Consistency of Support Vector Machines and Other Regularized Kernel Classifiers

Ingo Steinwart

***Abstract*—It is shown that various classifiers that are based on minimization of a regularized risk are universally consistent, i.e., they can asymptotically learn in every classification task. The role of the loss functions used in these algorithms is considered in detail. As an application of our general framework, several types of support vector machines (SVMs) as well as regularization networks are treated. Our methods combine techniques from stochastics, approximation theory, and functional analysis.**

***Index Terms*—Computational learning theory, kernel methods, pattern recognition, regularization, support vector machines (SVMs), universal consistency.**

I. Introduction

We treat the statistical classification problem which have been studied in both statistics and machine learning (cf. [1] for a throughout treatment). For recalling this problem let X be a nonempty set, and $Y := \{-1, 1\}$. A *classifier* C is a rule that assigns to every *training set*

$$T = ((x_1, y_1), \dots (x_n, y_n)) \in (X \times Y)^n$$

a measurable function $f_T : X \to \mathbb{R}$. Here, it is always assumed that T is independent and identically distributed (i.i.d.) with respect to an unknown distribution P on $X \times Y$. 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 an example (x, y) drawn from P independently to T. Here, misclassification means $\operatorname{sign} f_T(x) \neq y$. To make this precise the risk of a measurable function $f : X \to \mathbb{R}$ is defined by

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

The smallest achievable risk

$$\mathcal{R}_P := \inf\big\{\mathcal{R}_P(f) \mid f : X \to \mathbb{R} \text{ measurable}\big\}$$

is called the *Bayes risk* of P. A classifier is said to be *universally consistent* if

$$\lim_{n\to\infty} \mathcal{R}_P(f_T) = \mathcal{R}_P \tag{1}$$

holds in probability for all distributions P on $X \times Y$. It is strongly universally consistent if (1) even holds almost surely.

Manuscript received September 12, 2002; revised June 29, 2004. This work was supported in part by the DFG under Grant Ca 179/4-1.

The author is with the Los Alamos National Laboratory, Los Alamos, NM 87545 USA (e-mail: ingo@lanl.gov).

Communicated by A. B. Nobel, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2004.839514

The type of classifiers for which we will establish consistency results is based on one of the following optimization problems:

$$\arg\min_{f\in H} \Omega\big(\lambda_n, \|f\|_H\big) + \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \tag{2}$$

or

$$\arg\min_{\substack{f\in H\\ b\in\mathbb{R}}} \Omega\big(\lambda_n, \|f\|_H\big) + \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i) + b\big) \tag{3}$$

respectively. Here, $T = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in (X \times Y)^n$ is a training set, $\lambda_n > 0$ is a regularization parameter, H is a reproducing kernel Hilbert space (RKHS), Ω is a regularization function, and L is a loss function (cf. the Appendix and Section II for precise definitions). The additional term b in (3) is called *offset*. The corresponding decision functions of the considered classifiers are f_{T,λ_n} or $\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}$, respectively, where $f_{T,\lambda_n} \in H$ and $(\tilde{f}_{T,\lambda_n}, \tilde{b}_{T,\lambda_n}) \in H \times \mathbb{R}$ are *arbitrary* solutions of (2) and (3) (cf. Lemma 3.1 and 3.10 for the existence). Various recently proposed algorithms including *regularization networks* and several variants of support vector machines (SVMs) belong to this type of classifiers (see the examples below). In particular, if $\Omega\big(\lambda_n, \|f\|_H\big) = \lambda\|f\|^2$ and L is the hinge loss $L(y, t) := \max\{0, 1 - yt\}$ then (3) becomes the well-known primal optimization problem

$$\begin{aligned} \text{minimize} \quad & \lambda\langle f, f\rangle + \frac{1}{n}\sum_{i=1}^{n}\xi_i, && f \in H,\ b \in \mathbb{R},\ \xi \in \mathbb{R}^n \\ \text{subject to} \quad & y_i\big(f(x_i) + b\big) \geq 1 - \xi_i, && i = 1, \dots, n \\ & \xi_i \geq 0, && i = 1, \dots, n \end{aligned} \tag{4}$$

of the so-called L1-SVM with offset. For L1-SVMs without offset, we only have to drop the b in the constraints. Furthermore, if one considers the squared hinge loss function (L2-SVM) instead, then in the above sum ξ_i has to be replace by ξ_i^2.

Instead of minimizing over the whole space H in (2) and (3), it suffices to consider a small subspace. Indeed, by the representer theorem [2, proof of Theorem 1] there exists a solution f_{T,λ_n} of (2) which has the form

$$f_{T,\lambda_n} = \sum_{i=1}^{n} \alpha_i^* k(x_i, \cdot)$$

where $k : X \times X \to \mathbb{R}$ is the reproducing kernel of H. Hence, it suffices to consider (2) over the subspace $\{\sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : \alpha_1, \dots, \alpha_n \in \mathbb{R}\}$. Furthermore, we have

$$\|f\|_H^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$

for $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$. Thus, once one has found a minimizer $\alpha_1^*, \ldots, \alpha_n^* \in \mathbb{R}$ of

$$\Omega\Bigl(\lambda, \sqrt{\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)}\Bigr) + \frac{1}{n}\sum_{i=1}^n L\Bigl(y_i, \sum_{j=1}^n \alpha_j k(x_i, x_j)\Bigr)$$

a solution of (2) is given by $f_{T,\lambda_n} = \sum_{i=1}^n \alpha_i^* k(x_i, \cdot)$. A similarly argument can be employed for (3). However, for $\Omega(\lambda, t) = \lambda t^2$ and specific convex loss functions the dual of (2) or (3) is usually solved instead (cf. [3] and [4]). For example, if L is the hinge loss then the dual problem becomes

$$\text{maximize} \quad \sum_{i=1}^n \alpha_i - \frac{1}{4\lambda}\sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j k(x_i, x_j), \qquad \alpha \in \mathbb{R}^n$$

$$\text{subject to} \quad 0 \le \alpha_i \le \frac{1}{n}, \qquad i = 1, \ldots, n. \tag{5}$$

Recall that if (3) is considered instead, then the additional constraint $\sum_{i=1}^n y_i \alpha_i = 0$ appears in (5). Finally, for the squared hinge loss function the dual problem is similar (see [3]).

Since the above introduced framework is very general, we first give some examples that show how the theory developed in this paper can be applied to many algorithms of practical interest. Throughout these examples, we assume that X is a compact metric space, e.g., a bounded and closed subset of $\mathbb{R}^d$. The first group of examples we consider consists of almost classical types of SVMs. These classifiers have been introduced by Vapnik and his coauthors between 1992 and 1995. A good introduction to these SVMs is provided in [3]. We begin with the most common SVM.

Example 1.1: Let $\Omega(\lambda, t) := \lambda t^2$, L be the hinge loss function $L(y, t) := \max\{0, 1 - yt\}$, $y \in Y$, $t \in \mathbb{R}$, and (λ_n) be a positive sequence with $\lambda_n \to 0$. Then the classifiers based on either (2) or (3) are called L1-SVMs (cf. [5]). We will show that the L1-SVM based on (2) is universally consistent whenever a universal kernel k (see the following examples and Section II for a definition) is used and (λ_n) satisfies $n\lambda_n^2 \to \infty$. If the L1-SVM is based on (3) we can only establish consistency under the slightly stronger condition $n\lambda_n^2 / \log n \to \infty$. Note, that this condition actually ensures strong universal consistency for L1-SVMs with and without offset. Furthermore, for both classifiers these conditions can be improved by using smoothness properties of k. In particular, if k is even a universal C^∞-kernel on a closed ball of $\mathbb{R}^d$—e.g., the Gaussian radial basis function (RBF) kernel $k(x, x') := \exp(-\sigma\|x - x'\|_2^2)$ or Vovk's infinite polynomial kernel $k(x, x') := (1 - \langle x, x'\rangle)^{-\alpha}$ (see [6] and Example 3.7)—it suffices to use a sequence (λ_n) with $n\lambda_n^{1+\varepsilon} \to \infty$ for some arbitrary small $\varepsilon > 0$ in order to ensure strong universal consistency. For a Gaussian RBF kernel, the latter condition can be further weakened to $n\lambda_n |\log \lambda_n|^{-d-1} \to \infty$ if one is only interested in universal consistency.

As mentioned, in practice (3) is treated by solving the dual problem (5). In this case, the arising equality constraint in the dual problem can cause some algorithmic difficulties. Instead of omitting the offset as in (2), it is also possible to consider the optimization problem of the form

$$\min_{\substack{f \in H \\ b \in \mathbb{R}}} \Omega\bigl(\lambda_n, \|f\|_H + b\bigr) + \frac{1}{n}\sum_{i=1}^n L\bigl(y_i, f(x_i) + b\bigr). \tag{6}$$

It is easy to see that this approach is equivalent (2) using the kernel $k + 1$ instead of k.

Finally, note that all the conditions on (λ_n) presented in this example are derived from our general results using the facts $\delta_\lambda = \sqrt{2/\lambda}$, $|L_\lambda|_1 = 1$, and $\|L_\lambda\|_\infty \sim 1/\sqrt{\lambda}$ for $\lambda \to 0$ (see Section III for a definition of these quantities and the mentioned results).

The above conditions on (λ_n) are stronger than those derived in [7] for the L1-SVM *without* offset. However, they significantly improve the only known conditions (cf. [8]) for the L1-SVM *with* offset. Furthermore, for both SVMs the conditions almost coincide with the condition in [7] if the used kernel is smooth.

Recall, that there exists another well-known variant, the so-called ν-SVM, which is also based on the hinge loss function. However, the regularization of this classifier is different to that of the algorithms in consideration. Indeed, the (almost) optimal value for the regularization parameter ν is determined by the Bayes risk of the underlying measure P (see [9]).

Let us now treat the so-called L2-SVM.

Example 1.2: Let $\Omega(\lambda, t) := \lambda t^2$, L be the squared hinge loss function $L(y, t) := (\max\{0, 1 - yt\})^2$, and (λ_n) be a positive sequence with $\lambda_n \to 0$. Then the classifiers based on either (2) or (3) are called L2-SVMs. The L2-SVM based on (2) is universally consistent whenever a universal kernel k is used and (λ_n) satisfies $n\lambda_n^4 \to \infty$. If the L2-SVM is based on (3) we can only establish consistency under the slightly stronger condition $n\lambda_n^4 / \log n \to \infty$. Note that this condition actually ensures strong universal consistency for L2-SVMs with and without offset. Furthermore, for both classifiers these conditions can be improved by using smoothness properties of k. In particular, if k is even a C^∞-kernel it suffices to use a sequence (λ_n) with $n\lambda_n^{2+\varepsilon} \to \infty$ for some arbitrary small $\varepsilon > 0$ in order to ensure strong universal consistency. Again, for a Gaussian RBF kernel on $X \subset \mathbb{R}^d$ this condition can be further weakened to $n\lambda_n^2 |\log \lambda_n|^{-d-1} \to \infty$ if one is only interested in universal consistency.

Moreover, in [10], Mangasarian and Musicant proposed another variation of the theme, called Lagrangian SVM, which is based on the optimization problem (6) with $\Omega(\lambda, t) := \lambda t^2$ and $L(y, t) := \bigl(\max\{0, 1 - yt\}\bigr)^2$. The corresponding consistency conditions on (λ_n) are obvious. As indicated in [4, Sec. 10.6.2] it is open whether this modification has a significant influence on the generalization performance. Our work shows that asymptotically there is only a small difference in the kernel-independent conditions on (λ_n) which ensures consistency. Furthermore, considering smooth kernels this difference vanishes.

Finally, note that all the conditions on (λ_n) presented are derived from our general results using $\delta_\lambda = \sqrt{2/\lambda}$, $|L_\lambda|_1 \sim 1/\sqrt{\lambda}$, and $\|L_\lambda\|_\infty \sim 1/\lambda$.

The two classifiers in the next example are based on the square loss function. Interestingly, the first has been inspired by the SVM approach but the second has been introduced independently from SVMs.

Example 1.3: Least square SVMs (LS-SVMs) proposed in [11] are based on the minimization problem (3) with $\Omega(\lambda, t) :=$

λt^2 and $L(y,t) := (1-yt)^2$. The analogous classifiers based on (2) are called regularization networks or kernel ridge regression classifiers and were introduced in [12]. Since $\delta_\lambda = \sqrt{2/\lambda}$, $|L_\lambda|_1 \sim 1/\sqrt{\lambda}$, and $\|L_\lambda\|_\infty \sim 1/\lambda$ for $\lambda \to 0$, the conditions for both classifiers coincide with those of the corresponding L2-SVMs.

Since many SVMs for classification are actually based on a regression-like approach, it is rather natural to use the following SVM variants that were originally constructed for regression problems for classification tasks. These algorithms are treated in the following example.

Example 1.4: Support vector machines with 1-regression loss function (R1-SVMs) are based on the minimization problem (3) with $\Omega(\lambda,t) := \lambda t^2$ and

$$L_\varepsilon(y,t) := \max\{0, |y-t| - \varepsilon\}$$

for some fixed $0 \le \varepsilon < 1$. Since $\delta_\lambda \sim \sqrt{1/\lambda}$, $|L_\lambda|_1 = 1$, and $\|L_\lambda\|_\infty \sim 1/\sqrt{\lambda}$ for $\lambda \to 0$, the conditions ensuring universal consistency coincide with those of the L1-SVMs with offset. It is clear, that one can also consider this type of classifier without offset.

