Consistency of support vector machines and other regularized kernel classifiers

Ingo Steinwart*
Modeling, Algorithms, and Informatics Group, CCS-3
Mail Stop B256
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
ingo@lanl.gov

June 6, 2003

Abstract

We show that various classifiers that are based on a minimization of a regularized risk are universally consistent, i.e. they can asymptotically learn in every classification task. In particular we focus on the role of the loss functions used in these algorithms. As an application of our general framework, several types of support vector machines as well as regularization networks are treated. Our methods combine techniques from stochastics, approximation theory and functional analysis.

MSC Classification: 68T10
CR Classification: I.5.1., G.3
Key Words: Computational learning theory, Pattern recognition, Support vector machines, Kernel methods, Regularization, Universal consistency

*Research was partially supported by the DFG grant *Ca 179/4-1*.

I Introduction

We treat the statistical classification problem which have been studied in both statistics and machine learning (cf. [1] for a throughout treatment). Let (X,τ) be a nonempty, compact topological (Hausdorff) space, e.g. a closed, bounded subset of $\mathbb{R}^d$, and $Y := \{-1,1\}$. A *classifier* $\mathcal{C}$ is a rule that assigns to every *training set* $T \in (X\times Y)^n$ a measurable function $f_T : X \to \mathbb{R}$. Here, it is always assumed that T is i.i.d. according to an unknown Borel probability measure 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) generated with distribution 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

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

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

The type of classifiers for which we shall 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 these 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 III.1 and III.2 for the existence). Various recently proposed algorithms including *regularization networks* and several variants of *support vector machines (SVM's)* belong to this type of classifiers (cf. the examples below).

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 Thm. 1] there exists a solution f_{T,λ_n} of (2) which has the form $f_{T,\lambda_n} = \sum_{i=1}^n \alpha_i^* k(x_i,.)$. Hence it suffices to consider (2) over the subspace $\{\sum_{i=1}^n \alpha_i k(x_i,.) : \alpha_1,\dots,\alpha_n \in \mathbb{R}\}$. Furthermore, we have $\|f\|_H^2 = \sum_{i,j=1}^n \alpha_i\alpha_j k(x_i,x_j)$ for $f = \sum_{i=1}^n \alpha_i k(x_i,.)$. Thus, once one have found a solution $\alpha_1^*,\dots,\alpha_n^* \in \mathbb{R}$ of

$$\arg\min_{\alpha_i\in\mathbb{R}} \Omega\left(\lambda, \sqrt{\sum_{i,j=1}^{n} \alpha_i\alpha_j k(x_i,x_j)}\right) + \frac{1}{n}\sum_{i=1}^{n} L\Big(y_i, \sum_{j=1}^{n}\alpha_j k(x_i,x_j)\Big)$$

a solution of (2) can be computed by $f_{T,\lambda_n} = \sum_{i=1}^n \alpha_i^* k(x_i,.)$. A similarly argument can be employed for (3). For $\Omega(\lambda,t) = \lambda t^2$ and specific convex loss functions such as the hinge loss $L(y,t) := \max\{0, 1-yt\}$ the dual of this problem is solved in practice (cf. [3] and [4]).

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. The first group of examples we consider consists of almost classical types of support vector machines. These classifiers have been introduced by Vapnik and his co-authors between 1992 and 1995. A good introduction to these SVM's is provided in [3]. We begin with the most common SVM:

Example I.1 Let $\Omega(\lambda, t) := \lambda t^2$, L be the hinge loss function $L(y,t) := \max\{0, 1-yt\}$ and (λ_n) be a positive sequence with $\lambda_n \to 0$. Then the classifiers based on either (2) or (3) are called L1-SVM's (cf. [5]). The L1-SVM based on (2) is universally consistent whenever a universal kernel k (cf. the examples below 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 the slightly stronger condition $n\lambda_n^2/\log n \to \infty$. Note, that this condition actually ensures strong universal consistency for L1-SVM's 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 RBF kernel $k(x,x') := \exp(-\sigma\|x-x'\|_2^2)$ or V. Vovk's infinite polynomial kernel $k(x,x') := (1-\langle x, x'\rangle)^{-\alpha}$ (cf. [6] and Example III.8)—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 can be further weakened to $n\lambda_n|\log\lambda_n|^{-d-1} \to \infty$ if one is only interested in universal consistency.

In practice, (3) is treated by solving the dual problem (cf. [3]). Then it turns out that the offset b gives rise to an equality constraint in the dual problem which causes some algorithmic trouble. Instead of omitting the offset as in (2) it is also possible to consider the optimization problem of the form

$$\min_{\substack{f\in H\\ b\in\mathbb{R}}} \; \Omega\big(\lambda_n, \|f\|_H + b\big) + \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i) + b\big) \; . \tag{4}$$

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

Note, that all of these conditions 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 $\lambda \to 0$ (cf. Section III for a definition of these quantities).

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, in both cases they almost coincide with the results in [7] if the used kernel is smooth.

Recall, that there also exists another well-known variant, the so-called ν-SVM, which is 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 (cf. [9]). Our next example treats the so-called L2-SVM:

Example I.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-SVM's (cf. [5]). 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 L1-SVM is based on (3) we can only establish the slightly stronger condition $n\lambda_n^4/\log n \to \infty$. Note, that this condition actually ensures strong universal consistency for L2-SVM's 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) and $n\lambda_n^{2+\varepsilon} \to \infty$ for some arbitrary small $\varepsilon > 0$ in order to ensure strong universal consistency. For a Gaussian RBF kernel on $X \subset \mathbb{R}^d$ the latter 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 (4) with $\Omega(\lambda, t) := \lambda t^2$ and $L(y,t) := \big(\max\{0, 1-yt\}\big)^2$. The corresponding consistency conditions on (λ_n) are obvious. As indicated in [4, Sect. 10.6.2] it is open whether this modification has a significant influence on the generalization performance. However, 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.

These conditions 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 SVM's:

Example I.3 *Least square support vector machines* (LS-SVM's) proposed in [11] are based on the minimization problem (3) with $\Omega(\lambda, t) := \lambda t^2$ 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 $|L_\lambda|_1 \sim 1/\sqrt{\lambda}$ and $\|L_\lambda\|_\infty \sim 1/\lambda$ for $\lambda \to 0$ the conditions for both classifiers coincide with those of the corresponding L2-SVM's.

