Adaptive Learning Rates for Support Vector Machines Working on Data with Low Intrinsic Dimension

Thomas Hamm and Ingo Steinwart

March 18, 2020

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

Abstract

We derive improved regression and classification rates for support vector machines using Gaussian kernels under the assumption that the data has some low-dimensional intrinsic structure that is described by the box-counting dimension. Under some standard regularity assumptions for regression and classification we prove learning rates, in which the dimension of the ambient space is replaced by the box-counting dimension of the support of the data generating distribution. In the regression case our rates are minimax optimal, whereas in the classification case our rates are of the form of the best known. Furthermore, we show that a training validation approach for choosing the hyperparameters of an SVM in a data dependent way achieves the same rates adaptively, that is without any knowledge on the data generating distribution.

Keywords: curse of dimensionality, support vector machines, learning rates, regression, classification

1. Introduction

Theoretical bounds on rates of convergence of nonparametric methods for regression and classification are known to be heavily influenced by the dimension of the input space, a phenomenon usually called the curse of dimensionality in statistical learning theory. Therefore, a considerable amount of effort has been put into coming up with notions of intrinsic dimension of data with the goal to improve the depence on the dimension of the ambient space of some well-established learning rates. The apparent aim of this

idea is to prove much faster rates of convergence under the assumption that the gap between the intrinsic dimension of the data and the dimension of its ambient space is large. It is safe to say, that this assumption also has a practical relevance for real life datasets considering the amount of techniques developed for dimensonality reduction. It is therefore interesting to theoretically analyze the performance of common learning methods under an assumption of low intrinsic dimension of the data.

The probably widest-spread notion to describe the intrinsic dimension of data is to assume that the data generating distribution is supported on some low-dimensional smooth submanifold $X \subset \mathbb{R}^d$, e.g. [3, 27, 35, 36, 37]. Our notion of intrinsic dimension is instead based on the box-counting dimension of the support of the data generating distribution, which considerably generalizes the manifold assumption. A related concept of dimensionality is the so-called doubling dimension, e.g. employed in [17] for tree-based regressors. While in view of regularity constraints on the input space, the doubling dimension is comparable with the box-counting dimension, its great disadvantage however is, that the value of the doubling dimension is not consistent with classical notions of dimension. So far, existing results mainly cover rather basic learning algorithms under some of the notions for low intrinsic dimension, for example local polynomial regression [3], k-nearest neighbor [16, 20], kernel regression [18], or decision trees [27].

In this work we derive least-squares regression and classification rates for Gaussian SVMs under the assumption that the marginal of the data generating distribution on the input space $X \subset \mathbb{R}^d$ is supported on a set of upper box-counting dimension $\varrho \in (0, d]$ and under some standard regularity assumptions. In the regression case we assume that the target function is C^α for some $\alpha \in (0, \infty)$. In this setting we derive a rate of convergence of order $n^{-2\alpha/(2\alpha+\varrho)}$ up to a logarithmic factor. That is, in the special case $\varrho = d$ we reproduce the well-known minimax optimal rate [32]. To the best of our knowledge the only result on regression rates for SVMs under a low intrinsic dimension assumption is given in [36], where the authors derive under the assumption that X is a compact smooth ϱ-dimensional manifold and that the target function is α Hölder continuous for $\alpha \in (0, 1]$ the learning rate $(\log^2(n)/n)^{\alpha/(8\alpha+4\varrho)}$. That is, we significantly improve the existing rates under much more general assumptions. For classification our regularity assumptions are stated in terms of margin/noise conditions. We again reproduce the best known rate up to a logarithmic factor in the special case $\varrho = d$. As for regression, the only existing result [37] on Gaussian SVMs for classification is based on the assumption, that X is some compact smooth manifold. Unfortunately, their regularity assumptions are not comparable to ours and are very restrictive in addition.

As the application of a Gaussian SVM requires the choice of a regularization parameter λ and a bandwidth parameter γ, for which the asymptotically optimal choices depend on unknown characteristics of the data generating distribution, we also consider a training validation approach for chosing the hyperparameters adaptively in a data dependent way. We show, that the resulting learning method achieves the same rates for regression and classification without knowledge on the data generating distribution. Therefore, the overall conclusions of this paper can be summarized as follows: Standard, unmodified Gaussian SVMs using a training validation scheme for hyperparameter selection are, in terms of their generalization performance, adaptive to the intrinisc dimension of the data

and the regularity of the target function.

Additionally, as an overall contribution, we refine the results for adaptive SVM learning rates by showing that the size of the sets of candidate values for the hyperparameters only need to grow logarithmically in the sample size instead of linear, as for example in [29, Theorem 7.24]. This gives a significant improvement on the time complexity for attaining optimal rates adaptively, since this only adds a factor $\log n$ to the time complexity, instead of a factor of n.

The rest of this paper is organized as follows: In Section 2 we give a precise description of the setting we consider. The Sections 3 and 4 contain our main results for regression and classification, respectively. Sections 3 and 4 are each concluded with a comparison to existing results on learning rates under an assumption of low intrinsic dimensional data, which is comparable to ours. All proofs are deferred to the Appendix A including some auxiliary results on entropy numbers for Gaussian RKHSs and an oracle inequality for Gaussian SVMs for general loss functions.

2. Preliminaries

Let $\mathbf{P}$ be a distribution on $X \times Y$, where $X \subset \mathbb{R}^d$ is the input space and $Y \subset \mathbb{R}$ is the output space. We denote the marginal distribution of $\mathbf{P}$ on X by $\mathbf{P}_X$. One goal of nonparametric statistics is to learn a functional relationship between X and Y. That is, given a dataset $((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ sampled from $\mathbf{P}^n$ we want to find a decision function $f_D : X \to Y$ such that $f_D(x)$ is a good prediction for y given x with high probability. To assess the quality of a decision function we fix a loss function, i.e. a measurable function $L : Y \times \mathbb{R} \to [0, \infty)$. In this work, we will consider the following three loss functions: The least-squares loss $L_{\mathrm{LS}}(y, t) := (y - t)^2$, the classification loss $L_{\mathrm{class}}(y, t) := \mathbf{1}_{(-\infty, 0]}(y \operatorname{sgn} t)$ and because of the non-convexity[1] of L_{class}, the hinge loss $L_{\mathrm{hinge}}(y, t) := \max\{0, 1 - yt\}$ as a surrogate. Finally, we measure the quality of a decision function by its risk

$$\mathcal{R}_{L,\mathbf{P}}(f_D) := \int_{X \times Y} L(y, f_D(x)) \, \mathrm{d}\mathbf{P}(x, y).$$

We further call the minimum possible risk, denoted by $\mathcal{R}_{L,\mathbf{P}}^* := \inf_{f : X \to \mathbb{R}} \mathcal{R}_{L,\mathbf{P}}(f)$, the Bayes risk and a function $f_{L,\mathbf{P}}^*$ satisfying $\mathcal{R}_{L,\mathbf{P}}(f_{L,\mathbf{P}}^*) = \mathcal{R}_{L,\mathbf{P}}^*$ a Bayes decision function w.r.t. L and $\mathbf{P}$.

We say that a learning method $D \mapsto f_D$ learns with rate ε_n if there exists a constant $C > 0$ such that

$$\mathbf{P}^n \left(D \in (X \times Y)^n : \mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}_{L,\mathbf{P}}^* \leq C\tau\varepsilon_n \right) \geq 1 - \mathrm{e}^{-\tau} \tag{1}$$

for all $\tau \geq 1$. Many authors alternatively consider expectation bounds of the excess risk, that is $\mathbf{E}_{D \sim \mathbf{P}^n}(\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}_{L,\mathbf{P}}^*) \leq C\varepsilon_n$ for some constant C instead of (1). Note that the exponential high probability in (1) makes our notion of a learning rate stronger than convergence in expectation.

---

[1] We call a loss function $L : Y \times \mathbb{R} \to [0, \infty)$ convex if $L(y, \cdot)$ is convex for all $y \in Y$.

Our notion of the intrinsic dimension of the data is based on the upper box-counting dimension of the support of the marginal distribution $\mathbf{P}_X$. To this end, recall that the support of a Borel measure μ on a subset of $\mathbb{R}^d$, denoted by $\operatorname{supp} \mu$, is the complement of the largest open μ-zero set. To introduce the (upper) box-counting dimension we require the concept of covering numbers.

Definition 2.1. Given a normed space E and a subset $A \subset E$ we say that the points $x_1, \dots, x_m \in E$ are an ε-net of A, if

$$A \subset \bigcup_{j=1}^m (x_j + \varepsilon B_E),$$

where B_E is the closed unit ball in E. Given an $\varepsilon > 0$ the covering number $\mathcal{N}_E(A, \varepsilon)$ of A is defined as the minimum cardinality of an ε-net of A. We may also write $\mathcal{N}(A, \varepsilon) := \mathcal{N}_E(A, \varepsilon)$, if the ambient space E is known from the context. Finally, given a second normed space F and a bounded, linear operator $T : E \to F$, the covering numbers of T are defined by $\mathcal{N}(T, \varepsilon) := \mathcal{N}_F(TB_E, \varepsilon)$.

Now, our central assumption to describe the intrinsic dimension of the data in this paper reads as follows:

(DIM) There exist constants $C_{\dim} > 0$ and $\varrho > 0$ such that for all $\varepsilon \in (0, 1)$ we have

$$\mathcal{N}_{\ell_\infty^d}(\operatorname{supp} \mathbf{P}_X, \varepsilon) \leq C_{\dim} \varepsilon^{-\varrho}.$$

We further say that this assumption is satisfied exactly, if the opposite inequality is fulfilled for the same ϱ and a smaller constant c.

Note, that **(DIM)** implies that $\operatorname{supp} \mathbf{P}_X$ is bounded. The infimum over all ϱ, such that **(DIM)** is fulfilled for ϱ and some finite constant $C_{\dim}$ concides with the so-called upper box-counting dimension of $\operatorname{supp} \mathbf{P}_X$, which is defined as

$$\limsup_{\varepsilon \to 0} \frac{\log \mathcal{N}_{\ell_\infty^d}(\operatorname{supp} \mathbf{P}_X, \varepsilon)}{\log \frac{1}{\varepsilon}}, \tag{2}$$

cf. [10, Section 3.1]. Analogously, the lower box-counting dimension is defined by substituting $\limsup$ with $\liminf$ in (2) and in case those values coincide, this common limit is the box-counting dimension of $\operatorname{supp} \mathbf{P}_X$. Also note, that we can consider in assumption **(DIM)** the covering numbers with respect to the ℓ_p^d-norm for any $p \in [1, \infty)$, as a change of the norm will only influence the constant $C_{\dim}$, but not ϱ. One strength of assumption **(DIM)** is, that the upper box-counting dimension is defined for *any* bounded subset $X \subset \mathbb{R}^d$ and is at most d, that is, we basically impose no regularity in terms of smoothness on X at all, in contrast to the wide-spread manifold assumption in the literature. The reason we formulate our assumption by **(DIM)**, instead in terms of the upper box-counting dimension only, is that the constant $C_{\dim}$ influences the constants in our learning rates and we wish to track this dependence. To get some intuition on this assumption we give a few examples. For these examples assume that $\operatorname{supp} \mathbf{P}_X = X$.

