Fast Rates for Support Vector Machines using Gaussian Kernels*†‡

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

February 1, 2005

Abstract

We establish learning rates up to the order of n^{-1} for support vector machines with hinge loss (L1-SVMs) and nontrivial distributions. For the stochastic analysis of these algorithms we use recently developed concepts such as Tsybakov's noise assumption and local Rademacher averages. Furthermore we introduce a new geometric noise condition for distributions that is used to bound the approximation error of Gaussian kernels in terms of their widths.

1 Introduction

In recent years support vector machines (SVMs) have been the subject of many theoretical considerations. Despite this effort, their learning performance on restricted classes of distributions is still widely unknown. In particular, it is unknown under which circumstances SVMs can guarantee *fast* rates with respect to the sample size n for their learning performance. The aim of this work is to use recently developed concepts like Tsybakov's noise assumption and local Rademacher averages to establish learning rates up to the order of n^{-1} for nontrivial distributions. In addition to these concepts which are used to deal with the stochastic part of the analysis we also introduce a geometric assumption for distributions that allows us to estimate the approximation properties of Gaussian kernels. Unlike many other concepts introduced for bounding the approximation error our geometric assumption is not in terms of smoothness but describes the concentration of the marginal distribution near the decision boundary.

Let us formally introduce the statistical classification problem. To this end assume for technical reasons that $X \subset \mathbb{R}^d$ is a compact subset. We write $Y := \{-1, 1\}$. Given a finite *training set* $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$ the classification task is to predict the *label* y of a new sample (x, y). In the standard batch model it is assumed that the samples (x_i, y_i) are i.i.d. according to an unknown (Borel) 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 $\operatorname{sign} f_T(x)$, where we choose a fixed definition of $\operatorname{sign}(0) \in \{-1, 1\}$. In order to "learn" from the samples of T the decision function $f_T : X \to \mathbb{R}$ should guarantee a small probability for the misclassification of the example (x, y).

*AMS 2000 subject classification: *primary* 68Q32, secondary 62G20, 62G99, 68T05, 68T10, 41A46, 41A99
†Los Alamos Unclassified Report 04-8796
‡Submitted to the *Annals of Statistics* on 12/23/04.

Here, misclassification means sign $f_T(x) \neq y$. 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 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* Borel probability measures P on $X \times Y$. Since $\mathcal{R}_P(f_T)$ 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 convergence rate in (2) for *all* distributions. Unfortunately, this is impossible by a result of Devroye (see [14, 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 $\eta(x) := P(y = 1|x)$ satisfies certain smoothness assumptions Yang showed in [40] that some plug-in rules achieve rates 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.
- Recently, for SVMs with hinge loss (L1-SVMs) Wu and Zhou [39] established rates under the assumption that η is contained in a Sobolev space. In particular, they obtained rates of the form $(\log n)^{-p}$ for some $p > 0$ if the L1-SVM uses a Gaussian kernel with *fixed* width.
- It is well known (see [14, Sec. 18.1]) 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 $n^{-1/2}$.
- Let P be a distribution with no noise regarding the labeling, i.e. P satisfies $\mathcal{R}_P = 0$, and $\mathcal{F}$ be a class with finite VC-dimension. If $\mathcal{F}$ contains a Bayes decision function then the learning rate of the ERM classifier over $\mathcal{F}$ is, up to a logarithmic factor, of the form n^{-1} (see e.g. [14, Sec. 12.7]).

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. Remarkably, Tsybakov recently closed this gap in [37] by showing that certain ERM-type classifiers learn with rates of the form $n^{-\frac{q+1}{q+pq+2}}$, where $0 \leq q \leq \infty$ is a parameter describing how well the noise in the labels, i.e. the function

$$x \mapsto \min\big\{1 - \eta(x), \eta(x)\big\} = \frac{1}{2} - \Big|\eta(x) - \frac{1}{2}\Big|\,, \tag{3}$$

is distributed around the critical level $1/2$ (see Definition 2.2 in the following section) and $0 < p < 1$ measures the complexity of the function class $\mathcal{F}$ the ERM method minimizes over. Furthermore, Tsybakov showed that for specific types of distributions, the above rates are actually optimal in a min-max sense. Unfortunately, the ERM-classifier he considered requires substantial knowledge on *how* to approximate the desired Bayes decision functions by $\mathcal{F}$. Moreover, ERM classifiers are based on combinatorial optimization problems and hence they are usually hard to implement and in general there exist no efficient algorithms.

On the one hand SVMs do not share the implementation issues of ERM since they are based on a convex optimization (see e.g. [12, 28] for algorithmic aspects). On the other hand, however, their

known learning rates are rather unsatisfactory since either the assumptions on the distributions are too restrictive as in [30] or the established learning rates are too slow as in [39]. Our aim is to give SVMs a better theoretical foundation by establishing fast learning rates for a wide class of distributions. To this end we propose a geometric noise assumption (see Definition 2.3) which describes the concentration of the measure $|2\eta - 1|dP_X$—where P_X is the marginal distribution of P with respect to X—near the decision boundary by a parameter $\alpha \in (0, \infty]$. This assumption is then used to determine the approximation properties of Gaussian kernels which are used in the SVMs we consider. Provided that the tuning parameters are optimally chosen our main result then shows that the resulting learning rates for these classifiers are essentially of the form

$$n^{-\frac{\alpha}{2\alpha+1}}$$

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

$$n^{-\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}}$$

if $\alpha > \frac{q+2}{2q}$. In particular, we obtain learning rates faster than $n^{-1/2}$ whenever $\alpha > \frac{3q+4}{2q}$.

The rest of this work is organized as follows: In Section 2 introduce all important concepts of this work and then present our main results. In Section 3 we recall some basic theory on reproducing kernel Hilbert spaces and prove a new covering number bound for Gaussian kernels that describes a trade-off between the kernel widths and the considered radii of the covering balls. We then show in Section 4 all results that are related to our proposed geometric noise assumption. The last sections of the work contain the actual proof of our rates: In Section 5 we establish a general bound for ERM-type classifiers involving local Rademacher averages which is used to bound the estimation error in our analysis of SVMs. In order to apply this result we need "variance bounds" for L1-SVMs which are established in Section 6. Interestingly, it turns out that sharp versions of these bounds depend on both Tsybakov's noise assumption and the approximation properties of the used kernel. Finally, we prove our learning rates in Section 7.

2 Definitions and Main Results

In this section we first recall some basic notions related to support vector machines which are needed throughout this text. In Subsection 2.2, we then present a covering number bound for Gaussian RBF kernels which will play an important role in our analysis of the estimation error of L1-SVMs. In Subsection 2.3 we recall Tsybakov's noise assumption which will allow us to establish learning rates faster than $n^{-1/2}$. Then, in Subsection 2.4, we introduce a new assumption which is used to estimate the approximation error for L1-SVMs with Gaussian RBF kernels. Finally, we present and discuss our learning rates in Subsection 2.5.

2.1 RKHSs, SVMs and basic definitions

For two functions f and g we use the notation $f(\lambda) \preceq g(\lambda)$ to mean that there exists a constant $C > 0$ such that $f(\lambda) \leq Cg(\lambda)$ over some specified range of values of λ. We also use the notation $\succeq$ with similar meaning and the notation $\sim$ when both $\preceq$ and $\succeq$ hold. In particular we use the same notation for sequences.

Recall (see e.g. [1, 6]) that every positive definite kernel $k : X \times X \to \mathbb{R}$ over a non-empty set X has a unique reproducing kernel Hilbert space H (RKHS) whose unit ball we denote by B_H. Although we sometimes use generic kernels and RKHSs we are mainly interested in Gaussian RBF kernels which are the most widely used kernels in practice. Recall that these kernels are of the form

$$k_\sigma(x, x') = \exp(-\sigma^2 \|x - x'\|_2^2), \qquad x, x' \in X,$$

where $X \subset \mathbb{R}^d$ is a (compact) subset and $\sigma > 0$ is a free parameter whose *inverse* $1/\sigma$ is called the *width* of k_σ. We usually denote the corresponding RKHSs which are thoroughly described in [34] by $H_\sigma(X)$ or simply H_σ.

Let us now recall the definition of SVMs. To this end let P be a distribution on $X \times Y$ and $l : Y \times \mathbb{R} \to [0, \infty)$ be the *hinge loss function*, i.e.

$$l(y, t) := \max\{0, 1 - yt\}$$

for all $y \in Y$ and $t \in \mathbb{R}$. Furthermore, we define the *l-risk* of a measurable function $f : X \to \mathbb{R}$ by

$$\mathcal{R}_{l,P}(f) := \mathbb{E}_{(x,y)\sim P} l(y, f(x)) \,.$$

Now let H be a RKHS over X consisting of measurable functions. For $\lambda > 0$ we denote a solution of

$$\arg\min_{\substack{f \in H \\ b \in \mathbb{R}}} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{l,P}(f + b) \Big) \tag{4}$$

by $(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})$. Recall that $\tilde{f}_{P,\lambda}$ is uniquely determined (see e.g. [32]) while in some situations this is not true for the *offset* $\tilde{b}_{P,\lambda}$. In general we thus assume that $\tilde{b}_{P,\lambda}$ is an arbitrary solution. However, for the (trivial) distributions that satisfy $P(\{y^*\}|x) = 1$ P_X-a.s. for some $y^* \in Y$ we explicitly set $\tilde{b}_{P,\lambda} := y^*$ in order to control the size of the offset. Furthermore, if P is an empirical distribution with respect to a training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$ we write $\mathcal{R}_{l,T}(f)$ and $(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda})$, i.e.

$$(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda}) \in \arg\min_{\substack{f \in H \\ b \in \mathbb{R}}} \Big(\lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} \max\big\{0, 1 - y_i(f(x_i) + b)\big\} \Big) \,.$$

Note that in this case the above condition under which we set $\tilde{b}_{T,\lambda} := y^*$ means that all labels y_i of T are equal to y^*. An algorithm that constructs $(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda})$ for every training set T is called *L1-SVM with offset*. Furthermore, for $\lambda > 0$ we denote the unique solution of

$$\arg\min_{f \in H} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{l,P}(f) \Big) \tag{5}$$

by $f_{P,\lambda}$ and for empirical distributions based on a training set T we again write $f_{T,\lambda}$. A corresponding algorithm is called *L1-SVM without offset*. Recall that under some assumptions on the used RKHS and the choice of the regularization parameter λ it can be shown that both L1-SVM variants are universally consistent (see [31, 41, 33]), however no satisfying result on convergence rates has been established, yet.

We also emphasize that in many theoretical papers only L1-SVMs without offset are considered. The reason for this is that the offset often causes serious technical problems and in some cases such as stability analysis the results are even false for L1-SVMs with offset (for an analysis on partially stable learning algorithms including L1-SVMs with offset which resolves many of these problems we refer to [16]). However, in practice usually L1-SVMs with offset are used and therefore we feel that these algorithms should be considered in theory, too. As we will see, our techniques can be applied for both variants. The resulting rates coincide.

2.2 Covering numbers for Gaussian RKHSs

In order to bound the estimation error of L1-SVMs we need a complexity measure for the used RKHSs which is introduced in this section. To this end let us first recall some notations: For a

subset $A \subset E$ of a Banach space E the *covering numbers* are defined by

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

where B_E denotes the closed unit ball of E. Moreover, for a bounded linear operator $S : E \to F$ between two Banach spaces E and F, the covering numbers are $\mathcal{N}(S, \varepsilon) := \mathcal{N}(SB_E, \varepsilon, F)$.

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 $f : X \times Y \to \mathbb{R}$ with norm

$$\|f\|_{L_2(T)} := \left(\frac{1}{n} \sum_{i=1}^{n} \big| f(x_i, y_i) \big|^2 \right)^{\frac{1}{2}} \tag{6}$$

by $L_2(T)$. In other words, $L_2(T)$ is a L_2-space with respect to the empirical measure of T. Note, that for a function $f : X \times Y \to \mathbb{R}$ a canonical representant in $L_2(T)$ is its restriction $f_{|T}$. Furthermore, we write $L_2(T_X)$ for the space of all (equivalence classes of) square integrable functions with respect to the empirical measure of $x_1, \ldots, x_n$.

The proof of our learning rates uses the behaviour of $\log \mathcal{N}(B_{H_\sigma(X)}, \varepsilon, L_2(T_X))$ in ε *and* σ in order to bound the estimation error. Unfortunately, all known results on covering numbers for Gaussian RBF kernels emphasize the role of ε and hence we will establish in Section 3 the following result, of its own interest, which describes a trade-off between the influence of ε and σ.

Theorem 2.1 *Let $\sigma \geq 1$, $X \subset \mathbb{R}^d$ be a compact subset with non-empty interior, and $H_\sigma(X)$ be the RKHS of the Gaussian RBF kernel k_σ on X. Then for all $0 < p \leq 2$ and all $\delta > 0$, there exists a constant $c_{p,\delta,d} > 0$ independent of σ such that for all $\varepsilon > 0$ we have*

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

2.3 Tsybakov's noise assumption

Without making assumption on the noise (3) it is impossible to obtain rates faster $n^{-1/2}$. In this section we hence recall Tsybakov's noise assumption which was used in [37] to establish rates up to the order of n^{-1} for certain ERM-type classifiers. It turns out in this work that this assumption also allows such fast rates for L1-SVMs.

In order to motivate Tsybakov's assumption let us first observe that by Equation (3) the function $|2\eta - 1|$ can be used to describe the noise in the labels of a distribution P. 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. The following modified version of Tsybakov's noise condition describes the size of the latter regions in terms of Lorentz spaces $L_{q,\infty}$ (see e.g. [5] for these spaces).

Definition 2.2 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 $(2\eta - 1)^{-1} \in L_{q,\infty}(P_X)$, i.e. there exists a constant $C > 0$ such that

$$P_X\big(\{x \in X : |2\eta(x) - 1| \leq t\}\big) \;\leq\; C \cdot t^q \tag{7}$$

for all sufficiently small $t > 0$.

It is easy to see that a distribution that has Tsybakov noise exponent q also has Tsybakov noise exponent q' for all $q' < q$. Furthermore, all distributions obviously 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, for $q < \infty$ the above condition is satisfied if and only if (7) holds for all $t > 0$ and a possibly different constant C. 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. (7) holds (see [37, Prop. 1]). As already mentioned in the introduction this condition and hence Tsybakov's noise exponent enables us to obtain fast classification rates for certain ERM algorithms (see [19, 37]). Furthermore, it can be used to improve inequalities between the excess classification risk and other excess risks (see [4]).

2.4 A new geometric assumption for distributions

In this section we introduce a condition for distributions that will us allow to estimate the approximation error for Gaussian RBF kernels. To this end let l be the hinge loss function and P be a distribution on X. Let

$$\mathcal{R}_{l,P} := \inf\bigl\{\mathcal{R}_{l,P}(f) \mid f : X \to \mathbb{R} \text{ measurable}\bigr\}$$

denote the smallest possible l-risk of P. Since functions achieving the minimal l-risk occur in many situations we denote them by $f_{l,P}$ if no confusion regarding the non-uniqueness of this symbol can be expected. Furthermore, recall that $f_{l,P}$ has a shape similar to the Bayes decision function $\operatorname{sign} f_P$ (see e.g. [32]). Now, given a RKHS H over X we define the *approximation error function* with respect to H and P by

$$a(\lambda) \;:=\; \inf_{f \in H}\Bigl(\lambda\|f\|_H^2 + \mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P}\Bigr)\,, \qquad\qquad \lambda \geq 0\,. \tag{8}$$

Note that the obvious analogue of the approximation error function *with offset* is not greater than the above approximation error function *without offset* and hence we restrict our attention to the latter for simplicity.

For $\lambda > 0$, the solution $f_{P,\lambda}$ of (5) obviously satisfies $a(\lambda) = \lambda\|f_{P,\lambda}\|_H^2 + \mathcal{R}_{l,P}(f_{P,\lambda}) - \mathcal{R}_{l,P}$ and hence $a(\lambda)$ describes how well $\lambda\|f_{P,\lambda}\|_H^2 + \mathcal{R}_{l,P}(f_{P,\lambda})$ approximates $\mathcal{R}_{l,P}$. Recall that it was shown in [33] that if X is a compact metric space and 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. Since $H_\sigma(X)$ is always dense in $C(X)$ for compact $X \subset \mathbb{R}^d$ these statements are in particular true for the approximation error functions $a_\sigma(.)$ of Gaussian RBF kernels with fixed width $1/\sigma$. However, we are not aware of any weak condition on η or P that ensures $a_\sigma(\lambda) \preceq \lambda^\beta$ for $\lambda \to 0$ and some $\beta > 0$, and the results of [29] indicate that such behaviour of $a_\sigma(.)$ may actually require very restrictive conditions.