If one replaces the R1-SVM loss function by

$$L_\varepsilon(y,t) := (\max\{0, |y-t| - \varepsilon\})^2$$

for some $0 \le \varepsilon < 1$ one obtains SVMs with 2-regression loss function (R2-SVMs). Since $\delta_\lambda \sim \sqrt{1/\lambda}$, $|L_\lambda|_1 \sim 1/\sqrt{\lambda}$, and $\|L_\lambda\|_\infty \sim 1/\lambda$ for $\lambda \to 0$, the conditions ensuring universal consistency coincide with those of the L2-SVMs with offset. Again, one can also consider this type of classifier without offset or with optimized offset.

The following example provides an SVM that is not universally consistent.

Example 1.5: Support vector machines with asymmetric loss function are sometimes proposed in order to penalize errors in each class differently. In particular, this approach is sometimes recommended for unbalanced classification problems. Let us first consider an asymmetric hinge loss function, i.e., $L(y,t) := c_y \max\{0, 1-yt\}$ for $c_y > 0$, $y = \pm 1$. Unfortunately, it can be easily checked that L is admissible (see Section II) if and only if $c_{-1} = c_1$, i.e., if and only if L is *symmetric*. It will thus turn out, that the classifiers using L cannot be universally consistent if $c_{-1} \ne c_1$.

However, there are various other asymmetric loss functions that are admissible. For example, L defined by

$$L(1,t) := 2\max\{0, 1-t\} \text{ and } L(-1,t) := (\max\{0, 1+t\})^2$$

is such a function. It is easily seen that the corresponding conditions on (λ_n) ensuring universal consistency coincide with those of the L2-SVM.

The following two examples consider the logistic and the AdaBoost loss function, respectively.

Example 1.6: The logistic loss function $L(y,t) := \log(1 + \exp(-yt))$ is a convex regular loss function. For $\Omega(\lambda,t) := \lambda t^2$ we have $\delta_\lambda \sim \sqrt{1/\lambda}$, $|L_\lambda|_1 \sim 1$, and $\|L_\lambda\|_\infty \sim 1/\sqrt{\lambda}$ if $\lambda \to 0$. Therefore the conditions on (λ_n), for classifiers based on (2) or (3) with respect to Ω and L are equal to the conditions for the corresponding L1-SVM.

Example 1.7: Another example of a convex and admissible loss function is the AdaBoost loss function $L(y,t) := \exp(-yt)$. Obviously, for $\Omega(\lambda,t) := \lambda t^2$ we have

$$|L_\lambda|_1 \sim \exp(\sqrt{2/\lambda}K) \text{ and } \|L_\lambda\|_\infty \sim \exp(\sqrt{2/\lambda}K)$$

if $\lambda \to 0$. Classifiers based on (2) with L, Ω, and a universal kernel are therefore universally consistent by Theorem 3.20 if (λ_n) fulfills $\lambda_n \to 0$ and $n\lambda_n^2/\exp(\sqrt{32/\lambda_n}K) \to \infty$. An easy calculation shows that the fastest decreasing sequences (λ_n) satisfying these conditions are essentially of the form $\lambda_n = c(\log n)^{-2}$ for $c > 32K^2$. By our techniques, this rate cannot be significantly improved for smooth kernels. Finally, our techniques can also be adapted to classifiers based on (3). The resulting conditions are very similar.

Several other loss function can also be treated by our results including (see, e.g., [13] and [14]) the sigmoid loss function, a truncated hinge loss function, and some smooth approximations of the margin loss functions of the above examples.

The following last two examples are mainly of theoretical interest. Historically, they were considered in order to "explain" the generalization performance of SVMs.

Example 1.8: In order to motivate SVMs, the regularization function Ω defined by $\Omega(\lambda,t) := \infty \cdot \mathbf{1}_{(1/\sqrt{\lambda},\infty]}(t)$ in combination with the structural risk minimization method with respect to the hinge loss function $L(y,t) := \max\{0, 1-yt\}$ was considered in [15]. Using Proposition 3.3, this approach actually yields universal consistency for classifiers with the above Ω in terms of structural risk minimization. Moreover, our results also yield another method making these classifiers universally consistent. Indeed, since $\delta_\lambda \sim 1/\sqrt{\lambda}$, all of the above results on SVMs which do not depend on Theorem 3.20 can be directly applied to this modification.

Example 1.9: In [16, Sec. 10.2] SVMs were interpreted as an approximation of the minimization of the number of misclassified samples: the admissible loss function $L_\sigma(y,t) := (\max\{0, 1-yt\})^\sigma$, $\sigma > 0$, approximates the loss function $L_0(y,t) := \mathbf{1}_{\{y \ne \operatorname{sign} t\}}$ for $\sigma \to 0$. Since error minimization in H with respect to L_0 is NP-hard whenever the training set cannot be linearly separated (cf. [17], [18]), it was proposed to replace L_0 by L_σ for small $\sigma > 0$. Moreover, in order to apply results on empirical risk minimization it was assumed to use $\Omega(\lambda,t) := \infty \cdot \mathbf{1}_{(1/\sqrt{\lambda},\infty]}(t)$, or—as an approximation—$\Omega(\lambda,t) := \lambda t^2$. Unfortunately, using universal kernels on infinite spaces X the motivation in [16] cannot work since the corresponding function classes always have infinite Vapnik–Chervonenkis (VC) dimension. However, the classifiers based on L_σ are universally consistent for suitable sequences (λ_n). This can be seen by our theorems and the fact that L_σ is $\min\{1,\sigma\}$-Hölder-continuous.

As discussed in [1], there cannot exist a uniform rate of convergence in (1) for any classifier. Hence, there are roughly speaking two alternatives for investigating the generalization performance of classifiers: an asymptotic approach which treats

universal consistency for specific classifiers and a sample dependent approach that estimates the risk of decision functions in terms of observed data on the training set. For SVMs, mainly the latter approach has been followed so far. Unfortunately, it is shown in [19] that the existing bounds cannot explain the generalization performance of SVMs and thus, a sample-dependent theory still has to be developed (cf. [20] for the best known results in this direction). On the other hand, there are only a few works dealing with consistency of classifiers based on (2) or (3). Some preliminary results in [21] and [22] show consistency of L1-SVMs for *restricted* classes of distributions. Furthermore, there exist two results establishing universal consistency for classifiers based on (2) or (3). As already mentioned, the first result [8] in this direction showed that L1-SVMs with offset are universally consistent if the used kernel is universal and the regularization parameter λ_n tends "very slowly" to 0. In [7], this condition on (λ_n) was significantly improved for classifiers based on (2) and specific continuous convex loss functions including some standard choices listed in the above examples. In this work, we establish universal consistency for very general functions Ω and L. Namely, we prove both kernel-independent and kernel-dependent conditions on (λ_n) that ensure universal consistency of the corresponding classifiers. Unlike [7], our results neither make any convexity assumption on L nor they are restricted to (2). On the one hand, the shortcomings of this generality are sometimes stronger conditions on (λ_n). On the other hand, however, this generality gives us the opportunity to select Ω and L with respect to computational issues.

The algorithms treated in this work are based on a regularization approach. Regularization techniques are well known and have a broad range of applications. In particular, for statistical problems they have been intensively studied in the literature (cf., e.g., [23] and the references therein). However, the classifiers based on (2) or (3) differ from the commonly considered regularization scenarios in statistics.

- In general, using the loss function of interest in (2) or (3), i.e., $L_0(y,t) := \mathbf{1}_{\{y \neq \operatorname{sign} t\}}$, leads either to overfitting (cf. [22]) or to combinatorial optimization problems which are hardly solved efficiently (cf. [17], [18]). Therefore, one usually solves (2) or (3) with respect to a loss function L different from L_0. However, the classifier should still be universally consistent with respect to L_0. In order to guarantee this, a necessary condition on L is that the target function f^* with respect to L should have the same sign as the optimal Bayes decision function

$$x \mapsto \operatorname{sign}(2P(y = 1 \mid x) - 1).$$

  Surprisingly, we can show in this work that for continuous loss functions this is even a sufficient condition.

- Common techniques such as those used in [23] assume that the target function f^* is in a space $\tilde{H}$ which is related to H in order to estimate corresponding norms of $\|f_{T,\lambda_n} - f^*\|_{\tilde{H}}$. Unlike in regression, this assumption is too restrictive in classification. Indeed, considering, e.g., continuous kernels, H and $\tilde{H}$ only contain continuous functions. However, for the hinge loss function, the target function (almost) coincides with the Bayes decision function which, in general, is far from being continuous.

These differences show that the existing techniques of [23] cannot be applied to the classifiers based on (2) or (3). Therefore, in this work we develop a new ansatz. Although this ansatz is rather simple, it might be hidden by technical issues, and hence we like to give a brief roadmap, now. Since for both (2) and (3) the techniques are almost identical besides technical details we restrict ourselves to (2). Then given a loss function L and a probability measure P we define the *L-risk* of a measurable function $f : X \to \mathbb{R}$ by

$$\mathcal{R}_{L,P}(f) := \mathbb{E}_{(x,y)\sim P} L(y, f(x)). \tag{7}$$

The smallest possible L-risk is denoted by $\mathcal{R}_{L,P}$. Furthermore, given a regularization function Ω and an RKHS H the *regularized L-risk* is defined by

$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f) := \Omega(\lambda, \|f\|_H) + \mathcal{R}_{L,P}(f) \tag{8}$$

for all $f \in H$ and all $\lambda > 0$. If P is an empirical measure with respect to $T \in (X \times Y)^n$, we write $\mathcal{R}_{L,T}(\cdot)$ and $\mathcal{R}^{\text{reg}}_{L,T,\lambda}(\cdot)$, respectively. Note, that $\mathcal{R}^{\text{reg}}_{L,T,\lambda}(\cdot)$ is the objective function in (2). Now, the first step in our approach is to show that there always exists an element $f_{P,\lambda} \in H$ minimizing the regularized L-risk (see Lemma 3.1). In the next step (see Proposition 3.2) we show

$$\lim_{\lambda \to 0} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \mathcal{R}_{L,P} \tag{9}$$

if the used RKHS is rich enough, i.e., universal. We then prove that approximating the minimal L-risk is sufficient to approximately achieve the Bayes risk. More precisely, we show (see Proposition 3.3) that for all sequences of measurable functions $f_n : X \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \mathcal{R}_{L,P}(f_n) = \mathcal{R}_{L,P} \quad \Rightarrow \quad \lim_{n\to\infty} \mathcal{R}_P(f_n) = \mathcal{R}_P \tag{10}$$

if L is admissible. In the final step, we correlate the L-risk of $f_{T,\lambda}$ with the empirical L-risk of $f_{T,\lambda}$ by certain concentration inequalities. If, e.g., L is the hinge loss, $\Omega(\lambda, t) := \lambda t^2$ and $0 < \lambda_n \leq 1$ we have (see Lemma 3.21 and 3.22)

$$P^n\Big(T : |\mathcal{R}_{L,T}(f_{T,\lambda_n}) - \mathcal{R}_{L,P}(f_{T,\lambda_n})| > \varepsilon + \frac{2K^2}{n\lambda_n}\Big) \leq 2e^{-\frac{\varepsilon^2 n\lambda_n^2}{2(\sqrt{2}K+1)^4}} \tag{11}$$

where K is a constant depending on the kernel k. Hence,

$$|\mathcal{R}_{L,T}(f_{T,\lambda_n}) - \mathcal{R}_{L,P}(f_{T,\lambda_n})| \to 0$$

in probability whenever $n\lambda_n^2 \to \infty$. The rest of the proof consists of plugging (9)–(11) together (see the proof of Theorem 3.5). Note, that instead of using (11) we can and will also employ several other concentration inequalities. Unlike (11) that is based on a stability argument due to [20], most of them depend on covering numbers of certain operators related to H. It turns out that each tuple of concentration inequality, loss function, and RKHS gives a condition on (λ_n) ensuring $|\mathcal{R}_{L,T}(f_{T,\lambda_n}) - \mathcal{R}_{L,P}(f_{T,\lambda_n})| \to 0$. Some of these conditions

have already been listed in the preceding examples. The general theory can be found in Section III.

Our approach is somehow modular: if, e.g., there is a new concentration inequality for specific situations, one simply can use it together with (9) and (10) to obtain new consistency results. Furthermore, if one is interested in problems of the form (2) or (3) for function spaces H different from RKHSs, one only has to establish (9) and a concentration inequality in the spirit of (11). Finally, note that (10) is completely algorithm independent and thus it can be used in many other settings as well.

The rest of this work is organized as follows: In Section II, we introduce some notions for H, Ω, and L which are essential for our further work. In Section III, we present our general theory which leads to the examples discussed in the Introduction. The proofs of this theory can be found in Section IV. Finally, there is a small Appendix explaining kernels and some concepts from functional analysis.

II. Preliminaries

In the following, let $\overline{\mathbb{R}} := [-\infty, \infty]$, $\mathbb{R}^+ := [0, \infty)$, and $\overline{\mathbb{R}}^+ := [0, \infty]$. Given two functions $g, h : (0, \infty) \to (0, \infty)$ we write $g \preceq h$ if there exists a constant $c > 0$ with $g(\varepsilon) \leq c\, h(\varepsilon)$ for all sufficiently small $\varepsilon > 0$. We write $g \sim h$ if both $g \preceq h$ and $h \preceq g$. Analogously, we write $a_n \preceq b_n$ for two positive sequences (a_n) and (b_n) if there exists a constant $c > 0$ such that $a_n \leq c b_n$ for all $n \geq 1$. Again, $a_n \sim b_n$ means that both $a_n \preceq b_n$ and $b_n \preceq a_n$ hold. Furthermore, we always assume $0 \cdot \infty := 0$.

