# Fast Rates for Support Vector Machines

Ingo Steinwart and Clint Scovel

CCS-3, Los Alamos National Laboratory, Los Alamos NM 87545, USA
{ingo, jcs}@lanl.gov

**Abstract.** We establish learning rates to the Bayes risk for support vector machines (SVMs) using a regularization sequence $\lambda_n = n^{-\alpha}$, where $\alpha \in (0,1)$ is arbitrary. Under a noise condition recently proposed by Tsybakov these rates can become faster than $n^{-1/2}$. In order to deal with the approximation error we present a general concept called the approximation error function which describes how well the infinite sample versions of the considered SVMs approximate the data-generating distribution. In addition we discuss in some detail the relation between the "classical" approximation error and the approximation error function. Finally, for distributions satisfying a geometric noise assumption we establish some learning rates when the used RKHS is a Sobolev space.

## 1 Introduction

The goal in binary classification is to predict labels $y \in Y := \{-1, 1\}$ of unseen data points $x \in X$ using a training set $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$. As usual we assume that both the training samples $(x_i, y_i)$ and the new sample $(x, y)$ are i.i.d. drawn from an unknown distribution $P$ on $X \times Y$. Now given a *classifier* $\mathcal{C}$ that assigns to every $T$ a function $f_T : X \to \mathbb{R}$ the prediction of $\mathcal{C}$ for $y$ is $\operatorname{sign} f_T(x)$, where we choose a fixed definition of $\operatorname{sign}(0) \in \{-1, 1\}$. In order to "learn" from $T$ the decision function $f_T : X \to \mathbb{R}$ should guarantee a small probability for the misclassification, i.e. $\operatorname{sign} f_T(x) \neq y$, of the example $(x, y)$. To make this precise the risk of a measurable function $f : X \to \mathbb{R}$ is defined by

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

and the smallest achievable risk $\mathcal{R}_P := \inf\{\mathcal{R}_P(f) \mid f : X \to \mathbb{R} \text{ measurable}\}$ is known as the *Bayes risk* of $P$. A function $f_P$ 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. To make this precise, we say that a classifier is *universally consistent* if

$$\mathbb{E}_{T \sim P^n} \mathcal{R}_P(f_T) - \mathcal{R}_P \;\to\; 0 \qquad \text{for } n \to \infty. \tag{1}$$

Unfortunately, it is well known that no classifier can guarantee a convergence rate in (1) that simultaneously holds for all distributions (see [1–Thm. 7.2]). However, if one restricts considerations to suitable smaller classes of distributions

P. Auer and R. Meir (Eds.): COLT 2005, LNAI 3559, pp. 279–294, 2005.

such rates exist for various classifiers (see e.g. [2, 3, 1]). One interesting feature of these rates is that they are not faster than $n^{-1/2}$ if the considered distributions $P$ are allowed to be noisy in the sense of $\mathcal{R}_P > 0$. On the other hand, if one restricts considerations to noise-free distributions $P$ in the sense of $\mathcal{R}_P = 0$ then some empirical risk minimization (ERM) methods can actually learn with rate $n^{-1}$ (see e.g. [1]). Remarkably, it was only recently discovered (see [4, 5]) that there also exists classes of noisy distributions which can be learned with rates between $n^{-1/2}$ and $n^{-1}$. The key property of these classes is that their noise level $x \mapsto 1/2 - |\eta(x) - 1/2|$ with $\eta(x) := P(y = 1|x)$ is well-behaved in the sense of the following definition.

**Definition 1.** *A distribution $P$ on $X \times Y$ has* Tsybakov noise exponent $q \in [0, \infty]$*, if there exists a $C > 0$ such that for all sufficiently small $t > 0$ we have*

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

Obviously, all distributions have at least noise exponent 0. At the other extreme, (2) is satisfied for $q = \infty$ if and only if the conditional probability $\eta$ is bounded away from the critical level $1/2$. In particular this shows that noise-free distributions have exponent $q = \infty$.

The aim of this work is to establish learning rates for support vector machines (SVMs) under Tsybakov's noise assumption which are comparable to the rates of [4, 5]). Therefore let us now recall these classification algorithms: let $X$ be a compact metric space and $H$ be a RKHS over $X$ with continuous kernel $k$. Furthermore, let $l : Y \times \mathbb{R} \to [0, \infty)$ be the hinge loss which is defined by $l(y, t) := \max\{0, 1 - yt\}$. Then given a training set $T \in (X \times Y)^n$ and a regularization parameter $\lambda > 0$ SVMs solve the optimization problems

$$(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda}) := \arg\min_{\substack{f \in H \\ b \in \mathbb{R}}} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} l\big(y_i, f(x_i) + b\big) \,, \tag{3}$$

or

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

respectively. Furthermore, in order to control the size of the offset we always choose $\tilde{b}_{T,\lambda} := y^*$ if all samples of $T$ have label $y^*$. As usual we call algorithms solving (3) *L1-SVMs with offset* and algorithms solving (4) *L1-SVMs without offset*. For more information on these methods we refer to [6].

The rest of this work is organized as follows: In Section 2 we first introduce two concepts which describe the richness of RKHSs. We then present our main result and discuss it. The following sections are devoted to the proof of this result: In Section 3 we recall some results from [7] which are used for the analysis of the estimation error, and in Section 4 we then prove our main result. Finally, the relation between the approximation error and infinite sample SVMs which is of its own interest is discussed in the appendix.

## 2 Definitions and Results

For the formulation of our results we need two notions which deal with the richness of RKHSs. While the first notion is a complexity measure in terms of covering numbers which is used to bound the estimation error, the second one describes the approximation properties of RKHSs with respect to distributions.