Example 2.2. Let $X = [-1, 1]^d$. By a simple argument we see, that for $\varepsilon = 1/m$, where $m \in \mathbb{N}$, we have $\mathcal{N}_{\ell_\infty^d}(X, \varepsilon) = \varepsilon^{-d}$. Moreover, by [7, Proposition 1.3.1] there exist constants $c, C > 0$, such that

$$c\varepsilon^{-d} \leq \mathcal{N}_{\ell_\infty^d}(X, \varepsilon) \leq C\varepsilon^{-d}$$

for alle $\varepsilon \in (0, 1)$. That is, **(DIM)** is fulfilled exactly for $\varrho = d$. More generally we have for any bounded $X \subset \mathbb{R}^d$ with non-empty interior that **(DIM)** is fulfilled exactly for $\varrho = d$.

Example 2.3. Let $X \subset \mathbb{R}^d$ be a bounded d'-dimensional differentiable manifold. Then **(DIM)** is fulfilled for $\varrho = d'$. This follows from Example 2.2 and the fact, that the box-counting dimension is invariant under bi-Lipschitzian maps, cf. [10, Section 3.2]. Our assumption **(DIM)** therefore includes the manifold assumption commonly used in the literature.

Example 2.4. The *attractor* of a dynamical system is, loosely speaking, a set in the phase space of the dynamical system to which it tends to converge to, based on the initial conditions [23]. It is not unusual for such attractors of dynamical systems describing physical systems to exhibit a fractal structure, whose complexity is, amongst others, measured by their box-counting dimension [11]. A famous example is given by the Lorenz attractor associated to the dynamical system

$$\begin{aligned} x' &= -\sigma x + \sigma y \\ y' &= -xy + rx - y \\ z' &= xy - bz \end{aligned}$$

for real parameters σ, r, b, and was originally used to describe atmospherical convection, see [25] as well as for other examples. The Lorenz attractor is estimated to have a box-counting dimension of approximately 1.98 for certain values of σ, r, b, see [22]. This shows, that our assumptions allowing for non-integer dimensions is not only a mathematical quirk, but is also relevant for real-world datasets. Suppose, for example, the feature vectors x_i of the dataset D are generated by observing the state of such a dynamical systems at independent, random time steps.

A concept related to the box-counting dimension, but with far less farovable properties is the doubling dimension. Recall that the doubling dimension of a set $X \subset \mathbb{R}^d$, considered as a metric space, is the smallest constant c, such that every ball of radius $r > 0$ in X can be covered by 2^c balls of radius $r/2$. It is used, for example, in [17] to describe the intrinsic dimension of the data. We quickly sketch, that the box-counting dimension never exceeds the doubling dimension. To this end, let $X \subset \mathbb{R}^d$ be a bounded set with doubling dimension c. For simplicity, let us assume that $X \subset B_{\ell_2^d}$, that is $\mathcal{N}_{\ell_2^d}(X, 1) = 1$. By assumption X can be covered by 2^c balls of radius $1/2$ which in turn can each be covered by 2^c balls of radius $1/4$. Inductively this gives us $\mathcal{N}(X, 1/2^k) \leq 2^{ck}$ for all $k \in \mathbb{N}_0$. An simple argument then gives us $\mathcal{N}(X, \varepsilon) \leq 2^c \varepsilon^{-c}$ for all $\varepsilon \in (0, 1)$, and

hence we have $c \geq \varrho$, whenever **(DIM)** is fulfilled exactly for ϱ. In fact, we often have $c > \varrho$. For example, as a consequence of [2, Satz 2] the optimal value for the doubling dimension of the unit disc $B_{\ell_2^2}$ is given by $\log_2 7 \approx 2.81$, while its box-counting dimension is 2. Doubling dimensions of sets, that actually have a standard notion of dimension, can rarely be computed explicitly. For example, a ϱ-dimensional manifold has doubling dimensoin $\mathcal{O}(\varrho)$, where the proportionality constant depends on the curvature of X, see e.g. [8, Theorem 22].

Finally, we introduce our learning algorithm of interest. For a subset $X \subset \mathbb{R}^d$ we denote the Gaussian RKHS of width $\gamma > 0$ by $H_\gamma(X)$, that is the RKHS of the kernel $k_\gamma(x,y) = \exp(-\gamma^{-2}\|x-y\|^2)$ for $x, y \in X$. Recall, that a support vector machine (SVM) with Gaussian kernel solves the regularized empirical risk minimization problem

$$f_{D,\lambda,\gamma} := \underset{f \in H_\gamma(X)}{\arg\min}\ \lambda\|f\|_{H_\gamma(X)}^2 + \frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i)).$$

Note that by [29, Theorem 5.5], $f_{D,\lambda,\gamma}$ exists and is unique for convex loss functions L. To deal with some technical difficulties arising from the unboundedness of many loss functions, we also need to introduce the clipping operation. To this end, we say that a loss function L can be clipped at some value $M > 0$, if $L(y, \widehat{t}) \leq L(y, t)$ for all $y \in Y$ and $t \in \mathbb{R}$, where

$$\widehat{t} := \begin{cases} -M & \text{for } t < -M, \\ t & \text{for } t \in [-M, M], \\ M & \text{for } t > M. \end{cases}$$

is the clipped value of t at M, see [29, Definition 2.22]. Our final predictor is then $\widehat{f}_{D,\lambda,\gamma}$. If $Y \subset [-M, M]$ for some $M > 0$, the least-squares loss L_{LS} can be clipped at M. In the scenario of classification, our output space is given by $\{-1, 1\}$. In this case the hinge loss can be clipped at 1. In Sections 3 and 4 let $\widehat{\cdot}$ denote the clipped value at the appropriate value.

The definition of $\widehat{f}_{D,\lambda,\gamma}$ involves a choice for the regularization parameter λ and the width γ. Usually, the asymptotically optimal choices for λ and γ depend on properties of the unknown distribution **P**. To tackle this problem, we also introduce a training validation approach, that choses (approximately) optimal hyper-parameters λ and γ adaptively. To this end, we split our dataset $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ into a training set $D_1 := ((x_1, y_1), \ldots, (x_m, y_m))$ and a validation set $D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n))$, where $m := \lfloor n/2 \rfloor + 1$. Furthermore, we fix finite sets of candidate values $\Lambda_n \subset (0, 1]$ and $\Gamma_n \subset (0, 1]$ for λ and γ. In the first step we then compute the SVM solutions $f_{D_1,\lambda,\gamma}$ for all $(\lambda, \gamma) \in \Lambda_n \times \Gamma_n$. In the second step choose the parameters $(\lambda_{D_2}, \gamma_{D_2})$ that have the smallest empirical error on the validation set, that is

$$\sum_{i=m+1}^n L\left(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}(x_i), y_i\right) = \min_{(\lambda,\gamma) \in \Lambda_n \times \Gamma_n} \sum_{i=m+1}^n L\left(\widehat{f}_{D_1,\lambda,\gamma}(x_i), y_i\right).$$

Our final estimator is then $\widehat{f}_{D_1, \lambda_{D_2}, \gamma_{D_2}}$ and we call this learning method a training validation support vector machine (TV-SVM).

3. Learning Rates for Regression

In this section we derive learning rates for the least-squares loss function $L = L_{\mathrm{LS}}$ under suitable smoothness assumptions on $f^*_{L,\mathbf{P}}$. First of all, recall that for the least-squares loss, the Bayes decision function is given by the conditional mean function $f^*_{L,\mathbf{P}}(x) = \mathbf{E}(Y|X = x)$, see e.g. [14, Section 1.1]. We begin by introducing some tools for our notion of smoothness.

For a function $f : \mathbb{R}^d \to \mathbb{R}$ and $h \in \mathbb{R}^d$ the difference operator Δ_h is defined by $\Delta_h f(x) := f(x + h) - f(x)$. The s-fold application of Δ_h has the explicit expansion

$$\Delta_h^s f(x) = \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} f(x + jh).$$

Given a measure μ on $X \subset \mathbb{R}^d$ we further define the s-th modulus of smoothness by

$$\omega_{s, L_2(\mu)}(f, t) := \sup_{\|h\| \leq t} \|\Delta_h^s f\|_{L_2(\mu)}. \tag{3}$$

Finally, given an $\alpha > 0$ we set $s := \lfloor \alpha \rfloor + 1$ and define the semi-norm

$$|f|_{B_{2,\infty}^\alpha(\mu)} := \sup_{t > 0} t^{-\alpha} \omega_{s, L_2(\mu)}(f, t). \tag{4}$$

Remark 3.1. The so-called Besov spaces $B_{2,\infty}^\alpha(\mathbb{R}^d)$, commonly used in approximation theory, can be defined by means of real interpolation [1, Chapter 3] of Sobolev spaces, more precisely

$$B_{2,\infty}^\alpha(\mathbb{R}^d) := \left(W^{k_0,2}(\mathbb{R}^d), W^{k_1,2}(\mathbb{R}^d) \right)_{\theta,\infty} \quad \text{with } \alpha = k_0(1 - \theta) + k_1 \theta,$$

cf. [33, Section 1.6.4]. For $\alpha > d/2$ we can define an equivalent norm on $B_{2,\infty}^\alpha(\mathbb{R}^d)$ by

$$\| \cdot \|_{L_2(\mathbb{R}^d)} + | \cdot |_{B_{2,\infty}^\alpha(\mathbb{R}^d)},$$

where $|\cdot|_{B_{2,\infty}^\alpha(\mathbb{R}^d)}$ is defined with respect to the Lebesgue measure on $\mathbb{R}^d$ in (3) and (4), cf. [33, Section 2.6.1], explaining our notation. Furthermore, we see that $|f|_{B_{2,\infty}^\alpha(\mathbf{P}_X)} < \infty$, whenever $f \in B_{2,\infty}^\alpha(\mathbb{R}^d)$ and $\mathbf{P}_X$ has a bounded Lebesgue-density.

In light of the remark above, we would like to have an explicit description of functions f with $|f|_{B_{2,\infty}^\alpha(\mathbf{P}_X)} < \infty$ also for the case $\varrho < d$. The following proposition gives a sufficient condition and shows that this class of functions is indeed very rich. To formulate it, we first introduce some additional tools.

Given a function $f : \mathbb{R}^d \to \mathbb{R}$ we define, for a subset $\Omega \subset \mathbb{R}^d$ and $0 < \beta \le 1$, the Hölder seminorm by

$$|f|_{\beta,\Omega} := \sup_{\substack{x,y \in \Omega \\ x \neq y}} \frac{|f(x) - f(y)|}{\|x - y\|^\beta}. \tag{5}$$

If f is k-times continuously differentiable we denote for a multi-index $\nu = (\nu_1, \ldots, \nu_d) \in \mathbb{N}_0^d$ with $|\nu| := \nu_1 + \ldots + \nu_d = k$ the higher-order partial derivative

$$\partial^\nu f(x) = \frac{\partial^{|\nu|} f}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}}(x).$$

Furthermore, let $X^{+\delta} := \{x \in \mathbb{R}^d : \operatorname{dist}(x, X) < \delta\}$ be the open δ-neighborhood of X.

Proposition 3.2. *Let $X \subset \mathbb{R}^d$ be a non-empty set and for $\alpha > 0$ let $f : \mathbb{R}^d \to \mathbb{R}$ be a bounded function that is $\lfloor\alpha\rfloor$-times continuously differentiable on $X^{+\delta}$ for some $\delta > 0$. Additionally, assume that one of the following conditions is satisfied:*

(i) *α is an integer and $\partial^\nu f$ is bounded on $X^{+\delta}$ for all $|\nu| = \alpha$.*

(ii) *α is not an integer and for $\beta = \alpha - \lfloor\alpha\rfloor$ and all $|\nu| = \lfloor\alpha\rfloor$ we have $|\partial^\nu f|_{\beta, X^{+\delta}} < \infty$.*

Then $|f|_{B_{2,\infty}^\alpha(\mu)} < \infty$ for every measure μ with $\operatorname{supp} \mu \subset X$.

