Fast Rates for Support Vector Machines*

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

December 24, 2003

Abstract

We establish learning rates to the Bayes risk for support vector machines with hinge loss (L1-SVM's). Since a theorem of Devroye states that no learning algorithm can learn with a uniform rate to the Bayes risk for *all* probability distributions we have to restrict the class of considered distributions: in order to obtain fast rates we assume a noise condition recently proposed by Tsybakov and an approximation condition in terms of the distribution and the reproducing kernel Hilbert space used by the L1-SVM. For Gaussian RBF kernels with varying widths we propose a *geometric* noise assumption on the distribution which ensures the approximation condition. This geometric assumption is not in terms of smoothness but describes the concentration of the marginal distribution near the decision boundary. In particular we are able to describe nontrivial classes of distributions for which L1-SVM's using a Gaussian kernel can learn with almost linear rate.

We use various new and recently introduced techniques for establishing our results: the analysis of the estimation error is based on Talagrand's concentration inequality and local Rademacher averages. We furthermore develope a shrinking technique which allows us to control the typical size of the norm of the L1-SVM solution. It turns out that the above mentioned approximation assumption has a crucial impact on both the application of Talagrand's inequality and the shrinking technique. Moreover, for Gaussian kernels we develope a smoothing technique which allows us to treat the approximation error in a way directly linked to the classification problem. Finally, we prove some new bounds on covering numbers related to Gaussian RBF kernels.

1 Introduction

In recent years support vector machines (SVM's) have been the subject of many theoretical considerations. However, their learning *performance* on restricted classes of distributions is widely unknown. In particular, it is unknown under which circumstances SVM's can guarantee *fast* rates with respect to the sample size n for their learning performance. In recent years two concepts have revolutionized the learning theory community: Tsybakov's noise exponent for distributions which gives a sufficient condition for certain theoretical classifiers to learn with a rate faster than $n^{-\frac{1}{2}}$, and local Rademacher averages as a powerful new tool for bounding the estimation error of empirical risk minimization (ERM)-like algorithms. The aim of this paper is to apply these concepts to SVM's in order to obtain fast rates on their learning performance. Unlike many other works we

*AMS 2000 subject classification: primary 68Q32, secondary 62G20, 62G99, 68T05, 68T10, 41A46, 41A99 also address the *approximation error* by introducing a *geometric noise* condition for distributions. In particular we are able to describe distributions such that SVM's with Gaussian kernel learn almost linearly, i.e. with rate $n^{-1+\varepsilon}$ for all $\varepsilon > 0$, even though the Bayes classifier is *not* in the corresponding reproducing kernel Hilbert space (RKHS).

Let us formally introduce the statistical classification problem. To this end assume for technical reasons that X is a compact metric space. 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 T is 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 $f_T(x)$. 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). Here, misclassification means $\operatorname{sign} f_T(x) \neq y$ where we choose a fixed definition of $\operatorname{sign}(0) \in \{-1, 1\}$. To make this precise the risk of a measurable function $f : X \to \mathbb{R}$ is defined by

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

The smallest achievable risk $\mathcal{R}_P := \inf\big\{\mathcal{R}_P(f) \mid f : X \to \mathbb{R} \text{ measurable}\big\}$ is called the *Bayes risk* of P. A function 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 $\mathcal{R}(f_T)$ is bounded between $\mathcal{R}_P$ and 1 the convergence in (1) holds if and only if

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

The next naturally arising question is whether there are classifiers which guarantee a specific rate of convergence in (1) or (2) for *all* distributions. Unfortunately, this is impossible by a result of Devroye (see [13, Thm. 7.2]). However, if one restricts considerations to certain smaller classes of distributions such rates exist for various classifiers, e.g.:

- Assuming that the conditional probability $\eta(x) := P(1|x)$ satisfies certain smoothness assumptions Yang showed in [35] that some plug-in rules achieve rates for (2) which are of the form $n^{-\alpha}$ for some $0 < \alpha < 1/2$ depending on the assumed smoothness. He also showed that these rates are optimal in the sense that no classifier can obtain faster rates under the proposed smoothness assumptions.

- Recently, for SVM's with hinge loss (L1-SVM's) Wu and Zhou [34] established rates for (1) under the assumption that η is contained in a Sobolev space. In particular, he proved rates of the form $(\log n)^{-p}$ for some $p > 0$ if the L1-SVM uses a Gaussian kernel.

- It is well know (see [13, Thm. 18.3]) that using structural risk minimization over a sequence of hypothesis classes with finite VC-dimension every distribution which has a Bayes decision function in one of the hypothesis classes can be learned with rate $\sqrt{\frac{\log n}{n}}$.

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

Restricting the class of distributions always raises up the question whether it is likely that these restrictions are met in real world problems. Of course, the assumption of a noisefree distribution is almost never satisfied in practice. Furthermore, assuming that the conditional probability is smooth, say k-times continuously differentiable, seems to be very unlikely in real world classification problems. Therefore, the above listed rates are established for situations which are rarely met in practice.

Considering the ERM classifier and hypothesis classes $\mathcal{F}$ containing a Bayes decision function there is a large gap in the rates for noise-free and noisy distributions. In [32] Tsybakov closed this gap: he showed that certain ERM-type classifiers can obtain rates in (2) which are of the form $n^{-\frac{q+1}{q+pq+2}}$, where $0 \leq q \leq \infty$ is a parameter describing how well the noise is distributed (see Definition 2.1 in the following section) and $0 < p < 1$ measures the complexity of the hypothesis class. Unlike the above mentioned restrictions on the class of distributions Tsybakov's condition on the noise seems to be reasonable for many real world situations since it does not impose any kind of smoothness. Furthermore, Tsybakov showed that for specific types of distributions having classes with "smooth" boundaries the above rates are optimal in the sense that there is no classifier that has uniformly faster rates for these types of probability measures. Unfortunately, the ERM-classifier he considered is usually hard to implement and in general there exists no efficient algorithm. Furthermore, his classifier requires substantial knowledge on *how* to approximate the Bayes decision rules of the considered distributions. Of course, such knowledge is rarely present in practice.

In this work we will establish rates in (1) for SVM's and distributions satisfying Tsybakov's noise condition for some $0 \leq q \leq \infty$. Furthermore, these rates also incorporate the approximation properties of the used RKHS. Namely, we will show in Theorem 2.4 that the L1-SVM can learn with rate

$$n^{-\frac{4\beta(q+1)}{(2q+pq+4)(1+\beta)}+\varepsilon} \tag{3}$$

for all $\varepsilon > 0$ provided that the regularization sequence (λ_n) is suitably chosen. Here $0 < p < 2$ is the *complexity exponent* of the RKHS H (see Definition 2.3) which differs from Tsybakov's complexity measure. Furthermore, $0 < \beta \leq 1$ is the *approximation exponent* of H and P (see Definition 2.2) which describes how well H can approximate P with respect to the hinge loss. In the best case $\beta = 1$ which describes RKHS's containing a Bayes classifier the rate (3) is essentially equal to $n^{-\frac{2(q+1)}{2q+pq+4}}$. Furthermore, if the RKHS consists of C^∞ functions we may choose p arbitrarily close to 0. In this case our rate is essentially of the form $n^{-\frac{q+1}{q+2}}$. In particular, these considerations hold for the Gaussian RBF kernels. However, in this case the assumption $\beta = 1$ can essentially only hold for distributions which satisfy $\eta(x) \in \{0, 1/2, 1\}$ P_X-a.s. and have non-touching classes. Of course, these assumptions are rarely met in practice.

To overcome this problem for Gaussian kernels we treat the width $\sigma > 0$ of the kernel as a second regularization parameter which changes with the sample size. We then introduce a geometric noise condition which allows us to describe nontrivial classes of distributions which can be well-approximated by Gaussian kernels with changing widths. One amazing aspect of these approximation rates is the fact that Gaussian kernels poorly approximate smooth functions (cf. [27]) and hence plug-in rules based on Gaussian kernels may have a bad performance under smoothness assumptions on η. In particular, many types of SVM's including L2-SVM's and LS-SVM's are plug-in rules and therefore, their approximation properties under smoothness assumptions on η may be poor if a Gaussian kernel is used. However, L1-SVM's are not plug-in rules since their decision functions approximate the Bayes decision function (see [29]). Intuitively, we therefore only need a condition that measures the cost of approximating the "bump" of the Bayes decision function at the "decision boundary". We propose such a (*geometric noise*) condition parameterized by $0 \leq \alpha \leq \infty$ which does not measure any smoothness but *describes* how the noise and the marginal distribution is distributed near the "boundary". Every probability measure satisfies this assumption for $\alpha = 0$, and for $\alpha = \infty$ the condition describes distributions which have classes that are extremely concentrated on sets with strictly positive distance. For other α the condition describes intermediate assumptions.

Assuming such a geometric noise condition with parameter $0 < \alpha \leq \infty$ and using a covering number bound established in Theorem 2.15 we establish rates for the L1-SVM in Theorem 10.2 which are of the form

$$n^{-\frac{4\alpha(q+1)}{(2\alpha+1)(2q+pq+4)+2(2-p)(q+1)}+\varepsilon} \tag{4}$$

for all $\varepsilon > 0$ if both (λ_n) and (σ_n) are suitably chosen. In these rates $0 < p < 2$ is a *free* parameter. In particular, for $\alpha \leq \frac{q+2}{2q}$ we should choose p close to 2 in order to optimize this rate. This yields rates of the form $n^{-\frac{\alpha}{2\alpha+1}+\varepsilon}$ for all $\varepsilon > 0$. In the other case $\alpha > \frac{q+2}{2q}$ the parameter p should be close to 0. Then our rate becomes $n^{-\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}+\varepsilon}$.

In order to prove our rates we need various techniques: we use Zhang's [36] inequality (see equation 15) to bound excess classification risk in terms of excess hinge loss risk. The approximation rates are obtained by smoothing the Bayes decision function with the integral operator of the Gaussian RBF kernel—the Gauss-Weierstrass heat operator. This approximation result is unlike any we have found in the approximation theory literature. Indeed, much is known about the approximation properties of the Gauss-Weierstrass operator but these results are with respect to the continuous function norm or L_p spaces with Lebesgue measure (see e.g. [9]). In our situation we do not need to approximate whole classes of functions; we only need to approximate the Bayes decision function with respect to the X-marginal measure for which we assume a geometric noise condition. Although the results in Hush et al. [15] can be used to bound the estimation error, these rates would be at best $n^{-\frac{1}{2}}$. Consequently, as mentioned at the beginning, the estimation error is instead treated with concentration inequalities involving local Rademacher averages. This technique requires a certain "variance bound" which in the case of the L1-SVM depends on Tsybakov's noise exponent q. However, unlike for standard ERM-algorithms our variance bound also depends on the above mentioned *approximation exponent.* As a result, estimation and approximation error are intimately interwoven in our bounds and thus no classical decomposition of the learning rate into an estimation and an approximation part is possible. The situation becomes even more complicated by another observation: Usually, we can only assume that the objective function of the SVM optimization problem is minimized over the closed ball $\frac{1}{\sqrt{\lambda}}B_H$ of the RKHS. However, it turns out that assuming a nontrivial approximation exponent α the radius $\frac{1}{\sqrt{\lambda}}$ can be essentially replaced by $\lambda^{-\alpha}$ for some $\alpha < \frac{1}{2}$. Since this radius has a crucial impact on the estimation error the latter depends on the approximation exponent because of two independent reasons. Finally, since we use the Gaussian kernel width σ as a regularization parameter we require bounds on the covering numbers of the RKHS in terms of σ. As a consequence, the proofs of our results are rather technical.

The rest of this work is organized as follows: In Section 2 we define the approximation exponent for RKHS's, and introduce the noise concepts for distributions. We then present some examples of classes of distributions which are met by these concepts, and state our main results. Furthermore, here we establish notation. In Section 3 we consider some structural properties of the introduced approximation exponent. General bounds for ERM-type classifiers involving local Rademacher averages are established in Section 4. In the following section we prove "variance bounds" for L1-SVM's which depend on both Tsybakov's noise exponent and the approximation exponent. Local Rademacher averages for RKHS's are bounded in Section 5 using certain covering number bounds. These are used to reformulate our ERM-type classifier result of Section 4. In the following section we prove the rate (3) for general L1-SVM's using an iterative shrinking technique for the typical size of the norm of the L1-SVM decision function. The remaining parts of the work are devoted to L1-SVM's with Gaussian kernels. In Section 8 we prove approximation rates for Gaussian RKHS's and distributions that satisfy our geometric noise condition. Lorentz norms of covering numbers of Gaussian RKHS in terms of the Gaussian width σ are shown in Section 9. In the following section we prove the rates (4) by the above mentioned shrinking technique.

2 Definitions and Results

For two functions a and b we use the notation $a(\lambda) \preceq b(\lambda)$ to mean that there exists a constant $C > 0$ such that $a(\lambda) \leq Cb(\lambda)$ 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 addition we use the same notation for sequences.

Given a probability measure P on $X \times Y$ with conditional probability $\eta(x) := P(1|x)$, $x \in X$ 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 $X_0 := \{x \in X : \eta(x) = \frac{1}{2}\}$. It is easy to see that the behaviour of a function $f : X \to \mathbb{R}$ on X_0 has no influence on its risk $\mathcal{R}_P(f)$. Therefore, it is sometimes convenient to consider the restriction $\hat{P}_X$ of P_X onto $X_{-1} \cup X_1$. We sometimes use the notation $\Pr^*$ for outer measures to avoid measurability considerations.

As already mentioned in the introduction Tsybakov's noise exponent enables us to obtain fast classification rates. Let us recall its definition, which can be expressed in terms of Lorentz spaces $L_{q,\infty}$ (see e.g. [5] for these spaces):

Definition 2.1 Let $0 \leq q \leq \infty$ and P be a probability measure on $X \times Y$. We say that P *has Tsybakov noise exponent* q if $(2\eta - 1)^{-1} \in L_{q,\infty}(\hat{P}_X)$, i.e. there exists a constant $C > 0$ such that

$$P_X\big(0 < |2\eta - 1| \leq t\big) \ \leq \ C \cdot t^q \tag{5}$$

for all $t > 0$.

All distributions have at least noise exponent 0. In the other extreme case $q = \infty$ the conditional probability η is bounded away from $\frac{1}{2}$ on $X_{-1} \cup X_1$. Note that Tsybakov's noise condition does not require $P_X(X_0) = 0$.

The second important concept describes how well a given RKHS H can approximate a distribution P. Since this quantity is closely related to the definition of L1-SVM's we first recall the latter. To this end let $l(y,t) := \max\{0, 1 - yt\}$, $y \in Y$, $t \in \mathbb{R}$ be the hinge loss function. For a given distribution P on $X \times Y$ and a function $f : X \to \mathbb{R}$ the *l-risk* of f is defined by $\mathcal{R}_{l,P}(f) := \mathbb{E}_{(x,y)\sim P}\, l(y, f(x))$. For $\lambda > 0$ we denote a minimizer

$$(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda}) \in \arg \min_{(f,b)\in H\times\mathbb{R}} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{l,P}(f + b)\Big)\,. \tag{6}$$

If P is an empirical distribution with respect to a training set T we write $\mathcal{R}_{l,T}(f)$ and $(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda})$. In order to be able to control the size of the offset we always assume that we choose $b_{P,\lambda} := y^*$ if $P_X\big(x \in X : P(y^*|x) = 1\big) = 1$ for some $y^* \in Y$. Note that for empirical distributions based on T the latter condition means that all labels of T are equal to y^*. An algorithm that solves (6) with an empirical distribution is called *L1-SVM with offset*. Analogously, without the offset we denote a minimizer

$$f_{P,\lambda} \in \arg \min_{f\in H} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{l,P}(f)\Big)\,. \tag{7}$$

For empirical distributions we again write $f_{T,\lambda}$. An algorithm that solves (7) with an empirical distribution is called *L1-SVM without offset*. We emphasize that in many theoretical papers only L1-SVM's 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-SVM's with offset (for an analysis on partially stable learning algorithms including L1-SVM's with offset which resolves many of these problems we refer to [15]). However, in practice usually L1-SVM's 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.

Let us return to the approximation properties of H. Let $\mathcal{R}_{l,P} := \inf\{\mathcal{R}_{l,P}(f) \mid f : X \to \mathbb{R}\}$ denote the smallest possible l-risk. 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. Now, we define the *approximation error function* of the L1-SVM without offset by

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

Note that for $\lambda > 0$, the solution $f_{P,\lambda}$ of (7) satisfies

$$a(\lambda) = \lambda\|f_{P,\lambda}\|_H^2 + \mathcal{R}_{l,P}(f_{P,\lambda}) - \mathcal{R}_{l,P}.$$

In addition the obvious analogue of the approximation error function *with offset* is no greater than the approximation error function *without offset* so we restrict our attention to the latter. With the help of the approximation error function we define

Definition 2.2 Let H be a RKHS over X and P be a probability measure on $X \times Y$. We say that *H approximates P with exponent* $0 \leq \beta \leq 1$ if there exists a constant $C > 0$ such that

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

for all $\lambda > 0$.

Note, that H approximates P with exponent $\beta = 0$ for all pairs (H, P). We will see in the following section that the other extremal case $\beta = 1$ is equivalent to the fact that the minimal l-risk can be achieved by an element $f_{l,P} \in H$. Because of the specific structure of the approximation error function values $\beta > 1$ are only possible for distributions with $P_X(X_0) = 1$. The latter are uninteresting for classification considerations.

