Journal of Machine Learning Research 2 (2001) 67-93 Submitted 08/01; Published 12/01

On the Influence of the Kernel on the Consistency of Support Vector Machines

Ingo Steinwart STEINWART@MINET.UNI-JENA.DE
Mathematisches Institut
Friedrich-Schiller-Universität
Ernst-Abbe-Platz 1-4
07743 Jena, Germany

Editor: Bernhard Schölkopf (for HTTP://WWW.KERNEL-MACHINES.ORG)

Abstract

In this article we study the generalization abilities of several classifiers of support vector machine (SVM) type using a certain class of kernels that we call universal. It is shown that the soft margin algorithms with universal kernels are consistent for a large class of classification problems including some kind of noisy tasks provided that the regularization parameter is chosen well. In particular we derive a simple sufficient condition for this parameter in the case of Gaussian RBF kernels. On the one hand our considerations are based on an investigation of an approximation property—the so-called universality—of the used kernels that ensures that all continuous functions can be approximated by certain kernel expressions. This approximation property also gives a new insight into the role of kernels in these and other algorithms. On the other hand the results are achieved by a precise study of the underlying optimization problems of the classifiers. Furthermore, we show consistency for the maximal margin classifier as well as for the soft margin SVM's in the presence of large margins. In this case it turns out that also constant regularization parameters ensure consistency for the soft margin SVM's. Finally we prove that even for simple, noise free classification problems SVM's with polynomial kernels can behave arbitrarily badly.

Keywords: Computational learning theory, pattern recognition, PAC model, support vector machines, kernel methods

1. Introduction

Support vector machines comprise a class of learning algorithms originally introduced for pattern recognition problems. Although their development was motivated by results of statistical learning theory the known bounds on their generalization performance are not fully satisfactory. In particular, the influence of the chosen kernel is far from being completely understood. The aim of this paper is to give a new insight into the role of the kernels. Our considerations are mainly based on a certain approximation property of various standard kernels that generate function classes with infinite VC-dimension. Since in this case classical Vapnik-Chervonenkis theory fails to be applicable for support vector machines, other concepts such as data dependent structural risk minimization, e.g. in terms of the observed margin, were introduced (cf. Shawe-Taylor et al., 1998; Bartlett & Shawe-Taylor, 1999; Cristianini & Shawe-Taylor, 2000, chap. 2). The latter usually needs large margins on the training sets to provide good bounds. It is, however, open which distributions and kernels guarantee this assumption. A systematical study of this question is the starting point of this paper. The resulting techniques allow us to show new bounds for the generalization performance of several standard support vector classifiers.

We begin with a description of the problem of pattern recognition (cf. Vapnik, 1998; Cristianini & Shawe-Taylor, 2000). Let (X,d) be a compact metric space[1], $Y := \{-1,1\}$ and P be a probability measure on $X \times Y$, where X is equipped with the Borel σ-algebra. By disintegration (cf. Dudley, 1989, Lem. 1.2.1.) 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. We call $P(\,.\,|.)$, which is in fact a regular conditional probability, the *supervisor*. A classifier is an algorithm that constructs to every *training set* $T = ((x_1,y_1),\dots,(x_n,y_n)) \in (X \times Y)^n$ a *decision function* $f_T : X \to Y$. In our context it is always assumed that T is i.i.d. according to P, which itself is unknown. Then the decision function $f_T : X \to Y$ constructed by the classifier should guarantee a small probability for the misclassification of an example (x,y) randomly generated according to P. Here, misclassification means $f(x) \neq y$. To make this precise, for a measurable function $f : X \to \{-1,1\}$ we define the risk of f by

$$\mathcal{R}_P(f) \;:=\; \int\limits_{X\times Y} \mathbf{1}_{\{f(x)\neq y\}}\, P(dx,dy) \;=\; P\big(\{(x,y) : f(x) \neq y\}\big)\;.$$

When considering noisy supervisors we cannot expect that we obtain zero risk. Indeed, for

$$\begin{aligned} B_1(P) &:= \big\{x \in X \;:\; P(y=1|x) > P(y=-1|x)\big\} \\ B_{-1}(P) &:= \big\{x \in X \;:\; P(y=1|x) < P(y=-1|x)\big\} \\ B_0(P) &:= \big\{x \in X \;:\; P(y=1|x) = P(y=-1|x)\big\} \end{aligned}$$

and a function $f_0 : X \to \{-1,1\}$ with $f_0(x) = 1$ if $x \in B_1(P)$ and $f_0(x) = -1$ if $x \in B_{-1}(P)$ we have (cf. Devroye et al., 1997, Thm. 2.1.)

$$\mathcal{R}_P(f_0) \;=\; \inf\big\{\mathcal{R}_P(f) \;:\; f\!:X \to \{-1,1\} \text{ measurable}\big\} \;=\; \int\limits_X p(x)\, P_X(dx)\;, \qquad (1)$$

where the *noise level* $p : X \to \mathbb{R}$ is defined by $p(x) := P(y=-1|x)$ for $x \in B_1(P)$, $p(x) := P(y=1|x)$ for $x \in B_{-1}(P)$ and $p(x) = 1/2$ otherwise. Equation (1) shows that no function can yield less risk than f_0. The function f_0 is called an optimal Bayes decision rule and we write $\mathcal{R}_P := \mathcal{R}_P(f_0)$. Now, a classifier $\mathcal{C}$ should guarantee with high probability that $\mathcal{R}_P(f_T)$ is close to $\mathcal{R}_P$ provided that T is large enough. Here, f_T denotes the decision function constructed by $\mathcal{C}$ on the basis of T. Asymptotically, this means that

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

should hold in probability if $|T| \to \infty$. In this case the algorithm $\mathcal{C}$ is called *consistent* for P (cf. Devroye et al., 1997, Def. 6.1). If a classifier is consistent for all distributions on

[1] For mathematical notions see Section 2.

$X \times Y$ it is said to be universally consistent. Although several algorithms such as the k-nearest neighbour classifier for $k \to \infty$ and $k/|T| \to 0$ (cf. Devroye et al., 1997, Thm. 6.4) are universally consistent it is an open question whether support vector machines are universally consistent for a particular choice of the free parameters.

In this article we show that at least for a restricted class of distributions SVM's are consistent provided that the parameters are chosen in a specific manner. In particular our results cover both the noise free case and the case of constant noise level, i.e $p \equiv p^*$, $p^* \in [0, 1/2)$. Our results are based on an investigation of a certain approximation property of kernels. Recall that the ansatz of SVM's is to solve specific optimization problems over the class of functions $\{\langle w, \Phi(.)\rangle : w \in H\}$ (or $\{\langle w, \Phi(.)\rangle + b : w \in H, b \in \mathbb{R}\}$), where $\Phi : X \to H$ is a feature map of the used kernel. If this function class is dense in $C(X)$ we shall call the corresponding kernel *universal* (cf. Def. 4). Roughly speaking this notion enables us to approximate the Bayes decision rule in probability, a fact that is frequently used in our proofs of consistency. Using the approximation theorem of Stone-Weierstraß we show that kernels that can be expanded in certain types of Taylor or Fourier series are universal (cf. Cor. 10 and Cor. 11). In particular it turns out that the Gaussian RBF kernel is universal (cf. Ex. 1). Besides the importance of the notion of universality in the context of consistency it also turns out that this concept has strong implications for the geometric interpretation of the shape of the feature map (cf. Cor. 6 and the following remark).

Since the class of functions implemented by an SVM with universal kernel is very rich the problem of overfitting can always occur in the presence of noise. Thus it is very important to know how to chose the regularization parameter. The second part of this work is devoted to this question. Here, we show in particular that for a soft margin SVM with Gaussian RBF kernel on $X \subset \mathbb{R}^d$ the regularization sequence $c_n = n^{\beta-1}$, $0 < \beta < 1/d$, yields consistency for all problems with constant noise level (cf. Cor. 17 and Cor. 23). These results are of special interest since they show at the first time that SVM's are able to learn noisy problems arbitrarily well. Moreover, we prove that for problems that ensure a large margin it suffices to use universal kernels and a regularization parameter that is *independent* of the training set size (cf. Thm. 18, Thm. 19 and Thm. 24). For this class of problems we also prove consistency for the maximal margin classifier (cf. Thm. 25). Finally, it turns out that even for these simple, noise free classification problems SVM's with polynomial kernels can behave arbitrarily badly (cf. Prop. 20).

This work is organized as follows: we introduce some mathematical notions in Section 2. In the third section we study the concept of universal kernels. The following sections are devoted to applications of these kernels to support vector machines. We begin with the 2-soft margin classifier in Section 4. Here, consistency results for both noisy distributions and problems ensuring a large margin are proved. In the fifth section we show that these results also hold for the 1-soft margin algorithm. In the last section we discuss the maximal margin classifier in the presence of large margins.

2. Preliminaries

For a set X, a metric d on X is a function $d : X \times X \to [0, \infty)$ such that for all $x, y, z \in X$ we have $d(x, y) = d(y, x)$ and $d(x, y) \leq d(x, z) + d(z, y)$ as well as $d(x, y) = 0$ if and only if $x = y$. We denote the closed ball with radius ε and centre x by $B_d(x, \varepsilon) := \{y \in X : d(x, y) \leq \varepsilon\}$.

The covering numbers of X are defined by

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

for all $\varepsilon > 0$. The space (X,d) is precompact if and only if $\mathcal{N}((X,d),\varepsilon)$ is finite for all $\varepsilon > 0$. Moreover, X is called compact if every open covering of X has a finite subcovering. If the space (X,d) is complete, i.e. every Cauchy sequence converges in X, then X is compact if and only if X is precompact. For given $A, B \subset X$ we denote the distance of A and B by

$$d(A,B) \ := \ \inf_{\substack{x\in A\\ y\in B}} d(x,y).$$

We often use that if A is closed, B is compact and both sets are disjoint then $d(A,B) > 0$ holds.

A σ-algebra on a set X is a set of subsets of X that contains $\emptyset$ and is closed under elementary, countable set operations such as complements and countable intersections. For a metric space (X,d) the Borel σ-algebra is the smallest σ-algebra that contains all open sets. Let $\mathcal{A}$ be a σ-algebra on X. A subset A of X is called measurable if $A \in \mathcal{A}$. We say that a function $f : X \to \mathbb{R}$ is measurable if the pre-image of every Borel measurable $B \subset \mathbb{R}$ is in $\mathcal{A}$. Basic examples are the functions $\mathbf{1}_A$, where $A \in \mathcal{A}$ and $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ otherwise. A probability measure $P : \mathcal{A} \to \mathbb{R}^+$ is a σ-additive function with $P(\emptyset) = 0$ and $P(X) = 1$. If $\mathcal{A}$ is a Borel σ-algebra we call P a Borel probability measure. In this case P is said to be regular, if for all Borel measurable $B \subset X$ we have

$$P(B) \ = \ \sup\{P(K) : K \subset B,\, K \text{ compact}\} \ .$$

If (X,d) is compact, then every Borel probability measure on X is regular (cf. Dudley, 1989, p. 176).

In this paper H always denotes a Hilbert space, i.e. a complete, normed vector space endowed with a dot product $\langle .,. \rangle$ giving rise to its norm via $\|x\| = \sqrt{\langle x,x\rangle}$. Let $B_H := \{x \in H : \|x\| \leq 1\}$ be the closed unit ball of H and $S_H := \{x \in H : \|x\| = 1\}$ be its sphere. Recall, that every separable Hilbert space is isometrically isomorphic to the space of all 2-summable sequences ℓ_2.