In the case where X has a differentiable structure it is more desirable to express our regularity assumptions on $f_{L,\mathbf{P}}^*$ by this intrinsic structure of X. Under our general assumption **(DIM)** this is not possible, or at least there is no obvious way to do so. However, if X is a compact C^k-manifold we can define the class $C^k(X)$, see [28, Chapter 2]. Moreover, by Whitney's extension theorem [15, Theorem 2.3.6] every function $f \in C^k(X)$ has an extension to a function $\tilde{f} \in C^k(X^{+\delta})$ for some $\delta > 0$, eliminating the need to define our regularity assumptions by means of the ambient space in these cases. Especially this means, that the assumption $f_{L,\mathbf{P}}^* \in C^k(X)$, as in [35], is not more general than our assumption $f_{L,\mathbf{P}}^* \in C^k(X^{+\delta})$.

The following theorem establishes an oracle inequality for Gaussian SVMs under assumption **(DIM)**.

Theorem 3.3. *Assume that $\mathbf{P}$ satisfies **(DIM)** and $Y \subset [-M, M]$ as well as $f_{L,\mathbf{P}}^* \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $|f_{L,\mathbf{P}}^*|_{B_{2,\infty}^\alpha(\mathbf{P}_X)} < \infty$. Then there exists a constant $C > 0$ such that for all $\tau > 0, n > 1, \lambda \in (0,1)$, and $\gamma \in (0,1)$ we have*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}_{L,\mathbf{P}}^* \le C\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \lambda^{-1/\log n}\gamma^{-\varrho} n^{-1} \log^{d+1} n + \frac{\tau}{n}\right)$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$.

Using the theorem above we can easily derive learning rates by choosing specific values for the regularization parameter λ and the bandwidth γ. This leads to

Corollary 3.4. *Let the assumptions of Theorem 3.3 be satisfied. Choosing $\gamma_n = n^{-1/(2\alpha+\varrho)}$ and $\lambda_n = n^{-b}$ for some $b \geq (2\alpha + d)/(2\alpha + \varrho)$ there then exists a constant $C > 0$ such that for all $n > 1$ and $\tau \geq 1$ we have*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda_n,\gamma_n}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ not less than $1 - \mathrm{e}^{-\tau}$.

Remark 3.5. It is also possible to formulate Theorem 3.3 and Corollary 3.4 under alternative regularity assumptions. To briefly elaborate this, recall that in [37] the authors consider the case where X is a compact, connected and smooth submanifold of $\mathbb{R}^d$ without boundary and consider a convolution-type operator $S_\gamma : L_2(\mu) \to H_\gamma(X)$, where μ is the measure on X defined by the Riemannian volume form and derive the bounds

$$\|S_\gamma f\|^2_{H_\gamma(X)} \leq C_1 \|f\|^2_{L_2(\mu)} \gamma^{-d} \tag{6}$$

$$\|S_\gamma f - f\|^2_{L_2(\mu)} \leq C_2 \|f\|^2_{W^{2,2}(X)} \gamma^4 \tag{7}$$

$$\|S_\gamma f\|_{\ell_\infty(X)} \leq C_3 \|f\|_{\ell_\infty(X)} \tag{8}$$

[37, Lemma 4, Theorem 2, and Lemma 2], where $W^{2,2}(X)$ denotes the Sobolev space on X. Using these results one can easily derive a modification of Theorem 3.3 and Corollary 3.4 under the assumption that $f^*_{L,\mathbf{P}} \in W^{2,2}(X)$ is bounded and $\mathbf{P}_X$ has a bounded density with respect to μ and prove learning rates of the form $n^{-\frac{4}{4+\varrho}} \log^{d+1} n$. This is done by simply using (7) instead of Lemma A.12, (6) instead of Lemma A.13 and the supremum bound (8) instead of an analogous bound we derive in the proof of Theorem 3.3. Unfortunately, the authors also point out, that the order of approximation cannot be improved if we assume $f \in W^{m,2}(X)$ for some $m > 2$ using the operator S_γ.

The learning rates in Corollary 3.4 can only be achieved, if we know the intrinsic dimension ϱ of the data, as well as the regularity α of the Bayes decision function. However, this is highly unrealistic in practice. The following theorem therefore shows, that a TV-SVM with appropriately chosen candidate sets Λ_n and Γ_n achieves the same rate without knowledge on ϱ and α.

Theorem 3.6. *Let A_n be a minimal $1/\log n$-net of $(0, 1]$ with $1 \in A_n$ and let B_n be a minimal $1/\log n$-net of $[1, d]$ with $d \in B_n$. Set $\Gamma_n := \{n^{-a} : a \in A_n\}$ and $\Lambda_n := \{n^{-b} : b \in B_n\}$. Let the assumptions of Theorem 3.3 be satisfied and assume $\varrho \geq 1$. Then there exists a constant $C > 0$, such that for all $n > 1$ and $\tau \geq 1$ the TV-SVM using Λ_n and Γ_n satisfies*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{2\alpha}{2\alpha+\varrho}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ not less than $1 - \mathrm{e}^{-\tau}$.

Remark 3.7. The proof of Theorem 3.6 shows that the statement also holds if we pick as candidate set for λ the singleton $\Lambda_n = \{n^{-d}\}$. We decided to formulate the theorem

9

as it is, because this choice is closer to the practical usage of the training validation approach. To see why a singleton Λ_n is sufficient recall that to achieve optimal rates the regularization parameter λ_n only needs to satisfy $\lambda_n = n^{-b}$ for some $b \geq (2\alpha+d)/(2\alpha+\varrho)$. If we assume $\varrho \geq 1$ this bound on b is satisfied for $b = d$. This also shows that with no lower bound on ϱ the polynomial decay of the regularization parameter λ_n potentially needs to arbitrarily fast as $(2\alpha + d)/(2\alpha + \varrho)$ is unbounded for $\alpha, \varrho > 0$. On the one hand this shows why we need the additional constraint $\varrho \geq 1$ in Theorem 3.6, on the other hand we want to mention that the case $\varrho < 1$ is of little practical interest anyway. Nevertheless, the statement of Theorem 3.6 still holds with the rate $n^{-2\alpha/(2\alpha+1)} \log^{d+1} n$ in the case $\varrho < 1$.

We conclude this section with a comparison to existing results on learning rates for least-squares regression under the assumption, that the data has some low dimensional intrinisc structure, which is described by imposing constraints on $\operatorname{supp} \mathbf{P}_X$.

SVMs. In [36] the authors consider the case, where the input space $X \subset \mathbb{R}^d$ is a compact, smooth manifold of dimension ϱ and that the Bayes decision function is α-Hölder continuous with respect to the geodesic distance on X for some $\alpha \in (0, 1]$. They derive learning rates of the form $(\log^2(n)/n)^{\alpha/(8\alpha+4\varrho)}$. Under these assumptions on X, the assumption that the Bayes function is α-Hölder continuous with respect to the geodesic distance is equivalent to α-Hölder continuity with respect to the Euclidean distance, which is a consequence of [36, Lemma 1]. That is, Corollary 3.4 gives a significant improvement of the result in [36] under much less restrictive assumptions. Additionally, they do not consider adaptive parameter selection, i.e. the dimension ϱ of X and α need to be known.

Bayesian GP Regression. In [35] the authors consider the case of a compact ϱ-dimensional manifold $X \subset \mathbb{R}^d$, which is sufficiently regular and a C^α Bayes function, where $\alpha \leq 2$. For Bayesian regression using Gaussian processes with squared exponential covariance they derive the learning rate $n^{-2\alpha/(2\alpha+\varrho)} \log^{d+1} n$, which is identical to ours, but under much more restrictive assumptions on both α and $\operatorname{supp} \mathbf{P}_X$. Additionally, they present a training validation scheme for choosing the hyperparameters of the prior distribution. Under some additional technical assumption, which is hard to verify, they prove that this method achieves the same rates adaptively.

Local Polynomial Regression. In [3] the authors consider local linear regression in the setting of a differentiable ϱ-dimensional manifold $X \subset \mathbb{R}^d$ and a twice differentiable Bayes function. They further assume, that $\mathbf{P}_X$ has a differentiable density w.r.t. local charts and prove the learning rate $n^{-4/(4+\varrho)}$. They also state, that the result can be extended if the Bayes function is α-times differentiable using polynomials of degree $\alpha - 1$ to achieve the rate $n^{-2\alpha/(2\alpha+\varrho)}$. They propose a training validation scheme for bandwidth selection, but give no theoretical guarantees.

Tree-Based Regressor. In [17] the authors consider a tree-based locally constant regres-

sor. Based on the doubling dimension[2] ϱ of the input space X they prove that for a Lipschitz continuous target function their estimator achieves the learning rate $n^{-2/(2+k)}$ for a constant $k \in \mathcal{O}(\varrho \log \varrho)$. They achieve this rate adaptively using a training validation scheme, as well as a stopping criterion for building the tree.

k-**Nearest Neighbor.** In [20] the authors show, that under assumption **(DIM)** and for an α-Hölder continuous target function, where $0 < \alpha \le 1$, the k-NN rule achieves a learning rate of $n^{-2\alpha/(2\alpha+\varrho)}$. They only achieve this rate with knowledge on both, ϱ and α.

In the comparision above we only mentioned results, where the assumption on the intrinsic dimension is directly comparable to ours. For the sake of completeness we want to mention, that there are results, where the intrinsic dimension of the data is described by assuming that $\mathbf{P}_X$ is a so-called doubling measure. For example, this assumption is considered in [16] for the k-nearest neighbor method and in [18] for kernel regression. For an introduction to this notion of intrinsic dimensionality and the results built on it, we refer the reader to the cited articles.

As Remark 3.5 already indicates, some of our later results make it very easy to derive learning rates under slightly different regularity assumptions on the Bayes decision function under assumption **(DIM)**. Even more general, we derive an oracle inequality in Theorem A.10 for SVMs using Gaussian kernels under assumption **(DIM)** for generic loss functions L. Using Theorem A.10 it will then be an easy task to derive improved learning rates in the style of Corollary 3.4 for other regression tasks. For example, the conditional τ-quantile function can be estimated using the pinball loss

$$L_{\mathrm{pin}}(y,t) := \begin{cases} (1-\tau)(t-y), & \text{if } y < t, \\ \tau(y-t), & \text{if } t \ge y. \end{cases}$$

Using a variance bound, see Definition A.9, and a calibration inequality for the pinball loss from [30] and imposing some standard regularity assumptions on $f^*_{L_{\mathrm{pin}},\mathbf{P}}$ we can thus derive learning rates for the pinball loss, as well as bounds on the $L_q(\mathbf{P}_X)$ distance from $f_{D,\lambda,\gamma}$ to $f^*_{L_{\mathrm{pin}},\mathbf{P}}$ under assumption **(DIM)**. This would generalize some of the results from [9] in the sense, that we can substitute d with ϱ in their learning rates. The same is true for the conditional expectile function, which is estimated by the assymetric least-squares loss

$$L_{\mathrm{ALS}}(y,t) := \begin{cases} (1-\tau)(t-y)^2, & \text{if } y < t, \\ \tau(y-t)^2, & \text{if } t \ge y, \end{cases}$$

where $\tau \in (0,1)$, since in [12] the necessery variance bound and a calibration inequality for L_{ALS} are derived.

