the Borel-σ-algebra. Recall our notation for the constant $\kappa := \sum_{j=1}^s \sup_{x_j \in \mathcal{X}_j} \sqrt{k_j(x_j,x_j)} \ge \sqrt{\|k\|_\infty}$. By Assumptions 2.2 and 2.10, $\|f^*_{\mathcal{F},P}\|_{L_\infty(P_\mathcal{X})} < \infty$ and by Theorem 2.3,

$$\|f_{L,P,\lambda}\|_{L_\infty(P_\mathcal{X})} \le \kappa \|f_{L,P,\lambda}\|_H \le \kappa\sqrt{\mathcal{D}(\lambda)/\lambda} \le \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}} < \infty.$$

This in connection with the Lipschitz condition (1.1) for L tells us that the random variable ξ is bounded by

$$B_\lambda := |L|_1(\|f^*_{\mathcal{F},P}\|_{L_\infty(P_\mathcal{X})} + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}}).$$

By Assumption 2.10, we also know that its variance $\sigma^2(\xi)$ can be bounded as

$$\begin{aligned}\sigma^2(\xi) &\le \int_\mathcal{Z} (\xi(x,y))^2 dP(x,y) \\ &\le c_\theta(1 + \|f_{L,P,\lambda}\|_{L_\infty(P_\mathcal{X})})^{2-\theta}\{\mathcal{R}_{L^*,P}(f_{L,P,\lambda}) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})\}^\theta \\ &\le c_\theta(1 + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}})^{2-\theta}\{C_r\lambda^r\}^\theta \le c_\theta(1+\kappa\sqrt{C_r})^{2-\theta}C_r^\theta\lambda^{r-1+\frac{\theta(r+1)}{2}}.\end{aligned}$$

Now we apply the one-sided Bernstein inequality to ξ which asserts that, for all $\epsilon > 0$,

$$\text{Prob}\left(\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) - \mathbb{E}(\xi) > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\left(\sigma^2(\xi) + \frac{1}{3}B_\lambda\epsilon\right)}\right).$$

Solving the quadratic equation

$$\frac{n\epsilon^2}{2\left(\sigma^2(\xi) + \frac{1}{3}B_\lambda\epsilon\right)} = \log\frac{2}{\delta}$$

for $\epsilon > 0$, we see that with confidence $1 - \frac{\delta}{2}$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) - \mathbb{E}(\xi) \\
&\quad\leq \frac{\frac{1}{3}B_\lambda\log\frac{2}{\delta} + \sqrt{\left(\frac{1}{3}B_\lambda\log\frac{2}{\delta}\right)^2 + 2n\sigma^2(\xi)\log\frac{2}{\delta}}}{n} \\
&\quad\leq \frac{2B_\lambda\log\frac{2}{\delta}}{3n} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}\sigma^2(\xi)} \leq C_1'\log\frac{2}{\delta}\max\left\{\frac{\lambda^{\frac{r-1}{2}}}{n}, \frac{\lambda^{\frac{r-1}{2}+\frac{\theta(r+1)}{4}}}{\sqrt{n}}\right\},
\end{aligned}$$

where C_1' is the constant given by

$$C_1' = |L|_1(\|f_{\mathcal{F},P}^*\|_{L_\infty(P_\mathcal{X})} + \kappa\sqrt{C_r}) + \sqrt{2c_\theta}(1 + \kappa\sqrt{C_r})^{1-\frac{\theta}{2}}C_r^{\frac{\theta}{2}}. \tag{4.13}$$

But $\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) - \mathbb{E}(\xi) = \mathcal{S}_2$. So our conclusion follows. □

The term $\mathcal{S}_1$ involves the function $f_{L,\mathbb{D}_n,\lambda}$ which varies with the sample. Hence we need a concentration inequality to bound this term. We shall do so by applying the following concentration inequality [41] to the function set

$$\mathcal{G} = \{L^*(x, y, f(x)) - L^*(x, y, f_{\mathcal{F},P}^*(x)) : \ f \in H \text{ with } \|f\|_H \leq R\} \tag{4.14}$$

parametrized by the radius R involving the ℓ_2-empirical covering numbers of the function set.

