JOURNAL OF COMPLEXITY **18**, 768–791 (2002)
doi:10.1006/jcom.2002.0642

Support Vector Machines Are Universally Consistent

Ingo Steinwart[1]

Mathematisches Institut, Friedrich-Schiller-Universität, 07743 Jena, Germany
E-mail: steinwart@minet.unj-jena.de

Received June 15, 2001; revised December 12, 2001; accepted January 5, 2002;
published online July 2, 2002

We show that support vector machines of the 1-norm soft margin type are universally consistent provided that the regularization parameter is chosen in a distinct manner and the kernel belongs to a specific class—the so-called universal kernels—which has recently been considered by the author. In particular it is shown that the 1-norm soft margin classifier with Gaussian RBF kernel on a compact subset X of $\mathbb{R}^d$ and regularization parameter $c_n = n^{\beta-1}$ is universally consistent, if n is the training set size and $0 < \beta < 1/d$. © 2002 Elsevier Science (USA)

1. INTRODUCTION AND RESULTS

In recent years support vector machines (SVMs) have been successfully applied to many learning problems and they mostly outperformed neural networks. Even though their development was motivated by results from statistical learning theory the known bounds of their generalization performance are not fully satisfactory. In particular, it is open whether the support vector approach can yield sufficiently good results for *all* classification problems, or whether it only works fine for "benign" distributions. The aim of this work is to answer this question for the 1-norm soft margin classifier (1-SMC) equipped with several standard kernels like the Gaussian radial basis function (RBF) kernel.

Let us start with a description of the problem of pattern recognition or classification (cf. also [10, Chap. 1; 5, Chaps. 1 and 4]): assume that we have a set X which is split into two disjoint and *unknown* classes X_{-1} and X_1, i.e., $X = X_{-1} \cup X_1$. Obviously, these classes can be encoded by a function $f : X \to Y := \{-1, 1\}$ with $f^{-1}(\{-1\}) = X_{-1}$ and $f^{-1}(\{1\}) = X_1$. The classification task is to estimate f on the basis of finitely many *training samples*

$$((x_1, y_1), \ldots, (x_n, y_n)) \in X \times Y.$$

[1] Research was supported by DFG Grant Ca 179/4-1.

0885-064X/02 $35.00

Here, the ith *label* y_i contains information on the class membership of the point x_i. Note, that estimating f corresponds to reconstructing the classes X_{-1} and X_1 on the basis of the samples.

A typical example of a classification problem is to recognize handwritten letters by an algorithm that has seen some examples of these letters.

In the framework of statistical learning theory it is usually assumed that the training samples are drawn i.i.d. according to an unknown probability measure P on $X \times Y$. To simplify our considerations let us suppose in the following that X is a compact subset of $\mathbb{R}^d$ and P is a Borel probability measure. By disintegration (cf. [7, Lemma 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*. Since in this model the labels y_i are drawn according to the conditional probability $P(.|x_i)$ we may only expect noisy information, i.e., some of our labels may be incorrect. However, the noiseless case $P(.|x) \in \{0, 1\}$ for all $x \in X$ which is usually called agnostic learning model is also covered in this setting.

A classifier $\mathscr{C}$ is an algorithm that constructs to every training set $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ a (measurable) *decision function* $f_T : X \to Y$. Besides support vector machines which we shall introduce later on typical examples of classifiers are the nearest-neighbor algorithm and neural networks. In order to "learn" from the samples the decision function $f_T : X \to Y$ should guarantee a small probability for the misclassification of an example (x, y) generated with distribution P independently to T. Here, misclassification means $f(x) \neq y$. To make this precise we define the risk of f_T by

$$\mathscr{R}_P(f_T) := \int_{X \times Y} \mathbf{1}_{\{f_T(x) \neq y\}} P(dx, dy) = P(\{(x, y) : f_T(x) \neq y\}).$$