In the following we will therefore present a condition on distributions P that allows us to estimate $a_\sigma(\lambda)$ by λ *and* σ. In particular it will turn out that $a_\sigma(\lambda) \to 0$ with a polynomial rate in λ if we relate σ to λ in a certain manner. In order to introduce this assumption on P we define the classes of P by $X_{-1} := \{x \in X : \eta(x) < \frac{1}{2}\}$, $X_1 := \{x \in X : \eta(x) > \frac{1}{2}\}$, and $X_0 := \{x \in X : \eta(x) = \frac{1}{2}\}$ for some choice of η. Note, that a Tsybakov noise exponent $q > 0$ implies $P_X(X_0) = 0$. Now we define a distance function $x \mapsto \tau_x$ by

$$\tau_x \;:=\; \begin{cases} d(x, X_0 \cup X_1), & \text{if } x \in X_{-1}, \\ d(x, X_0 \cup X_{-1}), & \text{if } x \in X_1, \\ 0, & \text{otherwise}\,, \end{cases} \tag{9}$$

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

Definition 2.3 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

$$\int_X |2\eta(x) - 1| \exp\Bigl(-\frac{\tau_x^2}{t}\Bigr) P_X(dx) \;\le\; C t^{\frac{\alpha d}{2}} \tag{10}$$

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 (10) 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 2.4 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 (10) is satisfied whenever

$$\bigl(x \mapsto \tau_x^{-1}\bigr) \in L_{\alpha d}\bigl(|2\eta - 1|dP_X\bigr)\,. \tag{11}$$

Now, let us suppose $X_0 = \emptyset$ for a moment. In this case 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 (11) 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 $H_\sigma(X)$ for any $\sigma > 0$. Note, that from (11) it is obvious that the parameter α in (11) 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.

We now describe a regularity condition on η near the decision boundary that can be used to guarantee a geometric noise exponent.

Definition 2.5 Let $X \subset \mathbb{R}^d$, P be a distribution on $X \times Y$, and $\gamma > 0$. We say that P has an *envelope of order* γ if there is a constant $c_\gamma > 0$ such that for P_X-almost all $x \in X$ the regular conditional probability $\eta(x) := P(y = 1|x)$ satisfies

$$|2\eta(x) - 1| \le c_\gamma \tau_x^\gamma\,. \tag{12}$$

Obviously, if P has an envelope of order γ then the graph of $x \mapsto 2\eta(x) - 1$ lies in a multiple of the envelope defined by τ_x^γ at the top $-\tau_x^\gamma$ at the bottom and hence η can be very irregular away from the decision boundary but cannot be discontinuous when crossing it. The rate of convergence of $\eta(x) \to 1/2$ for $\tau_x \to 0$ is described by γ.

Interestingly, for distributions having both an envelope of order γ and and a Tsybakov noise exponent q we can bound the geometric noise exponent as the following theorem, which is proved in Section 4, shows.

Theorem 2.6 *Let $X \subset \mathbb{R}^d$ be compact and P be a distribution on $X \times Y$ that has an envelope of order $\gamma > 0$ and a Tsybakov noise exponent $q \ge 0$. Then P has geometric noise exponent $\frac{q+1}{d}\gamma$ if $q \ge 1$, and geometric noise exponent α for all $\alpha < \frac{q+1}{d}\gamma$ otherwise.*

Now the main result of this subsection which is proved in Section 4 shows that for distributions having a nontrivial geometric noise exponent we can bound the approximation error function for Gaussian RBF kernels.

Theorem 2.7 *Let $\sigma > 0$, X be the closed unit ball of the Euclidean space $\mathbb{R}^d$, and $a_\sigma(.)$ be the approximation error function with respect to $H_\sigma(X)$. Furthermore, let P be a distribution on $X \times Y$ that has geometric noise exponent $0 < \alpha < \infty$ with constant C in (10). Then there is a constant $c_d > 0$ depending only on the dimension d such that for all $\lambda > 0$ we have*

$$a_\sigma(\lambda) \;\le\; c_d\Big(\sigma^d\lambda + C(4d)^{\frac{\alpha d}{2}}\sigma^{-\alpha d}\Big)\,. \tag{13}$$

In order to let the right side of (13) converge to zero it is necessary to assume both $\lambda \to 0$ and $\sigma \to \infty$. An easy consideration shows that the fastest convergence rate is achieved if $\sigma(\lambda) := \lambda^{-\frac{1}{(\alpha+1)d}}$. In this case we have $a_{\sigma(\lambda)}(\lambda) \preceq \lambda^{\frac{\alpha}{\alpha+1}}$. In particular we can obtain 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 $B_{H_{\sigma(\lambda)}}$ as we have seen in Theorem 2.1.

2.5 Learning rates for L1-SVMs using Gaussian RBF kernels

With the help of the geometric noise assumption we can now formulate our main result for L1-SVMs using Gaussian RBF kernels.

Theorem 2.8 *Let X be the closed unit ball of $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $q \in [0,\infty]$ and geometric noise exponent $\alpha \in (0,\infty)$. We define*

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

and $\sigma_n := \lambda_n^{-\frac{1}{(\alpha+1)d}}$ in both cases. Then for all $\varepsilon > 0$ there exists a $C > 0$ such that for all $x \ge 1$ and $n \ge 1$ the L1-SVM without offset using the Gaussian RBF kernel k_{σ_n} satisfies

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

if $\alpha \le (q+2)/2q$. Here $\Pr^*$ denotes the outer probability measure of P^n in order to avoid measurability considerations. Analogously, in the case $\alpha > \frac{q+2}{2q}$ we have

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

If $\alpha = \infty$ the latter concentration inequality holds if $\sigma_n = \sigma$ is a constant with $\sigma > 2\sqrt{d}$. Furthermore, all results hold for the L1-SVM with offset if $q > 0$.

Remark 2.9 The rates in the above theorem are faster than the "parametric" rate $n^{-1/2}$ if and only if $\alpha > \frac{3q+4}{2q}$. In particular, for $q = \infty$ this condition becomes $\alpha > \frac{3}{2}$ and in a "typical intermediate" case $q = 1$ (cf. [37]) it becomes $\alpha > \frac{7}{2}$.

Remark 2.10 It is important to note that our techniques can also be used to establish rates for other definitions of the sequences (λ_n) and (σ_n). In fact, Theorem 2.7 guarantees $a_{\sigma_n}(\lambda_n) \to 0$ (which is necessary for our techniques to produce any rate) if $\sigma_n \to \infty$ and $\sigma_n^d\lambda_n \to 0$. In particular, if $\lambda_n := n^{-\iota}$ and $\sigma_n := n^\kappa$ for some $\iota, \kappa > 0$ with $\kappa d < \iota$ these conditions are satisfied and a conceptually easy but technically involved modification of our proof can produce rates for certain ranges of ι (and thus κ). In order to keep the presentation as short as possible we omitted the details and focused on the best possible rates.

Remark 2.11 Unfortunately, the choice of λ_n and σ_n that yield the optimal rates within our techniques, require to know the values of α and q which are typically not available. We have not investigated how to adaptively choose λ_n and σ_n yet, but this important question will be attacked in future research.

Remark 2.12 Another interesting but open question is whether the obtained rates are optimal for the class of considered distributions. In order to approach this question let us consider the case $\alpha = \infty$, which roughly speaking describes the case of almost no approximation error. In this case our rates are essentially of the form $n^{\frac{q+1}{q+2}}$ which coincide with the rates Tsybakov (see [37]) achieved for certain ERM classifiers based on hypothesis classes of small complexity. The latter rates in turn cannot be improved in a min-max sense for certain classes of distributions as it was also shown in [37]. This discussion indicates that the techniques used for the stochastic part of our analysis may be strong enough to produce optimal results. However, if we consider the case $\alpha < \infty$ then the approximation error function described in Theorem 2.7 and its influence on the estimation error (see our proofs, in particular Section 5 and Section 7) have a significant impact on the obtained rates. Since the sharpness of Theorem 2.7 is unclear to us we make no conjecture regarding the optimality of our rates in the general case.

Acknowledgment. We thank V. Koltchinskii and O. Bousquet for suggesting the local Rademacher averages as a way to obtain good performance bounds for SVMs and D. Hush for suggesting that "we are now in a position to obtain rates to Bayes".

3 Proof of Theorem 2.1

The main goal of this section is to prove Theorem 2.1 in Subsection 3.2. To this end we provide some RKHS theory which is used throughout this work in Subsection 3.1.

3.1 Some basic RKHS theory

For the proofs of this section we have to recall some basic facts from the theory of RKHSs. To this end let $X \subset \mathbb{R}^d$ be a compact subset and $k : X \times X \to \mathbb{R}$ be a continuous and positive semi-definite kernel with RKHS H. Then H consists of continuous functions on X and for $f \in H$ we have

$$\|f\|_\infty \leq K\|f\|_H$$

where

$$K := \sup_{x\in X} \sqrt{k(x,x)}. \tag{14}$$

Consequently if we denote the embedding of the RKHS H into the space of continuous functions $C(X)$ by

$$J_H : H \to C(X) \tag{15}$$

we have $\|J_H\| \leq K$. Furthermore, let us recall the representation of H based on Mercer's theorem (see [13]). To this end let $K_X : L_2(X) \to L_2(X)$ be the integral operator that is defined by

$$K_X f(x) := \int_X k(x,x')f(x')dx' , \qquad f \in L_2(X), x \in X , \tag{16}$$

where $L_2(X)$ denotes the L_2-space on X with respect to the Lebesgue measure. Then it was shown in [13] that the unique square root $K_X^{\frac{1}{2}}$ of K_X is an isometric isomorphism between $L_2(X)$ and H, and hence we have

$$H = K_X^{\frac{1}{2}} L_2(X) \qquad \text{and} \qquad \|K_X^{\frac{1}{2}} f\|_H = \|f\|_{L_2(X)}, \qquad f \in L_2(X) .$$

3.2 Proof of Theorem 2.1

In order to prove Theorem 2.1 we need the following result which bounds the covering numbers of $H_\sigma(X)$ with respect to $C(X)$.

Theorem 3.1 *Let $\sigma \geq 1$, $0 < p < 2$ and $X \subset \mathbb{R}^d$ be a compact subset with non-empty interior. Then there is a constant $c_{p,d} > 0$ independent of σ such that for all $\varepsilon > 0$ we have*

$$\log \mathcal{N}\big(B_{H_\sigma(X)}, \epsilon, C(X)\big) \leq c_{p,d}\sigma^{(1-\frac{p}{4})d}\epsilon^{-p}.$$

Proof: Let B_d be the closed unit ball of the Euclidean space $\mathbb{R}^d$ and $\mathring{B}_d$ be its interior. Then there exists an $r \geq 1$ such that $X \subset rB_d$. Now, it was recently shown in [34] that both restrictions $H_\sigma(rB_d) \to H_\sigma(X)$ and $H_\sigma(rB_d) \to H_\sigma(\mathring{B}_d)$ are isometric isomorphisms. Consequently, in the following we assume without loss of generality that $X = B_d$ or $X = \mathring{B}_d$ and do not concern ourselves with the distinction of both cases.

Now let us write $H_\sigma := H_\sigma(X)$ and $J_\sigma := J_{H_\sigma} : H_\sigma \to C(X)$ in order to simplify notations. Furthermore, let $K_\sigma : L_2(X) \to L_2(X)$ be the integral operator of k_σ defined as in (16), and $\|\cdot\|$ denote the norm in $L_2(X)$. According to [13, Thm. 3, p. 27], for any $f \in H_\sigma$, we obtain

$$\inf_{\|K_\sigma^{-1}h\| \leq R} \|f - h\| \;\leq\; \frac{1}{R}\left\|K_\sigma^{-\frac{1}{2}} f\right\|^2 \;=\; \frac{1}{R}\,\|f\|_{H_\sigma}^2\,,$$

where we use the convention $\|K_\sigma^{-1}h\| = \infty$ if $h \notin K_\sigma L_2(X)$. Suppose now that $\mathcal{H} \subset L_2(X)$ is a dense Hilbert space with $\|h\| \leq \|h\|_{\mathcal{H}}$, and that we have $K_\sigma : L_2(X) \to \mathcal{H} \subset L_2(X)$ with $\|K_\sigma : L_2(X) \to \mathcal{H}\| \leq c_{\sigma,\mathcal{H}} < \infty$ for some constant $c_{\sigma,\mathcal{H}} > 0$. It follows that

$$\inf_{\|h\|_{\mathcal{H}} \leq c_{\sigma,\mathcal{H}} R} \|f - h\| \;\leq\; \inf_{\|K_\sigma^{-1}h\| \leq R} \|f - h\| \;\leq\; \frac{1}{R}\|f\|_{H_\sigma}^2$$

and hence

$$\inf_{\|h\|_{\mathcal{H}} \leq R} \|f - h\| \;\leq\; \frac{c_{\sigma,\mathcal{H}}}{R}\|f\|_{H_\sigma}^2.$$

By [29, Thm. 3.1] it follows that f is in the real interpolation space $(L_2(X), \mathcal{H})_{\frac{1}{2},\infty}$ (see [7] for the definition of interpolation spaces) and that its norm in this space satisfies

$$\|f\|_{\frac{1}{2},\infty} \leq 2\sqrt{c_{\sigma,\mathcal{H}}}\|f\|_{H_\sigma}.$$

Therefore we obtain a continuous embedding

$$\Upsilon_1 : H_\sigma \to (L_2(X), \mathcal{H})_{\frac{1}{2},\infty}$$

with $\|\Upsilon_1\| \leq 2\sqrt{c_{\sigma,\mathcal{H}}}$. If in addition a subset inclusion $(L_2(X), \mathcal{H})_{\frac{1}{2},\infty} \subset C(X)$ exists which defines a continuous embedding

$$\Upsilon_2 : (L_2(X), \mathcal{H})_{\frac{1}{2},\infty} \to C(X)$$

we have a factorization $J_\sigma = \Upsilon_2\Upsilon_1$ and can conclude

$$\log \mathcal{N}\big(B_{H_\sigma(X)}, \epsilon, C(X)\big) \;=\; \log \mathcal{N}(J_\sigma, \epsilon) \;\leq\; \log \mathcal{N}\Big(\Upsilon_2, \frac{\epsilon}{2\sqrt{c_{\sigma,\mathcal{H}}}}\Big). \tag{17}$$

Consequently to bound $\log \mathcal{N}(J_\sigma, \epsilon)$ we need to select an $\mathcal{H}$, compute $c_{\sigma,\mathcal{H}}$, and bound $\log \mathcal{N}(\Upsilon_2, \epsilon)$. To that end let $\mathcal{H} := W^m(\mathring{X})$ be the Sobolev space with norm

$$\|f\|_m^2 = \sum_{|\alpha| \leq m} \|D^\alpha f\|^2 ,$$

where $|\alpha| := \sum_{i=1}^d \alpha_i$, $D^\alpha := \prod_{i=1}^d \partial_i^{\alpha_i}$, and $\partial_i^{\alpha_i}$ denotes the α_i-th partial derivative in the i-th coordinate of $\mathbb{R}^d$. By the Cauchy-Schwartz inequality we have

$$\begin{aligned} \|D^\alpha K_\sigma f\|^2 \;=\; \int_{\mathring{X}} \Big| \int_{\mathring{X}} D_x^\alpha k_\sigma(x, \acute{x}) f(\acute{x}) d\acute{x} \Big|^2 dx \;&\leq\; \int_X \Big(\int_X \big| D_x^\alpha k_\sigma(x, \acute{x}) \big|^2 d\acute{x} \int_{\mathring{X}} f^2(\acute{x}) d\acute{x} \Big) dx \\ &\leq\; \|f\|^2 \int_X \int_X \big| D_x^\alpha k_\sigma(x, \acute{x}) \big|^2 d\acute{x} dx \,, \end{aligned} \tag{18}$$

where the notation D_x^α indicates that the differentiation takes place in the x variable. To address the term $D_x^\alpha k_\sigma(x, \acute{x})$ we note that

$$D_x^\alpha (e^{-|x|^2}) = (-1)^{|\alpha|} e^{-\frac{|x|^2}{2}} h_\alpha(x) ,$$

where the multivariate Hermite functions $h_\alpha(x) = \prod_{i=1}^d h_{\alpha_i}(x_i)$ are products of the univariate. Since $\int_{\mathbb{R}} h_k^2(x) dx = 2^k k! \sqrt{\pi}$ (see e.g. [11]) we obtain

$$\int_{\mathbb{R}^d} \Big| D_x^\alpha (e^{-|x|^2}) \Big|^2 dx = \int_{\mathbb{R}^d} e^{-|x|^2} h_\alpha^2(x) dx \leq \int_{\mathbb{R}^d} h_\alpha^2(x) dx = 2^{|\alpha|} \alpha! \pi^{\frac{d}{2}} , \tag{19}$$

where we used the definition $\alpha! := \prod_{i=1}^d \alpha_i!$. Applying the translation invariance of k_σ we obtain

$$\int_{\mathbb{R}^d} \big| D_x^\alpha k_\sigma(x, \acute{x}) \big|^2 d\acute{x} = \int_{\mathbb{R}^d} \big| D_{\acute{x}}^\alpha k_\sigma(0, \acute{x}) \big|^2 d\acute{x} = \int_{\mathbb{R}^d} \Big| D_{\acute{x}}^\alpha (e^{-\sigma^2 |\acute{x}|^2}) \Big|^2 d\acute{x}$$

and by a change of variables we can apply inequality (19) to the integral on the right side

$$\int_{\mathbb{R}^d} \Big| D_{\acute{x}}^\alpha (e^{-\sigma^2 |\acute{x}|^2}) \Big|^2 d\acute{x} = \sigma^{2|\alpha|-d} \int_{\mathbb{R}^d} \Big| D_{\acute{x}}^\alpha (e^{-|\acute{x}|^2}) \Big|^2 d\acute{x} \leq \sigma^{2|\alpha|-d} 2^{|\alpha|} \alpha! \pi^{\frac{d}{2}}$$