Proposition 4.4 ([41, Proposition 6]). *Let $\mathcal{G}$ be a set of measurable functions on $\mathcal{Z}$, and $B, c > 0, \theta \in [0, 1]$ be constants such that each function $f \in \mathcal{G}$ satisfies $\|f\|_\infty \leq B$ and $\mathbb{E}(f^2) \leq c(\mathbb{E}f)^\theta$. If for some $a > 0$ and $p \in (0, 2)$,*

$$\sup_{\ell\in\mathbb{N}}\sup_{\mathbf{z}\in\mathcal{Z}^\ell}\log\mathcal{N}_{2,\mathbf{z}}(\mathcal{G},\epsilon) \leq a\epsilon^{-p}, \quad \forall\,\epsilon > 0,$$

then there exists a constant c_p' depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \le \frac{1}{2}\eta^{1-\theta}(\mathbb{E}f)^{\theta} + c_p'\eta + 2\left(\frac{ct}{n}\right)^{1/(2-\theta)} + \frac{18Bt}{n}, \quad \forall\, f \in \mathcal{G},$$

where

$$\eta := \max\left\{ c^{\frac{2-p}{4-2\theta+p\theta}} \left(\frac{a}{n}\right)^{\frac{2}{4-2\theta+p\theta}}, B^{\frac{2-p}{2+p}} \left(\frac{a}{n}\right)^{\frac{2}{2+p}} \right\}.$$

Lemma 4.5. *Under Assumptions* 2.5 *and* 2.10, *for any $R \ge 1$, $0 < \lambda \le 1$ and $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$, we have*

$$\begin{aligned} &\{\mathcal{R}_{L^*,P}(f) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})\} - \{\mathcal{R}_{L^*,\mathbb{D}_n}(f) - \mathcal{R}_{L^*,\mathbb{D}_n}(f^*_{\mathcal{F},P})\} \\ &\le C_2' R^{1-\theta} n^{-\frac{2(1-\theta)}{4-2\theta+\zeta\theta}} (\mathcal{R}_{L^*,P}(f) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P}))^{\theta} \\ &\quad + C_2'' \log\frac{2}{\delta} R n^{-\frac{2}{4-2\theta+\zeta\theta}}, \quad \forall\, \|f\|_H \le R, \end{aligned} \tag{4.15}$$

where C_2', C_2'' are constants independent of R, δ, n or λ and given explicitly in the proof. In particular, $C_2' = \frac{1}{2}$ when $\theta = 1$.

Proof. Consider the function set $\mathcal{G}$ defined by (4.14). Each function takes the form $g(x,y) = L^*(x,y,f(x)) - L^*(x,y,f^*_{\mathcal{F},P}(x))$ with $\|f\|_H \le R$. It satisfies

$$\|g\|_\infty \le |L|_1 \|f - f^*_{\mathcal{F},P}\|_\infty \le |L|_1(\kappa + \|f^*_{\mathcal{F},P}\|_{L_\infty(P_{\mathcal{X}})})R =: B$$

and by Assumption 2.10 and the condition $R \ge 1$,

$$\mathbb{E}(g^2) \le (1+\kappa)^{2-\theta} c_\theta R^{2-\theta} (\mathbb{E}g)^{\theta}.$$

Moreover, the Lipschitz property (1.1) and Theorem 2.6 imply that for any $\epsilon > 0$ there holds

$$\sup_{\ell\in\mathbb{N}} \sup_{\mathbf{z}\in\mathcal{Z}^\ell} \log \mathcal{N}_{2,\mathbf{z}}(\mathcal{G},\epsilon) \le \log \mathcal{N}\left(\{f \in H : \|f\|_H \le R\}, \frac{\epsilon}{|L|_1}\right) \le s c_\zeta \left(\frac{s|L|_1 R}{\epsilon}\right)^{\zeta}.$$

Thus all the conditions of Proposition 4.4 are satisfied with $p = \zeta$ and we see that with confidence at least $1 - \frac{\delta}{2}$, there holds

$$\begin{aligned} \mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \le \frac{1}{2}\eta^{1-\theta}(\mathbb{E}g)^{\theta} + c_\zeta'\eta + 2\left(\frac{c\log(2/\delta)}{n}\right)^{1/(2-\theta)} \\ + \frac{18B\log(2/\delta)}{n}, \qquad \forall\, g \in \mathcal{G}, \end{aligned} \tag{4.16}$$

where $c = (1+\kappa)^{2-\theta} c_\theta R^{2-\theta}$, $a = s c_\zeta (s|L|_1 R)^{\zeta}$ and

$$\eta = \max\left\{ c^{\frac{2-\zeta}{4-2\theta+\zeta\theta}} \left(\frac{a}{n}\right)^{\frac{2}{4-2\theta+\zeta\theta}}, B^{\frac{2-\zeta}{2+\zeta}} \left(\frac{a}{n}\right)^{\frac{2}{2+\zeta}} \right\}.$$