In order to state our first rate for L1-SVM's we finally need a complexity measure for RKHS's. To this end we have to recall some notations. For a subset $A \subset E$ of a Banach space E the *covering numbers* are defined by

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

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

Given a training set $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X\times Y)^n$ we denote the space of all equivalence classes of functions $f : X \times Y \to \mathbb{R}$ with norm

$$\|f\|_{L_2(T)} := \left(\frac{1}{n}\sum_{i=1}^{n}\big|f(x_i,y_i)\big|^2\right)^{\frac{1}{2}} \tag{9}$$

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 the 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 $x_1, \dots, x_n$. The complexity measure we need in our considerations is based on the spaces $L_2(T_X)$:

Definition 2.3 Let H be a RKHS over X and B_H its closed unit ball. We say that H *has complexity exponent* $0 < p \leq 2$ if there exists a constant $a_p > 0$ such that

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

for all $\varepsilon > 0$.

We will see in Section 9 that every RKHS has complexity exponent $p = 2$ by using the theory of absolutely 2-summing operators. However, for fast rates we need complexity exponents which are strictly smaller than 2. Furthermore, many SVM's use a parameterized family of RKHS's. For such SVM's the constant a_p may play a crucial role. We will see below, that this is in particular true for SVM's using a Gaussian RBF kernel.

Let us now consider learning rates for L1-SVM's. Our first result which establishes rates for L1-SVM's with general kernels reads as follows:

Theorem 2.4 *Let H be a RKHS of a continuous kernel on X with complexity exponent $0 < p < 2$, and let P be a probability measure on $X \times Y$ with Tsybakov noise exponent $0 \leq q \leq \infty$. Furthermore, assume that H approximates P with exponent $0 < \beta \leq 1$. Define $\lambda_n := n^{-\frac{4(q+1)}{(2q+pq+4)(1+\beta)}}$ and consider the L1-SVM without offset. Then for all $\varepsilon > 0$ there is a constant $C > 0$ such that for all $x \geq 1$ and all $n \geq 1$ we have*

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

Furthermore, the same result holds for the L1-SVM with offset if $q > 0$.

Remark 2.5 Using a tail bound of the form of Theorem 2.4 one can easily get convergence rates for (2). In the case of the above theorem these rates have the form $n^{-\frac{4\beta(q+1)}{(2q+pq+4)(1+\beta)}+\varepsilon}$ for all $\varepsilon > 0$. In other words the rates are exactly the terms in n in the above tail bounds. This is also true for the rates of L1-SVM's using Gaussian RBF kernels which are established below.

Remark 2.6 For brevity's sake our major aim was to show the best possible rates using our techniques. Therefore, the above theorem states rates for the L1-SVM under the assumption that (λ_n) optimizes the rates of the concentration inequalities we will apply in the proof of the theorem in Section 7. However, we emphasize, that the techniques of our proofs also give rates if (λ_n) is chosen in a different (and thus sub-optimal) way. This is also true for our results on L1-SVM's using Gaussian kernels.

Remark 2.7 If we assume a trivial Tsybakov exponent $q = 0$ we have $n^{-\frac{4\beta(q+1)}{(2q+pq+4)(1+\beta)}} = n^{-\frac{\beta}{1+\beta}}$. In other words, the rate of Theorem 2.4 is independent of the complexity exponent whenever H has a complexity exponent $p < 2$. We will show at the end of Section 9 that in this case actually no complexity condition on H is required. Recall that Tsybakov's rate in [32] is also essentially independent of the complexity of the used function class if $q = 0$.

Remark 2.8 In [32] it is assumed that a Bayes classifier is contained in the base function classes the algorithm minimizes over. This assumption corresponds to a perfect approximation of P by H, i.e. $\beta = 1$. In this case our rate is essentially of the form $n^{-\frac{2(q+1)}{2q+pq+4}}$. If we rescale the complexity exponent p from $(0,2)$ to $(0,1)$ and write p' for the new complexity measure this rate becomes essentially $n^{-\frac{q+1}{q+p'q+2}}$. This is exactly the *form* of Tsybakov's result in [32]. However, as far as we know our complexity measure cannot be compared to Tsybakov's.

Remark 2.9 By the nature of Theorem 2.4 it suffices to assume that P only satisfies Tsybakov's noise assumption for every $q' < q$. It also suffices to suppose that H approximates P with exponent β' for all $\beta' < \beta$, and that H has complexity exponent p' for all $p' > p$. As we will see in Section 3 the RKHS H has an approximation exponent $\beta = 1$ if and only if H contains a minimizer $f_{l,P}$ of the l-risk. In particular, if H has approximation exponent β for all $\beta < 1$ but not for $\beta = 1$ then H does not contain a minimizer $f_{l,P}$ but Theorem 2.4 can be applied for "$\beta = 1$". Furthermore, if the RKHS consists of C^∞ functions we can choose p arbitrarily close to 0. If both assumptions are true, we can hence obtain rates up to n^{-1} even though H does not contain a minimizer $f_{l,P}$ of the l-risk. For the Gaussian RBF kernel we can actually describe such distributions P in terms of their concentration near the "decision boundary" as we will see below.

The rest of this section is devoted to L1-SVM's using *Gaussian RBF kernels*, i.e. to kernels of the form $k_\sigma(x,x') = \exp(-\sigma^2\|x - x'\|_2^2)$, $x, x' \in X$, where $X \subset \mathbb{R}^d$ is a (compact) subset and $\sigma > 0$ is a free parameter called the *width*. We sometimes denote the corresponding RKHS by H_σ. The Gaussian RBF kernels are the most widely used kernels in practice. Of course we can apply Theorem 2.4 for these kernels, too. However, no smoothness condition on η or $f_P = \operatorname{sign} \circ (2\eta - 1)$ which ensures an approximation of P for some exponent $\beta > 0$ are known to us and the results of [27] indicate that such conditions must be very restrictive. We therefore choose another type of assumption on the distribution P. To this end we define the following function $x \mapsto \tau_x$ by

$$\tau_x \;:=\; \begin{cases} d(x, X_0 \cup X_1), & \text{if } x \in X_{-1}, \\ d(x, X_0 \cup X_{-1}), & \text{if } x \in X_1, \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

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

Definition 2.10 Let $X \subset \mathbb{R}^d$ be compact and P be a probability measure on $X \times Y$. We say that P has *geometric noise exponent* $\alpha > 0$ if there exists a constant $C > 0$ such that

$$\int_X |2\eta(x) - 1| \exp\Big(-\frac{\tau_x^2}{t}\Big) P_X(dx) \;\leq\; C t^{\frac{\alpha d}{2}} \tag{11}$$

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 (11) 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 2.11 Since $\exp(-t) \leq C_\alpha t^{-\alpha}$ holds for all $t > 0$ and a constant $C_\alpha > 0$ only depending on $\alpha > 0$ we easily see that (11) is satisfied whenever

$$\big(x \mapsto \tau_x^{-1}\big) \in L_{\alpha d}\big(|2\eta - 1| dP_X\big)\,. \tag{12}$$

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 $\big(x \mapsto \tau_x^{-1}\big) \in L_\infty\big(|2\eta - 1| dP_X\big)$ if and only if the two classes X_{-1} and X_1 have strictly positive distance! If (12) 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_σ for any $\sigma > 0$.

Note, that from (12) it is obvious that the parameter α in (12) 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.

The exponential function in (11) appears to be caused by the structure of the Gaussian kernel. Therefore, one can ask whether Definition 2.10 is taylored to the Gaussian kernel. The above example shows that condition (11) is actually a very general condition since distributions with (12) satisfies Definition 2.10. Obviously, (12) has no relation to the Gaussian RBF kernel. We now describe a regularity condition on η near the decision boundary that can be used to produce a geometric noise exponent. Like (12) this regularity condition does not have a relation to the Gaussian RBF kernel.

Definition 2.12 We say that η *is Hölder about* $\frac{1}{2}$ with exponent $\gamma > 0$ on $X \subset \mathbb{R}^d$ if there is a constant c_γ such that

$$|2\eta(x) - 1| \leq c_\gamma \tau_x^\gamma, \;\; \forall x \in X. \tag{13}$$

If η is Hölder about $\frac{1}{2}$ with exponent $\gamma > 0$, the graph of $2\eta(x) - 1$ lies in a multiple of the envelope defined by τ_x^γ at the top $-\tau_x^\gamma$ at the bottom. To be Hölder about $\frac{1}{2}$ it is sufficient that η is Hölder continuous, but it is far from being necessary. A function which is Hölder about $\frac{1}{2}$ can be very irregular away from X_0 but cannot jump across X_0 discontinuously. In addition a Hölder continuous function's exponent must satisfy $0 < \gamma \leq 1$ where being Hölder about $\frac{1}{2}$ only requires $\gamma > 0$. For distributions with Tsybakov noise exponent such that η is Hölder about $\frac{1}{2}$ we can bound the geometric noise exponent:

Theorem 2.13 *Let P be a probability measure on $X \times Y$ with $X \subset \mathbb{R}^d$ which has Tsybakov noise exponent $q \geq 0$ such that there exists a conditional probability $\eta(x) = P(y = 1|x)$ for P which is Hölder about $\frac{1}{2}$ with exponent $\gamma \geq 0$. Then when $q \geq 1$, P has geometric noise exponent $\alpha = \gamma\frac{q+1}{d}$ and when $0 \leq q < 1$, P has geometric noise exponent α for all $\alpha < \gamma\frac{q+1}{d}$.*

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

Theorem 2.14 *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and H_σ be the RKHS of the Gaussian RBF kernel k_σ on X with width $\sigma > 0$. We write $a_\sigma(.)$ for the approximation error function with respect to H_σ. Then there is a constant c_d depending only on d such that if P has geometric noise exponent $0 < \alpha < \infty$ with constant C, for all $\lambda > 0$ and all $\sigma > 0$ we have*

$$a_\sigma(\lambda) \;\leq\; c_d\Big(\sigma^d\lambda + C(4d)^{\frac{\alpha d}{2}}\sigma^{-\alpha d}\Big)\,. \tag{14}$$

In order to let the right hand side of (14) converge to zero it is necessary to assume both $\lambda \to 0$ and $\sigma \to \infty$. An easy consideration shows that the fastest rate of convergence can be achieved if $\sigma(\lambda) := \lambda^{-\frac{1}{(\alpha+1)d}}$. In this case we have $a_{\sigma(\lambda)}(\lambda) \leq 2C\lambda^{\frac{\alpha}{\alpha+1}}$. Roughly speaking this states that the family of spaces $H_{\sigma(\lambda)}$ approximates P with exponent $\frac{\alpha}{\alpha+1}$. Note, that we can obtain approximation rates up to linear order in λ for sufficiently benign distributions. The price for this good approximation property is, however, an increasing complexity of the hypothesis class $B_{H_{\sigma(\lambda)}}$ for $\sigma \to \infty$, i.e. $\lambda \to 0$. The following theorem estimates this in terms of the complexity exponent:

Theorem 2.15 *Let H_σ be the RKHS of the Gaussian RBF kernel k_σ on X and consider the evaluation map $I_\sigma : H_\sigma \to L_2(T_X)$ defined in (19) for an empirical distribution T. Then for all $0 < p \leq 2$ and $0 < \delta < \frac{2p}{8-4p}$, there exists a constant $c_{d,\delta} > 0$ such that for all $\varepsilon > 0$ and all $\sigma \geq 1$ we have*

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

In particular, Theorem 2.15 implies that for all $0 < p \leq 2$ and *all* $\delta > 0$ there exists a constant $c_{p,d,\delta} > 0$ such that for all $\varepsilon > 0$ and all $\sigma \geq 1$ we have

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

Having established both results for the approximation and complexity exponent we can now formulate our main result for L1-SVM's using Gaussian RBF kernels:

Theorem 2.16 *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 \leq q \leq \infty$ and geometric noise exponent $0 < \alpha < \infty$. We define*

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

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

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

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

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

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

Most of the remarks made after Theorem 2.4 also apply to the above theorem up to obvious modifications. In particular this is true for Remark 2.5, Remark 2.6, and Remark 2.9. Furthermore, Remark 2.8 applies if we assume "$p = 0$".

Acknowledgement:

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

3 Approximation error and the approximation error function

We need to control the classification risk $\mathcal{R}_P(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda})$ of a classifier built by optimizing a risk utilizing a loss function L on a training set T. Since the difference between the classifier built on the training T and that built on the measure P will be handled through the use of concentration inequalities, to consider the approximation error, we consider the classifier $(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda})$. Since this classifier uses a loss function L which is different from classification error we need to consider the price paid for minimizing L instead of classification error. Indeed, Steinwart [28] shows that when L is continuous and classification-calibrated that convergence to the minimal L-risk implies convergence to the Bayes risk. However to obtain rates it is useful to have a more quantitive estimate. For the hinge loss function l Zhang [36] proves that

$$\mathcal{R}_P(f) - \mathcal{R}_P \le 2(\mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P}) \tag{15}$$

for all measurable functions f. Zhang proves similar results for other common loss functions and Bartlett et al. [4] have provided a general framework for such inequalities. Therefore to have a quantitative bound on the excess classification risk it is sufficient to have a bound on the excess l-risk. As mentioned in Section 2 the obvious analogue of the approximation error function *with offset* is not greater than the approximation error function (8). Namely

$$\inf_{(f,b)\in H\times\mathbb{R}} \Big(\lambda\|f\|_H^2 + \mathcal{R}_{l,P}(f+b)\Big) - \mathcal{R}_{l,P} \ \le\ \inf_{f\in H} \Big(\lambda\|f\|_H^2 + \mathcal{R}_{l,P}(f)\Big) - \mathcal{R}_{l,P}.$$

Consequently, for all $\lambda > 0$ we have

$$\mathcal{R}_{l,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - \mathcal{R}_{l,P} \ \le\ \lambda\|\tilde{f}_{P,\lambda}\|_H^2 + \mathcal{R}_{l,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - \mathcal{R}_{l,P} \ \le\ a(\lambda).$$

Since $a(.)$ is defined as an infimum, we combine with Zhang's inequality (15) to produce the following chain of inequalities

$$\mathcal{R}_P(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) - \mathcal{R}_P \ \le\ 2a(\lambda) \ \le\ 2\big(\lambda\|f\|_H^2 + \mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P}\big), \qquad \forall f \in H \tag{16}$$

where, as we will see in Section 8, the last inequality allows the use of suboptimal test functions to bound the approximation error function.

One might guess that the addition of the $\lambda\|\tilde{f}_{P,\lambda}\|_H^2$ changing excess l-risk to the approximation error function might be too crude, but we show at the end of this section that this is not the case in most of the situations we consider. Along the way, we discuss the relationship between the approximation error, the approximation error function, and the map $\lambda \to \|f_{P,\lambda}\|$.

Here X denotes an arbitrary compact metric space, H a RKHS of continuous functions over X, and P a Borel probability measure on $X \times Y$. Unlike in the other sections of this paper, here L denotes an *arbitrary* convex loss function, that is a continuous function $L : Y \times \mathbb{R} \to [0, \infty)$ convex in its second variable. The corresponding L-risk $\mathcal{R}_{L,P}(f)$ of a function $f : X \to \mathbb{R}$ and its minimal value $\mathcal{R}_{L,P}$ are defined in the obvious way. For simplicity we also assume $\mathcal{R}_{L,P}(0) = 1$. Note that all the requirements are met by the hinge loss function.

We require the basic theory of RKHS as presented in [12]. Namely any continuous positive definite kernel $k(x, \acute{x})$ determines a Hilbert space of functions on X by

$$H := K_X^{\frac{1}{2}} L_2(X)$$

where $K_X^{\frac{1}{2}}$ is the unique square root of the integral operator $K_X : L_2(X) \to L_2(X)$ defined by

$$K_X f(x) := \int_X k(x, \acute{x}) f(\acute{x}) d\acute{x}\,, \qquad\qquad f \in L_2(X), x \in X\,,$$

and $L_2(X)$ denotes the L_2 space on X with Lebesgue measure. Note that the space $L_2(T_X)$ defined below (9) is something else. The norm on H is determined isometrically by

$$\|K_X^{\frac{1}{2}} f\|_H = \|f\|_{L_2(X)}.$$

The Hilbert space H consists of continuous functions on X and for $f \in H$ and $x \in X$ we have

$$|f(x)| \leq \|f\|_H \sqrt{k(x,x)}$$

giving rise to the inequality

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

where

$$K \ := \ \sup_{x \in X} \sqrt{k(x,x)}. \tag{17}$$

Consequently if we define

$$J_H : H \to C(X) \tag{18}$$

to be the embedding of the RKHS H into the continuous functions we have $\|J_H\| \leq K$. For universal kernels the range of J_H is dense in $C(X)$. For a training set T, consider the evaluation map $C(X) \to L_2(T_X) : f \mapsto f_{|T_X}$ defined below (9). This map has norm not greater than 1 and when composed with J_H produces the evaluation map of functions in H:

$$I_H : H \to L_2(T_X). \tag{19}$$

In Section 9, quantitative estimates on the compactness of I_H are provided to bound some Rademacher averages in Section 5.