In order to introduce the complexity measure let us recall that for a Banach space $E$ with closed unit ball $B_E$, the *covering numbers* of $A \subset E$ are defined by

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

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

$$\|f\|_{L_2(T)} := \left( \frac{1}{n} \sum_{i=1}^{n} \bigl| f(x_i, y_i) \bigr|^2 \right)^{\frac{1}{2}} \tag{5}$$

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 representative 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 $x_1, \ldots, x_n$. Now our complexity measure is:

**Definition 2.** *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 $c > 0$ such that for all $\varepsilon > 0$ we have*

$$\sup_{T_X \in X^n} \log \mathcal{N}\bigl(B_H, \varepsilon, L_2(T_X)\bigr) \;\leq\; c\varepsilon^{-p}\,.$$

By using the theory of absolutely 2-summing operators one can show that every RKHS has complexity exponent $p = 2$. However, for meaningful rates we need complexity exponents which are strictly smaller than 2.

In order to introduce the second notion describing the approximation properties of RKHSs we first have to recall the infinite sample versions of (3) and (4). To this end let $l$ be the hinge loss function and $P$ be a distribution on $X \times Y$. Then for $f : X \to \mathbb{R}$ the *$l$-risk* of $f$ is defined by $\mathcal{R}_{l,P}(f) := \mathbb{E}_{(x,y)\sim P} l(y, f(x))$. Now given a RKHS $H$ over $X$ and $\lambda > 0$ we define

$$(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda}) := \arg\min_{\substack{f \in H \\ b \in \mathbb{R}}} \Bigl( \lambda \|f\|_H^2 + \mathcal{R}_{l,P}(f + b) \Bigr) \tag{6}$$

and

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

(see [8] for the existence of these minimizers). Note that these definitions give the solutions $(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda})$ and $f_{T,\lambda}$ of (3) and (4), respectively, if $P$ is an empirical

distribution with respect to a training set $T$. In this case we write $\mathcal{R}_{l,T}(f)$ for the (empirical) $l$-risk.

With these notations in mind we define the *approximation error function* by

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

where $\mathcal{R}_{l,P} := \inf\{\mathcal{R}_{l,P}(f) \mid f : X \to \mathbb{R}\}$ denotes the smallest possible $l$-risk. Note that since the obvious variant of $a(.)$ that involves an *offset* is not greater than the above approximation error function, we restrict our attention to the latter. Furthermore, we discuss the relationship between $a(.)$ and the standard approximation error in the appendix.

The approximation error function quantifies how well an infinite sample L1-SVM with RKHS $H$ approximates the minimal $l$-risk. It was shown in [8] that if $H$ is dense in the space of continuous functions $C(X)$ then for *all* $P$ we have $a(\lambda) \to 0$ if $\lambda \to 0$. However, in non-trivial situations no rate of convergence which uniformly holds for all distributions $P$ is possible. The following definition characterizes distributions which guarantee certain polynomial rates:

**Definition 3.** *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 for all $\lambda > 0$ we have*

$$a(\lambda) \leq C\lambda^{\beta}\,.$$

Note, that $H$ approximates every distribution $P$ with exponent $\beta = 0$. We will see in the appendix that the other extremal case $\beta = 1$ is equivalent to the fact that the minimal $l$-risk can be achieved by an element $f_{l,P} \in H$.

With the help of the above notations we can now formulate our main result.

**Theorem 1.** *Let $H$ be a RKHS of a continuous kernel on a compact metric space $X$ 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 $0 < \beta \leq 1$. We define $\lambda_n := n^{-\alpha}$ for some $\alpha \in (0,1)$ and all $n \geq 1$. If $\alpha < \frac{4(q+1)}{(2q+pq+4)(1+\beta)}$ then there exists a $C > 0$ with*

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

*for all $n \geq 1$ and $x \geq 1$. Here $\Pr^*$ is the outer probability of $P^n$ in order to avoid measurability considerations. Furthermore, if $\alpha \geq \frac{4(q+1)}{(2q+pq+4)(1+\beta)}$ then for all $\varepsilon > 0$ there is a $C > 0$ such that for all $x \geq 1$, $n \geq 1$ we have*

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

*Finally, the same results hold for the L1-SVM with offset whenever $q > 0$.*

*Remark 1.* The best rates Theorem 1 can guarantee are (up to an $\varepsilon$) of the form

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

and an easy calculation shows that these rates are obtained for the value $\alpha := \frac{4(q+1)}{(2q+pq+4)(1+\beta)}$. This result has already been announced in [9] and presented in an earlier (and substantially longer) version of [7]. The main difference of Theorem 1 to its predecessors is that it does not require to choose $\alpha$ optimally. Finally note that unfortunately the optimal $\alpha$ is in terms of both $q$ and $\beta$, which are in general not accessible. At the moment we are not aware of any method which can adaptively find the (almost) optimal values for $\alpha$.

*Remark 2.* In [5] it is assumed that a Bayes classifier is contained in the base function classes the considered ERM method minimizes over. This assumption corresponds to a perfect approximation of $P$ by $H$, i.e. $\beta = 1$, as we will see in the apppendix. If in this case we rescale the complexity exponent $p$ from $(0, 2)$ to $(0, 1)$ and write $p'$ for the new complexity measure our optimal rate essentially becomes $n^{-\frac{q+1}{q+p'q+2}}$. Recall that this is exactly the *form* of Tsybakov's result in [5] which is known to be optimal in a minmax sense for some specific classes of distributions. However, as far as we know our complexity measure cannot be compared to Tsybakov's and thus the above reasoning *only indicates* that our optimal rates may be optimal in a minmax sense.