When considering noisy supervisors we cannot expect that we obtain zero risk. Indeed, let us define

$$B_1(P) := \{x \in X : P(y = 1|x) > P(y = -1|x)\},$$

$$B_{-1}(P) := \{x \in X : P(y = 1|x) < P(y = -1|x)\},$$

$$B_0(P) := \{x \in X : P(y = 1|x) = P(y = -1|x)\}.$$

Then for a function $f^*: X \to \{-1, 1\}$ with $f^*(x) = 1$ if $x \in B_1(P)$ and $f^*(x) = -1$ if $x \in B_{-1}(P)$ we have (cf. [6, Theorem 2.1])

$$\mathscr{R}_P(f^*) = \inf\{\mathscr{R}_P(f) : f : X \to \{-1, 1\} \text{ measurable}\} = \int_X s(x) P_X(dx), \qquad (1)$$

where the *noise level* $s: X \to \mathbb{R}$ is defined by $s(x) := P(y = -1|x)$ for $x \in B_1(P)$, $s(x) := P(y = 1|x)$ for $x \in B_{-1}(P)$ and $s(x) = \frac{1}{2}$ otherwise. Equation (1) shows that no function can yield less risk than f^*. The function f^* is called an *optimal Bayes decision rule* and we write $\mathscr{R}_P := \mathscr{R}_P(f^*)$ for the *Bayes risk*. Trying to obtain a risk close to $\mathscr{R}_P$ corresponds to reconstructing the classes $B_{-1}(P)$ and $B_1(P)$ in probability. Indeed, an easy computation similar to that of Lemma 4 yields

$$\mathscr{R}_P(f_T) = \mathscr{R}_P + \int_E (1 - 2s)\, dP_X,$$

where $E := \{x \in B_{-1}(P) : f_T(x) = 1\} \cup \{x \in B_1(P) : f_T(x) = -1\}$. Thus, the risk of f_T is close to $\mathscr{R}_P$ if and only if $P_X(E)$ is small. Note that in the described model, we try to imitate the supervisors response in order to recognize the underlying classes. In particular, we trust the supervisor in the sense that even though some labels may be incorrect we assume that for large sample sizes the supervisor tends to give more correct than incorrect information for every $x \in X$. It is doubtful whether one can learn without this assumption.

As indicated above, a classifier $\mathscr{C}$ should guarantee with high probability that $\mathscr{R}_P(f_T)$ is close to $\mathscr{R}_P$ provided that T is large enough. Asymptotically, this means that

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

should hold in probability if $|T| \to \infty$. In this case the algorithm $\mathscr{C}$ is called *consistent* for the distribution P (cf. [6, Definition 6.1]). If a classifier is consistent for all distributions on $X \times Y$ it is said to be *universally consistent*. Although several algorithms such as the k-nearest-neighbor classifier for $k \to \infty$ and $k/|T| \to 0$ are universally consistent (cf. [6, Theorem 6.4]) it is an open question whether SVMs are universally consistent for a particular choice of the free parameters. Having proved universal consistency for a classifier $\mathscr{C}$ does not guarantee that $\mathscr{C}$ works well for a specific classification task. Actually, for every classifier and every decreasing null sequence $(a_n) \subset (0, \frac{1}{16}]$ there exists a distribution P with $\mathscr{R}_P = 0$ and

$$\mathbb{E}_{P^n} \mathscr{R}_P(f_{((x_1,y_1),\ldots,(x_n,y_n))}) \geq a_n$$

for all $n \geq 1$ (cf. [6, Theorem 7.2]). From this one easily deduces that for no classifier there exists a positive, increasing and unbounded sequence (α_n) and a real number $p > 0$ such that

$$P^n(T \in (X \times Y)^n : |\mathcal{R}_P(f_T) - \mathcal{R}_P| \geq \varepsilon) \leq e^{-c\varepsilon^p \alpha_n} \tag{2}$$

holds for all distributions P on $X \times Y$ and all $n \geq 1$ even if $c > 0$ depends on P. In particular, this shows that for no classifier there exists a uniform rate of convergence. Thus, every study on the rate of convergence of a specific classifier must restrict the class of considered distributions. For examples that demonstrate that these restrictions are severe we refer to [6, Chap. 7].

The ansatz of support vector machines is based on the *generalized portrait algorithm* of [11] which we briefly describe now: suppose that we have a linearly separable training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$, i.e., there exists an element $w \in S_{\ell_2^d} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and a real number $b \in \mathbb{R}$ with

$$\langle w, x_i \rangle + b > 0 \qquad \text{for all } i \text{ with } y_i = 1,$$

$$\langle w, x_i \rangle + b < 0 \qquad \text{for all } i \text{ with } y_i = -1.$$

Geometrically, this means that T can be correctly separated by the affine linear hyperplane that is described by w and b. Now, the generalized portrait algorithm constructs the correctly separating hyperplane (w_T, b_T) that has maximal distance to the training points. The resulting decision function is defined by

$$f_T(x) := \operatorname{sign}(\langle w_T, x \rangle + b_T) \qquad \text{for all } x \in X. \tag{3}$$