Hence we obtain

$$\int_X \int_X |D_x^\alpha k_\sigma(x, \acute{x})|^2 d\acute{x} dx \leq \theta(d) \sigma^{2|\alpha|-d} 2^{|\alpha|} \alpha! \pi^{\frac{d}{2}} ,$$

where $\theta(d)$ is the volume of X. Since $\sum_{|\alpha| \leq m} \alpha! \leq d^m m!^d$ and $\|K_\sigma f\|_m^2 = \sum_{|\alpha| \leq m} \|D^\alpha K_\sigma f\|^2$ we can therefore infer from (18) that for $\sigma \geq 1$ we have

$$\|K_\sigma\| \;\leq\; \sqrt{\theta(d)} (2d)^{\frac{m}{2}} m!^{\frac{d}{2}} \sigma^{m - \frac{d}{2}} \;=:\; c_{\sigma,\mathcal{H}} . \tag{20}$$

Now let us consider $\Upsilon_2 : (L_2(X), W^m(\mathring{X}))_{\frac{1}{2},\infty} \to C(X)$. According to Triebel [36, p. 267] we have

$$\big(L_2(X), W^m(\mathring{X})\big)_{\frac{1}{2},\infty} = \big(L_2(\mathring{X}), W^m(\mathring{X})\big)_{\frac{1}{2},\infty} = B_{2,\infty}^{\frac{m}{2}}(\mathring{X})$$

isomorphically. Furthermore

$$\log \mathcal{N}\Big(B_{2,\infty}^{\frac{m}{2}}(\mathring{X}) \to C(X), \epsilon \Big) \leq c_{m,d}\, \epsilon^{-\frac{2d}{m}} \tag{21}$$

for $m > d$ follows from a similar result of Birman and Solomyak's ([8], cf. also [36]) for Slobodeckij (i.e. fractional Sobolev) spaces, where the constant $c_{m,d}$ depends only on m and d. Consequently we obtain from (17), (20) and (21) that

$$\log \mathcal{N}(J_\sigma, \epsilon) \;\leq\; c_{m,d}\left(\frac{\epsilon}{2\sqrt{c_{\sigma,\mathcal{H}}}}\right)^{-\frac{2d}{m}} \;=\; c_{m,d}\,(4c_{\sigma,\mathcal{H}})^{\frac{d}{m}}\,\epsilon^{-\frac{2d}{m}} \;=\; \tilde{c}_{m,d}\,\sigma^{d-\frac{d^2}{2m}}\,\epsilon^{-\frac{2d}{m}}$$

for all $m > d$ and new constants $\tilde{c}_{m,d}$ only depending on m and d. Setting $m := 2d/p$ finishes the proof of Theorem 3.1. ∎

Proof of Theorem 2.1: As in the previous proof we write $H_\sigma := H_\sigma(X)$ and $J_\sigma := J_{H_\sigma} : H_\sigma \to C(X)$ in order to simplify notations. Furthermore recall that for a training set $T \in (X \times Y)^n$ the space $L_2(T_X)$ was introduced in Subsection 2.2. Now let $R_{T_X} : C(X) \to L_2(T_X)$ be the restriction map defined by $f \mapsto f_{|T_X}$. Obviously, we have $\|R_{T_X}\| \leq 1$. Furthermore we define $I_\sigma := R_{T_X} \circ J_\sigma$ so that $I_\sigma : H_\sigma \to L_2(T_X)$ is the evaluation map. Then Theorem 3.1 and the product rule for covering numbers imply that

$$\sup_{T \in Z^n} \log \mathcal{N}(I_\sigma, \varepsilon) \;\leq\; c_{q,d}\,\sigma^{(1-\frac{q}{4})d}\varepsilon^{-q} \tag{22}$$

for all $0 < q < 2$. To complete the proof of Theorem 2.1 we derive another bound on the covering numbers and interpolate the two. To that end observe that $I_\sigma : H_\sigma \to L_2(T_X)$ factors through $C(X)$ with both factors J_s and R_{T_X} having norm not greater than 1. Hence Proposition 17.3.7 in [23] implies that I_σ is absolutely 2-summing with 2-summing norm not greater than 1. By König's theorem ([24, Lem. 2.7.2]) we obtain for the approximation numbers $(a_k(I_\sigma))$ of I_σ that $\sum_{k\geq 1} a_k^2(I_\sigma) \leq 1$ for all $\sigma > 0$. Since the approximation numbers are decreasing it follows that $\sup_k k^{\frac{1}{2}} a_k(I_\sigma) \leq 1$. Using Carl's inequality between approximation and entropy numbers (see Theorem 3.1.1 in [10]) we thus find a constant $\tilde{c} > 0$ such that

$$\sup_{T \in Z^n} \log \mathcal{N}(I_\sigma, \varepsilon) \;\leq\; \tilde{c}\varepsilon^{-2} \tag{23}$$

for all $\varepsilon > 0$ and all $\sigma > 0$. Let us now interpolate the bound (23) with the bound (22). Since $\|I_\sigma : H_\sigma \to L_2(T_X)\| \leq 1$ we only need to consider $0 < \varepsilon \leq 1$. Let $0 < q < p < 2$ and $0 < a \leq 1$. Then for $0 < \varepsilon < a$ we have

$$\log \mathcal{N}(I_\sigma, \varepsilon) \;\leq\; c_{q,d}\sigma^{\left(1-\frac{q}{4}\right)d}\varepsilon^{-q} \;\leq\; c_{q,d}\sigma^{\left(1-\frac{q}{4}\right)d}a^{p-q}\varepsilon^{-p}\,,$$

and for $a \leq \varepsilon \leq 1$ we find

$$\log \mathcal{N}(I_\sigma, \varepsilon) \;\leq\; \tilde{c}\varepsilon^{-2} \;\leq\; \tilde{c}a^{p-2}\varepsilon^{-p}\,.$$

Since $\sigma \geq 1$ we can set $a := \sigma^{-\frac{4-q}{8-4q}\cdot d}$ and obtain

$$\log \mathcal{N}(I_\sigma, \varepsilon) \;\leq\; \tilde{c}_{q,d}\sigma^{(1-\frac{p}{2})\cdot\frac{8-2q}{8-4q}\cdot d}\varepsilon^{-p}\,,$$

where $\tilde{c}_{q,d}$ is a constant depending only on q, d. The proof is finished by choosing $q := \frac{4\delta}{1+2\delta}$ when $\delta < \frac{2p}{8-4p}$ and q just smaller than p otherwise. ∎

4 Proof of the Theorems 2.7 and 2.6

In this section we prove the Theorems 2.7 and 2.6 which both deal with the geometric noise exponent.

4.1 Proof of Theorem 2.7

Before we prove Theorem 2.7 we briefly explain the main idea of this proof. To this end let f_P be a Bayes decision function with values in $\{-1,1\}$ and $\hat{K}_{\Omega,\sigma}$ be a normalized variant (see definition below (25)) of the integral operator associated to k_σ on $\Omega \subset \mathbb{R}^d$. Then smoothing f_P by $\hat{K}_{\Omega,\sigma}$ gives a function $\hat{K}_{\Omega,\sigma} f_P \in H_\sigma(\Omega)$ whose RKHS norm will be computed in Lemma 4.1. Furthermore, we will obtain $-1 \leq \hat{K}_{\Omega,\sigma} f_P \leq 1$ and hence the equation

$$\mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P} = \mathbb{E}_{P_X}\left(|2\eta - 1|\,|f - f_P|\right), \qquad f : X \to [-1,1], \tag{24}$$

shown by Zhang [41] can be used to estimate the excess l-risk $\mathcal{R}_{l,P}(\hat{K}_{\Omega,\sigma} f_P) - \mathcal{R}_{l,P}$ of $\hat{K}_{\Omega,\sigma} f_P$. Finally, we use the geometric noise exponent to bound $|\hat{K}_{\Omega,\sigma} f_P(x) - f_P(x)|$.

Let us now introduce some notations for the proof of Theorem 2.7. To this end let us first denote the Euclidean norm on $\mathbb{R}^d$ by $|\cdot|$ in order to avoid confusions with other arising norms. Since it will be useful to consider the integral operators and their associated RKHSs on more general sets than the closed unit ball, let $\Omega \subset \mathbb{R}^d$ be measurable and let $K_{\Omega,\sigma} : L_2(\Omega) \to L_2(\Omega)$ denote the integral operator associated to the restriction of k_σ to Ω. Furthermore let $i_\Omega : L_2(\Omega) \to L_2(\mathbb{R}^d)$ denote the extension of a function on Ω by zero to the rest of $\mathbb{R}^d$ and $r_\Omega : L_2(\mathbb{R}^d) \to L_2(\Omega)$ denote the restriction of a function on $\mathbb{R}^d$ to the set Ω. Obviously we have $\|i_\Omega\| = 1$, $\|r_\Omega\| \leq 1$ and

$$K_{\Omega,\sigma} = r_\Omega K_{\mathbb{R}^d,\sigma} i_\Omega. \tag{25}$$

It will also be useful to consider the normalized Gaussian kernel

$$\hat{k}_\sigma(x,x') := \sigma^d \pi^{-\frac{d}{2}} k_\sigma(x,x') = \sigma^d \pi^{-\frac{d}{2}} e^{-\sigma^2|x-x'|^2}\,.$$

Recall that its associated integral operator $\hat{K}_{\mathbb{R}^d,\sigma}$ is known as the *Gauss-Weierstraß integral operator*. Finally, we define $\hat{K}_{\Omega,\sigma}$ analogously to $K_{\Omega,\sigma}$ and hence in particular we obtain an equation analogous to (25).

Let us first compute the RKHS norm of functions mapped from $L_2(\Omega)$ to $H_\sigma(\Omega)$ by $\hat{K}_{\Omega,\sigma}$.

Lemma 4.1 *For $g \in L_2(\Omega)$ we have $\hat{K}_{\Omega,\sigma} g \in H_\sigma(\Omega)$ and*

$$\|\hat{K}_{\Omega,\sigma} g\|_{H_\sigma(\Omega)} \leq \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \|g\|_{L_2(\Omega)}\,.$$

Proof: Since

$$\hat{K}_{\Omega,\sigma} g = \hat{K}^{\frac{1}{2}}_{\Omega,\sigma} \hat{K}^{\frac{1}{2}}_{\Omega,\sigma} g = \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} K^{\frac{1}{2}}_{\Omega,\sigma} \hat{K}^{\frac{1}{2}}_{\Omega,\sigma} g$$

and $\hat{K}^{\frac{1}{2}}_{\Omega,\sigma} g \in L_2(\Omega)$ we observe from the discussion on RKHS in Subsection 3.1 that the first assertion is proved. Using the shorthand notation $\|\cdot\|_\sigma$ for $\|\cdot\|_{H_\sigma(\Omega)}$, we also obtain

$$\begin{aligned}
\|\hat{K}_{\Omega,\sigma} g\|_\sigma &= \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \|K^{\frac{1}{2}}_{\Omega,\sigma} \hat{K}^{\frac{1}{2}}_{\Omega,\sigma} g\|_\sigma \\
&= \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \|\hat{K}^{\frac{1}{2}}_{\Omega,\sigma} g\|_{L_2(\Omega)} \\
&\leq \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \|\hat{K}^{\frac{1}{2}}_{\Omega,\sigma}\| \|g\|_{L_2(\Omega)}\,.
\end{aligned}$$

The continuous functional calculus theorem for self adjoint operators (see e.g. [25]) implies that $\|\hat{K}^{\frac{1}{2}}_{\Omega,\sigma}\| = \|\hat{K}_{\Omega,\sigma}\|^{\frac{1}{2}}$. Therefore to finish the proof we only need to show that $\hat{K}_{\Omega,\sigma}$ is a contraction on $L_2(\Omega)$. To that end, recall that Young's inequality [26] states that for convolutions we have

$$\|f * g\|_{L_2(\mathbb{R}^d)} \leq \|f\|_{L_1(\mathbb{R}^d)} \|g\|_{L_2(\mathbb{R}^d)}$$

and since the Gauss-Weierstraß integral operator $\hat{K}_{\mathbb{R}^d,\sigma}$ is a convolution and $\int \sigma^d \pi^{-\frac{d}{2}} e^{-\sigma^2|x|^2} dx = 1$ it follows that $\hat{K}_{\mathbb{R}^d,\sigma}$ is a contraction. From (25) we have $\hat{K}_{\Omega,\sigma} = r_\Omega \hat{K}_{\mathbb{R}^d,\sigma} i_\Omega$ and since $\|i_\Omega\| = 1$ and $\|r_\Omega\| \leq 1$ it follows that $\|\hat{K}_{\Omega,\sigma}\| \leq 1$. ∎

Proof of Theorem 2.7: We utilize the trivial estimate

$$a_\sigma(\lambda) \leq \lambda \|f\|_\sigma^2 + \mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P}, \quad f \in H_\sigma(X) \tag{26}$$

to bound the approximation error function through a judicious choice of function $\hat{f} \in H_\sigma(X)$. To this end let f_P be any Bayes decision 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 a extension $\acute{f}_P$ of f_P to $\acute{X} := 3X$. To do so first consider the extension $\acute{\eta}$ of η that is constant in the outward radial direction, i.e.

$$\acute{\eta}(x) = \begin{cases} \eta(x), & \text{if } |x| \leq 1 \\ \eta\big(\frac{x}{|x|}\big), & \text{otherwise.} \end{cases} \tag{27}$$

Let us also define $\acute{X}_{-1} := \{x \in \acute{X} : \acute{\eta}(x) < \frac{1}{2}\}$ and $\acute{X}_1 := \{x \in \acute{X} : \acute{\eta}(x) > \frac{1}{2}\}$. The following lemma in which $B(x,r)$ denotes the open ball of radius r about x in $\mathbb{R}^d$ shows that this extension cooperates well with τ_x.

Lemma 4.2 *For $x \in X_1$, we have $B(x,\tau_x) \subset \acute{X}_1$ and for $x \in X_{-1}$, we have $B(x,\tau_x) \subset \acute{X}_{-1}$.*

Proof: Let $x \in X_1$ and $x' \in B(x,\tau_x)$. If $x' \in X$ we have $|x - x'| < \tau_x$ which implies $\eta(x) > \frac{1}{2}$ by the definition of τ_x. This shows $x' \in \acute{X}_1$. Now let us assume $|x'| > 1$. Since $|\langle x, x'\rangle| \leq |x'|$ and Pythagoras theorem we then obtain

$$\begin{aligned} \left|\frac{x'}{|x'|} - x\right|^2 &= \left|\frac{x'}{|x'|} - \frac{\langle x,x'\rangle x'}{|x'|^2}\right|^2 + \left|\frac{\langle x,x'\rangle x'}{|x'|^2} - x\right|^2 \leq \left|x' - \frac{\langle x,x'\rangle x'}{|x'|^2}\right|^2 + \left|\frac{\langle x,x'\rangle x'}{|x'|^2} - x\right|^2 \\ &= |x' - x|^2 . \end{aligned}$$

Therefore, we have $\left|\frac{x'}{|x'|} - x\right| < \tau_x$ which implies $\acute{\eta}(x') = \eta\big(\frac{x'}{|x'|}\big) > \frac{1}{2}$. ∎

In order to proceed with the proof of Theorem 2.7 let $\acute{f}_P : \acute{X} \to [-1,1]$ be a measurable extension of f_P that satisfies $\acute{f}_P = 1$ on $\acute{X}_1$ and $\acute{f}_P = -1$ on $\acute{X}_{-1}$. We define $\hat{f} := r_X \hat{K}_{\acute{X},\sigma} \acute{f}_P$.

Let us first determine the RKHS norm of $\hat{f}$. To this end recall that according to Aronszajn [1] we have $r_X H_\sigma(\acute{X}) \subset H_\sigma(X)$ and

$$\|r_X f\|_{H_\sigma(X)} \leq \|f\|_{H_\sigma(\acute{X})}, \qquad f \in H_\sigma(\acute{X}).$$

Therefore by Lemma 4.1 applied to $\Omega := \acute{X}$ we obtain $\hat{f} = r_X \hat{K}_{\acute{X},\sigma} \acute{f}_P \in H_\sigma(\acute{X})$ and

$$\|\hat{f}\|_{H_\sigma(X)} \leq \|\hat{K}_{\acute{X},\sigma} \acute{f}_P\|_{H_\sigma(\acute{X})} \leq \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \|\acute{f}_P\|_{L_2(\acute{X})} \leq \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \operatorname{vol}(\acute{X}) = \sigma^{\frac{d}{2}} \Big(\frac{81}{\pi}\Big)^{\frac{d}{4}} \theta(d), \tag{28}$$

where $\theta(d) = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$ is the volume of X.