---

[2]cf. the short discussion after Example 2.4

4. Learning Rates for Classification

Throughout this section let $Y = \{-1, 1\}$ and fix a version $\eta : X \to [0, 1]$ of the posterior probability of $\mathbf{P}$, that is the probability measures $\mathbf{P}(\,\cdot\,|x)$ on Y defined by $\mathbf{P}(\{1\}|x) = \eta(x)$ for $x \in X$ fulfill

$$\mathbf{P}(A \times B) = \int_A \mathbf{P}(B|x) \, \mathrm{d}\mathbf{P}_X(x)$$

for all measurable sets $A \subset X$ and $B \subset Y$. With the posterior probability η the optimal labeling strategy $f^*_{L_{\mathrm{class}}, \mathbf{P}}$ can be expressed by $f^*_{L_{\mathrm{class}}, \mathbf{P}}(x) = \mathrm{sign}(2\eta(x) - 1)$, cf. [29, Example 2.4]. Regularity assumptions on $\mathbf{P}$ for classification can be described by smoothness of η. However, this is rather restrictive, since irregularities in a region, where η is close to 1 do not make the classification task harder. Additionally, smoothness assumptions on η are often combined with the assumption, that $\mathbf{P}_X$ has a Lebesgue-density. However, these types of assumptions are typically motivated by considering plug-in classification rules, which implicitly treat the binary classification problem as a regression problem. To avoid the latter, we first observe that it is intuitively hard to correctly predict a label y for $x \in X$ with $\eta(x) \approx 1/2$. Similarly, it may be harder to make a correct prediction for points $x \in X$ that are located near the decision boundary $\{\eta = 1/2\}$. The following defintion captures this intuition by imposing constraints on the mass and/or location of the critical level $\eta(x) \approx 1/2$.

Definition 4.1. Let $X_{-1} := \{x \in X : \eta(x) < 1/2\}$ and $X_1 := \{x \in X : \eta(x) > 1/2\}$ and define

$$\Delta(x) := \begin{cases} \mathrm{dist}(x, X_1) & \text{if } x \in X_{-1}, \\ \mathrm{dist}(x, X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{else}, \end{cases}$$

where $\mathrm{dist}(x, A) := \inf_{y \in A} \|x - y\|$. We say that $\mathbf{P}$ has *noise exponent* $q \in [0, \infty]$, if there exists a constant $c_{\mathrm{NE}} > 0$ such that

$$\mathbf{P}_X\big(\{x \in X : |2\eta(x) - 1| < t\}\big) \leq (c_{\mathrm{NE}} t)^q$$

for all $t \geq 0$. We further say that $\mathbf{P}$ has *margin-noise exponent* $\beta \in [0, \infty)$, if there exists a constant $c_{\mathrm{MNE}} > 0$ such that

$$\int_{\{x \in X : \Delta(x) < t\}} |2\eta(x) - 1| \, \mathrm{d}\mathbf{P}_X(x) \leq c_{\mathrm{MNE}} t^\beta$$

for all $t \geq 0$.

Note that neither the noise, nor the margin-noise exponent require the existence of a density of $\mathbf{P}_X$. For this reason they are perfectly suited to be combined with our assumption **(DIM)**. Also note, that the noise exponent goes back to [21] and since then has been widely adapted and modified in the context of non-parametric classification.

A collection of distributions having non-trivial (margin-) noise exponent can be found in [29, Sections 8.2 and 8.3]. Margin/noise conditions have already been studied in the context of SVMs, for example [31, 5]. See also [5, Section 3.2] and [4, Section 4] for a detailed exposition on the relation of margin/noise conditions to other notions of regularity and a comparison of resulting learning rates for various learning methods.

Before presenting the main results of this section, we briefly address the issue of the non-convexity of L_{class}. By Zhang's inequality, cf. [29, Theorem 2.31] we have

$$\mathcal{R}_{L_{\text{class}},\mathbf{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathbf{P}} \leq \mathcal{R}_{L_{\text{hinge}},\mathbf{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathbf{P}}$$

for all $f : X \to \mathbb{R}$. That is, by deriving learning rates for the loss L_{hinge}, we implicitly derive the same learning rates for the loss L_{class}.

Our main results of this section are organized as in the previous section: We first present a general oracle inequality and then derive learning rates by a suitable choice of hyperparameters. Finally, we show that a TV-SVM achieves the same rates adaptively.

Theorem 4.2. *Assume $\mathbf{P}$ satisfies **(DIM)** and that $\mathbf{P}$ has noise exponent $q \in [0, \infty]$ as well as margin-noise exponent $\beta \in [0, \infty)$. Then for the SVM using the hinge loss $L = L_{\text{hinge}}$ there exists a constant $C > 0$ such that for all $\tau > 0, n > 1, \lambda \in (0, 1)$, and $0 < \gamma^{-\varrho} \leq n$ we have*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C \left(\lambda \gamma^{-d} + \gamma^{\beta} + \lambda^{-1/\log n} \left(\frac{\gamma^{-\varrho}}{n} \right)^{\frac{q+1}{q+2}} \log^{d+1} n + \left(\frac{\tau}{n} \right)^{\frac{q+1}{q+2}} + \frac{\tau}{n} \right)$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$.

Using the theorem above, we can easily derive classification rates by choosing the hyperparameters λ and γ appropriately.

Corollary 4.3. *Let the assumptions of Theorem 3.3 be satisfied and set $\gamma_n = n^{-a}$ and $\lambda_n = n^{-b}$ with*

$$a = \frac{q+1}{\beta(q+2) + \varrho(q+1)} \quad and \quad b \geq \frac{(d+\beta)(q+1)}{\beta(q+2) + \varrho(q+1)}.$$

Then there exists a constant $C > 0$ such that for all $n > 1$ and $\tau \geq 1$ we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda_n,\gamma_n}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{\beta(q+1)}{\beta(q+2)+\varrho(q+1)}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ not less than $1 - \mathrm{e}^{-\tau}$.

The following theorem shows, that the rates of the corollary above can also be achieved adaptively without knowledge on $\mathbf{P}$.

Theorem 4.4. *Let A_n be a minimal $1/\log n$-net of $(0, 1]$ with $1 \in A_n$ and let B_n be a minimal $1/\log n$-net of $(0, d]$ with $d \in B_n$. Set $\Gamma_n := \{n^{-a} : a \in A_n\}$ and $\Lambda_n := \{n^{-b} : b \in B_n\}$. Let the assumptions of Theorem 4.2 be satisfied and assume $\varrho \geq 1$. Then there*

exists a constant $C > 0$, such that the TV-SVM using Λ_n and Γ_n satisfies for all $n > 1$ and $\tau \geq 1$,

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L,\mathbf{P}} \leq C\tau\, n^{-\frac{\beta(q+1)}{\beta(q+2)+\varrho(q+1)}} \log^{d+1} n$$

with probability $\mathbf{P}^n$ not less than $1 - \mathrm{e}^{-\tau}$.

The statements of Remark 3.7 on the size of the candidate set Λ_n and the need for the constraint $\varrho \geq 1$ apply one-to-one for Theorem 4.4.

We again conclude the section with a comparison to existing results on classification rates under a low intrinsic dimensionality assumption similar to ours, although they are much rarer than for regression.

SVMs. In [37] the authors assume that the input space $X \subset \mathbb{R}^d$ is a compact, smooth ϱ-dimensional manifold without boundary. They further assume, that $\mathrm{sgn}(2\eta - 1)$ is contained in the interpolation space $(L_1(X), W^{2,1}(X))_\theta$ for some $\theta \in (0,1]$, where $W^{2,1}(X)$ is a Sobolev-space on the manifold X and derive the learning rate $(\log^2(n)/n)^{\theta/(6\theta+\varrho)}$ for Gaussian SVMs. Of course, such a regularity assumption for a discrete-valued function is very restrictive and by no means comparable to our assumptions. In addition, their fastest possible rate is given by $n^{-1/7}$, while we derive learning rates up to n^{-1}.

Dyadic Decision Trees. In [27] dyadic decision trees are considered. They assume, that for a partition $\mathcal{P}_m$ of the input space $X = [0,1]^d$ into cubes of sidelength $1/m$, where m is a dyadic integer, every $A \in \mathcal{P}_m$ satisfies $\mathbf{P}_X(A) \leq c_0 m^{-\varrho}$. Additionally they assume, that the number of cubes in $\mathcal{P}_m$, that intersect the decision boundary $\{x \in X : \eta(x) = 1/2\}$ is bounded by $c_1 m^{\varrho-1}$, that is they impose an assumption similar to **(DIM)** on the decision boundary. They derive a learning rate of $(\log(n)/n)^{1/\varrho}$. Remarkably, the optimal choice of their only hyperparameter, the depth of the tree, does not depend on ϱ.

Acknowledgments

A. Proofs and Auxiliary Results

In the following we present the proofs of the statements in the main sections of this work. We first state and prove some auxiliary auxiliary results on entropy numbers for Gaussian RKHSs in Section A.1. Based on these results we derive an oracle inequality for Gaussian SVMs for generic loss functions under assumption **(DIM)** in Section A.2. Sections A.3 and A.4 contain the proofs of Sections 3 and 4, respectively.

A.1. Entropy Estimates

Our main tool for bounding the statistical error is based on the entropy numbers of $H_\gamma(X)$ with respect to the empirical $L_2(\mathbf{D})$-norm, where $\mathbf{D} := n^{-1} \sum_{i=1}^n \delta_{x_i}$ is the empirical measure associated to a sample $D = (x_1, \ldots, x_n) \in X^n$ drawn from $\mathbf{P}_X^n$. To this end, we first give a definition of entropy numbers, which are the inverse concept of covering numbers.

Definition A.1. Given normed spaces E, F and a bounded, linear operator $T : E \to F$, for $i \in \mathbb{N}$ the i-th dyadic entropy number of T is defined as

$$e_i(T) := \inf \left\{ \varepsilon > 0 : \exists x_1, \ldots, x_{2^{i-1}} \in F \text{ such that } TB_E \subset \bigcup_{j=1}^{2^{i-1}} (x_j + \varepsilon B_F) \right\}.$$

As one would expect, there is a close connection between entropy numbers and covering numbers. For example we have $e_i(T) \in \mathcal{O}(i^{-1/q})$ as $i \to \infty$ for some $q > 0$ if and only if $\log \mathcal{N}(T, \varepsilon) \in \mathcal{O}(\varepsilon^q)$ as $\varepsilon \to 0$, cf. [29, Lemma 6.21 and Exercise 6.8].
We are interested in bounding the averaged entropy numbers of the embedding id : $H_\gamma(X) \to L_2(\mathbf{D})$. More precisely, we need a polynomial bound of the form

$$\mathbf{E}_{D \sim \mathbf{P}_X^n} e_i(H_\gamma(X) \to L_2(\mathbf{D})) \le a\, i^{-\frac{1}{2p}} \quad \text{for all} \quad i \in \mathbb{N} \tag{9}$$

for some constants $a > 0$ and $p \in (0,1)$. A bound of the form (9) in turn is a tool for bounding Rademacher averages of $H_\gamma(X)$ using a standard symmetrization procedure and Dudley's chaining, see [29, Section 7.3]. We will derive (9) from ℓ_∞-covering numbers of the unit ball in $H_\gamma(X)$. Covering numbers for Gaussian RKHSs with respect to $\|\cdot\|_\infty$ are well understood, for example [19, Theorem 3] showed that

$$\log \mathcal{N}(\text{id} : H_\gamma([0,1]^d) \to \ell_\infty([0,1]^d), \varepsilon) \asymp \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log \log \frac{1}{\varepsilon}\right)^d} \quad \text{as} \quad \varepsilon \to 0. \tag{10}$$

However, we will rely on the slightly suboptimal bound

$$\log \mathcal{N}(\text{id} : H_\gamma([0,1]^d) \to \ell_\infty([0,1]^d), \varepsilon) \lesssim \log^{d+1} \frac{1}{\varepsilon} \quad \text{as} \quad \varepsilon \to 0, \tag{11}$$

since it is very hard to make use of the extra double logarithmic factor in (10). One can show that (11) actually implies super-polynomial decay of $\mathbf{E}_{D \sim \mathbf{P}_X^n} e_i(\text{id} : H_\gamma(X) \to L_2(\mathbf{D}))$, meaning (9) is fulfilled for arbitrarily small $p > 0$ with a constant a necessarily depending on p. For our statistical analysis we need to explicitly track the dependence of a on p and γ. This is done in the following theorem and its corollary, whose proofs can be found later in this subsection after we have established some auxiliary lemmas.