Let us now proceed towards analyzing the approximation error function. We use the shorthand $\|\cdot\|$ for $\|\cdot\|_H$ when no confusion should arise. Analogously to the situation for the hinge loss we can define $f_{P,\lambda}$ if we replace the l-risk by the L-risk in (7). Furthermore, we define $f^*_{P,\lambda}$ through a set of intermediate minimizers $\hat{f}_{P,\lambda}$ defined as follows:

$$\hat{f}_{P,\lambda} \in \arg \min_{\|f\| \leq \frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f)\,, \tag{20}$$

Then $f^*_{P,\lambda}$ is defined as the unique element $f^*_{P,\lambda} \in \arg\min_{\|f\| \leq \frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f)$ with $\|f^*_{P,\lambda}\| \leq \|\hat{f}_{P,\lambda}\|$ for all $\hat{f}_{P,\lambda}$ satisfying (20). We need to prove

Lemma 3.1 *$f^*_{P,\lambda}$ is well defined.*

Proof: We first show that the set A of all solutions $\hat{f}_{P,\lambda}$ of (20) is nonempty. To that end consider a sequence (f_n) such that $\mathcal{R}_{L,P}(f_n) \to \inf_{\|f\| \leq \frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f)$. By the Eberlein-Smulyan theorem we can assume without loss of generality that there exists an f^* with $\|f^*\| \leq \frac{1}{\sqrt{\lambda}}$ such that $f_n \to f^*$ weakly. Using the fact that weak convergence in RKHS's imply pointwise convergence Lebesgue's theorem then gives

$$\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}(f^*)$$

by the continuity of L providing a solution of (20). We now proceed to show that there is a unique point in A with minimal norm.

Existence: Let $f_n \in A$ with

$$\|f_n\| \to \inf_{f \in A} \|f\|\,.$$

Like in the proof that A is not empty, we can conclude the existence of an $f^* \in A$ with

$$\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}(f^*).$$

This shows $f^* \in A$. Furthermore, we always have

$$\|f^*\| \le \liminf_{n\to\infty} \|f_n\| = \inf_{f\in A} \|f\|.$$

Uniqueness: Suppose we have two such elements f and g with $f \neq g$. By convexity we find $\frac{1}{2}(f+g) \in \arg\min_{\|f\|\le\frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f)$. However, H is strictly convex which gives $\|\frac{1}{2}(f+g)\| < \|f\|$. ∎

We can now define the approximation error and the approximation error function. In order to treat non-universal kernels we define the minimal L-risk of functions in H, i.e. the quantity

$$\mathcal{R}_{L,P,H} := \inf_{f\in H} \mathcal{R}_{L,P}(f).$$

Then we define

$$A(\lambda) := \inf_{f\in H} \Big(\lambda\|f\|^2 + \mathcal{R}_{L,P}(f)\Big) - \mathcal{R}_{L,P,H} \tag{21}$$

$$A^*(\lambda) := \inf_{\|f\|\le\frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P,H}. \tag{22}$$

Note that for $\lambda > 0$ we have

$$A(\lambda) = \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,H}$$

and

$$A^*(\lambda) = \mathcal{R}_{L,P}(f^*_{P,\lambda}) - \mathcal{R}_{L,P,H}.$$

Recall, that for universal kernels $\mathcal{R}_{L,P,H} = \mathcal{R}_{L,P}$ holds. Therefore, $A(.)$ equals the approximation error function $a(.)$ in this case. Furthermore, for these kernels, $A^*(\lambda)$ is the "classical" approximation error of the hypothesis class $\frac{1}{\sqrt{\lambda}}B_H$.

The following theorem (proven in Section 11) establishes some basic structure of these functions.

Theorem 3.2 *Consider the approximation error function $A(.)$ and the approximation error $A^*(.)$. We have $A(0) = A^*(0) = 0$, $A^*(.)$ is increasing, and $A(.)$ is increasing, concave, and continuous. In addition, we have*

$$A^*(\lambda) \le A(\lambda), \qquad \forall \lambda \ge 0$$

and for any function $h : (0,\infty) \to (0,\infty)$ such that $A^*(\lambda) \le h(\lambda)$ for all $\lambda > 0$, we have

$$A\big(\lambda h(\lambda)\big) \le 2h(\lambda), \qquad \forall \lambda > 0.$$

As a consequence, we note that $A(.)$ is a concave majorant of $A^*(.)$ and

$$\begin{aligned} \lambda A(1) \le A(\lambda) & \qquad \text{for all } 0 < \lambda \le 1, \\ A(\lambda) \le A(c\lambda) \le cA(\lambda) & \qquad \text{if } c \ge 1, \\ cA(\lambda) \le A(c\lambda) \le A(\lambda) & \qquad \text{if } 0 < c \le 1. \end{aligned}$$

We now turn to the main theorem of this section which establishes a relationship between the approximation error, the approximation error function, and $\lambda \to \|f_{P,\lambda}\|$. The proof appears in Section 11.

Theorem 3.3 *The function $\lambda \mapsto \|f_{P,\lambda}\|$ is bounded on $(0,\infty)$ if and only if $A(\lambda) \preceq \lambda$. In this case there exists an $f_{L,P,H} \in H$ minimizing the L-risk in H and we have $\lambda A(1) \leq A(\lambda) \leq \lambda \|f_{L,P,H}\|^2$. Moreover for all $\alpha > 0$ we have*

$$A^*(\lambda) \preceq \lambda^\alpha \qquad \textit{if and only if} \qquad A(\lambda) \preceq \lambda^{\frac{\alpha}{\alpha+1}}.$$

If one of the estimates is true we additionally have $\|f_{P,\lambda}\|^2 \preceq \lambda^{-\frac{1}{\alpha+1}}$ and $\mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P} \preceq \lambda^{\frac{\alpha}{\alpha+1}}$. Furthermore, if $\lambda^{\alpha+\varepsilon} \preceq A^*(\lambda) \preceq \lambda^\alpha$ for some $\alpha > 0$ and $\varepsilon \geq 0$ then we have

$$\lambda^{-\frac{\alpha}{(\alpha+\varepsilon)(\alpha+1)}} \preceq \|f_{P,\lambda}\|^2 \preceq \lambda^{-\frac{1}{\alpha+1}} \qquad \textit{and} \qquad \lambda^{\frac{\alpha+\varepsilon}{\alpha+1}} \preceq \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P} \preceq \lambda^{\frac{\alpha}{\alpha+1}}$$

and hence in particular $\lambda^{\frac{\alpha+\varepsilon}{\alpha+1}} \preceq A(\lambda) \preceq \lambda^{\frac{\alpha}{\alpha+1}}$.

The above theorem shows that under the assumption that $A^*(\lambda)$ behaves essentially like λ^α, both the excess L-risk and the approximation error function behave essentially like $\lambda^{\frac{\alpha}{\alpha+1}}$ supporting our claim that not much is lost in going from excess risk to the approximation error function in (16).

4 Bounding the estimation error of ERM-type classifiers using local Rademacher averages

In this section we will prove 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 which is closely related to local Rademacher averages. The latter have been intensively studied in learning theory in recent years (see [19], [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 $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha$ for constants $\alpha > 0$, $c > 0$, and certain functions g. However, for L1-SVM's and distributions P satisfying Tsybakov's noise condition for some $0 < q \leq \infty$ the "sharpest" variance bounds we will be able to show are of the form $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$ with $\delta > 0$. These bounds will be established in Section 6. We will also see there that 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. 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

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

The modulus of continuity will serve us as a complexity measure in the main theorem of this section. In Section 5 we will bound $\omega_n(\mathcal{F},\varepsilon)$ by local Rademacher averages which themselves are treated by certain covering numbers.

Before we state our main result we have to introduce some notation 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 4.1 *Let $\mathcal{F}$ be a convex set of bounded measurable functions from Z to $\mathbb{R}$ which is separable with respect to $\|.\|_\infty$ 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*

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

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 $g \in \mathcal{G}$. Let $n \geq 1$, $x > 0$ 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 4.2 Theorem 4.1 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-SVM's 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 4.1 is based on Talagrand's concentration inequality in [30] and its refinements in [25], [16], [18]. The below version of this inequality is derived from Bousquet's result in [8] using a little trick presented in [3, Lem. 2.5]:

Theorem 4.3 *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 4.4 *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*

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

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

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

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

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

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, our assumption on $\mathcal{G}$ yields

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

In order to bound the last probability we will apply Theorem 4.3. 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

$$\varepsilon \;\geq\; 10\mathbb{E}_{T' \sim P^n}\Big(\sup_{\substack{g \in \mathcal{G}, \\ \mathbb{E}_P g^2 \leq c\varepsilon^\alpha + \delta}} |\mathbb{E}_P g - \mathbb{E}_{T'} g|\Big) \;\geq\; 10\mathbb{E}_{T' \sim P^n} \sup_{h \in \mathcal{H}} \mathbb{E}_{T'} h\,. \tag{23}$$

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

$$\varepsilon \;\geq\; 10\left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}} \;\geq\; 10\left(\frac{1}{10} \cdot \left(\frac{60}{19}\right)^2\right)^{\frac{1}{2-\alpha}} \left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}} \;\geq\; \left(\frac{60}{19}\right)^{\frac{2}{2-\alpha}} \left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}} \tag{24}$$

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

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

This 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 (23) and $\varepsilon \geq \frac{10Bx}{n}$ we then find

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

Applying Theorem 4.3 then yields

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

■

With the help of the above lemma we can now prove the main result of this section, that is Theorem 4.1:

Proof of Theorem 4.1: We want to apply Lemma 4.4 to the class $\mathcal{G}$. It suffices to show the richness condition on $\mathcal{G}$ of Lemma 4.4. 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

$$\begin{aligned}\mathbb{E}_T(L\circ f_t - L\circ f_{P,\mathcal{F}}) \;&=\; \mathbb{E}_T(L\circ(tf+(1-t)f_{P,\mathcal{F}}) - L\circ f_{P,\mathcal{F}})\\ &\leq\; \mathbb{E}_T(tL\circ f + (1-t)L\circ f_{P,\mathcal{F}} - L\circ f_{P,\mathcal{F}})\\ &\leq\; t\mathbb{E}_T(L\circ f - L\circ f_{P,\mathcal{F}})\\ &\leq\; \varepsilon/20\,.\end{aligned}$$

Now, let $\varepsilon>0$ with $\varepsilon \geq 10\max\big\{\omega_n(\mathcal{G}, c\varepsilon^\alpha+\delta), \big(\frac{\delta x}{n}\big)^{\frac{1}{2}}, \big(\frac{4cx}{n}\big)^{\frac{1}{2-\alpha}}, \frac{Bx}{n}\big\}$. Then, by Lemma 4.4 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

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

we obtain the assertion. ■

5 Bounding the local Rademacher averages

The aim of this section is to bound the modulus of continuity of the class $\mathcal{G}$ in Theorem 4.1. 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 4.1.

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

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

Rademacher averages have been intensively used in empirical process theory. For more information we refer to [33]. 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. By symmetrization the modulus of continuity can be estimated by the local Rademacher average. More precisely, we always have (see [33])

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

Given a real number $a > 0$ we immediately obtain $\mathrm{Rad}(a\mathcal{F},n) = a\mathrm{Rad}(\mathcal{F},n)$. The following simple lemma describes how the local Rademacher averages behave under scaling the function class:

Lemma 5.1 *For all $a > 0$ we have*

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

Proof: Given a function class $\mathcal{G}$ we write $\mathcal{G}_\varepsilon := \{g : g\in\mathcal{G} \text{ and } \mathbb{E}_P g^2 \leq \varepsilon\}$. Obviously we have

$$(a\mathcal{F})_\varepsilon \;=\; \{f : f\in a\mathcal{F} \text{ and } \mathbb{E}_P f^2\leq\varepsilon\} \;=\; \{af : f\in\mathcal{F} \text{ and } \mathbb{E}_P(af)^2\leq\varepsilon\} \;=\; a\mathcal{F}_{a^{-2}\varepsilon} .$$

Since $\mathrm{Rad}(\mathcal{F},n,\varepsilon) = \mathrm{Rad}(\mathcal{F}_\varepsilon,n)$ we then obtain the assertion. ■

In the following we estimate Rademacher averages in terms of covering numbers using the path of [19]. Since we are interested in the arising constants and the extension of Theorem 2.4 described in Remark 2.7, we add the proofs for the sake of completeness. We begin by recalling an extension of a theorem of Dudley to subgaussian processes proven in [33]. For the formulation we also refer to [19]:

Theorem 5.2 *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*

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

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

The next theorem due to Talagrand [30] estimates the expected diameter of $\mathcal{F}$ when interpreted as a subset of $L_2(T)$:

Theorem 5.3 *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 [19]:

Proposition 5.4 *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*

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

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

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

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 5.2 and Theorem 5.3 to $\mathcal{F}_\varepsilon$ yields

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

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

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

This implies

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

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 4.1 by an assumption on the covering numbers of $\mathcal{G}$. As in Section 4 we assume that all minimizers exist. Then the corresponding result reads as follows:

Theorem 5.5 *Let $\mathcal{F}$ be a convex set of bounded measurable functions from Z to $\mathbb{R}$ which is separable with respect to $\|.\|_\infty$ and $L : \mathcal{F} \times Z \to [0,\infty)$ 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 there are constants $a\geq 1$ and $0<p<2$ with

$$\sup_{T\in Z^n} \log\mathcal{N}(B^{-1}\mathcal{G},\varepsilon, L_2(T)) \;\leq\; a\varepsilon^{-p} \tag{25}$$

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 $x>0$ 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

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

Proof: By Lemma 5.1 and Proposition 5.4 we find

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

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

$$\varepsilon^* = 2c_pB^{\frac{p}{2}}\bigl(c(\varepsilon^*)^\alpha+\delta\bigr)^{\frac12-\frac{p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12}\,. \tag{26}$$

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

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

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 4.1. In order to show the assertion it therefore suffices to bound ε from above. To this end let us first assume that

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

Then we have $\varepsilon = 2c_p B^{\frac{p}{2}} (c\varepsilon^\alpha + \delta)^{\frac{2-p}{4}} (\frac{a}{n})^{\frac{1}{2}}$. Since ε^* is the largest solution of this equation we hence find $\varepsilon \leq \varepsilon^*$. This shows that we always have

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

Hence it suffices to bound ε^* from above. To this end let us first assume $c(\varepsilon^*)^\alpha \geq \delta$. This implies

$$\varepsilon^* \;\leq\; 4c_p B^{\frac{p}{2}} \big(c \cdot (\varepsilon^*)^\alpha\big)^{\frac{2-p}{4}} \left(\frac{a}{n}\right)^{\frac{1}{2}}.$$

It is easy to see that this yields

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

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

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

Therefore we can obtain the assertion. ■

In this work we are mainly interested in L1-SVM's. Since Theorem 5.5 will be one of the main tools for the investigation of these algorithms we have to ensure that these classifiers fit into the framework of Theorem 5.5, i.e. that they are ERM-type algorithms. 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

$$L(f, x, y) \;:=\; \lambda \|f\|_H^2 + l(y, f(x)) \tag{27}$$

and

$$L(f, b, x, y) \;:=\; \lambda \|f\|_H^2 + l(y, f(x) + b) \tag{28}$$

for all $f \in H$, $b \in \mathbb{R}$, $x \in X$, and $y \in Y$. Since $\mathcal{R}_{L,T}(.)$ and $\mathcal{R}_{L,T}(.,.)$ coincide with the objective functions of the L1-SVM formulations we see that the L1-SVM's implement an empirical L-risk minimization. Furthermore note, that it is shown in [28] that all needed minimizers exist.

Below we will establish a simple lemma that estimates the covering numbers of the class $\mathcal{G}$ in Theorem 5.5 with the help of the covering numbers of B_H. Since for L1-SVM's the class $\mathcal{F}$ depends on the size of the offset $\tilde{b}_{P,\lambda}$ of the minimizer, we first have to bound this size. This is done in the following lemma which will be a crucial tool in investigating the L1-SVM's with offset. This lemma has been proved in [14] for empirical distributions. Although its generalization to general probability measures is straight forward we include the proof for completeness.

Lemma 5.6 *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 $P_X\big(x \in X : P(y^*|x) = 1\big) = 1$ for some $y^* \in Y$ there is nothing to be proved since $\tilde{b}_{P,\lambda} = y^*$ by our assumption on L1-SVM's 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 $\mathcal{R}_{l,P}(\tilde{f}_{P,\lambda}(x) + \tilde{b}_{P,\lambda}) > \mathcal{R}_{l,P}(\tilde{f}_{P,\lambda}(x) + b^*_{P,\lambda})$ by using the assumption on P. It is easily seen that this inequality contradicts the definition of $(\tilde{f}_{P,\lambda}(x), \tilde{b}_{P,\lambda})$. ■

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-SVM's Lemma 5.6 holds.

Recalling the definition of K in (17), we can state our announced covering number bound. For brevity's sake it only treats the case of L1-SVM's with offset. The other case can be shown completely analogously.

Lemma 5.7 *Let H be a RKHS over X, P be a probability measure on $X \times Y$, $\lambda > 0$, and L be defined by (28). Furthermore, let $1 \le \gamma \le \lambda^{-1}$ and*

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

Defining $B := 2\gamma K + 2$ 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 $\|g\|_\infty \le B$ for all $g \in \mathcal{G}$. Here $(f_{P,\mathcal{F}}, b_{P,\mathcal{F}})$ denotes a L-risk minimizer in $\mathcal{F}$. Assume that there are constants $a \ge 1$ and $0 < p < 2$ with

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

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

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

Proof: Let us write $\hat{\mathcal{G}} := \big\{L \circ (f, b) \;:\; (f, b) \in \mathcal{F}\big\}$ and $\mathcal{H} := \big\{l \circ (f + b) \;:\; (f, b) \in \mathcal{F}\big\}$. We then have

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

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

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

From this we easily deduce the assertion. ■

Note that for $\gamma := \lambda^{-\frac{1}{2}}$ the above lemma gives covering number bounds for L1-SVM's by Lemma 5.6. It will turn out in Sections 7 and 10 that in many situations it even can be applied for (slightly modified) L1-SVM's if γ is significantly smaller than $\lambda^{-\frac{1}{2}}$. In order to ensure that the above lemma is a non-void statement in this case we have to check that the minimizer $(f_{P,\mathcal{F}}, b_{P,\mathcal{F}})$ exists. This can be shown by an argument based on the weak compactness of closed balls in Hilbert spaces. Since this argument is only a small modification of the proof for the $\gamma = \lambda^{-\frac{1}{2}}$ case which was worked out in [28] we do not provide details here.