Let us now bound the term $\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P}$ in the right-hand side of inequality (26). To this end observe that $-1 \leq \acute{f}_P \leq 1$ obviously implies $-1 \leq i_{\acute{X}} \acute{f}_P \leq 1$, and hence a well-known property of the Gauss-Weierstraß integral operator yields

$$-1 \leq \hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}} \acute{f}_P \leq 1.$$

Since $\hat{K}_{\acute{X},\sigma} = r_{\acute{X}} \hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}}$ and P_X has support in X, Zhang's equation (24) implies

$$\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P} \;=\; \mathcal{R}_{l,P}\big(\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}} \acute{f}_P\big) - \mathcal{R}_{l,P} \;=\; \mathbb{E}_{P_X}\big(|2\eta - 1| \cdot |\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}} \acute{f}_P - f_P|\,\big). \tag{29}$$

In order to bound $|\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}} \acute{f}_P(x) - f_P(x)|$ for $x \in X_1$ we observe

$$\begin{aligned}
\hat{f}(x) = \int_{\acute{X}} \hat{k}_\sigma(x,x') \acute{f}_P(x')dx' \;&=\; \int_{\mathbb{R}^d} \hat{k}_\sigma(x,x') i_{\acute{X}} \acute{f}_P(x')dx' \\
&=\; \int_{\mathbb{R}^d} \hat{k}_\sigma(x,x')\big(i_{\acute{X}} \acute{f}_P(x') + 1\big)dx' - 1 \\
&\geq\; \int_{B(x,\tau_x)} \hat{k}_\sigma(x,x')\big(i_{\acute{X}} \acute{f}_P(x') + 1\big)dx' - 1\,.
\end{aligned} \tag{30}$$

Now remember that Lemma 4.2 showed $B(x,\tau_x) \subset \acute{X}_1$ for all $x \in X_1$ so that (30) implies

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

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 [18, inequality (3.5) on p. 59] we have $P_{\gamma_\sigma}(|u| \geq r) \leq 4e^{-\sigma^2 r^2/4d}$ and consequently we obtain

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

for all $x \in X_1$. Since for $x \in X_{-1}$ we analogously obtain $-1 \leq \hat{f}(x) \leq -1 + 8e^{-\sigma^2\tau_x^2/4d}$ we conclude

$$\big|\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}} \acute{f}_P(x) - f_P(x)\big| \;\leq\; 8e^{-\sigma^2\tau_x^2/4d}$$

for all $x \in X_1 \cup X_{-1}$. Consequently (29) and the geometric noise assumption for $t := \frac{4d}{\sigma^2}$ yields

$$\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P} \;\leq\; 8\mathbb{E}_{x\sim P_X}\Big(|2\eta(x) - 1|e^{-\sigma^2\tau_x^2/4d}\Big) \;\leq\; 8C(4d)^{\frac{\alpha d}{2}} \sigma^{-\alpha d}\,, \tag{31}$$

where C is the constant in (10). Now using (31) and (28) the estimate (26) applied to $\hat{f}$ implies the assertion. ∎

4.2 Proof of Theorem 2.6

In this subsection, all Lebesgue and Lorentz spaces (see e.g. [5]) and their norms are with respect to the measure P_X.

Proof of Theorem 2.6: Let us first consider the case $q \geq 1$ where we can apply the Hölder inequality for Lorentz spaces ([22]) which states

$$\|fg\|_1 \leq \|f\|_{q,\infty} \|g\|_{q',1}$$

for all $f \in L_{q,\infty}$, $g \in L_{q',1}$ and q' defined by $\frac{1}{q} + \frac{1}{q'} = 1$. Applying this inequality gives

$$\begin{aligned}
\mathbb{E}_{x\sim P_X}\Big(|2\eta(x) - 1|e^{-\tau_x^2/t}\Big) \;&\leq\; \|(2\eta - 1)^{-1}\|_{q,\infty} \Big\| x \mapsto \big(2\eta(x) - 1\big)^2 e^{-\frac{\tau_x^2}{t}} \Big\|_{q',1} \\
&\leq\; C \Big\| \big(2\eta - 1\big)^2 e^{-\left(\frac{|2\eta-1|}{c_\gamma}\right)^{\frac{2}{\gamma}} t^{-1}} \Big\|_{q',1}
\end{aligned} \tag{32}$$

where in the last estimate we used the Tsybakov assumption (7) and the fact that P has an envelope of order γ. Let us write $h(x) := |2\eta(x)-1|^{-1}$, $x \in X$, and $b := t(c_\gamma)^{\frac{2}{\gamma}}$ so that

$$|2\eta(x)-1|^2 e^{-\left(\frac{|2\eta-1|}{c_\gamma}\right)^{\frac{2}{\gamma}} t^{-1}} = g\big(h(x)\big)\,,$$

where $g(s) := s^{-2}e^{-s^{-\frac{2}{\gamma}}b}$ for all $s \geq 1$. Now it is easy to see that $g : [1,\infty) \to [0,\infty)$ is strictly increasing whenever $0 < b \leq \frac{2}{3\gamma}$ and hence we can extend g to a strictly increasing, continuous and invertible function on $[0,\infty)$ in this case. Let us denote such an extension also by g. Then for this extension we have

$$P_X\big(g\circ h > \tau\big) = P_X\big(h > g^{-1}(\tau)\big). \tag{33}$$

Now for a function $f : X \to [0,\infty)$ recall the non-increasing rearrangement

$$f^*(u) \;:=\; \inf\big\{\sigma \geq 0 : P_X(f > \sigma) \leq u\big\}\,, \qquad\qquad u > 0$$

of f which can be used to define Lorentz norms (see e.g. [5]). For $u > 0$ equation (33) then yields

$$(g\circ h)^*(u) = \inf\big\{\sigma : P_X\big(h > g^{-1}(\sigma)\big) \leq u\big\} = g\Big(\inf\Big\{g^{-1}(\sigma) : P_X\big(h > g^{-1}(\sigma)\big) \leq u\Big\}\Big) = g\circ h^*(u)$$

Now, inequality (7) implies $P_X\big(h \geq \big(\frac{C}{u}\big)^{1/q}\big) \leq u$ for all $u > 0$. Therefore, we find

$$h^*(u) \;=\; \inf\big\{\sigma \geq 0 : P_X(h > \sigma) \leq u\big\} \;\leq\; \inf\big\{\sigma \geq 0 : P_X(h \geq \sigma) \leq u\big\} \;\leq\; \Big(\frac{C}{u}\Big)^{\frac{1}{q}}$$

for all $0 < u < 1$. Since $(g\circ h)^* = g\circ h^*$ and g is increasing we hence have

$$(g\circ h)^*(u) \;\leq\; g\Big(\Big(\frac{C}{u}\Big)^{\frac{1}{q}}\Big)$$

for all $0 < u < 1$. Now, for fixed $\hat\alpha > 0$ the bound $e^{-x} \preceq \frac{x^{-\hat\alpha}}{\ln^2(x)+1}$ on $(0,\infty)$ implies

$$g(s) \;\preceq\; b^{\hat\alpha}\frac{s^{2(\hat\alpha/\gamma-1)}}{\ln^2\big(s^{-2/\gamma}b^{-1}\big)+1}$$

for $s \in [1,\infty)$. Using that $(g\circ h)^*(u) = 0$ holds for all $u \geq 1$, we hence obtain

$$(g\circ h)^*(u) \;\preceq\; b^{\hat\alpha}\frac{u^{\frac{2}{q}\left(1-\frac{\hat\alpha}{\gamma}\right)}}{\ln^2\big(\big(\frac{u}{C}\big)^{\frac{2}{q\gamma}}b^{-1}\big)+1}\,.$$

for $u > 0$ if we assume without loss of generality that $C \geq 1$. Let us define $\hat\alpha := \gamma\frac{q+1}{2}$. Then we find $\frac{1}{q'} + \frac{2}{q}(1-\frac{\hat\alpha}{\gamma}) = 0$ and consequently for $b \leq \frac{2}{3\gamma}$, i.e. $t \leq \frac{2}{3\gamma(c_\gamma)^{2/\gamma}}$, we obtain

$$\|g\circ h\|_{q',1} \;=\; \int_0^\infty u^{\frac{1}{q'}-1}(g\circ h)^*(u)du \;\preceq\; b^{\hat\alpha}\int_0^\infty \frac{u^{-1}}{\ln^2\big(\big(\frac{u}{C}\big)^{\frac{2}{q\gamma}}b^{-1}\big)+1}du \;\preceq\; t^{\gamma\frac{q+1}{2}} \tag{34}$$

by the definition on b. Since we also have $\mathbb{E}_{P_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \leq 1$ for all $t > 0$ estimate (32) together the definition of g and (34) yields the assertion in the case $q \geq 1$.

Let us now consider the case $0 \le q < 1$ where the Hölder inequality in Lorentz space cannot be used. Then for all $t, \tau \ge 0$ we have

$$\begin{aligned}\mathbb{E}_{x\sim P_X}\Big(|2\eta(x)-1|e^{-\frac{\tau_x^2}{t}}\Big) &= \int_{|2\eta-1|\le\tau}|2\eta(x)-1|e^{-\frac{\tau_x^2}{t}}P_X(dx) + \int_{|2\eta-1|>\tau}|2\eta(x)-1|e^{-\frac{\tau_x^2}{t}}P_X(dx)\\ &\le C\tau^{q+1} + \exp\Big(-\Big(\frac{\tau}{c_\gamma}\Big)^{\frac{2}{\gamma}}t^{-1}\Big)\,, \qquad (35)\end{aligned}$$

where we used the Tsybakov assumption (7) and the fact that P has an envelope of order γ. Let us define τ by $\tau^{q+1} := \exp(-(\frac{\tau}{c_\gamma})^{2/\gamma}t^{-1})$. For $\hat{a} := (c_\gamma)^{2/\gamma}(q+1)$ and small t this definition implies

$$\tau \le \Big(\frac{\hat{a}\gamma}{2}\Big)^{\frac{\gamma}{2}}\Big(t\ln\frac{1}{\hat{a}t}\Big)^{\frac{\gamma}{2}}$$

and hence the assertion follows from (35) for the case $0 < q < 1$. ∎

5 The estimation error of ERM-type classifiers

In order to bound the estimation error in the proof of Theorem 2.8 we now establish a concentration inequality for ERM-type algorithms which is based on a variant of Talagrand's concentration inequality. Our approach is inspired by a similar result of [4] which uses a complexity measure closely related to local Rademacher averages. The latter have been intensively studied in learning theory in recent years (see [21], [2], and [3]). One of the main features of the concentration inequalities using local Rademacher averages is that they all need a so-called "variance bound" of the form $\mathbb{E}_P g^2 \le c\,(\mathbb{E}_P g)^\alpha$ for constants $\alpha > 0$, $c > 0$, and certain functions g. However, for L1-SVMs and distributions P satisfying Tsybakov's noise condition for some $0 < q \le \infty$ the "sharpest" variance bounds we can show in Section 6 are of the form $\mathbb{E}_P g^2 \le c\,(\mathbb{E}_P g)^\alpha + \delta$ with $\delta > 0$, where both c and δ depend on the regularization parameter λ. Since the latter changes with $n \to \infty$ the above mentioned theory must be adapted to this more general situation in order to obtain a full control over the crucial values c and δ.

This section is organized as follows: In Subsection 5.1 we present the required modification of the result of [4]. Then in Subsection 5.2 we bound the arising local Rademacher averages.

5.1 Bounding the estimation error using local Rademacher averages

We first have to introduce some notations. To this end let $\mathcal{F}$ be a class of bounded measurable functions from Z to $\mathbb{R}$. In order to avoid measurability considerations we always assume that $\mathcal{F}$ *is separable with respect to* $\|.\|_\infty$. Given a probability measure P on Z we define the modulus of continuity of $\mathcal{F}$ by

$$\omega_n(\mathcal{F},\varepsilon) \;:=\; \omega_{P,n}(\mathcal{F},\varepsilon) \;:=\; \mathbb{E}_{T\sim P^n}\Big(\sup_{\substack{f\in\mathcal{F},\\ \mathbb{E}_P f^2\le\varepsilon}}|\mathbb{E}_P f - \mathbb{E}_T f|\Big)\,,$$

where we emphasize that the supremum is, as a function from Z to $\mathbb{R}$, measurable by the separability assumption on $\mathcal{F}$. The modulus of continuity will serve as a complexity measure in the main theorem of this section. In Subsection 5.2 we will then bound $\omega_n(\mathcal{F},\varepsilon)$ by local Rademacher averages which themselves are treated by certain covering numbers.

We also need some notations related to ERM-type algorithms: let $\mathcal{F}$ be as above and $L : \mathcal{F} \times Z \to [0, \infty)$ be a function. We call L a *loss function* if $L \circ f := L(f, .)$ is measurable for all $f \in \mathcal{F}$. Given a probability measure P on Z we denote by $f_{P,\mathcal{F}} \in \mathcal{F}$ a minimizer of

$$f \mapsto \mathcal{R}_{L,P}(f) := \mathbb{E}_{z\sim P} L(f, z).$$

Throughout this paper $\mathcal{R}_{L,P}(f)$ is called the L-risk of f. If P is an empirical measure with respect to $T \in Z^n$ we write $f_{T,\mathcal{F}}$ and $\mathcal{R}_{L,T}(.)$ as usual. For simplicity, we assume throughout this section that $f_{P,\mathcal{F}}$ and $f_{T,\mathcal{F}}$ do exist. Furthermore, although there may be multiple solutions we use a single symbol for them whenever no confusion regarding the non-uniqueness of this symbol can be expected. An algorithm that produces solutions $f_{T,\mathcal{F}}$ is called an *empirical L-risk minimizer*. Moreover, if $\mathcal{F}$ is convex, we say that L is convex if $L(., z)$ is convex for all $z \in Z$. Finally, L is called *line-continuous* if for all $z \in Z$ and all $f, \hat{f} \in \mathcal{F}$ the function $t \mapsto L(tf + (1 - t)\hat{f}, z)$ is continuous on $[0, 1]$. If $\mathcal{F}$ is a vector space then every convex L is line-continuous. Now the main result of this section reads as follows:

Theorem 5.1 *Let $\mathcal{F}$ be a convex set of bounded measurable functions from Z to $\mathbb{R}$, and let $L : \mathcal{F} \times Z \to [0, \infty)$ be a convex and line-continuous loss function. For a probability measure P on Z we define*

$$\mathcal{G} \;:=\; \left\{ L \circ f - L \circ f_{P,\mathcal{F}} \;:\; f \in \mathcal{F} \right\}.$$

Suppose that there are constants $c \geq 0$, $0 < \alpha \leq 1$, $\delta \geq 0$ and $B > 0$ with $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ and $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$. Furthermore, assume that $\mathcal{G}$ is separable with respect to $\|.\|_\infty$. Let $n \geq 1$, $x \geq 1$ and $\varepsilon > 0$ with

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

Then we have

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

Remark 5.2 Theorem 5.1 has been proved in [4] 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 [4]. If $\delta > 0$ the above theorem *apparently cannot* produce rates faster than $n^{-\frac{1}{2}}$. However, in order to decrease the approximation error, the class $\mathcal{F}$ and (thus $\mathcal{G}$) increases with n for many algorithms. If for such sequences $(\mathcal{F}_n)$ 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 in Section 6 this phenomenon actually occurs for L1-SVMs and distributions satisfying Tsybakov's noise assumption for some exponent $q > 0$. Namely, we will show that the rate of $\delta_n \to 0$ and the values of both c and B are determined by the approximation error function. In particular, in our analysis approximation properties of H will heavily influence the estimation error. As far as we know such an interweaving of approximation and estimation error has never been observed or analyzed before.

As already mentioned, the proof of Theorem 5.1 is based on Talagrand's concentration inequality in [35] and its refinements in [27], [17], [20]. The below version of this inequality is derived from Bousquet's result in [9] using a little trick presented in [3, Lem. 2.5]:

Theorem 5.3 *Let P be a probability measure on Z and $\mathcal{H}$ be a set of bounded measurable functions from Z to $\mathbb{R}$ which is separable with respect to $\|.\|_\infty$ and satisfies $\mathbb{E}_P h = 0$ for all $h \in \mathcal{H}$.*

Furthermore, let $b > 0$ and $\tau \geq 0$ be constants with $\|h\|_\infty \leq b$ and $\mathbb{E}_P h^2 \leq \tau$ for all $h \in \mathcal{H}$. Then for all $x \geq 1$ and all $n \geq 1$ we have

$$P^n\left(T \in Z^n : \sup_{h\in\mathcal{H}} \mathbb{E}_T h > 3\mathbb{E}_{T'\sim P^n} \sup_{h\in\mathcal{H}} \mathbb{E}_{T'} h + \sqrt{\frac{2x\tau}{n}} + \frac{bx}{n}\right) \;\leq\; e^{-x}\,.$$

This concentration inequality is used to prove the following lemma which is a generalized version of Lemma 13 in [4]:

Lemma 5.4 *Let P be a probability measure on Z and $\mathcal{G}$ be a set of bounded measurable functions from Z to $\mathbb{R}$ which is separable with respect to $\|.\|_\infty$. Let $c \geq 0$, $0 < \alpha \leq 1$, $\delta \geq 0$ and $B > 0$ be constants with $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ and $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$. Furthermore, assume that for all $T \in Z^n$ and all $\varepsilon > 0$ for which for some $g \in \mathcal{G}$ we have*

$$\mathbb{E}_T g \leq \varepsilon/20 \qquad \text{and} \qquad \mathbb{E}_P g \geq \varepsilon$$

there is a $g^* \in \mathcal{G}$ which satisfies

$$\mathbb{E}_T g^* \leq \varepsilon/20 \qquad \text{and} \qquad \mathbb{E}_P g^* = \varepsilon\,.$$

Then for all $n \geq 1$, $x \geq 1$, and all $\varepsilon > 0$ satisfying

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

we have

$$\Pr^*\Big(T \in Z^n :\ \text{for all } g \in \mathcal{G} \text{ with } \mathbb{E}_T g \leq \varepsilon/20 \text{ we have } \mathbb{E}_P g < \varepsilon\Big) \;\geq\; 1 - e^{-x}\,.$$

