When do Support Vector Machines learn fast?

Ingo Steinwart* and Clint Scovel
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
ingo@lanl.gov
jcs@lanl.gov

May 11, 2004

Abstract

We establish learning rates to the Bayes risk for support vector machines (SVM's) with hinge loss. Since a theorem of Devroye states that no learning algorithm can learn with a uniform rate to the Bayes risk for *all* probability distributions we have to restrict the class of considered distributions: in order to obtain fast rates we assume a noise condition recently proposed by Tsybakov and an approximation condition in terms of the distribution and the reproducing kernel Hilbert space used by the SVM. For Gaussian RBF kernels with varying widths we propose a *geometric* noise assumption on the distribution which ensures the approximation condition. This geometric assumption is not in terms of smoothness but describes the concentration of the marginal distribution near the decision boundary. In particular we are able to describe nontrivial classes of distributions for which SVM's using a Gaussian kernel can learn with almost linear rate.

AMS classification: primary 68Q32, secondary 62G20, 62G99, 68T05, 68T10, 41A46, 41A99

1 Introduction

In recent years support vector machines (SVM's) have been the subject of many theoretical considerations. In particular, it was recently shown ([10], [16], and [11]) that SVM's can learn for all data-generating distributions. However, these results are purely asymptotic, i.e. no performance guarantees can be given in terms of the number n of samples. In this paper we will establish such guarantees. Since by the no-free-lunch theorem of Devroye (see [4]) performance guarantees are impossible without assumptions on the data-generating distribution we will restrict our considerations to specific classes of distributions. In particular, we will present a geometric condition which describes how distributions behave close to the decision boundary. This condition is then used to establish learning rates for SVM's. To obtain learning rates faster than $n^{-\frac{1}{2}}$ we also employ a noise condition which was recently introduced by Tsybakov (see [13]). Combining both concepts we are in particular able to describe distributions such that SVM's with Gaussian kernel learn almost linearly, i.e. with rate $n^{-1+\varepsilon}$ for all $\varepsilon > 0$, even though the Bayes classifier cannot be represented by the SVM.

Let us now formally introduce the statistical classification problem. To this end assume that X is a set. We write $Y := \{-1, 1\}$. Given a finite *training set* $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$

*Corresponding Author

the classification task is to predict the *label* y of a new sample (x, y). In the standard batch model it is assumed that T is i.i.d. according to an unknown probability measure P on $X \times Y$. Furthermore, the new sample (x, y) is drawn from P independently of T. Given a *classifier* $\mathcal{C}$ that assigns to every training set T a measurable function $f_T : X \to \mathbb{R}$ the prediction of $\mathcal{C}$ for y is $f_T(x)$. In order to "learn" from the samples of T the decision function f_T should guarantee a small probability for the misclassification of the example (x, y). Here, misclassification means $\operatorname{sign} f_T(x) \neq y$ where we choose a fixed definition of $\operatorname{sign}(0) \in \{-1, 1\}$. To make this precise the risk of a measurable function $f : X \to \mathbb{R}$ is defined by

$$\mathcal{R}_P(f) \;:=\; P\big(\{(x, y) : \operatorname{sign} f(x) \neq y\}\big) \;.$$

The smallest achievable risk $\mathcal{R}_P := \inf\big\{\mathcal{R}_P(f) \mid f : X \to \mathbb{R} \text{ measurable}\big\}$ is called the *Bayes risk* of P. A function $f_P : X \to Y$ attaining this risk is called a *Bayes decision function*. Obviously, a good classifier should produce decision functions whose risks are close to the Bayes risk with high probability. This leads to the definition: a classifier is called *universally consistent* if

$$\mathcal{R}_P(f_T) \to \mathcal{R}_P \tag{1}$$

in probability for *all* probability measures P on $X \times Y$. Since $\mathcal{R}(f_T)$ is bounded between $\mathcal{R}_P$ and 1 the convergence in (1) holds if and only if

$$\mathbb{E}_{T \sim P^n} \mathcal{R}_P(f_T) - \mathcal{R}_P \;\to\; 0 \,. \tag{2}$$

The next naturally arising question is whether there are classifiers which guarantee a specific rate of convergence in (1) or (2) for *all* distributions. Unfortunately, this is impossible by the so-called no-free-lunch theorem of Devroye (see [4, Thm. 7.2]). However, if one restricts considerations to certain smaller classes of distributions such rates exist for various classifiers, e.g.:

- Assuming that the conditional probability $\eta(x) := P(1|x)$ satisfies certain smoothness assumptions Yang showed in [15] that some plug-in rules achieve rates for (2) which are of the form $n^{-\alpha}$ for some $0 < \alpha < 1/2$ depending on the assumed smoothness. He also showed that these rates are optimal in the sense that no classifier can obtain faster rates under the proposed smoothness assumptions.

- It is well know (see [4, Thm. 18.3]) that using structural risk minimization over a sequence of hypothesis classes with finite VC-dimension every distribution which has a Bayes decision function in one of the hypothesis classes can be learned with rate $\sqrt{\frac{\log n}{n}}$.

- Let P be a noisefree distribution, i.e. $\mathcal{R}_P = 0$ and $\mathcal{F}$ be a class with finite VC-dimension. If $\mathcal{F}$ contains a Bayes decision function then the rate of convergence of the ERM classifier over $\mathcal{F}$ is n^{-1}.

Restricting the class of distributions for classification always raises the question of whether it is likely that these restrictions are met in real world problems. Of course, in general this question cannot be answered. However, experience shows that the assumption that the distribution is noise-free is almost never satisfied in practice. Furthermore, it is also rather unrealistic to assume that a Bayes decision function can be represented by the algorithm. Finally, assuming that the conditional probability is smooth, say k-times continuously differentiable, seems to be unlikely for many real world classification problems. This discussion shows that the above listed rates are established for situations which are rarely met in practice.