A commutative algebra A is a vector space equipped with an additional associative and commutative multiplication $\cdot : A \times A \to A$ such that

$$\begin{aligned} x\cdot(y+z) &\ = \ x\cdot y + x\cdot z \\ \lambda(x\cdot y) &\ = \ (\lambda x)\cdot y \end{aligned}$$

holds for all $x, y, z \in A$ and $\lambda \in \mathbb{R}$. A classical example of an algebra is the space $C(X)$ of all continuous functions $f : X \to \mathbb{R}$ on the compact metric space (X,d) endowed with the usual supremum norm

$$\|f\|_\infty \ := \ \sup_{x\in X} |f(x)| \ .$$

The following well-known approximation theorem of Stone-Weierstraß (cf. Pedersen, 1988, Cor. 4.3.5.) states that certain subalgebras of $C(X)$ generate the whole space. This result will be the key tool when considering approximation properties of kernels in the next section:

Theorem 1 *Let (X, d) be a compact metric space and $A \subset C(X)$ be an algebra. Then A is dense in $C(X)$ if both A does not vanish, i.e. for all $x \in X$ there exists an $f \in A$ with $f(x) \neq 0$, and A separates points, i.e. for all $x, y \in X$ with $x \neq y$ there exists an $f \in A$ with $f(x) \neq f(y)$.*

3. Kernels

In the following let (X, d) be a metric space. A function $k : X \times X \to \mathbb{R}$ is called a kernel on X if there exists a Hilbert space H and a map $\Phi : X \to H$ with

$$k(x, y) \ = \ \langle \Phi(x), \Phi(y) \rangle$$

for all $x, y \in X$. We call Φ a *feature map* and H a *feature space* of k. Note, that both H and Φ are far from being unique. However, for a given kernel there exists a canonical feature space (with associated feature map), which is the so-called reproducing kernel Hilbert space (RKHS) (cf. Cristianini & Shawe-Taylor, 2000, Ch. 3). Since our ansatz is mainly based on specific series expansions of certain kernels we do not need to consider these spaces.

Let k be a kernel on X and $\Phi : X \to H$ be a feature map of k. A function $f : X \to \mathbb{R}$ is *induced* by k if there exists an element $w \in H$ such that $f = \langle w, \Phi(.) \rangle$. The next lemma shows that this definition is independent of Φ and H:

Lemma 2 *Let $k : X \times X \to \mathbb{R}$ be a kernel and $\Phi_1 : X \to H_1$, $\Phi_2 : X \to H_2$ be two feature maps of k. Then for all $w_1 \in H_1$ there exist $w_2 \in H_2$ with $\|w_2\| \leq \|w_1\|$ and*

$$\langle w_1, \Phi_1(x) \rangle \ = \ \langle w_2, \Phi_2(x) \rangle \qquad \textit{for all } x \in X.$$

Proof Let $H_1^* := \overline{\operatorname{span} \Phi_1(X)}$ and $\tilde{H}_1$ its orthogonal complement in H_1. Then $w_1 \in H_1$ can be written as $w_1 = w_1^* + \tilde{w}_1$ with $w_1^* \in H_1^*$ and $\tilde{w}_1 \in \tilde{H}_1$. Given an $x \in X$ we have $\langle \tilde{w}_1, \Phi_1(x) \rangle = 0$ and therefore we obtain $\langle w_1^*, \Phi_1(x) \rangle = \langle w_1, \Phi_1(x) \rangle$ for all $x \in X$. Now by the definition of H_1^* there exists a sequence $(w_n^{(1)}) \subset \operatorname{span} \Phi_1(X)$ with $w_n^{(1)} = \sum_{m=1}^{m_n} \lambda_m^{(n)} \Phi_1(x_m^{(n)})$ and $w_1^* = \sum_{n=1}^{\infty} w_n^{(1)}$. Then for $w_n^{(2)} := \sum_{m=1}^{m_n} \lambda_m^{(n)} \Phi_2(x_m^{(n)})$ and $l_2 \geq l_1 \geq 1$ we obtain

$$\begin{aligned}
\Big\| \sum_{n=l_1}^{l_2} w_n^{(1)} \Big\|^2 \ &= \ \sum_{n=l_1}^{l_2} \sum_{m=1}^{m_n} \sum_{i=l_1}^{l_2} \sum_{j=1}^{m_i} \lambda_m^{(n)} \lambda_j^{(i)} \langle \Phi_1(x_m^{(n)}), \Phi_1(x_j^{(i)}) \rangle \\
&= \ \sum_{n=l_1}^{l_2} \sum_{m=1}^{m_n} \sum_{i=l_1}^{l_2} \sum_{j=1}^{m_i} \lambda_m^{(n)} \lambda_j^{(i)} \langle \Phi_2(x_m^{(n)}), \Phi_2(x_j^{(i)}) \rangle \\
&= \ \Big\| \sum_{n=l_1}^{l_2} w_n^{(2)} \Big\|^2 .
\end{aligned}$$

Therefore, $(\sum_{n=1}^{m} w_n^{(2)})_{m \geq 1}$ is a Cauchy sequence and hence converges to $w_2 := \sum_{n=1}^{\infty} w_n^{(2)} \in H_2$. Clearly, we then have $\|w_2\| = \|w_1^*\| \leq \|w_1\|$. Moreover, an easy calculation similar to the consideration above shows $\langle w_1, \Phi_1(x) \rangle = \langle w_2, \Phi_2(x) \rangle$ for all $x \in X$. ■

In the following we only consider continuous kernels. The following lemma provides some useful properties of this class:

Lemma 3 *Let k be a kernel on the metric space (X, d) and $\Phi : X \to H$ be a feature map of k. Then k is continuous if and only if Φ is continuous. In this case*

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

defines a semi-metric on X such that the identity map $\mathrm{id} : (X, d) \to (X, d_k)$ *is continuous. If Φ is injective d_k is even a metric.*

Proof Let us first suppose that k is continuous. Since

$$d_k(x, y) \ = \ \sqrt{k(x, x) - 2k(x, y) + k(y, y)}$$

we observe that $d_k(x, .) : (X, d) \to \mathbb{R}$ is continuous for every $x \in X$. In particular, $\{y \in X : d_k(x, y) < \varepsilon\}$ is open with respect to d and therefore $\mathrm{id} : (X, d) \to (X, d_k)$ is continuous. Furthermore, $\Phi : (X, d_k) \to H$ is continuous and hence $\Phi : (X, d) \to H$ is also continuous. Conversely, assume that Φ is continuous. Since for all $x, x', y, y' \in X$ we have

$$\begin{aligned} |k(x, y) - k(x', y')| \ &\leq \ |\langle \Phi(x), \Phi(y) - \Phi(y') \rangle| + |\langle \Phi(x) - \Phi(x'), \Phi(y') \rangle| \\ &\leq \ \|\Phi(x)\| \cdot \big\|\Phi(y) - \Phi(y')\big\| + \big\|\Phi(y')\big\| \cdot \big\|\Phi(x) - \Phi(x')\big\| \end{aligned}$$

it is easily verified that k is also continuous. ■

The metric d_k enjoys the property that every induced function $\langle w, \Phi(.) \rangle$ is Lipschitz continuous with respect to d_k and the Lipschitz constant is bounded from above by $\|w\|$. This fact turns out to be very important in the proof of Theorem 12 since it allows us to control the behaviour of solutions of SVM's on subsets of small diameters.

From the last lemma we know in particular that for a continuous kernel every induced function is continuous. The following definition plays a central role throughout this paper:

Definition 4 *A continuous kernel k on a compact metric space (X, d) is called* universal *if the space of all functions induced by k is dense in $C(X)$, i.e. for every function $f \in C(X)$ and every $\varepsilon > 0$ there exists a function g induced by k with*

$$\|f - g\|_\infty \ \leq \ \varepsilon \ .$$

We also need a weaker concept. Let $\Phi : X \to H$ be a feature map of k and A, B be disjoint subsets of X. We say that k *separates A and B with margin $\gamma \geq 0$* if $\Phi(A)$ and $\Phi(B)$ can be separated by a hyperplane with margin γ, i.e. if there exists a pair $(w, b) \in S_H \times \mathbb{R}$ such that

$$\begin{aligned} \langle w, \Phi(x) \rangle + b \ &> \ \gamma \qquad && \text{for all } x \in A \text{ and} \\ \langle w, \Phi(y) \rangle + b \ &< \ -\gamma \qquad && \text{for all } y \in B \ . \end{aligned}$$

If $\gamma = 0$ we simply say that k separates A and B. In this case the restriction $w \in S_H$ is superfluous. By Lemma 2 both definitions are independent of the feature map Φ. We say that the kernel k separates all finite, resp. compact subsets if it separates all finite, resp.

compact disjoint subsets of X. Note, that if k separates *compact* sets A and B then it automatically separates them with a suitable margin $\gamma > 0$. Moreover, it was shown in Steinwart (2001a, Ex. 3.13) that there exists a continuous kernel that separates all finite subsets but fails to separate all compact subsets.

Before we investigate which kernels are universal we collect some useful properties of these kernels. Firstly, let (X, d) and (X', d') be compact metric spaces, k be a universal kernel on X and $\iota : X' \to X$ be a continuous and injective map. Then one easily checks that $k(\iota(.), \iota(.))$ is a universal kernel on X'. Moreover, we have $k(x, x) > 0$ for all $x \in X$ since $k(y, y) = 0$ implies $g(y) = 0$ for all induced functions g. Since all feature maps of k are continuous and X is compact we may also restrict ourselves to separable feature spaces of k. The next proposition is fundamental for our considerations of support vector machines:

Proposition 5 *Let (X, d) be a compact metric space and k be a universal kernel on X. Then for all compact and mutually disjoint subsets $K_1, \dots, K_n \subset X$, all $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and all $\varepsilon > 0$ there exists a function g induced by k with $\|g\|_\infty \le \max_i |\alpha_i| + \varepsilon$ such that*

$$\Big\| g_{|K} - \sum_{i=1}^{n} \alpha_i \mathbf{1}_{K_i} \Big\|_\infty \le \varepsilon \ ,$$

where $K := \bigcup_{i=1}^{n} K_i$ and $g_{|K}$ denotes the restriction of g to K.

Proof Since $d(K_i, K_j) > 0$ for all $i \neq j$ we obtain $\sum_{i=1}^{n} \alpha_i \mathbf{1}_{K_i} \in C(K)$. Since this function can be extended to a continuous function f on X with $\|f\|_\infty \le \max_i |\alpha_i|$ (by the Lemma of Urysohn or by a direct construction with the help of d) the assertion follows. ■

Corollary 6 *Every universal kernel separates all compact subsets.*

Proof Let (X, d) be a compact metric space and k be a universal kernel on X with feature map $\Phi : X \to H$. Given two compact and disjoint subsets K_1 and K_{-1} of X there exists an induced function $g = \langle w, \Phi(.)\rangle$ with $\big\| g_{|K_{-1} \cup K_1} - (\mathbf{1}_{K_1} - \mathbf{1}_{K_{-1}}) \big\|_\infty < 1/2$. This implies that $(\|w\|^{-1} w, 0)$ separates K_1 and K_{-1} with margin $\frac{1}{2\|w\|}$. ■

Although the previous corollary is an almost trivial consequence of the notion of universality it has surprising implications for the geometric interpretation of the shape of the feature map: let us suppose that we have a finite subset $\{x_1, \dots, x_n\}$ of X. Then the above corollary ensures that for every sequence of signs $y_1, \dots, y_n$ the corresponding training set can be correctly separated by a hyperplane in the feature space. Moreover, this can even be done by a hyperplane that has almost the same distance to every point of $\{x_1, \dots, x_n\}$. Therefore, any finite dimensional interpretation of the geometric situation in a feature space of a universal kernel must fail. In particular this holds for 2- or 3-dimensional drawings. (Actually, the shape of the feature map is even more complicated since not only all finite subsets but every pair of compact disjoint subsets can be separated.)

The following corollary ensures in particular that the semi-metric d_k induced by a universal kernel k is in fact a metric:

Corollary 7 *Every feature map of a universal kernel is injective.*

Proof Finite subsets are compact and thus the assertion follows by the previous corollary. ■

Proposition 8 *Let (X, d) be a compact metric space and k be a universal kernel on X. Then*

$$k^*(x,y) \;:=\; \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

defines a universal kernel on X.