Theorem A.2. *There exists a universal constant K only depending on d, such that*

$$e_i(\text{id} : H_\gamma(X) \to \ell_\infty(X)) \le K^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \mathcal{N}_{\ell_\infty^d}(X, \gamma)^{\frac{1}{2p}} i^{-\frac{1}{2p}}$$

holds for all $i \in \mathbb{N}$, $p \in (0,1)$ and $\gamma > 0$.

Combining Theorem A.2 with assumption **(DIM)** easily gives a bound of the form (9) as desired.

Corollary A.3. *Let* **P** *satisfy* **(DIM)**. *Then there exists a constant K only depending on d, such that the bound*

$$\mathbf{E}_{D\sim\mathbf{P}_X^n} e_i(\mathrm{id} : H_\gamma(X) \to L_2(\mathbf{D})) \leq (C_{\dim}K)^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \gamma^{-\frac{\varrho}{2p}} i^{-\frac{1}{2p}}$$

holds for all $i \in \mathbb{N}$, $p \in (0, 1)$ and $\gamma \in (0, 1)$.

Note, that the behavior of the bandwidth γ in the entropy estimate above now depends on ϱ instead of d, compared to e.g. [34, Lemma 4.5], which is a crucial improvement.

As already mentioned, the proofs of Theorem A.2 and Corollary A.3 need some preparatory lemmas on kernels and their reproducing kernel Hilbert spaces. To this end, we recall some basic definitions and facts, which can be found in [29, Section 4]. Given a non-empty set X, a symmetric function $k : X \times X \to \mathbb{R}$ is a *kernel*, if it is positive definite, that is if for all $n \in \mathbb{N}$ and all choices $x_1, \ldots, x_n \in X$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ we have

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_j, x_i) \geq 0.$$

A Hilbert space H of functions $f : X \to \mathbb{R}$ is called a *reproducing kernel Hilbert space* (RKHS), if the evaluation functional $H \to \mathbb{R}$ defined by $f \mapsto f(x)$ is continuous for every $x \in X$. Every RKHS H has a unique *reproducing kernel* $k : X \times X \to \mathbb{R}$, i.e. $k(\cdot, x) \in H$ for all $x \in X$ and

$$\langle f, k(\cdot, x) \rangle_H = f(x) \quad \text{for all} \quad f \in H, x \in X, \tag{12}$$

and k is a kernel in the sense above. Property (12) is called the *reproducing property*. Conversely, for every kernel k there exists a unique RKHS H for which it is the reproducing kernel.

Lemma A.4. *Let k be a kernel on X with RKHS H and let $\psi : Y \to X$ be a map. Then $k_\psi(\cdot, \cdot) := k(\psi(\cdot), \psi(\cdot))$ is a kernel on Y with RKHS $H_\psi = \{f \circ \psi : f \in H\}$ and the map $V : H \to H_\psi$ defined by $f \mapsto f \circ \psi$ is a metric surjection. The norm in H_ψ can be computed by*

$$\|g\|_{H_\psi} = \inf\{\|f\|_H : f \text{ with } g = f \circ \psi\}.$$

If ψ is bijective, then V is an isometric isomorphism.

Proof. Let $\Phi : X \to H, x \mapsto k(x, \cdot)$ be the canonical feature map of k and define $\Phi_\psi : Y \to H, y \mapsto \Phi(\psi(y))$. Then by construction we have $\langle \Phi_\psi(y), \Phi_\psi(y') \rangle_H = k_\psi(y, y')$ for all $y, y' \in Y$, that is, Φ_ψ is a feature map of k_ψ. The first two assertions now follow from [29, Theorem 4.21]. For the third assertion additionally apply this result on ψ^{-1}. $\square$

Corollary A.5. *Let k be a kernel on $X \subset \mathbb{R}^d$, H its RKHS and $Y \subset X$. Then $H|_Y :=$ $\{f|_Y : f \in H\}$ is the RKHS of $k|_{Y \times Y}$ and the restriction $H \to H|_Y$ is a metric surjection.*

Proof. This follows from Lemma A.4 with $\psi : Y \to X$ being the inclusion. $\qquad\square$

A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called radial if there exists a function $\kappa : [0, \infty) \to \mathbb{R}$ such that $k(x, y) = \kappa(\|x - y\|)$ for all $x, y \in \mathbb{R}^d$. Radial kernels are a special case of translation invariant kernels, which by Bochner's theorem [24, Theorem IX.9] are characterized as the inverse Fourier transform of a finite Borel measure on $\mathbb{R}^d$. Similarly, by Schoenberg's theorem [26] $\kappa : [0, \infty) \to \mathbb{R}$ defines a radial kernel on $\mathbb{R}^d$ via $k(\cdot, \cdot) = \kappa(\| \cdot - \cdot \|)$ for every $d \in \mathbb{N}$, if and only if there exists a finite Borel measure μ on $[0, \infty)$ such that

$$\kappa(t) = \int_0^\infty e^{-xt^2} \, d\mu(x)$$

for all $t \in [0, \infty)$, i.e. radial kernels are mixtures of Gaussians. The following corollary shows, that RKHSs of radial kernels are in some sense translational and rotational invariant.

Corollary A.6. *Let k be a radial kernel on $\mathbb{R}^d$ and for $X \subset \mathbb{R}^d$ denote the restriction of k onto $X \times X$ by k_X and its RKHS by $H(X)$. Fix an $a \in \mathbb{R}^d$ and an orthogonal matrix $U \in \mathbb{R}^{d \times d}$. Then the operator $T : H(X) \to H(a + UX)$ defined by $Tf(x) = f(U^{-1}(x - a))$ is well-defined and an isometric isomorphism.*

Proof. This follows from Lemma A.4 with the map $\psi : a + UX \to X$ defined by $x \mapsto U^{-1}(x - a)$, since $k_X(\psi(\cdot), \psi(\cdot)) = k_{a+UX}(\cdot, \cdot)$. $\qquad\square$

Lemma A.7. *Let k be a kernel on X, H be its RKHS, and $X_1, X_2 \subset X$ disjoint subsets with $X_1 \cup X_2 = X$. Then for all $\varepsilon > 0$ we have*

$$\mathcal{N}_{\ell_\infty(X)}(B_H, \varepsilon) \le \mathcal{N}_{\ell_\infty(X_1)}(B_{H|_{X_1}}, \varepsilon) \cdot \mathcal{N}_{\ell_\infty(X_2)}(B_{H|_{X_2}}, \varepsilon)$$

for all $\varepsilon > 0$.

Proof. Let $f_1, \dots, f_n \in \ell_\infty(X_1)$ be a minimal ε-net of $B_{H|_{X_1}}$ and let $g_1, \dots, g_m \in \ell_\infty(X_2)$ be a minimal ε-net of $B_{H|_{X_2}}$. Let $f \in B_H$. Then by Corollary A.5 we have $f|_{X_l} \in H|_{X_l}$ with $\|f|_{X_l}\|_{H|_{X_l}} \le \|f\|_H \le 1$ for $l = 1, 2$. Hence, there exist i, j with $\|f|_{X_1} - f_i\|_{\ell_\infty(X_1)} \le \varepsilon$ and $\|f|_{X_2} - g_j\|_{\ell_\infty(X_2)} \le \varepsilon$. If we denote the zero-extensions of f_i, g_j to X by $\hat{f}_i, \hat{g}_j$, we see that $\{\hat{f}_i + \hat{g}_j : i = 1, \dots, n, \, j = 1, \dots, m\}$ is a ε-net of B_H with cardinality $n \cdot m$. $\qquad\square$

Corollary A.8. *For all $\gamma > 0$ and $\varepsilon > 0$ we have*

$$\log \mathcal{N}(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X), \varepsilon)$$
$$\le \mathcal{N}_{\ell_\infty^d}(X, \gamma) \cdot \log \mathcal{N}\left(\mathrm{id} : H_1([-1, 1]^d) \to \ell_\infty([-1, 1]^d), \varepsilon\right).$$

Proof. Let $x_1, \dots, x_n \in \mathbb{R}^d$ be a minimal γ-net of X w.r.t. ℓ_∞^d. We partition X into $X_1, \dots, X_n$, where X_j consists of the points $x \in X$ that are ℓ_∞^d-closest to x_j. Here we break ties, for example, in favor of a smaller index j. Combining Lemma A.7, Corollary A.5 and Corollary A.6 we get

$$
\begin{aligned}
\log \mathcal{N}_{\ell_\infty}\left(B_{H_\gamma(X)}, \varepsilon\right) &\leq \sum_{j=1}^n \log \mathcal{N}_{\ell_\infty}\left(B_{H_\gamma(X_j)}, \varepsilon\right) \\
&\leq \sum_{j=1}^n \log \mathcal{N}_{\ell_\infty}\left(B_{H_\gamma(x_j + \gamma[-1,1]^d)}, \varepsilon\right) \\
&= n \log \mathcal{N}_{\ell_\infty}\left(B_{H_\gamma(\gamma[-1,1]^d)}, \varepsilon\right).
\end{aligned}
$$

The result now follows from [29, Proposition 4.37], which states that the scaling operator $\tau_\gamma : H_\gamma(\gamma[-1,1]^d) \to H_1([-1,1]^d)$ defined by $\tau_\gamma f(x) = f(\gamma x)$ is an isometric isomorphism. $\qquad\square$

Proof of Theorem A.2. For $f \in B_{H_1([-1,1]^d)}$, the reproducing property together with the Cauchy–Schwarz inequality implies

$$
\begin{aligned}
|f(x)| = |\langle f, k_1(x, \cdot)\rangle_{H_1([-1,1]^d)}| &\leq \|f\|_{H_1([-1,1]^d)} \|k_1(x, \cdot)\|_{H_1([-1,1]^d)} \\
&\leq \sqrt{k_1(x,x)} = 1
\end{aligned}
$$

for all $x \in [-1,1]^d$ and consequently we find $\mathcal{N}(B_{H_1([-1,1]^d)}, \varepsilon) = 1$ for all $\varepsilon \geq 1$. Furterhmore, by [19, Theorem 3] there exists a constant $K \geq 1$ such that

$$
\log \mathcal{N}\left(B_{H_1([-1,1]^d)}, \varepsilon\right) \leq K \log^{d+1} \frac{2}{\varepsilon}
$$

for all $\varepsilon \in (0,1]$. Some elementary calculations show that

$$
\sup_{\varepsilon \in (0,2)} \varepsilon^q \log^{d+1} \frac{2}{\varepsilon} = 2^q \left(\frac{d+1}{eq}\right)^{d+1},
$$

which combined with Corollary A.8 and the estimate above gives us

$$
\begin{aligned}
\log \mathcal{N}(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X), \varepsilon) &\leq K \mathcal{N}(X, \gamma) \log^{d+1} \frac{2}{\varepsilon} \\
&\leq 4 K \mathcal{N}(X, \gamma) \left(\frac{d+1}{eq}\right)^{d+1} \varepsilon^{-q}
\end{aligned} \tag{13}
$$

for $\varepsilon > 0$ and $q \in (0,2)$. As a final step we convert the latter bound on the covering numbers of $\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)$ into a bound on the entropy numbers. To this end, we fix an $i \geq 2$ and define $\varepsilon > 0$ by $\exp(a/\varepsilon^q) = 2^{i-1}$, where

$$
a := 4 K \mathcal{N}(X, \gamma) \left(\frac{d+1}{eq}\right)^{d+1}.
$$

By (13) this implies

$$e_i(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq \left(\frac{(i-1)\log 2}{a} \right)^{-\frac{1}{q}} \leq \left(\frac{2a}{\log 2} \right)^{\frac{1}{q}} i^{-\frac{1}{q}}$$
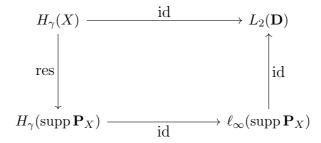
for all $i \geq 2$. Since $e_1(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq 1$ we get

$$e_i(\mathrm{id} : H_\gamma(X) \to \ell_\infty(X)) \leq \left(\frac{\mathcal{N}(X,\gamma)8K}{\log 2} \left(\frac{d+1}{eq} \right)^{d+1} \right)^{\frac{1}{q}} i^{-\frac{1}{q}}$$

for all $i \in \mathbb{N}$. Now substitute $2p = q$ and absorb all irrelevant constants into a new constant K. $\qquad \square$