Since SVM's 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 I.4 *Support vector machines with 1-regression loss function* (R1-SVM's) are based on the minimization problem (3) with $\Omega(\lambda, t) := \lambda t^2$ and $L_\varepsilon(y,t) := \max\{0, |y-t|-\varepsilon\}$. Let $0 \le \varepsilon < 1$. Since $|L_\lambda|_1 = 1$ and $\|L_\lambda\|_\infty \sim 1/\sqrt{\lambda}$ for $\lambda \to 0$ the conditions ensuring universal consistency coincide with those of the L1-SVM's with offset. It is clear, that one can also consider this type of classifier without offset or with optimized offset.

If one replaces the loss function of R1-SVM's by $L_\varepsilon(y,t) := (\max\{0, |y-t|-\varepsilon\})^2$ for some $0 \le \varepsilon < 1$ one obtains *support vector machines with 2-regression loss function* (R2-SVM's). Since $|L_\lambda|_1 \sim 1/\sqrt{\lambda}$ and $\|L_\lambda\|_\infty \sim 1/\lambda$ for $\lambda \to 0$ the conditions ensuring universal consistency coincide with those of the L2-SVM's with offset. Again, one can also consider this type of classifier without offset or with optimized offset.

The following example provides a support vector machine that is not universally consistent:

Example I.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 if and only if $c_{-1} = c_1$, i.e. if and only if L is *symmetric*. As indicated above, the classifier corresponding to L is thus not universally consistent if $c_{-1} \neq c_1$.

However, there are various other asymmetric loss functions that are admissible. For example, L defined by $L(1,t) := 2\max\{0, 1-t\}$ 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 considers the logistic and the AdaBoost loss function, respectively:

Example I.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 $|L_\lambda|_1 \sim 1$ 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 I.7 The loss function $L(y,t) := \exp(-yt)$ known from the AdaBoost algorithm is another example of a convex and admissible loss function. Obviously for $\Omega(\lambda, t) := \lambda t^2$ we have $|L_\lambda|_1 \sim \exp(\sqrt{2/\lambda}K)$ 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 III.19 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 are also possible and some of them have already been used. Common examples include (cf. 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 two examples are mainly of theoretical interest. Historically, they were considered in order to "explain" the generalization performance of support vector machines.

Example I.8 In order to motivate support vector machines 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 III.4 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 support vector machines which do not depend on Theorem III.19 can be directly applied to this modification.

Example I.9 In [16, Sect. 10.2] support vector machines 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 $L_0(y,t) := \mathbf{1}_{\{y \neq \operatorname{sign} t\}}$ 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 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 (cf. also [8]). Hence there are roughly speaking two alternatives for investigating the generalization performance of classsifers: 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 SVM's 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 SVM's 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-SVM's 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 shows that L1-SVM's with offset are universally consistent if the used kernel is universal and the regularization parameter λ_n tends "very slowly" to 0. In [7] the latter were 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:

- Using the loss function of interest in (2) or (3), i.e. $L_0(y,t) := \mathbf{1}_{\{y \neq \mathrm{sign}\, t\}}$, usually 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 $x \mapsto \mathrm{sign}(2P(1|x) - 1)$. Surprisingly, we can show in this work that for continuous loss functions this is even a sufficient condition.

- Common techniques such at 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. Similar arguments holds for every other reasonable loss function.

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

$$\mathcal{R}_{L,P}(f) \;:=\; \mathbb{E}_{(x,y)\sim P} L(y, f(x)) \tag{5}$$

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

$$\mathcal{R}^{reg}_{L,P,\lambda}(f) \;:=\; \Omega(\lambda, \|f\|_H) + \mathcal{R}_{L,P}(f) \tag{6}$$

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 $\mathcal{R}_{L,T}(.)$, $\mathcal{R}_{L,T}$ and $\mathcal{R}^{reg}_{L,T,\lambda}(.)$, respectively. Note, that $\mathcal{R}^{reg}_{L,T,\lambda}(.)$ 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 (cf. Lemma III.1). In the next step (cf. Proposition III.3) we show

$$\lim_{\lambda \to 0} \mathcal{R}^{reg}_{L,P,\lambda}(f_{P,\lambda}) \;=\; \mathcal{R}_{L,P} \tag{7}$$

if the used RKHS is rich enough, i.e. is universal. We then prove that approximating the minimal L-risk is sufficient to approximately achieve the Bayes risk. More precisely, we show (cf. Proposition III.4) that for all sequences of functions $f_n : X \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \mathcal{R}_{L,P}(f_n) = \mathcal{R}_{L,P} \qquad \Longrightarrow \qquad \lim_{n\to\infty} \mathcal{R}_P(f_n) = \mathcal{R}_P\,. \tag{8}$$

In the final step we correlate the risk of $f_{T,\lambda}$ with the empirical risk of $f_{T,\lambda}$ by certain concentration inequalities. If e.g. L is the hinge loss, $\Omega(\lambda,t) := \lambda t^2$ and $\lambda_n \le 1$ we have (cf. Lemma III.20 and III.21):

$$P^n\Big(T \in (X\times Y)^n : |\mathcal{R}_{L,T}(f_{T,\lambda_n}) - \mathcal{R}_{L,P}(f_{T,\lambda_n})| > \varepsilon + \frac{2K^2}{n\lambda_n}\Big) \;\le\; 2e^{-\frac{\varepsilon^2 n\lambda_n^2}{2(\sqrt{2}K+1)^4}}\,, \tag{9}$$

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 (7), (8) and (9) together (cf. the proof of Theorem III.6). Note, that instead of using (9) we can employ several other concentration inequalities. Unlike (9) that is based on a stability argument due to [20] most of them depend on covering numbers of certain operators related to H. 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 above 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 (7) and (8) to obtain new consistency results. Furthermore, if one is interested in problems of the form (2) or (3) for function spaces H different from RKHS's one only has to establish (7) and a concentration inequality in the spirirt of (9). Finally note that (8) is algorithm independent and thus can be used in many other settings, too.

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. We also recall the definition of covering numbers. 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 $g(\varepsilon) \le c\,h(\varepsilon)$ for all sufficiently small $\varepsilon > 0$. We write $g \sim h$ if both $g \preceq h$ and $h \preceq g$. Furthermore, we always assume $0\cdot\infty := 0$.

For a positive definite 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 $\Phi : X \to H$, $x \mapsto k(x,.)$ fulfills $k(.,.) = \langle \Phi(.), \Phi(.)\rangle_H$ by the reproducing property. We will often use the quantity $K := \sup\{\sqrt{k(x,x)} : x \in X\}$. Note, 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 $Iw := \langle w, \Phi(.)\rangle_H$, $w \in H$. 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, Sect. 3]).