Proof Let $\Phi : X \to H$ be a feature map of k and $\alpha(x) := k(x,x)^{-1/2}$. Clearly, $\alpha\Phi : X \to H$ is a feature map of k and thus k is a kernel. To show that k^* is universal we fix a function $f \in C(X)$ and $\varepsilon > 0$. For $a := \|\alpha\|_\infty$ we then get an induced function $g = \langle w, \Phi(.)\rangle$ with $\left\|\alpha^{-1} f - g\right\|_\infty \leq \frac{\varepsilon}{a}$. This yields

$$\|f - \langle w, \alpha\Phi(.)\rangle\|_\infty \;\leq\; \|\alpha\|_\infty \left\|\alpha^{-1} f - g\right\|_\infty \;\leq\; \varepsilon$$

and thus the assertion is proved. ■

Up to now we do not know whether there exist universal kernels. To attack this question we begin with a simple criterion that makes it possible to check whether a given kernel is universal:

Theorem 9 *Let (X, d) be a compact metric space and k be a continuous kernel on X with $k(x,x) > 0$ for all $x \in X$. Suppose that we have an injective feature map $\Phi : X \to \ell_2$ of k with $\Phi(x) = (\Phi_n(x))_{n\in\mathbb{N}}$. If $A := \operatorname{span}\{\Phi_n : n \in \mathbb{N}\}$ is an algebra then k is universal .*

Proof Because of $k(x,x) > 0$ for all $x \in X$ the algebra A does not vanish. Since k is continuous every $\Phi_n : X \to \mathbb{R}$ is continuous by Lemma 3 and hence $A \subset C(X)$. Moreover, A is even dense in $C(X)$ since the injectivity of Φ implies that A separates points and thus Theorem 1 can be applied. Now we fix $f \in C(X)$ and $\varepsilon > 0$. Then there exists a function

$$g = \sum_{j=1}^{n} \lambda_j \cdot (\Phi_{n_j}) \;\in\; A$$

such that $\|f - g\|_\infty \leq \varepsilon$. However, if we define $w_n := \lambda_j$ for $n = n_j$ and $w_n := 0$ otherwise, we have $w := (w_n) \in \ell_2$ and $\langle w, \Phi(.)\rangle = g$. ■

The following corollaries give various examples of universal kernels. We begin with kernels that can be expressed by a Taylor series:

Corollary 10 *Let $0 < r \leq \infty$ and $f : (-r, r) \to \mathbb{R}$ be a C^∞-function that can be expanded into its Taylor series in 0, i.e.*

$$f(x) \;=\; \sum_{n=0}^{\infty} a_n x^n \qquad \text{for all } x \in (-r, r)\;.$$

Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\}$. If we have $a_n > 0$ for all $n \geq 0$ then $k(x,y) := f(\langle x,y\rangle)$ defines a universal kernel on every compact subset of X.

Proof Since $|\langle x,y\rangle| \leq \|x\|_2 \|y\|_2 < r$ for all $x,y \in X$ we see that k is well-defined. We also have

$$\begin{aligned} k(x,y) &= \sum_{n=0}^{\infty} a_n \Big(\sum_{k=1}^{d} x_k y_k\Big)^n \\ &= \sum_{n=0}^{\infty} a_n \sum_{\substack{k_1+\cdots+k_d=n \\ k_1,\dots,k_d \geq 0}} c_{k_1,\dots,k_d} \prod_{i=1}^{d} (x_i y_i)^{k_i} \\ &= \sum_{k_1,\dots,k_d \geq 0} a_{k_1+\cdots+k_d} c_{k_1,\dots,k_d} \prod_{i=1}^{d} x_i^{k_i} \prod_{i=1}^{d} y_i^{k_i} \ , \end{aligned}$$

where $c_{k_1,\dots,k_d} := (\prod_{i=1}^{d} k_i!)^{-1} (\sum_{i=1}^{d} k_i)!$ (cf. also Poggio, 1975, Lem. 2.1). Note, that the series can be rearranged since it is absolutely summable. In particular, for $x = y$ we obtain that $\Phi : X \to \ell_2(\mathbb{N}_0^d)$ is well defined by

$$\Phi(x) := \Big(\sqrt{a_{k_1+\cdots+k_d} c_{k_1,\dots,k_d}} \prod_{i=1}^{d} x_i^{k_i}\Big)_{k_1,\dots,k_d \geq 0} .$$

The above equation also shows that $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$ holds for all $x,y \in X$ and hence k is indeed a kernel. Moreover, $a_0 > 0$ implies $k(x,x) > 0$ for all $x \in X$ and trivially, Φ is injective. Since $A := \operatorname{span}\{\Phi_{k_1,\dots,k_d} : k_1,\dots,k_d \geq 0\}$ is an algebra we thus obtain by Theorem 9 that k is universal. ■

Instead of Taylor series one can also consider Fourier expansions. The result reads as follows:

Corollary 11 *Let $f : [0,2\pi] \to \mathbb{R}$ be a continuous function that can be expanded in a pointwise absolutely convergent Fourier series of the form*

$$f(t) = \sum_{n=0}^{\infty} a_n \cos(nt) \ . \tag{2}$$

If $a_n > 0$ holds for all $n \geq 0$ then $k(x,y) := \prod_{i=1}^{d} f(|x_i - y_i|)$ defines a universal kernel on every compact subset of $[0,2\pi)^d$.

Recall, that every function $f : [0,2\pi] \to \mathbb{R}$ that can be extended to a continuous, symmetric, periodic and piecewise continuously differentiable function on $\mathbb{R}$ has a Fourier series of the form (2).

Proof By induction and the Cauchy product of series we may restrict ourselves to $d = 1$. Then

$$k(x,y) = a_0 + \sum_{n=1}^{\infty} a_n \sin(nx)\sin(ny) + \sum_{n=1}^{\infty} a_n \cos(nx)\cos(ny)$$

holds for all $x, y \in [0, 2\pi)$ and hence $\Phi = (\Phi_n)_{n\geq 0}$ defined by $\Phi_0(x) := a_0$ and $\Phi_{2n-1}(x) := \sqrt{a_n}\sin(nx)$, $\Phi_{2n}(x) := \sqrt{a_n}\cos(nx)$ for $n \geq 1$ is an injective feature map of k with image in ℓ_2. Moreover, $A := \operatorname{span}\big(\{\sqrt{a_n}\sin(n\cdot .) : n \geq 1\} \cup \{\sqrt{a_0}\cos(n\cdot .) : n \geq 0\}\big)$ is an algebra and since $a_0 > 0$ implies $k(x,x) > 0$ for all $x \in X$ we obtain that k is universal. ■

The following examples show that various well-known kernels are universal:

Example 1 The kernels $\exp(-\sigma^2\|. - .\|_2^2)$ and $\exp(\langle ., .\rangle)$ are universal on every compact subset of $\mathbb{R}^d$.

Proof The universality of $\exp(\langle ., .\rangle)$ is due to Corollary 10. Therefore, by Proposition 8 and $\exp(-\sigma^2\|x - y\|_2^2) = \exp(-\|\sigma x\|_2^2)\exp(-\|\sigma y\|_2^2)\exp(\langle\sqrt{2}\sigma x, \sqrt{2}\sigma y\rangle)$ the assertion follows for the RBF kernel. ■

Example 2 Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$ and $\alpha > 0$. Then V. Vovk's (cf. Saunders et al., 1998, p. 15) infinite polynomial kernel $k(x,y) := (1 - \langle x, y\rangle)^{-\alpha}$, $x, y \in X$, is universal on every compact subset of X.

Proof To check the assertion we use that $(1-t)^{-\alpha} = \sum_{n=0}^{\infty}\binom{-\alpha}{n}(-1)^n t^n$ holds for $|t| < 1$. Since $\binom{-\alpha}{n}(-1)^n > 0$ for all $n \geq 0$, the assertion then follows by Corollary 10. ■

Example 3 Let $0 < q < 1$ and $f(t) := (1 - q^2)/(2 - 4q\cos t + 2q^2)$, $t \in \mathbb{R}$. Then the stronger regularized Fourier kernel $k(x,y) := \prod_{i=1}^{d} f(x_i - y_i)$ considered by Vapnik (1998, p. 470) and Saunders et al. (1998, p. 15) is universal on every compact subset of $[0, 2\pi)^d$.

Proof The assertion can be seen using Corollary 11 and $f(t) = 1/2 + \sum_{n=1}^{\infty} q^n\cos(nt)$ (cf. Gradstein & Ryshik, 1981, p. 68). ■

Example 4 Let $0 < q < \infty$ and $f(t) := \pi\cosh\big((\pi - |t|)/q\big)/\big(2q\sinh(\pi/q)\big)$ for all t with $-2\pi \leq t \leq 2\pi$. Then the weaker regularized Fourier kernel $k(x,y) := \prod_{i=1}^{d} f(x_i - y_i)$ considered by Vapnik (1998, p. 470/1) and Saunders et al. (1998, p. 15) is universal on every compact subset of $[0, 2\pi)^d$.

Proof To obtain the assertion we use $f(t) = 1/2 + \sum_{n=1}^{\infty}\cos(nt)/(1 + q^2n^2)$ (cf. Gradstein & Ryshik, 1981, p. 68). ■

4. The 2-norm soft margin classifier

Let k be a kernel on X and $\Phi : X \to H$ be a feature map of k. For a training set $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ and $c_n > 0$ we denote the unique solution of the optimization problem

$$\begin{array}{lll}\text{minimize} & \mathcal{W}(w,b,\xi) := \langle w,w\rangle + c_n \sum_{i=1}^{n} \xi_i^2 & \text{over } w,b,\xi \\ \text{subject to} & y_i(\langle w,\Phi(x_i)\rangle + b) \geq 1-\xi_i, & i=1,\dots,n\end{array} \tag{3}$$

by $(w_T^{2,k,c_n}, b_T^{2,k,c_n}) \in H \times \mathbb{R}$. An algorithm $\mathcal{C}_k^{2,(c_n)}$ that provides the decision function

$$f_T^{2,k,c_n}(x) := \operatorname{sign}(\langle w_T^{2,k,c_n}, \Phi(x)\rangle + b_T^{2,k,c_n})\ , \qquad x \in X$$

for every training set T is called a 2-norm soft margin classifier (2-SMC) with kernel k and parameter sequence (c_n). Note, that in order to have a small set of free parameters one usually fixes $c_n := c$ for all $n \geq 1$. In this section it turns out that this is not suitable for problems that do no guarantee a large margin. Instead one should use sequences $c_n = cn^{\beta-1}$ where $\beta > 0$ is a parameter a-priori determined by the kernel and c is a new free parameter (cf. Cor. 17). Of course, for fixed training set sizes both parameterizations are equivalent, i.e. they can be transformed into each other.

By Lemma 2 the decision function is independent of the choice of the feature map Φ. Moreover, f_T^{2,k,c_n} can be expressed by

$$f_T^{2,k,c_n}(x) \;=\; \sum_{i=1}^{n} y_i \alpha_i k(x_i,x) + b_T^{2,k,c_n},$$

where $\alpha_i \geq 0$ are suitable constants depending on T and b_T^{2,k,c_n} can also be computed with the help of the kernel (cf. Cristianini & Shawe-Taylor, 2000; Vapnik, 1998; Schölkopf et al., 2001). Note, that if k is a kernel on X which separates all finite sets and X has infinitely many elements then the function class represented by the 2-SMC has infinite VC-dimension. For more information on this we refer to Vapnik (1998, Ch. 4), Cristianini & Shawe-Taylor (2000, Ch. 4) and van der Vaart & Wellner (1996, Ch. 2.6).