For the proof of Corollary A.3 we will use the basic fact that entropy numbers are dominated by decompositions, that is, if we can decompose $T : E \to F$ into $T = RS$ with bounded, linear operators $S : E \to \tilde{F}$, $R : \tilde{F} \to F$, and an intermediate normed space $\tilde{F}$, then we have $e_i(T) \leq \|R\| e_i(S)$ as well as $e_i(T) \leq e_i(R)\|S\|$ for all $i \in \mathbb{N}$, c.f. [7, Section 1.3].

Proof of Corollary A.3. Consider the decomposition of $\mathrm{id} : H_\gamma(X) \to L_2(\mathbf{D})$ for a sample $D \in (\mathrm{supp}\,\mathbf{P}_X)^n$ described by the commutative diagram

$$
\begin{array}{ccc}
H_\gamma(X) & \xrightarrow{\quad\mathrm{id}\quad} & L_2(\mathbf{D}) \\
\downarrow{\scriptstyle\mathrm{res}} & & \uparrow{\scriptstyle\mathrm{id}} \\
H_\gamma(\mathrm{supp}\,\mathbf{P}_X) & \xrightarrow{\quad\mathrm{id}\quad} & \ell_\infty(\mathrm{supp}\,\mathbf{P}_X)
\end{array}
$$

where $\mathrm{res} : H_\gamma(X) \to H_\gamma(\mathrm{supp}\,\mathbf{P}_X)$ is the restriction operator. We have $\|\mathrm{res} : H_\gamma(X) \to H_\gamma(\mathrm{supp},\mathbf{P}_X)\| \leq 1$ by Corollary A.5 and trivially also $\|\mathrm{id} : \ell_\infty(\mathrm{supp}\,\mathbf{P}_X) \to L_2(\mathbf{D})\| \leq 1$. Theorem A.2 then implies

$$e_i(\mathrm{id} : H_\gamma(X) \to L_2(\mathbf{D})) \leq K^{\frac{1}{2p}} p^{-\frac{d+1}{2p}} \mathcal{N}_{\ell_\infty^d}(\mathrm{supp}\,\mathbf{P}_X, \gamma)^{\frac{1}{2p}} i^{-\frac{1}{2p}}$$

for all $i \in \mathbb{N}$, $\gamma > 0$ and $p \in (0,1)$ for $\mathbf{P}_X^n$-almost all $D \in (X \times Y)^n$. Now combine the above bound with assumption **(DIM)** and take the expectation w.r.t. $D \sim \mathbf{P}_X^n$. $\qquad \square$

A.2. A General Oracle Inequality

In this section we derive a general oracle inequality for bounding the excess risk of our estimator $f_{D,\lambda,\gamma}$. A main ingredient for this oracle inequality is a so-called variance bound, which has been proven to be a useful tool for deriving fast learning rates, see for example [6], especially the discussion in Section 5.2 therein.

Definition A.9. Let L be a loss that can be clipped at $M > 0$ and let $\mathcal{F}$ be some function class of measurable functions $f : X \to \mathbb{R}$. Assume there exists a Bayes decision function $f^*_{L,\mathbf{P}} : X \to [-M, M]$. We say, that the supremum bound is satisfied, if there exists a constant $B > 0$, such that $L(y,t) \leq B$ for all $(y,t) \in Y \times [-M, M]$. We further say, that the variance bound is satisfied, if there exist $\vartheta \in [0,1]$ and $V \geq B^{2-\vartheta}$, such that

$$\mathbf{E}(L \circ \widehat{f} - L \circ f^*_{L,\mathbf{P}})^2 \leq V \cdot \left(\mathbf{E}\, L \circ \widehat{f} - L \circ f^*_{L,\mathbf{P}}\right)^{\vartheta}$$

for all $f \in \mathcal{F}$.

Intuitively, the variance bound ensures a small variance of the excess risk, whenever our estimator is close to the optimum.

The main results of Sections 3 and 4 are based on the following theorem, which may also serve as a blueprint for deriving learning rates for Gaussian SVMs under assumption **(DIM)** using other loss functions than the ones considered in this paper. To this end we first introduce some minimum regularity assumption, that we need to impose on the loss function L. We say that a loss function $L : Y \times \mathbb{R} \to [0, \infty)$ is locally Lipschitz continuous if for every $a > 0$ the functions $L(y, \cdot)|_{[-a,a]}, y \in Y$ are uniformly Lipschitz continuous, that is

$$|L|_{a,1} := \sup_{\substack{s,t \in [-a,a], s \neq t \\ y \in Y}} \frac{|L(y,t) - L(y,s)|}{|t-s|} < \infty.$$

The following theorem establishes an oracle inequality for Gaussian SVMs under assumption **(DIM)** for generic loss functions.

Theorem A.10. *Assume L is a locally Lipschitz continuous loss that can be clipped at $M > 0$ and that the supremum and variance bounds are satisfied for constants $B > 0, \vartheta \in [0,1]$, and $V \geq B^{2-\vartheta}$. Furthermore, assume X satisfies **(DIM)** for $C_{\dim} > 0$ and $0 < \varrho \leq d$ and fix an $f_0 \in H_\gamma(X)$ and a $B_0 \geq B$ with $\|L \circ f_0\|_\infty \leq B_0$. Then there exists a constant K such that for all $n \in \mathbb{N}, \gamma \in (0,1), \lambda > 0, p \in (0, 1/2]$ and $\tau > 0$ we have*

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq 9(\lambda\|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}})$$
$$+ K\left(\frac{p^{-d-1}\gamma^{-\varrho}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$.

Proof. By [29, Theorem 7.23] we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq 9(\lambda\|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}})$$
$$+ K(p)\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}$$

with probability not less than $1 - 3\mathrm{e}^{-\tau}$, where

$$a = \max\left\{(C_{\dim}K)^{\frac{1}{2p}}p^{-\frac{d+1}{2p}}\gamma^{-\frac{\varrho}{2p}}, B\right\} \leq p^{-\frac{d+1}{2p}}\gamma^{-\frac{\varrho}{2p}}\max\left\{(C_{\dim}K)^{\frac{1}{2p}}, B\right\}$$

with the constant K from Corollary A.3 and $K(p)$ is given by

$$K(p) = \max\left\{2B, 3\left(30 \cdot 2^p C_1(p)|L|_{M,1}^p V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta-\vartheta p}},\right.$$
$$\left. 90 \cdot 120^p C_2^{1+p}(p)|L|_{M,1}^p B^{1-p}\right\}.$$

The constants $C_1(p)$ and $C_2(p)$ are given by

$$C_1(p) := \frac{2\sqrt{\log 256}C_p^p}{(\sqrt{2}-1)(1-p)2^{\frac{p}{2}}}, \quad C_2(p) = \left(\frac{8\sqrt{\log 16}C_p^p}{(\sqrt{2}-1)(1-p)4^p}\right)^{\frac{2}{1+p}},$$

where

$$C_p = \frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}} \cdot \frac{1-p}{p},$$

which can be tracked in [29, proof of Theorem 7.16]. It was shown in [12, Proof of Theorem 7] that $C_1(p)$ and $C_2^{1+p}(p)$ are uniformly bounded in $p \in (0, \frac{1}{2}]$, more precisely, we have

$$\sup_{p \in (0,\frac{1}{2}]} C_1(p) \leq 46\mathrm{e} \quad \text{and} \quad \sup_{p \in (0,\frac{1}{2}]} C_2^{1+p}(p) \leq 1035\mathrm{e}^2,$$

which implies $\tilde{K} := \sup_{p \in (0,1/2]} K(p) < \infty$. This yields the estimate

$$K(p)\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} \leq \tilde{K}\max\left\{C_{\dim}K, B^{2p}\right\}^{\frac{1}{2-p-\vartheta+\vartheta p}}\left(\frac{p^{-d-1}\gamma^{-\varrho}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}.$$

The result now follows since $\sup_{\vartheta \in [0,1], p \in (0,1/2]} \max\{C_{\dim}K, B^{2p}\}^{1/(2-p-\vartheta+\vartheta p)} < \infty$. $\qquad\square$

A.3. Proofs related to Section 3

Lemma A.11. *Let $f : (a,b) \to \mathbb{R}$ be s-times continuously differentiable. Furthermore, fix $x \in (a,b)$, $h > 0$, and $k \in \{1,\ldots,s\}$ with $(x, x+sh) \subset (a,b)$. Then there exists a $\xi \in (x, x+kh)$ such that $h^{-s}\Delta_h^s f(x) = h^{k-s}\Delta_h^{s-k}f^{(k)}(\xi)$.*

Proof. Because of $\Delta_h^s = \Delta_h \Delta_h^{s-1}$ we can apply the mean value theorem to the function $h^{-1}\Delta_h^{s-1}f$, which gives us $h^{-s}\Delta_h^s f(x) = h^{1-s}\Delta_h^{s-1}f'(\xi)$ for some $\xi \in (x, x+h)$. Note that in the last step we used $\frac{\mathrm{d}}{\mathrm{d}x}\Delta_h f(x) = \Delta_h f'(x)$. That is, we have proven the assertion for $k = 1$. Now we can iterate this argument by applying the mean value theorem to $h^{-1}\Delta_h^{s-2}f'$ and so on. $\qquad\square$

Proof of Proposition 3.2. Fix an $x \in X$ and an $h \in \mathbb{R}^d$ with $\|h\| < \delta/(s-1)$, where $s := \lfloor \alpha \rfloor + 1$, and define the univariate function $F(t) := f(x + th)$. Then we have $\Delta_h^s f(x) = \Delta_1^s F(0) = \Delta_1 F^{(s-1)}(\xi)$ by Lemma A.11 for some $\xi \in (0, s-1)$. The $(s-1)$-th derivative of F is given by

$$\frac{\mathrm{d}^{s-1}}{\mathrm{d}t^{s-1}} F(t) = \sum_{|\nu| = s-1} \frac{(s-1)!}{\nu!} h^\nu \, \partial^\nu f(x + th).$$