6 Variance bounds for L1-SVM's

In this section we prove some "variance bounds" in the sense of Theorem 4.1 and Theorem 5.5 for L1-SVM's. 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-SVM's—the general case

As already announced we will establish variance bounds for L1-SVM's for general probability measures in this section. Unfortunately, since our techniques heavily rely on the strict convexity of the RKHS norm it turns out that they can only be used for L1-SVM's *without* offset.

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 (27). 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 "pointwise Lipschitz continuous" with respect to $d_{x,y}$, i.e. we have

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

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

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

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

Lemma 6.1 *Let $0 < \lambda < 1$ and $\varepsilon > 0$. Then with the above notation we have*

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

Proof: Let $x \in X$, $y \in Y$ and $f, g \in \mathcal{F}$ with $d_{x,y}(f,g) \geq \varepsilon$. Then we find

$$\varepsilon \;\leq\; 2\sqrt{\lambda}\|f-g\|_H + |f(x)-g(x)| \;\leq\; (2+K)\|f-g\|_H\,.$$

Since l is convex and the norm $\|.\|$ of the RKHS satisfies the parallelogram law we also have

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

■

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

$$\delta_P(\varepsilon) \;:=\; \inf\biggl\{\frac{\mathcal{R}_{L,P}(f)+\mathcal{R}_{L,P}(g)}{2} - \mathcal{R}_{L,P}\Bigl(\frac{f+g}{2}\Bigr) : f,g \in \mathcal{F} \text{ with } d_P(f,g) \geq \varepsilon\biggr\}.$$

Again, it is easy to see that $\delta_P(\varepsilon) \geq 0$ for all $\varepsilon > 0$ by the convexity of L. The next lemma which is based on Lemma 6.1 significantly improves this:

Lemma 6.2 *With the above notations we have*

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

for all $0 < \lambda < 1$, $\varepsilon > 0$, *and all distributions* P *on* $X \times Y$.

Proof: Let f and g with $d_P(f,g) \geq \varepsilon$. Then by Lemma 6.1 we find

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

■

Now we can prove the main result of this subsection which states a "variance bound" for the class $\mathcal{G}$ defined in Theorem 4.1 for L1-SVM's without offset:

Proposition 6.3 *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 (27) 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*

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

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

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

for all $f \in \mathcal{F}$. Here, we used Lemma 6.2 in the last inequality. 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 pointwise Lipschitz-continuous with respect to $d_{x,y}$ we find

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

■

Remark 6.4 Proposition 6.3 establishes a variance bound of the form $\mathbb{E}_P g^2 \le c\,(\mathbb{E}_P g)^\alpha + \delta$ with $\alpha = 1$, $c = \frac{(4+2K)^2}{2\lambda}$, and $\delta = 0$. Recall, that by substituting α by 1 the term $\varepsilon := \varepsilon(n, a, B, c, \delta, x)$ in Theorem 5.5 becomes

$$\varepsilon(n, a, B, c, \delta, x) = B^{\frac{2p}{2+p}} c^{\frac{2-p}{2+p}} \Big(\frac{a}{n}\Big)^{\frac{2}{2+p}} + xB\Big(\frac{a}{n}\Big)^{\frac{2}{2+p}} + \frac{cx}{n}\,. \tag{29}$$

6.2 Bounding the variance for L1-SVM's—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 4.1. Besides the fact that this bound was only established for L1-SVM's without offset it appears to be sharp since it has the "optimal" values $\alpha = 1$ and $\delta = 0$, and actually this bound will enable us to prove the case $q = 0$ in Theorem 2.4. However, 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 $q > 0$. As already mentioned, it will turn out that these variance bounds *depend on the approximation error function!* An additional benefit of the approach of this subsection is that it can also be used for L1-SVM's *with* offset. In fact besides slightly larger constants the result are the same.

As in the last subsection l denotes the hinge loss. If no confusion can arise we denote a minimizer of $\mathcal{R}_{l,P}$ by $f_{l,P}$. For the shape of these minimizers which depend on $\eta := P(1|\,.\,)$ we refer to [36] and [29]. We begin with a variance bound for the empirical l-risk minimizer:

Lemma 6.5 *Let P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 < q \le \infty$. Then for all bounded measurable functions $f : X \to \mathbb{R}$ there exists a minimizer $f_{l,P}$ mapping into $[-1, 1]$ such that*

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

Proof: Given a fixed $x \in X$ we write $p := P(1|x)$ and $t := f(x)$. We first consider the case $p = 1/2$. Let $f_{l,P}$ be a minimizer with $f_{l,P}(x) = t$ if $t \in [-1, 1]$, $f_{l,P}(x) = 1$ if $t > 1$, and $f_{l,P}(x) = -1$ otherwise. Let us show

$$\begin{aligned} &\frac{\big(l(1,t) - l(1, f_{l,P}(x))\big)^2}{2} + \frac{\big(l(-1,t) - l(-1, f_{l,P}(x))\big)^2}{2} \\ \le\ & |t|\left(\frac{l(1,t) - l(1, f_{l,P}(x))}{2} + \frac{l(-1,t) - l(-1, f_{l,P}(x))}{2}\right). \end{aligned} \tag{30}$$

Obviously, this estimate is trivially satisfied if $t \in [-1, 1]$. If $t > 1$ we have $l(1, t) = l(1, f_{l,P}(x)) = 0$, $l(-1, t) = 1 + t$ and $l(-1, f_{l,P}(x)) = 2$. Therefore, (30) reduces to

$$\frac{(t-1)^2}{2} \le |t|\Big(\frac{t-1}{2}\Big)\,,$$

which is true for all $t > 1$. The case $t < -1$ can be shown analogously. Now, let us treat the case $p \neq 1/2$. We will show

$$\begin{aligned} & p\big(l(1,t) - l(1, f_{l,P}(x))\big)^2 + (1-p)\big(l(-1,t) - l(-1, f_{l,P}(x))\big)^2 \\ \leq\ & \Big(|t| + \frac{2}{|2p-1|}\Big)\Big(p\big(l(1,t) - l(1, f_{l,P}(x))\big) + (1-p)\big(l(-1,t) - l(-1, f_{l,P}(x))\big)\Big)\,. \end{aligned} \tag{31}$$

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 (31) reduces to

$$p\, l^2(1,t) + (1-p)\big(l(-1,t) - 2\big)^2 \ \leq\ \Big(|t| + \frac{2}{2p-1}\Big)\Big(p\, l(1,t) + (1-p)\big(l(-1,t) - 2\big)\Big)\,. \tag{32}$$

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, (32) 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

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

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 (32). Furthermore, for $p < \frac{1}{2}$ the proof of (31) is completely analogous and therefore (31) 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 (30) implies $h_2(x) \leq \|f\|_\infty h_1(x)$ for all x with $\eta(x) = 1/2$. Similarly, (31) 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

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

Now let t be defined 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. ■

With the help of Lemma 6.5 we can now show a variance bound for the L1-SVM. For brevity's sake we only state and prove the result for L1-SVM's with offset. Therefore, the loss function L is defined as in (28). Considering the proof it is immediately clear that the following variance bound also holds for the L1-SVM without offset.

Proposition 6.6 *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*

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

Proof: Let us define $\hat{C} := K\gamma + 1$. By Lemma 5.6 we then see $|\tilde{b}_{P,\lambda}| \leq \hat{C}$. For fixed $f + b$ we choose a minimizer $f_{l,P}$ according to Lemma 6.5. We first observe

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

By Lemma 6.5 and $a^p+b^p\le 2(a+b)^p$ for all $a,b\ge 0$, $0<p\le 1$ we find

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

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

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

■

Remark 6.7 Proposition 6.6 establishes a variance bound of the form $\mathbb{E}_P g^2\le c\,(\mathbb{E}_P g)^\alpha+\delta$ with $\alpha=\frac{q}{q+1}$, $c=(64+32\|(2\eta-1)^{-1}\|_{q,\infty})B^{\frac{q+2}{q+1}}$, and $\delta=(128+64\|(2\eta-1)^{-1}\|_{q,\infty})B^{\frac{q+2}{q+1}}a^{\frac{q}{q+1}}(\lambda)$. Recall, that by substituting α by $\frac{q}{q+1}$ the term $\varepsilon:=\varepsilon(n,a,B,c,\delta,x)$ in Theorem 5.5 becomes

$$\varepsilon=B^{\frac{2p(q+1)}{2q+pq+4}}c^{\frac{(2-p)(q+1)}{2q+pq+4}}\Big(\frac{a}{n}\Big)^{\frac{2(q+1)}{2q+pq+4}}+B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}\Big(\frac{a}{n}\Big)^{\frac{1}{2}}+B\Big(\frac{a}{n}\Big)^{\frac{2}{2+p}}+\sqrt{\frac{\delta x}{n}}+\Big(\frac{cx}{n}\Big)^{\frac{q+1}{q+2}}+\frac{Bx}{n}. \quad (33)$$

Of course, we can also replace c and δ by the above estimates. However, we will see in Section 7 and Section 10 that the above form is slightly easier to control.

7 Combining noise, complexity and approximation exponent: proof of Theorem 2.4

In this section we prove Theorem 2.4. Since our variance bounds have different forms for the cases $q=0$ and $q>0$ we prove the theorem for these cases separately. We begin with the case

$q = 0$. However, before we begin with the proof we explain its main idea which contains a method *fundamental* for all our results on rates. For simplicity we only consider the L1-SVM without offset in the following explanation. Recall that for these classifiers it is well known that $\|f_{T,\lambda}\| \leq \frac{1}{\sqrt{\lambda}}$ holds for all training sets T and all $\lambda > 0$. Now, let us assume that H approximates P with exponent $0 < \beta \leq 1$ and that H has complexity exponent $0 < p < 2$. Obviously we have $B \preceq \frac{1}{\sqrt{\lambda}}$. Furthermore, Proposition 6.3 shows $c \sim \frac{1}{\lambda}$ and $\delta = 0$. With the help of Remark 6.4 we then see that the term $\varepsilon(n, a, B, c, \delta, x)$ in Theorem 5.5 becomes

$$\varepsilon(n, a, B, c, \delta, x) \;=\; \lambda_n^{-\frac{p}{2+p}} \lambda_n^{-\frac{2-p}{2+p}} n^{-\frac{2}{2+p}} + x\lambda_n^{-\frac{1}{2}} n^{-\frac{2}{2+p}} + x\lambda_n^{-1} n^{-1} \preceq x\lambda_n^{-\frac{2}{2+p}} n^{-\frac{2}{2+p}} \tag{34}$$

if $\lambda_n n \to \infty$. Note that the latter is also a necessary condition for $\varepsilon(n, a, B, c, \delta, x) \to 0$. Now recall that $\mathcal{R}_P(f) - \mathcal{R}_P \leq 2\mathcal{R}_{l,P}(f) - 2\mathcal{R}_{l,P}$ holds for all measurable functions $f : X \to \mathbb{R}$ and the hinge loss function l as shown in [4] and [36]. Using Theorem 5.5 we then find

$$\mathcal{R}_P(f_{T,\lambda_n}) - \mathcal{R}_P \;\leq\; 2\big(\lambda\|f_{T,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda_n}) - \mathcal{R}_{l,P}\big) \;\leq\; 2a(\lambda_n) + c_p x\lambda_n^{-\frac{2}{2+p}} n^{-\frac{2}{2+p}} \tag{35}$$

with probability not less than $1 - e^{-x}$. Since $a(\lambda) \preceq \lambda^\beta$ it is easily checked that the fastest rate of convergence on the right side of (35) can be achieved for $\lambda_n := n^{-\frac{2}{2+2\beta+\beta p}}$. In this case the right side of (35) converges to zero with order $n^{-\frac{2\beta}{2+2\beta+\beta p}}$. Unfortunately, this is significantly worse than the result of Theorem 2.4! The reason for this bad rate is that we only used the trivial estimate $\|f_{T,\lambda}\| \leq \frac{1}{\sqrt{\lambda}}$. However, $\lambda\|f_{P,\lambda}\|^2 \leq a(\lambda) \preceq \lambda^\beta$ immediately implies $\|f_{P,\lambda}\| \preceq \lambda^{\frac{\beta-1}{2}}$. Now let us assume for a moment that we could prove such a bound for the empirical solutions $f_{T,\lambda}$, too. The term $\varepsilon(n, a, B, c, \delta, x)$ in Theorem 5.5 would then become

$$\begin{aligned}
\varepsilon(n, a, B, c, \delta, x) \;&=\; \lambda_n^{\frac{\beta-1}{2}\cdot\frac{2p}{2+p}} \lambda_n^{-\frac{2-p}{2+p}} n^{-\frac{2}{2+p}} + x\lambda_n^{\frac{\beta-1}{2}} n^{-\frac{2}{2+p}} + x\lambda_n^{-1} n^{-1} \\
&\preceq\; x\lambda_n^{\frac{\beta p-2}{2+p}} n^{-\frac{2}{2+p}} + x\lambda_n^{-1} n^{-1}\,.
\end{aligned}$$

Therefore, by Theorem 5.5 we would find

$$\mathcal{R}_P(f_{T,\lambda_n}) - \mathcal{R}_P \;\preceq\; \lambda_n^\beta + x\lambda_n^{\frac{\beta p-2}{2+p}} n^{-\frac{2}{2+p}} + x\lambda_n^{-1} n^{-1} \tag{36}$$

with high probability as in (35). It can be easily checked that the fastest rate of convergence on the right side of (36) would be achieved for $\lambda_n := n^{-\frac{1}{\beta+1}}$. In this case the right side of (36) would converge to zero with order $n^{-\frac{\beta}{\beta+1}}$, i.e. essentially with the order of Theorem 2.4 in the case $q = 0$. Unfortunately, we are not able to prove $\|f_{T,\lambda}\| \leq \lambda^{\frac{\beta-1}{2}}$. However, we will show that this bound "almost" holds with high probability. The idea of the proof is as follows: We define $\lambda_n := n^{-\frac{1}{\beta+1}}$ and begin with the trivial estimate $\|f_{T,\lambda}\| \leq \frac{1}{\sqrt{\lambda}}$. Then by estimate (35) and Theorem 5.5 we find a constant $C > 0$ such that for all $n \geq 1$ and all $x \geq 1$ the probability of

$$\begin{aligned}
\lambda_n\|f_{T,\lambda_n}\|^2 \;&\leq\; \lambda_n\|f_{T,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{T,\lambda_n}) - \mathcal{R}_{l,P} \\
&\leq\; \lambda_n\|f_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda_n}) - \mathcal{R}_{l,P} + Cx\lambda_n^{-\frac{2}{2+p}} n^{-\frac{2}{2+p}} \\
&\leq\; C\lambda_n^\beta + Cx\lambda_n^{\frac{2\beta}{2+p}}
\end{aligned}$$

is not smaller than $1 - e^{-x}$. For such training sets T we obtain $\|f_{T,\lambda_n}\| \leq Cx\lambda_n^{\frac{\beta}{2+p}-\frac{1}{2}} \leq Cx\lambda_n^{\frac{\beta}{4}-\frac{1}{2}}$. In other words, with high probability we have a nontrivial bound on $\|f_{T,\lambda_n}\|$ for large sample sizes.