Given a Borel probability measure P on $X \times Y$ with noise level p we denote the nondeterministic part of the supervisor by $X^+ := \{x \in X : p(x) > 0\}$. If $P_X(X^+) > 0$ we write $q^* := \inf_{x\in X} p(x)$ and $p^* := \sup_{x \in X} p(x)$. Due to technical reasons we define $q^* := p^* := 1/4$ otherwise. We begin with a preliminary result:

Theorem 12 *Let (X,d) be a compact metric space and k be a universal kernel on X. Then for all Borel probability measures P on $X \times Y$ with $q^*, p^* \in (0, 1/2)$ and all $\varepsilon > 0$ there exist $c^* > 0$ and $\delta^* > 0$ such that for all $c \geq c^*$, $0 < \delta \leq \delta^*$ and all $n \geq 1$ we have*

$$\Pr{}^*\Big(\Big\{T \in (X\times Y)^n : \mathcal{R}_P(f_T^{2,k,c/n}) \leq \mathcal{R}_P + 4\frac{p^*-q^*}{1-2q^*}P_X(X^+) + \varepsilon\Big\}\Big) \;\geq\; 1 - 3Me^{-2\left(\frac{\delta}{M}\right)^2 n}\ ,$$

where $M := \mathcal{N}\big((X,d_k), \frac{\delta}{\sqrt{c}}\big)$ is the covering number of X with respect to the metric d_k which is induced by the kernel k. Moreover, $\Pr^*$ *denotes the outer probability measure of P^n.*

Note, that in order to avoid the (probably very difficult) question whether the sets

$$\big\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{2,k,c/n}) \leq \alpha\big\}$$

are measurable we consider the outer probability measure, only.

Since the proof of Theorem 12 is very technical we like to explain the basic idea of the proof firstly. Let us suppose that the supervisor has a constant level of noise $p \in [0, 1/2)$. Moreover, we assume that we have an induced function $\langle w^*, \Phi(.)\rangle$ which has the constant values $1-2p$, resp. $-(1-2p)$ on $B_1(P)$, resp. $B_{-1}(P)$. Now let us take a "representative" training set T of length n. Then one easily checks (cf. estimate (4)) that

$$\langle w_T^{k,c/n}, w_T^{k,c/n}\rangle + \frac{c}{n}\sum_{l=1}^{n} \xi_l^2 \quad \lesssim \quad \langle w^*, w^*\rangle + 4cp(1-p)$$

holds. Here $\lesssim$ means that the relation $\leq$ only holds "approximately". On the other hand, by the continuity of the decision function $w_T^{k,c/n}$ a misclassified (compared with the optimal Bayes decision rule) element z forces the sum of those slack variables, which belong to samples in the "neighbourhood" of z, to be "approximately" greater than their cardinality (cf. inequality (6)). Conversely, for a correctly classified element the corresponding sum of the slack variables is "approximately" larger than $4p(1-p)$ times their cardinality (cf. inequality (7)). Combining these considerations we obtain

$$\begin{aligned} c(1-2p)^2 P_X(E) + 4cp(1-p) \quad &= \quad c\big(P_X(E) + 4p(1-p)P_X(X\setminus E)\big) \\ &\lesssim \quad \langle w_T^{k,c/n}, w_T^{k,c/n}\rangle + \frac{c}{n}\sum_{l=1}^{n}\xi_l^2 \\ &\lesssim \quad \langle w^*, w^*\rangle + 4cp(1-p)\ , \end{aligned}$$

where E denotes the set of misclassified (compared with the optimal Bayes decision rule) elements. Thus $P_X(E)$ must be "small" if we have chosen c "large enough".

The difficulty of the proof below is firstly, to transfer the idea to the general case and secondly, to give exact formulations of "representative", "neighbourhood" and "approximately".

Proof of Theorem 12 Without loss of generality we may assume $\varepsilon \in (0, 1]$. Let $X^0 := X \setminus X^+$ be the deterministic part of the supervisor and $X^0_{-1} := X^0 \cap B_{-1}(P)$, $X^0_1 := X^0 \cap B_1(P)$ be the parts of the classes $B_{-1}(P)$, $B_1(P)$ in X^0. Furthermore, let $X^+_{-1} := X^+ \cap B_{-1}(P)$ and $X^+_1 := X^+ \cap B_1(P)$ be the parts of the classes $B_{-1}(P)$ and $B_1(P)$ in X^+. We define $\delta^* := \min\{2p^*, \frac{\varepsilon}{74}q^*(1-2q^*)\}$ and fix $\delta \leq \delta^*$. Since P_X is regular (cf. Dudley, 1989, p. 176) there exist compact subsets K_i^j of X_i^j with $P_X(X_i^j \setminus K_i^j) \leq \delta/4$ for $i \in \{-1, 1\}$ and $j \in \{0, +\}$. Moreover, for a fixed feature map $\Phi : X \to H$ of k Proposition 5 ensures the existence of an element $w^* \in H$ with

$$\langle w^*, \Phi(x)\rangle \quad \in \quad \begin{cases} [1, 1+\delta] & \text{if } x \in K_1^0 \\ [-(1+\delta), -1] & \text{if } x \in K_{-1}^0 \\ [1-2p^*-\delta, 1-2p^*] & \text{if } x \in K_1^+ \\ [-(1-2p^*), -(1-2p^*-\delta)] & \text{if } x \in K_{-1}^+ \\ [-(1+\delta), 1+\delta] & \text{otherwise.} \end{cases}$$

We define $c^* := \frac{2}{\varepsilon(1-2q^*)}\|w^*\|_2^2$ and for fixed $c \geq c^*$ let $\sigma := \frac{\delta}{\sqrt{c}}$. By Lemma 13 we then obtain partitions $\tilde{\mathcal{P}}_i^j$ of K_i^j with $\mathrm{diam}_{d_k}(A) \leq \sigma$ for all $A \in \tilde{\mathcal{P}}_i^j$ and

$$\Big| \bigcup_{\substack{i\in\{-1,1\}\\ j\in\{0,+\}}} \tilde{\mathcal{P}}_i^j \Big| \;\leq\; \mathcal{N}\big((X,d_k),\sigma\big) \;=\; M\ .$$

Let $\mathcal{P}_i^j := \{A \in \tilde{\mathcal{P}}_i^j : P_X(A) \geq \frac{\delta}{q^*M}\}$ for $i \in \{-1,1\}$ and $j \in \{0,+\}$. To construct "representative" training sets we define

$$F_{n,A} := \Big\{ \big((x_1,y_1),\dots,(x_n,y_n)\big) \;:\; \big|\{l : x_l \in A\}\big| \geq \Big(P_X(A) - \frac{\delta}{M}\Big)\, n \Big\}$$

for all $A \in \mathcal{P}_i^j$, $i \in \{-1,1\}$ and $j \in \{0,+\}$. Moreover, for $A \in \mathcal{P}_i^+$, $i \in \{-1,1\}$ let

$$\begin{aligned}
F_{n,A}^+ &:= \Big\{ \big((x_1,y_1),\dots,(x_n,y_n)\big) \;:\; \big|\{l : x_l \in A \text{ and } y_l = i\}\big| \geq \Big((1-p^*)P_X(A) - \frac{\delta}{M}\Big)\, n \Big\}\\
F_{n,A}^- &:= \Big\{ \big((x_1,y_1),\dots,(x_n,y_n)\big) \;:\; \big|\{l : x_l \in A \text{ and } y_l \neq i\}\big| \geq \Big(q^*P_X(A) - \frac{\delta}{M}\Big)\, n \Big\}\\
F_n &:= \bigcap_{A\in\mathcal{P}_{-1}^0\cup\mathcal{P}_1^0} F_{n,A} \quad\cap \bigcap_{A\in\mathcal{P}_{-1}^+\cup\mathcal{P}_1^+} \big(F_{n,A}\cap F_{n,A}^+ \cap F_{n,A}^-\big)\ .
\end{aligned}$$

Lemma 14 yields $P^n(F_n) \geq 1 - 3Me^{-2(\frac{\delta}{M})^2 n}$ and thus it suffices to show that

$$\mathcal{R}_P(f_T^{2,k,c/n}) \leq \mathcal{R}_P + 4\frac{p^*-q^*}{1-2q^*}P_X(X^+) + \varepsilon$$

holds for all $T \in F_n$. Therefore, let us fix a training set $T = \big((x_1,y_1),\dots,(x_n,y_n)\big) \in F_n$. For $c_n := c/n$ we denote the solution of (3) by (w_T, b_T, ξ_T). Our first step is to estimate $\mathcal{W}(w_T,b_T,\xi_T)$ from above by comparing it with $\mathcal{W}(w^*,0,\xi^*)$, where ξ^* is an admissible slack vector of $(w^*,0)$. Hence we have to construct ξ^*. For this let us first assume that we have a sample $(x_l,y_l) \in K_1^0$. Then we observe that

$$y_l\langle w^*, \Phi(x_l)\rangle \;=\; \langle w^*, \Phi(x_l)\rangle \;\geq\; 1$$

and thus we may define $\xi_l^* := 0$. Analogous considerations yield

$$\xi_l^* \quad := \quad \begin{cases} 0 & \text{if } x_l \in K_i^0\\ 2p^*+\delta & \text{if } x_l \in K_i^+, y_l = i\\ 2-2p^* & \text{if } x_l \in K_i^+, y_l \neq i\\ 2+\delta & \text{otherwise.}\end{cases}$$

Moreover, our construction of F_n guarantees that there are at most

$$n - n\sum_{\substack{i\in\{-1,1\}\\ j\in\{0,+\}}}\sum_{A\in\mathcal{P}_i^j}\Big(P_X(A) - \frac{\delta}{M}\Big) \;\leq\; n\Big(1 - \sum_{\substack{i\in\{-1,1\}\\ j\in\{0,+\}}} P_X(X_i^j) + \Big(2 + \frac{1}{q^*}\Big)\delta\Big) \;=\; \frac{2}{q^*}\delta n$$

samples which are not elements of a suitable K_i^j. Furthermore, since there are at most

$$n - n \sum_{i\in\{-1,1\}} \sum_{A\in\mathcal{P}_i^0} \Big(P_X(A) - \frac{\delta}{M}\Big) \;\le\; n\Big(P_X(X^+) + \frac{2}{q^*}\delta\Big)$$

samples in $K^+ := K_1^+ \cup K_{-1}^+$ we also obtain that there are at most

$$n\Big(P_X(X^+) + \frac{2}{q^*}\delta\Big) - n \sum_{i\in\{-1,1\}} \sum_{A\in\mathcal{P}_i^+} \Big((1-p^*)P_X(A) - \frac{\delta}{M}\Big) \;\le\; n\Big(p^* P_X(X^+) + \frac{4}{q^*}\delta\Big)$$

samples in K^+ which have incorrect labels. Since these have larger slack variables with respect to $(w^*, 0)$ than the correctly labeled samples in K^+ we obtain

$$\begin{aligned}
&\mathcal{W}(w_T, b_T, \xi_T)\\
\le\;& \mathcal{W}(w^*, 0, \xi^*)\\
\le\;& \langle w^*, w^*\rangle + \frac{c}{n} \sum_{i\in\{-1,1\}} \sum_{\substack{x_l\in K_i^+\\ y_l=i}} (\xi_l^*)^2 + \frac{c}{n} \sum_{i\in\{-1,1\}} \sum_{\substack{x_l\in K_i^+\\ y_l\neq i}} (\xi_l^*)^2 + 2\delta c(2+\delta)^2\\
\le\;& \langle w^*, w^*\rangle + c(1-p^*)P_X(X^+)(2p^*+\delta)^2 + c\Big(p^*(P_X(X^+) + \frac{4}{q^*}\delta\Big)(2-2p^*)^2 + 2\delta c(2+\delta)^2\\
\le\;& \langle w^*, w^*\rangle + c\Big(4p^*(1-p^*)P_X(X^+) + \frac{27}{q^*}\delta\Big)\,. \qquad (4)
\end{aligned}$$