Let us finally present an example which shows how the developed theory can be used to establish learning rates for specific types of kernels and distributions.

*Example 1 (SVMs using Sobolev spaces).* Let $X \subset \mathbb{R}^d$ be the closed unit Euclidian ball, $\Omega$ be the centered open ball of radius 3, and $W^m(\Omega)$ be the Sobolev space of order $m \in \mathbb{N}$ over $\Omega$. Recall that $W^m(\Omega)$ is a RKHS of a continuous kernel if $m > d/2$ (see e.g. [10]). Let us write $H_m := \{f_{|X} : f \in W^m(\Omega)\}$ for the restriction of $W^m(\Omega)$ onto $X$ endowed with the induced RKHS norm. Then (see again [10]) the RKHS $H_m$ has complexity exponent $p := d/m$ if $m > d/2$.

Now let $P$ be a distribution on $X \times Y$ which has geometric noise exponent $\alpha \in (0, \infty]$ in the sense of [7], and let $k_\sigma(x, x') := \exp(-\sigma^2\|x-x'\|)$, $x, x' \in \Omega$, be a Gaussian RBF kernel with associated integral operator $T_\sigma : L_2(\Omega) \to L_2(\Omega)$, where $L_2(\Omega)$ is with respect to the Lebesgue measure. Then by the results in [7–Secs. 3 & 4] there exist constants $c_d, c_{\alpha,m,d} \geq 1$ such that for all $\sigma > 0$ there exists an $f_\sigma \in L_2(\Omega)$ with $\|f_\sigma\|_{L_2(\Omega)} = c_d\sigma^d$, $\mathcal{R}_{l,P}((T_\sigma f_\sigma)_{|X}) - \mathcal{R}_{l,P} \leq c_{\alpha,m,d}\sigma^{-\alpha d}$, and $\|(T_\sigma f_\sigma)_{|X}\|_{H_m} \leq c_{\alpha,m,d}\sigma^{m-d/2}\|f_\sigma\|_{L_2(\Omega)}$. This yields a constant $c > 0$ with

$$a(\lambda) \leq c\big(\lambda\sigma^{2m+d} + \sigma^{-\alpha d}\big)$$

for all $\sigma > 0$ and all $\lambda > 0$. Minimizing with respect to $\sigma$ then shows that $H_m$ approximates $P$ with exponent $\beta := \frac{\alpha d}{(\alpha+1)d+2m}$. Consequently we can use Theorem 1 to obtain learning rates for SVMs using $H_m$ for $m > d/2$. In particular the resulting optimal rates in the sense of Remark 1 are (essentially) of the form

$$n^{-\frac{4\alpha dm(q+1)}{(2mq+dq+4m)(2\alpha d+d+2m)}}.$$

## 3 Prerequisites

In this section we recall some important notions and results that we require in the proof of our main theorem. To this end let $H$ be a RKHS over $X$ that has a continuous kernel $k$. Then recall that every $f \in H$ is continuous and satisfies

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

where we use

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

The rest of this section recalls some results from [7] which will be used to bound the estimation error of L1-SVMs. Before we state these results we have to recall some notation from [7]: let $\mathcal{F}$ be a class of bounded measurable functions from a set $Z$ to $\mathbb{R}$, and let $L : \mathcal{F} \times Z \to [0,\infty)$ be a function. We call $L$ a *loss function* if $L \circ f := L(f,.)$ is measurable for all $f \in \mathcal{F}$. 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 $[0,1]$. Note that if $\mathcal{F}$ is a vector space then every convex $L$ is line-continuous. Now, given a probability measure $P$ on $Z$ we denote by $f_{P,\mathcal{F}} \in \mathcal{F}$ a minimizer of the $L$-risk

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

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. Also note that although there may exist multiple solutions we use a single symbol for them whenever no confusion regarding the non-uniqueness of this symbol can be expected. Furthermore, an algorithm that produces solutions $f_{T,\mathcal{F}}$ for all possible $T$ is called an *empirical $L$-risk minimizer*.

Now the main result of this section, shown in [7], reads as follows:

**Theorem 2.** *Let $\mathcal{F}$ be a convex set of bounded measurable functions from $Z$ to $\mathbb{R}$ and let $L : \mathcal{F} \times Z \to [0,\infty)$ be a convex and line-continuous loss function. For a probability measure $P$ on $Z$ we define*

$$\mathcal{G} \;:=\; \big\{ L\circ f - L\circ f_{P,\mathcal{F}} \;:\; f \in \mathcal{F} \big\}. \tag{9}$$

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

$$\sup_{T\in Z^n} \log \mathcal{N}\big(B^{-1}\mathcal{G}, \varepsilon, L_2(T)\big) \;\leq\; a\varepsilon^{-p} \tag{10}$$

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

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

*where*

$$\varepsilon(n,B,c,\delta,x) := B^{\frac{2p}{4-2\alpha+\alpha p}} c^{\frac{2-p}{4-2\alpha+\alpha p}} n^{-\frac{2}{4-2\alpha+\alpha p}} + B^{\frac{p}{2}} \delta^{\frac{2-p}{4}} n^{-\frac{1}{2}} + Bn^{-\frac{2}{2+p}} + \Big(\frac{\delta x}{n}\Big)^{\frac{1}{2}} + \Big(\frac{cx}{n}\Big)^{\frac{1}{2-\alpha}} + \frac{Bx}{n}\,.$$

Let us now recall some variance bounds of the form $\mathbb{E}_P g^2 \le c\,(\mathbb{E}_P g)^\alpha + \delta$ for SVMs proved in [7]. To this end let $H$ be a RKHS of a continuous kernel over $X$, $\lambda > 0$, and $l$ be the hinge loss function. We define

$$L(f,x,y) \;:=\; \lambda\|f\|_H^2 + l\big(y, f(x)\big) \tag{11}$$

and

$$L(f,b,x,y) \;:=\; \lambda\|f\|_H^2 + l\big(y, f(x)+b\big) \tag{12}$$

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-SVMs actually implement an empirical $L$-risk minimization in the sense of Theorem 2. Now the first variance bound from [7] does not require any assumptions on $P$.