When α is not an integer this leads us to the estimate

$$
\begin{aligned}
|\Delta_1 F^{(s-1)}(\xi)| &= |F^{(s-1)}(\xi + 1) - F^{(s-1)}(\xi)| \\
&= \left| \sum_{|\nu| = s-1} \frac{(s-1)!}{\nu!} h^\nu \big(\partial^\nu f(x + (\xi+1)h) - \partial^\nu f(x + \xi h) \big) \right| \\
&\leq \max_{|\nu| = s-1} |\partial^\nu f|_{\beta, X+\delta} \|h\|^\beta \left| \sum_{|\nu| = s-1} \frac{(s-1)!}{\nu!} \prod_{j=1}^{d} |h_j|^{\nu_j} \right| \\
&= \max_{|\nu| = s-1} |\partial^\nu f|_{\beta, X+\delta} \|h\|^\beta \|h\|_{\ell_1^d}^{s-1} \\
&\leq d^{\frac{s-1}{2}} \max_{|\nu| = s-1} |\partial^\nu f|_{\beta, X+\delta} \|h\|^\alpha
\end{aligned}
$$

using the definition of the Hölder seminorm (5) in the first inequality, the multinomial theorem in the next step, and $\beta + s - 1 = \alpha$ in the last step. When α is an integer we analogously get

$$|\Delta_1 F^{(s-1)}(\xi)| \leq 2 d^{\frac{s-1}{2}} \max_{|\nu| = s-1} \|\partial^\nu f|_{X+\delta}\|_\infty \|h\|^\alpha.$$

In both cases this implies

$$\omega_{s, L_2(\mu)}(f, t) = \sup_{\|h\| \leq t} \|\Delta_h^s f\|_{L_2(\mu)} \leq C t^\alpha$$

for $t < \delta/(s-1)$. For $t \geq \delta/(s-1)$ we can estimate

$$
\begin{aligned}
t^{-\alpha} \omega_{s, L_2(\mu)}(f, t) &\leq \left(\frac{\delta}{s-1} \right)^{-\alpha} \sup_{\|h\| \leq t} \|\Delta_h^s f\|_{L_2(\mu)} \\
&\leq \left(\frac{\delta}{s-1} \right)^{-\alpha} \sup_{\|h\| \leq t} \sum_{j=0}^{s} \binom{s}{j} \|f(\cdot + jh)\|_{L_2(\mu)} \\
&\leq \left(\frac{\delta}{s-1} \right)^{-\alpha} 2^s \mu(X) \|f\|_\infty
\end{aligned}
$$

which gives us $|f|_{B_{2,\infty}^\alpha(\mu)} = \sup_{t>0} t^{-\alpha} \omega_{s, L_2(\mu)}(f, t) < \infty$. $\qquad\square$

To prove Theorem 3.3 we need a suitable function $f_0 \in H_\gamma(X)$ bounding the approximation error. To this end, we first collect some facts on Gaussian RKHSs, which are a

summary of [29, Theorem 4.21, Lemma 4.45, and Proposition 4.46]. By introducing the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ defined by

$$K_\gamma(x) := \left(\frac{2}{\gamma\sqrt{\pi}} \right)^{\frac{d}{2}} \exp(-2\gamma^{-2}\|x\|^2), \qquad x \in \mathbb{R}^d,$$

the Gaussian RKHS $H_\gamma(X)$ can be characterized as the image of the convolution operator $L_2(\mathbb{R}^d) \to H_\gamma(X)$ defined by $g \mapsto K_\gamma * g$. The $H_\gamma(X)$-norm can be computed by $\|f\|_{H_\gamma(X)} = \inf\{\|g\|_{L_2(\mathbb{R}^d)} : f = K_\gamma * g\}$. Furthermore, for $0 < \gamma_1 < \gamma_2 < \infty$ the space $H_{\gamma_2}(X)$ is continuously embedded into $H_{\gamma_1}(X)$ with

$$\|\mathrm{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\| \leq \left(\frac{\gamma_2}{\gamma_1} \right)^{\frac{d}{2}}. \tag{14}$$

We will further make use of integration in spherical coordinates [13, Theorem 2.49]. Namely, for $f \in L_1(\mathbb{R}^d)$ or $f \geq 0$ we have

$$\int_{R^d} f(x) \, \mathrm{d}x = \int_0^\infty \int_{\mathbb{S}^{d-1}} f(r\omega) r^{d-1} \, \mathrm{d}\sigma(\omega) \, \mathrm{d}r, \tag{15}$$

where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and σ is the surface measure on $\mathbb{S}^{d-1}$. For radial functions f, that is $f(x) = g(\|x\|)$ Equation (15) simplifies to

$$\int_{\mathbb{R}^d} f(x) \, \mathrm{d}x = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty g(r) r^{d-1} \, \mathrm{d}r, \tag{16}$$

since $\sigma(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$, see e.g. [13, Proposition 2.54]. Using (16) one can easily check, that

$$\int_{\mathbb{R}^d} (\gamma\sqrt{\pi})^{-\frac{d}{2}} K_\gamma(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} \gamma^{-d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|x\|^2 \right) \, \mathrm{d}x = 1, \tag{17}$$

which we will rely on later. Finally, we define

$$G := \sum_{j=1}^s \binom{s}{j} (-1)^{1-j} (j\gamma\sqrt{\pi})^{-\frac{d}{2}} K_{j\gamma}. \tag{18}$$

The following lemma bounds the approximation error of a Gaussian SVM using the least-squares loss.

Lemma A.12. *For $f \in L_2(\mathbb{R}^d)$ with $|f|_{B^\alpha_{2,\infty}(\mathbf{P}_X)} < \infty$ we have*

$$\|G * f - f\|^2_{L_2(\mathbf{P}_X)} \leq |f|^2_{B^\alpha_{2,\infty}(\mathbf{P}_X)} 2^{-\alpha} \left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right)^2 \gamma^{2\alpha}.$$

Proof. First compute

$$G * f(x) = \int_{\mathbb{R}^d} \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} (j\gamma)^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2(j\gamma)^{-2}\|y\|^2\right) f(x+y) \, \mathrm{d}y$$

$$= \int_{\mathbb{R}^d} \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x+jh) \, \mathrm{d}h.$$

Using equation (17) this implies

$$G * f(x) - f(x) = G * f(x) - \int_{\mathbb{R}^d} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x) \, \mathrm{d}h$$

$$= \int_{\mathbb{R}^d} \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x+jh) \, \mathrm{d}h$$

$$- \int_{\mathbb{R}^d} \gamma^{-d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) f(x) \, \mathrm{d}h$$

$$= \int_{\mathbb{R}^d} (-1)^{1-s} \left(\frac{2}{\gamma^2 \pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) \Delta_h^s f(x) \, \mathrm{d}h.$$

With this identity we can bound our desired $L_2(\mathbf{P}_X)$-norm by

$$\|G * f - f\|_{L_2(\mathbf{P}_X)}^2$$

$$= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2 \pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) \Delta_h^s f(x) \, \mathrm{d}h \right)^2 \mathrm{d}\mathbf{P}_X(x)$$

$$\leq \left(\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2 \pi}\right)^d \left(\exp\left(-2\gamma^{-2}\|h\|^2\right) \Delta_h^s f(x) \right)^2 \mathrm{d}\mathbf{P}_X(x) \right)^{\frac{1}{2}} \mathrm{d}h \right)^2$$

$$= \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2 \pi}\right)^{\frac{d}{2}} \exp\left(-2\gamma^{-2}\|h\|^2\right) \|\Delta_h^s f\|_{L_2(\mathbf{P}_X)} \, \mathrm{d}h \right)^2$$

using Minkowski's integral inequality. With our assumptions on f we can further bound this by

$$\|G * f - f\|_{L_2(\mathbf{P}_X)}^2 \leq \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\gamma^{-2}\|h\|^2\right) \omega_{s,L_2(\mathbf{P}_X)}(f, \|h\|) \, \mathrm{d}h \right)^2$$

$$\leq |f|_{B_{2,\infty}^\alpha(\mathbf{P}_X)}^2 \left(\frac{2}{\gamma^2 \pi}\right)^d \left(\int_{\mathbb{R}^d} \exp\left(-2\gamma^{-2}\|h\|^2\right) \|h\|^\alpha \, \mathrm{d}h \right)^2$$

which leaves us with computing the integral in the last step. This is done using spherical

coordinates, which gives us

$$\int_{\mathbb{R}^d} \exp\left(-2\gamma^{-2}\|h\|^2\right)\|h\|^\alpha \, \mathrm{d}h = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \exp\left(-2\gamma^{-2}r^2\right) r^{\alpha+d-1} \, \mathrm{d}r$$

$$= \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \frac{1}{2}\left(\frac{\gamma}{\sqrt{2}}\right)^{\alpha+d} \mathrm{e}^{-u} u^{\frac{\alpha+d}{2}-1} \, \mathrm{d}u$$

$$= \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \left(\frac{\gamma}{\sqrt{2}}\right)^{\alpha+d} \Gamma\left(\frac{\alpha+d}{2}\right).$$

Combining these considerations we get

$$\|G * f - f\|^2_{L_2(\mathbf{P}_X)} \leq |f|^2_{B^\alpha_{2,\infty}(\mathbf{P}_X)} 2^{-\alpha} \left(\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2 \gamma^{2\alpha}.$$

$\square$

The following lemma bounds the regularization term.

Lemma A.13. *For $f \in L_2(\mathbb{R}^d)$ we have $\|G * f\|_{H_\gamma(X)} \leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} 2^s \|f\|_{L_2(\mathbb{R}^d)}$.*

Proof. Because of the embedding property (14) we have

$$\|G * f\|_{H_\gamma(X)} \leq \sum_{j=1}^s \binom{s}{j} (j\gamma\sqrt{\pi})^{-\frac{d}{2}} \|K_{j\gamma} * f\|_{H_\gamma(X)}$$

$$\leq \sum_{j=1}^s \binom{s}{j} (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|K_{j\gamma} * f\|_{H_{j\gamma}(X)}$$

$$\leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|f\|_{L_2(\mathbb{R}^d)} \sum_{j=1}^s \binom{s}{j}$$

$$\leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} 2^s \|f\|_{L_2(\mathbb{R}^d)}.$$

$\square$

Proof of Theorem 3.3. For $Y = [-M, M]$ the least-squares loss satisfies the supremum/ variance bound for the constants $B = 4M^2, V = 16M^2$ and $\vartheta = 1$ by [29, Example 7.3]. Theorem A.10 therefore gives us for $\lambda > 0, \gamma \in (0, 1)$, and $p \in (0, 1/2]$

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathbf{P}} \leq 9(\lambda\|f_0\|^2_{H_\gamma(X)} + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}^*_{L,\mathbf{P}})$$

$$+ Kp^{-d-1}\gamma^{-\varrho}\lambda^{-p}n^{-1} + \frac{(3456M^2 + 15B_0)\tau}{n}$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$. To bound the approximation error we set $f_0 := G * f^*_{L,\mathbf{P}}$, where G is defined by (18). First we determine B_0, i.e. a bound on $\sup_{(x,y) \in X \times Y} |y - f_0(x)|^2$. By Young's convolution inequality we have

$$\|f_0\|_{L_\infty(\mathbb{R}^d)} = \|G * f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \leq \|f^*_{L,\mathbf{P}}\|_{L_\infty(\mathbb{R}^d)} \cdot \|G\|_{L_1(\mathbb{R}^d)}$$

and by using Equation (17) we get

$$\|G\|_{L_1(\mathbb{R}^d)} \leq \sum_{j=1}^{s} \binom{s}{j} (j\gamma\sqrt{\pi})^{-\frac{d}{2}} \|K_{j\gamma}\|_{L_1(\mathbb{R}^d)} = \sum_{j=1}^{s} \binom{s}{j} \leq 2^s.$$