Considering the ERM classifier and hypothesis classes $\mathcal{F}$ containing a Bayes decision function there is a large gap in the rates for noise-free and noisy distributions. In [13] Tsybakov proposed a condition on the noise which describes intermediate situations. In order to present this condition we write $\eta(x) := P(1|x)$, $x \in X$ for the conditional probability and P_X for the marginal distribution of P on X. With the help of η the noise can be described by the function $|2\eta - 1|$. Indeed, in regions where this function is close to 1 there is only a small amount of noise, whereas function values close to 0 only occur in regions with a high noise. We will use the following modified version of Tsybakov's noise condition which describes the size of the latter regions:

Definition 1.1 Let $0 \leq q \leq \infty$ and P be a probability measure on $X \times Y$. We say that P *has Tsybakov noise exponent* q if there exists a constant $C > 0$ such that

$$P_X\big(\,|2\eta - 1| \leq t\big) \;\leq\; C \cdot t^q \tag{3}$$

for all $t > 0$.

All distributions have at least noise exponent 0. In the other extreme case $q = \infty$ the conditional probability η is bounded away from $\frac{1}{2}$. In particular this means that noise-free distributions have exponent $q = \infty$. Furthermore, it suffices to assume that (3) holds for all $0 < t < t_0$, where t_0 can be arbitrarily small. This shows that the Tsybakov noise exponent only measures the size of regions with high noise. Finally note, that Tsybakov's original noise condition assumed $P_X(f \neq f_P) \leq c(\mathcal{R}_P(f) - \mathcal{R}_P)^{\frac{q}{1+q}}$ for all $f : X \to Y$ which is satisfied if e.g. (3) holds (see [13, Prop. 1]).

In [13] Tsybakov showed that if P has a noise exponent q then ERM-type classifiers can obtain rates in (2) which are of the form $n^{-\frac{q+1}{q+pq+2}}$, where $0 < p < 1$ measures the complexity of the hypothesis class. In particular, rates faster than $n^{-\frac{1}{2}}$ are possible whenever $q > 0$ and $p < 1$. Unfortunately, the ERM-classifier he considered is usually hard to implement and in general there exists no efficient algorithm. Furthermore, his classifier requires substantial knowledge on *how* to approximate the Bayes decision rules of the considered distributions. Of course, such knowledge is rarely present in practice.

In this paper we will use the Tsybakov noise exponent to establish rates for SVM's which are very similar to the above rates of Tsybakov. We begin by recalling the definition of SVM's. To this end let H be a reproducing kernel Hilbert space (RKHS) of a kernel $k : X \times X \to \mathbb{R}$ (cf. [1], [3]), i.e. H is a Hilbert space consisting of functions from X to $\mathbb{R}$ such that the evaluation functionals are continuous, and k is symmetric and positive definite. Throughout this paper we assume that X is a compact metric space and that k is continuous, i.e. H contains only continuous functions. In order to avoid cumbersome notations we additionally assume $\|k\|_\infty \leq 1$. Now given a regularization parameter $\lambda > 0$ the decision function of an SVM is

$$(f_{T,\lambda}, b_{T,\lambda}) \;:=\; \arg\min_{\substack{f \in H \\ b \in \mathbb{R}}} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n l\big(y_i(f(x_i) + b)\big)\,, \tag{4}$$

where $l(t) := \max\{0, 1 - t\}$ is the so-called hinge loss. Instead of solving (4) directly one usually solves the quadratic dual problem

$$\begin{array}{lll} \text{maximize} & \sum\limits_{i=1}^n \alpha_i - \frac{1}{4\lambda} \sum\limits_{i,j=1}^n y_i y_j \alpha_i \alpha_j k(x_i, x_j) & \text{for } \alpha \in \mathbb{R}^n \\ \text{subject to} & \sum\limits_{i=1}^n y_i \alpha_i = 0, & \\ & 0 \leq \alpha_i \leq \frac{1}{n}, & i = 1, \ldots, n\,. \end{array} \tag{5}$$

instead. Then if $(\alpha_1^*, \ldots, \alpha_n^*) \in \mathbb{R}$ denotes a solution of (5) we have $f_{T,\lambda} = \frac{1}{2\lambda} \sum_{i=1}^n y_i \alpha_i^* k(x_i, .)$ and $b_{T,\lambda}$ can be computed using an α_i^* with $0 < \alpha_i^* < \frac{1}{n}$ (see [3] for details). Note, that solving (5) has the remarkable property that H only occurs implicitely via its kernel.

Only a few results on learning rates for SVM's are known: In [9] it was shown that SVM's can learn with linear rate if the distribution is noise-free and the two classes can be strictly separated by the RKHS. For RKHS which are dense in the space $C(X)$ of continuous functions the latter condition is satisfied if the two classes have strictly positive ditance in the input space. Of course, these assumptions are too strong for almost all real-world problems. Furthermore, Wu and Zhou (see [14]) recently established rates for (1) under the assumption that η is contained in a Sobolev space. In particular, he proved rates of the form $(\log n)^{-p}$ for some $p > 0$ if the SVM uses a Gaussian kernel. Of course, these rates are much too slow to be of practical interest and the problems with smoothness assumptions have already been discussed above.

In order to state our first result which is much stronger than the above mentioned results we need two concepts both of which deal with the involved RKHS. The first concept describes how well a given RKHS H can approximate a distribution P. In order to introduce it we define the *l-risk* of a function $f : X \to \mathbb{R}$ by $\mathcal{R}_{l,P}(f) := \mathbb{E}_{(x,y)\sim P} l(yf(x))$. The smallest posssible l-risk is denoted by $\mathcal{R}_{l,P} := \inf\{\mathcal{R}_{l,P}(f) \mid f : X \to \mathbb{R}\}$. Furthermore, we define the *approximation error function* by

$$a(\lambda) \;:=\; \inf_{f \in H} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{l,P}(f) \Big) - \mathcal{R}_{l,P}\,, \qquad\qquad \lambda \geq 0\,. \tag{6}$$

The approximation error function quantifies how well an infinite sample SVM with RKHS H approximates the minimal l-risk. It was shown in [11] that if H is dense in the space of continuous functions $C(X)$ then for *all* P we have $a(\lambda) \to 0$ if $\lambda \to 0$. However, in non-trivial situations no rate of convergence which uniformly holds for all distributions P is possible. The following definition characterizes distributions which guarantee certain polynomial rates:

Definition 1.2 Let H be a RKHS over X and P be a probability measure on $X \times Y$. We say that *H approximates P with exponent* $0 < \beta \leq 1$ if there exists a constant $C > 0$ such that

$$a(\lambda) \leq C\lambda^\beta$$

for all $\lambda > 0$.