Proof: We define $\mathcal{H} := \{\mathbb{E}_P g - g : g \in \mathcal{G}, \mathbb{E}_P g = \varepsilon\}$. Obviously, we have $\mathbb{E}_P h = 0$, $\|h\|_\infty \leq 2B$, and $\mathbb{E}_P h^2 = \mathbb{E}_P g^2 - (\mathbb{E}_P g)^2 \leq c\varepsilon^\alpha + \delta$ for all $h \in \mathcal{H}$. Moreover, since it is also easy to verify that $\mathcal{H}$ is separable with respect to $\|.\|_\infty$, our assumption on $\mathcal{G}$ yields

$$\begin{aligned}
& \Pr^*\big(T \in Z^n : \exists g \in \mathcal{G} \text{ with } \mathbb{E}_T g \leq \varepsilon/20 \text{ and } \mathbb{E}_P g \geq \varepsilon\big) \\
\leq\ & \Pr^*\big(T \in Z^n : \exists g \in \mathcal{G} \text{ with } \mathbb{E}_T g \leq \varepsilon/20 \text{ and } \mathbb{E}_P g = \varepsilon\big) \\
=\ & \Pr^*\big(T \in Z^n : \exists g \in \mathcal{G} \text{ with } \mathbb{E}_P g - \mathbb{E}_T g \geq 19\varepsilon/20 \text{ and } \mathbb{E}_P g = \varepsilon\big) \\
\leq\ & P^n\Big(T \in Z^n : \sup_{\substack{g\in\mathcal{G}\\ \mathbb{E}_P g = \varepsilon}} (\mathbb{E}_P g - \mathbb{E}_T g) \geq 19\varepsilon/20\Big) \\
=\ & P^n\big(T \in Z^n : \sup_{h\in\mathcal{H}} \mathbb{E}_T h \geq 19\varepsilon/20\big)\,.
\end{aligned}$$

Note, that since $\mathcal{H}$ is separable with respect to $\|.\|_\infty$ the sets in the last two lines are actually measurable. In order to bound the last probability we will apply Theorem 5.3. To this end we have to show $\frac{19\varepsilon}{20} > 3\mathbb{E}_{T'\sim P^n} \sup_{h\in\mathcal{H}} \mathbb{E}_{T'} h + \sqrt{\frac{2x\tau}{n}} + \frac{bx}{n}$. Our assumptions on ε imply

$$\varepsilon \;\geq\; 10\mathbb{E}_{T'\sim P^n}\Big(\sup_{\substack{g\in\mathcal{G},\\ \mathbb{E}_P g^2 \leq c\varepsilon^\alpha + \delta}} |\mathbb{E}_P g - \mathbb{E}_{T'} g|\Big) \;\geq\; 10\mathbb{E}_{T'\sim P^n} \sup_{h\in\mathcal{H}} \mathbb{E}_{T'} h\,. \tag{36}$$

Furthermore, since $10 \geq \left(\frac{60}{19}\right)^2$ and $0 < \alpha \leq 1$ we have

$$\varepsilon \;\geq\; 10\left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}} \;\geq\; 10\left(\frac{1}{10}\cdot\left(\frac{60}{19}\right)^2\right)^{\frac{1}{2-\alpha}}\left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}} \;\geq\; \left(\frac{60}{19}\right)^{\frac{2}{2-\alpha}}\left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}} \tag{37}$$

If $\delta \leq c\varepsilon^\alpha$ we hence find

$$\varepsilon \;\geq\; \left(\frac{60}{19}\right)^{\frac{2}{2-\alpha}}\left(\frac{2(c\varepsilon^\alpha+\delta)x}{\varepsilon^\alpha n}\right)^{\frac{1}{2-\alpha}}\,,$$

which implies $\frac{19}{60}\varepsilon \geq \sqrt{\frac{2(c\varepsilon^\alpha+\delta)x}{n}}$. Furthermore, if $\delta > c\varepsilon^\alpha$ the assumptions of the theorem shows

$$\varepsilon \;\geq\; 10\sqrt{\frac{\delta x}{n}} \;\geq\; \frac{60}{19}\sqrt{\frac{4\delta x}{n}} \;\geq\; \frac{60}{19}\sqrt{\frac{2(c\varepsilon^\alpha+\delta)x}{n}}\,.$$

Hence we have $\frac{19}{60}\varepsilon \geq \sqrt{\frac{2(c\varepsilon^\alpha+\delta)x}{n}}$ for all ε satisfying the assumptions of the theorem. Now let $\tau := c\varepsilon^\alpha + \delta$ and $b := 2B$. By (36) and $\varepsilon \geq \frac{10Bx}{n}$ we then find

$$\frac{19\varepsilon}{20} \;\geq\; \frac{19}{6}\mathbb{E}_{T'\sim P^n}\sup_{h\in\mathcal{H}}\mathbb{E}_{T'}h + \sqrt{\frac{2(c\varepsilon^\alpha+\delta)x}{n}} + \frac{19Bx}{6n} \;>\; 3\mathbb{E}_{T'\sim P^n}\sup_{h\in\mathcal{H}}\mathbb{E}_{T'}h + \sqrt{\frac{2x\tau}{n}} + \frac{bx}{n}\,.$$

Applying Theorem 5.3 then yields

$$\begin{aligned}
&\Pr^*\bigl(T\in Z^n : \exists g\in\mathcal{G} \text{ with } \mathbb{E}_T g \leq \varepsilon/20 \text{ and } \mathbb{E}_P g \geq \varepsilon\bigr)\\
\leq\;& P^n\bigl(T\in Z^n : \sup_{h\in\mathcal{H}}\mathbb{E}_T h \geq 19\varepsilon/20\bigr)\\
\leq\;& P^n\biggl(T\in Z^n : \sup_{h\in\mathcal{H}}\mathbb{E}_T h > 3\mathbb{E}_{T'\sim P^n}\sup_{h\in\mathcal{H}}\mathbb{E}_{T'}h + \sqrt{\frac{2x\tau}{n}} + \frac{bx}{n}\biggr)\\
\leq\;& e^{-x}\,.
\end{aligned}$$

■

With the help of the above lemma we can now prove the main result of this section, that is Theorem 5.1:

Proof of Theorem 5.1: In order to apply Lemma 5.4 to the class $\mathcal{G}$ it obviously suffices to show the richness condition on $\mathcal{G}$ of Lemma 5.4. To this end let $f \in \mathcal{F}$ with

$$\begin{aligned}
\mathbb{E}_T(L\circ f - L\circ f_{P,\mathcal{F}}) &\leq \varepsilon/20\\
\mathbb{E}_P(L\circ f - L\circ f_{P,\mathcal{F}}) &\geq \varepsilon\,.
\end{aligned}$$

For $t\in[0,1]$ we define $f_t := tf + (1-t)f_{P,\mathcal{F}}$. Since $\mathcal{F}$ is convex we have $f_t \in \mathcal{F}$ for all $t\in[0,1]$. By the line-continuity of L and Lebesgue's theorem we find that the map $h : t \mapsto \mathbb{E}_P(L\circ f_t - L\circ f_{P,\mathcal{F}})$ which maps from $[0,1]$ to $[0,B]$ is continuous. Since $h(0) = 0$ and $h(1) \geq \varepsilon$ there is a $t\in(0,1]$ with

$$\mathbb{E}_P(L\circ f_t - L\circ f_{P,\mathcal{F}}) \;=\; h(t) \;=\; \varepsilon$$

by the intermediate value theorem. Moreover, for this t we have

$$\mathbb{E}_T(L\circ f_t - L\circ f_{P,\mathcal{F}}) \;\leq\; \mathbb{E}_T\Bigl(tL\circ f + (1-t)L\circ f_{P,\mathcal{F}} - L\circ f_{P,\mathcal{F}}\Bigr) \;\leq\; \varepsilon/20\,.$$

Now, let $\varepsilon > 0$ with $\varepsilon \geq 10 \max\{\omega_n(\mathcal{G}, c\varepsilon^\alpha + \delta), \left(\frac{\delta x}{n}\right)^{\frac{1}{2}}, \left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}}, \frac{Bx}{n}\}$. Then, by Lemma 5.4 we find that with probability at least $1 - e^{-x}$ every $f \in \mathcal{F}$ with $\mathbb{E}_T(L \circ f - L \circ f_{P,\mathcal{F}}) \leq \varepsilon/20$ satisfies $\mathbb{E}_P(L \circ f - L \circ f_{P,\mathcal{F}}) < \varepsilon$. Since we always have

$$\mathbb{E}_T\big(L \circ f_{T,\mathcal{F}} - L \circ f_{P,\mathcal{F}}\big) \;\leq\; 0 \;<\; \varepsilon/20$$

we obtain the assertion. ∎

5.2 Bounding the local Rademacher averages

The aim of this subsection is to bound the modulus of continuity of the class $\mathcal{G}$ in Theorem 5.1. To this end we will first relate the modulus of continuity to local Rademacher averages. Then we will bound these averages with the help of covering numbers associated to $\mathcal{G}$ and reformulate Theorem 5.1.

Let us first recall the definition of (local) Rademacher averages. To this end let $\mathcal{F}$ be a class of bounded measurable functions from Z to $\mathbb{R}$ which is separable with respect to $\|.\|_\infty$. Furthermore, let P be a probability measure on Z and (ε_i) be 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. The *Rademacher average* of $\mathcal{F}$ is

$$\mathrm{Rad}_P(\mathcal{F}, n) \;:=\; \mathrm{Rad}(\mathcal{F}, n) \;:=\; \mathbb{E}_{P^n}\mathbb{E}_\mu \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(z_i)\right| .$$

Rademacher averages have been intensively used in empirical process theory. For more information we refer to [38]. Now for $\varepsilon > 0$ the *local Rademacher average* of $\mathcal{F}$ is defined by

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

Obviously, the local Rademacher average is a Rademacher average of a restricted function class. Furthermore for a given real number $a > 0$ we immediately obtain $\mathrm{Rad}(a\mathcal{F}, n) = a\mathrm{Rad}(\mathcal{F}, n)$ and

$$\mathrm{Rad}(a\mathcal{F}, n, \varepsilon) \;=\; a\mathrm{Rad}(\mathcal{F}, n, a^{-2}\varepsilon) . \tag{38}$$

Finally, by symmetrization the modulus of continuity can be estimated by the local Rademacher average. More precisely, we always have (see [38])

$$\omega_{P,n}(\mathcal{F}, \varepsilon) \;\leq\; 2\mathrm{Rad}_P(\mathcal{F}, n, \varepsilon) .$$

In the following we estimate Rademacher averages in terms of covering numbers using the path of [21]. Since we are interested in the arising constants, we add the proofs for the sake of completeness. We begin by recalling an extension of a theorem of Dudley to subgaussian processes proved in [38]. For the formulation we also refer to [21]:

Theorem 5.5 *There exists a universal constant $C > 0$ such that for all $\|.\|_\infty$-separable sets $\mathcal{F}$ of measurable functions from Z to $[-1, 1]$, all probability measures P on Z, and all $n \geq 1$ we have*

$$\mathrm{Rad}(\mathcal{F}, n) \;\leq\; \frac{C}{\sqrt{n}}\,\mathbb{E}_{T\sim P^n} \int_0^{\delta_T} \sqrt{\log \mathcal{N}\big(\mathcal{F}, \varepsilon, L_2(T)\big)}\, d\varepsilon ,$$

where $\delta_T := \sup_{f\in\mathcal{F}} \|f\|_{L_2(T)}$.

The next theorem due to Talagrand [35] estimates the expected diameter of $\mathcal{F}$ when interpreted as a subset of $L_2(T)$:

Theorem 5.6 *Let $\mathcal{F}$ be a class of measurable functions from Z to $[-1,1]$ which is separable with respect to $\|.\|_\infty$ and P be a probability measure on Z. Then we have*

$$\mathbb{E}_{T\sim P^n} \sup_{f\in\mathcal{F}} \|f\|^2_{L_2(T)} \;\leq\; 8\mathrm{Rad}(\mathcal{F},n) + \sup_{f\in\mathcal{F}} \mathbb{E}_P f^2 \,.$$

With the help of the above theorems we now can establish the following bound on the local Rademacher averages which is a slight modification of a result in [21]:

Proposition 5.7 *Let $\mathcal{F}$ be a class of measurable functions from Z to $[-1,1]$ which is separable with respect to $\|.\|_\infty$ and let P be a probability measure on Z. Assume there are constants $a>0$ and $0<p<2$ with*

$$\sup_{T\in Z^n} \log \mathcal{N}\big(\mathcal{F},\varepsilon,L_2(T)\big) \;\leq\; a\varepsilon^{-p}$$

for all $\varepsilon>0$. Then there exists a constant $c_p>0$ depending only on p with

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

Proof: We write $\mathcal{F}_\varepsilon := \{f : f\in\mathcal{F} \text{ and } \mathbb{E}_P f^2 \leq \varepsilon\}$ and $\delta_T := \sup_{f\in\mathcal{F}_\varepsilon} \|f\|_{L_2(T)}$. Then applying Theorem 5.5 and Theorem 5.6 to $\mathcal{F}_\varepsilon$ yields

$$\begin{aligned}
\mathrm{Rad}(\mathcal{F},n,\varepsilon) \;\leq\; \frac{C}{\sqrt{n}}\,\mathbb{E}_{T\sim P^n}\int_0^{\delta_T}\sqrt{\log\mathcal{N}\big(\mathcal{F}_\varepsilon,\delta,L_2(T)\big)}d\delta \;&\leq\; \frac{C\sqrt{a}}{\sqrt{n}}\,\mathbb{E}_{T\sim P^n}\int_0^{\delta_T}\delta^{-p/2}d\delta \\
&\leq\; \frac{c_p\sqrt{a}}{\sqrt{n}}\,\mathbb{E}_{T\sim P^n}\delta_T^{1-p/2} \\
&\leq\; \frac{c_p\sqrt{a}}{\sqrt{n}}\,\big(\mathbb{E}_{T\sim P^n}\delta_T^2\big)^{1/2-p/4} \\
&\leq\; \frac{c_p\sqrt{a}}{\sqrt{n}}\,\Big(8\mathrm{Rad}(\mathcal{F},n,\varepsilon)+\varepsilon\Big)^{1/2-p/4},
\end{aligned}$$

where $c_p>0$ is a constant depending only on p. If $\varepsilon \geq \mathrm{Rad}(\mathcal{F},n,\varepsilon)$ we hence find

$$\mathrm{Rad}(\mathcal{F},n,\varepsilon) \;\leq\; c_p'\sqrt{a}\,\varepsilon^{1/2-p/4}n^{-1/2}\,,$$

where $c_p' := 9^{1/2-p/4}c_p$. Conversely, if $\varepsilon < \mathrm{Rad}(\mathcal{F},n,\varepsilon)$ we obtain

$$\mathrm{Rad}(\mathcal{F},n,\varepsilon) \;\leq\; \frac{c_p'\sqrt{a}}{\sqrt{n}}\,\Big(\mathrm{Rad}(\mathcal{F},n,\varepsilon)\Big)^{1/2-p/4},$$

which implies

$$\mathrm{Rad}(\mathcal{F},n,\varepsilon) \;\leq\; c_p''\Big(\frac{a}{n}\Big)^{2/(2+p)},$$

where $c_p''>0$ is a constant depending only on p. ∎

Using the above proposition we may now replace the modulus of continuity in Theorem 5.1 by an assumption on the covering numbers of $\mathcal{G}$. As in Section 5 we assume that all minimizers exist. Then the corresponding result reads as follows:

Theorem 5.8 *Let $\mathcal{F}$ be a convex set of bounded measurable functions from Z to $\mathbb{R}$ and let $L : \mathcal{F} \times Z \to [0,\infty)$ be a convex and line-continuous loss function. For a probability measure P on Z we define*

$$\mathcal{G} \;:=\; \bigl\{L\circ f - L\circ f_{P,\mathcal{F}} \;:\; f\in\mathcal{F}\bigr\}\,.$$

Suppose that there are constants $c\geq 0$, $0<\alpha\leq 1$, $\delta\geq 0$ and $B>0$ with $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ and $\|g\|_\infty \leq B$ for all $g\in\mathcal{G}$. Furthermore, assume that $\mathcal{G}$ is separable with respect to $\|.\|_\infty$ and that there are constants $a\geq 1$ and $0<p<2$ with

$$\sup_{T\in Z^n} \log \mathcal{N}\bigl(B^{-1}\mathcal{G},\varepsilon,L_2(T)\bigr) \;\leq\; a\varepsilon^{-p} \tag{39}$$

for all $\varepsilon>0$. Then there exists a constant $c_p>0$ depending only on p such that for all $n\geq 1$ and all $x\geq 1$ we have

$$\Pr^*\Bigl(T\in Z^n : \mathcal{R}_{L,P}(f_{T,\mathcal{F}}) > \mathcal{R}_{L,P}(f_{P,\mathcal{F}}) + c_p\,\varepsilon(n,a,B,c,\delta,x)\Bigr) \leq e^{-x}\,,$$

where

$$\begin{aligned}\varepsilon(n,a,B,c,\delta,x) \;:=\; & B^{\frac{2p}{4-2\alpha+\alpha p}} c^{\frac{2-p}{4-2\alpha+\alpha p}} \Bigl(\frac{a}{n}\Bigr)^{\frac{2}{4-2\alpha+\alpha p}} + B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12} + B\Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}} \\ & + \sqrt{\frac{\delta x}{n}} + \Bigl(\frac{cx}{n}\Bigr)^{\frac{1}{2-\alpha}} + \frac{Bx}{n}\,.\end{aligned}$$