**Proposition 1.** *Let $0 < \lambda < 1$, $H$ be a RKHS over $X$, and $\mathcal{F} \subset \lambda^{-\frac{1}{2}} B_H$. Furthermore, let $L$ be defined by (11), $P$ be a probability measure and $\mathcal{G}$ be defined as in (9). Then for all $g \in \mathcal{G}$ we have*

$$\mathbb{E}_P g^2 \;\le\; 2\lambda^{-1}(2+K)^2 \mathbb{E}_P g\,.$$

Finally, the following variance bound from [7] shows that the previous bound can be improved if one assumes a non-trivial Tsybakov exponent for $P$.

**Proposition 2.** *Let $P$ be a distribution on $X \times Y$ with Tsybakov noise exponent $0 < q \le \infty$. Then there exists a constant $C > 0$ such that for all $\lambda > 0$, all $0 < r \le \lambda^{-1/2}$ satisfying $\tilde{f}_{P,\lambda} \in rB_H$, all $f \in rB_H$, and all $b \in \mathbb{R}$ with $|b| \le Kr + 1$ we have*

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

*Furthermore, the same result holds for SVMs without offset.*

## 4 Proof of Theorem 1

In this section we prove Theorem 1. To this end we write $f(x) \preceq g(x)$ for two functions $f, g : D \to [0,\infty)$, $D \subset (0,\infty)$, if there exists a constant $C > 0$ such that $f(x) \le Cg(x)$ holds over some range of $x$ which usually is implicitly defined by the context. However for sequences this range is always $\mathbb{N}$. Finally we write $f(x) \sim g(x)$ if both $f(x) \preceq g(x)$ and $g(x) \preceq f(x)$ for the same range.

Since our variance bounds have different forms for the cases $q = 0$ and $q > 0$ we have to prove the theorem for these cases separately. We begin with the case $q = 0$ and an important lemma which describes a "shrinking technique".

**Lemma 1.** *Let $H$ and $P$ be as in Theorem 1. For $\gamma > -\beta$ we define $\lambda_n := n^{-\frac{1}{1+\beta+\gamma}}$. Now assume that there are constants $0 \le \rho < \beta$ and $C \ge 1$ such that*

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

*for all $n \ge 1$, $x \ge 1$. Then there is another constant $\hat{C} \ge 1$ such that for $\hat{\rho} := \min\{\beta, \frac{\rho+\beta+\gamma}{2}, \beta+\gamma\}$ 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 (11). By our assumption we have $\hat{f}_{T,\lambda_n} = f_{T,\lambda_n}$ with probability not less than $1 - e^{-x}$ since $f_{T,\lambda_n}$ is unique for every training set $T$ by the strict convexity of $L$. We will show that for some $\tilde{C} > 0$ and all $n \ge 1$, $x \ge 1$ the improved bound

$$\|\hat{f}_{T,\lambda_n}\| \le \tilde{C} x \lambda_n^{\frac{\hat{\rho}-1}{2}} \tag{13}$$

holds with probability not less than $1 - e^{-x}$. Consequently, $\|f_{T,\lambda_n}\| \le \tilde{C} x \lambda_n^{\frac{\hat{\rho}-1}{2}}$ will hold with probability not less than $1 - 2e^{-x}$. Obviously, the latter implies the assertion. In order to establish (13) we will apply Theorem 2 to the modified L1-SVM classifier that produces $\hat{f}_{T,\lambda_n}$. To this end we first observe that the separability condition of Theorem 2 is satisfied since $H$ is separable and continuously embedded into $C(X)$. Furthermore it was shown in [7] that the covering number condition holds and by Proposition 1 we may choose $c$ such that $c \sim x\lambda_n^{-1}$, and $\delta = 0$. Additionally, we can obviously choose $B \sim \lambda_n^{(\rho-1)/2}$. The term $\varepsilon(n, B, c, \delta, x)$ in Theorem 2 can then be estimated by

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

Now for $\rho \le \beta + \gamma$ we have $\frac{\rho+\beta+\gamma}{2} \le \frac{p\rho+2\beta+2\gamma}{2+p}$, and hence we obtain

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

Furthermore, if $\rho > \beta + \gamma$ we have both $\beta + \gamma < \frac{p\rho+2\beta+2\gamma}{2+p}$ and $\beta + \gamma < \frac{\rho+\beta+\gamma}{2}$, and thus we again find

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

Now, in both cases Theorem 2 gives 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

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

holds with probability not less than $1 - e^{-x}$. Furthermore, by Theorem 4 we obtain $\|f_{P,\lambda_n}\| \leq \lambda_n^{(\rho-1)/2} \leq Cx\lambda_n^{(\rho-1)/2}$ for large $n$ which gives $f_{P,\lambda_n} = \hat{f}_{P,\lambda_n}$ for such $n$. With probability not less than $1 - e^{-x}$ we hence have