But

$$\mathbb{E}g = \mathcal{R}_{L^*,P}(f) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})$$

and

$$\frac{1}{n}\sum_{i=1}^{n} g(z_i) = \mathcal{R}_{L^*,\mathbb{D}_n}(f) - \mathcal{R}_{L^*,\mathbb{D}_n}(f^*_{\mathcal{F},P}).$$

Notice from the inequality $4 - 2\theta + \zeta\theta \geq 2 + \zeta$ that

$$\eta \leq C_3' R n^{-\frac{2}{4-2\theta+\zeta\theta}},$$

where C_3' is the constant given by

$$\begin{aligned} C_3' := {} & ((1+\kappa)^{2-\theta} c_\theta)^{\frac{2-\zeta}{4-2\theta+\zeta\theta}} (s c_\zeta (s|L|_1)^\zeta)^{\frac{2}{4-2\theta+\zeta\theta}} \\ & + (|L|_1(\kappa + \|f^*_{\mathcal{F},P}\|_{L_\infty(P_\mathcal{X})}))^{\frac{2-\zeta}{2+\zeta}} (s c_\zeta (s|L|_1)^\zeta)^{\frac{2}{2+\zeta}}. \end{aligned}$$

Then our desired bound holds true with the constants given by

$$C_2'' = \max\left\{ \frac{1}{2}(C_3')^{1-\theta}, c_\zeta' C_3', 2(1+\kappa)(c_\theta)^{1/(2-\theta)} + 18|L|_1(\kappa + \|f^*_{\mathcal{F},P}\|_{L_\infty(P_\mathcal{X})}) \right\}$$

and

$$C_2' = \begin{cases} C_2'' & \text{if } 0 \leq \theta < 1, \\ \frac{1}{2} & \text{if } \theta = 1. \end{cases}$$

Here the case $\theta = 1$ can be seen directly from (4.16). This completes the proof. □

Combining all the above results yields the following error bounds. For $R \geq 1$, we denote a sample set

$$\mathcal{W}(R) = \{\mathbf{z} \in \mathcal{Z}^n : \|f_{L,\mathbb{D}_n,\lambda}\|_H \leq R\}. \tag{4.17}$$

Proposition 4.6. *Under Assumptions* 2.2, 2.5 *and* 2.10*, let* $R \geq 1$, $0 < \lambda \leq 1$ *and* $0 < \delta < 1$. *Then there exists a subset* $\mathcal{V}_R$ *of* $\mathcal{Z}^n$ *with probability at most* δ *such that*

$$\begin{aligned} & \mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} + \lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2 \\ & \quad \leq C_2' R^{1-\theta} n^{-\frac{2(1-\theta)}{4-2\theta+\zeta\theta}} (\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P}))^\theta \\ & \qquad + C_1' \log\frac{2}{\delta} \max\left\{ \frac{\lambda^{\frac{r-1}{2}}}{n}, \frac{\lambda^{\frac{r-1}{2}+\frac{\theta(r+1)}{4}}}{\sqrt{n}} \right\} \\ & \qquad + C_2'' \log\frac{2}{\delta} R n^{-\frac{2}{4-2\theta+\zeta\theta}} + C_r \lambda^r, \quad \forall\, \mathbf{z} \in \mathcal{W}(R)\backslash\mathcal{V}_R. \end{aligned} \tag{4.18}$$

To apply the above analysis we need a radius R which bounds the norm of the function $f_{L,\mathbb{D}_n,\lambda}$.

Lemma 4.7. *If $L(x,y,0)$ is bounded by a constant $|L|_0$ almost surely, then we have almost surely*

$$\|f_{L,\mathbb{D}_n,\lambda}\|_H \le \sqrt{|L|_0/\lambda}.$$

Proof. By the definition of the function $f_{L,\mathbb{D}_n,\lambda}$, we have

$$\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda}) + \lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2 \le \mathcal{R}_{L^*,\mathbb{D}_n}(0) + \lambda\|0\|_H^2 = 0.$$

Hence we have almost surely

$$\begin{aligned}\lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2 &\le -\mathcal{R}_{L^*,\mathbb{D}_n}(f_{L,\mathbb{D}_n,\lambda})\\ &\le \frac{1}{n}\sum_{i=1}^n L(x_i,y_i,0) \le |L|_0.\end{aligned}$$

Then our desired bound follows. □

Applying Proposition 4.6 to $R = \sqrt{|L|_0/\lambda}$ gives a learning rate. But we can do better by an iteration technique. However, we will first give the proof of Theorem 2.6.