Although we are mainly interested in 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 I.8 and I.9). Therefore, let us fix the properties of Ω which we shall need in the following:

Definition II.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 $x \mapsto P(\,.\,|x)$ from X into the set of all probability measures on Y such that P is the joint distribution of $(P(\,.\,|x))_x$ and of the marginal distribution P_X of P on X (cf. [26, Lem. 1.2.1.]).

In order to treat the 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 (5) of a function $f : X \to \overline{\mathbb{R}}$ by

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

Roughly speaking, it turns out that the solutions of (2) or (3) tend to a function f^* that minimizes this L-risk. 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 II.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

$$C(\alpha, t_\alpha) \;=\; \min_{t \in \overline{\mathbb{R}}} C(\alpha, t) \;=:\; M(\alpha) \qquad (10)$$

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 [27]. Besides the asymmetric hinge loss function discussed in Example I.5 all loss functions treated in the examples of the introduction are admissible. We will see in Lemma IV.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 above 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 $P(y = 1) = \alpha$.

An admissible loss function L is called *convex* if $L(y, .)$ is convex for $y = \pm 1$. It is said to be *1-Hölder-continuous* if

$$|L|_1 \;:=\; \sup\left\{ \frac{|L(y,t) - L(y,t')|}{|t - t'|} : y \in Y,\ t, t' \in \mathbb{R},\ t \neq t' \right\} \;<\; \infty\ .$$

Analogously, L is *locally* 1-Hölder-continuous if $L_{|Y\times[-a,a]}$ is 1-Hölder-continuous for all $a > 0$. Recall, that continuous convex functions on $\mathbb{R}$ are locally 1-Hölder-continuous and hence convex loss functions are locally 1-Hölder-continuous. Although most of our loss functions are convex we like to develop a general theory. Thus, we have to refine the above notion: the *inverted modulus of continuity* of L is defined by

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

where

$$\omega(L, \delta) \;:=\; \sup\big\{|L(y,t) - L(y,t')| : y \in Y \text{ and } t, t' \in \mathbb{R} \text{ with } |t - t'| \le \delta\big\}$$

denotes the modulus of continuity of L. Analogously, the (inverted) modulus of continuity of $L_{|Y\times[-a,a]}$ is defined. Note, that for locally 1-Hölder-continuous L we have $\omega(L_\lambda, \delta) \le \delta |L_\lambda|_1$ for all $\delta > 0$ and hence $\varepsilon/|L_\lambda|_1 \le \omega^{-1}(L_\lambda, \varepsilon)$ for all $\varepsilon > 0$. Furthermore, in our specific situation the modulus of continuity is continuous in δ (cf. [28, Prop. 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 $Y \times [-a, a]$, $a > 0$.

All of the loss functions considered in the Examples I.1-I.6 satisfy the conditions of the following definition which are needed to treat classifiers that are based on (3):

Definition II.3 An admissible loss function L is called *regular* if L is locally 1-Hölder-continuous, $L(1,.)$ is monotone decreasing and unbounded on $(-\infty,0]$, $L(-1,.)$ 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

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

Finally, recall that for a metric space (M,d) the *covering numbers* of M are defined by

$$\mathcal{N}\big((M,d),\varepsilon\big) := \min\Big\{n\in\mathbb{N}\cup\{\infty\} : \exists x_1,\dots,x_n \text{ with } M\subset\bigcup_{i=1}^n B_d(x_i,\varepsilon)\Big\} .$$

Here $B_d(x,\varepsilon)$ denotes the closed ball with centre x and radius $\varepsilon>0$. For a bounded linear operator T between Banach spaces E and F we define $\mathcal{N}(T,\varepsilon):=\mathcal{N}(TB_E,\varepsilon)$, where B_E denotes the closed unit ball of E. Usually, it is more convenient to consider $\mathcal{H}\big((M,d),\varepsilon\big):=\ln\mathcal{N}\big((M,d),\varepsilon\big)$ and $\mathcal{H}(T,\varepsilon):=\ln\mathcal{N}(T,\varepsilon)$, instead. Recall, that there also exists a concept—the so-called entropy numbers—which is inverse to the above notions. For details we refer to [28].

Furthermore, we also need a similar notion. Let ℓ_∞^n denotes 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

$$\mathcal{N}\big(\mathcal{F},n,\varepsilon\big) := \sup\Big\{\mathcal{N}\big((\mathcal{F}_{|X_0},\ell_\infty^{|X_0|}),\varepsilon\big) : X_0\subset X,\ |X_0|\leq n\Big\}$$

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 we define $\mathcal{H}\big(\mathcal{F},n,\varepsilon\big)$and $\mathcal{H}(T,n,\varepsilon)$, where T is an operator mapping into a space of functions.

III Results

In this section we develop the general theory which leads to the examples discussed in the introduction. This section is organized as follows: In Subsection III.A we mainly formalize the results (7) and (8) discussed in the roadmap. In Subsection 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 consitency. Finally, for convex loss functions and classifiers without offset the latter is improved in Subsection III.D.

III.A Results of the roadmap

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

$$\begin{aligned} \delta_\lambda &:= \sup\big\{t:\Omega(\lambda,t)\leq L(1,0)+L(-1,0)\big\} \\ L_\lambda &:= L_{|Y\times[-\delta_\lambda K,\delta_\lambda K]} \end{aligned}$$

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) and (2) as we shall see in the following lemmas:

Lemma III.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*

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

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

The proof of this lemma can be found in Section IV. Note, that in particular Lemma III.1 ensures that there always exists a solution of (2). The next lemma states the result of Lemma III.1 for modified regularized risks which corresponds to (3). Again, its proof can be found in Section IV.

Lemma III.2 *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*

$$\Omega\big(\lambda, \|\tilde{f}_{P,\lambda}\|\big) + \mathcal{R}_{L,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) \;=\; \inf_{\substack{f \in H \\ b \in \mathbb{R}}} \Omega\big(\lambda, \|f\|\big) + \mathcal{R}_{L,P}(f + b)\ . \tag{11}$$

Moreover, for all such pairs we have $\|\tilde{f}_{P,\lambda}\| \leq \delta_\lambda$.

The following proposition formalizes (7). A proof can be found in Section IV.