It can be shown (see [7]) that the extremal case $\beta = 1$ is equivalent to the fact that the minimal l-risk can be achieved by an element of H. Because of the specific structure of the approximation error function values $\beta > 1$ are only possible for distributions with $\eta \equiv \frac{1}{2}$. The latter are uninteresting for classification considerations.

Finally, we need a complexity measure for RKHS's. To this end we have to recall some notations. For a subset $A \subset E$ of a Banach space E the *covering numbers* are defined by

$$\mathcal{N}(A, \varepsilon, E) \;:=\; \min\Big\{ n \geq 1 : \exists x_1, \ldots, x_n \in E \text{ with } A \subset \bigcup_{i=1}^n (x_i + \varepsilon B_E) \Big\} \qquad\qquad \varepsilon > 0\,,$$

where B_E denotes the closed unit ball of E. Furthermore, given a training set $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ we denote the space of all equivalence classes of functions $f : X \to \mathbb{R}$ with norm

$$\|f\|_{L_2(T)} := \left(\frac{1}{n} \sum_{i=1}^n \big|f(x_i)\big|^2 \right)^{\frac{1}{2}} \tag{7}$$

by $L_2(T)$. In other words, $L_2(T)$ is a L_2-space with respect to the empirical measure of $(x_1, \ldots, x_n)$. Note, that for a function $f : X \to \mathbb{R}$ a canonical representative in $L_2(T)$ is the restriction $f_{|\{x_1,\ldots,x_n\}}$. Now our complexity measure is:

Definition 1.3 Let H be a RKHS over X and B_H its closed unit ball. We say that H *has complexity exponent* $0 < p \leq 2$ if there exists a constant $a_p > 0$ such that

$$\sup_{T \in (X \times Y)^n} \log \mathcal{N}(B_H, \varepsilon, L_2(T)) \;\leq\; a_p \varepsilon^{-p}$$

for all $\varepsilon > 0$.

It was shown in [7] that every RKHS has complexity exponent $p = 2$ by using the theory of absolutely 2-summing operators. However, for interesting rates we need complexity exponents which are strictly smaller than 2. For many RKHS such results are known (see e.g. [11] and [7]) Furthermore, many SVM's use a parameterized family of RKHS's. For such SVM's the constant a_p may play a crucial role. We will see below, that this is in particular true for SVM's using a Gaussian RBF kernel.

Now we are in the position to formulate our first rate which applies to SVM's using general kernels:

Theorem 1.4 *Let H be a RKHS of a continuous kernel on X with complexity exponent $0 < p < 2$, and let P be a probability measure on $X \times Y$ with Tsybakov noise exponent $0 < q \leq \infty$. Furthermore, assume that H approximates P with exponent $0 < \beta \leq 1$. We define $\lambda_n := n^{-\frac{4(q+1)}{(2q+pq+4)(1+\beta)}}$. Then for all $\varepsilon > 0$ there is a constant $C > 0$ such that for all $x \geq 1$ and all $n \geq 1$ we have*

$$\Pr^*\Big(T \in (X \times Y)^n : \mathcal{R}_P(f_{T,\lambda_n} + b_{T,\lambda_n}) \leq \mathcal{R}_P + Cx^2 n^{-\frac{4\beta(q+1)}{(2q+pq+4)(1+\beta)}+\varepsilon}\Big) \;\geq\; 1 - e^{-x}\,.$$

Remark 1.5 Using a tail bound of the form of Theorem 1.4 one can easily get convergence rates for (2). In the case of the above theorem these rates have the form $n^{-\frac{4\beta(q+1)}{(2q+pq+4)(1+\beta)}+\varepsilon}$ for all $\varepsilon > 0$. In other words the rates are exactly the terms in n in the above tail bounds. This is also true for the rates of SVM's using Gaussian RBF kernels which are established below.

Remark 1.6 For brevity's sake our major aim was to show the best possible rates using our techniques. Therefore, the above theorem states rates for the SVM under the assumption that (λ_n) is optimally chosen. However, we emphasize, that the techniques of [7] also give rates if (λ_n) is chosen in a different (and thus sub-optimal) way. This is also true for our results on SVM's using Gaussian kernels.

Remark 1.7 In [13] it is assumed that a Bayes classifier is contained in the base function classes the algorithm minimizes over. This assumption corresponds to a perfect approximation of P by H, i.e. $\beta = 1$. In this case our rate is essentially of the form $n^{-\frac{2(q+1)}{2q+pq+4}}$. If we rescale the complexity exponent p from $(0, 2)$ to $(0, 1)$ and write p' for the new complexity exponent this rate becomes essentially $n^{-\frac{q+1}{q+p'q+2}}$. This is exactly the *form* of Tsybakov's result in [13]. However, as far as we know our complexity measure cannot be compared to Tsybakov's.

Remark 1.8 By the nature of Theorem 1.4 it suffices to assume that P only satisfies Tsybakov's noise assumption for every $q' < q$. It also suffices to suppose that H approximates P with exponent β' for all $\beta' < \beta$, and that H has complexity exponent p' for all $p' > p$. It is shown in [7] that the

RKHS H has an approximation exponent $\beta = 1$ if and only if H contains a minimizer of the l-risk. In particular, if H has approximation exponent β for all $\beta < 1$ but not for $\beta = 1$ then H does not contain such a minimizer but Theorem 1.4 gives the same result as for $\beta = 1$.