For later purposes we note that we also have $\mathcal{W}(w_T, b_T, \xi_T) \le \mathcal{W}\big(0, 0, (1, \ldots, 1)\big) \le c$ and thus $\|w_T\|_2 \le \sqrt{c}$. In the second step of the proof we estimate $\mathcal{W}(w_T, b_T, \xi_T)$ from below. For this let us denote the set of misclassified points in X_i^j by

$$E_i^j \;:=\; \{x \in X_i^j : f_T(x) \neq i\}\;.$$

For brevity's sake we also write $E^j := E_1^j \cup E_{-1}^j$ and $E := E^0 \cup E^+$. Let us first consider an $A \in \mathcal{P}_i^0$ with $A \cap E \neq \emptyset$. Without loss of generality we may assume that $i = 1$. Then for $x_l \in A$ and $z \in A \cap E$ we obtain

$$\begin{aligned}
1 - \xi_T(l) \;&\le\; y_l\big(\langle w_T, \Phi(x_l)\rangle + b_T\big)\\
&=\; \langle w_T, \Phi(x_l) - \Phi(z)\rangle + \langle w_T, \Phi(z)\rangle + b_T\\
&\le\; \|w_T\|\; d_k(x_l, z)\\
&\le\; \delta,
\end{aligned}$$

i.e. $\xi_T^2(l) \ge (1-\delta)^2 \ge 1 - 2\delta$. Since the same estimate holds in the case $i = -1$ our construction of F_n implies

$$\frac{1}{n} \sum_{i\in\{-1,1\}} \sum_{\substack{A\in\mathcal{P}_i^0\\ A\cap E\neq\emptyset}} \sum_{x_l\in A} \xi_T^2(l) \;\ge\; \sum_{i\in\{-1,1\}} \sum_{\substack{A\in\mathcal{P}_i^0\\ A\cap E\neq\emptyset}} (1-2\delta)\Big(P_X(A) - \frac{\delta}{M}\Big) \;\ge\; P_X(E^0) - \frac{3}{q^*}\delta\,. \qquad (5)$$

Now let us consider an $A \in \mathcal{P}_i^+$ with $A \cap E \neq \emptyset$. Without loss of generality we may assume that $i = 1$ again. Then for fixed $z \in A \cap E$ and $a := -\big(\langle w_T, \Phi(z)\rangle + b_T\big) \geq 0$ we obtain

$$\xi_T(l) \quad \geq \quad \begin{cases} 1-\delta+a & \text{for } x_l \in A \text{ with } y_l = 1 \\ \max\{0, 1-\delta-a\} & \text{for } x_l \in A \text{ with } y_l = -1 \end{cases}$$

analogously to the above considerations. We first treat the case $1-\delta-a \geq 0$. Since there are at least $\big(P_X(A) - \delta/M\big)n$ samples in A and at least $\big((1-p^*)P_X(A) - \delta/M\big)n$ correctly labeled samples in A we get

$$\begin{aligned} \frac{1}{n}\sum_{x_l \in A} \xi_T^2(l) \quad &\geq \quad (1-\delta+a)^2\Big((1-p^*)P_X(A) - \frac{\delta}{M}\Big) + (1-\delta-a)^2 p^* P_X(A) \\ &\geq \quad P_X(A)\big((1-\delta+a)^2(1-p^*) + (1-\delta-a)^2 p^*\big) - \frac{4\delta}{M}\ . \end{aligned}$$

Now an easy minimization with respect to $a \in [0, 1-\delta]$ yields

$$\frac{1}{n}\sum_{x_l \in A} \xi_T^2(l) \ \geq \ P_X(A)\big((1-\delta)^2(1-p^*) + (1-\delta)^2 p^*\big) - \frac{4\delta}{M} \ \geq \ (1-2\delta)P_X(A) - \frac{4\delta}{M}\ .$$

On the other hand, if $1-\delta-a < 0$ we have $1-\delta+a > 2-2\delta$ and thus the same inequality follows:

$$\frac{1}{n}\sum_{x_l \in A} \xi_T^2(l) \ \geq \ (1-\delta+a)^2\Big((1-p^*)P_X(A) - \frac{\delta}{M}\Big) \ \geq \ (1-2\delta)P_X(A) - \frac{4\delta}{M}\ .$$

Therefore, we obtain

$$\frac{1}{n}\sum_{i \in \{-1,1\}} \sum_{\substack{A \in \mathcal{P}_i^+ \\ A \cap E \neq \emptyset}} \sum_{x_l \in A} \xi_T^2(l) \ \geq \ \sum_{i \in \{-1,1\}} \sum_{\substack{A \in \mathcal{P}_i^+ \\ A \cap E \neq \emptyset}} \Big((1-2\delta)P_X(A) - \frac{4\delta}{M}\Big)\ . \tag{6}$$

Finally, an analogous consideration yields

$$\frac{1}{n}\sum_{i \in \{-1,1\}} \sum_{\substack{A \in \mathcal{P}_i^+ \\ A \cap E = \emptyset}} \sum_{x_l \in A} \xi_T^2(l) \ \geq \ \sum_{i \in \{-1,1\}} \sum_{\substack{A \in \mathcal{P}_i^+ \\ A \cap E = \emptyset}} \Big((1-2\delta)4q^*(1-q^*)P_X(A) - \frac{4\delta}{M}\Big)\ . \tag{7}$$

Now, the estimates (6) and (7) imply

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\{-1,1\}}\sum_{\substack{A\in\mathcal{P}_i^+\\ A\cap E\neq\emptyset}}\sum_{x_l\in A}\xi_T^2(l)+\frac{1}{n}\sum_{i\in\{-1,1\}}\sum_{\substack{A\in\mathcal{P}_i^+\\ A\cap E=\emptyset}}\sum_{x_l\in A}\xi_T^2(l)\\
\geq\;& \sum_{i\in\{-1,1\}}\sum_{\substack{A\in\mathcal{P}_i^+\\ A\cap E\neq\emptyset}}\Bigl((1-2\delta)P_X(A)-\frac{4\delta}{M}\Bigr)+\sum_{i\in\{-1,1\}}\sum_{\substack{A\in\mathcal{P}_i^+\\ A\cap E=\emptyset}}\Bigl((1-2\delta)4q^*(1-q^*)P_X(A)-\frac{4\delta}{M}\Bigr)\\
\geq\;& \sum_{i\in\{-1,1\}}\sum_{\substack{A\in\mathcal{P}_i^+\\ A\cap E\neq\emptyset}}(1-2\delta)(1-2q^*)^2P_X(A)+\sum_{\substack{i\in\{-1,1\}\\ A\in\mathcal{P}_i^+}}(1-2\delta)4q^*(1-q^*)P_X(A)-4\delta\\
\geq\;& (1-2\delta)(1-2q^*)^2P_X(E^+)+(1-2\delta)4q^*(1-q^*)P_X(X^+)-6\delta-\frac{2}{q^*}\delta\\
\geq\;& (1-2q^*)^2P_X(E^+)+4q^*(1-q^*)P_X(X^+)-\frac{7}{q^*}\delta\ .
\end{aligned}$$

The latter inequality together with (5) yields

$$\frac{1}{n}\sum_{l=1}^{n}\xi_T^2(l)\geq P_X(E^0)+(1-2q^*)^2P_X(E^+)+4q^*(1-q^*)P_X(X^+)-\frac{10}{q^*}\delta\ .$$

Combining this estimate with (4) we now obtain

$$\begin{aligned}
&\langle w^*,w^*\rangle\\
\geq\;& c\Bigl(P_X(E^0)+(1-2q^*)^2P_X(E^+)+4q^*(1-q^*)P_X(X^+)-4p^*(1-p^*)P_X(X^+)-\frac{37}{q^*}\delta\Bigr)\\
\geq\;& c^*\Bigl(P_X(E^0)+(1-2q^*)^2P_X(E^+)-4(p^*-q^*)P_X(X^+)-\frac{37}{q^*}\delta\Bigr)\ . \qquad (8)
\end{aligned}$$

Moreover, a simple calculation shows

$$\mathcal{R}_P(f_T^{2,k,c/n})\;=\;\mathcal{R}_P+\int\limits_E(1-2p)\,dP_X\;\leq\;\mathcal{R}_P+\frac{1}{1-2q^*}\Bigl(P_X(E^0)+(1-2q^*)^2P_X(E^+)\Bigr)$$

and thus $P_X(E^0)+(1-2q^*)^2P_X(E^+)\geq(1-2q^*)(\mathcal{R}_P(f_T^{2,k,c/n})-\mathcal{R}_P)$. With this and (8) we find

$$\langle w^*,w^*\rangle\;\geq\;\frac{2\,\langle w^*,w^*\rangle}{\varepsilon(1-2q^*)}\Bigl((1-2q^*)\bigl(\mathcal{R}_P(f_T^{2,k,c/n})-\mathcal{R}_P\bigr)-4(p^*-q^*)P_X(X^+)-\frac{37}{q^*}\delta\Bigr)\ .$$

The latter inequality finally implies

$$\begin{aligned}
\mathcal{R}_P(f_T^{2,k,c/n})-\mathcal{R}_P\;&\leq\;\frac{\varepsilon}{2}+4\frac{p^*-q^*}{1-2q^*}P_X(X^+)+\frac{37}{(1-2q^*)q^*}\,\frac{\varepsilon(1-2q^*)q^*}{74}\\
&=\;4\frac{p^*-q^*}{1-2q^*}P_X(X^+)+\varepsilon\ .
\end{aligned}$$

Thus the assertion follows. ■

We have to prove the remaining lemmas now. We begin with the lemma that constructs the partitions $\mathcal{P}_i^j$ of the above proof:

Lemma 13 *Using the notations of the proof of Theorem 12 there exist partitions $\mathcal{P}_i^j$ of K_i^j, $i \in \{-1,1\}$, $j \in \{0,+\}$, such that* $\operatorname{diam}_{d_k}(A) \le \sigma$ *for all $A \in \mathcal{P}_i^j$, $i \in \{-1,1\}$, $j \in \{0,+\}$, and*

$$\Big| \bigcup_{\substack{i\in\{-1,1\}\\ j\in\{0,+\}}} \mathcal{P}_i^j \Big| \;\le\; \mathcal{N}\big((X,d_k),\sigma\big)\ . \tag{9}$$

Proof By the definition of the covering numbers there exists a partition $\mathcal{P}$ of X with $\operatorname{diam}_{d_k}(A) \le \sigma$ for all $A \in \mathcal{P}$ and $|\mathcal{P}| \le \mathcal{N}\big((X,d_k),\sigma\big)$. Let us define $\tilde{\mathcal{P}}_i^j := \{A \in \mathcal{P} : A \cap K_i^j \neq \emptyset\}$ and $\mathcal{P}_i^j := \{A \cap K_i^j : A \in \tilde{\mathcal{P}}_i^j\}$. Therefore, to prove (9) we have to show that the $\tilde{\mathcal{P}}_i^j$'s are mutually disjoint. Assume the converse, i.e. there exists $A \in \tilde{\mathcal{P}}_i^j \cap \tilde{\mathcal{P}}_{i'}^{j'}$ with $i \neq i'$ or $j \neq j'$. By the definition of the partitions there exist $z_1, z_2 \in A$ with $z_1 \in K_i^j$ and $z_2 \in K_{i'}^{j'}$. Now on the one hand, we obtain

$$|\langle w^*, \Phi(z_1) - \Phi(z_2)\rangle| \;\le\; \|w^*\|\, d_k(z_1,z_2) \;\le\; \|w^*\|\,\sigma \;\le\; \|w^*\| \frac{\delta}{\sqrt{c^*}} \;<\; \delta$$

but on the other hand we also have

$$|\langle w^*, \Phi(z_1) - \Phi(z_2)\rangle| \;=\; |\langle w^*, \Phi(z_1)\rangle - \langle w^*, \Phi(z_2)\rangle| \;\ge\; 1 - (1-2p^*) \;\ge\; \delta^* \;\ge\; \delta\ .$$