Proposition III.3 *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*

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

The next result shows that it suffices to approximate the minimal L-risk in order to approximate the Bayes risk, i.e. it formalizes (8). Again, the proof is worked out in Section IV.

Proposition III.4 *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 $\mathcal{R}_{L,P}(f) \leq \mathcal{R}_{L,P} + \delta$ we have $\mathcal{R}_P(f) \leq \mathcal{R}_P + \varepsilon$.*

III.B Consistency results based on covering numbers

So far, we have established (7) and (8) of the roadmap. As described in the introduction we finally need some concentration inequalities in the spirit of (9). In this subsection we use concentration inequalities depending on covering numbers. We begin with a result for classifiers based on (2). Its proof which is similar to the methods of [29] for the square loss function in regression scenarios is in Section IV.

Lemma III.5 *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 $n \geq 1$ we have*

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

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

Now, we can establish our first consistency result:

Theorem III.6 *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 the optimization problem (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.

Proof: Let us define $c := \sup\big\{|L(y,t) - L(y,t')| : y \in Y,\ t,t' \in [-\hat{\delta}K, \hat{\delta}K]\big\}$, where $\hat{\delta}$ is defined as above. For later purpose we note that

$$\sup\big\{\omega^{-1}(L_\lambda, \varepsilon) : \lambda \in (0,1],\ \varepsilon \in (0,c)\big\} \ < \ \infty\ . \tag{12}$$

Now, let $\varepsilon \in (0,c)$. We fix $\delta > 0$ according to Proposition III.4. Then by Proposition III.3 there exists an integer $n_0 \geq 1$ such that for all $n \geq n_0$ we have

$$\big|\mathcal{R}^{reg}_{L,P,\lambda_n}(f_{P,\lambda_n}) - \mathcal{R}_{L,P}\big| \leq \delta/3\ . \tag{13}$$

Moreover, (12) 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 III.5 and Hoeffding's inequality applied to $\mathcal{R}_{L,T}(f_{P,\lambda_n})$ we may additionally assume

$$\Pr{}^*\Big(T \in (X \times Y)^n : |\mathcal{R}_{L,T}(f_{T,\lambda_n}) - \mathcal{R}_{L,P}(f_{T,\lambda_n})| + |\mathcal{R}_{L,T}(f_{P,\lambda_n}) - \mathcal{R}_{L,P}(f_{P,\lambda_n})| \leq \delta/3\Big) \ \geq \ 1 - \varepsilon \tag{14}$$

for all $n \geq n_0$. If T belongs to set of samples considered in inequality (14) 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 (13). Now, the definition of δ gives the first assertion. The second assertion follows by the specific form of the tail bound in Lemma III.5 and Hoeffding's inequality. ■

For various "smooth" kernels (cf. the appendix for the different notions of smoothness) there exist bounds on the covering numbers of the embedding $I : H \to C(X)$ (cf. Lemma IV.2). Applying these bounds to Theorem III.6 we obtain the following corollary whose proof can be found in Section IV:

Corollary III.7 *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*

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

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 [38]. Here, we only consider a specific class of universal kernels:

Example III.8 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 V. 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 III.9 As already mentioned the Gaussian RBF kernel on $X \subset \mathbb{R}^d$ also fits into the framework of Example III.8. For this kernel, however, a sharper upper bound for the covering numbers of I was recently shown in [30]. Namely, there was proved

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

In particular, the classifier considered in Corollary III.7 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 III.5 by a suitable concentration inequality. The main difficulty here is to derive bounds for the offset (cf. Lemma IV.3) in order to establish such concentration inequalities. As a consequence the assumptions on L have to be slightly sharper:

Lemma III.10 *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*

$$\Pr{}^*\big(T \in (X\times Y)^n : |\mathcal{R}_{L,T}(\tilde f_{T,\lambda} + \tilde b_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde f_{T,\lambda} + \tilde b_{T,\lambda})| \geq \varepsilon\big) \leq 2e^{2\mathcal{H}\left(I, \frac{c\varepsilon}{|L_\lambda|_1 \delta_\lambda}\right) - \frac{c\varepsilon^2 n}{\|L_\lambda\|_\infty^2}} .$$

The proof of this lemma can be found in Section IV. Proceeding as in the proof of Theorem III.6, the result for classifiers based on (3) now reads as follows:

Theorem III.11 *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 the optimization problem (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 IV.2 we can immediately derive the results of Corollary III.7 for classifiers with offset and regular loss functions.

III.C Consistency results based on lacalized covering numbers

Instead of using a concentration inequality that is based on the covering numbers of $I : H \to C(X)$ we can apply the following result which is based on localized covering numbers:

Lemma III.12 *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 $n \geq 1$ we have*

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

The proof of this lemma which is mainly due to [31] 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 III.13 *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*

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

for all $\varepsilon > 0$ Then the classifier based on the optimization problem (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 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 III.14 *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*

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

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 III.15 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 [32] that

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

In particular, the classifier considered in Corollary III.14 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 concentration inequality is

Lemma III.16 *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*

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

The proof of this lemma is in Section IV. The consistency result corresponding to Lemma III.16 is:

Theorem III.17 *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 the optimization problem (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 III.14 and Example III.15 for classifiers with offset. Here, we only state:

Corollary III.18 *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.

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 III.14.

Theorem III.19 *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. [33]): 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 i-th 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

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

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 [33] shows:

Lemma III.20 *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*

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

Hence, for the proof of Theorem III.19 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 III.21 *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 III.19: 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 III.5 by another suitable result. To make this precise, by Lemma III.20, Lemma III.21 and the proof of Theorem III.6 we only have to ensure

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

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) in the spirit of Theorem III.19. The reason for this lack is that Theorem III.19 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. Ex. I.1) on $X = \{0\}$.