Throughout the paper, let X be a compact metric space. For a positive semidefinite kernel $k : X \times X \to \mathbb{R}$, we denote the corresponding RKHS (cf. [24], [25, Ch. 3], and the Appendix) by H_k or simply H. For its closed unit ball we write B_H. Recall that the map $\Phi : X \to H$, $x \mapsto k(x, \cdot)$ fulfills $k(\cdot, \cdot) = \langle \Phi(\cdot), \Phi(\cdot) \rangle_H$ by the reproducing property. We will often use the quantity

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

Recall that KB_H is the smallest ball in H centered at the origin that contains the image of X under Φ. Moreover, k is continuous if and only if Φ is. In this case, H can be continuously embedded into the space of all continuous functions $C(X)$ via $I : H \to C(X)$ defined by $Iw := \langle w, \Phi(\cdot) \rangle_H$, $w \in H$. Since we always assume that k is continuous, we often identify elements of H as continuous functions on X. If the embedding $I : H \to C(X)$ has a dense image we call k a *universal* kernel. Besides many other examples, the Gaussian RBF kernel

$$k(x, x') := \exp(-\sigma^2 \|x - x'\|_2^2)$$

is universal for fixed $\sigma > 0$ on every compact, i.e., closed and bounded, subset of $\mathbb{R}^d$ (cf. [22, Sec. 3]).

Besides the regularization function $\Omega(\lambda, t) := \lambda t^2$, we sometimes also consider $\Omega(\lambda, t) := \infty \cdot \mathbf{1}_{(1/\sqrt{\lambda}, \infty]}(t)$ (cf. Example 1.8 and 1.9), and therefore, let us fix the properties of Ω which we will need in the following.

Definition 2.1: Let $\Omega : [0, \infty) \times \overline{\mathbb{R}}^+ \to \overline{\mathbb{R}}^+$ be an increasing function which is continuous in 0 with respect to the first variable and unbounded with respect to the second variable. Moreover, let us assume that for all $\lambda > 0$ there exists a $t > 0$ such that $\Omega(\lambda, t) < \infty$. We call Ω a *regularization function* if for all $\lambda > 0$, $s \in \mathbb{R}^+$, $t \in \overline{\mathbb{R}}^+$, and for all sequences $(t_n) \subset \overline{\mathbb{R}}^+$ with $t_n \to t$ and $\Omega(\lambda, t_n) < \infty$ we have $\Omega(\lambda, 0) = \Omega(0, s) = 0$ and $\Omega(\lambda, t_n) \to \Omega(\lambda, t)$.

Recall, that for a given Borel probability measure P on $X \times Y$ there exists a map $x \mapsto P(\cdot \mid x)$ from X into the set of all probability measures on Y such that P is the joint distribution of $(P(\cdot \mid x))_x$ and the marginal distribution P_X of P (cf. [26, Lemma 1.2.1.]).

In order to treat the L-risk for a given loss function $L : Y \times \overline{\mathbb{R}} \to \overline{\mathbb{R}}^+$ we write $C(\alpha, t) := \alpha L(1, t) + (1 - \alpha) L(-1, t)$ for $\alpha \in [0, 1]$ and $t \in \overline{\mathbb{R}}$. This function can be used to compute the L-risk (7) of a function $f : X \to \overline{\mathbb{R}}$ by

$$\mathcal{R}_{L,P}(f) = \int_X C\big(P(1 \mid x), f(x)\big)\, P_X(dx).$$

Roughly speaking, it turns out that the solutions of (2) or (3) tend to a function f^* that minimizes this L-risk (see [27] for details). Hence, the following definition is fundamental in order to guarantee that these solutions tend to have the same sign as the Bayes decision rule.

Definition 2.2: A continuous function $L : Y \times \overline{\mathbb{R}} \to \overline{\mathbb{R}}^+$ with $L(Y, \mathbb{R}) \subset \mathbb{R}^+$ is called an admissible loss function if for every $\alpha \in [0, 1]$ and every $t_\alpha \in \overline{\mathbb{R}}$ with

$$C(\alpha, t_\alpha) = \min_{t \in \overline{\mathbb{R}}} C(\alpha, t) =: M(\alpha) \tag{12}$$

we have $t_\alpha < 0$ if $\alpha < 1/2$ and $t_\alpha > 0$ if $\alpha > 1/2$.

A similar notion together with some sufficient conditions can be found in [28]. Besides the asymmetric hinge loss function discussed in Example 1.5, all loss functions treated in the examples of the Introduction are admissible. We will see in Lemma 4.1 that there always exists a measurable version of $\alpha \mapsto t_\alpha$. For specific loss functions, such minimizing functions can be found in [13]. As indicated earlier, the admissibility of L is necessary in order to get universally consistent classifiers based on (2) or (3). To see this, it suffices to consider a space X which only consists of one point. Then, assuming that the admissibility condition is violated for some $\alpha \neq 1/2$ it is easy to check that there exists a classifier based on (2) or (3), respectively, that is not consistent for the probability measure P with $P(y = 1) = \alpha$.

III. Results

In this section, we develop the general theory which leads to the examples discussed in the Introduction. It is organized as follows: In Section III-A, we mainly formalize the results (9) and (10) discussed in the roadmap. In Section III-B, we present consistency results in terms of covering numbers. These results are kernel dependent. In the following subsection, we establish consistency results based on localized covering numbers that lead to a kernel-independent condition on (λ_n) ensuring universal consistency. Finally, for convex loss functions and classifiers without offset the latter condition is improved in Section III-D.

A. Some Results Mentioned in the Roadmap

To simplify notations we fix some technical definitions: let k be a positive semidefinite kernel, L be an admissible loss function, and Ω a regularization function. For $\lambda > 0$ we define

$$\delta_\lambda := \sup\{t : \Omega(\lambda, t) \le L(1,0) + L(-1,0)\}$$
$$L_\lambda := L_{|Y\times[-\delta_\lambda K, \delta_\lambda K]}.$$

Note, that we have $0 < \delta_\lambda < \infty$ for all $\lambda > 0$ and

$$\hat{\delta} := \inf\{\delta_\lambda : \lambda \in (0,1]\} > 0.$$

Moreover, we even obtain $\delta_\lambda \to \infty$ for $\lambda \to 0$. The main purpose of δ_λ which is a key quantity throughout this work is that it gives a simple upper bound on the norm of the solutions of (2) as we see in the following lemma.

Lemma 3.1: Let L be an admissible loss function, Ω a regularization function, and k be a continuous kernel on X. Then for all Borel probability measures P on $X \times Y$ and all $\lambda > 0$, there is an $f_{P,\lambda} \in H$ with

$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \inf_{f\in H} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f).$$

Moreover, for all such $f_{P,\lambda} \in H$ we have $\|f_{P,\lambda}\| \le \delta_\lambda$.

The proof of this lemma can be found in Section IV. Note, that Lemma 3.1 in particular ensures that there always exists a solution of (2). The following proposition formalizes (9). Again a proof can be found in Section IV.

Proposition 3.2: Let k be a universal kernel on X, L be an admissible loss function, and Ω be a regularization function. Then for every Borel probability measure P on $X \times Y$ we have

$$\lim_{\lambda\to 0} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \mathcal{R}_{L,P}.$$

The next result shows that it suffices to approximate the minimal L-risk in order to approximate the Bayes risk, i.e., it formalizes (10). Again, the proof is worked out in Section IV.

Proposition 3.3: Let L be an admissible loss function and P be a Borel probability measure on $X \times Y$. Then for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all measurable $f : X \to \overline{\mathbb{R}}$ with $\mathcal{R}_{L,P}(f) \le \mathcal{R}_{L,P} + \delta$ we have $\mathcal{R}_P(f) \le \mathcal{R}_P + \varepsilon$.

B. Consistency Results Based on Covering Numbers

So far, we have established (9) and (10) of the roadmap. As described in the Introduction, we finally need some concentration inequalities in the spirit of (11). To this end, let us first recall that for a metric space (M,d) the *covering numbers* of M are defined by

$$\mathcal{N}((M,d),\varepsilon) := \min\Big\{n \in \mathbb{N} \,\Big|\, x_1,\ldots,x_n : M \subset \bigcup_{i=1}^n B(x_i,\varepsilon)\Big\}.$$

Here $B(x,\varepsilon)$ denotes the closed ball with center x and radius $\varepsilon > 0$. For a bounded linear operator S between Banach spaces E and F we define $\mathcal{N}(S,\varepsilon) := \mathcal{N}(SB_E,\varepsilon)$ where B_E denotes the closed unit ball of E. Usually, it is more convenient to consider the logarithmic covering numbers

$$\mathcal{H}\big((M,d),\varepsilon\big) := \ln \mathcal{N}\big((M,d),\varepsilon\big)$$

and

$$\mathcal{H}(T,\varepsilon) := \ln \mathcal{N}(T,\varepsilon)$$

instead. Finally, recall, that there also exists a concept—the so-called entropy numbers—which is "inverse" to the above notions. For details we refer to [29].

In the following, we also have to measure the continuity of a given loss function L. To this end, we use the inverted modulus of continuity which is defined by

$$\omega^{-1}(L,\varepsilon) := \sup\big\{\delta > 0 : \omega(L,\delta) \le \varepsilon\big\}$$

where

$$\omega(L,\delta) := \sup_{\substack{y\in Y, t,t'\in\mathbb{R}\\ |t-t'|\le\delta}} |L(y,t) - L(y,t')|$$

denotes the modulus of continuity of L. Analogously, the (inverted) modulus of continuity of $L_{|Y\times[-a,a]}$ is defined. Note that in our specific situation the modulus of continuity is continuous in δ (see [29, Proposition 5.4.2]). In particular, the supremum in the definition of ω^{-1} is attained, i.e., we have $\omega\big(L,\omega^{-1}(L,\varepsilon)\big) \le \varepsilon$ for all $\varepsilon > 0$. This also holds for restrictions of L to $Y \times [-a,a]$, $a > 0$.

Now we can state a concentration inequality for classifiers based on (2). The proof of this inequality which is similar to the methods of [30] for the square loss function in regression scenarios can be found in Section IV.

Lemma 3.4: Let k be a continuous kernel on X, L be an admissible loss function, and Ω be a regularization function. Then for all Borel probability measures P on $X \times Y$, all $\varepsilon > 0$, $\lambda > 0$, and all $n \ge 1$ we have

$$\Pr^*\big(T \in (X\times Y)^n : |\mathcal{R}_{L,T}(f_{T,\lambda}) - \mathcal{R}_{L,P}(f_{T,\lambda})| \ge \varepsilon\big) \le 2e^{\mathcal{H}\left(\delta_\lambda I, \omega^{-1}(L_\lambda,\varepsilon/3)\right) - \frac{2\varepsilon^2 n}{9\|L_\lambda\|_\infty^2}}$$

where $I : H \to C(X)$ denotes the canonical embedding and $\Pr^*$ is the outer probability measure of P^n.

Now, we can establish our first consistency result.

Theorem 3.5: Let k be a universal kernel on X, L be an admissible loss function, and Ω be a regularization function. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n}\, \mathcal{H}\big(\delta_{\lambda_n} I, \omega^{-1}(L_{\lambda_n},\varepsilon)\big) \to 0$$

for all $\varepsilon > 0$. Then the classifier based on (2) with respect to k, Ω, L, and (λ_n) is universally consistent. If we additionally have

$$\sum_{n=1}^\infty \exp\Big(-\varepsilon n/\|L_{\lambda_n}\|_\infty^2\Big) < \infty$$

for all $\varepsilon > 0$, then the classifier is even strongly universally consistent.