An easy calculation (cf. [5, Chap. 6]) shows that up to normalization (w_T, b_T) is the unique solution of the optimization problem

$$\begin{array}{lll} \text{minimize} & \langle w, w \rangle & \text{over } w, b \\ \text{subject to} & y_i(\langle w, x_i \rangle + b) \geq 1 & i = 1, \ldots, n. \end{array} \tag{4}$$

Obviously, this ansatz has two shortcomings: firstly, a linear decision function may be unsuitable to distinct between the classes $B_{-1}(P)$ and $B_1(P)$. In particular, training sets may occur that are not linearly separable and thus (w_T, b_T) may not exist. Secondly, even if we have a linearly separable training set, in the presence of noise it can happen that any good decision function must classify some examples incorrectly.

To avoid the first problem SVMs map the input data $x_1, \ldots, x_n$ into a (possibly infinite dimensional) Hilbert space—the so-called *feature space*—by a nonlinear *feature map* $\Phi : X \to H$. Necessary properties of Φ are discussed below. Now, the ansatz of the generalized portrait algorithm is

implemented in H instead of X, i.e., we simply replace x and x_i in (3) and (4) by $\Phi(x)$ and $\Phi(x_i)$ and the vector w in (4) is chosen from H. The corresponding algorithm is called *maximal margin classifier* and was the first classifier of SVM type (cf. [1]).

To avoid the second problem the linear constraints in (4) are relaxed to $y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$, $\xi_i \geq 0$. Then, in order to prevent trivial solutions the objective function also has to take the *slack variables* ξ_i into account. Combining both modifications can lead to the following quadratic optimization problem:

$$\begin{array}{lll} \text{minimize} & \langle w, w\rangle + c\sum_{i=1}^{n} \xi_i & \text{for } w, b, \xi \\ \text{subject to} & y_i(\langle w, \Phi(x_i)\rangle + b) \geq 1 - \xi_i, & i = 1, \ldots, n, \\ & \xi_i \geq 0, & i = 1, \ldots, n, \end{array} \tag{5}$$

where $c > 0$ is a free parameter which is usually tuned heuristically. Note that, due to the special form of the supplemented term $c\sum_{i=1}^{n} \xi_i$, the objective function is still convex. In the following we denote a solution of (5) by $(w_T^{\Phi,c}, b_T^{\Phi,c}) \in H \times \mathbb{R}$. Recently, it was shown in [2] that this solution is not unique in general. However, an algorithm $\mathscr{C}^{\Phi,c}$ that provides the decision function

$$f_T^{\Phi,c} := \operatorname{sign}\left(\langle w_T^{\Phi,c}, \Phi(.)\rangle + b_T^{\Phi,c}\right) \tag{6}$$

for every training set T is called 1-*norm soft margin classifier* (1-SMC) with feature map Φ and parameter c. The 1-SMC was introduced in [4] and its excellent learning ability has been proved in several experiments since then (cf. the brief surveys in [5, Chap. 8, 10, Chap. 12]).

To treat the above optimization problem algorithmically one usually consider the Wolfe dual (cf. [5, Chap. 6] for a derivation):

$$\begin{array}{lll} \text{maximize} & \sum_{i=1}^{n} \alpha_i - \frac{1}{4}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j)\rangle & \text{for } \alpha_i,\ i = 1, \ldots, n \\ \text{subject to} & \sum_{i=1}^{n} y_i \alpha_i = 0, & i = 1, \ldots, n. \\ & 0 \leq \alpha_i \leq c, & \end{array} \tag{7}$$

If $(\alpha_1^*, \ldots, \alpha_n^*)$ denotes a solution of (7) the solution vector $w_T^{\Phi,c}$ of (5) can be computed by

$$w_T^{\Phi,c} = \frac{1}{2}\sum_{i=1}^{n} y_i \alpha_i^* \Phi(x_i)$$

and the corresponding *bias* $b_T^{\Phi,c}$ by

$$b_T^{\Phi,c} = y_j - \frac{1}{2}\sum_{i=1}^{n} y_i \alpha_i^* \langle \Phi(x_i), \Phi(x_j) \rangle$$

for every α_j^* with $0 < \alpha_j^* < c$. Note that in both the optimization problem (7) and in the evaluation of the resulting decision function (6) only inner products of Φ with itself occur. Thus, instead of computing the feature map directly, it suffices to know the function $\langle \Phi(.), \Phi(.) \rangle : X \times X \to \mathbb{R}$. This leads to the following definition:

DEFINITION 1. A function $k : X \times X \to \mathbb{R}$ is said to be 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. [5, Chap. 3]). As indicated above the decision function only depends on the kernel. Thus we denote it by $f_T^{k,c}$ in the following.