IV Proofs

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

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

Since $\Omega(\lambda, \|f_\varepsilon\|) \leq \mathcal{R}^{reg}_{L,P,\lambda}(f_\varepsilon) \leq \mathcal{R}^{reg}_{L,P,\lambda}(0) + \varepsilon \leq 2\big(L(1,0)+L(-1,0)\big)$ there exists a $\delta > 0$ with $\|f_\varepsilon\| \leq \delta$. Now by Alaoglu's 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 $\|f_{\varepsilon_n}\|_\infty \leq \delta K$ and hence $\big|L(.,f_{\varepsilon_n}(.))\big|$ is bounded by the continuity of L. Therefore Lebesgue's theorem implies

$$\mathcal{R}_{L,P}(f_{\varepsilon_n}) \to \mathcal{R}_{L,P}(f_{P,\lambda}) \ . \tag{15}$$

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

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

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

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

Moreover, we always have $\|f_{P,\lambda}\| \leq \liminf_{n\to\infty} \|f_{\varepsilon_n}\| = c$ (cf. [34, Ch. 1 Cor. 2.6.]) and thus $\Omega(\lambda, \|f_{P,\lambda}\|) \leq \Omega(\lambda, c) = \lim_{n\to\infty} \Omega(\lambda, \|f_{\varepsilon_n}\|)$. This together with (15) yields $\mathcal{R}^{reg}_{L,P,\lambda}(f_{\varepsilon_n}) \to \mathcal{R}^{reg}_{L,P,\lambda}(f_{P,\lambda})$. Since the construction of f_ε implies $\mathcal{R}^{reg}_{L,P,\lambda}(f_{\varepsilon_n}) \to \inf_{f\in H} \mathcal{R}^{reg}_{L,P,\lambda}(f)$ the first assertion follows. The second assertion is trivial. ■

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

$$P_X\big(x \in X : P(y|x) = 1\big) = 1 \ . \tag{16}$$

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

Proof of Lemma III.2: Analogously to the proof of Lemma III.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. (16) does not hold. Then for $A_y := \{x \in X : P(y|x) > 0\}$ we find $P_X(A_y) > 0$ for all $y \in Y$. In particular, there exists a $\rho > 0$ such that for $A^\rho_y := \{x \in X : P(y|x) > \rho\}$ we have $P_X(A^\rho_y) > 0$ for all $y \in Y$. Moreover, for $y = 1$ we find

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

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

$$L(1, f_\varepsilon(x_\varepsilon) + b_\varepsilon) \ \leq \ \frac{2\big(L(1,0)+L(-1,0)\big)}{\rho P_X(A^\rho_1)} \ . \tag{17}$$

Since L is regular and the right hand side of (17) 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 III.1.
Now let us assume that P is y-degenerated for some $y \in Y$. Then one easily checks that $(\tilde{f}_{P,\lambda}, \tilde{b}_{P,\lambda}) := \big(0, f^*(P(1|x))\big) \in H \times \overline{\mathbb{R}}$ is a pair fulfilling (11) where f^* denotes a function according to Lemma IV.1. Finally, the bound on $\|\tilde{f}_{P,\lambda}\|$ for general solutions of (11) is trivial. ■

For the proof of Proposition III.3 we need the following technical lemma:

Lemma IV.1 *There exists 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*

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

Proof: Let (α_i) be a dense sequence in $[0,1]$ with $\alpha_1 = 0$ and $\alpha_i \neq \alpha_j$ if $i \neq j$. We fix a solution $t_i \in \overline{\mathbb{R}}$ of (10) for every α_i, $i \geq 1$ 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

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

i.e. $f(\alpha)$ is a solution of (10) for α. Since M is concave it is also continuous for all but at most countably many points of $[0,1]$ (cf. [35, Thm. 1.16]) and thus we only have to modify f on countably many points in order to construct f^*. ■

Proof of Proposition III.3: Let $\varepsilon > 0$ and $f_\varepsilon \in H$ with $\mathcal{R}_{L,P}(f_\varepsilon) \leq \inf\{\mathcal{R}_{L,P}(f) : f \in H\} + \varepsilon$. Since $\Omega(., \|f_\varepsilon\|)$ is continuous in 0 there exists λ_0 such that $\Omega(\lambda, \|f_\varepsilon\|) \leq \varepsilon$ for all $\lambda \leq \lambda_0$. This together with the definition of f_ε yields

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

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

$$\inf_{f\in H} \mathcal{R}_{L,P}(f) \;=\; \inf\big\{\, \mathcal{R}_{L,P}(f) \;:\; f : X \to \mathbb{R} \text{ bounded and measurable}\big\} \;=\; \mathcal{R}_{L,P} \;. \tag{18}$$

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

$$P_X\big(\{x \in X : |h(x) - f(x)| \geq \varepsilon\}\big) \;\leq\; \delta$$

and $\|f\|_\infty \leq \|h\|_\infty$. 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 $|f^*(\alpha)| \leq n$, $f_n(\alpha) := n \operatorname{sign} f^*(\alpha)$ if $|f^*(\alpha)| = \infty$ and $f_n(\alpha) = 0$ otherwise. Moreover, we define

$$M_n(\alpha) \;:=\; \alpha L\big(1, f_n(\alpha)\big) + (1-\alpha)L\big(-1, f_n(\alpha)\big) \;=\; C\big(\alpha, f_n(\alpha)\big) \;.$$

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

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

Since we always have $M(\alpha) \leq \alpha L(1,0) + (1-\alpha)L(-1,0)$ 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

$$|M_n(\alpha) - M(\alpha)| \;\leq\; \alpha|L(1,n) - L(1,\infty)| + (1-\alpha)|L(-1,n) - L(-1,\infty)| \tag{19}$$

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

$$|M_n(\alpha) - M(\alpha)| \;\leq\; \alpha|L(1,-n) - L(1,-\infty)| + (1-\alpha)|L(-1,-n) - L(-1,-\infty)| \tag{20}$$

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

$$\alpha L(1,\infty) + (1-\alpha)L(-1,\infty) \;=\; \inf_{t\in\mathbb{R}} \alpha L(1,t) + (1-\alpha)L(-1,t) \;\leq\; \alpha L(1,0) + (1-\alpha)L(-1,0) \;<\; \infty$$

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, too. 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 find

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

■

Proof of Proposition III.4: For $f : X \to \overline{\mathbb{R}}$ the set of points misclassified by f is defined

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

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

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

In order to estimate the second term from below let $f_* : [0,1] \to \overline{\mathbb{R}}$ be a function with $C(\alpha, f_*(\alpha)) = \inf\big\{C(\alpha,t) : t \geq 0\big\}$ if $\alpha < 1/2$ and $C(\alpha, f_*(\alpha)) = \inf\big\{C(\alpha,t) : t \leq 0\big\}$ if $\alpha > 1/2$. Using the technique of the proof of Lemma IV.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