Proof: Let us define

$$c := \sup\{|L(y,t) - L(y,t')| : y \in Y,\ t,t' \in [-\hat{\delta}K, \hat{\delta}K]\}$$

where $\hat{\delta}$ is defined as in Section III-A. For later purpose we note that

$$\sup\{\omega^{-1}(L_\lambda,\varepsilon) : \lambda \in (0,1],\ \varepsilon \in (0,c)\} < \infty. \qquad (13)$$

Now, let $\varepsilon \in (0, c)$. We fix $\delta > 0$ according to Proposition 3.3. Then by Proposition 3.2, there exists an integer $n_0 \geq 1$ such that for all $n \geq n_0$ we have

$$\left|\mathcal{R}^{\text{reg}}_{L,P,\lambda_n}(f_{P,\lambda_n}) - \mathcal{R}_{L,P}\right| \leq \delta/3. \tag{14}$$

Moreover, (13) guarantees that we can assume without loss of generality that there exists a $\rho > 0$ such that for all $n \geq n_0$ we have

$$\mathcal{H}\big(\delta_{\lambda_n} I, \omega^{-1}(L_{\lambda_n}, \varepsilon)\big) \geq \rho.$$

Therefore, by our assumption on (λ_n) together with Lemma 3.4 and Hoeffding's inequality applied to $\mathcal{R}_{L,T}(f_{P,\lambda_n})$ we may additionally assume

$$\Pr^*\Big(T : \begin{matrix}|\mathcal{R}_{L,T}(f_{T,\lambda_n}) - \mathcal{R}_{L,P}(f_{T,\lambda_n})| > \delta/3 \text{ or} \\ |\mathcal{R}_{L,T}(f_{P,\lambda_n}) - \mathcal{R}_{L,P}(f_{P,\lambda_n})| > \delta/3\end{matrix}\Big) \leq \varepsilon \tag{15}$$

for all $n \geq n_0$. Now, if T belongs to set of samples considered in inequality (15) we find

$$\begin{aligned}\mathcal{R}_{L,P}(f_{T,\lambda_n}) &\leq \Omega(\lambda, \|f_{T,\lambda_n}\|) + \mathcal{R}_{L,P}(f_{T,\lambda_n}) \\ &\leq \Omega(\lambda, \|f_{T,\lambda_n}\|) + \mathcal{R}_{L,T}(f_{T,\lambda_n}) + \delta/3 \\ &\leq \Omega(\lambda, \|f_{P,\lambda_n}\|) + \mathcal{R}_{L,T}(f_{P,\lambda_n}) + \delta/3 \\ &\leq \Omega(\lambda, \|f_{P,\lambda_n}\|) + \mathcal{R}_{L,P}(f_{P,\lambda_n}) + 2\delta/3 \\ &\leq \mathcal{R}_{L,P} + \delta\end{aligned}$$

where the last estimate is due to (14). The definition of δ then gives the first assertion. The second assertion follows by the specific form of the tail bound in Lemma 3.4 and Hoeffding's inequality. □

Our next goal is to simplify the above theorem. To this end, let us first introduce some notions for loss functions: we say that an admissible loss function L is *convex* if $L(y, \cdot)$ is convex for $y = \pm 1$. Note that almost all loss functions considered in the examples are convex. Moreover, L is said to be 1-*Hölder-continuous* if

$$\sup\left\{\frac{|L(y,t) - L(y,t')|}{|t - t'|} : y \in Y,\ t, t' \in \mathbb{R},\ t \neq t'\right\} < \infty.$$

In this case, we denote the supremum by $|L|_1$. Analogously, we say that L is *locally* 1-Hölder-continuous if $L_{|Y\times[-a,a]}$ is 1-Hölder-continuous for all $a > 0$. Recall, that convex functions on $\mathbb{R}$ are locally 1-Hölder-continuous and hence convex loss functions are locally 1-Hölder-continuous. Furthermore note, that for locally 1-Hölder-continuous L we have

$$\omega(L_{|Y\times[-a,a]}, \delta) \leq \delta |L_{|Y\times[-a,a]}|_1, \quad \text{for all } a, \delta > 0$$

and hence,

$$\varepsilon\, |L_{|Y\times[-a,a]}|_1^{-1} \leq \omega^{-1}(L_{|Y\times[-a,a]}, \varepsilon), \quad \text{for all } a, \varepsilon > 0.$$

In view of the announced simplification we also have to recall that for various "smooth" kernels (see the Appendix for the different notions of smoothness) there exist bounds on the covering numbers of the embedding $I : H \to C(X)$ (see Lemma 4.2). Applying these bounds to Theorem 3.5 we now obtain the following corollary whose proof can be found in Section IV.

Corollary 3.6: Let k be a universal kernel on the closure of a bounded C^∞-domain $X \subset \mathbb{R}^d$, L be an admissible loss function which is locally 1-Hölder-continuous, and Ω a regularization function. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\begin{aligned}&\frac{\|L_{\lambda_n}\|_\infty^2}{n} \cdot \big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)^{\frac{2d}{d+2\beta}} \to 0 \text{ if } k \text{ is } \beta\text{-Hölder-continuous} \\ &\frac{\|L_{\lambda_n}\|_\infty^2}{n} \cdot \big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)^{\frac{d}{r}} \to 0 \text{ if } k \in C^{r,r}(\overline{X}, \overline{X}), r \in \mathbb{N} \\ &\frac{\|L_{\lambda_n}\|_\infty^2}{n} \cdot \big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)^{\frac{d}{s}} \to 0 \text{ if } H_k \cong W_p^s(X),\ s > 0 \\ &\frac{\|L_{\lambda_n}\|_\infty^2}{n} \cdot \big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)^{\alpha} \to 0 \text{ for some } \alpha > 0 \text{ if } k \in C^{\infty,\infty}(\overline{X}, \overline{X}).\end{aligned}$$

Then the classifier based on (2) with respect to k, Ω, L, and (λ_n) is strongly universally consistent.

Examples of kernels that satisfy one of the above smoothness assumptions can be found in [31]. Here, we only consider a specific class of universal kernels:

Example 3.7: Let $r > 0$ and $f : (-r, r) \to \mathbb{R}$ be a function that can be expressed by its Taylor series in 0, i.e.,

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

for all $x \in (-r, r)$. Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\}$. If we have $a_n > 0$ for all $n \geq 0$ then $k(x, x') := f(\langle x, x'\rangle)$ defines a universal kernel on every compact subset of X (cf. [22]). Moreover, we obviously have $k \in C^{\infty,\infty}(X, X)$. Furthermore, these statements also hold for the normalized version

$$k^*(x, x') := k(x, x')/\sqrt{k(x, x)k(x', x')}$$

of k.

Besides the Gaussian RBF kernel, Vovk's infinite polynomial kernel $k(x, x') := (1 - \langle x, x'\rangle)^{-\alpha}$ is a well-known example from this class of kernels (cf. [6] and [22]).

Example 3.8: As already mentioned, the Gaussian RBF kernel on $X \subset \mathbb{R}^d$ also fits into the framework of Example 3.7. For this kernel, however, a sharper upper bound for the covering numbers of I was recently shown in [32]. Namely, there was proved that

$$\mathcal{H}(I, \varepsilon) \preceq \left(\log \frac{1}{\varepsilon}\right)^{d+1}.$$

Consequently, the classifier considered in Corollary 3.6 is universal consistent if $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n} \cdot \Big(\log\big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)\Big)^{d+1} \to 0.$$

Finally, we consider classifiers based on (3). As already indicated, we only have to replace Proposition 3.4 by a suitable concentration inequality. It turns out that the main difficulty for this is to derive bounds for the offset (cf. Lemma 4.3). To this end we need the following definition:

Definition 3.9: An admissible loss function L is called *regular* if L is locally 1-Hölder-continuous, $L(1, \cdot)$ is monotone decreasing and unbounded on $(-\infty, 0]$, $L(-1, \cdot)$ is monotone

increasing and unbounded on $[0,\infty)$, and for all $\gamma > 0$ there exists a constant $c_\gamma > 0$ such that for all $a > 0$ we have

$$\begin{aligned}\bigl|L_{|Y\times[-\gamma a,\gamma a]}\bigr|_1 &\le c_\gamma \bigl|L_{|Y\times[-a,a]}\bigr|_1\\ \bigl\|L_{|Y\times[-\gamma a,\gamma a]}\bigr\|_\infty &\le c_\gamma \bigl\|L_{|Y\times[-a,a]}\bigr\|_\infty .\end{aligned}$$

Note that all loss functions considered in the Examples 1.1–1.6 are regular. The next lemma states the result of Lemma 3.1 for modified regularized risks of the form (3). Its proof can be found in Section IV.

Lemma 3.10: Let L be a regular loss function, Ω a regularization function, and k be a continuous kernel on X. Then for all Borel probability measures P on $X \times Y$ and all $\lambda > 0$ there is a pair $(\tilde f_{P,\lambda}, \tilde b_{P,\lambda}) \in H \times \overline{\mathbb{R}}$ with

$$\begin{aligned}\Omega\bigl(\lambda, \|\tilde f_{P,\lambda}\|\bigr) + \mathcal{R}_{L,P}(\tilde f_{P,\lambda} + \tilde b_{P,\lambda})\\ = \inf_{\substack{f\in H\\ b\in\mathbb{R}}} \Omega\bigl(\lambda, \|f\|\bigr) + \mathcal{R}_{L,P}(f+b). \quad (16)\end{aligned}$$

Moreover, for all such pairs we have $\|\tilde f_{P,\lambda}\| \le \delta_\lambda$.

Now we can state the announced concentration inequality for classifiers based on (3).

Lemma 3.11: Let k be a universal kernel on X, L be a regular loss function, and Ω be a regularization function. Then for all Borel probability measures P on $X \times Y$ there exists a constant $c > 0$ such that for all $\varepsilon \in (0,1]$, $\lambda \in (0,1]$, and all $n \ge 1$ we have

$$\begin{aligned}\Pr^*\bigl(T : |\mathcal{R}_{L,T}(\tilde f_{T,\lambda} + \tilde b_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde f_{T,\lambda} + \tilde b_{T,\lambda})| \ge \varepsilon\bigr)\\ \le 2e^{2\mathcal{H}\left(I, \frac{c\varepsilon}{|L_\lambda|_1\delta_\lambda}\right) - \frac{c\varepsilon^2 n}{\|L_\lambda\|_\infty^2}}\end{aligned}$$

where again $\Pr^*$ is the outer probability measure of P^n.

The proof of this lemma can be found in Section IV. Proceeding as in the proof of Theorem 3.5, the result for classifiers based on (3) now reads as follows.

Theorem 3.12: Let k be a universal kernel on X, L be a regular loss function, and Ω be a regularization function. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n}\,\mathcal{H}\left(I, \frac{\varepsilon}{\delta_{\lambda_n}|L_{\lambda_n}|_1}\right) \to 0$$

for all $\varepsilon > 0$. Then the classifier based on (3) with respect to k, Ω, L, and (λ_n) is universally consistent. If we additionally have

$$\sum_{n=1}^{\infty} \exp(-\varepsilon n/\|L_{\lambda_n}\|_\infty^2) < \infty$$

for all $\varepsilon > 0$, then the classifier is even strongly universally consistent.

Using known estimates for covering numbers which are collected in Lemma 4.2 we can immediately derive the results of Corollary 3.6 for classifiers with offset and regular loss functions.

C. Consistency Results Based on Localized Covering Numbers

Instead of using a concentration inequality that is based on the covering numbers of $I : H \to C(X)$, we can also use concentration inequalities which are based on localized covering numbers. Let us first recall their definition. To this end, let ℓ_∞^n denote the space $\mathbb{R}^n$ equipped with the maximum norm. Then, for a given set $\mathcal{F}$ of functions from X to $\mathbb{R}$ the localized covering numbers of $\mathcal{F}$ are defined by

$$\mathcal{N}(\mathcal{F}, n, \varepsilon) := \sup\Bigl\{\mathcal{N}\bigl((\mathcal{F}_{|X_0}, \ell_\infty^{|X_0|}), \varepsilon\bigr) : X_0 \subset X,\ |X_0| \le n\Bigr\}$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$. Here, $\mathcal{F}_{|X_0} := \{f_{|X_0} : f \in \mathcal{F}\}$ is considered as a subset of $\ell_\infty^{|X_0|}$. Analogously to the definition of covering numbers, we also define $\mathcal{H}\bigl(\mathcal{F}, n, \varepsilon\bigr)$ and $\mathcal{H}(S, n, \varepsilon)$, where S is an operator mapping into a space of functions. Now, the announced concentration inequality reads as follows.

Lemma 3.13: Let k be a continuous kernel on X, L be an admissible loss function, and Ω be a regularization function. Then for all Borel probability measures P on $X \times Y$, all $\varepsilon > 0$, $\lambda > 0$, and all $n \ge 1$ we have

$$\begin{aligned}\Pr^*\bigl(T \in (X\times Y)^n : |\mathcal{R}_{L,T}(f_{T,\lambda}) - \mathcal{R}_{L,P}(f_{T,\lambda})| \ge \varepsilon\bigr)\\ \le 12n\, e^{\mathcal{H}\left(\delta_\lambda I, 2n, \omega^{-1}(L_\lambda, \varepsilon/6)\right) - \frac{\varepsilon^2 n}{36\|L_\lambda\|_\infty^2}}.\end{aligned}$$

The proof of this lemma which is mainly due to [33] can be found in Section IV. The consistency result that is derived from the above lemma is stated in the following theorem. Its proof is obvious.

Theorem 3.14: Let k be a universal kernel on X, L be an admissible loss function, and Ω be a regularization function. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n}\Bigl(\log n + \mathcal{H}\bigl(\delta_{\lambda_n} I, 2n, \omega^{-1}(L_{\lambda_n}, \varepsilon)\bigr)\Bigr) \to 0$$

for all $\varepsilon > 0$ Then the classifier based on (2) with respect to k, Ω, L, and (λ_n) is universally consistent. If we additionally have

$$\sum_{n=1}^{\infty} \exp(-\varepsilon n/\|L_{\lambda_n}\|_\infty^2) < \infty$$

for all $\varepsilon > 0$ then the classifier is even strongly universally consistent.

Using the dual version of the Maurey-Carl inequality (see [34]) we obtain the next corollary which provides a kernel independent condition on (λ_n). Details of its proof can be found in Section IV.

Corollary 3.15: Let k be a universal kernel, L be an admissible, locally 1-Hölder-continuous loss function, and Ω be a regularization function. Suppose we have a positive null sequence (λ_n) with

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n}\,\bigl(\delta_{\lambda_n}|L_{\lambda_n}|_1\bigr)^2 \log n \to 0.$$

Then the classifier based on (2) with respect to k, Ω, L, and (λ_n) is strongly universally consistent.

Let us now demonstrate how specific properties of the kernel can be used to derive sharper bounds on the localized covering numbers of the corresponding embedding $I : H \to C(X)$.

Example 3.16: Let k be a universal kernel that can be expressed by a series of the form

$$k(x,y) = \sum_{n=0}^{\infty} a_n \Phi_n(x) \Phi_n(y)$$

where $\Phi_n : X \to \mathbb{R}$ are continuous functions that are uniformly bounded with respect to the $\|.\|_\infty$-norm and $(a_n) \in \ell_1$ is a strictly positive sequence. Examples of such kernels can be found in [22]. If we even have $(a_n) \in \ell_p$ for some $0 < p < 1$ then it was shown in [35] that

$$\mathcal{H}(I, n, \varepsilon) \preceq \left(\frac{\log n}{\varepsilon^2}\right)^p.$$

In particular, the classifier considered in Corollary 3.15 is strongly universally consistent if $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n} \left(\delta_{\lambda_n} |L_{\lambda_n}|_1\right)^{2p} \log n \to 0.$$

We now present consistency results using localized covering numbers for classifiers that are based on (3). The used concentration inequality is as follows.