Proof: By (38) and Proposition 5.7 we find

$$\begin{aligned}\operatorname{Rad}(\mathcal{G},n,\varepsilon) \;=\; B\operatorname{Rad}(B^{-1}\mathcal{G},n,B^{-2}\varepsilon) \;&\leq\; c_p B\,\max\biggl\{B^{-1+\frac{p}{2}}\varepsilon^{\frac12-\frac{p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12}, \Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}}\biggr\} \\ &=\; c_p\max\biggl\{B^{\frac{p}{2}}\varepsilon^{\frac12-\frac{p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12}, B\Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}}\biggr\}\,.\end{aligned}$$

We assume without loss generality that $c_p\geq 5$. Let $\varepsilon^*>0$ be the largest real number that satisfies

$$\varepsilon^* = 2c_p B^{\frac{p}{2}}\bigl(c(\varepsilon^*)^\alpha+\delta\bigr)^{\frac12-\frac{p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12}\,. \tag{40}$$

Furthermore, let $\varepsilon>0$ be a such that

$$\varepsilon \;=\; 2c_p\max\biggl\{B^{\frac{p}{2}}(c\varepsilon^\alpha+\delta)^{\frac{2-p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12}, B\Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}}, \sqrt{\frac{\delta x}{n}}, \Bigl(\frac{4cx}{n}\Bigr)^{\frac{1}{2-\alpha}}, \frac{Bx}{n}\biggr\}\,.$$

It is easy to see that both ε and ε^* exist. Moreover, our above considerations show $\varepsilon \geq 10\max\bigl\{\omega_n(\mathcal{G},c\varepsilon^\alpha+\delta), \bigl(\frac{\delta x}{n}\bigr)^{\frac12}, \bigl(\frac{4cx}{n}\bigr)^{\frac{1}{2-\alpha}}, \frac{Bx}{n}\bigr\}$, i.e. ε satisfies the assumptions of Theorem 5.1. In order to show the assertion it therefore suffices to bound ε from above. To this end let us first assume that

$$B^{\frac{p}{2}}(c\varepsilon^\alpha+\delta)^{\frac{2-p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12} \;\geq\; \max\biggl\{B\Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}}, \sqrt{\frac{\delta x}{n}}, \Bigl(\frac{4cx}{n}\Bigr)^{\frac{1}{2-\alpha}}, \frac{Bx}{n}\biggr\}\,.$$

Then we have $\varepsilon = 2c_pB^{\frac{p}{2}}(c\varepsilon^\alpha+\delta)^{\frac{2-p}{4}}(\frac{a}{n})^{\frac12}$. Since ε^* is the largest solution of this equation we hence find $\varepsilon\leq\varepsilon^*$. This shows that we always have

$$\varepsilon \;\leq\; \varepsilon^* + 2c_p\biggl(B\Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}} + \sqrt{\frac{\delta x}{n}} + \Bigl(\frac{4cx}{n}\Bigr)^{\frac{1}{2-\alpha}} + \frac{Bx}{n}\biggr)\,.$$

Hence it suffices to bound ε^* from above. To this end let us first assume $c(\varepsilon^*)^\alpha \geq \delta$. This implies $\varepsilon^* \leq 4c_p B^{p/2}\bigl(c\cdot(\varepsilon^*)^\alpha\bigr)^{1/2-p/4}\bigl(\frac{a}{n}\bigr)^{1/2}$, and hence we find

$$\varepsilon^* \;\leq\; 16c_p^2 B^{\frac{2p}{4-2\alpha+\alpha p}} c^{\frac{2-p}{4-2\alpha+\alpha p}} \biggl(\frac{a}{n}\biggr)^{\frac{2}{4-2\alpha+\alpha p}}.$$

Conversely, if $c(\varepsilon^*)^\alpha < \delta$ holds then we immediately obtain

$$\varepsilon^* \;<\; 4c_p\, B^{\frac{p}{2}} \delta^{\frac{2-p}{4}} \biggl(\frac{a}{n}\biggr)^{\frac{1}{2}}.$$

■

6 Variance bounds for L1-SVMs

In this section we prove some "variance bounds" in the sense of Theorem 5.1 and Theorem 5.8 for L1-SVMs.

Let us first ensure that these classifiers are ERM-type algorithms that fit into the framework of Theorem 5.8. To this end let H be a RKHS of a continuous kernel over X, $\lambda > 0$, and $l : Y \times \mathbb{R} \to [0,\infty)$ be the hinge loss function. We define

$$L(f,x,y) \;:=\; \lambda\|f\|_H^2 + l(y,f(x)) \tag{41}$$

and

$$L(f,b,x,y) \;:=\; \lambda\|f\|_H^2 + l(y,f(x)+b) \tag{42}$$

for all $f \in H$, $b \in \mathbb{R}$, $x \in X$, and $y \in Y$. Then $\mathcal{R}_{L,T}(.)$ and $\mathcal{R}_{L,T}(.,.)$ obviously coincide with the objective functions of the L1-SVM formulations and hence we see that the L1-SVMs implement an empirical L-risk minimization. Furthermore note, that all above minimizers exist (see [33]) and thus the L1-SVM formulations in terms of L actually fit into the framework of Theorem 5.8.

The rest of this section is organized as follows: in the first subsection we establish a variance bound which holds for all distributions P on $X \times Y$. In the second subsection we will improve this variance bound for probability measures having some Tsybakov noise exponent $q > 0$.

6.1 Bounding the variance for L1-SVMs—the general case

Let us begin with stating the main result of this subsection which gives a "variance bound" for the class $\mathcal{G}$ defined in Theorem 5.1 for L1-SVMs without offset:

Proposition 6.1 *Let $0 < \lambda < 1$, H be a RKHS over X, and $\mathcal{F} \subset \lambda^{-\frac{1}{2}} B_H$. Furthermore, let L be defined by (41) and P be a probability measure. We write*

$$\mathcal{G} \;:=\; \bigl\{L\circ f - L\circ f_{P,\mathcal{F}} \;:\; f \in \mathcal{F}\bigr\}.$$

Then for all $g \in \mathcal{G}$ we have

$$\mathbb{E}_P g^2 \;\leq\; \frac{(4+2K)^2}{2\lambda}\mathbb{E}_P g\,.$$

Remark 6.2 Proposition 6.1 establishes a variance bound of the form $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ with $\alpha = 1$, $c = \frac{(4+2K)^2}{2\lambda}$, and $\delta = 0$. In particular, by substituting α by 1 and for $x \geq 1$ the term $\varepsilon(n,a,B,c,\delta,x)$ in Theorem 5.8 can be estimated by

$$\varepsilon(n,a,B,c,\delta,x) \;\preceq\; B^{\frac{2p}{2+p}} c^{\frac{2-p}{2+p}} \biggl(\frac{a}{n}\biggr)^{\frac{2}{2+p}} + xB\biggl(\frac{a}{n}\biggr)^{\frac{2}{2+p}} + \frac{cx}{n}\,. \tag{43}$$

Remark 6.3 Unfortunately, our techniques heavily rely on the strict convexity of the RKHS norm and hence it turns out that they can only be used for L1-SVMs *without* offset.

For the proof of the above proposition we need some notations. To this end let $\lambda > 0$, H be a RKHS over X, and $\mathcal{F} \subset \lambda^{-\frac{1}{2}} B_H$. Furthermore, we assume that l denotes—as usual—the hinge loss and L is defined by (41). We define the "metric"

$$d_{x,y}(f,g) \;:=\; 2\sqrt{\lambda}\|f-g\|_H + |f(x)-g(x)|$$

for all $(x,y) \in X \times Y$ and all $f,g \in \mathcal{F}$. Note that L is "point-wise Lipschitz continuous" with respect to $d_{x,y}$, i.e. we have

$$\big|L(f,x,y) - L(g,x,y)\big| \;\le\; d_{x,y}(f,g)$$

for all $(x,y) \in X \times Y$ and all $f,g \in \mathcal{F}$. Our ansatz is a modification of the idea presented in [4] which uses a modulus of convexity in order to quantify the convexity of the loss function. In our situation the strict convexity of L is due to the RKHS norm of the regularization term. This is reflected in the definition of $d_{x,y}$ as well as in the following definition: for $\varepsilon > 0$ the "modulus of convexity of L" is defined by

$$\delta(\varepsilon) \;:=\; \inf\left\{ \frac{L(f,x,y)+L(g,x,y)}{2} - L\left(\frac{f+g}{2},x,y\right) : (x,y) \in X\times Y, f,g \in \mathcal{F} \text{ with } d_{x,y}(f,g) \ge \varepsilon \right\}.$$

Since L is convex in f it is easy to see that $\delta(\varepsilon) \ge 0$ for all $\varepsilon > 0$. In the next lemma we establish a stronger lower estimate of $\delta(.)$.

Lemma 6.4 *Let $0 < \lambda < 1$ and $\varepsilon > 0$. Then with the above notation we have*

$$\delta(\varepsilon) \;\ge\; \frac{\lambda\varepsilon^2}{(4+2K)^2}\,.$$

Proof: Let $x \in X$, $y \in Y$ and $f,g \in \mathcal{F}$ with $d_{x,y}(f,g) \ge \varepsilon$. Then we find $\varepsilon \le 2\sqrt{\lambda}\|f-g\|_H + |f(x)-g(x)| \le (2+K)\|f-g\|_H$. Since l is convex and the norm $\|.\|$ of the RKHS satisfies the parallelogram law we then have

$$\begin{aligned}
& \frac{L(f,x,y)+L(g,x,y)}{2} - L\left(\frac{f+g}{2},x,y\right) \\
= \;& \lambda\frac{\|f\|^2+\|g\|^2}{2} - \lambda\left\|\frac{f+g}{2}\right\|^2 + \frac{l(y,f(x))+l(y,g(x))}{2} - l\left(y,\frac{f(x)+g(x)}{2}\right) \\
\ge \;& \lambda\left\|\frac{f-g}{2}\right\|^2 \\
\ge \;& \frac{\lambda\varepsilon^2}{(4+2K)^2}\,.
\end{aligned}$$

■

Let us now define a "modulus of continuity" for the L-risk $f \mapsto \mathcal{R}_{L,P}(f)$. To this end we write $d_P(f,g) := \big(\mathbb{E}_{(x,y)\sim P} d_{x,y}^2(f,g)\big)^{1/2}$ for all $f,g \in \mathcal{F}$, and

$$\delta_P(\varepsilon) \;:=\; \inf\left\{ \frac{\mathcal{R}_{L,P}(f)+\mathcal{R}_{L,P}(g)}{2} - \mathcal{R}_{L,P}\left(\frac{f+g}{2}\right) : f,g \in \mathcal{F} \text{ with } d_P(f,g) \ge \varepsilon \right\}.$$

Again, it is easy to see that $\delta_P(\varepsilon) \ge 0$ for all $\varepsilon > 0$ by the convexity of L. The next lemma which is based on Lemma 6.4 significantly improves this lower bound on $\delta_P(\varepsilon)$.

Lemma 6.5 *Let $0 < \lambda < 1$, $\varepsilon > 0$, and P be a distribution on $X \times Y$. Then with the above notation we have*

$$\delta_P(\varepsilon) \;\geq\; \frac{\lambda\varepsilon^2}{(4+2K)^2}\,.$$

Proof: Let f and g with $d_P(f,g) \geq \varepsilon$. Then by Lemma 6.4 we find

$$\begin{aligned}
\frac{\mathcal{R}_{L,P}(f) + \mathcal{R}_{L,P}(g)}{2} - \mathcal{R}_{L,P}\Big(\frac{f+g}{2}\Big) &= \mathbb{E}_{(x,y)\sim P}\Big(\frac{L(f,x,y)+L(g,x,y)}{2} - L\Big(\frac{f+g}{2},x,y\Big)\Big)\\
&\geq \mathbb{E}_{(x,y)\sim P}\delta(d_{x,y}(f,g))\\
&\geq \frac{\lambda\varepsilon^2}{(4+2K)^2}\,.
\end{aligned}$$

■

Proof of Proposition 6.1: By the definition of the modulus of convexity δ_P, the definition of $f_{P,\mathcal{F}}$ and Lemma 6.5 we obtain

$$\begin{aligned}
\frac{\mathcal{R}_{L,P}(f) + \mathcal{R}_{L,P}(f_{P,\mathcal{F}})}{2} &\geq \mathcal{R}_{L,P}\Big(\frac{f+f_{P,\mathcal{F}}}{2}\Big) + \delta_P\big(d_P(f,f_{P,\mathcal{F}})\big)\\
&\geq \mathcal{R}_{L,P}(f_{P,\mathcal{F}}) + \delta_P\big(d_P(f,f_{P,\mathcal{F}})\big)\\
&\geq \mathcal{R}_{L,P}(f_{P,\mathcal{F}}) + \frac{\lambda d_P^2(f,f_{P,\mathcal{F}})}{(4+2K)^2}
\end{aligned}$$

for all $f \in \mathcal{F}$. For $g := L\circ f - L\circ f_{P,\mathcal{F}}$ we hence have

$$\mathbb{E}_P g \;=\; \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}(f_{P,\mathcal{F}}) \;\geq\; 2\frac{\lambda d_P^2(f,f_{P,\mathcal{F}})}{(4+2K)^2}\,.$$

Furthermore, since L is point-wise Lipschitz-continuous with respect to $d_{x,y}$ we find

$$\mathbb{E}_P g^2 = \mathbb{E}_P\big(L\circ f - L\circ f_{P,\mathcal{F}}\big)^2 \;\leq\; \mathbb{E}_{(x,y)\sim P} d_{x,y}^2(f,f_{P,\mathcal{F}})\big) \;=\; d_P^2(f,f_{P,\mathcal{F}})\,.$$

■

6.2 Bounding the variance for L1-SVMs—Tsybakov's noise condition

As we have seen in the previous subsection we always have a variance bound for the L1-SVM in the sense of Theorem 5.1. Besides the fact that this bound was only established for L1-SVMs without offset it appears to be sharp since it has the "optimal" values $\alpha = 1$ and $\delta = 0$. However, it turns out that if we want to show rates faster than $n^{-\frac{1}{2}}$ we need a variance bound which is less sensitive to the regularization parameter λ. In this subsection we will establish such bounds for underlying distributions P satisfying Tsybakov's noise assumption for some exponent $q > 0$. An additional benefit of the approach of this subsection is that it can also be used for L1-SVMs *with* offset. In fact besides slightly larger constants the results are the same.

As in the last subsection l denotes the hinge loss. If no confusion can arise, $f_{l,P}$ denotes a minimizer of $\mathcal{R}_{l,P}$. For the shape of these minimizers which depend on $\eta := P(y=1|\,.\,)$ we refer to [41] and [32]. We begin with a variance bound which can be used when considering the empirical l-risk minimizer:

Lemma 6.6 *Let P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 < q \leq \infty$. Then there exists a minimizer $f_{l,P}$ mapping into $[-1,1]$ such that for all bounded measurable functions $f : X \to \mathbb{R}$ we have*

$$\mathbb{E}_P\big(l \circ f - l \circ f_{l,P}\big)^2 \;\leq\; \big(\|(2\eta-1)^{-1}\|_{q,\infty} + 2\big)\,\big(\|f\|_\infty + 1\big)^{\frac{q+2}{q+1}}\,\Big(\mathbb{E}_P\big(l \circ f - l \circ f_{l,P}\big)\Big)^{\frac{q}{q+1}}\,.$$

Proof: Given a fixed $x \in X$ we write $p := P(1|x)$ and $t := f(x)$. Since Tsybakov's noise assumption implies $P_X(X_0) = 0$ we can restrict our considerations to $p \neq 1/2$. We will show

$$\begin{aligned} & p\big(l(1,t) - l(1,f_{l,P}(x))\big)^2 + (1-p)\big(l(-1,t) - l(-1,f_{l,P}(x))\big)^2 \\ \leq\; & \Big(|t| + \frac{2}{|2p-1|}\Big)\,\Big(p\big(l(1,t) - l(1,f_{l,P}(x))\big) + (1-p)\big(l(-1,t) - l(-1,f_{l,P}(x))\big)\Big)\,. \end{aligned} \tag{44}$$

Without loss of generality we may assume $p > 1/2$. Then we may set $f_{l,P}(x) := 1$ and thus we have $l(1, f_{l,P}(x)) = 0$ and $l(-1, f_{l,P}(x)) = 2$. Therefore (44) reduces to

$$p\,l^2(1,t) + (1-p)\big(l(-1,t) - 2\big)^2 \;\leq\; \Big(|t| + \frac{2}{2p-1}\Big)\Big(p\,l(1,t) + (1-p)\big(l(-1,t) - 2\big)\Big)\,. \tag{45}$$

Let us first consider the case $t \in [-1,1]$. Since we then have $l(1,t) = 1 - t$ and $l(-1,t) = 1 + t$ we find

$$p\,l^2(1,t) + (1-p)\big(l(-1,t) - 2\big)^2 \;=\; p(1-t)^2 + (1-p)(t-1)^2 \;=\; (1-t)^2$$

and

$$p\,l(1,t) + (1-p)\big(l(-1,t) - 2\big) \;=\; p(1-t) + (1-p)(t-1) \;=\; (2p-1)(1-t)\,.$$

Therefore, (45) reduces to

$$(1-t)^2 \;\leq\; \Big(|t| + \frac{2}{2p-1}\Big)(2p-1)(1-t)\,.$$

Obviously, the latter inequality is equivalent to $1 - t \leq (2p-1)|t| + 2$ which is always satisfied for $t \in [-1,1]$ and $p \geq 1/2$.