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

for some constants $\tilde{C}_1, \tilde{C}_2 > 0$ independent of $n$ and $x$. From this we easily obtain that (13) holds for *all* $n \geq 1$ with probability not less than $1 - e^{-x}$. □

*Proof (of Theorem 1 for $q = 0$).* We first observe that there exists a $\gamma > -\beta$ with $\alpha = \frac{4(q+1)}{(2q+pq+4)(1+\beta+\gamma)}$. We fix this $\gamma$ and define $\rho_0 := 0$ and $\rho_{i+1} := \min\{\beta, \frac{\rho_i+\beta+\gamma}{2}, \beta+\gamma\}$. Then it is easy to check that this definition gives

$$\rho_i = \min\Big\{\beta, (\beta+\gamma)\sum_{j=1}^{i} 2^{-j}, \beta+\gamma\Big\} = \min\big\{\beta, (\beta+\gamma)(1-2^{-i})\big\}\,.$$

Now, iteratively applying Lemma 2 gives a sequence of constants $C_i > 0$ with

$$\Pr^*\Big(T \in (X \times Y)^n : \|f_{T,\lambda_n}\| \leq C_i x \lambda_n^{\frac{\rho_i - 1}{2}}\Big) \;\geq\; 1 - e^{-x} \tag{14}$$

for all $n \geq 1$ and all $x \geq 1$. Let us first consider the case $-\beta < \gamma \leq 0$. Then we have $\rho_i = (\beta+\gamma)(1-2^{-i})$, and hence (14) shows that for all $\varepsilon > 0$ there exists a constant $C > 0$ such that

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

for all $n \geq 1$ and all $x \geq 1$. We write $\rho := (1-\varepsilon)(\beta+\gamma)$. As in the proof of Lemma 1 we denote a minimizer of $\mathcal{R}_{L,T}$ on $Cx\lambda_n^{\frac{\rho-1}{2}} B_H$ 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 2 to the modified optimization problem which defines $\hat{f}_{T,\lambda_n}$. To this end we first see as in the proof of Lemma 1 that

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

where in the last two estimates we used the definition of $\rho$. Furthermore, we have already seen in the proof of Lemma 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 $n$. Therefore, applying Theorem 2 and an inequality of Zhang (see [11]) between the excess classification risk and the excess $l$-risk we find that for all $n \geq 1$ we have with probability not less than $1 - e^{-x}$:

$$\begin{aligned}\mathcal{R}_P(\hat{f}_{T,\lambda_n}) - \mathcal{R}_P &\leq 2\lambda_n \|\hat{f}_{T,\lambda_n}\|^2 + 2\mathcal{R}_{l,P}(\hat{f}_{T,\lambda_n}) - 2\mathcal{R}_{l,P} \\ &\leq 2\lambda_n \|\hat{f}_{P,\lambda_n}\|^2 + 2\mathcal{R}_{l,P}(\hat{f}_{P,\lambda_n}) - 2\mathcal{R}_{l,P} + \tilde{C}_1 x^2 \lambda_n^{\beta+\gamma-\varepsilon} \\ &\leq \tilde{C}_2 \lambda_n^{\beta+\gamma-\varepsilon}\,, \end{aligned} \tag{15}$$

where $\tilde{C}_1, \tilde{C}_2 > 0$ are constants independent of $n$ and $x$. Now, from (15) we easily deduce the assertion using the definition of $\lambda_n$ and $\gamma$.

Let us finally consider the case $\gamma > 0$. Then for large integers $i$ we have $\rho_i = \beta$, and hence (14) gives a $C > 0$ such that for all $n \geq 1$, $x \geq 1$ we have

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

Proceeding as for $\gamma \leq 0$ we get $\varepsilon(n, B, c, \delta, x) \preceq x^2\lambda_n^{\frac{p\beta+2\beta+2\gamma}{2+p}} + x^2\lambda_n^{\beta+\gamma} \preceq x^2\lambda_n^{\beta}$, from which we easily obtain the assertion using the definition of $\lambda_n$ and $\gamma$. □

In the rest of this section we will prove Theorem 1 for $q > 0$. We begin with a lemma which is similar to Lemma 1.

**Lemma 2.** *Let $H$ and $P$ be as in Theorem 1. For $\gamma > -\beta$ we define $\lambda_n := n^{-\frac{4(q+1)}{(2q+pq+4)(1+\beta+\gamma)}}$. Now assume that there are $\rho \in [0, \beta)$ and $C \geq 1$ with*

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

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

Now, let $L$ be defined by (12). Analogously to the proof of Lemma 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 (see [7]) we have $|\hat{b}_{T,\lambda_n}| \leq Cx\lambda_n^{\frac{\rho-1}{2}}(K+1)$ with probability not less than $1 - e^{-x}$ for all possible values of the offset. In addition, 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 $T$ 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 1 it now suffices to show the existence of a $\tilde{C} > 0$ such that $\|\hat{f}_{T,\lambda_n}\| \leq \tilde{C}x\lambda_n^{\frac{\hat{\rho}-1}{2}}$ with probability not less than $1 - e^{-x}$. To this end we first observe by Proposition 2 that we may choose $B$, $c$ and $\delta$ such that

$$B \sim x\lambda_n^{\frac{\rho-1}{2}}, \qquad c \sim x^{\frac{q+2}{q+1}}\lambda_n^{\frac{\rho-1}{2}\cdot\frac{q+2}{q+1}}, \qquad \text{and} \qquad \delta \sim x^{\frac{q+2}{q+1}}\lambda_n^{\frac{\rho-1}{2}\cdot\frac{q+2}{q+1}+\frac{\beta q}{q+1}}\,.$$