Using kernels instead of computing feature maps directly also works in several other situations and is known as the so-called "kernel-trick" (cf. [8]). In fact, every algorithm that is based on inner products only, can be "kernelized." The advantage of this ansatz is that kernels often enlarge the considered class of functions without changing the design of an algorithm.

Obviously, not every kernel is a good kernel, e.g., for the kernel with feature map $\Phi = id_{\mathbb{R}^d}$ kernelizing has no effect and for the kernel with feature map $\Phi \equiv 1$ the 1-SMC cannot learn at all. Hence, it is natural to ask whether there are kernels that fit to every classification problem. Fortunately, such kernels actually exist. To introduce them let $k : X \times X \to \mathbb{R}$ be a kernel and let $\Phi : X \to H$ be a feature map of k. A function $f : X \to \mathbb{R}$ is *induced* by the kernel k if there exists an element $w \in H$ such that $f = \langle w, \Phi(.) \rangle$. We know from [9, Lemma 2] that this notion is independent of Φ and H. The following definition made in [9] is fundamental:

DEFINITION 2. A continuous kernel $k : X \times X \to \mathbb{R}$ is called *universal* if the set of all induced functions is dense in $C(X)$, i.e., for all $g \in C(X)$ and all $\varepsilon > 0$ there exists a function f induced by k with $\|f - g\|_\infty \leqslant \varepsilon$.

In [9, Theorem 9] it was shown that k is universal if $k(x, x) > 0$ for all $x \in X$ and $\operatorname{span}\{\Phi_n : n \geqslant 1\}$ forms a sub-algebra of $C(X)$ for a suitable feature map

$\Phi : X \to \ell_2$ with $\Phi(x) = (\Phi_n(x))_{n \geq 1}$. In particular, it turned out that the following kernels were universal (cf. [9, Sect. 3]):

- the Gaussian RBF kernel $\exp(-\sigma^2 \|. - .\|_2^2)$ for all $\sigma > 0$ and all compact $X \subset \mathbb{R}^d$.
- the kernel $\exp(\langle ., . \rangle)$ for all compact subsets $X \subset \mathbb{R}^d$.
- Vovk's real infinite polynomial $(1 - \langle ., . \rangle)^{-\alpha}$ for all $\alpha > 0$ and all compact subsets $X \subset \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$.
- the stronger regularized Fourier kernel $k(x, y) := \prod_{i=1}^d \frac{1-q^2}{2(1-2q\cos(x_i - y_i)+q^2)}$ for all $0 < q < 1$ and all compact $X \subset [0, 2\pi)^d$.
- the weaker regularized Fourier kernel $k(x, y) := \prod_{i=1}^d \frac{\pi}{2q \sinh(\pi/q)} \cosh\left(\frac{\pi - |x_i - y_i|}{q}\right)$ for all $0 < q < \infty$ and all compact $X \subset [0, 2\pi)^d$.

In [9] it was also shown that using universal kernels the 1-SMC is consistent for all classification problems with constant level of noise provided that the regularization parameter c is chosen in a specific manner that depends on the sample size n and the noise level. In this article, we show that these classifiers are even universally consistent provided that the parameters are chosen in this manner. To prepare this result recall that the covering numbers of a metric space (X, d) are defined by

$$\mathcal{N}((X, d), \varepsilon) := \inf\left\{ n \in \mathbb{N} : \exists x_1, \ldots, x_n \text{ with } X \subset \bigcup_{i=1}^n B_d(x_i, \varepsilon) \right\}$$

for all $\varepsilon > 0$. The space (X, d) is precompact if and only if $\mathcal{N}((X, d), \varepsilon)$ is finite for all $\varepsilon > 0$. We also need the following result which has been proved in [9, Lemma 3 and Corollary 7]:

LEMMA 1. *Let $k : X \times X \to \mathbb{R}$ be a universal kernel on a compact subset X of $\mathbb{R}^d$ and $\Phi : X \to H$ be a feature map of k. Then Φ is continuous and*

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

defines a metric on X such that $\mathrm{id} : (X, |.|) \to (X, d_k)$ *is continuous. In particular, $\mathcal{N}((X, d_k), \varepsilon)$ is finite for all $\varepsilon > 0$.*

Now, our first result which almost states universal consistency for the 1-SMC reads as follows:

THEOREM 1. *Let $X \subset \mathbb{R}^d$ be compact and $k : X \times X \to \mathbb{R}$ be a universal kernel. Then for all Borel probability measures P on $X \times Y$ and all $\varepsilon > 0$ there exists a constant $c^* > 0$ such that for all $c \geq c^*$ and all $n \geq 1$ we have*

$$\Pr^*(\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{k,c/n}) \leq \mathcal{R}_P + \varepsilon\}) \geq 1 - 2Me^{-(\varepsilon^6/2^{29}M^2)n},$$

where $\Pr^*$ *is the outer probability of P^n and $M := \frac{64}{\varepsilon}\mathcal{N}\left((X, d_k), \frac{\varepsilon}{32\sqrt{c}}\right)$.*

This theorem shows that given a classification problem, a universal kernel and an accuracy ε we just have to choose the parameter c "large enough" to obtain asymptotically a risk which is optimal up to ε. It turns out that the universal consistency of the 1-SMC which is stated in the following theorem is a direct consequence of Theorem 1:

THEOREM 2. *Let $X \subset \mathbb{R}^d$ be compact and k be a universal kernel on X with $\mathcal{N}((X, d_k), \varepsilon) \in \mathcal{O}(\varepsilon^{-\alpha})$ for some $\alpha > 0$. Suppose that we have a positive sequence (c_n) with $nc_n \to \infty$ and $c_n \in \mathcal{O}(n^{\beta-1})$ for some $0 < \beta < \frac{1}{\alpha}$. Then for all Borel probability measures P on $X \times Y$ and all $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \Pr^*(\{T \in (X \times Y)^n : \mathcal{R}_P(f_T^{k,c_n}) \leqslant \mathcal{R}_P + \varepsilon\}) = 1.$$

Since for no classifier there is a uniform rate of convergence we have not estimated the probability in the above equation asymptotically. Moreover note, that by (2) the constant c^* of Theorem 1 cannot be of the form $c^* = c_P \varepsilon^{-q}$, where $c_P > 0$ depends on the distribution P and $q > 0$ does not depend on it. In other words, the influence of the *unknown* measure P on c^* is rather strong and thus, it is almost useless to determine the (asymptotic) behavior of c^* with respect to P and ε in the general case. If we only consider noiseless problems which additionally guarantee a large margin, i.e., the classes $B_{-1}(P)$ and $B_1(P)$ have strictly positive distance, then c^* neither depends on the distribution nor on ε (cf. [9]). In particular, we obtain a uniform rate of convergence, namely $1 - e^{-c\varepsilon n}$, where $c > 0$ only depends on the margin.

For the Gaussian RBF kernel which is one of the most important kernels we immediately obtain the following corollary:

COROLLARY 1. *Let $X \subset \mathbb{R}^d$ be compact and k be a Gaussian RBF kernel on X. Moreover, let $c_n := n^{\beta-1}$ for some $0 < \beta < \frac{1}{d}$ and all $n \geqslant 1$. Then the 1-SMC with kernel k and sequence (c_n) is universally consistent.*

2. PROOFS OF THE THEOREMS

Before we prove Theorem 1 we would like to explain the basic idea of the proof. For this, let us suppose that we have an induced function $\langle w^*, \Phi(.)\rangle$ which has the constant values 1 and -1 on $B_1(P)$ and $B_{-1}(P)$, respectively. Moreover, we assume that the supervisor has a constant level of noise $p \in [0, \frac{1}{2})$. Now let us take a "representative" training set T of length n. Then one easily checks (cf. Lemma 5) that

$$\left\langle w_T^{\Phi,c/n}, w_T^{\Phi,c/n} \right\rangle + \frac{c}{n} \sum_{l=1}^{n} \xi_l \lesssim \langle w^*, w^* \rangle + 2cp.$$

Here $\lesssim$ means that the relation $\leqslant$ only holds "approximately." On the other hand, by the continuity of the decision function $w_T^{\Phi,c/n}$ a misclassified (compared with the optimal Bayes decision rule) element z forces the sum of those slack variables, which belong to samples in the "neighborhood" of z, to be "approximately" greater than their cardinality (cf. (13) in the proof of Lemma 6). Conversely, for a correctly classified element the corresponding sum of the slack variables is "approximately" larger than $2p$ times their cardinality (cf. (14) in the proof of Lemma 6). Combining these considerations we obtain

$$\begin{aligned} c(1-2p)P_X(E) + 2cp &= c(P_X(E) + 2pP_X(X\backslash E)) \\ &\lesssim \left\langle w_T^{\Phi,c/n}, w_T^{\Phi,c/n} \right\rangle + \frac{c}{n}\sum_{l=1}^{n} \xi_l \\ &\lesssim \langle w^*, w^* \rangle + 2cp, \end{aligned}$$

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," "neighborhood" and "approximately." Thus, we firstly concentrate ourselves on the constructive part of the proof, which specifies these notions. Necessary, but lengthy computations are worked out in several lemmas in the next section.