The main idea of our shrinking technique which is used in all our proofs is to iterate the above step in order to successively improve the bound on $\|f_{T,\lambda_n}\|$. The following lemma works out a single step for $q = 0$ in the situation of Theorem 2.4:

Lemma 7.1 *Let H be a RKHS of a continuous kernel on X with complexity exponent $0 < p < 2$. Furthermore, let P be a probability measure on $X \times Y$ that is approximated by H with exponent $0 < \beta \le 1$. Define $\lambda_n := n^{-\frac{1}{\beta+1}}$ and assume that there are constants $0 \le \rho < \beta$ and $C \ge 1$ such that*

$$\Pr{}^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \le Cx\lambda_n^{\frac{\rho-1}{2}}\Big) \;\ge\; 1 - e^{-x}$$

for all $n \ge 1$ and all $x \ge 1$. Then there is another constant $\hat{C} \ge 1$ such that for $\hat{\rho} := \frac{\rho+\beta}{2}$ and for all $n \ge 1$, $x \ge 1$ we have

$$\Pr{}^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \le \hat{C}x\lambda_n^{\frac{\hat{\rho}-1}{2}}\Big) \;\ge\; 1 - e^{-x}\,.$$

Proof: Let $\hat{f}_{T,\lambda_n}$ be a minimizer of $\mathcal{R}_{L,T}$ on $Cx\lambda_n^{\frac{\rho-1}{2}} B_H$, where L is defined by (27). 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 $n \ge 1$, $x \ge 1$ the improved bound

$$\|\hat{f}_{T,\lambda_n}\| \le \tilde{C}x\lambda_n^{\frac{\hat{\rho}-1}{2}} \tag{37}$$

holds with probability not less than $1-e^{-x}$. Consequently, $\|f_{T,\lambda_n}\| \le \tilde{C}x\lambda_n^{\frac{\hat{\rho}-1}{2}}$ holds with probability not less than $1 - 2e^{-x}$. Obviously, the latter implies the assertion. In order to establish (37) we will apply Theorem 5.5 to the modified L1-SVM classifier which produces $\hat{f}_{T,\lambda_n}$. To this end we first observe that by Proposition 6.3 we may choose B, c, δ such that

$$\begin{aligned} B &\sim x\lambda_n^{\frac{\rho-1}{2}} \\ c &\sim x\lambda_n^{-1} \\ \delta &= 0\,. \end{aligned}$$

Furthermore, we can choose $a \sim 1$. By Remark 6.4 we then see that the term $\varepsilon(n, a, B, c, \delta, x)$ in Theorem 5.5 becomes

$$\begin{aligned} \varepsilon(n,a,B,c,\delta,x) &\preceq x\lambda_n^{\frac{(\rho-1)p}{2+p}}\lambda_n^{-\frac{2-p}{2+p}}n^{-\frac{2}{2+p}} + x^2\lambda_n^{\frac{\rho-1}{2}}n^{-\frac{2}{2+p}} + x\lambda_n^{-1}n^{-1} \\ &= x\lambda_n^{\frac{p\rho+2\beta}{2+p}} + x^2\lambda_n^{\frac{p\rho+2\rho+2-p+4\beta}{4+2p}} + x\lambda_n^{-1}\lambda_n^{\beta+1} \\ &\preceq x^2\lambda_n^{\frac{\rho+\beta}{2}}\,. \end{aligned}$$

By Theorem 5.5 there is therefore a constant $\tilde{C}_1 > 0$ independent of n and x such that for all $n \ge 1$ and all $x \ge 1$ the estimate

$$\begin{aligned} \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 &\le \lambda_n\|\hat{f}_{T,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{T,\lambda_n}) - \mathcal{R}_{l,P} \\ &\le \lambda_n\|\hat{f}_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1x^2\lambda_n^{\frac{\rho+\beta}{2}} \end{aligned}$$

holds with probability not less than $1-e^{-x}$. Now recall that our considerations in Section 3 showed $\|f_{P,\lambda_n}\| \preceq \lambda_n^{\frac{\beta-1}{2}}$. Since $\rho < \beta$ this implies $\|f_{P,\lambda_n}\| \le \lambda_n^{\frac{\rho-1}{2}} \le Cx\lambda_n^{\frac{\rho-1}{2}}$ for large n. In other words, for large n we have $f_{P,\lambda_n} = \hat{f}_{P,\lambda_n}$. With probability not less than $1 - e^{-x}$ this gives

$$\begin{aligned}
\lambda_n \|\hat{f}_{T,\lambda_n}\|^2 &\leq \lambda_n \|f_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(f_{P,\lambda_n}) - \mathcal{R}_{l,P} + \tilde{C}_1 x^2 \lambda_n^{\frac{\rho+\beta}{2}} \\
&\leq \tilde{C}_2 \lambda_n^{\beta} + \tilde{C}_1 x^2 \lambda_n^{\frac{\rho+\beta}{2}} \\
&\leq \tilde{C}_3 x^2 \lambda_n^{\frac{\rho+\beta}{2}}
\end{aligned}$$

for some constants $\tilde{C}_2, \tilde{C}_3 > 0$ independent of n and x. From this we easily obtain that (37) holds for *all* $n \geq 1$ with probability not less than $1 - e^{-x}$. ■

Proof of Theorem 2.4 for distributions with Tsybakov exponent $\mathbf{q = 0}$***:*** We define $\rho_0 := 0$ and $\rho_{i+1} := \frac{\rho_i + \beta}{2}$. Iteratively applying Lemma 7.2 gives a sequence of constants $C_i > 0$ with

$$\Pr{}^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq C_i x \lambda_n^{\frac{\rho_i - 1}{2}}\Big) \;\geq\; 1 - e^{-x}$$

for all $n \geq 1$ and all $x \geq 1$. Since an easy induction shows $\rho_i = (1 - 2^{-n})\beta$ we hence find a constant $C > 0$ such that

$$\Pr{}^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq C x \lambda_n^{\frac{(1-\varepsilon)\beta - 1}{2}}\Big) \;\geq\; 1 - e^{-x}$$

for all $n \geq 1$ and all $x \geq 1$. We write $\rho := (1 - \varepsilon)\beta$. As in the proof of Lemma 7.1 we denote a minimizer of $\mathcal{R}_{L,T}$ on $C x \lambda_n^{\frac{\rho-1}{2}} B_H$ by $\hat{f}_{T,\lambda_n}$. We have just seen that $\hat{f}_{T,\lambda_n} = f_{T,\lambda_n}$ with probability not less than $1 - e^{-x}$. Therefore, we only have to apply Theorem 5.5 to the modified optimization problem which defines $\hat{f}_{T,\lambda_n}$. As in the proof of Lemma 7.1 we see

$$\varepsilon(n, a, B, c, \delta, x) \;\preceq\; x^2 n^{-\frac{\rho+\beta}{2\beta+2}} .$$

Applying Theorem 5.5 we then obtain that with probability not less than $1 - e^{-x}$ we have

$$\begin{aligned}
\mathcal{R}_P(\hat{f}_{T,\lambda_n}) - \mathcal{R}_P &\leq 2\lambda_n \|\hat{f}_{T,\lambda_n}\|^2 + 2\mathcal{R}_{l,P}(\hat{f}_{T,\lambda_n}) - \mathcal{R}_{l,P} \\
&\leq 2\lambda_n \|\hat{f}_{P,\lambda_n}\|^2 + 2\mathcal{R}_{l,P}(\hat{f}_{P,\lambda_n}) - \mathcal{R}_{l,P} \\
&\leq a(\lambda_n) + \tilde{C}_1 x^2 n^{-\frac{\rho+\beta}{2\beta+2}} ,
\end{aligned} \tag{38}$$

where $\tilde{C}_1 > 0$ is a constant independent of n and x. Furthermore, we have already seen in the proof of Lemma 7.1 that $\lambda_n \|\hat{f}_{P,\lambda_n}\|^2 + \mathcal{R}_{l,P}(\hat{f}_{P,\lambda_n}) - \mathcal{R}_{l,P} \leq a(\lambda_n)$ holds for large n. Now, from (38) we easily deduce the assertion using the definition of ρ. ■

In the rest of this section we will prove Theorem 2.4 for distributions having a Tsybakov noise exponent $q > 0$. Since for such distributions our variance bound Proposition 6.6 significantly differs from Proposition 6.3 which has been used for $q = 0$ we first have to establish a new shrinking lemma:

Lemma 7.2 *Let H be a RKHS of a continuous kernel on X with complexity exponent $0 < p < 2$, and let P be a distribution with Tsybakov noise exponent $0 < q \leq \infty$. Furthermore, assume that H approximates P with exponent $0 < \beta \leq 1$. Define $\lambda_n := n^{-\frac{4(q+1)}{(2q+pq+4)(1+\beta)}}$ and assume that there are constants $0 \leq \rho < \beta$ and $C \geq 1$ such that*

$$\Pr{}^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq C x \lambda_n^{\frac{\rho-1}{2}}\Big) \;\geq\; 1 - e^{-x}$$

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

$$\Pr{}^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq \hat{C} x \lambda_n^{\frac{\hat{\rho}-1}{2}}\Big) \;\geq\; 1 - e^{-x}\,.$$

The same result holds for L1-SVM's with offset.

Proof: For brevity's sake we only prove this Lemma for L1-SVM's *with offset.* The proof for L1-SVM's without offset is almost identical. Therefore, let L be defined by (28). Analogously to the proof of Lemma 7.1 we denote a minimizer of $\mathcal{R}_{L,T}(.,.)$ on $Cx\lambda_n^{\frac{\rho-1}{2}}(B_H \times [-K-1, K+1])$ by $(\hat{f}_{T,\lambda_n}, \hat{b}_{T,\lambda_n})$. By our assumption Lemma 5.6 shows $|\tilde{b}_{T,\lambda_n}| \leq Cx\lambda_n^{\frac{\rho-1}{2}}(K+1)$ with probability not less than $1 - e^{-x}$ for all possible values of the offset. Therefore, for such training sets we have $\hat{f}_{T,\lambda_n} = \tilde{f}_{T,\lambda_n}$ since the RKHS component $\tilde{f}_{T,\lambda_n}$ of L1-SVM solutions is unique for every training set T by the strict convexity of L in f. Furthermore, by the above considerations we may define $\hat{b}_{T,\lambda_n} := \tilde{b}_{T,\lambda_n}$ for such training sets. As in the proof of Lemma 7.1 it now suffices to show the existence of a constant $\tilde{C} > 0$ which satisfies

$$\|\hat{f}_{T,\lambda_n}\| \leq \tilde{C} x \lambda_n^{\frac{\hat{\rho}-1}{2}} \tag{39}$$

with probability not less than $1 - e^{-x}$. To this end we first observe by Proposition 6.6 that we may choose B, c and δ such that

$$\begin{aligned}
B &\sim x\lambda_n^{\frac{\rho-1}{2}} \\
c &\sim x^{\frac{q+2}{q+1}} \lambda_n^{\frac{\rho-1}{2}\cdot\frac{q+2}{q+1}} \\
\delta &\sim x^{\frac{q+2}{q+1}} \lambda_n^{\frac{\rho-1}{2}\cdot\frac{q+2}{q+1}+\frac{\beta q}{q+1}}\,.
\end{aligned}$$

Furthermore, we can obviously choose $a \sim 1$. With these relations Remark 6.7 tells us

$$\begin{aligned}
\varepsilon(n,a,B,c,\delta,x) &\preceq x\lambda_n^{\frac{\rho-1}{2}} n^{-\frac{2(q+1)}{2q+pq+4}} + x\lambda_n^{\frac{(\rho-1)}{2}\cdot\frac{2q+pq+4}{4(q+1)}+\frac{2-p}{4}\cdot\frac{\beta q}{q+1}} n^{-\frac{1}{2}} + x\lambda_n^{\frac{\rho-1}{2}} n^{-\frac{2}{2+p}} \\
&\quad + x^2\lambda_n^{\frac{\rho-1}{4}\cdot\frac{q+2}{q+1}+\frac{\beta q}{2(q+1)}} n^{-\frac{1}{2}} + x^2\lambda_n^{\frac{(\rho-1)}{2}} n^{-\frac{q+1}{q+2}} + x^2\lambda_n^{\frac{(\rho-1)}{2}} n^{-1} \\
&\preceq x^2\lambda_n^{\frac{\rho-1}{2}} n^{-\frac{2(q+1)}{2q+pq+4}} + x^2\lambda_n^{\frac{(\rho-1)}{2}\cdot\frac{2q+pq+4}{4(q+1)}+\frac{2-p}{4}\cdot\frac{\beta q}{q+1}} n^{-\frac{1}{2}} \\
&\sim x^2\lambda_n^{\frac{\rho-1}{2}}\lambda_n^{\frac{1+\beta}{2}} + x^2\lambda_n^{\frac{(\rho-1)}{2}\cdot\frac{2q+pq+4}{4(q+1)}+\frac{2-p}{4}\cdot\frac{\beta q}{q+1}}\lambda_n^{\frac{(2q+pq+4)(1+\beta)}{8(q+1)}} \\
&\sim x^2\lambda_n^{\frac{\rho+\beta}{2}} + x^2\lambda_n^{\frac{(\rho+\beta)(2q+pq+4)+2\beta q(2-p)}{8(q+1)}}\,.
\end{aligned}$$

Now observe that $\beta > \rho \geq 0$ and $p < 2$ imply

$$\frac{\rho+\beta}{2} \leq \frac{(\rho+\beta)(2q+pq+4)+2\beta q(2-p)}{8(q+1)}\,.$$

Therefore we obtain

$$\varepsilon(n,a,B,c,\delta,x) \;\preceq\; x^2\lambda_n^{\frac{\rho+\beta}{2}}\,.$$

The rest of the proof is analogous to the proof of Lemma 7.1 ■

***Proof of Theorem 2.4 for distributions with Tsybakov exponent* $\mathbf{q > 0}$:** By using Lemma 7.2 the proof in the case $q > 0$ is completely analogous to the case $q = 0$. ■

8 The approximation error function for Gaussian kernels

We now consider the approximation error function for the RKHS H_σ on the closed unit ball $X \subset \mathbb{R}^d$ defined by the Gaussian kernel

$$k_\sigma(x, \acute{x}) = e^{-\sigma^2|x-\acute{x}|^2},$$

where we denote the Euclidian norm on $\mathbb{R}^d$ by $|\cdot|$. In the following it will be useful to consider the integral operators and their associated RKHS's on more general sets than X. Let $\Omega \subset \mathbb{R}^d$ be measurable and let $K_{\Omega,\sigma}$ denote the integral operator with kernel k_σ on $L_2(\Omega)$ and when necessary denote the corresponding RKHS, discussed in Section 3, by $H_\sigma(\Omega)$. If we denote $i_\Omega : L_2(\Omega) \to L_2(\mathbb{R}^d)$ the extension of a function on Ω by zero to the rest of $\mathbb{R}^d$ and by $r_\Omega : L_2(\mathbb{R}^d) \to L_2(\Omega)$ the restriction of a function on $\mathbb{R}^d$ to the set Ω, then $\|i_\Omega\| = 1$ and $\|r_\Omega\| \leq 1$ and

$$K_{\Omega,\sigma} = r_\Omega K_{\mathbb{R}^d,\sigma} i_\Omega. \tag{40}$$

It will also be useful to consider the normalized Gaussian kernel

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

and call them normalized since integration with respect to x or $\acute{x}$ over $\mathbb{R}^d$ produces unity. We also consider the corresponding *Gauss-Weierstrass integral operator* $\hat{K}_{\mathbb{R}^d,\sigma}$ and the normalized operators $\hat{K}_{\Omega,\sigma}$. In particular (40) also holds for the normalized operators.

We need a preparatory lemma.

Lemma 8.1 *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 Section 3 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. [23]) 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 [24] states that for convolutions

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

and since the Gauss-Weierstrass integral operator $\hat{K}_{\mathbb{R}^d,\sigma}$ is a convolution and $\int \sigma^d \pi^{-\frac{d}{2}} e^{-\sigma^2|u|^2} du = 1$ it follows that $\hat{K}_{\mathbb{R}^d,\sigma}$ is a contraction. From (40) 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 $\|\hat{K}_{\Omega,\sigma}\| \leq 1$. ∎

When the distribution has a nontrivial geometric noise exponent, we can now esablish upper bounds on the approximation error function for Gaussian RKHS in terms of the parameter σ.

Proof of Theorem 2.14: We utilize the righthand side of (16), i.e.

$$a(\lambda) \leq \lambda \|f\|_\sigma^2 + \mathcal{R}_{l,P}(f) - \mathcal{R}_{l,P}, \quad f \in H_\sigma(X) \tag{41}$$

to bound the approximation error function through a judicious choice of function $\hat{f} \in H_\sigma(X)$. Let $\eta(x) = P(y=1|x)$ be any regular conditional distribution for P and let f_P be any Bayes function with values in $[-1,1]$ such that $f_P = 1$ on X_1 and $f_P = -1$ on X_{-1}. We will choose a function $\hat{f}$ by smoothing the extension $\acute{f}_P$ of f_P to $\acute{X} := 3X$. To do so first consider the extension of η to be constant in the outward radial direction

$$\acute{\eta}(x) = \begin{cases} \eta(x), & |x| \leq 1 \\ \eta(\frac{x}{|x|}), & |x| > 1 \end{cases} \tag{42}$$

and define $\acute{X}_{-1} := \{x \in \acute{X} : \acute{\eta}(x) < \frac{1}{2}\}$, $\acute{X}_1 := \{x \in \acute{X} : \acute{\eta}(x) > \frac{1}{2}\}$. 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 8.2 *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 $\acute{\eta}(x') = \eta\left(\frac{x'}{|x'|}\right) > \frac{1}{2}$. ∎

Let $\acute{f}_P$ be a measurable function with values in $[-1,1]$ which coincides with f_P on X such that $\acute{f}_P = 1$ on $\acute{X}_1$ and $\acute{f}_P = -1$ on $\acute{X}_{-1}$. Consider the function $\hat{f} = r_X \hat{K}_{\acute{X},\sigma} \acute{f}_P$. We first need to show that $\hat{f} \in H_\sigma(X)$. In addition we will bound the first term $\lambda\|\hat{f}\|^2$ in the righthand side of inequality (41). According to Aronszajn [1] we have $r_X H_\sigma(\acute{X}) \subset H_\sigma(X)$ and

$$\|r_X f\|_{H_\sigma(X)} \leq \|f\|_{H_\sigma(\acute{X})}$$

for all $f \in H_\sigma(\acute{X})$. Consequently to show that $\hat{f} = r_X \hat{K}_{\acute{X},\sigma} \acute{f}_P \in H_\sigma(X)$ it suffices to show that $\hat{K}_{\acute{X},\sigma} \acute{f}_P \in H_\sigma(\acute{X})$. We apply Lemma 8.1 with $\Omega = \acute{X}$ to obtain

$$\begin{aligned} \|\hat{f}\|_{H_\sigma(X)} = \|r_X \hat{K}_{\acute{X},\sigma} \acute{f}_P\|_{H_\sigma(X)} \leq \|\hat{K}_{\acute{X},\sigma} \acute{f}_P\|_{H_\sigma(\acute{X})} \leq \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} \|\acute{f}_P\|_{L_2(\acute{X})} &\leq \sigma^{\frac{d}{2}} \pi^{-\frac{d}{4}} vol(\acute{X}) \\ &= \sigma^{\frac{d}{2}} (\frac{81}{\pi})^{\frac{d}{4}} \theta(d) , \end{aligned} \tag{43}$$