Lemma 3.17: Let k be a universal kernel on X, L be a regular loss function, and Ω be a regularization function. Then for all Borel probability measures P on $X \times Y$ there exists a constant $c > 0$ such that for all $\varepsilon \in (0,1]$, $\lambda \in (0,1]$, and all $n \geq 1$ we have

$$\Pr^*\left(T : |\mathcal{R}_{L,T}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda})| \geq \varepsilon\right) \leq 12n\, e^{2\mathcal{H}\left(I, 2n, \frac{c\varepsilon}{|L_\lambda|_1 \delta_\lambda}\right) - \frac{c\varepsilon^2 n}{\|L_\lambda\|_\infty^2}}.$$

The proof of this lemma is in Section IV. The consistency result corresponding to Lemma 3.17 is as follows.

Theorem 3.18: Let k be a universal kernel on X, L be a regular loss function, and Ω be a regularization function. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n} \left(\log n + \mathcal{H}\left(I, 2n, \frac{\varepsilon}{\delta_{\lambda_n} |L_{\lambda_n}|_1}\right)\right) \to 0$$

for all $\varepsilon > 0$. Then the classifier based on (3) with respect to k, Ω, L, and (λ_n) is universally consistent. If we additionally have

$$\sum_{n=1}^{\infty} \exp(-\varepsilon n / \|L_{\lambda_n}\|_\infty^2) < \infty$$

for all $\varepsilon > 0$ then the classifier is even strongly universally consistent.

Again it is possible to show the results of Corollary 3.15 and Example 3.16 for classifiers with offset. Here, we only state the following.

Corollary 3.19: Let k be a universal kernel on X, L be a regular loss function, and Ω a regularization function. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\frac{\|L_{\lambda_n}\|_\infty^2}{n} \left(\delta_{\lambda_n} |L_{\lambda_n}|_1\right)^2 \log n \to 0.$$

Then the classifier based on (3) with respect to k, Ω, L, and (λ_n) is strongly universally consistent.

D. Consistency Results Based on Stability

In practice, one usually considers convex loss functions and the regularization function $\omega(\lambda, t) = \lambda t^2$ in order to solve (2) efficiently. Since in this case the corresponding classifier is *stable* (cf. the definition below) it turns out that there also exists a *kernel-independent* condition for the regularization sequence which is usually slightly milder than that of Corollary 3.15.

Theorem 3.20: Let k be a universal kernel on X, L be a convex admissible loss function, and $\Omega(\lambda, t) = \lambda t^2$. Suppose we have a positive sequence (λ_n) with $\lambda_n \to 0$ and

$$\frac{n\lambda_n^2}{|L_{\lambda_n}|_1^4} \to \infty.$$

Then the classifier based on (2) with respect to k, Ω, L, and (λ_n) is universally consistent.

In order to prove this kernel-independent condition on (λ_n) we have to recall the notion of stable classifiers (cf. [36]): let

$$T := \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$$

and $(x, y) \in X \times Y$. Moreover, let $T_{i,(x,y)}$ denote the training set that is identical to T apart from the ith sample which is replaced by (x, y). A classifier $\mathcal{C}$ based on the optimization problem (2) is called *stable* with respect to the sequence (β_n) if for all $n \geq 1$, $T \in (X \times Y)^n$, $(x, y), (x', y') \in X \times Y$, and all $i = 1, \ldots, n$, we have

$$\left|L(y', f_T(x')) - L(y', f_{T_{i,(x,y)}}(x'))\right| \leq \beta_n.$$

For stable classifiers there exists a kernel-independent way of estimating the deviation of $\mathcal{R}_{L,T}(f_T)$ from $\mathcal{R}_{L,P}(f_T)$ as the following result in [36] shows.

Lemma 3.21: Let $\mathcal{C}$ be a (β_n)-stable classifier based on (2) with respect to k, L, Ω, and (λ_n). Then for all $n \geq 1$ we have

$$\Pr^*\left(T \in (X \times Y)^n : |\mathcal{R}_{L,T}(f_T) - \mathcal{R}_{L,P}(f_T)| > \varepsilon + \beta_n\right) \leq 2e^{-\frac{\varepsilon^2 n}{2(n\beta_n + \|L_{\lambda_n}\|_\infty)^2}}.$$

Hence, for the proof of Theorem 3.20 we mainly have to check whether our classifiers are (β_n)-stable for a suitable sequence (β_n). This is done in the following lemma which is proved in Section IV.

Lemma 3.22: Let $\mathcal{C}$ be a classifier based on (2) with respect to a convex loss function L, $\Omega(\lambda, t) := \lambda t^2$, and regularization sequence (λ_n). Then $\mathcal{C}$ is $2K^2 |L_{\lambda_n}|_1^2 / (n\lambda_n)$-stable.

Proof of Theorem 3.20: Again, it suffices to replace the tail-bound of the probability of $|\mathcal{R}_{L,T}(f_T) - \mathcal{R}_{L,P}(f_T)| \geq \varepsilon$ in Lemma 3.4 by another suitable result. To make this precise, by Lemma 3.21, Lemma 3.22, and the proof of Theorem 3.5 we only have to ensure

$$\frac{n}{\left(\frac{|L_{\lambda_n}|_1^2}{\lambda_n} + \|L_{\lambda_n}\|_\infty\right)^2} \to \infty.$$

Since we always have $\|L_{\lambda_n}\|_\infty \preceq \delta_{\lambda_n} |L_{\lambda_n}|_1$ and $\delta_{\lambda_n} \sim \frac{1}{\sqrt{\lambda_n}}$ this follows by the assumption on (λ_n). □

Unfortunately, we cannot prove a kernel-independent condition on (λ_n) for classifiers based on (3) in the spirit of Theorem 3.20. The reason for this lack is that Theorem 3.20 is based on the notion of stability. Unlike classifiers that are based on (2), classifiers based on (3) are not sufficiently stable, in general. This can be easily checked for the L1-SVM with offset (cf. Example 1.1) on $X = \{0\}$.

IV. Proofs

Proof of Lemma 3.1: For all $\varepsilon \in \big(0, L(1,0) + L(-1,0)\big]$ we fix an element $f_\varepsilon \in H$ with

$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_\varepsilon) \le \inf_{f\in H} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f) + \varepsilon.$$

Since

$$\begin{aligned}\Omega(\lambda, \|f_\varepsilon\|) &\le \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_\varepsilon)\\ &\le \mathcal{R}^{\text{reg}}_{L,P,\lambda}(0) + \varepsilon \le 2\big(L(1,0) + L(-1,0)\big)\end{aligned}$$

there exists a $\delta > 0$ with $\|f_\varepsilon\| \le \delta$. Now by the Eberlein–Smulyan theorem and the Bolzano–Weierstraß theorem there exist elements $f_{P,\lambda} \in \delta B_H$, $c \in [0,\delta]$, and a sequence (f_{ε_n}) such that $\|f_{\varepsilon_n}\| \to c$ and $f_{\varepsilon_n} \to f_{P,\lambda}$ weakly. In particular, by the continuity of L and the reproducing property of H we obtain

$$L\big(y, f_{\varepsilon_n}(x)\big) \to L\big(y, f_{P,\lambda}(x)\big)$$

for all $(x,y) \in X \times Y$. Moreover, we have $\|f_{\varepsilon_n}\|_\infty \le \delta K$ and hence $\big|L(\cdot, f_{\varepsilon_n}(\cdot))\big|$ is bounded by the continuity of L. Therefore, Lebesgue's theorem implies

$$\mathcal{R}_{L,P}(f_{\varepsilon_n}) \to \mathcal{R}_{L,P}(f_{P,\lambda}). \tag{17}$$

Thus, for a fixed $\rho > 0$, there exists an index n_0 such that for all $n \ge n_0$ we have both $\varepsilon_n \le \rho$ and

$$\begin{aligned}\Omega(\lambda, \|f_{\varepsilon_n}\|) &+ \mathcal{R}_{L,P}(f_{P,\lambda}) - \rho\\ &\le \Omega(\lambda, \|f_{\varepsilon_n}\|) + \mathcal{R}_{L,P}(f_{\varepsilon_n})\\ &\le \Omega(\lambda, \|f_{P,\lambda}\|) + \mathcal{R}_{L,P}(f_{P,\lambda}) + \varepsilon_n,\end{aligned}$$

where for the latter inequality the definition of f_{ε_n} was used. Hence, we find

$$\lim_{n\to\infty} \Omega(\lambda, \|f_{\varepsilon_n}\|) \le \Omega(\lambda, \|f_{P,\lambda}\|).$$

Moreover, we always have $\|f_{P,\lambda}\| \le \liminf_{n\to\infty} \|f_{\varepsilon_n}\| = c$ (cf. [37, Ch. 1 Corollary 2.6.]) and thus,

$$\Omega(\lambda, \|f_{P,\lambda}\|) \le \Omega(\lambda, c) = \lim_{n\to\infty} \Omega(\lambda, \|f_{\varepsilon_n}\|).$$

This together with (17) yields

$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{\varepsilon_n}) \to \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}).$$

Since the construction of f_ε implies

$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{\varepsilon_n}) \to \inf_{f\in H} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f)$$

the first assertion follows. The second assertion is trivial. □

For the proof of Proposition 3.2 we need the following technical lemma.

Lemma 4.1: There is a measurable function $f^* : [0,1] \to \overline{\mathbb{R}}$ with $M(\alpha) = C(\alpha, f^*(\alpha))$ for all $\alpha \in [0,1]$. In particular, we have

$$\begin{aligned}\mathcal{R}_{L,P} &= \inf\big\{\, \mathcal{R}_{L,P}(f) \mid f : X \to \overline{\mathbb{R}} \text{ measurable}\big\}\\ &= \int_X M\big(f^*(P(1 \mid x))\big)\, P_X(dx).\end{aligned}$$

Proof: Let (α_i) be a dense sequence in $[0,1]$ with $\alpha_1 = 0$ and $\alpha_i \ne \alpha_j$ if $i \ne j$. We fix a solution $t_i \in \overline{\mathbb{R}}$ of (12) for every α_i, $i \ge 1$, and define $f_n(\alpha) := t_i$ if i is the index such that α_i is the closest lower bound of α in the set $\{\alpha_1, \dots, \alpha_n\}$. Then $f : [0,1] \to \overline{\mathbb{R}}$ defined by $f(\alpha) := \liminf f_n(\alpha)$ is measurable. Now fix a real number $\alpha \in (0,1)$ at which M is continuous. By our construction, there exists subsequences (α_{i_j}) and (t_{i_j}) with $a_{i_j} \to \alpha$ and $t_{i_j} \to f(\alpha)$. The continuity of L implies

$$\begin{aligned}M(\alpha) &= \lim_{j\to\infty} M(\alpha_{i_j})\\ &= \lim_{j\to\infty} \alpha_{i_j} L(1, t_{i_j}) + (1-\alpha_{i_j}) L(-1, t_{i_j})\\ &= \alpha L(1, f(\alpha)) + (1-\alpha) L(-1, f(\alpha))\end{aligned}$$

i.e., $f(\alpha)$ is a solution of (12) for α. Since M is concave it is also continuous for all but at most countably many points of $[0,1]$ (cf. [38, Theorem 1.16]) and, thus, we only have to modify f on countably many points in order to construct f^*. □

Proof of Proposition 3.2: Let $\varepsilon > 0$ and $f_\varepsilon \in H$ with

$$\mathcal{R}_{L,P}(f_\varepsilon) \le \inf\{\mathcal{R}_{L,P}(f) : f \in H\} + \varepsilon.$$

Since $\Omega(\cdot, \|f_\varepsilon\|)$ is continuous in 0 there exists λ_0 such that $\Omega(\lambda, \|f_\varepsilon\|) \le \varepsilon$ for all $\lambda \le \lambda_0$. This together with the definition of f_ε yields

$$\lim_{\lambda\to 0} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \inf_{f\in H} \mathcal{R}_{L,P}(f).$$

Thus, in order to prove the assertion, it suffices to show

$$\inf_{f\in H} \mathcal{R}_{L,P}(f) = \inf\big\{\mathcal{R}_{L,P}(f) \colon f \in L_\infty(P_X)\big\} = \mathcal{R}_{L,P}. \tag{18}$$

For a proof of the first identity recall that k is universal. Thus, for all $\varepsilon, \delta > 0$ and every bounded measurable function $h : X \to \mathbb{R}$ there exists an element $f \in H$ with

$$P_X\big(\{x \in X : |h(x) - f(x)| \ge \varepsilon\}\big) \le \delta$$

and $\|f\|_\infty \le \|h\|_\infty$. Since L is uniformly continuous on $Y \times [-\|h\|_\infty, \|h\|_\infty]$ the first identity then follows.