4.5. *Proofs of the main results in Sec.* 2

Proof of Theorem 2.6. By the definition of the ℓ_2-empirical covering number, for every $j \in \{1,\dots,s\}$ and $\mathbf{x}^{(j)} \in (\mathcal{X}_j)^n$, there exists a set of functions $\{f_i^{(j)} : i = 1,\dots,\mathcal{N}^{(j)}\}$ with $\mathcal{N}^{(j)} = \mathcal{N}(\{f \in H_j : \|f\|_{H_j} \le 1\},\epsilon)$ such that for every $f^{(j)} \in H_j$ with $\|f^{(j)}\|_{H_j} \le 1$ we can find some $i_j \in \{1,\dots,\mathcal{N}^{(j)}\}$ satisfying $d_{2,\mathbf{x}^{(j)}} \times (f^{(j)}, f_{i_j}^{(j)}) \le \epsilon$.

Now every function $f \in H$ with $\|f\|_H \le 1$ can be written as $f = f^{(1)} + \cdots + f^{(s)}$ with $\|f^{(j)}\|_{H_j} \le 1$. Also, every $\mathbf{x} = (x_\ell)_{\ell=1}^n \in (\mathcal{X})^n$ can be expressed as $x_\ell = (x_\ell^{(1)},\dots,x_\ell^{(s)})$ with $\mathbf{x}^{(j)} = (x_\ell^{(j)})_{\ell=1}^n \in (\mathcal{X}_j)^n$. By taking the function $f_{i_1,\dots,i_s} = f_{i_1}^{(1)} + \cdots + f_{i_s}^{(s)}$, we see that

$$\begin{aligned}d_{2,\mathbf{x}}(f, f_{i_1,\dots,i_s}) &= \left\{\frac{1}{n}\sum_{\ell=1}^n (f(x_\ell) - f_{i_1,\dots,i_s}(x_\ell))^2\right\}^{1/2}\\ &= \left\{\frac{1}{n}\sum_{\ell=1}^n ((f^{(1)}(x_\ell^{(1)}) + \cdots + f^{(s)}(x_\ell^{(s)}))\right.\\ &\qquad \left. - (f_{i_1}^{(1)}(x_\ell^{(1)}) + \cdots + f_{i_s}^{(s)}(x_\ell^{(s)})))^2\right\}^{1/2}\\ &\le \sum_{j=1}^s \left\{\frac{1}{n}\sum_{\ell=1}^n (f^{(j)}(x_\ell^{(j)}) - f_{i_j}^{(j)}(x_\ell^{(j)}))^2\right\}^{1/2}\\ &= \sum_{j=1}^s d_{2,\mathbf{x}^{(j)}}(f^{(j)}, f_{i_j}^{(j)}) \le s\epsilon.\end{aligned}$$

The number of functions of the form $f_{i_1,\ldots,i_s}$ is $\Pi_{j=1}^{s}\mathcal{N}^{(j)}$. Therefore,

$$\log\mathcal{N}(\{f\in H:\|f\|_H\le 1\},s\epsilon)\le\sum_{j=1}^{s}\log\mathcal{N}(\{f\in H_j:\|f\|_{H_j}\le 1\},\epsilon)$$
$$\le sc_\zeta\left(\frac{1}{\epsilon}\right)^{\zeta}.$$

Then our desired statement follows by scaling R to 1. □

We are now in a position to prove our main results stated in Sec. 2. Theorem 2.12 is proved by applying Proposition 4.6 iteratively. The iteration technique for analyzing regularization schemes has been well developed in the literature [29, 41, 12, 13].

Proof of Theorem 2.12. Take $R^{[0]}=\max\{\sqrt{|L|_0},1\}\frac{1}{\sqrt{\lambda}}$. Lemma 4.7 tells us that $\mathcal{W}(R^{[0]})=\mathcal{Z}^n$. We apply an iteration technique with a sequence of radii $\{R^{[\ell]}\ge 1\}_{\ell\in\mathbb{N}}$ to be defined below.