Some calculations then show that $\varepsilon(n, B, c, \delta, x)$ in Theorem 2 satisfies

$$\varepsilon(n, B, c, \delta, x) \;\preceq\; x^2\lambda_n^{\frac{\rho+\beta+\gamma}{2}} + x^2\lambda_n^{\frac{(\rho+\beta+\gamma)(2q+pq+4)+2\beta q(2-p)}{8(q+1)}}\,.$$

Furthermore observe that we have $\rho \leq \beta - \gamma$ if and only if $\rho + \beta + \gamma \leq \frac{(\rho+\beta+\gamma)(2q+pq+4)+2\beta q(2-p)}{4(q+1)}$. Now let us first consider the case $\rho \leq \beta - \gamma$. Then the above considerations show

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

Furthermore, we obviously have $\lambda_n^\beta \leq \lambda_n^{\frac{\rho+\beta+\gamma}{2}}$. As in the proof of Lemma 1 we hence find a constant $\tilde{C} > 0$ such that for all $x \geq 1$, $n \geq 1$ we have

$$\lambda \|\hat{f}_{T,\lambda_n}\|^2 \leq \tilde{C} x^2 \lambda_n^{\frac{\rho+\beta+\gamma}{2}}$$

with probability not less than $1 - e^{-x}$. On the other hand if $\rho > \beta - \gamma$ we have

$$\varepsilon(n, a, B, c, \delta, x) \preceq x^2 \lambda_n^{\frac{(\rho+\beta+\gamma)(2q+pq+4)+2\beta q(2-p)}{8(q+1)}} \leq x^2 \lambda_n^\beta,$$

so that we get $\lambda \|\hat{f}_{T,\lambda_n}\|^2 \leq \tilde{C} x^2 \lambda_n^\beta$ in the above sense. □

*Proof (of Theorem 1 for $q > 0$).* By using Lemma 2 the proof in the case $q > 0$ is completely analogous to the case $q = 0$. □

## References


1. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
2. Yang, Y.: Minimax nonparametric classification—part I and II. IEEE Trans. Inform. Theory **45** (1999) 2271–2292
3. Wu, Q., Zhou, D.X.: Analysis of support vector machine classification. Tech. Report, City University of Hong Kong (2003)
4. Mammen, E., Tsybakov, A.: Smooth discrimination analysis. Ann. Statist. **27** (1999) 1808–1829
5. Tsybakov, A.: Optimal aggregation of classifiers in statistical learning. Ann. Statist. **32** (2004) 135–166
6. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
7. Steinwart, I., Scovel, C.: Fast rates for support vector machines using Gaussian kernels. Ann. Statist. **submitted** (2004) `http://www.c3.lanl.gov/~ingo/publications/ann-04a.pdf`.
8. Steinwart, I.: Consistency of support vector machines and other regularized kernel machines. IEEE Trans. Inform. Theory **51** (2005) 128–142
9. Steinwart, I., Scovel, C.: Fast rates to bayes for kernel machines. In Saul, L.K., Weiss, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA (2005) 1345–1352
10. Edmunds, D., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators. Cambridge University Press (1996)
11. Zhang, T.: Statistical behaviour and consistency of classification methods based on convex risk minimization. Ann. Statist. **32** (2004) 56–134
12. Rockafellar, R.: Convex Analysis. Princeton University Press (1970)

## Appendix

Throughout this section $P$ denotes a Borel probability measure on $X \times Y$ and $H$ denotes a RKHS of continuous functions over $X$. We use the shorthand $\|\cdot\|$ for $\|\cdot\|_H$ when no confusion should arise. 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. Furthermore, let us define $f_{P,\lambda}$ by replacing $\mathcal{R}_{l,P}$ by $\mathcal{R}_{L,P}$ in (7). In addition we write

$$f^*_{P,\lambda} = \arg\min\left\{ \|f\| : f \in \arg\min_{\|f'\| \le \frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f') \right\}. \tag{16}$$

Of course, we need to prove the existence and uniqueness of $f^*_{P,\lambda}$ which is done in the following lemma.

**Lemma 3.** *Under the above assumptions $f^*_{P,\lambda}$ is well defined.*

*Proof.* Let us first show that there exists an $f' \in \lambda^{-1/2} B_H$ which minimizes $\mathcal{R}_{L,P}(.)$ in $\lambda^{-1/2} B_H$. To that end consider a sequence $(f_n)$ in $\lambda^{-1/2} B_H$ such that $\mathcal{R}_{L,P}(f_n) \to \inf_{\|f\| \le \lambda^{-1/2}} \mathcal{R}_{L,P}(f)$. By the Eberlein-Smulyan theorem we can assume without loss of generality that there exists an $f^*$ with $\|f^*\| \le \lambda^{-1/2}$ and $f_n \to f^*$ weakly. Using the fact that weak convergence in RKHS's imply pointwise convergence, Lebesgue's theorem and the continuity of $L$ then give

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

Hence there is a minimizer of $\mathcal{R}_{L,P}(.)$ in $\frac{1}{\sqrt{\lambda}} B_H$, i.e. we have

$$A := \left\{ f : f \in \arg\min_{\|f'\| \le \frac{1}{\sqrt{\lambda}}} \mathcal{R}_{L,P}(f') \right\} \neq \emptyset.$$

We now show that there is exactly one $f^* \in A$ having minimal norm.