We now show the second identity of (18): let $f_n : [0,1] \to \overline{\mathbb{R}}$ be defined by $f_n(\alpha) := f^*(\alpha)$ if $|f^*(\alpha)| \le n$, $f_n(\alpha) := n \operatorname{sign} f^*(\alpha)$ if $|f^*(\alpha)| = \infty$, and $f_n(\alpha) = 0$ otherwise. Moreover, we define

$$\begin{aligned}M_n(\alpha) &:= \alpha L\big(1, f_n(\alpha)\big) + (1-\alpha) L\big(-1, f_n(\alpha)\big)\\ &= C\big(\alpha, f_n(\alpha)\big).\end{aligned}$$

Note, that for all $\alpha \in [0,1]$ with $|f^*(\alpha)| < \infty$ this definition yields

$$\begin{aligned}M_n(\alpha) = \mathbf{1}_{[-n,n]}\big(f^*(\alpha)\big) M(\alpha)\\ + \mathbf{1}_{\mathbb{R}\setminus[-n,n]}\big(f^*(\alpha)\big)\big(\alpha L(1,0) + (1-\alpha) L(-1,0)\big).\end{aligned}$$

Since we always have $M(\alpha) \le \alpha L(1,0)+(1-\alpha)L(-1,0)$, we thus find that $M_n(\alpha) \searrow M(\alpha)$ for all $\alpha \in [0,1]$ with $|f^*(\alpha)| < \infty$. Now, let $\alpha \in [0,1]$ with $|f^*(\alpha)| = \infty$. By the definition, we obtain

$$|M_n(\alpha) - M(\alpha)| \le \alpha|L(1,n) - L(1,\infty)| + (1-\alpha)|L(-1,n) - L(-1,\infty)| \quad (19)$$

if $f^*(\alpha) = \infty$ and

$$|M_n(\alpha) - M(\alpha)| \le \alpha|L(1,-n) - L(1,-\infty)| + (1-\alpha)|L(-1,-n) - L(-1,-\infty)| \quad (20)$$

if $f^*(\alpha) = -\infty$. Since L is assumed to be continuous on $X \times \overline{\mathbb{R}}$, we observe that the right-hand sides of (19) and (20) converge to 0 if they are finite for all n. To check the latter let us first consider the case $\alpha \in (0,1)$ with $f^*(\alpha) = \infty$. Then we have

$$\begin{aligned} &\alpha L(1,\infty) + (1-\alpha)L(-1,\infty) \\ &\quad = \inf_{t\in\mathbb{R}} \alpha L(1,t) + (1-\alpha)L(-1,t) \\ &\quad \le \alpha L(1,0) + (1-\alpha)L(-1,0) \\ &\quad < \infty \end{aligned}$$

and thus, $L(1,\infty), L(-1,\infty) \in \mathbb{R}$. Hence, $|M_n(\alpha) - M(\alpha)| \to 0$ in this case. Obviously, this also holds for $\alpha \in (0,1)$ with $f^*(\alpha) = -\infty$. Recalling our convention $0 \cdot \infty = 0$ we also have $L(1,\infty) < \infty$ or $L(-1,-\infty) < \infty$ if $f^*(1) = \infty$ or $f^*(0) = -\infty$, respectively, and thus, we can ensure the above convergence in these cases as well. Because of the specific form of the right-hand sides of (19) and (20), we even obtain that M_n converges uniformly to M on $\{\alpha \in [0,1] : |f^*(\alpha)| = \infty\}$. Together with the dominated convergence on the complement we therefore find

$$\begin{aligned} \mathcal{R}_{L,P} &= \int_X M\big(P(1 \mid x)\big) P_X(dx) \\ &= \lim_{n\to\infty} \int_X M_n\big(P(1 \mid x)\big) P_X(dx) \\ &\ge \inf\big\{\mathcal{R}_{L,P}(h) \mid h : X \to \mathbb{R} \text{ bounded, measurable}\big\}. \end{aligned}$$ □

Proof of Proposition 3.3: For $f : X \to \overline{\mathbb{R}}$, the set of points misclassified by f is defined as

$$E_f := \big\{x \in X : P(1|x) < \frac{1}{2} \text{ and } f(x) > 0\big\} \cup \big\{x \in X : P(1|x) > \frac{1}{2} \text{ and } f(x) < 0\big\}.$$

Given a measurable function $f^* : [0,1] \to \overline{\mathbb{R}}$ according to Lemma 4.1 we obtain

$$\begin{aligned} &\mathcal{R}_{L,P}(f) \\ &\ge \int_{X\setminus E_f} \sum_{y\in Y} L\big(y, f^*(P(1 \mid x))\big) P(y \mid x) P_X(dx) \\ &\quad + \int_{E_f} \sum_{y\in Y} L\big(y, f(x)\big) P(y \mid x) P_X(dx) \\ &= \mathcal{R}_{L,P} + \int_{E_f} C\big(P(1 \mid x), f(x)\big) - M\big(P(1 \mid x)\big) P_X(dx). \end{aligned}$$

In order to estimate the second term from below let $f_* : [0,1] \to \overline{\mathbb{R}}$ be a function with $C(\alpha, f_*(\alpha)) = \inf\big\{C(\alpha,t) : t \ge 0\big\}$ if $\alpha < 1/2$ and $C(\alpha, f_*(\alpha)) = \inf\big\{C(\alpha,t) : t \le 0\big\}$ if $\alpha > 1/2$. Using the technique of the proof of Lemma 4.1, we may additionally assume that f_* is measurable. Moreover, since L is admissible we find $C(\alpha, f_*(\alpha)) - M(\alpha) > 0$ for all $\alpha \in [0,1] \setminus \{1/2\}$ and therefore, $\Delta : X \to \mathbb{R}$ defined by

$$\Delta(x) := C\big(P(1 \mid x), f_*(P(1 \mid x))\big) - M\big(P(1 \mid x)\big)$$

is a strictly positive function on $\hat{X} := \{x \in X : P(1 \mid x) \ne 1/2\}$. Furthermore, recall that

$$\mathcal{R}_P(f) = \mathcal{R}_P + \int_E \big(1 - 2s(x)\big) P_X(dx)$$

holds (cf. [1, p. 10]), where s denotes the noise level of P, i.e., $s(x) = \min\{P(1 \mid x), P(-1 \mid x)\}$. Since $1 - 2s$ is also strictly positive on $\hat{X}$ the measures ΔdP_X, $(1-2s)dP_X$, and P_X are absolutely continuous to each other on $\hat{X}$. This yields the assertion. □

Proof of Lemma 3.4: We write $\mathcal{F} := \big\{L(\cdot, f(\cdot)) : f \in \delta_\lambda I B_H\big\}$. By the definition of the modulus of continuity every ε-net $f_1, \ldots, f_n$ of $\delta_\lambda I B_H$ defines an $\omega(L_\lambda, \varepsilon)$-net $L\big(\cdot, f_1(\cdot)\big), \ldots, L\big(\cdot, f_n(\cdot)\big)$ of $\mathcal{F}$ with respect to the supremum norm. Therefore, we have

$$\mathcal{N}(\mathcal{F}, \varepsilon) \le \mathcal{N}\big(\delta_\lambda I, \omega^{-1}(L_\lambda, \varepsilon)\big).$$

Moreover, $\mathcal{F}$ is a subset of $C(X \times Y)$ of nonnegative functions that are bounded by $\|L_\lambda\|_\infty$. Thus, applying Hoeffding's inequality (cf. [1, Theorem 8.1]) yields

$$\begin{aligned} &\Pr^*\Big(T : \sup_{f\in\delta_\lambda B_H} \big|\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)\big| \ge \varepsilon\Big) \\ &\quad \le 2\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{3}\right) e^{-\frac{2\varepsilon^2 n}{9\|L_\lambda\|_\infty^2}} \\ &\quad \le 2e^{\mathcal{H}\left(\delta_\lambda I, \omega^{-1}(L_\lambda, \varepsilon/3)\right) - \frac{2\varepsilon^2 n}{9\|L_\lambda\|_\infty^2}}. \end{aligned}$$

Since Lemma 3.1 guarantees $f_{T,\lambda} \in \delta_\lambda B_H$, the assertion now follows. □

In order to prove Corollary 3.6, we have to recall some estimates for the covering numbers of the embedding $I : H \to C(X)$ which are summarized in the following lemma.

Lemma 4.2: Given a continuous kernel k on the closure of a bounded C^∞-domain $X \subset \mathbb{R}^d$ we have

$$\mathcal{H}(I, \varepsilon) \preceq \begin{cases} \varepsilon^{-\frac{2d}{d+2\beta}}, & \text{if } k \text{ is } \beta\text{-Hölder-continuous} \\ \varepsilon^{-\frac{d}{r}}, & \text{if } k \in C^{r,r}(\overline{X}, \overline{X}) \\ \varepsilon^{-\frac{d}{s}}, & \text{if } H_k \cong W_p^s(X) \\ \varepsilon^{-\alpha}. & \text{for all } \alpha > 0 \text{ if } k \in C^{\infty,\infty}(\overline{X}, \overline{X}). \end{cases}$$

Proof: In the first case, the embedding $I : H \to C(\overline{X}, d_k)$ is 1-Hölder-continuous and therefore the assertion follows by

$$\mathcal{N}\big((\overline{X}, d_k), \varepsilon^\beta\big) \preceq \mathcal{N}\big(\overline{X}, \varepsilon\big).$$

and a result in [39] (see also [29, Ch. 5] and [40]).

If $k \in C^{r,r}(\overline{X} \times \overline{X})$ the embedding I actually maps into $C^r(\overline{X})$ (see [31, p. 42]. Moreover, the embedding $I_r : C^r(X) \to$

$C(X)$ factors through $id : B^m_{\infty,\infty}(X) \to B^0_{\infty,1}(X)$ by embeddings (see [41, Secs. 3.2.4 and 3.3.1]). Thus, we find

$$\mathcal{H}(I_r, \varepsilon) \preceq \varepsilon^{-d/r}$$

by [42, Sec. 3.3.2], i.e., we have shown the second assertion.

In the third case we have to consider the embedding $I : W^s_p(\overline{X}) \to C(\overline{X})$. Corresponding estimates of the covering numbers of this embedding can be found in [42, Ch. 3.3].

Finally, the last assertion is a direct consequence of the second assertion. □

Proof of Corollary 3.6: By Lemma 4.2, we obtain that in every case there exists a $\gamma > 0$ with

$$\mathcal{H}(I, \varepsilon) \preceq \varepsilon^{-\gamma}.$$

Moreover, since $\omega(L_\lambda, \delta) \leq \delta |L_\lambda|_1$ we find $\omega^{-1}(L_\lambda, \varepsilon) \geq \varepsilon / |L_\lambda|_1$. This yields

$$\mathcal{H}\big(\delta_\lambda I, \omega^{-1}(L_\lambda, \varepsilon)\big) \leq \mathcal{H}\left(I, \frac{\varepsilon}{\delta_\lambda |L_\lambda|_1}\right) \preceq \left(\frac{\delta_\lambda |L_\lambda|_1}{\varepsilon}\right)^\gamma.$$

Hence, in order to satisfy the first condition of Theorem 3.5, it suffices to ensure

$$\frac{\|L_{\lambda_n}\|^2_\infty}{n} \cdot \big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)^\gamma \to 0.$$

This is done by the conditions of the corollary together with Lemma 4.2. Moreover, the above convergence yields

$$\big(\delta_{\lambda_n} |L_{\lambda_n}|_1\big)^\gamma \preceq \frac{n}{\|L_{\lambda_n}\|^2_\infty}$$

and since we always have $\|L_{\lambda_n}\|_\infty \preceq \delta_{\lambda_n} |L_{\lambda_n}|_1$ we find

$$n^{1-2/(2+\gamma)} \preceq \frac{n}{\|L_{\lambda_n}\|^2_\infty}.$$

This yields the second condition of Theorem 3.5. □

In order to treat classifiers that are based on (3) we say that a probability measure P on $X \times Y$ is *y-degenerated* for a $y \in Y$, if

$$P_X\big(x \in X : P(y \mid x) = 1\big) = 1. \tag{21}$$

If P is y-degenerated for some $y \in Y$ it is called *degenerated.*

Proof of Lemma 3.10: As in the proof of Lemma 3.1, we fix pairs $(f_\varepsilon, b_\varepsilon) \in H \times \mathbb{R}$ with

$$\Omega\big(\lambda, \|f_\varepsilon\|\big) + \mathcal{R}_{L,P}(f_\varepsilon + b_\varepsilon) \leq \inf_{\substack{f \in H \\ b \in \mathbb{R}}} \Omega\big(\lambda, \|f\|\big) + \mathcal{R}_{L,P}(f + b) + \varepsilon$$

for all $\varepsilon \in \big(0, L(1,0) + L(-1,0)\big]$. Again, we easily get a $\delta > 0$ such that $\|f_\varepsilon\| \leq \delta$ for all such ε. Let us first assume that P is not degenerated, i.e., (21) does not hold. Then, for $A_y := \{x \in X : P(y \mid x) > 0\}$ we find $P_X(A_y) > 0$ for all $y \in Y$. In particular, there exists a $\rho > 0$ such that for $A^\rho_y := \{x \in X : P(y \mid x) > \rho\}$ we have $P_X(A^\rho_y) > 0$ for all $y \in Y$. Moreover, for $y = 1$ we find

$$\begin{aligned}\int_{A^\rho_y} \Big(&P(1 \mid x) L(1, f_\varepsilon(x) + b_\varepsilon) \\ &+ P(-1 \mid x) L(-1, f_\varepsilon(x) + b_\varepsilon) P_X(dx)\Big) \\ &\leq \Omega\big(\lambda, \|f_\varepsilon\|\big) + \mathcal{R}_{L,P}(f_\varepsilon + b_\varepsilon) \\ &\leq 2\big(L(1,0) + L(-1,0)\big).\end{aligned}$$

Therefore, there exists an element $x_\varepsilon \in A^\rho_1$ with

$$L(1, f_\varepsilon(x_\varepsilon) + b_\varepsilon) \leq \frac{2\big(L(1,0) + L(-1,0)\big)}{\rho P_X(A^\rho_1)}. \tag{22}$$

Since L is regular and the right-hand side of (22) is independent of ε there exists a constant $c \in \mathbb{R}$ such that $f_\varepsilon(x_\varepsilon) + b_\varepsilon \geq c$ holds for all ε. Moreover, we have $|f_\varepsilon(x_\varepsilon)| \leq K \|f_\varepsilon\| \leq \delta K$ and thus we find $b_\varepsilon \geq c - \delta K$ for all $\varepsilon > 0$. Analogously, we can bound b_ε uniformly from above. The rest of the proof follows the lines of the proof of Lemma 3.1.