Furthermore, if the RKHS consists of C^∞ functions we can choose p arbitrarily close to 0. If both assumptions are true, we can hence obtain rates up to n^{-1} even though H does not contain a minimizer of the l-risk.

In view of Theorem 1.4 and the remarks concerning covering numbers it is often only necessary to estimate the approximation exponent. In particular this seems to be true for the most popular kernel, that is the Gaussian RBF kernel $k_\sigma(x, x') = \exp(-\sigma^2\|x - x'\|_2^2)$, $x, x' \in X$ on (compact) subsets X of $\mathbb{R}^d$ with width σ. However, to our best knowledge no non-trivial condition on η or $f_P = \operatorname{sign} \circ (2\eta - 1)$ which ensures an approximation exponent $\beta > 0$ for *fixed* width has been established and [8] shows that Gaussian kernels poorly approximate smooth functions. Hence plug-in rules based on Gaussian kernels may have a bad performance under smoothness assumptions on η. In particular, many types of SVM's using other loss functions are plug-in rules and therefore, their approximation properties under smoothness assumptions on η may be poor if a Gaussian kernel is used. However, our SVM's are not plug-in rules since their decision functions approximate the Bayes decision function (see [12]). Intuitively, we therefore only need a condition that measures the cost of approximating the "bump" of the Bayes decision function at the "decision boundary". We will now establish such a condition for Gaussian RBF kernels with *varying* widths σ_n. To this end let $X_{-1} := \{x \in X : \eta < \frac{1}{2}\}$ and $X_1 := \{x \in X : \eta > \frac{1}{2}\}$. Recall that these two sets are the classes which have to be learned. Since we are only interested in distributions P having a Tsybakov exponent $q > 0$ we always assume that $X = X_{-1} \cup X_1$ holds P_X-almost surely. Now we define

$$\tau_x := \begin{cases} d(x, X_1), & \text{if } x \in X_{-1}, \\ d(x, X_{-1}), & \text{if } x \in X_1, \\ 0, & \text{otherwise}. \end{cases} \tag{8}$$

Here, $d(x, A)$ denotes the distance of x to a set A with respect to the Euclidian norm. Roughly speaking τ_x measures the distance of x to the "decision boundary". With the help of this function we can define the following geometric condition for distributions:

Definition 1.9 Let $X \subset \mathbb{R}^d$ be compact and P be a probability measure on $X \times Y$. We say that P has *geometric noise exponent* $\alpha > 0$ if there exists a constant $C > 0$ such that

$$\int_X |2\eta(x) - 1| \exp\Big(-\frac{\tau_x^2}{t}\Big) P_X(dx) \;\le\; C t^{\frac{\alpha d}{2}} \tag{9}$$

holds for all $t > 0$. We say that P has geometric noise exponent $\alpha = \infty$ if it has geometric noise exponent α' for all $\alpha' > 0$.

Note, that in the above definition we make neither any kind of smoothness assumption nor do we assume a condition on P_X in terms of absolute continuity with respect to the Lebesgue measure. Instead, the integral condition (9) describes the concentration of the measure $|2\eta - 1|dP_X$ near the decision boundary. The less the measure is concentrated in this region the larger the geometric noise exponent can be chosen. The following examples illustrate this:

Example 1.10 Since $\exp(-t) \le C_\alpha t^{-\alpha}$ holds for all $t > 0$ and a constant $C_\alpha > 0$ only depending on $\alpha > 0$ we easily see that (9) is satisfied whenever

$$\big(x \mapsto \tau_x^{-1}\big) \in L_{\alpha d}\big(|2\eta - 1|dP_X\big)\,. \tag{10}$$

Now, recall that τ_x measures the distance to the class x does not belong to. In particular, we have $\bigl(x \mapsto \tau_x^{-1}\bigr) \in L_\infty\bigl(|2\eta - 1| dP_X\bigr)$ if and only if the two classes X_{-1} and X_1 have strictly positive distance! If (10) holds for some $0 < \alpha < \infty$ then the two classes may "touch", i.e. the decision boundary $\partial X_{-1} \cap \partial X_1$ is nonempty. Using this interpretation we easily can construct distributions which have geometric noise exponent ∞ and touching classes! In general for these distributions there is no Bayes classifier in the RKHS H_σ of k_σ for any $\sigma > 0$.

Note, that from (10) it is obvious that the parameter α in (10) describes the concentration of the measure $|2\eta - 1| dP_X$ near the decision boundary. For the distributions described above $|2\eta - 1| dP_X$ must have a very low concentration near the decision boundary.

Example 1.11 We say that η *is Hölder about* $\frac{1}{2}$ with exponent $\gamma > 0$ on $X \subset \mathbb{R}^d$ if there is a constant c_γ such that

$$|2\eta(x) - 1| \leq c_\gamma \tau_x^\gamma, \ \forall x \in X. \tag{11}$$

If η is Hölder about $\frac{1}{2}$ with exponent $\gamma > 0$, the graph of $2\eta(x) - 1$ lies in a multiple of the envelope defined by τ_x^γ at the top and by $-\tau_x^\gamma$ at the bottom. To be Hölder about $\frac{1}{2}$ it is sufficient that η is Hölder continuous, but it is far from being necessary. A function which is Hölder about $\frac{1}{2}$ can be very irregular away from the decision boundary but it cannot jump across the decision boundary discontinuously. In addition a Hölder continuous function's exponent must satisfy $0 < \gamma \leq 1$ where being Hölder about $\frac{1}{2}$ only requires $\gamma > 0$.

For distributions with Tsybakov noise exponent such that η is Hölder about $\frac{1}{2}$ we can bound the geometric noise exponent. Indeed, let P be a probability measure on $X \times Y$ which has Tsybakov noise exponent $q \geq 0$ and a conditional probability η which is Hölder about $\frac{1}{2}$ with exponent $\gamma \geq 0$. Then (see [7]) if $q \geq 1$, P has geometric noise exponent $\alpha = \gamma \frac{q+1}{d}$ and if $0 \leq q < 1$, P has geometric noise exponent α for all $\alpha < \gamma \frac{q+1}{d}$.