Now let us consider the case $t \leq -1$. Since we then have $l(1,t) = 1 - t$ and $l(-1,t) = 0$ we find

$$p\,l^2(1,t) + (1-p)\big(l(-1,t) - 2\big)^2 \;=\; p(1-t)^2 + 4(1-p)$$

and

$$p\,l(1,t) + (1-p)\big(l(-1,t) - 2\big) \;=\; p(1-t) - 2(1-p)\,.$$

Therefore, it suffices to show

$$p(1-t)^2 + 4(1-p) \;\leq\; \Big(-t + \frac{2}{2p-1}\Big)\big(p(1-t) + 2(p-1)\big)\,.$$

It is easy to check that this inequality is equivalent to

$$4 - 3p \;\leq\; -\frac{2p^2 - 3p + 2}{2p-1}t + \frac{6p-4}{2p-1}\,.$$

Since $\frac{6p-4}{2p-1} - 4 + 3p = \frac{6p^2-5p}{2p-1}$ we thus have to show

$$p^2(6-2t) - p(5-3t) - 2t \geq 0\,.$$

The left hand side is minimal if $p = \frac{5-3t}{12-4t}$. Therefore, we obtain

$$p^2(6-2t)-p(5-3t)-2t \;\geq\; \left(\frac{5-3t}{12-4t}\right)^2(6-2t)-\frac{(5-3t)^2}{12-4t}-2t \;=\; -\frac{(5-3t)^2}{24-8t}-2t \;=\; \frac{7t^2-18t-25}{24-8t}$$

and hence it suffices to show $7t^2 - 18t - 25 \geq 0$. However, the latter is true for all $t \leq -1$ since $t \mapsto 7t^2 - 18t - 25$ is decreasing on $(-\infty, -1]$.

Now, let us consider the third case $t > 1$. Since we then have $l(1,t) = 0$ and $l(-1,t) = 1+t$ we find

$$p\, l^2(1,t) + (1-p)\big(l(-1,t)-2\big)^2 \;=\; (1-p)(t-1)^2$$

and

$$p\, l(1,t) + (1-p)\big(l(-1,t)-2\big) \;=\; (1-p)(t-1)\,.$$

Therefore, it suffices to show

$$t-1 \;\leq\; t + \frac{2}{2p-1}\,.$$

Since this is always true we have proved (45). Furthermore, for $p < \frac{1}{2}$ the proof of (44) is completely analogous and therefore (44) holds.

Now, let us write $g(y,x) := l(y,f(x)) - l(y,f_{l,P}(x))$, $h_1(x) := \eta(x)g(1,x) + (1-\eta(x))g(-1,x)$, and $h_2(x) := \eta(x)g^2(1,x) + (1-\eta(x))g^2(-1,x)$. Then (44) yields $h_2(x) \leq \big(\|f\|_\infty + \frac{2}{|2\eta(x)-1|}\big)h_1(x)$ for all x with $\eta(x) \neq 1/2$. Hence for $t \geq 1$ we find

$$\begin{aligned}
\mathbb{E}_P g^2 &= \int\limits_{|2\eta-1|^{-1}<t} h_2\, dP_X + \int\limits_{t\leq|2\eta-1|^{-1}<\infty} h_2\, dP_X \\
&\leq \big(\|f\|_\infty + 2t\big) \int\limits_{|2\eta-1|^{-1}<t} h_1\, dP_X + \int\limits_{t\leq|2\eta-1|^{-1}<\infty} \big(\|f\|_\infty+1\big)^2\, dP_X \\
&\leq 2\big(\|f\|_\infty + t\big)\mathbb{E}_P g + \big(\|f\|_\infty+1\big)^2 P_X\Big(|2\eta-1|^{-1} \geq t\Big) \\
&\leq 2\,t\,\big(\|f\|_\infty+1\big)\mathbb{E}_P g + \big(\|f\|_\infty+1\big)^2 \|(2\eta-1)^{-1}\|_{q,\infty}\, t^{-q}\,.
\end{aligned}$$

Let us define t by $t^{q+1} := (\|f\|_\infty+1)(\mathbb{E}_P g)^{-1}$. Since $\mathbb{E}_P g \leq \|f\|_\infty + 1$ we have $t \geq 1$ and hence the above estimate yields the assertion. ∎

In the case of L1-SVMs with offset we also need the following lemma which bounds the size of the offset $\tilde{b}_{P,\lambda}$. This lemma has been proved in [15] for empirical distributions. Although its generalization to general probability measures is straight forward we include the proof for completeness.

Lemma 6.7 *Let P be a distribution on $X \times Y$ and $\lambda > 0$. Then for all possible pairs $(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda}) \in H \times \mathbb{R}$ we have*

$$|\tilde{b}_{P,\lambda}| \;\leq\; \|\tilde{f}_{P,\lambda}\|_\infty + 1\,.$$

Proof: If $P(y = y^*|x) = 1$ P_X-a.s. for some $y^* \in Y$ there is nothing to be proved since $\tilde{b}_{P,\lambda} = y^*$ by our assumption on L1-SVMs mentioned in Section 2. Now let us assume that $\tilde{b}_{P,\lambda} > \|\tilde{f}_{P,\lambda}\|_\infty + 1$ and that P is not degenerate in the above way. Then there exists a constant $\delta > 0$ such that $\tilde{b}_{P,\lambda} > \|\tilde{f}_{P,\lambda}\|_\infty + 1 + \delta$. This implies $\tilde{f}_{P,\lambda}(x) + \tilde{b}_{P,\lambda} > 1 + \delta$ for all $x \in X$. We define $b^*_{P,\lambda} := \tilde{b}_{P,\lambda} - \delta$. Obviously, we then find $l(1, \tilde{f}_{P,\lambda}(x) + \tilde{b}_{P,\lambda}) = 0 = l(1, \tilde{f}_{P,\lambda}(x) + b^*_{P,\lambda})$ and

$$l(1, \tilde{f}_{P,\lambda}(x) + \tilde{b}_{P,\lambda}) = 1 + \tilde{f}_{P,\lambda}(x) + \tilde{b}_{P,\lambda} = 1 + \tilde{f}_{P,\lambda}(x) + b^*_{P,\lambda} + \delta = l(-1, \tilde{f}_{P,\lambda}(x) + b^*_{P,\lambda}) + \delta$$

for all $x \in X$. Therefore we obtain $\mathcal{R}_{l,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) > \mathcal{R}_{l,P}(\tilde{f}_{P,\lambda} + b^*_{P,\lambda})$ by using the assumption on P. ∎

The proof of the above lemma can be easily generalized to a larger class of loss functions. In particular for the squared hinge loss function used in L2-SVMs Lemma 6.7 holds.

With the help of Lemma 6.6 we can now show a variance bound for the L1-SVM. For brevity's sake we only state and prove the result for L1-SVMs with offset. Therefore, the loss function L is defined as in (42). Considering the proof it is immediately clear that the following variance bound also holds for the L1-SVM without offset.

Proposition 6.8 *Let P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 < q \leq \infty$. Define $C := 16+8\|(2\eta-1)^{-1}\|_{q,\infty}$. Furthermore, let $\lambda > 0$ and $0 < \gamma \leq \lambda^{-1/2}$ such that $\tilde{f}_{P,\lambda} \in \gamma B_H$. Then for all $f \in \gamma B_H$ and all $b \in \mathbb{R}$ with $|b| \leq K\gamma + 1$ we have*

$$\begin{aligned}\mathbb{E}\big(L \circ (f,b) - L \circ (\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})\big)^2 \;&\leq\; 8C(K\gamma+1)^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(L \circ (f,b) - L \circ (\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})\big)\Big)^{\frac{q}{q+1}} \\ &\quad + 16C(K\gamma+1)^{\frac{q+2}{q+1}} a^{\frac{q}{q+1}}(\lambda)\,.\end{aligned}$$

Proof: Let us define $\hat{C} := K\gamma + 1$. By Lemma 6.7 we then see $|\tilde{b}_{P,\lambda}| \leq \hat{C}$. We fix $f+b$ and choose a minimizer $f_{l,P}$ according to Lemma 6.6. Using $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$ we first observe

$$\begin{aligned}&\mathbb{E}\big(L \circ (f,b) - L \circ (\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})\big)^2 \\ \leq\;& 2\mathbb{E}\big(\lambda\|f\|^2 - \lambda\|\tilde{f}_{P,\lambda}\|^2\big)^2 + 2\mathbb{E}\big(l \circ (f+b) - l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda})\big)^2 \\ \leq\;& 2\lambda^2\|f\|^4 + 2\lambda^2\|\tilde{f}_{P,\lambda}\|^4 + 2\mathbb{E}\big(l \circ (f+b) - l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda})\big)^2 \\ \leq\;& 4\mathbb{E}\big(l \circ (f+b) - l \circ f_{l,P}\big)^2 + 4\mathbb{E}\big(l \circ f_{l,P} - l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda})\big)^2 + 2\lambda^2\|f\|^4 + 2\lambda^2\|\tilde{f}_{P,\lambda}\|^4\,.\end{aligned}$$

By Lemma 6.6 and $a^p + b^p \leq 2(a+b)^p$ for all $a, b \geq 0$, $0 < p \leq 1$ we find

$$\begin{aligned}&\mathbb{E}\big(l \circ (f+b) - l \circ f_{l,P}\big)^2 + \mathbb{E}\big(l \circ f_{l,P} - l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda})\big)^2 \\ \leq\;& C\hat{C}^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(l \circ (f+b) - l \circ f_{l,P}\big) + \mathbb{E}\big(l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - l \circ f_{l,P}\big)\Big)^{\frac{q}{q+1}}\,.\end{aligned}$$

Since $\lambda^2\|f\|^4 \leq 1$ and $\lambda^2\|\tilde{f}_{P,\lambda}\|^4 \leq 1$ we hence obtain

$$\begin{aligned}&\mathbb{E}\big(L \circ (f,b) - L \circ (\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})\big)^2 \\ \leq\;& 4C\hat{C}^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(l \circ (f+b) - l \circ f_{l,P}\big) + \mathbb{E}\big(l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - l \circ f_{l,P}\big)\Big)^{\frac{q}{q+1}} + 2\lambda^2\|f\|^4 + 2\lambda^2\|\tilde{f}_{P,\lambda}\|^4 \\ \leq\;& 8C\hat{C}^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(l \circ (f+b) - l \circ f_{l,P}\big) + \mathbb{E}\big(l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - l \circ f_{l,P}\big) + \lambda^2\|f\|^4 + \lambda^2\|\tilde{f}_{P,\lambda}\|^4\Big)^{\frac{q}{q+1}} \\ \leq\;& 8C\hat{C}^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(l \circ (f+b) - l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda})\big) + 2\mathbb{E}\big(l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - l \circ f_{l,P}\big) + \lambda\|f\|^2 + \lambda\|\tilde{f}_{P,\lambda}\|^2\Big)^{\frac{q}{q+1}} \\ \leq\;& 8C\hat{C}^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(L \circ (f+b) - L \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda})\big) + 2\mathbb{E}\big(l \circ (\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - l \circ f_{l,P}\big) + 2\lambda\|\tilde{f}_{P,\lambda}\|^2\Big)^{\frac{q}{q+1}} \\ \leq\;& 8C\hat{C}^{\frac{q+2}{q+1}}\Big(\mathbb{E}\big(L \circ (f,b) - L \circ (\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})\big)\Big)^{\frac{q}{q+1}} + 16C\hat{C}^{\frac{q+2}{q+1}} a^{\frac{q}{q+1}}(\lambda)\,.\end{aligned}$$

∎

Remark 6.9 Proposition 6.8 establishes a variance bound of the form $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ with $\alpha = \frac{q}{q+1}$, $c = (128 + 64\|(2\eta-1)^{-1}\|_{q,\infty})B^{\frac{q+2}{q+1}}$, and $\delta = (256 + 128\|(2\eta-1)^{-1}\|_{q,\infty})B^{\frac{q+2}{q+1}} a^{\frac{q}{q+1}}(\lambda)$. Recall, that by substituting α by $\frac{q}{q+1}$ the term $\varepsilon := \varepsilon(n, a, B, c, \delta, x)$ in Theorem 5.8 can be estimated by

$$\varepsilon \;\preceq\; B^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \Big(\frac{a}{n}\Big)^{\frac{2(q+1)}{2q+pq+4}} + B^{\frac{p}{2}} \delta^{\frac{2-p}{4}} \Big(\frac{a}{n}\Big)^{\frac{1}{2}} + B\Big(\frac{a}{n}\Big)^{\frac{2}{2+p}} + \sqrt{\frac{\delta x}{n}} + \Big(\frac{cx}{n}\Big)^{\frac{q+1}{q+2}} + \frac{Bx}{n} \tag{46}$$

for $x \geq 1$. Of course, we can also replace c and δ by the above estimates. However, we will see in Section 7 that the above form is slightly easier to control.

7 Proof of Theorem 2.8

In this last section we prove our main result Theorem 2.8. Since the proof is rather complex we split it into 3 parts: in Subsection 7.1 we estimate some covering numbers related to L1-SVMs and Theorem 5.8. In Subsection 7.2 we then show that the trivial bound $f_{T,\lambda} \leq \lambda^{-1/2}$ can be significantly improved under the assumptions of Theorem 2.8. Finally, in Subsection 7.3 we prove Theorem 2.8.

7.1 Covering numbers related to SVMs

In this subsection we establish a simple lemma that estimates the covering numbers of the class $\mathcal{G}$ in Theorem 5.8 with the help of the covering numbers of B_H. For brevity's sake it only treats the case of L1-SVMs with offset. The other case can be shown completely analogously.

Lemma 7.1 *Let H be a RKHS over X, P be a probability measure on $X \times Y$, $\lambda > 0$, and L be defined by (42). Furthermore, let $1 \leq \gamma \leq \lambda^{-\frac{1}{2}}$, K be defined by (14), and*

$$\mathcal{F} := \{(f, b) \in H \times \mathbb{R} : \|f\|_H \leq \gamma \text{ and } |b| \leq \gamma K + 1\}\,.$$

Defining $B := 2\gamma K + 3$ and

$$\mathcal{G} \;:=\; \big\{L \circ (f, b) - L \circ (f_{P,\mathcal{F}}, b_{P,\mathcal{F}}) \;:\; (f, b) \in \mathcal{F}\big\}$$

then gives $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$, where $(f_{P,\mathcal{F}}, b_{P,\mathcal{F}})$ denotes a L-risk minimizer in $\mathcal{F}$. Assume that there are constants $a \geq 1$ and $0 < p < 2$ such that for all $\varepsilon > 0$ we have

$$\sup_{T \in Z^n} \log \mathcal{N}(B_H, \varepsilon, L_2(T_X)) \;\leq\; a\varepsilon^{-p}\,.$$

Then there exists a constant $c_p > 0$ depending only on p such that for all $\varepsilon > 0$ we have

$$\sup_{T \in Z^n} \log \mathcal{N}\big(B^{-1}\mathcal{G}, \varepsilon, L_2(T)\big) \;\leq\; c_p\, a\varepsilon^{-p}\,.$$

Proof: Let us write $\hat{\mathcal{G}} := \big\{L \circ (f, b) \;:\; (f, b) \in \mathcal{F}\big\}$ and $\mathcal{H} := \big\{l \circ (f + b) \;:\; (f, b) \in \mathcal{F}\big\}$. Furthermore, for brevity's sake we denote the set of all constant functions from X to $[a, b]$ by $[a, b]$. We then have

$$\mathcal{N}\big(B^{-1}\mathcal{G}, \varepsilon, L_2(T)\big) \;=\; \mathcal{N}\big(B^{-1}\hat{\mathcal{G}}, \varepsilon, L_2(T)\big) \;\leq\; \mathcal{N}\big([0, \lambda\gamma^2] + B^{-1}\mathcal{H}, \varepsilon, L_2(T)\big)$$

using the Lipschitz-continuity of the hinge loss function. By the sub-additivity of the log-covering numbers we hence find

$$\begin{aligned}
\log \mathcal{N}\big(B^{-1}\mathcal{G}, 3\varepsilon, L_2(T)\big) &\leq \log \mathcal{N}\big([0, \lambda\gamma^2], \varepsilon, \mathbb{R}\big) + \log \mathcal{N}\big(B^{-1}\mathcal{H}, 2\varepsilon, L_2(T)\big) \\
&\leq \log\Big(\frac{1}{\varepsilon} + 1\Big) + \log \mathcal{N}\big(B^{-1}(B \cdot B_H + [-B, B]), 2\varepsilon, L_2(T_X)\big) \\
&\leq 2\log\Big(\frac{2}{\varepsilon} + 1\Big) + \log \mathcal{N}\big(B_H, \varepsilon, L_2(T_X)\big)\,.
\end{aligned}$$

From this we easily deduce the assertion. ∎

7.2 Shrinking the size of the SVM minimizers

In this subsection we show that the trivial bound $\|f_{T,\lambda}\| \leq \lambda^{-1/2}$ can be significantly improved under the assumptions of Theorem 2.8. In view of Theorem 5.8 this improvement will have a substantial impact on rates of Theorem 2.8. In order to obtain a rather flexible result let us suppose that for all $0 < p < 2$ we can determine constants $c, \gamma > 0$ such that

$$\sup_{T \in Z^n} \log \mathcal{N}(B_{H_\sigma}, \varepsilon, L_2(T_X)) \;\leq\; c\sigma^{\gamma d}\varepsilon^{-p} \tag{47}$$

holds for all $\varepsilon > 0$, $\sigma \geq 1$. Recall, that by Theorem 2.1 we can actually choose $\gamma := (1 - \frac{p}{2})(1 + \delta)$ for all $\delta > 0$.