Now let us assume that P is y-degenerated for some $y \in Y$. Then, one easily checks that

$$(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda}) := \big(0, f^*(P(1 \mid x))\big) \in H \times \overline{\mathbb{R}}$$

is a pair fulfilling (16) where $f^*(\alpha) = t_\alpha$, $\alpha \in [0,1]$ is a function according to (12).

Finally, the bound on $\|\tilde{f}_{P,\lambda}\|$ for general solutions of (16) is trivial. □

In order to prove Lemma 3.11 we need the following lemma that shows that the bound for $\tilde{b}_{T,\lambda}$, which was obtained in the proof of Lemma 3.10, is rather sloppy in typical situations.

Lemma 4.3: Let L be a regular and admissible loss function and P be a nondegenerated probability measure on $X \times Y$. Then there exists a constant $c > 0$ such that for all $n \geq 1$ we have

$$\Pr^*\big(T \in (X \times Y)^n : |\tilde{b}_{T,\lambda}| \leq c + \delta_\lambda K \ \forall \lambda > 0\big) \geq 1 - 2e^{-cn}.$$

Considering very specific convex loss functions and regularization functions, the preceding lemma can be improved. Since we are mainly interested in general loss functions we omit the details.

Proof of Lemma 4.3: Since P is nondegenerated, there is a $\rho > 0$ such that for $A^\rho_y := \{x \in X : P(y \mid x) > \rho\}$, $y \in Y$ we have $P_X(A^\rho_y) > 0$. With $c_1 := \frac{\rho}{2} \min\{P_X(A^\rho_1), P_X(A^\rho_{-1})\}$ we obtain by Hoeffding's inequality that

$$P^n\big(T : |\{i : x_i \in A^\rho_y, y_i = y\}| \geq c_1 n\big) \geq 1 - e^{-2c_1^2 n}$$

holds for $y = \pm 1$. Now let $\lambda > 0$ and $T \in (X \times Y)^n$ be a training set with $|\{i : x_i \in A^\rho_1, y_i = 1\}| \geq c_1 n$. Then we have

$$\frac{1}{n} \sum_{\substack{x_i \in A^\rho_1 \\ y_i = 1}} L\big(1, \tilde{f}_{T,\lambda}(x_i) + \tilde{b}_{T,\lambda}\big) \leq L(1,0) + L(-1,0).$$

Since with $m := |\{i : x_i \in A^\rho_1, y_i = 1\}|$ and $l := L(1,0) + L(-1,0)$, this yields

$$\frac{1}{m} \sum_{\substack{x_i \in A^\rho_1 \\ y_i = 1}} L\big(1, \tilde{f}_{T,\lambda}(x_i) + \tilde{b}_{T,\lambda}\big) \leq \frac{nl}{m} \leq \frac{L(1,0) + L(-1,0)}{c_1}$$

there exists an index i with

$$L\big(1, \tilde{f}_{T,\lambda}(x_i) + \tilde{b}_{T,\lambda}\big) \le \frac{L(1,0) + L(-1,0)}{c_1}.$$

Recalling that L is regular and c_1 is independent of T and λ we thus find a constant $c_2 \in \mathbb{R}$ such that $\|\tilde{f}_{T,\lambda}\| + \tilde{b}_{T,\lambda} \ge c_2$ holds for all $\lambda > 0$ and all $T \in (X \times Y)^n$ with $|\{i : x_i \in A_1^\rho, y_i = 1\}| \ge c_1 n$. This yields the desired estimate from below by Lemma 3.10. The estimate from above can be proved analogously. □

The following lemma shows that for degenerated probability measures there is nothing to be done in view of our desired tail-bounds.

Lemma 4.4: Let $y \in Y$ and P be a y-degenerated probability measure on $X \times Y$. Moreover, let L be an admissible loss function and $\lambda > 0$. Then we have

$$P^n(T \in (X \times Y)^n : T \text{ is } y\text{-degenerated}) = 1$$

and $\mathcal{R}_{L,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) = L(y, f^*(y))$.

Proof: The first assertion is trivial. Now let us assume without loss of generality that P is 1-degenerated. Observing that

$$\begin{aligned}
&\Omega\big(\lambda, \|\tilde{f}_{P,\lambda}\|\big) + L(1, f^*(1)) \\
&= \Omega\big(\lambda, \|\tilde{f}_{P,\lambda}\|\big) + \inf_{\substack{f \in H \\ b \in \mathbb{R}}} \int_X L(1, f(x) + b) P_X(dx) \\
&\le \Omega\big(\lambda, \|\tilde{f}_{P,\lambda}\|\big) + \mathcal{R}_{L,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) \\
&\le L(1, f^*(1))
\end{aligned}$$

we can easily deduce the second assertion. □

Proof of Lemma 3.11: By Lemma 4.4, we may assume without loss of generality that P is not degenerated. Let $c_1 > 0$ be chosen accordingly to Lemma 4.3. We define

$$\mathcal{G} := \{f(\cdot) + b : f \in \delta_\lambda I B_H,\ |b| \le c_1 + \delta_\lambda K\}$$

and

$$\mathcal{F} := \{L(\cdot, g(\cdot)) : g \in \mathcal{G}\}.$$

Moreover, we write $a := c_1 + 2\delta_\lambda K$ and $\tilde{L}_\lambda := L_{|Y \times [-a,a]}$ for short. Now, an easy calculation similar to that of the proof of Lemma 3.4 shows that

$$\begin{aligned}
\mathcal{N}(\mathcal{F}, \varepsilon) &\le \mathcal{N}\big(\mathcal{G}, \omega^{-1}(\tilde{L}_\lambda, \varepsilon)\big) \\
&\le \mathcal{N}\left(I, \frac{\varepsilon}{2|\tilde{L}_\lambda|_1 \delta_\lambda}\right) \cdot \mathcal{N}\left([-a,a], \frac{\varepsilon}{2|\tilde{L}_\lambda|_1}\right)
\end{aligned}$$

where $\mathcal{F}$ and $\mathcal{G}$ are considered as subsets of $C(X)$. Hence, with the help of Lemma 4.3 and Hoeffding's inequality we obtain

$$\begin{aligned}
&\Pr^*\big(T : |\mathcal{R}_{L,T}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda})| \le \varepsilon\big) \\
&\qquad \ge 1 - 2\mathcal{N}\big(\mathcal{G}, \omega^{-1}(\tilde{L}_\lambda, \varepsilon/3)\big) e^{-\frac{2\varepsilon^2 n}{9\|\tilde{L}_\lambda\|_\infty^2}} - 2e^{-c_1 n}.
\end{aligned}$$

Since L is regular, there also exists a constant $c_2 > 0$ independent of λ such that $\|\tilde{L}_\lambda\|_\infty \le c_2 \|L_\lambda\|_\infty$ and $|\tilde{L}_\lambda|_1 \le c_2 |L_\lambda|_1$. Therefore, we get

$$\begin{aligned}
&\Pr^*\big(T : |\mathcal{R}_{L,T}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda})| \ge \varepsilon\big) \\
&\le 2e^{\mathcal{H}\left(I, \frac{\varepsilon}{6|\tilde{L}_\lambda|_1 \delta_\lambda}\right) + \mathcal{H}\left([-a,a], \frac{\varepsilon}{6|\tilde{L}_\lambda|_1}\right) - \frac{2\varepsilon^2 n}{9\|\tilde{L}_\lambda\|_\infty^2}} + 2e^{-c_1 n} \\
&\le 2e^{\mathcal{H}\left(I, \frac{c_3 \varepsilon}{|L_\lambda|_1 \delta_\lambda}\right) + \ln\left(\frac{|L_\lambda|_1 \delta_\lambda}{c_3 \varepsilon}\right) - \frac{c_3 \varepsilon^2 n}{\|L_\lambda\|_\infty^2}} + 2e^{-c_1 n}
\end{aligned}$$

for a suitable constant $c_3 > 0$ independent of λ, ε, and n. Since $X \ne \emptyset$ and k is universal, we get rank $I \ge 1$ and thus we have $\ln(1/\delta) \le \mathcal{H}(I, c_4\delta)$ for all $\delta > 0$ and a suitable constant $c_4 > 0$ independent of δ. Thus, we finally find a constant $c > 0$ for which the assertion holds. □

Proof of Lemma 3.13: As in the proof of Lemma 3.4, we define $\mathcal{F} := \{L(\cdot, f(\cdot)) : f \in \delta_\lambda I B_H\}$. Then Lemma 3.4 in [43] yields

$$\begin{aligned}
&\Pr^*\Big(T \in (X \times Y)^n : \sup_{f \in \delta_\lambda B_H} \big|\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)\big| \ge \varepsilon\Big) \\
&\qquad \le 12\, n\, \mathcal{N}(\mathcal{F}, 2n, \varepsilon/6)\, e^{-\frac{\varepsilon^2 n}{36\|L_\lambda\|_\infty^2}}.
\end{aligned}$$

Now the assertion follows by the arguments used in the proof of Lemma 3.4. □

Proof of Corollary 3.15: By the dual Maurey-Carl inequality (cf. [34], [44], and [40]) we have

$$\mathcal{H}(I, n, \varepsilon) \preceq \frac{\log n}{\varepsilon^2}.$$

Therefore, we find

$$\begin{aligned}
\mathcal{H}\big(\delta_{\lambda_n} I, 2n, \omega^{-1}(L_{\lambda_n}, \varepsilon)\big) &\le \mathcal{H}\left(I, 2n, \frac{\varepsilon}{\delta_{\lambda_n} |L_{\lambda_n}|_1}\right) \\
&\preceq \frac{\delta_{\lambda_n}^2 |L_{\lambda_n}|_1^2 \log n}{\varepsilon^2}.
\end{aligned}$$

Since $\delta_{\lambda_n} |L_{\lambda_n}|_1 \to \infty$, the first condition of Theorem 3.14 follows. The summability condition can be derived as in the proof of Corollary 3.6. □

Proof of Lemma 3.17: With the notions of the proof of Lemma 3.11 we find

$$\begin{aligned}
&\Pr^*\big(T : |\mathcal{R}_{L,T}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda})| \ge \varepsilon\big) \\
&\le 12n\mathcal{N}\big(\mathcal{G}, 2n, \omega^{-1}(\tilde{L}_\lambda, \varepsilon/6)\big) e^{-\frac{\varepsilon^2 n}{36\|\tilde{L}_\lambda\|_\infty^2}} + 2e^{-c_1 n} \\
&\le 12n\, e^{\mathcal{H}\left(I, 2n, \frac{c_3 \varepsilon}{|L_\lambda|_1 \delta_\lambda}\right) + \ln\left(\frac{|L_\lambda|_1 \delta_\lambda}{c_3 \varepsilon}\right) - \frac{c_3 \varepsilon^2 n}{\|L_\lambda\|_\infty^2}} + 2e^{-c_1 n}
\end{aligned}$$

by Lemma 3.4 in [43] (cf. the proof of Lemma 3.13). Again, there is also a constant $c_4 > 0$ with $\ln(1/\delta) \le \mathcal{H}(I, 2n, c_4\delta)$ for all $\delta > 0$, $n \ge 1$. Thus, we finally find a constant $c > 0$ for which the assertion holds. □

Proof of Lemma 3.22: For differentiable loss functions the assertion can be found in [4, Theorem 12.4]. In this book, there are also some remarks concerning nondifferentiable loss functions. For the sake of completeness, we present a proof which mainly follows the elegant approach of [45]. For this purpose, we need to recall that for a convex, continuous function $f : \mathbb{R}^d \to \mathbb{R}$ the subdifferential of f in $x \in \mathbb{R}^d$ is defined by

$$\partial f(x) := \{x^* \in \mathbb{R}^d : \langle x^*, y - x\rangle \le f(y) - f(x) \,\forall\, y \in \mathbb{R}^d\}.$$