For distributions having a non-trivial geometric noise exponent we can bound the approximation error function for Gaussian RBF kernels:

Theorem 1.12 *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and H_σ be the RKHS of the Gaussian RBF kernel k_σ on X with width $\sigma > 0$. We write $a_\sigma(.)$ for the approximation error function with respect to H_σ. Then there is a constant c_d depending only on d such that if P has geometric noise exponent $0 < \alpha < \infty$ with constant C, for all $\lambda > 0$ and all $\sigma > 0$ we have*

$$a_\sigma(\lambda) \ \leq \ c_d \Bigl(\sigma^d \lambda + C (4d)^{\frac{\alpha d}{2}} \sigma^{-\alpha d} \Bigr) . \tag{12}$$

In order to let the right hand side of (12) converge to zero it is necessary to assume both $\lambda \to 0$ and $\sigma \to \infty$. An easy consideration shows that the fastest rate of convergence can be achieved if $\sigma(\lambda) := \lambda^{-\frac{1}{(\alpha+1)d}}$. In this case we have $a_{\sigma(\lambda)}(\lambda) \leq 2C\lambda^{\frac{\alpha}{\alpha+1}}$. Roughly speaking this states that the family of spaces $H_{\sigma(\lambda)}$ approximates P with exponent $\frac{\alpha}{\alpha+1}$. Note, that we can obtain approximation rates up to linear order in λ for sufficiently benign distributions. The price for this good approximation property is, however, an increasing complexity of the hypothesis class $B_{H_{\sigma(\lambda)}}$ for $\sigma \to \infty$, i.e. $\lambda \to 0$. The following theorem estimates this in terms of the complexity exponent:

Theorem 1.13 *Let H_σ be the RKHS of the Gaussian RBF kernel k_σ on X. Then for all $0 < p \leq 2$ and $\delta > 0$, there exists a constant $c_{p,d,\delta} > 0$ such that for all $\varepsilon > 0$ and all $\sigma \geq 1$ we have*

$$\sup_{T \in Z^n} \log \mathcal{N}(B_{H_\sigma}, \varepsilon, L_2(T)) \ \leq \ c_{p,d,\delta} \, \sigma^{(1-\frac{p}{2})(1+\delta)d} \varepsilon^{-p}.$$

Having established both results for the approximation and complexity exponent we can now formulate our main result for SVM's using Gaussian RBF kernels:

Theorem 1.14 *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 < q \leq \infty$ and geometric noise exponent $0 < \alpha < \infty$. We define*

$$\lambda_n \;:=\; \begin{cases} n^{-\frac{\alpha+1}{2\alpha+1}} & \text{if } \alpha \leq \frac{q+2}{2q} \\ n^{-\frac{2(\alpha+1)(q+1)}{2\alpha(q+2)+3q+4}} & \text{otherwise}\,, \end{cases}$$

and $\sigma_n := \lambda_n^{-\frac{1}{(\alpha+1)d}}$ in both cases. Then for all $\varepsilon > 0$ there exists a constant $C > 0$ such that for all $x \geq 1$ and all $n \geq 1$ the SVM using λ_n and Gaussian RBF kernel with width σ_n satisfies

$$\Pr^*\Bigl(T \in (X \times Y)^n : \mathcal{R}_P(f_{T,\lambda_n}) \leq \mathcal{R}_P + Cx^2 n^{-\frac{\alpha}{2\alpha+1}+\varepsilon}\Bigr) \;\geq\; 1 - e^{-x}$$

if $\alpha \leq \frac{q+2}{2q}$ and

$$\Pr^*\Bigl(T \in (X \times Y)^n : \mathcal{R}_P(f_{T,\lambda_n}) \leq \mathcal{R}_P + Cx^2 n^{-\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}+\varepsilon}\Bigr) \;\geq\; 1 - e^{-x}$$

otherwise. If $\alpha = \infty$ the latter concentration inequality holds if $\sigma_n = \sigma$ is a constant with $\sigma > 2\sqrt{d}$.

Most of the remarks made after Theorem 1.4 also apply to the above theorem up to obvious modifications. In particular this is true for Remark 1.5, Remark 1.6, and Remark 1.8.

2 Idea of the proofs

In this section we provide a brief overview of the ideas which are used to prove our rates. The full proofs can be found in [7]. The basic idea of the proofs consists of an inequality between excess classification and excess l-risk, and a modification of the classical decomposition into estimation and approximation error. More precisely we have

$$\begin{aligned} \mathcal{R}_P(f_{T,\lambda} + b_{T,\lambda}) - \mathcal{R}_P \;&\leq\; 2\bigl(\mathcal{R}_{l,P}(f_{T,\lambda} + b_{T,\lambda}) - \mathcal{R}_{l,P}\bigr) \\ &=\; 2\bigl(\mathcal{R}_{l,P}(f_{T,\lambda} + b_{T,\lambda}) - \mathcal{R}_{l,P}(f_{P,\lambda} + b_{P,\lambda}) + \mathcal{R}_{l,P}(f_{P,\lambda} + b_{P,\lambda}) - \mathcal{R}_{l,P}\bigr) \\ &\leq\; 2\bigl(\lambda\|f_{T,\lambda}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda} + b_{T,\lambda}) - \lambda\|f_{P,\lambda}\|^2 - \mathcal{R}_{l,P}(f_{P,\lambda} + b_{P,\lambda}) + a(\lambda)\bigr)\,, \end{aligned}$$

where the first inequality is due to Zhang (see [16] and also [2]) and $(f_{P,\lambda}, b_{P,\lambda})$ minimizes the infinite-sample SVM problem, i.e. $(f_{P,\lambda}, b_{P,\lambda}) = \arg\min_{f \in H, b \in \mathbb{R}} \lambda\|f\|^2 + \mathcal{R}_{l,P}(f + b)$. Therefore, we can split the anaylsis into an estimation error part which deals with

$$\lambda\|f_{T,\lambda}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda} + b_{T,\lambda}) - \lambda\|f_{P,\lambda}\|^2 - \mathcal{R}_{l,P}(f_{P,\lambda} + b_{P,\lambda}) \tag{13}$$

and an approximation error part which deals with the approximation error function $\lambda \mapsto a(\lambda)$.