Therefore the assertion follows. ■

Lemma 14 *Using the notations of the proof of Theorem 12 we have*

$$P^n(F_n) \ge 1 - 3Me^{-2(\frac{\delta}{M})^2 n}\ .$$

Proof Let us recall Hoeffding's inequality (cf. Devroye et al., 1997, Thm. 8.1) which in particular states that for all i.i.d. random variables $z_i : (\Omega, \mathcal{A}, Q) \to \{0,1\}$ and all $\varepsilon > 0$, $n \ge 1$ we have

$$Q^n\Big(\sum_{i=1}^n z_i \le n(q-\varepsilon)\Big) \;\le\; e^{-2\varepsilon^2 n}\ ,$$

where $q := Q(z_i = 1)$. Thus for $A \in \mathcal{P}_i^+$ we get

$$\begin{aligned}
& P^n(F_{n,A}^+) \\
\ge\;& P^n\Big(\Big\{\big((x_1,y_1),\dots,(x_n,y_n)\big) : \big|\{l : x_l \in A, y_l = i\}\big| \ge \Big(\int_A (1-p)\, dP_X - \frac{\delta}{M}\Big)\, n\Big\}\Big) \\
\ge\;& 1 - P^n\Big(\Big\{\big((x_1,y_1),\dots,(x_n,y_n)\big) : \big|\{l : x_l \in A, y_l = i\}\big| \le \Big(\int_A (1-p)\, dP_X - \frac{\delta}{M}\Big)\, n\Big\}\Big) \\
\ge\;& 1 - e^{-2(\frac{\delta}{M})^2 n}\ ,
\end{aligned}$$

i.e. $P^n(Z^n \setminus F^+_{n,A}) \leq e^{-2(\frac{\delta}{M})^2 n}$, where $Z := X \times Y$. Analogously, we obtain $P^n(Z^n \setminus F^-_{n,A}) \leq e^{-2(\frac{\delta}{M})^2 n}$ for all $A \in \mathcal{P}^+_i$ and $P^n(Z^n \setminus F_{n,A}) \leq e^{-2(\frac{\delta}{M})^2 n}$ for all $A \in \mathcal{P}^j_i$. These estimates yield

$$\begin{aligned}
& P^n(F^n) \\
= \; & P^n\Big(\bigcap_{A \in \mathcal{P}^0_{-1} \cup \mathcal{P}^0_1} F_{n,A} \quad \cap \bigcap_{A \in \mathcal{P}^+_{-1} \cup \mathcal{P}^+_1} \big(F_{n,A} \cap F^+_{n,A} \cap F^-_{n,A}\big)\Big) \\
= \; & 1 - P^n\Big(\bigcup_{A \in \mathcal{P}^0_{-1} \mathcal{P}^0_1} (Z^n \setminus F_{n,A}) \quad \cup \bigcup_{A \in \mathcal{P}^+_{-1} \cup \mathcal{P}^+_1} \big((Z^n \setminus F_{n,A}) \cup (Z^n \setminus F^+_{n,A}) \cup (Z^n \setminus F^-_{n,A})\big)\Big) \\
\geq \; & 1 - \sum_{\substack{A \in \mathcal{P}^j_{-1} \cup \mathcal{P}^j_1 \\ j \in \{0,+\}}} P^n\big(Z^n \setminus F_{n,A}\big) - \sum_{A \in \mathcal{P}^+_{-1} \cup \mathcal{P}^+_1} P^n\big(Z^n \setminus F^+_{n,A}\big) - \sum_{A \in \mathcal{P}^+_{-1} \cup \mathcal{P}^+_1} P^n\big(Z^n \setminus F^-_{n,A}\big) \\
\geq \; & 1 - 3M e^{-2(\frac{\delta}{M})^2 n} \, .
\end{aligned}$$

■

With the help of Theorem 12 we can now investigate how to choose the parameter sequence (c_n) for a given universal kernel:

Theorem 15 *Let (X, d) be a compact metric space and k be a universal kernel on X such that the covering numbers of (X, d_k) fulfill $\mathcal{N}\big((X, d_k), \varepsilon\big) \in \mathcal{O}(\varepsilon^{-\alpha})$ for some $\alpha > 0$. Suppose that we have a positive sequence (c_n) with $(c_n) \in \mathcal{O}(n^{\beta-1})$ for some $0 < \beta < \frac{1}{\alpha}$ and $nc_n \to \infty$. Then for all Borel probability measures P on $X \times Y$ with $q^*, p^* \in (0, 1/2)$ and all $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \Pr{}^*\Big(\Big\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{2,k,c_n}) \leq \mathcal{R}_P + 4\frac{p^* - q^*}{1 - 2q^*} P_X(X^+) + \varepsilon\Big\}\Big) \; = \; 1 \; .$$

Proof Let $\gamma := \frac{1 - \alpha\beta}{4(1+\alpha)} > 0$ and $\delta_n := n^{-\gamma}$. By Theorem 12 there exist $c^* > 0$ and $\delta^* > 0$ such that for all $c \geq c^*$, $0 < \delta \leq \delta^*$ and all $n \geq 1$ we have

$$\Pr{}^*\Big(\Big\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{2,k,c/n}) \leq \mathcal{R}_P + 4\frac{p^* - q^*}{1 - 2q^*} P_X(X^+) + \varepsilon\Big\}\Big) \; \geq \; 1 - 3M e^{-2(\frac{\delta}{M})^2 n} \; ,$$

where $M := \mathcal{N}\big((X, d_k), \frac{\delta}{\sqrt{c}}\big)$. Since $\delta_n \to 0$ and $nc_n \to \infty$ we may fix an $n_0 \geq 1$ such that for all $n \geq n_0$ we have both $\delta_n \leq \delta^*$ and $nc_n \geq c^*$. This yields

$$\Pr{}^*\Big(\Big\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{2,k,c_n}) \leq \mathcal{R}_P + 4\frac{p^* - q^*}{1 - 2q^*} P_X(X^+) + \varepsilon\Big\}\Big) \; \geq \; 1 - 3M_n e^{-2\left(\frac{\delta_n}{M_n}\right)^2 n} \; ,$$

where $M_n := \mathcal{N}\big((X, d_k), \frac{\delta_n}{\sqrt{nc_n}}\big)$. This implies $M_n \in \mathcal{O}(n^{\alpha(\gamma + \beta/2)})$ and hence we easily check that $M_n e^{-\left(\frac{\delta_n}{M_n}\right)^2 n} \in \mathcal{O}(e^{-n^{2(1+\alpha)\gamma}})$ holds. Thus the assertion follows. ■

Corollary 16 *Under the assumptions of Theorem 15 the 2-SMC with sequence (c_n) is consistent for all Borel probability measures P on $X \times Y$ with $q^* = p^* < 1/2$. In particular this holds for all Borel probability measures with constant noise level.*

Corollary 17 *Let $X \subset \mathbb{R}^d$ be compact and k be a Gaussian RBF kernel on X. Let $0 < \beta < \frac{1}{d}$ and (c_n) be a positive sequence with $(c_n) \in \mathcal{O}(n^{\beta-1})$ and $nc_n \to \infty$. Then the 2-SMC with kernel k and sequence (c_n) is consistent for all Borel probability measures P on $X \times Y$ with $q^* = p^* < 1/2$.*

Proof Let $\sigma > 0$ and $k(x,y) := \exp(-\sigma^2 \|x-y\|_2^2)$. Since $1 - e^{-t} \leq t$ for all $t \geq 0$ we observe

$$d_k(x,y) \;=\; \sqrt{2 - 2\exp(-\sigma^2\|x-y\|_2^2)} \;\leq\; \sqrt{2}\sigma\|x-y\|_2 \; .$$

This yields $\mathcal{N}\big((X,d_k),\varepsilon\big) \leq \mathcal{N}\big((X,\|.\|_2),\frac{\varepsilon}{\sqrt{2}\sigma}\big)$ and thus $\mathcal{N}\big((X,d_k),\varepsilon\big) \in \mathcal{O}(\varepsilon^{-d})$ (cf. Carl & Stephani, 1990, p. 9). ■

For the classification problems we have considered up to now we usually may not expect that we obtain a large margin for sample sizes growing to infinity. In the following we restrict ourselves to distributions that guarantee a fixed and strictly positive margin for all training sets. Of course, these classification problems must have a deterministic supervisor, i.e. a noise level that vanishes (almost) everywhere. In general, additional assumptions are required. Using universal kernel these reduce to a simple geometric condition:

Theorem 18 *Let (X,d) be a compact metric space and k be a universal kernel on X. Suppose that we have a Borel probability measure P on $X \times Y$ with a deterministic supervisor and with classes $B_{-1}(P), B_1(P)$ which have strictly positive distance, i.e. $d\big(B_{-1}(P), B_1(P)\big) > 0$. Then k separates $B_{-1}(P)$ and $B_1(P)$ with margin $\gamma > 0$ and for all $c > 0$, $\varepsilon > 0$ and $n \geq mM$ we have*

$$\Pr{}^*\big(\{T \in (X \times Y)^n : \mathcal{R}_{P,S}(f_T^{2,k,c}) \leq \varepsilon\}\big) \;\geq\; 1 - M\; e^{-\frac{\varepsilon n}{2M} + m} \; .$$

Here, $M := \mathcal{N}\big((X,d_k),\gamma/2\big)$ is the covering number of (X,d_k) and $m := \big\lfloor \frac{4}{c\gamma^2} \big\rfloor + 1$.

Proof Since (X,d_k) is precompact and $d\big(B_{-1}(P), B_1(P)\big) > 0$ both $B_i(P)$ are compact, too. Thus they can be separated with margin $\gamma > 0$ by Corollary 6. In analogue to Lemma 13 we now construct partitions $\tilde{\mathcal{P}}_i$ of $B_i(P)$, $i \in \{-1,1\}$ such that $\operatorname{diam}_{d_k}(A) \leq \gamma/2$ for all $A \in \tilde{\mathcal{P}}_i$ and $\big|\tilde{\mathcal{P}}_{-1} \cup \tilde{\mathcal{P}}_1\big| \leq M$. We define $\mathcal{P}_i := \{A \in \tilde{\mathcal{P}}_i : P_X(A) > \frac{\varepsilon}{M}\}$ for $i \in \{-1,1\}$. Moreover, for $n \geq mM$ and $A \in \mathcal{P}_{-1} \cup \mathcal{P}_1$ let

$$\begin{aligned} F_{n,A} &:= \Big\{\big((x_1,y_1),\dots,(x_n,y_n)\big) \in (X \times Y)^n \;:\; \big|\{l : x_l \in A\}\big| \geq m\Big\} \\ F_n &:= \bigcap_{A \in \mathcal{P}_{-1} \cup \mathcal{P}_1} F_{n,A} \; . \end{aligned}$$

For $n \geq mM$ the Chernoff-Okamoto inequality (see e.g. Dudley, 1978) then yields

$$\begin{aligned} P^n\big((X \times Y)^n \setminus F_{n,A}\big) &\leq \exp\Big(-\frac{\big(n\frac{\varepsilon}{M} - (m-1)\big)^2}{2n\frac{\varepsilon}{M}\big(1-\frac{\varepsilon}{M}\big)}\Big) \\ &\leq \exp\Big(-\frac{n^2\big(\frac{\varepsilon}{M}\big)^2 - 2n\frac{\varepsilon}{M}(m-1) + (m-1)^2}{2n\frac{\varepsilon}{M}}\Big) \\ &\leq \exp\big(-\frac{\varepsilon n}{2M} + m\big) \end{aligned}$$

and thus $P^n(F_n) \geq 1 - M\ e^{-\frac{\varepsilon n}{2M}+m}$ for all $n \geq mM$. Hence it suffices to show that for all $T = \big((x_1,y_1),\dots,(x_n,y_n)\big) \in F_n$ the decision function $f_T^{2,k,c}$ classifies $\bigcup_{A\in\mathcal{P}_1} A$ and $\bigcup_{A\in\mathcal{P}_{-1}} A$ correctly. To see this we fix a feature map $\Phi : X \to H$ of k. For brevity's sake the unique solution of problem (3) is denoted by (w_T, b_T, ξ_T). Furthermore, $(w^*, b^*) \in S_H \times \mathbb{R}$ is a hyperplane that separates $\Phi\big(B_1(P)\big)$ and $\Phi\big(B_{-1}(P)\big)$ with margin $\gamma > 0$. Then we have $y_l\big(\langle w^*, \Phi(x_l)\rangle + b^*\big) \geq \gamma$ for all $l = 1,\dots,n$ and therefore we obtain

$$\langle w_T, w_T\rangle + c\sum_{l=1}^{n} \xi_l^2 \ \leq\ \Big\langle \frac{w^*}{\gamma}, \frac{w^*}{\gamma}\Big\rangle \ =\ \frac{1}{\gamma^2}\ .$$