For basic properties we refer to [37], [38], and in particular [46, Theorems 23.8 and 23.9]. In the following, $\mathbb{E}_T f$ denotes the expectation of f with respect to the empirical measure induced by T. We will also use this notation for H-valued functions f. Now, recall that the convexity of L implies that L_{λ_n} is locally 1-Hölder-continuous. We fix

$$T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$$

and write $S := T_{i,(x,y)}$ for some fixed $i = 1, \ldots, n$ and $(x, y) \in X \times Y$. Since we actually have to consider finite dimensional

subspaces of H only, an easy calculation using [46, Theorems 23.8 and 23.9] shows $\mathcal{R}^{\text{reg}}_{L,T,\lambda_n}(f) = 2\lambda_n f + D$, where

$$D := \{\mathbb{E}_T h\Phi : h(x_i, y_i) \in \partial L(y_i, f(x_i)) \,\forall\, i = 1, \ldots, n\}$$

and the subdifferential ∂L of L is with respect to the second variable only. Recalling that f_{T,λ_n} minimizes $\mathcal{R}^{\text{reg}}_{L,T,\lambda_n}$ we also observe $0 \in \partial\mathcal{R}^{\text{reg}}_{L,T,\lambda_n}(f_{T,\lambda_n})$ and thus, there exists $h(x_i, y_i) \in \partial L\big(y_i, f_{T,\lambda_n}(x_i)\big)$, $i = 1, \ldots, n$ with

$$0 = 2\lambda_n f_{T,\lambda_n} + \mathbb{E}_T h\Phi. \tag{23}$$

By the Lipschitz continuity of L_{λ_n} and the norm bound of f_{T,λ_n} which was shown in Lemma 3.1, we actually have $\|h\|_\infty \le |L_{\lambda_n}|_1$. Moreover, $h(x_i, y_i) \in \partial L(y_i, f_{T,\lambda_n}(x_i))$ implies

$$h(x_i, y_i)\big(f_{S,\lambda_n}(x_i) - f_{T,\lambda_n}(x_i)\big) \le L(y_i, f_{S,\lambda_n}(x_i)) - L(y_i, f_{T,\lambda_n}(x_i))$$

for all $i = 1, \ldots, n$. Recalling the reproducing property of Φ integration with respect to the empirical measure of S then yields

$$\mathbb{E}_S L\big(\cdot, f_{T,\lambda_n}(\cdot)\big) + \langle f_{S,\lambda_n} - f_{T,\lambda_n}, \mathbb{E}_S h\Phi\rangle \le \mathbb{E}_S L\big(\cdot, f_{S,\lambda_n}(\cdot)\big).$$

Now, we have

$$\lambda_n \|f_{T,\lambda_n}\|^2 + 2\lambda_n\langle f_{S,\lambda_n} - f_{T,\lambda_n}, f_{T,\lambda_n}\rangle + \lambda_n \|f_{T,\lambda_n} - f_{S,\lambda_n}\|^2 = \lambda_n \|f_{S,\lambda_n}\|^2$$

and we thus find

$$\mathcal{R}^{\text{reg}}_{L,S,\lambda_n}(f_{S,\lambda_n}) \ge \mathcal{R}^{\text{reg}}_{L,S,\lambda_n}(f_{T,\lambda_n}) + \lambda_n \|f_{T,\lambda_n} - f_{S,\lambda_n}\|^2 + \langle f_{S,\lambda_n} - f_{T,\lambda_n}, \mathbb{E}_S h\Phi + 2\lambda_n f_{T,\lambda_n}\rangle.$$

Moreover, f_{S,λ_n} minimizes $\mathcal{R}^{\text{reg}}_{L,S,\lambda_n}$ and hence we have

$$\mathcal{R}^{\text{reg}}_{L,S,\lambda_n}(f_{T,\lambda_n}) \ge \mathcal{R}^{\text{reg}}_{L,S,\lambda_n}(f_{S,\lambda_n}).$$

This yields

$$\begin{aligned}\lambda_n \|f_{T,\lambda_n} - f_{S,\lambda_n}\|^2 &\le \langle f_{T,\lambda_n} - f_{S,\lambda_n}, \mathbb{E}_S h\Phi + 2\lambda_n f_{T,\lambda_n}\rangle \\ &\le \|f_{T,\lambda_n} - f_{S,\lambda_n}\| \, \|\mathbb{E}_S h\Phi + 2\lambda_n f_{T,\lambda_n}\|.\end{aligned}$$

With the help of (23), we can replace $2\lambda_n f_{T,\lambda_n}$ by $-\mathbb{E}_T h\Phi$. Then we obtain

$$\|f_{T,\lambda_n} - f_{S,\lambda_n}\| \le \frac{1}{\lambda_n}\|\mathbb{E}_S h\Phi - \mathbb{E}_T h\Phi\| \le \frac{2K|L_{\lambda_n}|_1}{n\lambda_n}.$$

It can be easily seen that the latter estimate implies the assertion. □

APPENDIX
BANACH SPACES, OPERATORS, AND KERNELS

Let E be a Banach space and $A \subset E$ be an arbitrary subset. The set A is *closed* if for every sequence $(x_n) \subset A$ with $x_n \to x \in E$ we have $x \in A$. A set is *open* if its complement is closed. The closure $\overline{A}$ of A is the smallest closed set in E that contains A. We say that A is *dense* if $\overline{A} = E$. The set A is *compact* if every open covering of A has a finite subcovering. If E is finite dimensional then the compact subset of E are exactly the sets which are both bounded and closed. In contrast to this, for E being infinite dimensional, the closed unit ball B_E of E is closed and bounded but never compact. However, for some spaces such as Hilbert spaces there is a nontrivial topology on E (smaller than the norm topology) such that B_E is compact with respect to this topology (Alaoglu's theorem). Finally, a linear bounded operator $S : E \to F$ between Banach spaces E and F is *compact* if $\overline{SB_E}$ is compact. This holds if and only if its covering numbers are finite.

A symmetric function $k : X \times X \to \mathbb{R}$ is called *positive semidefinite* if for all $x_1, \ldots, x_n \in X$ and all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ we have

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

We say that k is *positive definite* if this inequality is strict for all mutually different $x_1, \ldots, x_n \in X$ and all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}\setminus\{0\}$. A symmetric, positive semidefinite function is called a *kernel*. Having a kernel one can construct an associated Hilbert space—the so-called RKHS. Indeed, let

$$H_0 := \Big\{\sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in X, i = 1, \ldots, n\Big\}.$$

Then by

$$\Big\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{m} \beta_j k(\hat{x}_j, \cdot)\Big\rangle := \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x_i, \hat{x}_j)$$

we can define an inner product on H_0. In general, the space H_0 equipped with this inner product fails to be complete, i.e., some Cauchy sequences may not converge in H_0. Therefore, instead of using H_0 one considers the completion H of H_0, i.e., the smallest Hilbert space containing H_0. The space H is called the RKHS of k. Obviously, the map $\Phi : X \to H$, $x \mapsto k(x, \cdot)$, is well defined. Furthermore, it satisfies the so-called reproducing property

$$\langle w, \Phi(x)\rangle = w(x)$$

for all $w \in H$, $x \in X$.

In view of Corollary 3.6 we also need some notions of smoothness. If k is a universal kernel

$$d_k(x, x') := \|\Phi(x) - \Phi(x')\|_H$$

$x, x' \in X$ defines a metric on X (cf. [22, Lemma 3]). Due to the continuity of k, the topology generated by d_k coincides with the topology of the original metric d (cf. [22, Lemma 3 and Corollary 7]). Moreover, the embedding $I : H \to C(X)$ is obviously 1-Hölder-continuous with respect to d_k, that is, IB_H is uniformly Lipschitz continuous (cf. [29, Ch. 5]). In particular, I is a compact operator and hence its covering numbers are finite.

A kernel k is called *β-Hölder-continuous* (with respect to d), $0 < \beta \le 1$ if there exists a constant $c > 0$ such that

$$k(x, x) - 2k(x, x') + k(x', x') \le c\, d^{2\beta}(x, x') \tag{24}$$

for all $x, x' \in X$ (cf. [31, p. 135]. Obviously, this condition means that $id : (X, d) \to (X, d_k)$ is β-Hölder continuous and hence I is β-Hölder continuous with respect to d.

In order to consider smooth kernels on bounded C^∞-domains $X \subset \mathbb{R}^d$ (cf. [41, Sec. 3.2]) we also write $C^{r,r}(X \times X)$, $r \in \mathbb{N}$, for the space of all functions $f : X \times X \to \mathbb{R}$ for which $\frac{\partial^{|\alpha_1|+|\alpha_2|} f}{\partial_{\alpha_1}\partial_{\alpha_2}}$ is continuous for all $\alpha_1, \alpha_2 \in \mathbb{N}_0^d$ with $|\alpha_1|, |\alpha_2| \le r$ (cf. [31, p. 40]). Recall, that the open Euclidian balls in $\mathbb{R}^d$ are

C^∞-domains. A function $f : \overline{X} \times \overline{X} \to \mathbb{R}$ is in $C^{r,r}(\overline{X} \times \overline{X})$ if and only if f is continuous and its restriction on $X \times X$ is in $C^{r,r}(X \times X)$. Furthermore, we write $f \in C^{\infty,\infty}(\overline{X} \times \overline{X})$ if $f \in C^{r,r}(\overline{X} \times \overline{X})$ for all $r \in \mathbb{N}$.

Finally, $W_p^s(X)$ denotes the Sobolev space (cf. [41]) and we write $f \in W_p^s(\overline{X})$ if $f : \overline{X} \to \mathbb{R}$ is continuous and $f_{|X} \in W_p^s(X)$.

REFERENCES

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

[2] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," in *Proc. 14th Annu. Conf. Computational Learning Theory*, vol. 2111, Lecture Notes in Artificial Intelligence, 2001, pp. 416–426.

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

[4] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[5] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[6] C. Saunders, M. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola, "Support Vector Machine—Reference Manual," Univ. London, Royal Holloway, Dept. Computer Sci., London, U.K., Tech. Rep. CSD-TR-98-03, 1998. Also available [Online] at http://eprints.ecs.soton.ac.uk/archive/00008959/01/SVM_Reference.pdf.

[7] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, pp. 56–134, 2004.

[8] I. Steinwart, "Support vector machines are universally consistent," *J. Complexity*, vol. 18, pp. 768–791, 2002.

[9] ——, "On the optimal parameter choice for ν-support vector machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 1274–1284, Oct. 2003.

[10] O. Mangasarian and D. Musicant, "Lagrangian support vector machines," *J. Mach. Learn. Res.*, vol. 1, pp. 161–177, 2001.

[11] J. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, pp. 293–300, 1999.

[12] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Comput.*, vol. 7, pp. 219–269, 1995.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[14] Y. Lee and O. Mangasarian, "SSVM: A smooth support vector machine for classification," *Comput. Optim. Appl.*, vol. 20, pp. 5–22, 2001.

[15] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 171–204.

[16] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[17] D. Johnson and F. Preparata, "The densest hemisphere problem," *Theor. Comput. Sci.*, vol. 6, pp. 93–107, 1978.

[18] K.-U. Höffgen and H.-U. Simon, "Robust trainability of single neurons," in *Proc. Computational Learning Theory (COLT) Conf.*, 1992, pp. 428–438.

[19] I. Steinwart. (2002) "Which Data–Dependent Bounds are Suitable for SVM's?". [Online]. Available: http://www.c3.lanl.gov/~ingo/publications/bounds.ps

[20] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, 2002.

[21] Y. Lin, "Support vector machines and the Bayes rule in classification," *Data Mining Knowledge Disc.*, vol. 6, pp. 259–275, 2002.

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

[23] D. Cox and F. O'Sullivan, "Asymptotic analysis of penalized likelihood and related estimators," *Ann. Statist.*, vol. 18, pp. 1676–1695, 1990.

[24] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

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

[26] R. Dudley, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, pp. 899–929, 1978.

[27] I. Steinwart, "Sparseness of support vector machines," *J. Mach. Learn. Res.*, vol. 4, pp. 1071–1105, 2003.

[28] Y. Lin, "A note on margin-based loss functions in classification," *Statist. Probab. Lett.*, vol. 68, pp. 73–82, 2004.

[29] B. Carl and I. Stephani, *Entropy, Compactness and the Approximation of Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

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

[31] K. Ritter, *Average-Case Analysis of Numerical Problems*. Berlin, Germany: Springer-Verlag, 2000, vol. 1733, Lecture Notes in Mathematics.

[32] D. Zhou, "The covering number in learning theory," *J. Complexity*, vol. 18, pp. 739–767, 2002.

[33] R. Williamson, A. Smola, and B. Schölkopf, "Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators," *IEEE Trans. Inf. Theory*, vol. 47, pp. 2516–2532, Sep. 2001.

[34] B. Carl and A. Pajor, "Gelfand numbers of operators with values in a Hilbert space," *Invent. Math.*, vol. 94, pp. 479–504, 1988.

[35] I. Steinwart, "Entropy numbers of convex hulls and an application to learning algorithms," *Arch. Math.*, vol. 80, pp. 310–318, 2003.

[36] O. Bousquet and A. Elisseeff, "Algorithmic stability and generalization performance," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, vol. 13, pp. 196–202.

[37] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*. Dordrecht, The Netherlands: Reidel, 1986.

[38] R. Phelps, *Convex Functions, Monotone Operators and Differentiability*. Berlin, Germany: Springer-Verlag, 1986, vol. 1364, Lecture Notes in Mathematics.

[39] B. Carl, S. Heinrich, and T. Kühn, "s–numbers of integral operators with Hölder-continuous kernels over metric compacta," *J. Funct. Anal.*, vol. 81, pp. 54–73, 1988.

[40] I. Steinwart, "Entropy of $C(K)$-valued operators," *J. Approx. Theory*, vol. 103, pp. 302–328, 2000.

[41] H. Triebel, *Theory of Function Spaces*. Leipzig, Germany: Akademische Verlagsgesellschaft Geest & Portig, 1983.

[42] D. Edmunds and H. Triebel, *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[43] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *J. Assoc. Comput. Mach.*, vol. 44, pp. 615–631, 1997.

[44] B. Carl, I. Kyrezi, and A. Pajor, "Metric entropy of convex hulls in Banach spaces," *J. London Math. Soc.*, vol. 60, pp. 871–896, 1999.

[45] T. Zhang, "Convergence of large margin separable linear classification," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, vol. 13, pp. 357–363.

[46] R. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.