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

We now proceed to bound the term $\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P}$ in the righthand side of inequality (41). For any function which satisfies $-1 \leq f \leq 1$, Zhang [36] shows that

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

Since $-1 \leq \acute{f}_P \leq 1$ it follows that $-1 \leq i_{\acute{X}}\acute{f}_P \leq 1$. It is well known for the Gauss-Weierstrass heat operator $\hat{K}_{\mathbb{R}^d,\sigma}$ that consequently

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

Since $\hat{K}_{\acute{X},\sigma} = r_{\acute{X}}\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}}$ follows from (40) and P_X has support in X we obtain

$$\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P} = \mathcal{R}_{l,P}(\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}}\acute{f}_P) - \mathcal{R}_{l,P} = \mathbb{E}_{P_X}(|2\eta - 1||\hat{K}_{\mathbb{R}^d,\sigma} i_{\acute{X}}\acute{f}_P - f_P|). \tag{44}$$

Now for $x \in X$ we have

$$\begin{aligned} \hat{f}(x) = \int_{\acute{X}} \hat{k}_\sigma(x,\acute{x})\acute{f}_P(\acute{x})d\acute{x} &= \int_{\mathbb{R}^d} \hat{k}_\sigma(x,\acute{x}) i_{\acute{X}}\acute{f}_P(\acute{x})d\acute{x} \\ &= \int_{\mathbb{R}^d} \hat{k}_\sigma(x,\acute{x})(i_{\acute{X}}\acute{f}_P(\acute{x})+1)d\acute{x} - 1 \\ &\geq \int_{B(x,\tau_x)} \hat{k}_\sigma(x,\acute{x})(i_{\acute{X}}\acute{f}_P(\acute{x})+1)d\acute{x} - 1. \end{aligned} \tag{45}$$

When $x \in X_1$, Lemma 8.2 showed that $B(x,\tau_x) \subset \acute{X}_1$ so that (45) implies

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

where $\gamma_\sigma = \sigma^d(\pi)^{-\frac{d}{2}}e^{-\sigma^2|u|^2}du$ is a spherical Gaussian in $\mathbb{R}^d$. According to the tail bound inequality [17, inequality 3.5, p. 59] for spherical Gaussians we have

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

Consequently, for $x \in X_1$ we obtain

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

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

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

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

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

Consequently from inequality (44) and letting $t = \frac{4d}{\sigma^2}$ in the geometric noise assumption we obtain

$$\mathcal{R}_{l,P}(\hat{f}) - \mathcal{R}_{l,P} \;\leq\; 8\mathbb{E}_{x\sim P_X}(|2\eta(x) - 1|e^{-\sigma^2\tau_x^2/4d}) \;\leq\; 8C(4d)^{\frac{\alpha d}{2}}\sigma^{-\alpha d}\,, \tag{46}$$

where α and C are the constants corresponding the geometric noise assumption. To finish the proof of Theorem 2.14, we apply the inequalities (46) and (43) to inequality (41) with the choice $\hat{f} = r_X\hat{K}_{\acute{X},\sigma}\acute{f}_P$. ■

8.1 Sufficient conditions for the geometric noise condition

In this section we provide a relationship between the geometric noise exponent and the Tsybakov noise and Hölder about $\frac{1}{2}$ exponents.

Proof of Theorem 2.13: When $\gamma = 0$ the theorem is trivially true so we assume $\gamma > 0$. In the following, all Lebesgue and Lorentz spaces (see e.g. [5]) are with respect to the measure $\hat{P}_X$. First let us consider the case $q \geq 1$ where we can apply the Hölder inequality for Lorentz spaces ([20])

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

where $\acute{q}$ is defined by $\frac{1}{q} + \frac{1}{\acute{q}} = 1$, to obtain that

$$\mathbb{E}_{x\sim P_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) = \mathbb{E}_{\hat{P}_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \leq \|(2\eta-1)^{-1}\|_{q,\infty}\|(2\eta-1)^2 e^{-\frac{\tau_x^2}{t}}\|_{\acute{q},1}.$$

The Hölder about $\frac{1}{2}$ assumption implies that

$$|2\eta(x)-1|^2 e^{-\frac{\tau_x^2}{t}} \leq |2\eta(x)-1|^2 e^{-\left(\frac{|2\eta(x)-1|}{c_\gamma}\right)^{\frac{2}{\gamma}} t^{-1}}$$

for all $x \in X$. Let $a = |2\eta - 1|^{-1}$ and $b = t(c_\gamma)^{\frac{2}{\gamma}}$ so that

$$|2\eta - 1|^2 e^{-\frac{\tau_x^2}{t}} \leq g(a)$$

where $g(a) = a^{-2}e^{-\frac{a^{-\frac{2}{\gamma}}}{b}}$. Since the range of a is constrained to $a \geq 1$ one can show for $\gamma > 0$ and $0 < b \leq \frac{2}{3\gamma}$ that g is strictly increasing and invertible on $a \geq 1$. Extend g to a strictly increasing and invertible function on $\mathbb{R}^+$ and denote this extension also by g. Then for this extension we have $\hat{P}_X(g(a) > g(\tau)) = \hat{P}_X(a > \tau)$ which amounts to

$$\hat{P}_X(g(a) > \tau) = \hat{P}_X(a > g^{-1}(\tau))$$

For a function f we utilize the non-increasing rearrangement

$$f^*(u) \;:=\; \inf\{\sigma : \hat{P}_X(f > \sigma) \leq u\},$$

of f which can be used to compute Lorentz norms (see e.g. [5]). The identity $(g \circ a)^* = g \circ a^*$ follows immediately:

$$\begin{aligned}
(g\circ a)^*(u) &= \inf\{\sigma : \hat{P}_X(g(a) > \sigma) \leq u\} \\
&= \inf\{\sigma : \hat{P}_X(a > g^{-1}(\sigma)) \leq u\} \\
&= g(\inf\{g^{-1}(\sigma) : \hat{P}_X(a > g^{-1}(\sigma)) \leq u\}) \\
&= g(a^*(u)) \\
&= g \circ a^*(u).
\end{aligned}$$

Now, inequality (5) implies $\hat{P}_X\big(a \leq (\frac{u}{C})^{\frac{1}{q}}\big) \leq u$ for all $u > 0$. Therefore, we find

$$a^*(u) \;=\; \inf\{\sigma : \hat{P}_X(a > \sigma) \leq u\} \;\leq\; \inf\{\sigma : \hat{P}_X(a \geq \sigma) \leq u\} \;\leq\; \left(\frac{u}{C}\right)^{-\frac{1}{q}}$$

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

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

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

$$g(a) \preceq b^{\hat\alpha} \frac{a^{2\left(\frac{\hat\alpha}{\gamma}-1\right)}}{\ln^2\left(a^{-\frac{2}{\gamma}} b^{-1}\right)+1}$$

and so

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

If we define $\hat\alpha := \gamma\frac{q+1}{2}$ then it follows that $\frac{1}{\acute q} + \frac{2}{q}(1-\frac{\hat\alpha}{\gamma}) = 0$. Consequently,

$$\begin{aligned}
\|(2\eta-1)^2 e^{-\frac{\tau_x^2}{t}}\|_{\acute q,1} \;\leq\; \|g(a)\|_{\acute q,1} \;=\; \int_0^\infty u^{\frac{1}{\acute q}}(g\circ a)^*(u)\frac{du}{u} \;&\preceq\; b^{\hat\alpha}\int_0^\infty \frac{1}{\ln^2\left(\left(\frac{u}{C}\right)^{\frac{2}{q\gamma}} b^{-1}\right)+1}\frac{du}{u} \\
&\preceq\; b^{\hat\alpha}\int_0^\infty \frac{1}{\ln^2 u+1}\frac{du}{u}
\end{aligned}$$

by a change of variables. Since

$$\int_0^\infty \frac{1}{\ln^2 u+1}\frac{du}{u} < \infty$$

we obtain that

$$\mathbb{E}_{P_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \;\preceq\; t^{\gamma\frac{q+1}{2}} \tag{47}$$

for $t \leq \frac{2}{3\gamma(c_\gamma)^{\frac{2}{\gamma}}}$. In addition, since $\mathbb{E}_{P_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \leq 1$ for all positive t estimate (47) holds for all $t > 0$. Since $t^{\gamma\frac{q+1}{2}} = t^{\frac{\alpha d}{2}}$ with $\alpha = \gamma\frac{q+1}{d}$, the Definition 2.10 of the geometric noise exponent implies the assertion for $q \geq 1$.

Now consider the case $0 \leq q < 1$ where the Hölder inequality in Lorentz space does not apply. Then

$$\begin{aligned}
& \mathbb{E}_{P_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \\
=\;& \mathbb{E}_{\hat{\mathcal{P}}_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \\
=\;& \mathbb{E}_{\hat{\mathcal{P}}_X}(\mathbf{1}_{|2\eta(x)-1|\leq\tau}|2\eta(x)-1|e^{-\tau_x^2/t}) + \mathbb{E}_{\hat{\mathcal{P}}_X}(\mathbf{1}_{|2\eta(x)-1|>\tau}|2\eta(x)-1|e^{-\tau_x^2/t}) \\
\leq\;& C\tau^{q+1} + \mathbb{E}_{\hat{\mathcal{P}}_X}(\mathbf{1}_{|2\eta(x)-1|>\tau}|2\eta(x)-1|e^{-\tau_x^2/t}).
\end{aligned}$$

Since η is Hölder about $\frac{1}{2}$ (inequality (13)) we obtain

$$\mathbb{E}_{P_X}(|2\eta(x)-1|e^{-\tau_x^2/t}) \leq C\tau^{q+1} + e^{-(\frac{\tau}{c_\gamma})^{\frac{2}{\gamma}}/t} \tag{48}$$

for all $t, \tau \geq 0$. We define τ by

$$\tau^{q+1} \;:=\; e^{-(\frac{\tau}{c_\gamma})^{\frac{2}{\gamma}}/t}.$$

For $\hat{a} := (c_\gamma)^{\frac{2}{\gamma}}(q+1)$ and small t this definition implies

$$\tau \leq \big(\frac{\hat{a}\gamma}{2}\big)^{\frac{\gamma}{2}}\big(t\ln\frac{1}{\hat{a}t}\big)^{\frac{\gamma}{2}}.$$

Since

$$\big(\frac{\hat{a}\gamma}{2}\big)^{\frac{\gamma(q+1)}{2}}\big(t\ln\frac{1}{\hat{a}t}\big)^{\frac{\gamma(q+1)}{2}} \preceq t^{\frac{\alpha d}{2}}$$

for all $\alpha < \gamma\frac{q+1}{d}$, inequality (48) and the Definition 2.10 of the geometric noise exponent implies the assertion for $0 < q < 1$. ■

9 Lorentz norms on the log covering numbers for Gaussian kernels

In this section we consider the map $I_{H_\sigma} : H_\sigma \to L_2(T_X)$ defined in (19) for the Gaussian RKHS H_σ defined by the Gaussian kernel

$$k_\sigma(x,\acute{x}) = e^{-\sigma^2|x-\acute{x}|^2}.$$

In particular we provide bounds on the covering numbers of I_{H_σ} needed in Section 5. Along the way we bound the covering numbers of the map J_{H_σ} defined in (18). We use the shorthand notation I_σ for I_{H_σ} and J_σ for J_{H_σ}. Since the bounds will be of the form $\log\mathcal{N}(\varepsilon) \leq C\varepsilon^{-\frac{1}{p}}$ for some p and C, and such an inequality implies that $\log\mathcal{N}(.)$ lies in the Lorentz space $L_{p,\infty}$ (see e.g. [5]) with norm not greater than C we refer to such bounds as bounds on the Lorentz norms of the log covering numbers. We begin by first considering the factor $J_\sigma : H_\sigma \to C(X)$ of I_σ.

Theorem 9.1 *Consider the embedding $J_\sigma : H_\sigma \to C(X)$ and let $0 < p < 2$. There is a constant $c_{p,d} > 0$ depending only on p and d such that for all $\varepsilon > 0$*

$$\log\mathcal{N}(J_\sigma,\epsilon) \leq c_{p,d}\sigma^{(1-\frac{p}{4})d}\epsilon^{-p}.$$

Proof: Since $H_\sigma = H_\sigma(X)$ consists of analytic functions, H_σ is isometrically isomorphic with $H_\sigma(\mathring{X})$ where $\mathring{X} \subset X \subset \mathbb{R}^d$ is the open unit ball ([1]). Consequently in the following we do not concern ourselves with the distinction between $H_\sigma(X)$ and $H_\sigma(\mathring{X})$. Let $K_\sigma : L_2(\mathring{X}) \to L_2(\mathring{X})$ denote the integral operator with kernel k_σ on the open unit ball $\mathring{X}$. Let $\|\cdot\|$ denote the norm in $L_2(\mathring{X})$. According to Cucker and Smale [12, Thm. 3, p. 27] we obtain

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

for all $f \in H_\sigma$ where $\|K_\sigma^{-1}h\| = \infty$ if $h \neq K_\sigma g$ for some $g \in L_2(\mathring{X})$.

Suppose now that $\mathcal{H} \subset L_2(\mathring{X})$ is a dense Hilbert space with $\|h\| \leq \|h\|_{\mathcal{H}}$, and that $K_\sigma : L_2(\mathring{X}) \to \mathcal{H} \subset L_2(\mathring{X})$ with $\|K_\sigma : L_2(\mathring{X}) \to \mathcal{H}\| \leq c_{\sigma,\mathcal{H}}$. It follows that

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

so that

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

By a result of Smale and Zhou [27, Thm. 3.1] it follows that f is in the real interpolation space $(L_2(\mathring{X}),\mathcal{H})_{\frac{1}{2},\infty}$ (see [6] for the definition of interpolation spaces) and

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

Therefore we obtain a continuous embedding

$$J_1 : H_\sigma \to (L_2(\mathring{X}),\mathcal{H})_{\frac{1}{2},\infty}$$

with $\|J_1\| \leq 2\sqrt{c_{\sigma,\mathcal{H}}}$. If in addition a subset inclusion $(L_2(\mathring{X}),\mathcal{H})_{\frac{1}{2},\infty} \subset C(\mathring{X})$ exists which extends to a continuous embedding

$$J_2 : (L_2(\mathring{X}),\mathcal{H})_{\frac{1}{2},\infty} \to C(X)$$

then we have a factorization $J_\sigma = J_2 J_1$ and can conclude

$$\log \mathcal{N}(J_\sigma,\epsilon) \leq \log \mathcal{N}\Big(J_2, \frac{\epsilon}{2\sqrt{c_{\sigma,\mathcal{H}}}}\Big). \tag{49}$$

Consequently to bound $\log \mathcal{N}(J_\sigma,\epsilon)$ we need to select an $\mathcal{H}$, compute $c_{\sigma,\mathcal{H}}$, and bound $\log \mathcal{N}(J_2,\epsilon)$ for the embedding $J_2 : (L_2(\mathring{X}),\mathcal{H})_{\frac{1}{2},\infty} \to C(X)$. To that end let $\mathcal{H}$ be the Sobolev space $\mathcal{H} = W^m(\mathring{X})$ 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

$$\begin{aligned}
\|D^\alpha K_\sigma f\|^2 &= \int_{\mathring{X}} \Big|\int_{\mathring{X}} D_x^\alpha k_\sigma(x,\acute{x}) f(\acute{x}) d\acute{x}\Big|^2 dx \\
&\leq \int_X \Big(\int_X \big|D_x^\alpha k_\sigma(x,\acute{x})\big|^2 d\acute{x} \int_{\mathring{X}} f^2(\acute{x}) d\acute{x}\Big) dx \\
&\leq \|f\|^2 \int_X \int_X \big|D_x^\alpha k_\sigma(x,\acute{x})\big|^2 d\acute{x} dx,
\end{aligned} \tag{50}$$

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

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

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

$$\int_{\mathbb{R}^d} |D_x^\alpha(e^{-|x|^2})|^2 dx = \int_{\mathbb{R}^d} e^{-|x|^2} h_\alpha^2(x) dx \leq \int_{\mathbb{R}^d} h_\alpha^2(x) dx = 2^{|\alpha|}\alpha! \pi^{\frac{d}{2}} \tag{51}$$

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

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

By a change of variables we can apply inequality (51) to the integral on the righthand side

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

Now, using the trivial estimate

$$\int_X\int_X |D_x^\alpha k_\sigma(x,\acute{x})|^2 d\acute{x}dx \leq \int_X\int_{\mathbb{R}^d} |D_x^\alpha k_\sigma(x,\acute{x})|^2 d\acute{x}dx$$

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 $\theta(d) = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$ 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 therefore obtain from (50) that

$$\|K_\sigma\| \leq \left(\frac{2\pi^d}{d\Gamma(\frac{d}{2})}\right)^{\frac{1}{2}}(2d)^{\frac{m}{2}}m!^{\frac{d}{2}}\sigma^{m-\frac{d}{2}}$$

for $\sigma \geq 1$. Therefore we can set $c_{\sigma,\mathcal{H}} = \left(\frac{2\pi^d}{d\Gamma(\frac{d}{2})}\right)^{\frac{1}{2}}(2d)^{\frac{m}{2}}m!^{\frac{d}{2}}\sigma^{m-\frac{d}{2}}$.