Apply Proposition 4.6 to $R=R^{[\ell]}$, and when $0\le\theta<1$, apply the elementary inequality

$$\frac{1}{q}+\frac{1}{q^*}=1\text{ with }q,q^*>1\Rightarrow a\cdot b\le\frac{1}{q}a^q+\frac{1}{q^*}b^{q^*},\quad\forall\,a,b\ge 0$$

with $q=\frac{1}{\theta},q^*=\frac{1}{1-\theta}$ and

$$a=2^{-\theta}(\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda})-\mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P}))^\theta,\quad b=2^\theta C_2'R^{1-\theta}n^{-\frac{2(1-\theta)}{4-2\theta+\zeta\theta}}.$$

We know that there exists a subset $\mathcal{V}_{R^{[\ell]}}$ of $\mathcal{Z}^n$ with measure at most δ such that

$$\begin{aligned}&\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda})-\mathcal{R}^*_{L^*,P,\mathcal{F}}+\lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2\\&\le\frac{1}{2}\{\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda})-\mathcal{R}_{L^*,P}(f^*_{\mathcal{F},P})\}\\&\quad+(2^\theta C_2')^{\frac{1}{1-\theta}}R^{[\ell]}n^{-\frac{2}{4-2\theta+\zeta\theta}}\\&\quad+C_1'\log\frac{2}{\delta}\max\left\{\frac{\lambda^{\frac{r-1}{2}}}{n},\frac{\lambda^{\frac{r-1}{2}+\frac{\theta(r+1)}{4}}}{\sqrt{n}}\right\}\\&\quad+C_2''\log\frac{2}{\delta}R^{[\ell]}n^{-\frac{2}{4-2\theta+\zeta\theta}}+C_r\lambda^r,\quad\forall\,\mathbf{z}\in\mathcal{W}(R^{[\ell]})\backslash\mathcal{V}_{R^{[\ell]}}.\end{aligned}$$

It follows that when $\lambda=n^{-\beta}$ for some $\beta>0$, we have

$$\begin{aligned}&\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda})-\mathcal{R}^*_{L^*,P,\mathcal{F}}+\lambda\|f_{L,\mathbb{D}_n,\lambda}\|_H^2\\&\quad\le\max\{a_{n,\delta}R^{[\ell]},b_{n,\delta}\},\quad\forall\,\mathbf{z}\in\mathcal{W}(R^{[\ell]})\backslash\mathcal{V}_{R^{[\ell]}},\end{aligned}\tag{4.19}$$

where

$$a_{n,\delta}:=\{4(2^\theta C_2')^{\frac{1}{1-\theta}}+4C_2''\}\log\frac{2}{\delta}n^{-\frac{2}{4-2\theta+\zeta\theta}}$$

and

$$b_{n,\delta}:=\{4C_1'+4C_r\}\log\frac{2}{\delta}n^{-\alpha'}$$

with

$$\alpha' := \min\left\{\frac{1}{2} + \beta\left(\frac{\theta(1+r)}{4} - \frac{1-r}{2}\right),\ r\beta\right\}.$$

Thus we have

$$\|f_{L,\mathbb{D}_n,\lambda}\|_H \le \max\{n^{\frac{\beta}{2}}\sqrt{a_{n,\delta}}\sqrt{R^{[\ell]}}, n^{\frac{\beta}{2}}\sqrt{b_{n,\delta}}\}, \quad \forall\, \mathbf{z} \in \mathcal{W}(R^{[\ell]})\backslash\mathcal{V}_{R^{[\ell]}}.$$

Hence

$$\mathcal{W}(R^{[\ell]}) \subseteq \mathcal{W}(R^{[\ell+1]}) \cup \mathcal{V}_{R^{[\ell]}}, \tag{4.20}$$

after we define the sequence of radii $\{R^{[\ell]} \ge 1\}_{\ell\in\mathbb{N}}$ by

$$R^{[\ell+1]} = \max\{n^{\frac{\beta}{2}}\sqrt{a_{n,\delta}}\sqrt{R^{[\ell]}}, n^{\frac{\beta}{2}}\sqrt{b_{n,\delta}}, 1\}. \tag{4.21}$$