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

is a strictly positive function on $\hat{X} := \{x \in X : P(1|x) \neq 1/2\}$. 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. $s(x) = \min\{P(1|x), P(-1|x)\}$. 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 III.5: Let $\mathcal{F} := \{L(.,f(.)) : f \in \delta_\lambda I B_H\}$. By the definition of the modulus of continuity every ε-net $f_1,\dots,f_n$ of $\delta_\lambda I B_H$ defines an $\omega(L_\lambda,\varepsilon)$-net $L(.,f_1(.)),\dots,L(.,f_n(.))$ 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 non-negative functions that are bounded by $\|L_\lambda\|_\infty$. Thus, applying Hoeffding's inequality (cf. [1, Thm. 8.1]) yields

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

Since Lemma III.1 guarantees $f_{T,\lambda}\in\delta_\lambda B_H$ the assertion now follows. ■

In order to prove Corollary III.7 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 IV.2 *Given a continuous kernel k on the closure of a bounded C^∞-domain $X\subset\mathbb{R}^d$ we have*

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

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 [36] (cf. also [28, Ch. 5] and [37]).
If $k\in C^{r,r}(\overline{X}\times\overline{X})$ the embedding I actually maps into $C^r(\overline{X})$ (cf. [38, 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 (cf. [39, Sect. 3.2.4 and 3.3.1]). Thus, we find

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

by [40, Sect. 3.3.2], i.e. we have shown the second assertion.
In the third case we have to consider the embedding $I : W_p^s(\overline{X})\to C(\overline{X})$. Corresponding estimates of the covering numbers of this embedding can be found in [40, Ch. 3.3].
Finally, the last assertion is a direct consequence of the second assertion. ■

Proof of Corollary III.7: By Lemma IV.2 we obtain that in every case there exists a $\gamma>0$ with

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

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

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

Hence, in order to satisfy the first condition of Theorem III.6 it suffices to ensure $\frac{\|L_{\lambda_n}\|_\infty^2}{n}\cdot\big(\delta_{\lambda_n}|L_{\lambda_n}|_1\big)^\gamma\to 0$. This is done by the conditions of the corollary together with Lemma IV.2. Moreover, the above convergence yields $\big(\delta_{\lambda_n}|L_{\lambda_n}|_1\big)^\gamma\preceq\frac{n}{\|L_{\lambda_n}\|_\infty^2}$ and since we always have $\|L_{\lambda_n}\|_\infty\preceq\delta_{\lambda_n}|L_{\lambda_n}|_1$ we find $n^{1-2/(2+\gamma)}\preceq\frac{n}{\|L_{\lambda_n}\|_\infty^2}$. This yields the second condition of Theorem III.6. ■

In order to prove Lemma III.10 we need the following lemma that shows that the bound for $\tilde{b}_{T,\lambda}$ which was obtained in the proof of Lemma III.2 is rather sloppy in typical situations:

Lemma IV.3 *Let L be a regular and admissible loss function and P be a non-degenerated probability measure on $X \times Y$. Then there exists a constant $c > 0$ such that for all $n \geq 1$ we have*

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

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

Proof of Lemma IV.3: Since P is non-degenerated there exists a $\rho > 0$ such that for $A_y^\rho := \{x \in X : P(y|x) > \rho\}$, $y \in Y$ we have $P_X(A_y^\rho) > 0$. With $c_1 := \frac{\rho}{2}\min\{P_X(A_1^\rho), P_X(A_{-1}^\rho)\}$ we obtain by Hoeffding's inequality that

$$P^n\big(T \in (X \times Y)^n : |\{i : x_i \in A_y^\rho, y_i = y\}| \geq c_1 n\big) \;\geq\; 1 - e^{-2c_1^2 n}$$

holds for $y = \pm 1$. Now let $\lambda > 0$ and $T \in (X \times Y)^n$ be a training set with $|\{i : x_i \in A_1^\rho, y_i = 1\}| \geq c_1 n$. Then we have

$$\frac{1}{n}\sum_{\substack{x_i \in A_1^\rho \\ y_i = 1}} L\big(1, \tilde{f}_{T,\lambda}(x_i) + \tilde{b}_{T,\lambda}\big) \;\leq\; \Omega\big(\lambda, \|\tilde{f}_{T,\lambda}\|\big) + \mathcal{R}_{L,T}(\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda}) \;\leq\; L(1,0) + L(-1,0) \;.$$

Since this yields

$$\frac{1}{|\{i : x_i \in A_1^\rho, y_i = 1\}|}\sum_{\substack{x_i \in A_1^\rho \\ y_i = 1}} L\big(1, \tilde{f}_{T,\lambda}(x_i) + \tilde{b}_{T,\lambda}\big) \;\leq\; \frac{n\big(L(1,0) + L(-1,0)\big)}{|\{i : x_i \in A_1^\rho, y_i = 1\}|} \;\leq\; \frac{L(1,0) + L(-1,0)}{c_1}$$

there exists an index i with

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

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 $\|\tilde{f}_{T,\lambda}\| + \tilde{b}_{T,\lambda} \geq c_2$ holds for all $\lambda > 0$ and all $T \in (X \times Y)^n$ with $|\{i : x_i \in A_1^\rho, y_i = 1\}| \geq c_1 n$. This yields the desired estimate from below by Lemma III.2. 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 IV.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*

$$\begin{aligned} P^n(T \in (X \times Y)^n : T \text{ is } y\text{-degenerated}) &= 1 &\quad (21)\\ \mathcal{R}_{L,P}(\tilde{f}_{P,\lambda} + \tilde{b}_{P,\lambda}) &= L(y, f^*(y)) \;. &\quad (22) \end{aligned}$$

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

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

we can easily deduce the second assertion. ■

Proof of Lemma III.10: By Lemma IV.4 we may assume without loss of generality that P is not degenerated. Let $c_1 > 0$ be chosen accordingly to Lemma IV.3. We define $\mathcal{G} := \{f(.)+b : f \in \delta_\lambda IB_H,\ |b| \leq c_1 + \delta_\lambda K\}$ and $\mathcal{F} := \{L(., g(.)) : g \in \mathcal{G}\}$. 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 III.5 shows