Proof of Theorem 1. For brevity's sake, let $s(x) := P(y = -1|x)$ for $x \in B_1(P)$, $s(x) := P(y = 1|x)$ for $x \in B_{-1}(P)$ and $s(x) = \frac{1}{2}$ otherwise. Then an easy computation (cf. Eq. (1)) shows

$$\mathscr{R}_P = \int_X s \, dP_X.$$

Trivially, we may assume without loss of generality, that $\varepsilon \in (0, 1]$. We define $\tau := \frac{\varepsilon}{32}$ and fix an integer m with $\frac{1}{2^m} \leqslant \tau \leqslant \frac{1}{2^{m-1}}$. Furthermore, let

$$X_i := \left\{ x \in X : \frac{i}{2^m} \leqslant s(x) < \frac{i+1}{2^m} \right\}, \qquad i = 0, \ldots, 2^{m-1} - 2,$$

$$X_{2^{m-1}-1} := \left\{ x \in X : \frac{1}{2} - \frac{1}{2^m} \leqslant s(x) \leqslant \frac{1}{2} \right\}.$$

Note, that this definition immediately yields

$$\sum_{i=0}^{2^{m-1}-1} \frac{i}{2^m} P_X(X_i) \leqslant \mathscr{R}_P \leqslant \sum_{i=0}^{2^{m-1}-1} \frac{i}{2^m} P_X(X_i) + \tau. \tag{8}$$

Due to the compactness of X the measure P_X is regular. Hence there exist compact subsets $\tilde{\boldsymbol{K}}_i^j \subset X_i^j := X_i \cap B_j(P)$, $i = 0, \ldots, 2^{m-1} - 2$, $j \in \{-1, 1\}$ and $\tilde{\boldsymbol{K}}_{2^{m-1}-1} \subset X_{2^{m-1}-1}$ such that

$$P_X(X_i^j \setminus \tilde{\boldsymbol{K}}_i^j) \leqslant \frac{\tau}{2^m}, \qquad i = 0, \ldots, 2^{m-1} - 2, \quad j \in \{-1, 1\},$$

$$P_X(X_{2^{m-1}-1} \setminus \tilde{\boldsymbol{K}}_{2^{m-1}-1}) \leqslant \frac{\tau}{2^m}.$$

For later purpose, we write $\tilde{\boldsymbol{K}}_{2^{m-1}-1}^1 := \tilde{\boldsymbol{K}}_{2^{m-1}-1} \cap (B_1(P) \cup B_0(P))$ and $\tilde{\boldsymbol{K}}_{2^{m-1}-1}^{-1} := \tilde{\boldsymbol{K}}_{2^{m-1}-1} \cap B_{-1}(P)$. Furthermore, let $\Phi : X \to H$ be a feature map of k. Since k is universal, Lemma 2 provides an element $w^* \in H$ such that

$$\langle w^*, \Phi(x) \rangle \in [1, 1 + \tau], \qquad x \in \bigcup_{i=0}^{2^{m-1}-2} \tilde{\boldsymbol{K}}_i^1,$$

$$\langle w^*, \Phi(x) \rangle \in [-(1 + \tau), -1], \qquad x \in \bigcup_{i=0}^{2^{m-1}-2} \tilde{\boldsymbol{K}}_i^{-1},$$

$$\langle w^*, \Phi(x) \rangle \in [-\tau, \tau], \qquad x \in K_{2^{m-1}-1}$$

hold and $\langle w^*, \Phi(.) \rangle$ only takes values between $-(1 + \tau)$ and $1 + \tau$. We define $c^* := \frac{34}{\varepsilon} \|w^*\|_H^2$ and for fixed $c \geqslant c^*$ we introduce $\sigma := \frac{\tau}{\sqrt{c}}$. Then for every $i = 0, \ldots, 2^{m-1} - 1$ and $j \in \{-1, 1\}$ there exists a finite partition $\tilde{\mathscr{A}}_i^j$ of $\tilde{\boldsymbol{K}}_i^j$ such that each $A \in \tilde{\mathscr{A}}_i^j$ has diameter less than or equal to σ with respect to the metric d_k introduced in Lemma 1. Moreover, by the definition of the covering numbers we can also ensure $|\tilde{\mathscr{A}}_i^j| \leqslant \mathscr{N}((X, d_k), \sigma)$. We define

$$\mathscr{A}_i^j := \left\{ A \in \tilde{\mathscr{A}}_i^j : P_X(A) \geqslant \frac{2\tau}{M} \right\}.$$