For any positive integer $J \in \mathbb{N}$, we have

$$\mathcal{Z}^m = \mathcal{W}(R^{[0]}) \subseteq \mathcal{W}(R) \cup \mathcal{V}_{R^{[0]}} \subseteq \cdots \subseteq \mathcal{W}(R^{[J]}) \cup (\cup_{\ell=0}^{J-1}\mathcal{V}_{R^{[\ell]}}),$$

which tells us that the set $\mathcal{W}(R^{[J]})$ has measure at least $1 - J\delta$. We also see iteratively from the definition (4.21) that

$$\begin{aligned}
R^{[J]} &\le \max\{n^{\frac{\beta}{2}}\sqrt{a_{n,\delta}}\sqrt{R^{[J-1]}}, n^{\frac{\beta}{2}}\sqrt{b_{n,\delta}}, 1\} \\
&\le \cdots \le \max\{(n^{\frac{\beta}{2}}\sqrt{a_{n,\delta}})^{1+\frac{1}{2}+\cdots+\frac{1}{2^{J-1}}}(R^{[0]})^{\frac{1}{2^J}}, n^{\frac{\beta}{2}}\sqrt{b_{n,\delta}}, 1, \ldots, \\
&\quad \times (n^{\frac{\beta}{2}}\sqrt{a_{n,\delta}})^{1+\frac{1}{2}+\cdots+\frac{1}{2^{J-1}}}(\max\{n^{\frac{\beta}{2}}\sqrt{b_{n,\delta}}, 1\})^{\frac{1}{2^{J-1}}}\} \\
&\le \{4(2^\theta C_2')^{\frac{1}{1-\theta}} + 4C_2'' + 4C_1' + 4C_r + 1\}\max\{\sqrt{|L|_0}, 1\}\log\frac{2}{\delta}n^{\alpha''},
\end{aligned}$$

where

$$\begin{aligned}
\alpha'' &= \max\left\{\left(2 - \frac{1}{2^{J-1}}\right)\left(\frac{\beta}{2} - \frac{1}{4-2\theta+\zeta\theta}\right) + \frac{\beta}{2^{J+1}}, \frac{\beta}{2} - \frac{\alpha'}{2},\right. \\
&\quad \times \left(\frac{\beta}{2} - \frac{1}{4-2\theta+\zeta\theta}\right)\frac{1}{2}\left(\frac{\beta}{2} - \frac{\alpha'}{2}\right), \ldots, \\
&\quad \left.\times \left(2 - \frac{1}{2^{J-1}}\right)\left(\frac{\beta}{2} - \frac{1}{4-2\theta+\zeta\theta}\right) + \frac{1}{2^{J-1}}\left(\frac{\beta}{2} - \frac{\alpha'}{2}\right)\right\} \\
&\le \max\left\{\beta - \frac{2}{4-2\theta+\zeta\theta}, \frac{\beta}{2} - \frac{\alpha'}{2}\right\} + \frac{1}{2^J} \\
&= \max\left\{\beta - \frac{2}{4-2\theta+\zeta\theta}, \frac{(1-r)\beta}{2}, \frac{(1-r)\beta}{2} + \frac{\beta(1+r)(1-\frac{\theta}{2})-1}{4}\right\} + \frac{1}{2^J}.
\end{aligned}$$

Denote

$$\alpha''' = \max\left\{\beta - \frac{2}{4-2\theta+\zeta\theta}, \frac{(1-r)\beta}{2}, \frac{(1-r)\beta}{2} + \frac{\beta(1+r)(1-\frac{\theta}{2})-1}{4}\right\}$$

and the constant

$$C_3 = \{4(2^\theta C_2')^{\frac{1}{1-\theta}} + 4C_2'' + 4C_1' + 4C_r\} \max\{\sqrt{|L|_0}, 1\}.$$

Choose J to be the smallest positive integer greater than or equal to $\log_2 \frac{1}{\epsilon}$. Then $\frac{1}{2^J} \leq \epsilon$ and

$$R^{[J]} \leq C_3 \log \frac{2}{\delta} n^{\alpha''' + \epsilon}.$$

Applying (4.19) to $\ell = J$, we know that for every $\mathbf{z} \in \mathcal{W}(R^{[J]}) \backslash \mathcal{V}_{R^{[J]}}$, there holds

$$\begin{aligned}\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} &\leq \max\{a_{n,\delta} R^{[J]}, b_{n,\delta}\} \\ &\leq (C_3\{4(2^\theta C_2')^{\frac{1}{1-\theta}} + 4C_2''\} + \{4C_1' + 4C_r\}) \\ &\quad \times \left(\log \frac{2}{\delta}\right)^2 n^{\epsilon - \min\{\frac{2}{4-2\theta+\zeta\theta} - \alpha''',\ \alpha'\}}.\end{aligned}$$