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

where $\mathcal{F}$ and $\mathcal{G}$ are considered as subsets of $C(X)$. Hence with the help of Lemma IV.3 and Hoeffding's inequality we obtain

$$\begin{aligned}
&\Pr{}^*\big(T \in (X\times Y)^n : |\mathcal{R}_{L,T}(\tilde{f}_{T,\lambda}+\tilde{b}_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde{f}_{T,\lambda}+\tilde{b}_{T,\lambda})| \leq \varepsilon\big)\\
\geq\ & \Pr{}^*\Big(T \in (X\times Y)^n : |\tilde{b}_{T,\lambda}| \leq a \ \text{ and } \sup_{\substack{f\in\delta_\lambda B_H\\ |b|\leq a}} |\mathcal{R}_{L,T}(f+b) - \mathcal{R}_{L,P}(f+b)| \leq \varepsilon\Big)\\
\geq\ & 1 - 2\mathcal{N}\big(\mathcal{G},\omega^{-1}(\tilde{L}_\lambda,\varepsilon/3)\big)e^{-\frac{2\varepsilon^2 n}{9\|\tilde{L}_\lambda\|_\infty^2}} - 2e^{-c_1 n} .
\end{aligned}$$

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

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

for a suitable constant $c_3 > 0$ independent of λ, ε and n. Since $X \neq \emptyset$ and k is universal we get rank $I \geq 1$ and thus we have $\ln(1/\delta) \leq \mathcal{H}(I, c_4\delta)$ 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 III.12: Analogously to the proof of Lemma III.5 we define $\mathcal{F} := \big\{L(., f(.)) : f \in \delta_\lambda IB_H\big\}$. Then Lemma 3.4 in [41] yields

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

Now the assertion follows by the arguments used in the proof of Lemma III.5. ∎

Proof of Corollary III.14: By the dual Maurey-Carl inequality (cf. [42], [43] and [37]) we have

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

Therefore, we find

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

Since $\delta_{\lambda_n}|L_{\lambda_n}|_1 \to \infty$ the first condition of Theorem III.13 follows. The summability condition can be derived as in the proof of Corollary III.7. ∎

Proof of Lemma III.16: With the notions of the proof of Lemma III.10 we find

$$\begin{aligned}
&\Pr{}^*\big(T \in (X\times Y)^n : |\mathcal{R}_{L,T}(\tilde{f}_{T,\lambda}+\tilde{b}_{T,\lambda}) - \mathcal{R}_{L,P}(\tilde{f}_{T,\lambda}+\tilde{b}_{T,\lambda})| \geq \varepsilon\big)\\
\leq\ & 12n\,\mathcal{N}\big(\mathcal{G}, 2n, \omega^{-1}(\tilde{L}_\lambda,\varepsilon/3)\big)e^{-\frac{\varepsilon^2 n}{36\|\tilde{L}_\lambda\|_\infty^2}} + 2e^{-c_1 n}\\
\leq\ & 12n\,e^{\mathcal{H}\left(I,2n,\frac{c_3\varepsilon}{|L_\lambda|_1\delta_\lambda}\right)+\ln\left(\frac{|L_\lambda|_1\delta_\lambda}{c_3\varepsilon}\right)-\frac{c_3\varepsilon^2 n}{\|L_\lambda\|_\infty^2}} + 2e^{-c_1 n}
\end{aligned}$$

by Lemma 3.4 in [41] (cf. the proof of Lemma III.12). Again there is also a constant $c_4 > 0$ with $\ln(1/\delta) \leq \mathcal{H}(I, 2n, c_4\delta)$ for all $\delta > 0$, $n \geq 1$. Thus we finally find a constant $c > 0$ for which the assertion holds. ∎

Proof of Lemma III.21: For differentiable loss functions the assertion can be found in [4, Thm. 12.4]. In this book there are also some remarks concerning non-differentiable loss functions. For sake of completeness we present a proof which mainly follows the elegant approach of [44]. For this purpose we need to recall that for a convex, continuous function $f : \mathbb{R}^d \to \mathbb{R}$ the sub-differential of f in $x \in \mathbb{R}^d$ is defined by

$$\partial f(x) \ := \ \{x^* \in \mathbb{R}^d \ : \ \langle x^*, y - x\rangle \leq f(y) - f(x) \ \text{ for all } y \in \mathbb{R}^d\} \ .$$

For basic properties we refer to [34], [35], and in particular [45, Thm. 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 [45, Thm. 23.8 and 23.9] shows

$$\partial \mathcal{R}^{reg}_{L,T,\lambda_n}(f) \ = \ 2\lambda_n f + \big\{\mathbb{E}_T h\Phi \ : \ h(x_i, y_i) \in \partial L(y_i, f(x_i)) \text{ for all } i = 1, \ldots, n\big\} \ ,$$

where the sub-differential ∂L of L is with respect to the second variable, only. Recalling that f_{T,λ_n} minimizes $\mathcal{R}^{reg}_{L,T,\lambda_n}$ we also observe $0 \in \partial \mathcal{R}^{reg}_{L,T,\lambda_n}(f_{T,\lambda_n})$ 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 III.1 we actually have $\|h\|_\infty \leq |L_{\lambda_n}|_1$. Moreover, $h(x_i, y_i) \in \partial L(y_i, f_{T,\lambda_n}(x_i))$ implies

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

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

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

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

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

Moreover, f_{S,λ_n} minimizes $\mathcal{R}^{reg}_{L,S,\lambda_n}$ and hence we have $\mathcal{R}^{reg}_{L,S,\lambda_n}(f_{S,\lambda_n}) \leq \mathcal{R}^{reg}_{L,S,\lambda_n}(f_{T,\lambda_n})$. This yields

$$\lambda_n \|f_{T,\lambda_n} - f_{S,\lambda_n}\|^2 \ \leq \ \langle f_{T,\lambda_n} - f_{S,\lambda_n}, \mathbb{E}_S h\Phi + 2\lambda_n f_{T,\lambda_n}\rangle \ \leq \ \|f_{T,\lambda_n} - f_{S,\lambda_n}\| \ \|\mathbb{E}_S h\Phi + 2\lambda_n f_{T,\lambda_n}\| \, .$$

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

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

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

V 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 sub-covering. 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 not compact. However, for some spaces such as Hilbert spaces there is a non-trivial 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 semi-definite* if for all $x_1, \ldots, x_n \in X$ and all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ we have

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

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

$$H_0 \;:=\; \Big\{ \sum_{i=1}^{n} \alpha_i k(x_i, .) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in X \Big\}\,.$$

Now let us assume for simplicity, that k is a positive definite kernel. Then we can define by

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

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 $\Phi : X \to H$, $x \mapsto k(x, .)$ 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 III.7 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, Lem. 3]). Due to the continuity of k the topology generated by d_k coincides with τ (cf. [22, Lem. 3 and Cor. 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. [28, Ch. 5]). In particular, I is a compact operator and hence its covering numbers are finite.