Let us first treat the estimation error part. To this end we write $L \circ (f,b)(x,y) := \lambda\|f\|^2 + l(y(f(x) + b))$ and $\mathcal{R}_{L,P}(f,b) := \mathbb{E}_P L \circ (f,b) = \lambda\|f\|^2 + \mathcal{R}_{l,P}(f + b)$. Furthermore, for $0 < \lambda \leq 1$ we define

$$\mathcal{G} \;:=\; \bigl\{L \circ (f,b) - L \circ (f_{P,\lambda}, b_{P,\lambda}) \;:\; (f,b) \in \gamma B_H \times (\gamma+1)B_{\mathbb{R}}\bigr\}\,,$$

where $1 \leq \gamma \leq \lambda^{-\frac{1}{2}}$ is a constant we discuss later and B_H, $B_{\mathbb{R}}$ denote the closed unit balls in H and $\mathbb{R}$. Furthermore, for $n \geq 1$ and $\varepsilon > 0$ we define the local Rademacher average of $\mathcal{G}$ by

$$\mathrm{Rad}_P(\mathcal{G}, n, \varepsilon) \;:=\; \mathbb{E}_{P^n}\mathbb{E}_\mu \sup_{\substack{g\in\mathcal{G},\\ \mathbb{E}_P g^2 \leq \varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(z_i) \right| ,$$

where (ε_i) is a sequence of i.i.d. Rademacher variables (that is, symmetric $\{-1, 1\}$-valued random variables) with respect to some probability measure μ on a set Ω. With this notations the following theorem from [7] which is a consequence of Talagrand's concentration inequality holds:

Theorem 2.1 *Let P be a probability measure on $X \times Y$ and $0 < \lambda \leq 1$. Suppose that there are constants $c \geq 0$, $0 < \alpha \leq 1$ and $\delta \geq 0$ with $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ for all $g \in \mathcal{G}$. Let $B := 2\gamma + 3$, $n \geq 1$, $x > 0$ and $\varepsilon > 0$ with*

$$\varepsilon \geq 10 \max\left\{ 2\mathrm{Rad}_P(\mathcal{G}, n, c\varepsilon^\alpha + \delta), \sqrt{\frac{\delta x}{n}}, \left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}}, \frac{Bx}{n} \right\}.$$

Then we have

$$\Pr^*\Big(T \in Z^n : \mathcal{R}_{L,P}(f_{T,\lambda}, b_{T,\lambda}) < \mathcal{R}_{L,P}(f_{P,\lambda}, b_{P,\lambda}) + \varepsilon\Big) \;\geq\; 1 - e^{-x}\,.$$

Theorem 2.1 has been proved in [2] for $\delta = 0$. In this case its main advantage compared to the "standard analysis" using uniform deviation bounds is that it can produce rates faster than $n^{-\frac{1}{2}}$ for risk deviations. For a further discussion of this issue we refer to [2]. If $\delta > 0$ the above theorem *apparently cannot* produce rates faster than $n^{-\frac{1}{2}}$. However, in order to decrease the approximation error function we have to choose $\lambda_n \to 0$ and thus the class $\mathcal{G} = \mathcal{G}_n$ increases with n. If for such sequences $(\mathcal{G}_n)$ we can show that $\delta_n \to 0$ then the term $\sqrt{\frac{\delta x}{n}}$ no longer prohibits rates faster than $n^{-\frac{1}{2}}$. As we will see below this phenomenon actually occurs for distributions satisfying Tsybakov's noise assumption for some exponent $q > 0$.

The above theorem bounds the estimation error part (13) by ε. Hence, in order to obtain rates we have to estimate ε, i.e. we have to provide a bound on the local Rademacher average, constants for the so-called *variance bound* $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$, and finally, a bound for B. Let us begin with the local Rademacher average which can be treated by the following proposition

Proposition 2.2 *Let H be a RKHS on X with complexity exponent $0 < p < 2$ and corresponding constant a. Then there exists a constant $c_p > 0$ such that for all $n \geq 1$ and all $\varepsilon > 0$ we have*

$$\mathrm{Rad}(\mathcal{G}, n, \varepsilon) \;\leq\; c_p \max\left\{ B^{\frac{p}{2}} \varepsilon^{\frac{1}{2}-\frac{p}{4}} \left(\frac{a}{n}\right)^{\frac{1}{2}}, B\left(\frac{a}{n}\right)^{\frac{2}{2+p}} \right\}.$$

Besides some technical details the proof of this proposition heavily relies on techniques of Mendelson (see [6]). Again we refer to [7]. Using Proposition 2.2 one can give upper bounds on the error ε in Theorem 2.1 in terms of the arising constants. More precisely, we obtain that besides a constant only depending on p and H we can choose

$$\varepsilon \;:=\; B^{\frac{2p}{4-2\alpha+\alpha p}} c^{\frac{2-p}{4-2\alpha+\alpha p}} \left(\frac{a}{n}\right)^{\frac{2}{4-2\alpha+\alpha p}} + B^{\frac{p}{2}} \delta^{\frac{2-p}{4}} \left(\frac{a}{n}\right)^{\frac{1}{2}} + B\left(\frac{a}{n}\right)^{\frac{2}{2+p}} \sqrt{\frac{\delta x}{n}} + \left(\frac{cx}{n}\right)^{\frac{1}{2-\alpha}} + \frac{Bx}{n}\,.$$

Let us now deal with the variance bound. It turns out that Tsybakov's noise exponent influences the corresponding constants c, α and δ in Theorem 2.1. Surprisingly, however, the approximation error function also influences δ. More precisely, we have the following result:

Proposition 2.3 *Let P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 < q \leq \infty$. Then there exists a constant $C > 0$ such that for all $0 < \lambda \leq 1$, $0 < \gamma \leq \lambda^{-1/2}$ and all $g \in \mathcal{G}$ we have*

$$\mathbb{E}g^2 \;\leq\; CB^{\frac{q+2}{q+1}}\big(\mathbb{E}g\big)^{\frac{q}{q+1}} + CB^{\frac{q+2}{q+1}}a^{\frac{q}{q+1}}(\lambda)\,,$$

in other words we have $\alpha = \frac{q}{q+1}$, $c = CB^{\frac{q+2}{q+1}}$, and $\delta = CB^{\frac{q+2}{q+1}}a^{\frac{q}{q+1}}(\lambda)$.