Since the set $\mathcal{W}(R^{[J]})$ has measure at least $1 - J\delta$ while the set $\mathcal{V}_{R^{[J]}}$ has measure at most δ, we know that with confidence at least $1 - (J+1)\delta$,

$$\mathcal{R}_{L^*,P}(f_{L,\mathbb{D}_n,\lambda}) - \mathcal{R}^*_{L^*,P,\mathcal{F}} \leq \widetilde{C} \left(\log \frac{2}{\delta}\right)^2 m^{\epsilon - \alpha},$$

where

$$\alpha = \min\left\{\frac{2}{4 - 2\theta + \zeta\theta} - \alpha''',\ \alpha'\right\}$$

and

$$\widetilde{C} = (C_3\{4(2^\theta C_2')^{\frac{1}{1-\theta}} + 4C_2''\} + \{4C_1' + 4C_r\}).$$

Scaling $J\delta$ to δ, and expressing α explicitly, we see that the conclusion of Theorem 2.12 holds true. □

It only remains to prove Theorem 2.9. We will do so by showing that Theorem 2.9 is a special case of Theorem 2.12.

Proof of Theorem 2.9. Since P has a τ-quantile of p-average type 2 for some $p \in (0, \infty]$, we know from [27] that Assumption 2.10 holds true with $\theta = \frac{p}{p+1}$. Since $\mathcal{X}_j \subset \mathbb{R}^{d_j}$ and $k_j \in C^\infty(\mathcal{X}_j \times \mathcal{X}_j)$, we know from [44] that Assumption 2.5 holds true for an arbitrarily small $\zeta > 0$. By inserting $r = \frac{1}{2}$, $\beta = \frac{4(p+1)}{3(p+2)}$ and $\theta = \frac{p}{p+1}$ into the expression of α in Theorem 2.12 and choosing ζ to be sufficiently small, we know that the conclusion of Theorem 2.9 follows from that of Theorem 2.12.

□

Acknowledgments

The work by A. Christmann described in this paper is partially supported by a grant of the Deutsche Forschungsgesellschaft [Project No. CH/291/2-1]. The work

by D.-X. Zhou described in this paper is supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 105011].

References

[1] F. Bach, Consistency of the group Lasso and multiple kernel learning, *J. Mach. Learn. Res.* **9** (2008) 1179–1225.
[2] B. E. Boser, I. Guyon and V. Vapnik, A training algorithm for optimal margin classifiers, in *Proc. Fifth Annual ACM Workshop on Computational Learning Theory* (ACM, Madison, WI, 1992), pp. 144–152.
[3] A. Christmann and R. Hable, Consistency of support vector machines using additive kernels for additive models, *Comput. Statist. Data Anal.* **56** (2012) 854–873.
[4] A. Christmann, A. Van Messem and I. Steinwart, On consistency and robustness properties of support vector machines for heavy-tailed distributions, *Stat. Interface* **2** (2009) 311–327.
[5] C. Cortes and V. Vapnik, Support vector networks, *Mach. Learn.* **20** (1995) 273–297.
[6] F. Cucker and D.-X. Zhou, *Learning Theory. An Approximation Theory Viewpoint* (Cambridge University Press, Cambridge, 2007).
[7] M. Eberts and I. Steinwart, Optimal regression rates for SVMs using Gaussian kernels, *Electron. J. Statist.* **7** (2013) 1–42.
[8] D. Edmunds and H. Triebel, *Function Spaces, Entropy Numbers, Differential Operators* (Cambridge University Press, Cambridge, 1996).
[9] T. Hastie and R. Tibshirani, Generalized additive models, *Statist. Sci.* **1** (1986) 297–318.
[10] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models* (CRC Press, 1990).
[11] T. Hofmann, B. Schölkopf and A. J. Smola, Kernel methods in machine learning, *Ann. Statist.* **36** (2008) 1171–1220.
[12] T. Hu, Online regression with varying Gaussians and non-identical distributions, *Anal. Appl.* **9** (2011) 395–408.
[13] T. Hu, J. Fan, Q. Wu and D.-X. Zhou, Regularization schemes for minimum error entropy principle, *Anal. Appl.*, published online (2014); DOI: 10.1142/S0219530514500110.
[14] P. J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, in *Proc. 5th Berkeley Symp. on Math. Statist. and Probab.*, Vol. 1 (1967), pp. 221–233.
[15] R. Koenker, *Quantile Regression* (Cambridge University Press, Cambridge, 2005).
[16] R. Koenker and G. Bassett, Regression quantiles, *Econometrica* **46** (1978) 33–50.
[17] V. Koltchinskii and M. Yuan, Sparse recovery in large ensembles of kernel machines, in *Proc. 21st Annual Conf. Learning Theory* (*COLT 2008*), Finland (2008), pp. 229–238.
[18] Y. Lin and H. H. Zhang, Component selection and smoothing in multivariate nonparametric regression, *Ann. Statist.* **34** (2006) 2272–2297.
[19] L. Meier, S. van de Geer and P. Bühlmann, High-dimensional additive modeling, *Ann. Statist.* **37** (2009) 3779–3821.
[20] H. Q. Minh, Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory, *Constr. Approx.* **32** (2010) 307–338.
[21] T. Poggio and F. Girosi, A theory of networks for approximation and learning, *Proc. IEEE* **78** (1990) 1481–1497.