In particular this implies $\|w_T\| \leq 1/\gamma$. Now, let us suppose that there exists a misclassified point z. Without loss of generality we may assume that $z \in \bigcup_{A\in\mathcal{P}_1} A$. Hence there is an $A \in \mathcal{P}_1$ with $z \in A$ and for this there exist mutually different indexes $l_1,\dots,l_m$ such that $x_{l_j} \in A$. In particular this yields $d_k(x_{l_j}, z) \leq \gamma/2$ for all $j = 1,\dots,m$. Since $\langle w_T, \Phi(z)\rangle + b_T \leq 0$ we thus obtain

$$1-\xi_{l_j} \ \leq\ \langle w_T, \Phi(x_{l_j})\rangle + b_T \ =\ \langle w_T, \Phi(x_{l_j}) - \Phi(z)\rangle + \langle w_T, \Phi(z)\rangle + b_T \ \leq\ \|w_T\|\, d_k(x_{l_j}, z) \ \leq\ \frac{1}{2}.$$

Hence we have $\xi_{l_j} \geq 1/2$ and this leads to the contradiction

$$\frac{1}{\gamma^2} \ \geq\ \langle w_T, w_T\rangle + c\sum_{l=1}^{n}\xi_l^2 \ \geq\ c\sum_{j=1}^{m}\xi_{l_j}^2 \ \geq\ \frac{cm}{4} \ =\ \frac{c}{4}\Big(\Big\lfloor\frac{4}{c\gamma^2}\Big\rfloor + 1\Big) \ >\ \frac{1}{\gamma^2}\ .$$

Therefore the assertion follows. ■

Theorem 18 shows that the 2-SMC works well with fixed weight factor c whenever it treats a classification problem that ensures a large margin. We believe that these distributions are also the only ones for which a fixed c is suitable. Our conjecture is based on the observation that the constant c controls the Lipschitz constant of the solution of (3) with respect to the metric d_k: if we have a classification problem that does not guarantee a large margin the Lipschitz constant may grow like n. The proofs of this section indicate that this may be too fast since for large sample sizes the solution need not be "almost" constant on each element of the partitions, i.e. overfitting may occur.

In the proof of the above theorem we only used elements of the partitions $\mathcal{P}_i$ whose probability was larger than or equal to $\frac{\varepsilon}{M}$. If we extend our considerations to all elements with strictly positive probability we obtain the following theorem:

Theorem 19 *Let (X, d) be a compact metric space and k a universal kernel on X. Suppose that we have a Borel probability measure P on $X \times Y$ with a deterministic supervisor and with classes $B_{-1}(P), B_1(P)$ which have strictly positive distance, i.e. $d\big(B_{-1}(P), B_1(P)\big) > 0$. Then k separates $B_{-1}(P)$ and $B_1(P)$ with margin $\gamma > 0$ and for all $c > 0$ there exists a constant $\rho > 0$ such that for all $n \geq \big(\lfloor \frac{4}{c\gamma^2} \rfloor + 1\big)\mathcal{N}\big((X, d_k), \gamma/2\big)$ we have*

$$\Pr{}^*\big(\{T \in (X \times Y)^n : \mathcal{R}_{P,S}(f_T^{2,k,c}) = 0\}\big) \ \geq \ 1 - e^{-\rho n} \ ,$$

Up to now we have only treated universal kernels. One may ask whether other classes of kernels are also suitable to treat with the classification problems considered in this work. One type often used are polynomial kernels of the form $(\langle .,. \rangle + c)^m$, $c \geq 0$, $m \in \mathbb{N}$, on a subset X of $\mathbb{R}^n$. For these kernels the functions generated by the 2-SMC are polynomials on X of degree up to m. Thus the next proposition shows that these kernels are not even capable to solve the problems of Theorems 18 and 19:

Proposition 20 *Let $\mathcal{P}_n^d$ be the set of all polynomials on $X := [0,1]^d$ whose degree is less than $n+1$. Then for all $\varepsilon > 0$ there exists a Borel probability measure P on $X \times Y$ with $d\big(B_1(P), B_{-1}(P)\big) > 0$, $\mathcal{R}_P = 0$ and*

$$\inf\{\mathcal{R}_P(\operatorname{sign} f) \ : \ f \in \mathcal{P}_n^d\} \ \geq \ \frac{1}{2} - \varepsilon \ .$$

Proof We first treat the case $d = 1$. We fix an integer $m \geq (3n+2)/\varepsilon$ and let $I_i := [(i+\varepsilon)/m, (i+1-\varepsilon)/m]$, $i = 0, \dots, m-1$. Denoting the Lebesgue measure on I_i by λ_{I_i} let $P_X := (1-2\varepsilon)^{-1} \sum_{i=0}^{m-1} \lambda_{I_i}$. Moreover, we define a deterministic supervisor $P(.|.)$ by $P(y=1|x) := 1$ for $x \in I_{2i}$, $i = 0, \dots, \lfloor (m+1)/2 \rfloor$, and $P(y=-1|x) := 1$ otherwise. For a fixed polynomial $f \in \mathcal{P}_n^d$ we denote its mutually different and ordered zeroes in $(0,1)$ by $x_1 < \dots < x_k$, $k \leq n$. For brevity's sake let $x_0 := 0$ and $x_{k+1} := 1$. Finally, we define $a_j := \big|\{i : I_i \subset [x_j, x_{j+1}]\}\big|$ for $j = 0, \dots, k$. Then by an easy observation we get $\sum_{j=0}^k a_j \ \geq \ m-k \ \geq \ m-n$. Moreover, at most $\lfloor (a_j+1)/2 \rfloor$ intervals I_i are correctly classified on $[x_j, x_{j+1}]$ by the function $\operatorname{sign} f$. Hence at least

$$\sum_{j=0}^{k} \Big\lfloor \frac{a_j}{2} \Big\rfloor \geq \sum_{j=0}^{k} \Big(\frac{a_j}{2} - 1 \Big) \geq \frac{m-n}{2} - (k+1) \geq \frac{m-3n-2}{2}$$

intervals I_i are not correctly classified on $[0,1]$ by $\operatorname{sign} f$. Since $P_X(I_i) = 1/m$ we thus obtain

$$\mathcal{R}_P(\operatorname{sign} f) \ \geq \ \frac{1}{m} \sum_{j=0}^{k} \Big\lfloor \frac{a_j}{2} \Big\rfloor \ \geq \ \frac{1}{2} - \frac{3n+2}{m} \ \geq \ \frac{1}{2} - \varepsilon \ .$$

To treat the case $d > 1$ we have to embed $[0,1]$ into $[0,1]^d$ via $t \mapsto (t, 0, \dots, 0)$. Considering the above distribution P embedded into $[0,1]^d$ we then get the assertion since polynomials in d variables on $[0,1]$ embedded into $[0,1]^d$ as above are essentially polynomials in one variable. ■

5. The 1-norm soft margin classifier

We now consider the 1-norm soft margin classifier. Again we fix a kernel k on X with feature map $\Phi : X \to H$. For a training set $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ and $c_n > 0$ we denote a solution of the optimization problem

$$\begin{array}{lll} \text{minimize} & \langle w, w\rangle + c_n \sum\limits_{i=1}^{n} \xi_i & \text{over } w, b, \xi \\ \text{subject to} & y_i(\langle w, \Phi(x_i)\rangle + b) \geq 1 - \xi_i, & i = 1, \dots, n \\ & \xi_i \geq 0, & i = 1, \dots, n \end{array} \tag{10}$$

by $(w_T^{1,k,c_n}, b_T^{1,k,c_n}) \in H \times \mathbb{R}$. An algorithm $\mathcal{C}_k^{1,(c_n)}$ that provides the decision function

$$f_T^{1,k,c}(x) := \operatorname{sign}(\langle w_T^{1,k,c_n}, \Phi(x)\rangle + b_T^{1,k,c_n}) \ , \qquad x \in X$$

for every training set T is called a 1-norm soft margin classifier (1-SMC) with kernel k and parameter sequence (c_n). For an introduction to the 1-SMC as well as for implementation techniques we refer to Cristianini & Shawe-Taylor (2000, Ch. 6 and 7) and Vapnik (1998, Ch. 10)

Burges & Crisp (2000) proved that in general the optimization problem (10) has no unique solution. Although the 1-SMC only has to construct an arbitrary *solution* we show in this section that it enjoys all the properties of the 2-SMC proven in this work. We begin with a statement which is analogous to Theorem 12:

Theorem 21 *Let (X, d) be a compact metric space and k be a universal kernel on X. Then for all Borel probability measures P on $X \times Y$ with $q^*, p^* \in (0, 1/2)$ and all $\varepsilon > 0$ there exist $c^* > 0$ and $\delta^* > 0$ such that for all $c \geq c^*$, $0 < \delta \leq \delta^*$ and all $n \geq 1$ we have*

$$\Pr{}^*\Big(\Big\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{1,k,c/n}) \leq \mathcal{R}_P + 4\frac{p^* - q^*}{1 - 2q^*} P_X(X^+) + \varepsilon\Big\}\Big) \ \geq \ 1 - 3Me^{-2\left(\frac{\delta}{M}\right)^2 n} \ ,$$

where $M := 4\mathcal{N}\big((X, d_k), \frac{\delta}{\sqrt{c}}\big)$ is essentially the covering number of X with respect to the metric d_k which is induced by the kernel k.

Sketch of the proof Since the proof is very similar to that of Theorem 12 we only point out the main differences besides adjusting the constants: Firstly the vector $w^* \in H$ has to be chosen in a different manner, namely it has to fulfill

$$\langle w^*, \Phi(x)\rangle \quad \in \quad \begin{cases} i\,[1, 1 + \delta] & \text{if } x \in K_i^j \\ [-(1 + \delta), 1 + \delta] & \text{otherwise.} \end{cases}$$

This condition also enforces the second modification since Lemma 13 does not hold anymore in this setting. Indeed, we cannot guarantee that the definition of the $\tilde{\mathcal{P}}_i^j$'s in Lemma 13 implies that they are mutually disjoint. Instead, we only obtain $|\tilde{\mathcal{P}}_i^j| \leq N\big((X, d_k), \frac{\delta}{\sqrt{c}}\big)$ and thus

$$\Big| \bigcup_{\substack{i \in \{-1,1\} \\ j \in \{0,+\}}} \mathcal{P}_i^j \Big| \ \leq \ 4\mathcal{N}\big((X, d_k), \sigma\big) \ .$$

The latter explains the definition of M, which is different to that of Theorem 12. ■

Following the proof of Theorem 15 we obtain an analogous result for the 1-SMC:

Theorem 22 *Let (X, d_k) be a compact metric space and k be a universal kernel on X such that the covering numbers of (X, d_k) fulfill $\mathcal{N}\big((X, d_k), \varepsilon\big) \in \mathcal{O}(\varepsilon^{-\alpha})$ for some $\alpha > 0$. Suppose that we have a positive sequence (c_n) with $(c_n) \in \mathcal{O}(n^{\beta-1})$ for some $0 < \beta < \frac{1}{\alpha}$ and $nc_n \to \infty$. Then for all Borel probability measures P on $X \times Y$ with $q^*, p^* \in (0, 1/2)$ and all $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \Pr^*\Big(\Big\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{1,k,c_n}) \leq \mathcal{R}_P + 4\frac{p^* - q^*}{1 - 2q^*} P_X(X^+) + \varepsilon\Big\}\Big) \;=\; 1$$

To complete our considerations of noisy problems we mention that for the 1-SMC with Gaussian RBF kernel we obtain the following consistency result which has already been proved for the 2-SMC:

Corollary 23 *Let $X \subset \mathbb{R}^d$ be compact and k be a Gaussian RBF kernel on X. Let $0 < \beta < \frac{1}{d}$ and (c_n) be a positive sequence with $(c_n) \in \mathcal{O}(n^{\beta-1})$ and $nc_n \to \infty$. Then the 1-SMC with kernel k and sequence (c_n) is consistent for all Borel probability measures P on $X \times Y$ with $q^* = p^* < 1/2$.*

In the presence of large margins it turns out that we may fix the weight factor analogously to the 2-SMC. For brevity's sake we only state a result that is similar to Theorem 18. However, a result that is analogous to Theorem 19 also holds.