*Existence:* Let $(f_n) \subset A$ with $\|f_n\| \to \inf_{f \in A} \|f\|$ for $n \to \infty$. Like in the proof establishing $A \neq \emptyset$, we can show that there exists an $f^* \in A$ with $f_n \to f^*$ weakly, and $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}(f^*)$. This shows $f^* \in A$. Furthermore, by the weak convergence 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 $\|\frac{1}{2}(f+g)\| < \|f\|$. □

In the following we will define the approximation error and the approximation error function for general $L$. In order to also treat non-universal kernels we first denote the minimal $L$-risk of functions in $H$ by

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

Furthermore, we say that $f \in H$ *minimizes the $L$-risk in $H$* if $\mathcal{R}_{L,P}(f) = \mathcal{R}_{L,P,H}$. Note that if such a minimizer exists then by Lemma 3 there actually 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 $H$. Moreover we have $\|f_{P,\lambda}\| \leq \|f^*_{L,P,H}\|$ for all $\lambda > 0$ since otherwise we find a contradiction by

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

Now, for $\lambda \geq 0$ we write

$$a(\lambda) := \lambda \|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,H} , \tag{17}$$

$$a^*(\lambda) := \mathcal{R}_{L,P}(f^*_{P,\lambda}) - \mathcal{R}_{L,P,H} . \tag{18}$$

Recall, that for universal kernels and the hinge loss function we have $\mathcal{R}_{L,P,H} = \mathcal{R}_{L,P}$ (see [8]), and hence in this case $a(.)$ equals the approximation error function defined in Section 2. Furthermore, for these kernels, $a^*(\lambda)$ is the "classical" approximation error of the hypothesis class $\lambda^{-1/2} B_H$. Our first theorem shows how to compare $a(.)$ and $a^*(.)$.

**Theorem 3.** *With the above notations we have $a(0) = a^*(0) = 0$. Furthermore, $a^*(.)$ is increasing, and $a(.)$ is increasing, concave, and continuous. In addition, we have*

$$a^*(\lambda) \leq a(\lambda) \qquad \text{for all } \lambda \geq 0,$$

*and for any $h : (0,\infty) \to (0,\infty)$ with $a^*(\lambda) \leq h(\lambda)$ for all $\lambda > 0$, we have*

$$a\big(\lambda h(\lambda)\big) \;\leq\; 2h(\lambda) \qquad \text{for all } \lambda > 0.$$

*Proof.* It is clear from the definitions (17) and (18) that $a(0) = a^*(0) = 0$ and $a^*(.)$ is increasing. Since $a(.)$ is an infimum over a family of linear increasing functions of $\lambda$ it follows that $a(.)$ is also concave and increasing. Consequently $a(.)$ is continuous for $\lambda > 0$ (see [12–Thm. 10.1]), and continuity at 0 follows from the proof of [8–Prop. 3.2]. To prove the second assertion, observe that $\|f_{P,\lambda}\|^2 \leq 1/\lambda$ implies $\mathcal{R}_{L,P}(f^*_{P,\lambda}) \leq \mathcal{R}_{L,P}(f_{P,\lambda})$ for all $\lambda > 0$ and hence we find $a^*(\lambda) \leq a(\lambda)$ for all $\lambda \geq 0$. Now let $\tilde{\lambda} := h(\lambda)\|f^*_{P,\lambda}\|^{-2}$. Then we obtain

$$\begin{aligned}\tilde{\lambda}\|f_{P,\tilde{\lambda}}\|^2 + \mathcal{R}_{L,P}(f_{P,\tilde{\lambda}}) \leq \tilde{\lambda}\|f^*_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f^*_{P,\lambda}) &\leq \tilde{\lambda}\|f^*_{P,\lambda}\|^2 + \mathcal{R}_{L,P,H} + h(\lambda) \\ &\leq \mathcal{R}_{L,P,H} + 2h(\lambda) .\end{aligned}$$

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

Our next goal is to show how the asymptotic behaviour of $a(.)$, $a^*(.)$ and $\lambda \mapsto \|f_{P,\lambda}\|$ are related to each other. Let us begin with a lemma that characterizes the existence of $f^*_{L,P,H} \in H$ in terms of the function $\lambda \mapsto \|f_{P,\lambda}\|$.

**Lemma 4.** *The minimizer $f^*_{L,P,H} \in H$ of the $L$-risk in $H$ exists if and only if there exists a constant $c > 0$ with $\|f_{P,\lambda}\| \leq c$ for all $\lambda > 0$. In this case we additionally have $\lim_{\lambda \to 0^+} \|f_{P,\lambda} - f^*_{L,P,H}\|_H = 0$.*

*Proof.* Let us first assume that $f^*_{L,P,H} \in H$ exists. Then we have already seen $\|f_{P,\lambda}\| \leq \|f^*_{L,P,H}\|$ for all $\lambda > 0$, so that it remains to show the convergence. To this end let $(\lambda_n)$ be a positive sequence converging to 0. By the boundedness of $(f_{P,\lambda_n})$ there then exists an $f^* \in H$ and a subsequence $(f_{P,\lambda_{n_i}})$ with $f_{P,\lambda_{n_i}} \to f^*$ weakly. This implies $\mathcal{R}_{L,P}(f_{P,\lambda_{n_i}}) \to \mathcal{R}_{L,P}(f^*)$ as in the proof of Lemma 3. Furthermore, we always have $\lambda_{n_i}\|f_{P,\lambda_{n_i}}\|^2 \to 0$ and thus

$$\mathcal{R}_{L,P,H} = \lim_{i \to \infty} \lambda_{n_i}\|f_{P,\lambda_{n_i}}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_{n_i}}) = \mathcal{R}_{L,P}(f^*)\,, \tag{19}$$

where the first equality can be shown as in [8] for universal kernels. In other words $f^*$ minimizes the $L$-risk in $H$ and hence we have

$$\|f_{P,\lambda_{n_i}}\| \leq \|f^*_{L,P,H}\| \leq \|f^*\| \leq \liminf_{j \to \infty} \|f_{P,\lambda_{n_j}}\|$$

for all $i \geq 1$. This shows both $\|f_{P,\lambda_{n_i}}\| \to \|f^*\|$ and $\|f^*_{L,P,H}\| = \|f^*\|$, and consequently we find $f^*_{L,P,H} = f^*$ by (19). In addition an easy calculation gives

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

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.