Consequently, we get

$$\sup_{(x,y)\in X\times Y} |f_0(x) - y|^2 \leq \left(2^s \|f_{L,\mathbf{P}}^*\|_{L_\infty(\mathbb{R}^d)} + M\right)^2,$$

i.e. we can set

$$B_0 := \max\left\{ \left(2^s \|f_{L,\mathbf{P}}^*\|_{L_\infty(\mathbb{R}^d)} + M\right)^2, 4M^2 \right\}.$$

Using Lemma A.12 and Lemma A.13 and setting $p = \log 2/(2\log n) \leq 1/2$ we get

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}_{L,\mathbf{P}}^* \leq 9c_1\lambda\gamma^{-d} + 9c_2\gamma^{2\alpha} + 3^{d+1}K\lambda^{-1/\log n}\gamma^{-\varrho}n^{-1}\log^{d+1}n + c_3\tau n^{-1}$$

for $\lambda, \gamma \in (0,1)$, where the constants c_1 and c_2 are given in Lemma A.12 and A.13 and $c_3 := 3456M^2 + 15B_0$. We deduce that with $C := \max\{9c_1, 9c_2, 3^{d+1}K, c_3\}$ our assertion holds. $\quad\square$

Proof of Corollary 3.4. With the choices of λ_n and γ_n as stated an application of Theorem 3.3 gives

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}_{L,\mathbf{P}}^* \leq C\left(n^{-b}n^{\frac{d}{2\alpha+\varrho}} + n^{-\frac{2\alpha}{2\alpha+\varrho}} + \mathrm{e}^b n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1}n + \frac{\tau}{n}\right)$$

$$\leq C\left(2n^{-\frac{2\alpha}{2\alpha+\varrho}} + \mathrm{e}^b n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1}n + \frac{\tau}{n}\right)$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$ for all $n \in \mathbb{N}$. A substitution of τ then easily proves the assertion. $\quad\square$

Proof of Theorem 3.6. We define $\gamma_n := n^{-1/(2\alpha+\varrho)}$ and $\lambda_n := n^{-(2\alpha+d)/(2\alpha+\varrho)}$. By [29, Theorem 7.2], which states an oracle inequality for empirical risk minimization over finite hypothesis sets, we have

$$\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}_{L,\mathbf{P}}^* \leq 6 \min_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \left(\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda,\gamma}) - \mathcal{R}_{L,\mathbf{P}}^*\right)$$

$$+ \frac{512M^2(\tau + \log(1 + |\Lambda_n \times \Gamma_n|))}{n - m}$$

$$\leq 6\left(\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_0,\gamma_0}) - \mathcal{R}_{L,\mathbf{P}}^*\right)$$

$$+ \frac{2048M^2(\tau + \log(1 + |\Lambda_n \times \Gamma_n|))}{n}$$

with probability $\mathbf{P}^{n-m}$ not less than $1 - \mathrm{e}^{-\tau}$, where in the last step we picked $\gamma_0 := n^{-a} \in \Gamma_n$ and $\lambda_0 := n^{-b} \in \Lambda_n$ for values a and b, which we will specify in a moment. An application of Theorem 3.3 gives us

$$
\begin{aligned}
\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_0,\gamma_0}) - \mathcal{R}_{L,\mathbf{P}}^* &\le C\left(\lambda_0\gamma_0^{-d} + \gamma_0^{2\alpha} + b^{d+1}\gamma_0^{-\varrho}m^{-1}\log^{d+1} n + \frac{\tau}{m}\right) \\
&\le C\left(\lambda_0\gamma_0^{-d} + \gamma_0^{2\alpha} + 2b^{d+1}\gamma_0^{-\varrho}n^{-1}\log^{d+1} n + \frac{2\tau}{n}\right)
\end{aligned}
$$

with probability $\mathbf{P}^m$ not less than $1 - 3\mathrm{e}^{-\tau}$. Now let $\lambda_0 = n^{-d}$ and let $a \in A_n$ satisfy $1/(2\alpha + \varrho) \le a \le 1/(2\alpha + \varrho) + 1/\log n$, which implies

$$
\begin{aligned}
\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_0,\gamma_0}) - \mathcal{R}_{L,\mathbf{P}}^* &\le C\left(\mathrm{e}^d\lambda_n\gamma_n^{-d} + \gamma_n^{2\alpha} + 2d^{d+1}\mathrm{e}^\varrho\gamma_n^{-\varrho}n^{-1}\log^{d+1} n + \frac{2\tau}{n}\right) \\
&= C\left(\mathrm{e}^d n^{-\frac{2\alpha}{2\alpha+\varrho}} + n^{-\frac{2\alpha}{2\alpha+\varrho}} + 2d^{d+1}\mathrm{e}^\varrho n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1} n + \frac{2\tau}{n}\right)
\end{aligned}
$$

with probability $\mathbf{P}^m$ not less than $1 - 3\mathrm{e}^{-\tau}$. Combining these inequalities and using that $|\Lambda_n \times \Gamma_n| \in \mathcal{O}(\log^2 n)$ we get

$$
\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}_{L,\mathbf{P}}^* \le c_1\left(n^{-\frac{2\alpha}{2\alpha+\varrho}}\log^{d+1} n + \frac{\tau}{n}\right) + c_2\left(\frac{\tau}{n} + \frac{\log n}{n}\right)
$$

with probability $\mathbf{P}^n$ not less than $(1 - \mathrm{e}^{-\tau})(1 - 3\mathrm{e}^{-\tau}) \ge 1 - 4\mathrm{e}^{-\tau}$ for all $n > 2$. $\qquad\square$

A.4. Proofs related to Section 4

Proof of Theorem 4.2. The distribution $\mathbf{P}$ obviously satisfies the supremum bound for $B = 2$ and by [29, Theorem 8.24] the variance bound is satisfied for $V = 6c_{\mathrm{NE}}^{q/(q+1)}$ and $\vartheta = q/(q + 1)$. Theorem A.10 then gives us for $\lambda > 0, \gamma \in (0, 1)$, and $p \in (0, 1/2]$

$$
\begin{aligned}
\mathcal{R}_{L,\mathbf{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}_{L,\mathbf{P}}^* \le C\bigg(&\lambda\|f_0\|_{H_\gamma(X)}^2 + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}_{L,\mathbf{P}}^* \\
&+ \left(\frac{p^{-d-1}\lambda^{-p}\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}\bigg) + \frac{15B_0\tau}{n}
\end{aligned}
$$

with probability not less than $1 - 3\mathrm{e}^{-\tau}$. In [29, Theorem 8.18] a function $f_0 \in H_\gamma(X)$ with $\|f_0\|_\infty \le 1$ and

$$
\lambda\|f_0\|_{H_\gamma(X)}^2 + \mathcal{R}_{L,\mathbf{P}}(f_0) - \mathcal{R}_{L,\mathbf{P}}^* \le c_1\lambda\gamma^{-d} + c_2 c_{\mathrm{MNE}}\gamma^\beta,
$$

is constructed, which implies for $\lambda \in (0, 1)$ and $0 < \gamma^{-\varrho}/n \le 1$

$$
\mathcal{R}_{L,\mathbf{P}}(f_{D,\lambda,\gamma}) - \mathcal{R}_{L,\mathbf{P}}^* \le \tilde{C}\left(\lambda\gamma^{-d} + \gamma^\beta + p^{-d-1}\lambda^{-p}\left(\frac{\gamma^{-\varrho}}{n}\right)^{\frac{q+1}{q+2}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} + \frac{\tau}{n}\right)
$$

with probability $\mathbf{P}^n$ not less than $1 - 3\mathrm{e}^{-\tau}$. Choosing $p = \log 2/(2\log n)$ and some elementary transformations then prove the assertion. $\qquad\square$

As the proof of Corollary 4.3 merely consists of plugging in the specified values, we will skip it at this point.

Proof of Theorem 4.4. Recall that our optimal choice for γ_n and λ_n was given by $\gamma_n = n^{-a}$ and $\lambda_n = n^{-b}$ with

$$a = \frac{q+1}{\beta(q+2) + \varrho(q+1)} \quad \text{and} \quad b \geq \frac{(d+\beta)(q+1)}{\beta(q+2) + \varrho(q+1)}.$$

Now note that for $\varrho \geq 1$ we have $a \leq 1$ and the choice $b = d$ is admissible. That is, by construction Γ_n and Λ_n cover a possible choice of γ_n and λ_n, which achieve optimal rates. The statement can now be proven exactly as in the proof of Theorem 3.6 by using Theorem 4.2. $\quad\square$

References

[1] J. Bergh and J. Löfström. *Interpolation Spaces. An introduction*, Springer-Verlag, Berlin-New York, 1976.

[2] K. Bezdek. Über einige Kreisüberdeckungen. *Beiträge Algebra Geom.*, pages 7–13, 1983.

[3] P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems*, volume 54 of *IMS Lecture Notes Monogr. Ser.*, pages 177–186. Inst. Math. Statist., Beachwood, OH, 2007.

[4] I. Blaschzyk and I. Steinwart. Improved classification rates under refined margin conditions. *Electron. J. Stat.*, 12:793–823, 2018.

[5] I. Blaschzyk and I. Steinwart. Improved classification rates for localized SVMs. *arXiv*, 1905.01502, 2019.

[6] O. Bousquet. New approaches to statistical learning theory. *Ann. Inst. Statist. Math.*, 55:371–389, 2003.

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

[8] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *STOC'08*, pages 537–546. ACM, New York, 2008.

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

[10] K. Falconer. *Fractal Geometry*, John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2003.

[11] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Phys. D*, 7:153–180, 1983.

[12] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108:203–227, 2019.

[13] G. B. Folland. *Real Analysis*, John Wiley & Sons, Inc., New York, second edition, 1999.

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

[15] L. Hörmander. *The Analysis of Linear Partial Differential Operators. I*, Springer-Verlag, Berlin, 2003.

[16] S. Kpotufe. k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011.

[17] S. Kpotufe and S. Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *J. Comput. System Sci.*, 78:1496–1515, 2012.

[18] S. Kpotufe and V. Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems 26*, pages 3075–3083. Curran Associates, Inc., 2013.

[19] T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *J. Complexity*, 27:489–499, 2011.

[20] S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory*, 41:1028–1039, 1995.

[21] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.

[22] M. J. McGuinness. The fractal dimension of the Lorenz attractor. *Phys. Lett. A*, 99:5–9, 1983.

[23] J. Milnor. On the concept of attractor. *Comm. Math. Phys.*, 99:177–195, 1985.

[24] M. Reed and B. Simon. *Methods of Modern Mathematical Physics II. Fourier Analysis, Self-Adjointness*, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975.

[25] D. Ruelle. Strange attractors. *Math. Intelligencer*, 2:126–137, 1979/80.

[26] I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math. (2)*, 39:811–841, 1938.

[27] C. Scott and R. D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inform. Theory*, 52:1335–1353, 2006.

[28] M. Spivak. *A Comprehensive Introduction to Differential Geometry. Vol. I*, Publish or Perish, Inc., Wilmington, Del., second edition, 1979.

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

[30] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17:211–225, 2011.

[31] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35:575–607, 2007.

[32] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

[33] H. Triebel. *Theory of Function Spaces II*, Birkhäuser Verlag, Basel, 1992.

[34] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37:2655–2675, 2009.

[35] Y. Yang and D. B. Dunson. Bayesian manifold regression. *Ann. Statist.*, 44:876–905, 2016.

[36] G.-B. Ye and D.-X. Zhou. Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.*, 29:291–310, 2008.

[37] G.-B. Ye and D.-X. Zhou. SVM learning and L^p approximation by Gaussians on Riemannian manifolds. *Anal. Appl. (Singap.)*, 7:309–339, 2009.