In view of Theorem 2.1 it hence remains to bound B, i.e. the constant γ. Since we always have

$$\lambda\|f_{T,\lambda}\|^2 \;\leq\; \lambda\|f_{T,\lambda}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda} + b_{T,\lambda}) \;\leq\; \mathcal{R}_{l,P}(0) \;=\; 1\,,$$

an obvious choice would be $\gamma = \lambda^{-\frac{1}{2}}$. However, considering the infinite-sample solution we observe

$$\lambda\|f_{P,\lambda}\|^2 \;\leq\; \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda} + b_{P,\lambda}) \;\leq\; a(\lambda)$$

which is a significantly better behaviour if (H, P) has some approximation exponent $\beta > 0$. Therefore, it is highly desirable to approximately establish this relation for the empirical solutions, too. Fortunately, it is indeed possible to prove this relation in a certain sense. To this end we first apply Theorem 2.1 for $\gamma = \lambda^{-\frac{1}{2}}$. This gives us

$$\lambda\|f_{T,\lambda}\|^2 \;\leq\; \lambda\|f_{T,\lambda}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda} + b_{T,\lambda}) \;\leq\; \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda} + b_{P,\lambda}) + \varepsilon \;\leq\; a(\lambda) + \varepsilon$$

with probability not less than $1 - e^{-x}$. Now for sequences (λ_n) say as in Theorem 1.4 we observe that $\varepsilon = \varepsilon_n$ dominates $a(\lambda_n)$ and therefore we cannot immediately establish the desired estimate. Nonetheless, we have $\varepsilon_n \to 0$ polynomially and the corresponding exponent can be directly controlled. The basic idea is now to use the estimate on $\gamma_n = \gamma$ which is implied by $\lambda\|f_{T,\lambda_n}\|^2 \leq 2\varepsilon_n$ for large n in order to again apply Theorem 2.1. In turns out that the arising sequence $(\hat{\varepsilon}_n)$ converges faster than (ε_n) and hence our estimate on γ_n is improved for large n. Iterating this procedure then finally gives the desired estimate. For details we again refer to [7].

Now let us turn to the approximation error part. To prove Theorem 1.12 we bound the approximation error function (6) by a judicious choice of function $\hat{f} \in H$ in the inequality

$$a(\lambda) \leq \lambda\|f\|_H^2 + \mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P}, \quad f \in H. \tag{14}$$

To that end, let $\eta(x) = P(y = 1|x)$ be any regular conditional distribution for P and let f_P be any Bayes function with values in $[-1, 1]$ such that $f_P = 1$ on X_1 and $f_P = -1$ on X_{-1}. We will choose a function $\hat{f}$ by smoothing an extension $\acute{f}_P$ of f_P to $\acute{X} := 3X$. To do so first consider the extension of η to be constant in the outward radial direction $\acute{\eta}(x) = \eta(x),\ |x| \leq 1$ and $\acute{\eta}(x) = \eta(\frac{x}{|x|}),\ |x| > 1$ and define $\acute{X}_{-1} := \{x \in \acute{X} : \acute{\eta}(x) < \frac{1}{2}\}$, $\acute{X}_1 := \{x \in \acute{X} : \acute{\eta}(x) > \frac{1}{2}\}$. It is easy to show that when $x \in X_1$, we have $B(x, \tau_x) \subset \acute{X}_1$ and when $x \in X_{-1}$, we have $B(x, \tau_x) \subset \acute{X}_{-1}$ where $B(x, \tau)$ denotes the open ball of radius τ about x. Let $\acute{f}_P$ be a measurable function with values in $[-1, 1]$ which coincides with f_P on X such that $\acute{f}_P = 1$ on $\acute{X}_1$, $\acute{f}_P = -1$ on $\acute{X}_{-1}$, and $\acute{f}_P(x) = 0$, $|x| > 3$. Let K_σ denote the integral operator associated with the Gaussian RBF kernel k_σ. Also consider the normalized Gaussian kernel $\hat{k}_\sigma = \sigma^d\pi^{-\frac{d}{2}}k_\sigma$ and the corresponding *Gauss-Weierstrass integral operator* $\hat{K}_\sigma$. We consider the function $\hat{f} = \hat{K}_\sigma\acute{f}_P$. We recall that the RKHS associated with k_σ on $\acute{X}$ is $H_\sigma(\acute{X}) = K_\sigma^{\frac{1}{2}}L_2(\acute{X})$ and the norm is defined by $\|K_\sigma^{\frac{1}{2}}g\|_{H_\sigma(\acute{X})} = \|g\|_{L_2(\acute{X})}$. Therefore

$$\|\hat{f}\|_{H_\sigma(\acute{X})} = \|\hat{K}_\sigma\acute{f}_P\|_{H_\sigma(\acute{X})} = \sigma^{\frac{d}{2}}\pi^{-\frac{d}{4}}\|K_\sigma^{\frac{1}{2}}\hat{K}_\sigma^{\frac{1}{2}}g\|_{H_\sigma(\acute{X})} = \sigma^{\frac{d}{2}}\pi^{-\frac{d}{4}}\|\hat{K}_\sigma^{\frac{1}{2}}g\|_{L_2(\acute{X})} \leq \sigma^{\frac{d}{2}}\pi^{-\frac{d}{4}}\|\hat{K}_\sigma^{\frac{1}{2}}\|\|g\|_{L_2(\acute{X})}\,.$$