Note that this immediately yields that the cardinality of the union of all $\mathscr{A}_i^j$ is smaller than or equal to M. For later purpose we write $K_i^j := \bigcup_{A \in \mathscr{A}_i^j} A$ for all $i = 0, \ldots, 2^{m-1} - 1$, $j \in \{-1, 1\}$. Now we construct "representative"

training sets. For this let

$$F^+_{n,A} := \Bigg\{((x_1,y_1),\ldots,(x_n,y_n)) : |\{l : x_l \in A, y_l = j\}| \geq (1-\tau)\left(1-\frac{i+1}{2^m}\right)P_X(A)n\Bigg\},$$

$$F^-_{n,A} := \left\{((x_1,y_1),\ldots,(x_n,y_n)) : |\{l : x_l \in A, y_l \neq j\}| \geqslant (1-\tau)\frac{i}{2^m}P_X(A)n\right\},$$

where $n \geqslant 1$, $i = 0,\ldots,2^{m-1}-2$, $j \in \{-1,1\}$ and $A \in \mathscr{A}^j_i$. Moreover, for $A \in \mathscr{A}^j_{2^{m-1}-1}$, $j \in \{-1,1\}$ let

$$F^+_{n,A} := \Bigg\{((x_1,y_1),\ldots,(x_n,y_n)) : |\{l : x_l \in A, y_l = 1\}| \geqslant (1-\tau)\left(\frac{1}{2}-\frac{1}{2^m}\right)P_X(A)n\Bigg\},$$

$$F^-_{n,A} := \Bigg\{((x_1,y_1),\ldots,(x_n,y_n)) : |\{l : x_l \in A, y_l = -1\}| \geqslant (1-\tau)\left(\frac{1}{2}-\frac{1}{2^m}\right)P_X(A)n.\Bigg\}$$

Furthermore, for $n \geqslant 1$ we denote by F_n the intersection of all of the above sets, i.e.,

$$F_n := \bigcap_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-1} \bigcap_{A\in\mathscr{A}^j_i} \left(F^+_{n,A} \cap F^-_{n,A}\right).$$

By Lemma 3 we obtain $P^n(F_n) \geqslant 1 - 2Me^{-2(\tau^6/M^2)n}$ for all $n \geqslant 1$. Therefore it suffices to show that $\mathscr{R}_P\left(f^{k,c/n}_T\right) \leqslant \mathscr{R}_P + \varepsilon$ holds for all $T \in F_n$. Let us assume the converse, i.e., there exists a training set $T = ((x_1,y_1),\ldots,(x_n,y_n)) \in F_n$ with

$$\mathscr{R}_P\left(f^{k,c/n}_T\right) > \mathscr{R}_P + \varepsilon. \tag{9}$$

Then for $i = 0,\ldots,2^{m-1}-2$ and $j \in \{-1,1\}$ we denote the set of misclassified points (compared with the optimal Bayes decision rule) in X^j_i

by E_i^j, i.e.,

$$E_i^j := \left\{ x \in X_i^j : \ f_T^{k,c/n}(x) \neq j \right\}.$$

Analogously, let

$$E_{2^{m-1}-1}^j := \left\{ x \in X_{2^{m-1}-1} \cap B_j(P) : \ f_T^{k,c/n}(x) \neq j \right\}.$$

Since we know by Lemma 4 that

$$\begin{aligned} \mathscr{R}_P\left(f_T^{k,c/n}\right) \leqslant \mathscr{R}_P &+ 2^{-m+1} P_X\left(E_{2^{m-1}-1}^1 \cup E_{2^{m-1}-1}^{-1}\right) \\ &+ \sum_{i=0}^{2^{m-1}-2} \left(1 - \frac{i}{2^{m-1}}\right) P_X\left(E_i^1 \cup E_i^{-1}\right) \end{aligned}$$

holds, our assumption (9) and $2^{-m+1} P_X\left(E_{2^{m-1}-1}^1 \cup E_{2^{m-1}-1}^{-1}\right) \leqslant 2^{-m+1} \leqslant 2\tau$ yield

$$\varepsilon - 2\tau < \sum_{i=0}^{2^{m-1}-2} \left(1 - \frac{i}{2^{m-1}}\right) P_X\left(E_i^1 \cup E_i^{-1}\right). \tag{10}$$