Lemma 7.2 *Let X be the closed unit ball of the Euclidean space $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 \leq q \leq \infty$ and geometric noise exponent $0 < \alpha < \infty$. Furthermore, let us assume that (47) is satisfied for some $0 < \gamma \leq 2$ and $0 < p < 2$. Given an $0 \leq \varsigma < \frac{1}{5}$ we define*

$$\lambda_n \;:=\; n^{-\frac{4(\alpha+1)(q+1)}{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)} \cdot \frac{1}{1-\varsigma}}$$

and

$$\sigma_n := \lambda_n^{-\frac{1}{(\alpha+1)d}}$$

Assume that for the L1-SVM without offset using the Gaussian RBF kernel with width σ_n there are constants $\frac{1}{2(\alpha+1)} + 4\varsigma < \rho \leq \frac{1}{2}$ and $C \geq 1$ such that

$$\Pr^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq Cx\lambda_n^{-\rho}\Big) \;\geq\; 1 - e^{-x}$$

for all $n \geq 1$ and all $x \geq 1$. Then there is another constant $\hat{C} \geq 1$ such that for $\hat{\rho} := \frac{1}{2}\big(\frac{1}{2(\alpha+1)} + 4\varsigma + \rho\big)$ and for all $n \geq 1$, $x \geq 1$ we have

$$\Pr^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq \hat{C}x\lambda_n^{-\hat{\rho}}\Big) \;\geq\; 1 - e^{-x}\,.$$

If $q > 0$ then the same result is true for L1-SVMs with offset.

Proof: We only prove the lemma for L1-SVMs without offset since the proof for L1-SVMs with offset is analogous. Now let $\hat{f}_{T,\lambda_n}$ be a minimizer of $\mathcal{R}_{L,T}$ on $Cx\lambda_n^{(\rho-1)/2}B_{H_{\sigma_n}}$, where L is defined by (41). By our assumption we have $\hat{f}_{T,\lambda_n} = f_{T,\lambda_n}$ with probability not less than $1 - e^{-x}$ since f_{T,λ_n} is unique for every training set T by the strict convexity of L. We show that for some constant $\tilde{C} > 0$ and all $n \geq 1$, $x \geq 1$ the improved bound

$$\|\hat{f}_{T,\lambda_n}\| \leq \tilde{C}x\lambda_n^{\frac{\hat{\rho}-1}{2}} \tag{48}$$

holds with probability not less than $1 - e^{-x}$. Consequently, $\|f_{T,\lambda_n}\| \leq \tilde{C}x\lambda_n^{(\hat{\rho}-1)/2}$ holds with probability not less than $1 - 2e^{-x}$. Obviously, the latter implies the assertion. In order to establish (48) we will apply Theorem 5.8 to the modified L1-SVM classifier which produces $\hat{f}_{T,\lambda_n}$. To this end we first remark that the infinite sample version $\hat{f}_{P,\lambda_n}$ which minimizes $\mathcal{R}_{L,P}$ on $Cx\lambda_n^{(\rho-1)/2}B_{H_{\sigma_n}}$ exists by a small modification of [33, Lem. 3.1].

Let us first treat the case $q > 0$. By Proposition 6.8 and assumption (47) we observe that we may choose B, a and c such that

$$\begin{aligned} B &\sim x\lambda_n^{-\rho} \\ a &\sim \lambda_n^{-\frac{\gamma}{\alpha+1}} \\ c &\sim x^{\frac{q+2}{q+1}}\lambda_n^{-\rho\cdot\frac{q+2}{q+1}} . \end{aligned}$$

Furthermore, Theorem 2.7 shows $a_{\sigma_n}(\lambda_n) \preceq \lambda_n^{\frac{\alpha}{\alpha+1}}$ and thus by Proposition 6.8 we may choose

$$\delta \sim x^{\frac{q+2}{q+1}}\lambda_n^{\frac{\alpha q-\rho(q+2)(\alpha+1)}{(\alpha+1)(q+1)}} .$$

Now Remark 6.9 and a rather time-consuming but simple calculation shows that

$$\varepsilon(n,a,B,c,\delta,x) \preceq x^2\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} .$$

By Theorem 5.8 there is therefore a constant $\tilde{C}_1 > 0$ independent of n and x such that for all $n \geq 1$ and all $x \geq 1$ the estimate

$$\begin{aligned} \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 &\leq \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{T,\lambda_n}) - \mathcal{R}_{l,P} \\ &\leq \lambda_n\|\hat{f}_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1x^2\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} \end{aligned}$$

holds with probability not less than $1-e^{-x}$. Now, $\lambda\|f_{P,\lambda}\|^2 \leq a_{\sigma_n}(\lambda_n) \preceq \lambda_n^{\frac{\alpha}{\alpha+1}}$ yields $\|f_{P,\lambda_n}\| \preceq \lambda_n^{-\frac{1}{2(\alpha+1)}}$ and hence $\rho > \frac{1}{2(\alpha+1)}$ implies $\|f_{P,\lambda_n}\| \leq \lambda_n^{-\rho} \leq Cx\lambda_n^{-\rho}$ for large n. In other words, for large n we have $f_{P,\lambda_n} = \hat{f}_{P,\lambda_n}$. Consequently, with probability not less than $1-e^{-x}$ we have

$$\begin{aligned} \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 &\leq \lambda_n\|f_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1x^2\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} \\ &\leq \tilde{C}_2\lambda_n^{\frac{\alpha}{\alpha+1}} + \tilde{C}_1x^2\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-4\varsigma} , \end{aligned}$$

which shows the assertion in the case $q > 0$.

Let us now prove the assertion for $q = 0$. By Proposition 6.1 and assumption (47) we observe that we may choose B, a and c such that

$$\begin{aligned} B &\sim x\lambda_n^{-\rho} \\ a &\sim \lambda_n^{-\frac{\gamma}{\alpha+1}} \\ c &\sim \lambda_n^{-1} , \end{aligned}$$

and thus Remark 6.2 and a simple calculation gives us

$$\varepsilon(n,a,B,c,\delta,x) \preceq x^2\lambda_n^{\frac{2\alpha-2\alpha p\rho-2p\rho+\alpha p+p}{(2+p)(\alpha+1)}-4\varsigma} .$$

By Theorem 5.8 there is therefore a constant $\tilde{C}_1 > 0$ independent of n and x such that for all $n \geq 1$ and all $x \geq 1$ the estimate

$$\begin{aligned} \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 &\leq \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{T,\lambda_n}) - \mathcal{R}_{l,P} \\ &\leq \lambda_n\|\hat{f}_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1x^2\lambda_n^{\frac{2\alpha-2\alpha p\rho-2p\rho+\alpha p+p}{(2+p)(\alpha+1)}-4\varsigma} \end{aligned}$$

holds with probability not less than $1-e^{-x}$. As in the case $q>0$ we find $f_{P,\lambda_n} = \hat{f}_{P,\lambda_n}$ for all large n. With probability not less than $1-e^{-x}$ this gives

$$\begin{aligned}\lambda_n \|\hat{f}_{T,\lambda_n}\|^2 &\leq \lambda_n \|f_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1 x^2 \lambda_n^{\frac{2\alpha-2\alpha p\rho-2p\rho+\alpha p+p}{(2+p)(\alpha+1)}-4\varsigma} \\ &\leq \tilde{C}_2 \lambda_n^{\frac{\alpha}{\alpha+1}} + \tilde{C}_1 x^2 \lambda_n^{\frac{2\alpha-2\alpha\rho-2\rho+1}{2(\alpha+1)}-4\varsigma} \\ &= \tilde{C}_2 \lambda_n^{\frac{\alpha}{\alpha+1}} + \tilde{C}_1 x^2 \lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-4\varsigma},\end{aligned}$$

where we used $\rho > \frac{1}{2(\alpha+1)}$ and $p<2$. From this we obtain the assertion for $q=0$. ∎

7.3 Proof of Theorem 2.8

The next theorem establishes almost the result of Theorem 2.8. We present this intermediate result because it clarifies the impact of covering number bounds of the form (47) on our rates.

Theorem 7.3 *Let X be the closed unit ball of the Euclidean space $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 \leq q \leq \infty$ and geometric noise exponent $0 < \alpha < \infty$. Finally, let us assume that we can bound the covering numbers by (47) for some $0 < \gamma \leq 2$ and $0 < p < 2$. Given an $0 \leq \varsigma < \frac{1}{5}$ we define*

$$\lambda_n := n^{-\frac{4(\alpha+1)(q+1)}{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}\cdot\frac{1}{1-\varsigma}}$$

and

$$\sigma_n := \lambda_n^{-\frac{1}{(\alpha+1)d}}$$

Then for all $\varepsilon > 0$ there is a constant $C > 0$ such that for all $x \geq 1$ and all $n \geq 1$ the L1-SVM without offset and with regularization parameter λ_n and Gaussian RBF kernel with width σ_n satisfies

$$\Pr^*\Big(T \in (X\times Y)^n : \mathcal{R}_P(f_{T,\lambda_n}) \leq \mathcal{R}_P + Cx^2 n^{-\frac{4\alpha(q+1)}{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}\cdot\frac{1}{1-\varsigma}+20\varsigma+\varepsilon}\Big) \;\geq\; 1-e^{-x}\,.$$

If $q>0$ then the same result is true for L1-SVMs with offset.

Proof: Iteratively using Lemma 7.2 we find a constant $C \geq 1$ such that for $\rho := \frac{1}{2(\alpha+1)} + 4\varsigma + \varepsilon$ and all $n \geq 1$, $x \geq 1$ we have

$$\Pr^*\Big(T \in (X\times Y)^n : \|f_{T,\lambda_n}\| \leq Cx\lambda_n^{-\rho}\Big) \;\geq\; 1-e^{-x}\,.$$

Repeating the calculations of Lemma 7.2 (distinguish between the cases $q>0$ and $q=0$) we hence find a constant $\tilde{C} > 0$ such that for all $n \geq 1$ and all $x \geq 1$ we have

$$\lambda_n \|f_{T,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda_n}) - \mathcal{R}_{l,P} \;\leq\; \lambda_n \|f_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1 x^2 \lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-4\varsigma}$$

with probability not less than $1-e^{-x}$. By the definition of ρ we obtain

$$\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-4\varsigma} \leq \lambda_n^{\frac{\alpha}{\alpha+1}-4\varsigma-\varepsilon-4\varsigma} \leq n^{-\frac{4\alpha(q+1)}{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}\cdot\frac{1}{1-\varsigma}+20\varsigma+3\varepsilon}\,.$$

From this we easily deduce the assertion. ∎

In order to prove Theorem 2.8 recall that by Theorem 2.1 we can choose $\gamma := (1-\frac{p}{2})(1+\delta)$ for all $\delta > 0$. The idea of the proof of Theorem 2.8 is to let $\delta \to 0$ while simultaneously adjusting ς. The resulting rate is then optimized with respect to p. Unfortunately, a rigorous proof requires to choose p a-priori. Therefore, the optimization step is somewhat hidden in the following proof:

Proof of Theorem 2.8: Let us first consider the case $\alpha \leq \frac{q+2}{2q}$. Our aim is to apply Theorem 7.3. To this end we write $p_\delta := 2-\delta$ and $\gamma_\delta := (1-\frac{p_\delta}{2})(1+\delta) = \frac{\delta}{2}(1+\delta)$ for $\delta > 0$. Furthermore, we define ς_δ by

$$\frac{4(\alpha+1)(q+1)}{(2\alpha+1)(4q-\delta q+4)+4\gamma_\delta(q+1)} \cdot \frac{1}{1-\varsigma_\delta} \;=\; \frac{\alpha+1}{2\alpha+1}\,.$$

Since $2\alpha q - q - 2 \leq 0 < 2\delta(q+1)$ we have $q(2\alpha+1) < 2(1+\delta)(q+1)$ and hence

$$4(2\alpha+1)(q+1) < 4(2\alpha+1)(q+1) - \delta q(2\alpha+1) + 2\delta(1+\delta)(q+1)\,.$$

This shows $\varsigma_\delta > 0$ for all $\delta > 0$. Furthermore, these definitions also imply $\varsigma_\delta \to 0$ and $\gamma_\delta \to 0$ whenever $\delta \to 0$. Now, Theorem 7.3 tells us that for all $\varepsilon > 0$ and all small enough $\delta > 0$ there exists a constant $C_{\delta,\varepsilon} \geq 1$ such that for all $n \geq 1$, $x \geq 1$ we have

$$\Pr^*\Bigl(T \in (X\times Y)^n : \mathcal{R}_P(f_{T,\lambda_n}) \leq \mathcal{R}_P + C_{\delta,\varepsilon}x^2 n^{-\frac{4\alpha(q+1)}{(2\alpha+1)(4q-\delta q+4)+4\gamma_\delta(q+1)}\cdot\frac{1}{1-\varsigma_\delta}+20\varsigma_\delta+\varepsilon}\Bigr) \;\geq\; 1-e^{-x}\,.$$

In particular, if we choose δ sufficiently small we find the assertion.

Let us now consider the case $\frac{q+2}{2q} < \alpha < \infty$. In this case we write $p_\delta := \delta$ and $\gamma_\delta := (1-\frac{p_\delta}{2})(1+\delta) = 1 + \frac{\delta}{2} - \frac{\delta^2}{2}$ for $\delta > 0$. Furthermore, we define ς_δ by

$$\frac{4(\alpha+1)(q+1)}{(2\alpha+1)(2q+\delta q+4)+4\gamma_\delta(q+1)} \cdot \frac{1}{1-\varsigma_\delta} \;=\; \frac{2(\alpha+1)(q+1)}{2\alpha(q+2)+3q+4}\,.$$

Since for $0 < \delta \leq 1$ we have $0 < \delta q(2\alpha+1) + 2\delta(q+1) - 2\delta^2(q+1)$ we easily check $\varsigma_\delta > 0$. Furthermore, the definitions ensure $\varsigma_\delta \to 0$ and $\gamma_\delta \to 1$ whenever $\delta \to 0$. The rest of the proof follows that of the first case.

Finally, let us treat the case $\alpha = \infty$. We define α_λ by $\log\lambda = \alpha_\lambda d \log\frac{2\sqrt{d}}{\sigma}$. Since $\sigma > 2\sqrt{d}$ we have $\alpha_\lambda > 0$ for all $0 < \lambda < 1$. Furthermore, applying Theorem 2.7 for α_λ we find $a(\lambda) \leq 2C_d\lambda$ for all $0 < \lambda < 1$ and a constant $C_d > 0$ depending only on the dimension d. Adapted versions of Lemma 7.2 and Theorem 7.3 then yield the assertion. ∎

References

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

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[3] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, to appear, 2005.

[4] 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.

[5] C. Bennett and R. Sharpley. *Interpolation of Operators.* Academic Press, Boston, 1988.

[6] C. Berg, J.P.R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, New York, 1984.

[7] J. Bergh and J. Löfström. *Interpolation Spaces, An Introduction*. Springer-Verlag, New York, 1976.

[8] M. Sh. Birman and M. Z. Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α (russian). *Mat. Sb.*, 73:331–355, 1967.

[9] O. Bousquet. A Bennet concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334:495–500, 2002.

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

[11] R. Courant and D. Hilbert. *Methods of Mathematical Physics*. Interscience Publishers, New York, first english edition, 1953.

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

[13] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49, 2002.

[14] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[15] J. Howse, D. Hush, and C. Scovel. Linking learning strategies and performance for support vector machines. `http://www.c3.lanl.gov/ml/pubs_select.shtml`, 2002.

[16] D. Hush, C. Scovel, and I. Steinwart. Stability of unstable learning algorithms. `http://www.c3.lanl.gov/~ingo/publications/ml-03.ps`, 2003.

[17] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C. R. Math. Acad. Sci. Paris*, 334:501–504, 2002.

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

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

[20] P. Massart. About the constants in Talagrand's concentration inequality for empirical processes. *Ann. Probab.*, 28:863–884, 2000.

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

[22] R. O'Neil. Convolution operators and L(p,q) spaces. *Duke Math. J.*, 30:129–142, 1963.

[23] A. Pietsch. *Operator Ideals*. North-Holland, Amsterdam, 1980.

[24] A. Pietsch. *Eigenvalues and s-Numbers*. Geest & Portig K.-G., Leipzig, 1987.

[25] M. R. Reed and B. Simon. *Methods of Modern Mathematical Physics, v.1*. Academic Press, New York, 1972.

[26] M. R. Reed and B. Simon. *Methods of Modern Mathematical Physics, v.4.* Academic Press, New York, 1972.

[27] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Related Fields*, 119:163–175, 2001.

[28] B. Schölkopf and A.J. Smola. *Learning with Kernels.* MIT Press, 2002.

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

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

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

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

[33] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inform. Theory*, 51:128–142, 2005.

[34] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert space of Gaussian RBF kernels. Technical report, Los Alamos National Laboratory, 2004.

[35] M. Talagrand. Sharper bounds for gaussian and empirical processes. *Ann. Probab.*, 22:28–76, 1994.

[36] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators.* North Holland, Amsterdam, 1978.

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

[38] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer, New York, 1996.

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

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

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