The continuous functional calculus theorem for self adjoint operators implies that $\|\hat{K}_\sigma^{\frac{1}{2}}\| = \|\hat{K}_\sigma\|^{\frac{1}{2}}$. Therefore to finish the proof we only need to show that $\hat{K}_\sigma$ is a contraction on $L_2(\acute{X})$. The latter follows from Young's inequality since the Gauss-Weierstrass integral operator $\hat{K}_\sigma$ is a convolution and $\int \sigma^d \pi^{-\frac{d}{2}} e^{-\sigma^2|u|^2} du = 1$. We therefore obtain $\|\hat{f}\|_{H_\sigma(\acute{X})} \leq \sigma^{\frac{d}{2}} (\frac{81}{\pi})^{\frac{d}{4}} \theta(d)$ where $\theta(d) = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$ is the volume of X. According to Aronszajn [1] we also have $\|f\|_{H_\sigma(X)} \leq \|f\|_{H_\sigma(\acute{X})}$ and so we have bounded the first term in the inequality (14) with the choice $\hat{f}$ and proved the first term in the inequality of Theorem 1.12.

We now proceed to bound the term $\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P}$ in (14). For any function which satisfies $-1 \leq f \leq 1$, Zhang [16] shows that

$$\mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P} = \mathbb{E}_{P_X}(|2\eta - 1||f - f_P|).$$

It is well known for the Gauss-Weierstrass heat operator $\hat{K}_\sigma$ that since $-1 \leq \acute{f}_P \leq 1$ it follows that $-1 \leq \hat{f} = \hat{K}_\sigma \acute{f}_P \leq 1$ and so we obtain

$$\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P} = \mathcal{R}_{l,P}(\hat{K}_\sigma \acute{f}_P) - \mathcal{R}_{l,P} = \mathbb{E}_{P_X}(|2\eta - 1||\hat{K}_\sigma \acute{f}_P - f_P|).$$

Now for $x \in X$ we have

$$\begin{aligned}
\hat{f}(x) = \int_{\acute{X}} \hat{k}_\sigma(x, \acute{x}) \acute{f}_P(\acute{x}) d\acute{x} &= \int_{\mathbb{R}^d} \hat{k}_\sigma(x, \acute{x}) \acute{f}_P(\acute{x}) d\acute{x} \\
&= \int_{\mathbb{R}^d} \hat{k}_\sigma(x, \acute{x})(\acute{f}_P(\acute{x}) + 1) d\acute{x} - 1 \\
&\geq \int_{B(x,\tau_x)} \hat{k}_\sigma(x, \acute{x})(\acute{f}_P(\acute{x}) + 1) d\acute{x} - 1.
\end{aligned}$$

When $x \in X_1$, we have observed that $B(x, \tau_x) \subset \acute{X}_1$ so that $\acute{f}_P(\acute{x}) = 1$ for all $\acute{x} \in B(x, \tau_x)$ and so obtain

$$\hat{f}(x) \;\geq\; 2 \int_{B(x,\tau_x)} \hat{k}_\sigma(x, \acute{x}) d\acute{x} - 1 \;=\; 2P_{\gamma_\sigma}(|u| < \tau_x) - 1 \;=\; 1 - 2P_{\gamma_\sigma}(|u| \geq \tau_x)\,,$$

where $\gamma_\sigma = \sigma^d(\pi)^{-\frac{d}{2}} e^{-\sigma^2|u|^2} du$ is a spherical Gaussian in $\mathbb{R}^d$. According to the tail bound inequality [5, Inequality 3.5, p. 59] for spherical Gaussians we have

$$P_{\gamma_\sigma}(|u| \geq r) \;\leq\; 4e^{-\sigma^2 r^2/4d}\,.$$

Consequently, for $x \in X_1$ we obtain

$$1 \;\geq\; \hat{f}(x) \;\geq\; 1 - 8e^{-\sigma^2 \tau_x^2/4d}.$$

For $x \in X_{-1}$ we analogously obtain that

$$-1 \;\leq\; \hat{f}(x) \;\leq\; -1 + 8e^{-\sigma^2 \tau_x^2/4d}$$

so that on $X_1 \cup X_{-1}$ we have

$$|\hat{K}_\sigma \acute{f}_P(x) - f_P(x)| \;\leq\; 8e^{-\sigma^2 \tau_x^2/4d}.$$

Since $\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P} = \mathbb{E}_{P_X}(|2\eta - 1||\hat{K}_\sigma \acute{f}_P - f_P|)$, combining this with the geometric noise assumption of Theorem 1.12 completes the proof of that theorem.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. `http://stat-www.berkeley.edu/tech-reports/638.pdf`, 2003.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1997.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, Berlin, 1991.

[6] S. Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, 48:1977–1991, 2002.

[7] C. Scovel and I. Steinwart. Fast rates for support vector machines. *Ann. Statist.*, submitted, 2003. `http://www.c3.lanl.gov/~ingo/publications/ann-03.ps`.

[8] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, 1:17–41, 2003.

[9] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2001.

[10] I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18:768–791, 2002.

[11] I. Steinwart. Consistency of support vector machines and other regularized kernel machine. *IEEE Trans. Inform. Theory*, accepted with minor revisions, 2003. `http://www.c3.lanl.gov/~ingo/publications/info-02.ps`.

[12] I. Steinwart. Sparseness of support vector machines. *J. Mach. Learn. Res.*, 4:1071–1105, 2003.

[13] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32:135–166, 2004.

[14] Q. Wu and D.-X. Zhou. Analysis of support vector machine classification. Tech. Report, City University of Hong Kong, 2003.

[15] Y. Yang. Minimax nonparametric classification—part I and II. *IEEE Trans. Inform. Theory*, 45:2271–2292, 1999.

[16] T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32:56–134, 2004.