Let us now assume $\|f_{P,\lambda}\| \leq c$ for some $c > 0$ and all $\lambda > 0$. Then there exists an $f^* \in H$ and a sequence $(f_{P,\lambda_n})$ with $f_{P,\lambda_n} \to f^*$ weakly. As in the first part of the proof we easily see that $f^*$ minimizes the $L$-risk in $H$. □

Note that 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 $f^*_{L,T,H} \in H$ exists and coincides with the *hard margin* SVM solution. Consequently, the above lemma 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 $\lambda \to 0$.

The following lemma which shows that the function $f_{P,\lambda}$ minimizes $\mathcal{R}_{L,P}(.)$ over the ball $\|f_{P,\lambda}\|B_H$ is somewhat well-known:

**Lemma 5.** *Let $\lambda > 0$ and $\gamma := 1/\|f_{P,\lambda}\|^2$. Then we have $f^*_{P,\gamma} = f_{P,\lambda}$.*

*Proof.* We first show that $f_{P,\lambda}$ minimizes $\mathcal{R}_{L,P}(.)$ over the ball $\|f_{P,\lambda}\|B_H$. To this end assume the converse $\mathcal{R}_{L,P}(f^*_{P,\gamma}) < \mathcal{R}_{L,P}(f_{P,\lambda})$. Since we also have $\|f^*_{P,\gamma}\| \leq 1/\sqrt{\gamma} = \|f_{P,\lambda}\|$ we then find the false inequality

$$\lambda\|f^*_{P,\gamma}\|^2 + \mathcal{R}_{L,P}(f^*_{P,\gamma}) \;<\; \lambda\|f_{P,\lambda}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda})\,, \tag{20}$$

and consequently $f_{P,\lambda}$ minimizes $\mathcal{R}_{L,P}(.)$ over $\|f_{P,\lambda}\|B_H$. 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 again find (20) and hence the assumption $f_{P,\lambda} \neq f^*_{P,\gamma}$ must be false. □

Let us now turn to the main theorem of this section which describes asymptotic relationships between the approximation error, the approximation error function, and the function $\lambda \mapsto \|f_{P,\lambda}\|$.

**Theorem 4.** *The function $\lambda \mapsto \|f_{P,\lambda}\|$ is bounded on $(0,\infty)$ if and only if $a(\lambda) \preceq \lambda$ and in this case we also have $a(\lambda) \sim \lambda$. Moreover for all $\alpha > 0$ we have*

$$a^*(\lambda) \preceq \lambda^\alpha \qquad \text{if and only if} \qquad a(\lambda) \preceq \lambda^{\frac{\alpha}{\alpha+1}}\,.$$

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

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

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

Theorem 4 shows that if $a^*(\lambda)$ behaves essentially like $\lambda^\alpha$ then the approximation error function behaves essentially like $\lambda^{\frac{\alpha}{\alpha+1}}$. Consequently we do not loose information when considering $a(.)$ instead of the approximation error $a^*(.)$.

*Proof (of Theorem 4).* If $\lambda \mapsto \|f_{P,\lambda}\|$ is bounded on $(0,\infty)$ the minimizer $f^*_{L,P,H}$ exists by Lemma 4 and hence we find

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

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 $\lambda > 0$. Moreover by Theorem 3 we easily find $\lambda a(1) \leq a(\lambda)$ for all $\lambda > 0$.

For the rest of the proof we observe that Theorem 3 gives $a(\lambda) \leq a(c\lambda) \leq c\,a(\lambda)$ for $\lambda > 0$ and $c \geq 1$, and $c\,a(\lambda) \leq a(c\lambda) \leq a(\lambda)$ for $\lambda > 0$ and $0 < c \leq 1$. Therefore we can ignore arising constants by using the "$\preceq$"–notation.

Now let us assume $a^*(\lambda) \preceq \lambda^\alpha$ for some $\alpha > 0$. Then from Theorem 3 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 5 we then obtain

$$a^*(\gamma) \;=\; \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,H} \;\leq\; a(\lambda) \;\preceq\; \lambda^{\frac{\alpha}{\alpha+1}} \;\preceq\; \|f_{P,\lambda}\|^{-2\alpha} \;=\; \gamma^\alpha\,.$$

Now, if $f^*_{L,P,H}$ does not exists then the function $\lambda \mapsto \|f_{P,\lambda}\|^{-2}$ tends to 0 if $\lambda \to 0$ and thus $a^*(\lambda) \preceq \lambda^\alpha$. In addition, if $f^*_{L,P,H}$ exists the assertion is trivial.

For the third assertion recall that Lemma 5 states $f_{P,\lambda} = f^*_{P,\gamma}$ with $\gamma := \|f_{P,\lambda}\|^{-2}$ and hence we find

$$a(\lambda) \;=\; \lambda\|f_{P,\lambda}\|^2 + a^*\big(\|f_{P,\lambda}\|^{-2}\big)\,. \tag{21}$$

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

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

Combining this with (21) yields the third assertion. □