Theorem 24 *Let (X, d) be a compact metric space and k a universal kernel on X. Suppose that we have a Borel probability measure P on $X \times Y$ with a deterministic supervisor and with classes $B_{-1}(P), B_1(P)$ which have strictly positive distance, i.e. $d\big(B_{-1}(P), B_1(P)\big) > 0$. Then k separates $B_{-1}(P)$ and $B_1(P)$ with margin $\gamma > 0$ and for all $c > 0$, $\varepsilon > 0$ and $n \geq mM$ we have*

$$\Pr^*\big(\{T \in (X \times Y)^n : \mathcal{R}_{P,S}(f_T^{1,k,c}) \leq \varepsilon\}\big) \;\geq\; 1 - M\; e^{-\frac{\varepsilon n}{2M} + m}\;,$$

where $M := \mathcal{N}\big((X, d_k), \gamma/2\big)$ is the covering number of (X, d_k) and $m := \big\lfloor \frac{2}{c\gamma^2} \big\rfloor + 1$.

Sketch of the proof The proof is completely analogous to that of Theorem 18. However, since the slack variables are not squared we obtain

$$c \sum_{j=1}^{m} \xi_{l_j} \;\geq\; \frac{cm}{2}$$

in the last estimate of the proof of Theorem 18 and this yields the slightly better value of m. ■

Finally we mention that using polynomial kernels Proposition 20 can also be applied. Thus problems that cannot be treated well with a fixed polynomial kernel and the 2-SMC cannot be treated well with the 1-SMC and the same kernel, either (and vice versa).

6. The maximal margin hyperplane classifier

We finally consider the maximal margin classifier. Again we fix a kernel k on X with feature map $\Phi : X \to H$. For a training set $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ we denote the unique solution of the optimization problem

$$\begin{array}{lll} \text{minimize} & \langle w, w\rangle & \text{over } w, b \\ \text{subject to} & y_i(\langle w, \Phi(x_i)\rangle + b) \geq 1, & i = 1, \dots, n \end{array} \tag{11}$$

by $(w_T^k, b_T^k) \in H \times \mathbb{R}$. An algorithm $\mathcal{C}_k$ that provides the decision function

$$f_T^k(x) := \operatorname{sign}(\langle w_T^k, \Phi(x)\rangle + b_T^k) \ , \qquad\qquad x \in X$$

for every training set T is called a maximal margin classifier (MMC) with kernel k. For an introduction to the MMC as well as for implementation techniques and a geometric motivation we again refer to Cristianini & Shawe-Taylor (2000, Ch. 6 and 7) and Vapnik (1998, Ch. 10).

The MMC is assumed to work poorly in the absence of large margins (cf. Tong Zhang, 2001). Thus we only consider the setting of Theorem 18. We begin with a result similar to Theorem 18 and Theorem 24:

Theorem 25 *Let (X, d) be a compact metric space and k a universal kernel on X. Suppose that we have a Borel probability measure P on $X \times Y$ with a deterministic supervisor and with classes $B_{-1}(P), B_1(P)$ which have strictly positive distance, i.e. $d\big(B_{-1}(P), B_1(P)\big) > 0$. Then k separates $B_{-1}(P)$ and $B_1(P)$ with margin $\gamma > 0$ and for all $\varepsilon > 0$ and $n \geq M := \mathcal{N}\big((X, d_k), \gamma/2\big)$ we have*

$$\Pr{}^*\big(\{T \in (X \times Y)^n : \mathcal{R}_{P,S}(f_T^k) \leq \varepsilon\}\big) \ \geq \ 1 - M \ e^{n \ln\left(1 - \frac{\varepsilon}{2M}\right)} \ .$$

Sketch of the proof We repeat the construction of the proof of Theorem 18 with $m := 1$. An easy calculation then shows

$$P^n(F_n) \ \geq \ 1 - M \ e^{n \ln\left(1 - \frac{\varepsilon}{2M}\right)} \ .$$

Now suppose that for $T \in F_n$ we have an element $z \in \bigcup_{A \in \mathcal{P}_1} A$ misclassified by f_T^k. Hence there exist an $A \in \mathcal{P}_1$ with $z \in A$ and a sample (x_l, y_l) of T with $x_l \in A$. Then the following estimate yields a contradiction:

$$0 \ \geq \ \langle w_T^k, \Phi(z)\rangle + b_T^k \ = \ \langle w_T^k, \Phi(z) - \Phi(x_{l_j})\rangle + \langle w_T^k, \Phi(x_{l_j})\rangle + b_T^k \ \geq \ \|w_T^k\| \ d_k(x_{l_j}, z) + 1 \ \geq \ \frac{1}{2} \ .$$

Therefore the assertion follows. ■

The above result and its proof indicate that in the presence of large margins it may be more suitable to use the MMC instead of a soft margin algorithm. We mention that an estimate which is very similar to Theorem 25 can be obtained using data-dependent margin-based bounds of Shawe-Taylor et al. (1998). To compare both we first state a corollary:

Corollary 26 *Let k_σ be the Gaussian RBF kernel with parameter σ on the unit ball $X := B_{\ell_2^d}$ of the d-dimensional Euclidean space. Let P be a Borel probability measure on $X \times Y$ which can be separated by k_σ with margin $\gamma > 0$. Then for all $\delta \in (0,1)$ and all $n \geq 2\big(\frac{16\sigma}{\gamma}\big)^d$ we have*

$$\Pr{}^*\Big(\Big\{T \in (X \times Y)^n : \mathcal{R}_{P,S}(f_T^k) \leq \frac{4d}{n}\Big(\frac{16\sigma}{\gamma}\Big)^d\Big(d \ln \frac{16\sigma}{\gamma} + \ln \frac{2}{\delta}\Big)\Big\}\Big) \;\geq\; 1 - \delta \;.$$

Proof As already shown in the proof of Corollary 17 we have

$$\mathcal{N}\big((X, d_k), \varepsilon\big) \;\leq\; \mathcal{N}\Big((X, \|.\|_2), \frac{\varepsilon}{\sqrt{2}\sigma}\Big) \;\leq\; 2 \cdot 5^d \Big(\frac{\varepsilon}{\sqrt{2}\sigma}\Big)^{-d} \;\leq\; 2(8\sigma)^d \varepsilon^{-d}$$

and thus $M := \mathcal{N}\big((X, d_k), \gamma/2\big) \leq 2(16\sigma)^d\gamma^{-d}$. Now let $\varepsilon := 2M(1 - \big(\frac{\delta}{M}\big)^{1/n})$ for $n \geq 2(16\sigma)^d\gamma^{-d}$, i.e. $\delta = M\big(1 - \frac{\varepsilon}{2M}\big)^n$. Since $\varepsilon < \frac{2M}{n} \ln \frac{M}{\delta}$ Theorem 25 yields the assertion. ■

Corollary 26 shows that the learning curve of the MMC is of order $\mathcal{O}(n^{-1})$ provided that P guarantees a large margin. These conditions also allow the application of margin-based bounds on generalization proved by Shawe-Taylor et al. (1998) (cf. also Bartlett & Shawe-Taylor, 1999; Cristianini & Shawe-Taylor, 2000). We then obtain that the learning curve is of order $\mathcal{O}(n^{-1}(\log n)^2)$. However, to compare both results one also has to consider the constants that arise since the sample size n only varies in a typical range. Here we observe that the influence of the margin γ is essentially of order $\mathcal{O}(\gamma^{-2})$ in the estimates of Shawe-Taylor et al. (1998) while the Corollary 26 shows that our estimates are essentially influenced by order $\mathcal{O}(\gamma^{-d})$, where d is the dimension of the *input* space X. We thus suppose that only for small input dimensions d our estimates are more suitable to treat realistic sample sizes.

7. Conclusions

The aim of this paper has been to investigate which kind of distributions could be classified well by support vector machines (SVM's). It has turned out that the ability of the kernel to approximate arbitrary continuous functions plays a fundamental role for this question. Since the resulting function classes represented by the classifier are very large there always exists the risk of *overfitting.* However, using soft margin support vector machines with specific sequences of regularization parameters this risk can be controlled at least for simple noise models, e.g. models with constant noise level. In particular, the restriction to large margin problems in Tong Zhang (2001, p. 442) has been significantly weakened.

Since the ansatz of this paper is new many questions remain open, and are worth for further investigations. Firstly it is interesting whether the soft margin algorithms yield arbitrarily good generalization for *all distributions.* Up to now our results only provide consistency if the noise level is constant. However, approximating an arbitrary noise level by step functions it seems possible that the soft margin algorithms with certain parameter sequences are universally consistent. Of course, the universality of kernels, which roughly speaking enables us to do “almost everything” with induced functions on compact subsets would play a crucial role for this ansatz, too. However, a solution is certainly also based on a deeper investigation of the underlying optimization problems in order to remove the restriction $q^*, p^* \in (0, 1/2)$. Moreover, we suppose that the universality of kernels is also very important for related algorithms such as ν-SVM's, linear programming SVM's and leave-one-out SVM's as well as all kind of SVM's used for regression, distribution estimation, etc.. However, since each treatment of one of these algorithms needs its own investigation of the corresponding optimization problem many open questions remain.

Remark *Using the techniques of this work the author was recently able to prove universal consistency for the 1-SMC (cf. Steinwart, 2001b).*

Acknowledgments

We would like to acknowledge support for this project from the Deutsche Forschungsgemeinschaft (DFG grant Ca 179/4-1).

References

P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 43-54, Cambridge, MA, 1999. MIT Press.

C.J.C. Burges and D.J. Crisp. Uniqueness of the SVM solution. In M.I. Jordan, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing 12*, pages 223-229, Cambridge, MA, 2000. MIT Press.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, 2000.

B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators.* Cambridge University Press, Cambridge, 1990.

L. Devroye, L. Györfi and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer, New York, 1997.

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

R.M. Dudley. *Real Analysis and Probability.* Chapman & Hall, New York, 1989.

S. Gradstein and I.M. Ryshik. *Summen- Produkt- und Integraltafeln * Tables of series, products and integrals.* Verlag Harri Deutsch, Frankfurt am Main, 1981.

G.K. Pedersen. *Analysis Now.* Springer, Berlin, 1988.

T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics,* 19(4):201-209, 1975.

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, Egham, UK. Available electronically at `http://www.dcs.rhbnc.ac.uk/research/compint/areas/comp_learn/sv/pub/report98-03.ps`, 1998.

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

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory* 44(5):1926-1940, 1998.

I. Steinwart. On the influence of the kernel on the generalization ability of support vector machines. Technical Report 01-01, Fakultät für Mathematik und Informatik, Friedrich-Schiller-Universität, Jena, Germany, available electronically at `http://www.minet.uni-jena.de/Math-Net/reports/sources/2001/01-01report.ps`, 2001a.

I. Steinwart. On the generalization ability of support vector machines. Submitted to *J. Complexity*, 2001b.

Tong Zhang. A leave-one-out cross validation bound for kernel methods with applications in learning. In *Proceedings of the 14th Annual Conference on Computational Learning Theory, Lecture Notes in Artificial Intelligence 2111*, pages 427-443, 2001.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer, New York, 1996.

V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.