Now let us consider $J_2 : (L_2(\mathring{X}), W^m(\mathring{X}))_{\frac{1}{2},\infty} \to C(X)$. According to Triebel [31, p. 267] we have

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

isomorphically and

$$\log\mathcal{N}(B_{2,\infty}^{\frac{m}{2}}(\mathring{X}) \to C(X), \epsilon) \leq c_{m,d}\epsilon^{-\frac{2d}{m}} \tag{52}$$

for $m > d$ follows from a similar result of Birman and Solomyak's ([7], cf. also [31]) for Slobodeckij (fractional Sobolev) spaces, where the constant $c_{m,d}$ depends only on m and d. Consequently we obtain from inequalities (49) and (52) that

$$\begin{aligned}
\log\mathcal{N}(J_\sigma,\epsilon) &\leq c_{m,d}\left(\frac{\epsilon}{2\sqrt{c_{\sigma,\mathcal{H}}}}\right)^{-\frac{2d}{m}} \\
&= c_{m,d}(4c_{\sigma,\mathcal{H}})^{\frac{d}{m}}\epsilon^{-\frac{2d}{m}} \\
&= c_{m,d}\left(\frac{32\pi^d}{d\Gamma(\frac{d}{2})}\right)^{\frac{d}{2m}}(2d)^{\frac{d}{2}}m!^{\frac{d^2}{2m}}\sigma^{d-\frac{d^2}{2m}}\epsilon^{-\frac{2d}{m}} \\
&= \tilde{c}_{m,d}\sigma^{d-\frac{d^2}{2m}}\epsilon^{-\frac{2d}{m}}
\end{aligned}$$

for all $m > d$. Setting $m = \frac{2d}{p}$ finishes the proof of Theorem 9.1. ■

Proof of Theorem 2.15: Since I_σ factors through J_σ and the evaluation map $C(X) \to L_2(T_X)$ and the latter has norm not greater than 1, Theorem 9.1 and the product rule for covering numbers imply that

$$\sup_{T\in Z^n}\log\mathcal{N}(I_\sigma,\varepsilon) \leq c_{q,d}\,\sigma^{(1-\frac{q}{4})d}\varepsilon^{-q} \tag{53}$$

for all $0 < q < 2$. To complete the proof of Theorem 2.15 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 having norm not greater than 1. Hence Proposition 17.3.7 in [21] implies that I_σ is absolutely 2-summing with 2-summing norm not greater than 1. By König's theorem ([22, Lem. 2.7.2]) we obtain for 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

$$\sup_{T\in Z^n} \log \mathcal{N}(I_\sigma, \varepsilon) \;\leq\; \tilde{c}\varepsilon^{-2} \tag{54}$$

for all $\varepsilon > 0$ and all $\sigma > 0$. We now interpolate the bound (54) with the bound (53). Since $\|I_\sigma : H_\sigma \to L_2(T_X)\| \leq 1$ we need only 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. ∎

Let us finally treat Remark 2.7. We have seen in the above proof that we always have

$$\|(a_k(I_H))\|_2 \;\leq\; 1\,.$$

By Carl's inequality we hence find

$$\int_0^\infty \sqrt{\log \mathcal{N}(I_H, \varepsilon)}\,d\varepsilon \;<\infty.$$

Therefore, by the proofs of Lemma 5.7 and Proposition 5.4 we obtain

$$\mathrm{Rad}(\mathcal{G}, n, \varepsilon) \;\leq\; c_p B\sqrt{\frac{a}{n}}\,,$$

where $\mathcal{G}$ is the function class considered in Theorem 5.5. The proof of Theorem 5.5 then shows that the concentration inequality of Theorem 5.5 holds for $p = 2$. Finally, in order to prove Remark 2.7 we have repeat the proofs of Section 7 for $q = 0$ and $p = 2$.

10 L1-SVM's with Gaussian kernels: proof of Theorem 2.16

In this section be prove Theorem 2.16. To this end let us suppose that for all $0 < p < 2$ we can determine constants $c, \gamma > 0$ such that

$$\sup_{T\in Z^n} \log \mathcal{N}(B_{H_\sigma}, \varepsilon, L_2(T_X)) \;\leq\; c\sigma^{\gamma d}\varepsilon^{-p} \tag{55}$$

holds for all $\varepsilon > 0$, $\sigma \geq 1$. Recall, that by Theorem 2.15 we can choose $\gamma := (1 - \frac{p}{2})(1 + \delta)$ for all $\delta > 0$.

As in the previous proofs the approximation properties of H with respect to P play an important role for downsizing the norm of the empirical L1-SVM solutions. This downsizing is again achieved by our shrinking technique.

Lemma 10.1 *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 \le q \le \infty$ and geometric noise exponent $0 < \alpha < \infty$. Furthermore, let us assume that we can bound the covering numbers by (55) for some $0 < \gamma \le 2$ and $0 < p < 2$. Given an $0 \le \varsigma < \frac{1}{5}$ 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 $\frac{1}{2(\alpha+1)} + 4\varsigma < \rho \le \frac{1}{2}$ and $C \ge 1$ such that

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

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

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

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

Proof: For brevity's sake we only prove the lemma for L1-SVM's without offset. Using the idea of the proof of Lemma 7.2 the proof of this lemma for L1-SVM's with offset is analogous. Therefore, let L be defined by (27). Furthermore, let $\hat{f}_{T,\lambda_n}$ be a minimizer of $\mathcal{R}_{L,T}$ on $Cx\lambda_n^{-\rho}B_H$. As in the proof of Lemma 7.1 it suffices to show the existence of a constant $\tilde{C} > 0$ which satisfies

$$\|\hat{f}_{T,\lambda_n}\| \le \tilde{C}x\lambda_n^{-\hat{\rho}} \tag{56}$$

with probability not less than $1 - e^{-x}$.
Let us first treat the case $q > 0$. For brevity's sake we write $\beta := \frac{2(q+1)}{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}$. By Proposition 6.6 and assumption (55) 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 2.14 shows $a_{\sigma_n}(\lambda_n) \preceq \lambda_n^{\frac{\alpha}{\alpha+1}}$ and thus by Proposition 6.6 we may choose

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

In order to apply Theorem 5.5 our first aim is to simplify the expression for $\varepsilon(n,a,B,c,\delta,x)$ given in Remark 6.7. Since the arising terms are quite complex we begin with some preliminary estimates. In order to estimate the first term $B^{\frac{2p(q+1)}{2q+pq+4}}c^{\frac{(2-p)(q+1)}{2q+pq+4}}\left(\frac{a}{n}\right)^{\frac{2(q+1)}{2q+pq+4}}$ we observe

$$B^{\frac{2p(q+1)}{2q+pq+4}}c^{\frac{(2-p)(q+1)}{2q+pq+4}}a^{\frac{2(q+1)}{2q+pq+4}} \sim x\lambda_n^{-\rho-\frac{2\gamma(q+1)}{(\alpha+1)(2q+pq+4)}} \sim x\lambda_n^{-\frac{\rho(\alpha+1)(2q+pq+4)+2\gamma(q+1)}{(\alpha+1)(2q+pq+4)}} \,. \tag{57}$$

The definition of λ_n gives $n = \lambda_n^{-\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{4(\alpha+1)(q+1)}\cdot(1-\varsigma)}$ and therefore, we have

$$\begin{aligned}
B^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \left(\frac{a}{n}\right)^{\frac{2(q+1)}{2q+pq+4}} &\sim x\lambda_n^{-\frac{\rho(\alpha+1)(2q+pq+4)+2\gamma(q+1)}{(\alpha+1)(2q+pq+4)}} n^{-\frac{2(q+1)}{2q+pq+4}} \\
&= x\lambda_n^{-\frac{\rho(\alpha+1)(2q+pq+4)+2\gamma(q+1)}{(\alpha+1)(2q+pq+4)}} \lambda_n^{\frac{q+1}{2q+pq+4}\cdot\frac{1-\varsigma}{(\alpha+1)\beta}} \\
&= x\lambda_n^{-\frac{2\rho(\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} \lambda_n^{\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}(1-\varsigma)} \\
&= \lambda_n^{\frac{2\alpha+1-2\rho(\alpha+1)}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} \\
&= x\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} . \qquad (58)
\end{aligned}$$

Furthermore, parts of the second term $B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}\left(\frac{a}{n}\right)^{\frac{1}{2}}$ of Remark 6.7 can be estimated by

$$\begin{aligned}
B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}a^{\frac{1}{2}} &\preceq x\lambda_n^{-\frac{p\rho}{2}} \lambda_n^{\frac{2-p}{4}\cdot\frac{\alpha q-\rho(q+2)(\alpha+1)}{(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}} \\
&\sim x\lambda_n^{\frac{2\alpha q-2\rho(q+2)(\alpha+1)-\alpha pq+p\rho(q+2)(\alpha+1)-2p\rho(\alpha+1)(q+1)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}} \\
&\sim x\lambda_n^{\frac{2\alpha q-2\rho(q+2)(\alpha+1)-\alpha pq-pq\rho(\alpha+1)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}} \\
&\sim x\lambda_n^{\frac{\alpha q(2-p)-\rho(\alpha+1)(2q+pq+4)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}} .
\end{aligned}$$

Using the definition of λ_n we hence obtain

$$\begin{aligned}
B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}\left(\frac{a}{n}\right)^{\frac{1}{2}} &\preceq x\lambda_n^{\frac{\alpha q(2-p)-\rho(\alpha+1)(2q+pq+4)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}} n^{-\frac{1}{2}} \\
&= x\lambda_n^{\frac{\alpha q(2-p)-\rho(\alpha+1)(2q+pq+4)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{2\gamma(q+1)}{4(\alpha+1)(q+1)}} \lambda_n^{\frac{1-\varsigma}{4(\alpha+1)\beta}} \\
&= x\lambda_n^{\frac{2\alpha q(2-p)-2\rho(\alpha+1)(2q+pq+4)}{8(\alpha+1)(q+1)}} \lambda_n^{-\frac{4\gamma(q+1)}{8(\alpha+1)(q+1)}} \lambda_n^{\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}(1-\varsigma)} \\
&= x\lambda_n^{\frac{2\alpha q(2-p)-2\rho(\alpha+1)(2q+pq+4)}{8(\alpha+1)(q+1)}} \lambda_n^{\frac{(2\alpha+1)(2q+pq+4)}{8(\alpha+1)(q+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}} \\
&= x\lambda_n^{\frac{4\alpha q-2\alpha pq-4\alpha\rho q-2\alpha\rho pq-8\alpha\rho-4\rho q-2\rho pq-8\rho}{8(\alpha+1)(q+1)}} \lambda_n^{\frac{4\alpha q+2\alpha pq+8\alpha+2q+pq+4}{8(\alpha+1)(q+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}} \\
&= x\lambda_n^{\frac{8\alpha q-4\alpha\rho q-2\alpha\rho pq-8\alpha\rho-4\rho q-2\rho pq-8\rho+8\alpha+2q+pq+4}{8(\alpha+1)(q+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}} \\
&= x\lambda_n^{\frac{8\alpha(q+1)-2\rho(\alpha+1)(2q+pq+4)+2q+pq+4}{8(\alpha+1)(q+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}} \\
&= x\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}\cdot\frac{2q+pq+4}{4(q+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}} . \qquad (59)
\end{aligned}$$

Let us compare the first and the second term of the expression for $\varepsilon(n, a, B, c, \delta, x)$ given in Remark 6.7: since $2q + pq + 4 \leq 4(q + 1)$ and $2\rho(\alpha + 1) - 1 > 0$ we have $\frac{2\rho(\alpha+1)-1}{2(\alpha+1)} \cdot \frac{2q+pq+4}{4(q+1)} \leq \frac{2\rho(\alpha+1)-1}{2(\alpha+1)}$ and $\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)} \leq \frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}$. This shows

$$\lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}\cdot\frac{2q+pq+4}{4(q+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{8(\alpha+1)(q+1)}} \leq \lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} , \qquad (60)$$

and therefore (58) and (59) implies $B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}\left(\frac{a}{n}\right)^{\frac{1}{2}} \preceq B^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \left(\frac{a}{n}\right)^{\frac{2(q+1)}{2q+pq+4}}$, i.e. the first term dominates the second term. Let us now treat the third term $B\left(\frac{a}{n}\right)^{\frac{2}{2+p}}$ in Remark 6.7. Since

$$Ba^{\frac{2}{2+p}} \sim x\lambda_n^{-\rho}\lambda_n^{-\frac{\gamma}{\alpha+1}\cdot\frac{2}{2+p}} \sim x\lambda_n^{-\frac{\rho(\alpha+1)(2+p)+2\gamma}{(\alpha+1)(2+p)}}$$

we find

$$\begin{aligned}
B\Big(\frac{a}{n}\Big)^{\frac{2}{2+p}} &\sim x\lambda_n^{-\frac{\rho(\alpha+1)(2+p)+2\gamma}{(\alpha+1)(2+p)}} n^{-\frac{2}{2+p}} \\
&= x\lambda_n^{-\frac{\rho(\alpha+1)(2+p)+2\gamma}{(\alpha+1)(2+p)}} \lambda_n^{\frac{1-\varsigma}{(\alpha+1)(2+p)\beta}} \\
&= x\lambda_n^{-\frac{2\rho(\alpha+1)(2+p)(q+1)+4\gamma(q+1)}{2(\alpha+1)(2+p)(q+1)}} \lambda_n^{(1-\varsigma)\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2+p)(q+1)}} \\
&= x\lambda_n^{\frac{(1-\varsigma)(2\alpha+1)(2q+pq+4)-4\varsigma\gamma(q+1)-2\rho(\alpha+1)(2+p)(q+1)}{2(\alpha+1)(2+p)(q+1)}} .
\end{aligned}$$

Since $\varsigma \leq \frac{1}{5}$ we have $(1-\varsigma)(2q+4)-8\varsigma(q+1) > 0$ and hence $(1-\varsigma)(2\alpha+1)(2q+pq+4)-4\varsigma\gamma(q+1) > 0$. Therefore, we obtain

$$\begin{aligned}
\lambda_n^{-\frac{\rho(\alpha+1)(2+p)+2\gamma}{(\alpha+1)(2+p)}} n^{-\frac{2}{2+p}} &= \lambda_n^{\frac{(1-\varsigma)(2\alpha+1)(2q+pq+4)-4\varsigma\gamma(q+1)-2\rho(\alpha+1)(2+p)(q+1)}{2(\alpha+1)(2+p)(q+1)}} \\
&\leq \lambda_n^{\frac{(1-\varsigma)(2\alpha+1)(2q+pq+4)-4\varsigma\gamma(q+1)-2\rho(\alpha+1)(2q+pq+4)}{2(\alpha+1)(2q+pq+4)}} \\
&= \lambda_n^{\frac{\alpha}{\alpha+1}-\frac{2\rho(\alpha+1)-1}{2(\alpha+1)}-\varsigma\cdot\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{2(\alpha+1)(2q+pq+4)}} .
\end{aligned}$$

Using (58) and $p < 2$ this shows $B\big(\frac{a}{n}\big)^{\frac{2}{2+p}} \preceq B^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \big(\frac{a}{n}\big)^{\frac{2(q+1)}{2q+pq+4}}$, i.e. the first term dominates the third term. Furthermore, for the fourth term $\sqrt{\frac{\delta x}{n}}$ we obtain

$$\begin{aligned}
\sqrt{\delta x} \preceq x^2\lambda_n^{\frac{\alpha q-\rho(q+2)(\alpha+1)}{2(\alpha+1)(q+1)}} &\preceq x^2\lambda_n^{\frac{2\alpha q-2\rho(q+2)(\alpha+1)-\alpha pq-pq\rho(\alpha+1)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}} \\
&\sim x^2\lambda_n^{\frac{\alpha q(2-p)-\rho(\alpha+1)(2q+pq+4)}{4(\alpha+1)(q+1)}} \lambda_n^{-\frac{\gamma}{2(\alpha+1)}}
\end{aligned}$$

by a crude estimate. As in (59) and (60) we hence see that the fourth term is dominated by the first term combined with an additional factor x, i.e. $\sqrt{\frac{\delta x}{n}} \preceq xB^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \big(\frac{a}{n}\big)^{\frac{2(q+1)}{2q+pq+4}}$. Moreover, parts of the fifth term $\big(\frac{cx}{n}\big)^{\frac{q+1}{q+2}}$ become

$$c^{\frac{q+1}{q+2}} \sim x\lambda_n^{-\rho} \preceq x\lambda_n^{-\frac{\rho(\alpha+1)(2q+pq+4)+2\gamma(q+1)}{(\alpha+1)(2q+pq+4)}} ,$$

which shows $\big(\frac{cx}{n}\big)^{\frac{q+1}{q+2}} \preceq xB^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \big(\frac{a}{n}\big)^{\frac{2(q+1)}{2q+pq+4}}$ by (57), (58), and $\frac{q+1}{q+2} > \frac{2(q+1)}{2q+pq+4}$. Finally, the sixth term $\frac{Bx}{n}$ is obviously dominated by the third term in the sense of $\frac{Bx}{n} \preceq Bx\big(\frac{a}{n}\big)^{\frac{2}{2+p}}$ and thus $\frac{Bx}{n} \preceq xB^{\frac{2p(q+1)}{2q+pq+4}} c^{\frac{(2-p)(q+1)}{2q+pq+4}} \big(\frac{a}{n}\big)^{\frac{2(q+1)}{2q+pq+4}}$. Putting the above considerations together Remark 6.7 gives us

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

By Theorem 5.5 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

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

holds with probability not less than $1-e^{-x}$. Now, it is easy to see that $\lambda\|f_{P,\lambda}\|^2 \leq a_{\sigma_n}(\lambda_n) \preceq \lambda_n^{\frac{\alpha}{\alpha+1}}$ yields $\|f_{P,\lambda_n}\| \preceq \lambda_n^{-\frac{1}{2(\alpha+1)}}$. Since $\rho > \frac{1}{2(\alpha+1)}$ this 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 $f_{P,\lambda_n} = \hat{f}_{P,\lambda_n}$ as in the previous proofs. Consequently, with probability not less than $1 - e^{-x}$ we have

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

and hence

$$\|\hat{f}_{T,\lambda_n}\| \leq \tilde{C}_3 x \lambda_n^{\frac{\alpha}{2(\alpha+1)} - \frac{2\rho(\alpha+1)-1}{4(\alpha+1)} - \frac{1}{2} - 2\varsigma} = \tilde{C}_3 x \lambda_n^{-\frac{1}{4(\alpha+1)} - \frac{\rho}{2} - 2\varsigma} = \tilde{C}_3 x \lambda_n^{-\hat{\rho}}\,.$$

Let us now prove the assertion for $q = 0$. By Proposition 6.3 and assumption (55) we observe that we may choose B, a and c such that

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

In order to apply Theorem 5.5 our first aim is to simplify the expression for $\varepsilon(n, a, B, c, \delta, x)$ given in Remark 6.4. Using $n = \lambda_n^{-\frac{(2\alpha+1)(2q+pq+4)+4\gamma(q+1)}{4(\alpha+1)(q+1)}\cdot(1-\varsigma)}$ the first term can be bounded by

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

Furthermore, the second term can be estimated by

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

Finally, the third term can be bounded by

$$\frac{cx}{n} \sim x\lambda_n^{-1} \lambda_n^{\frac{2\alpha+1+\gamma}{\alpha+1} - \varsigma\cdot\frac{2\alpha+1+\gamma}{\alpha+1}} \preceq x \lambda_n^{\frac{\alpha+\gamma}{\alpha+1} - 4\varsigma} \preceq x^2 \lambda_n^{\frac{2\alpha-2\alpha p\rho-2p\rho+\alpha p+p}{(2+p)(\alpha+1)} - 4\varsigma}\,,$$

where in the last step we used $\rho > \frac{1}{2(\alpha+1)}$. Putting the above considerations together Remark 6.4 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 5.5 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

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

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 as for $q > 0$. ■

The next theorem establishes almost the result of Theorem 2.16. We present this intermediate result because it clarifies the impact of covering number bounds of the form (55) on our rates.