[22] G. Raskutti, M. J. Wainwright and B. Yu, Minimax-optimal rates for sparse additive models over kernel classes via convex programming, *J. Mach. Learn. Res.* **13** (2012) 389–427.
[23] B. Schölkopf and A. J. Smola, *Learning with Kernels* (MIT Press, Cambridge, MA, 2002).
[24] B. Schölkopf, A. J. Smola, R. C. Williamson and P. L. Bartlett, New support vector algorithms, *Neural Comput.* **12** (2000) 1207–1245.
[25] S. Smale and D.-X. Zhou, Shannon sampling II. Connections to learning theory, *Appl. Comput. Harmonic Anal.* **19** (2005) 285–302.
[26] I. Steinwart and A. Christmann, *Support Vector Machines* (Springer, New York, 2008).
[27] I. Steinwart and A. Christmann, How SVMs can estimate quantiles and the median, *Adv. Neural Inf. Process. Syst.* **20** (2008) 305–312.
[28] I. Steinwart and A. Christmann, Estimating conditional quantiles with the help of the pinball loss, *Bernoulli* **17** (2011) 211–225.
[29] I. Steinwart and C. Scovel, Fast rates for support vector machines using Gaussian kernels, *Ann. Statist.* **35** (2007) 575–607.
[30] C. J. Stone, Additive regression and other nonparametric models, *Ann. Statist.* **13** (1985) 689–705.
[31] H. Sun and Q. Wu, Indefinite kernel network with dependent sampling, *Anal. Appl.* **11** (2013) 1350020, 15 pp.
[32] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, *Least Squares Support Vector Machines* (World Scientific, Singapore, 2002).
[33] T. Suzuki and M. Sugiyama, Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness, *Ann. Statist.* **41** (2013) 1381–1405.
[34] I. Takeuchi, Q. V. Le, T. D. Sears and A. J. Smola, Nonparametric quantile estimation, *J. Mach. Learn. Res.* **7** (2006) 1231–1264.
[35] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
[36] V. N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
[37] V. N. Vapnik and A. Lerner, Pattern recognition using generalized portrait method, *Autom. Remote Control* **24** (1963) 774–780.
[38] G. Wahba, Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, in *Advances in Kernel Methods — Support Vector Learning*, eds. B. Schölkopf, C. J. C. Burges and A. J. Smola (MIT Press, Cambridge, MA, 1999), pp. 69–88.
[39] H. Wendland, *Scattered Data Approximation* (Cambridge University Press, Cambridge, 2005).
[40] Q. Wu, Y. M. Ying and D.-X. Zhou, Learning rates of least square regularized regression, *Found. Comput. Math.* **6** (2006) 171–192.
[41] Q. Wu, Y. M. Ying and D.-X. Zhou, Multi-kernel regularized classifiers, *J. Complexity* **23** (2007) 108–134.
[42] D. H. Xiang, Conditional quantiles with varying Gaussians, *Adv. Comput. Math.* **38** (2013) 723–735.
[43] D. H. Xiang and D.-X. Zhou, Classification with Gaussians and convex loss, *J. Mach. Learn. Res.* **10** (2009) 1447–1468.
[44] D.-X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory* **49** (2003) 1743–1752.