If X itself has a metric d a kernel k is called *β-Hölder-continuous* (with respect to d), $0 < \beta \leq 1$ if there exists a constant $c > 0$ such that

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

for all $x, x' \in X$ (cf. [38, 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. [39, Sect. 3.2]) we also define

$$C^{r,r}(X \times X) \;:=\; \Big\{ f : X \times X \to \mathbb{R} : \frac{\partial^{|\alpha_1| + |\alpha_2|}}{\partial_{\alpha_1} \partial_{\alpha_2}} f \text{ continuous for all } \alpha_1, \alpha_2 \in \mathbb{N}_0^d \text{ with } |\alpha_1|, |\alpha_2| \leq r \Big\}$$

for $r \in \mathbb{N}$ (cf. [38, p.40]). Recall, that the open euklidian 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. [39]) 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", Springer, New York, 1997

[2] B. Schölkopf, R. Herbrich, and A.J. Smola, A generalized representer theorem, In "Proceedings of the 14th Annual Conference on Computational Learning Theory", *Lecture Notes in Artificial Intelligence* **2111** (2001), 416-426

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

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

[5] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning* **20** (1995), 273-297

[6] C. Saunders, M.O. Stitson, J. Weston, L. Bottou, B. Schölkopf and A. Smola, Support vector machine — reference manual, *Technical Report* **CSD-TR-98-03**, Department of Computer Science, Royal Holloway, University of London, available at `http://www.cs.rhul.ac.uk/~craig/psfiles/report98-03.ps.gz`, 1998

[7] T. Zhang, Statistical behaviour and consistency of classification methods based on convex risk minimization, *Ann. Statist*, to appear

[8] I. Steinwart, Support vector machines are universally consistent, *J. Complexity* **18** (2002), 768-791

[9] I. Steinwart, On the optimal parameter choice for ν-support vector machines, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, accepted

[10] O.L. Mangasarian and D.R. Musicant, Lagrangian support vector machines, *Journal of Machine Learning Research* **1** (2001), 161-177

[11] J.A.K. Suykens and J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters* **9** (1999), 293-300

[12] F. Girosi, M. Jones and T. Poggio, Regularization theory and neural networks architectures, *Neural Computation* **7** (1995), 219-269

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

[14] Y.J. Lee and O.L. Mangasarian, SSVM: a smooth support vector machine for classification, *Comput. Optim. Appl.* **20** (2001), 5-22

[15] T. Evgeniou, M. Pontil and T. Poggio, Regularization networks and support vector machines, in A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, "Advances in Large Margin Classifiers", 171-204, MIT Press, 2000

[16] V.N. Vapnik, "Statistical Learning Theory", Wiley, 1998

[17] D.S. Johnson and F.P. Preparata, The densest hemisphere problem, *Theoretical Computer Science* **6** (1978), 93-107

[18] K–U. Höffgen and Hans-U. Simon, Robust trainability of single neurons, In "Proceedings of the 5th Annual ACM Conference on Computational Learning Theory", 428-438, ACM Press, 1992

[19] I. Steinwart, Which data-dependent bounds are suitable for SVM's?, available at `http://www.c3.lanl.gov/~ingo/publications/bounds.ps`

[20] O. Bousquet and A. Elisseeff, Stability and generalization, *Journal of Machine Learning Research* **2** (2002), 499-526

[21] Y. Lin, Support vector machines and the Bayes rule in classification, *Data Mining and Knowledge Discovery* **6** (2002), 259-275

[22] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, *Journal of Machine Learning Research* **2** (2001), 67-93

[23] D.D. Cox and F. O'Sullivan, Asymptotic analysis of penalized likelihood and related estimators, *Ann. Statist.* **18** (1990), 1676-1695

[24] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337-404

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

[26] R.M. Dudley, Central limit theorems for empirical measures, *Ann. Probab.* **6** (1978), 899-929

[27] Y. Lin, A Note on Margin-based Loss Functions in Classification, TR 1044r, Department of Statistics, University of Wisconsin, available at `ftp://ftp.stat.wisc.edu/pub/yilin/tr1044r.ps`

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

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

[30] D.X. Zhou, The covering number in learning theory, *J. Complexity*, to appear

[31] R. Williamson, A.J. Smola and B. Schölkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, *IEEE Trans. Inform. Theory* **47** (2001), 2516-2532

[32] I. Steinwart, Entropy numbers of convex hulls and an application to learning algorithms, *Arch. Math.*, to appear, available at `http://www.c3.lanl.gov/~ingo/publications/archiv-02.ps`

[33] O. Bousquet and A. Elisseeff, Algorithmic stability and generalization performance, in T.K. Leen, T.G. Dietterich and V. Tresp, editors, "Advances in Neural Information Processing Systems 13", 196-202, MIT Press, 2001

[34] V. Barbu and Th. Precupanu, "Convexity and Optimization in Banach Spaces", D. Reidel Publishing Company, Dordrecht, 1986

[35] R.R. Phelps, Convex Functions, Monotone Operators and Differentiability, *Lecture Notes in Math.* **1364** (1993)

[36] B. Carl, S. Heinrich and T. Kühn, $s-$numbers of integral operators with Hölder-continuous kernels over metric compacta, *J. Funct. Anal.* **81** (1988), 54-73

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

[38] K. Ritter, Average-Case Analysis of Numerical Problems, *Lecture Notes in Math.* **1733** (2000)

[39] H. Triebel, "Theory of Function Spaces", Akademische Verlagsgesellschaft Geest & Portig, Leipzig, 1983

[40] D.E. Edmunds and H. Triebel, "Function Spaces, Entropy Numbers, Differential Operators", Cambridge University Press, 1996

[41] N. Alon, S. Ben-David, N. Cesa-Bianchi and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, *Journal of the ACM* **44** (1997), 615-631

[42] B. Carl and A. Pajor, Gelfand numbers of operators with values in a Hilbert space, *Invent. Math.* **94** (1988), 479-504

[43] B. Carl, I. Kyrezi and A. Pajor, Metric entropy of convex hulls in Banach spaces, *J. London Math. Soc.* **60** (1999), 871-896.

[44] Tong Zhang, Convergence of Large Margin Separable Linear Classification, in T.K. Leen, T.G. Dietterich and V. Tresp, editors, "Advances in Neural Information Processing Systems 13", 357-363, MIT Press, 2001.

[45] R.T. Rockafellar, "Convex Analysis", Princeton University Press, 1970