Theorem 10.2 *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $0 \leq q \leq \infty$ and geometric noise exponent $0 < \alpha < \infty$. Finally, let us assume that we can bound the covering numbers by (55) 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-SVM's with offset.

Proof: Since the proof is very similar to the proof of Theorem 2.4 we only sketch it. Iteratively using Lemma 10.1 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 10.1 (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

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

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 2.16 recall that by Theorem 2.15 we can choose $\gamma := (1-\frac{p}{2})(1+\delta)$ for all $\delta > 0$. The idea of the proof of Theorem 2.16 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 2.16: Let us first consider the case $\alpha \leq \frac{q+2}{2q}$. Our aim is to apply Theorem 10.2. 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 10.1 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

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

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 2.14 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. It is easy to see that we are hence in the situation of Theorem 2.4 for "$\beta = 1$" (in the sense of Remark 2.9) and p arbitrarily close to 0. ∎

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. `http://www.stat.berkeley.edu/~bartlett/papers/bbm-lrc-02b.pdf`, 2002.

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

[7] 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.

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

[9] P. L. Butzer and H. Berens. *Semi-groups of operators and approximation.* Springer-Verlag, New York, 1967.

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

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

[14] 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.

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

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

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

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

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

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

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

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

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

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

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

[26] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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

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

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

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

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

[32] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32, 2004.

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

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

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

[36] T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32, 2004.

11 Appendix

In this section we prove the theorems of Section 3. We begin by deriving some elementary but useful properties of the approximation error function. In particular we are interested in the question how the approximation error function influences the map $\lambda \mapsto \|f_{P,\lambda}\|$.

Recall that Lemma 3.1 shows that $f^*_{P,\lambda}$ is well defined. We now compare $f^*_{P,\lambda}$ and $f_{P,\lambda}$ in terms of their L risk and their norm:

Lemma 11.1 *For $\lambda > 0$ we have $\mathcal{R}_{L,P}(f^*_{P,\lambda}) \leq \mathcal{R}_{L,P}(f_{P,\lambda})$ and $\|f_{P,\lambda}\| \leq \|f^*_{P,\lambda}\|$.*

Proof: The first assertion follows from $\|f_{P,\lambda}\|^2 \leq 1/\lambda$. Then using the first assertion we find

$$\lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f^*_{P,\lambda}) \leq \lambda\|f_{P,\lambda}\|^2_H + \mathcal{R}_{L,P}(f_{P,\lambda}) \leq \lambda\|f^*_{P,\lambda}\|^2_H + \mathcal{R}_{L,P}(f^*_{P,\lambda}),$$

i.e. $\lambda\|f_{P,\lambda}\|^2 \leq \lambda\|f^*_{P,\lambda}\|^2$. ■

In the following we say that $f \in H$ *minimizes the L-risk in H* if $\mathcal{R}_{L,P}(f) = \mathcal{R}_{L,P,H}$. If no confusion can occur we denote such functions by $f_{L,P,H}$. The next lemmas describe the situation in which such a minimizer exists. We begin with a simple observation:

Lemma 11.2 *Assume that there is a minimizer $f_{L,P,H} \in H$. Then there exists a unique element $f^*_{L,P,H} \in H$ minimizing the L-risk in H with $\|f^*_{L,P,H}\| \leq \|f\|$ for all $f \in H$ minimizing the L risk in H. Furthermore, we have $\|f_{P,\lambda}\| \leq \|f^*_{L,P,H}\|$ for all $\lambda > 0$.*

Proof: The first assertion is a direct consequence of Lemma 3.1 for $\lambda = 1/\|f_{L,P}\|^2$ and the second assertion follows from Lemma 11.1. ■

The next lemma shows that for $\lambda \to 0$ through positive values the infinite sample versions $f_{P,\lambda}$ converge to $f^*_{L,P,H} \in H$ whenever the latter exists. If H is a universal kernel, i.e. it is dense in $C(X)$, P is an empirical distribution based on a training set T, and L is the (squared) hinge loss function then $f^*_{L,T,H} \in H$ coincides with the *hard margin* SVM solution. This shows that both the L1-SVM and the L2-SVM solutions $f_{T,\lambda}$ converge to the hard margin solution if T is *fixed* and $\lambda \to 0$.

Lemma 11.3 *Assume that there is a minimizer $f_{L,P,H} \in H$. Then for all positive sequences $\lambda_n \to 0$ we have $f_{P,\lambda_n} \to f^*_{L,P}$ with respect to the norm of H.*

Proof: By Lemma 11.2 we have $\|f_{P,\lambda_n}\| \leq \|f^*_{L,P,H}\|$ and thus there exists an $f^* \in H$ and a subsequence $(f_{P,\lambda_{n_i}})$ with $f_{P,\lambda_{n_i}} \to f^*$ weakly. This implies $\mathcal{R}_{L,P}(f_{P,\lambda_{n_i}}) \to \mathcal{R}_{L,P}(f^*)$. Furthermore, we always have $\lambda_{n_i}\|f_{P,\lambda_{n_i}}\|^2 \to 0$ and thus

$$\mathcal{R}_{L,P,H} = \lim_{i\to\infty} \lambda_{n_i}\|f_{P,\lambda_{n_i}}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_{n_i}}) = \mathcal{R}_{L,P}(f^*)\,.$$

Here, the first equality can be shown as in [28] for universal kernels. In other words f^* minimizes the L-risk in H. Hence by Lemma 11.2 we find

$$\|f_{P,\lambda_{n_i}}\| \leq \|f^*\| \leq \liminf_{i\to\infty} \|f_{P,\lambda_{n_i}}\|\,,$$

i.e. $\|f_{P,\lambda_{n_i}}\| \to \|f^*\|$. This yields

$$\|f_{P,\lambda_{n_i}} - f^*\|^2 \;=\; \|f_{P,\lambda_{n_i}}\|^2 - 2\langle f_{P,\lambda_{n_i}}, f^*\rangle + \|f^*\|^2 \;\to\; \|f^*\|^2 - 2\|f^*\|^2 + \|f^*\|^2 \;=\; 0\,.$$

Furthermore, $\|f_{P,\lambda_{n_i}}\| \to \|f^*\|$ together with $\|f_{P,\lambda_{n_i}}\| \leq \|f^*_{L,P,H}\|$ implies $\|f^*\| \leq \|f^*_{L,P,H}\|$, i.e. $f^* = f^*_{L,P,H}$ by Lemma 11.2. Now assume that $f_{P,\lambda_n} \not\to f^*_{L,P,H}$. Then there exists a $\delta > 0$ and a subsequence $(f_{P,\lambda_{n_j}})$ with $\|f_{P,\lambda_{n_j}} - f^*_{L,P,H}\| > \delta$. On the other hand applying the above reasoning to this subsequence gives a sub-subsequence converging to $f^*_{L,P,H}$ and hence we have found a contradiction. ■

The next lemma characterizes the existence of $f^*_{L,P,H} \in H$ in terms of the function $\lambda \mapsto \|f_{P,\lambda}\|$:

Lemma 11.4 *There exists an element $f_{L,P,H} \in H$ minimizing the L-risk in H if and only if there exists a constant $c > 0$ with $\|f_{P,\lambda}\| \leq c$ for all $\lambda > 0$.*

Proof: If there exists an element $f_{L,P} \in H$ minimizing the L-risk we can set $c := \|f^*_{L,P}\|$. On the other hand if $\|f_{P,\lambda}\| \leq c$ for some $c > 0$ and all $\lambda > 0$ there exists an $f^* \in H$ and a sequence (f_{P,λ_n}) with $f_{P,\lambda_n} \to f^*$ weakly. As in the first part of the proof of Lemma 11.3 we easily see that f^* minimizes the L-risk in H. ■

The following lemma which shows that $f_{P,\lambda}$ is a solution of (20) for a suitably chosen size of the underlying ball is somewhat well known:

Lemma 11.5 *When $\gamma := 1/\|f_{P,\lambda}\|^2$, we have $f^*_{P,\gamma} = f_{P,\lambda}$.*

Proof: We first show that $f_{P,\lambda}$ minimizes (20) for regularization parameter γ. Assume that this does not hold. Then we have

$$\mathcal{R}_{L,P}(f^*_{P,\gamma}) \;<\; \mathcal{R}_{L,P}(f_{P,\lambda})\,.$$

Since we also have $\|f^*_{P,\gamma}\| \leq 1/\sqrt{\gamma} = \|f_{P,\lambda}\|$ we find

$$\lambda\|f^*_{P,\gamma}\|^2 + \mathcal{R}_{L,P}(f^*_{P,\gamma}) \;<\; \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda})$$

which contradicts the definition of $f_{P,\lambda}$. Hence $f_{P,\lambda}$ minimizes (20) for regularization parameter γ. Now assume that $f_{P,\lambda} \neq f^*_{P,\gamma}$, i.e. $\|f_{P,\lambda}\| > \|f^*_{P,\gamma}\|$. Since $\mathcal{R}_{L,P}(f^*_{P,\gamma}) = \mathcal{R}_{L,P}(f_{P,\lambda})$ we then have

$$\lambda\|f^*_{P,\gamma}\|^2 + \mathcal{R}_{L,P}(f^*_{P,\gamma}) \;<\; \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda})$$

which again contradicts the definition of $f_{P,\lambda}$. ■

The next lemma compares the size of the norms of solutions for different λ with the corresponding L-risks:

Lemma 11.6 *For all $\lambda_1, \lambda_2 > 0$ we have*

$$\|f_{P,\lambda_1}\| \geq \|f_{P,\lambda_2}\| \qquad \textit{if and only if} \qquad \mathcal{R}_{L,P}(f_{P,\lambda_1}) \leq \mathcal{R}_{L,P}(f_{P,\lambda_2})\,.$$

Proof: Assume that $\|f_{P,\lambda_1}\| \geq \|f_{P,\lambda_2}\|$ but $\mathcal{R}_{L,P}(f_{P,\lambda_1}) > \mathcal{R}_{L,P}(f_{P,\lambda_2})$. Then we find

$$\lambda_1\|f_{P,\lambda_2}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_2}) < \lambda_1\|f_{P,\lambda_1}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_1})$$

which contradicts the definition of f_{P,λ_1}. Analogously, if $\mathcal{R}_{L,P}(f_{P,\lambda_1}) \leq \mathcal{R}_{L,P}(f_{P,\lambda_2})$ but $\|f_{P,\lambda_1}\| < \|f_{P,\lambda_2}\|$ we find

$$\lambda_2\|f_{P,\lambda_1}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_1}) < \lambda_2\|f_{P,\lambda_2}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_2})$$

which contradicts the definition of f_{P,λ_2}. ■

Proof of Theorem 3.2: It is clear from the definitions (21) that $A(0) = A^*(0) = 0$ and $A^*(.)$ is increasing. Since $A(.)$ is an infimum over a family of concave (linear) increasing functions of λ it follows that $A(.)$ is also concave and increasing. Consequently Theorem 10.1 in [26] on the continuity of concave functions implies that $A(.)$ is continuous for $\lambda > 0$. Continuity at 0 follows from the proof of Proposition III.3 in [28] completing the proof of the first assertion. To prove the second assertion, observe that Lemma 11.1 implies $A^*(\lambda) \leq A(\lambda)$ for all $\lambda > 0$ and since $A(0) = A^*(0)$ we obtain $A^*(\lambda) \leq A(\lambda)$ for all $\lambda \geq 0$. Now let $\varepsilon := h(\lambda)$ and $\tilde\lambda := \varepsilon\|f^*_{P,\lambda}\|^{-2}$. Then we find

$$\tilde\lambda\|f_{P,\tilde\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\tilde\lambda}) \leq \tilde\lambda\|f^*_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f^*_{P,\lambda}) = \tilde\lambda\|f^*_{P,\lambda}\|^2 + \mathcal{R}_{L,P,H} + \varepsilon \leq \mathcal{R}_{L,P,H} + 2\varepsilon\,.$$

This shows $A(\tilde\lambda) \leq 2h(\lambda)$. Furthermore we have $\lambda h(\lambda) \leq \varepsilon\|f^*_{P,\lambda}\|^{-2} = \tilde\lambda$ and thus the assertion follows since $A(.)$ is an increasing function. ■

Proof of Theorem 3.3: If $\lambda \mapsto \|f_{P,\lambda}\|$ is bounded on $(0,\infty)$ there exists an $f_{L,P,H} \in H$ minimizing the L-risk in H by Lemma 11.4. This yields

$$A(\lambda) \;=\; \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,H} \;\leq\; \lambda\|f^*_{L,P,H}\|^2 + \mathcal{R}_{L,P}(f^*_{L,P,H}) - \mathcal{R}_{L,P,H} \;=\; \lambda\|f^*_{L,P,H}\|^2\,.$$

Conversely, if there exists a constant $C > 0$ with $A(\lambda) \leq C\lambda$ we find

$$\lambda\|f_{P,\lambda}\|^2 \;\leq\; A(\lambda) \;\leq\; C\lambda$$

which shows $\|f_{P,\lambda}\| \;\leq\; \sqrt{C}$ for all $\lambda > 0$ proving the first assertion.

Now let us assume $A^*(\lambda) \preceq \lambda^\alpha$ for some $\alpha > 0$. Then from Theorem 3.2 we know $A(\lambda^{1+\alpha}) \preceq \lambda^\alpha$ which leads to $A(\lambda) \preceq \lambda^{\frac{\alpha}{\alpha+1}}$. The latter immediately implies $\|f_{P,\lambda}\|^2 \preceq \lambda^{-\frac{1}{\alpha+1}}$. Conversely, if $A(\lambda) \preceq \lambda^{\frac{\alpha}{\alpha+1}}$ we define $\gamma := \|f_{P,\lambda}\|^{-2}$. By Lemma 11.5 we then obtain

$$A^*(\gamma) \;=\; \mathcal{R}_{L,P}(f^*_{P,\gamma}) - \mathcal{R}_{L,P} \;=\; \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P} \;\leq\; c_1\lambda^{\frac{\alpha}{\alpha+1}} \;\leq\; c_2\|f_{P,\lambda}\|^{-2\alpha} \;=\; \gamma^\alpha$$

for some constants $c_1, c_2 > 0$ independent of γ. Now, if there is no $f_{L,P,H} \in H$ minimizing the L-risk in H the function $\lambda \mapsto \|f_{P,\lambda}\|^{-2}$ tends to 0 if $\lambda \to 0$ and thus $A^*(\lambda) \preceq \lambda^\alpha$. If there is an $f_{L,P,H} \in H$ minimizing the L-risk in H the assertion is trivial.

For the third assertion recall that Lemma 11.5 states $f_{P,\lambda} = f^*_{P,\gamma}$ with $\gamma := \|f_{P,\lambda}\|^{-2}$ and hence we find

$$A(\lambda) \;=\; \lambda \|f_{P,\lambda}\|^2 + A^*\big(\|f_{P,\lambda}\|^{-2}\big)\,. \tag{61}$$

Furthermore, we have already seen $\|f_{P,\lambda}\|^{-2} \succeq \lambda^{\frac{1}{\alpha+1}}$. By our assumption we hence get

$$\lambda^{\frac{\alpha}{\alpha+1}} \;\succeq\; \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P} \;=\; A^*\big(\|f_{P,\lambda}\|^{-2}\big) \;\succeq\; \|f_{P,\lambda}\|^{-2(\alpha+\varepsilon)} \;\succeq\; \lambda^{\frac{\alpha+\varepsilon}{\alpha+1}}\,.$$

Combining this with (61) yields the third assertion. ■