Now let us denote the slack variables, which correspond to a fixed solution $(w_T^{\Phi,c/n}, b_T^{\Phi,c/n})$ of our optimization problem (5), by $\xi_1, \ldots, \xi_n$. Then Lemma 5 yields

$$\left\langle w_T^{\Phi,c/n}, w_T^{\Phi,c/n} \right\rangle + \frac{c}{n} \sum_{l=1}^{n} \xi_l \leqslant \langle w^*, w^* \rangle + 2c(1-\tau)(\mathscr{R}_P + 9\tau). \tag{11}$$

On the other hand, by Lemma 6 and inequality (10) we obtain

$$\begin{aligned} \frac{c}{n} \sum_{l=1}^{n} \xi_l &\geqslant (1-\tau)^2 c \left(2\mathscr{R}_P + \sum_{i=0}^{2^{m-1}-2} \left(1 - \frac{i}{2^{m-1}}\right) P_X\left(E_i^1 \cup E_i^{-1}\right) - 9\tau \right) \\ &> (1-\tau)^2 c(2\mathscr{R}_P + \varepsilon - 11\tau) \\ &= c(1-\tau)(2\mathscr{R}_P + \varepsilon - 11\tau - 2\tau\mathscr{R}_P - \varepsilon\tau - 11\tau^2) \\ &> c(1-\tau)(2\mathscr{R}_P + \varepsilon - 13\tau). \end{aligned}$$

Therefore our assumption (9) must be false since inequality (11) yields

$$\begin{aligned}\|w^*\|_H^2 &> c(1-\tau)(2\mathscr{R}_P + \varepsilon - 13\tau) - c(1-\tau)(2\mathscr{R}_P + 18\tau)\\ &= c(1-\tau)(\varepsilon - 31\tau)\\ &\geqslant c\left(1 - \frac{1}{32}\right)\frac{\varepsilon}{32}\\ &\geqslant \|w^*\|_H^2. \quad \blacksquare\end{aligned}$$

Proof of Theorem 2. Since $nc_n \to \infty$ there exists an integer n_0 such that $nc_n \geqslant c^*$ for all $n \geqslant n_0$. Thus for $n \geqslant n_0$ Theorem 1 yields

$$\Pr^*\left(\left\{T:\ \mathscr{R}_P\left(f_T^{k,c_n}\right) \leqslant \mathscr{R}_P + \varepsilon\right\}\right) \geqslant 1 - 2M_n e^{-(\varepsilon^6/2^{29}M_n^2)n},$$

where $M_n := \frac{64}{\varepsilon}\mathscr{N}\left((X,d_k), \frac{\varepsilon}{32\sqrt{nc_n}}\right)$. Moreover, by the assumption on the covering numbers of (X,d_k) we obtain $M_n^2 \in \mathscr{O}((nc_n)^\alpha)$ and thus $nM_n^{-2} \to \infty$. $\blacksquare$

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

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

This yields $\mathscr{N}((X,d_k),\varepsilon) \leqslant \mathscr{N}\left((X,\|.\|_2), \frac{\varepsilon}{\sqrt{2}\sigma}\right)$ and thus $\mathscr{N}((X,d_k),\varepsilon) \in \mathscr{O}(\varepsilon^{-d})$ (cf. [3, p. 9]). $\blacksquare$

3. PROOFS OF THE LEMMAS

In this section, we show the lemmas used in the proof of Theorem 1. We begin with the following result which is needed to construct an almost optimal decision function:

LEMMA 2. *Let $X \subset \mathbb{R}^n$ be compact and $k: X \times X \to \mathbb{R}$* be a *universal kernel. Then for all $\varepsilon > 0$ and all pairwise disjoint and compact subsets K_{-1}, K_0 and K_1 there exists an induced function $f: X \to [-(1+\varepsilon), 1+\varepsilon]$ such that*

$$\begin{aligned} f(x) &\in [1, 1+\varepsilon], && x \in K_1,\\ f(x) &\in [-(1+\varepsilon), -1], && x \in K_{-1},\\ f(x) &\in [-\varepsilon, \varepsilon], && x \in K_0. \end{aligned}$$

Proof. It suffices to show that there exists a continuous function g on X with values in $[-(1+\varepsilon/2), 1+\varepsilon/2]$ such that $g(x) = 1+\varepsilon/2$ if $x \in K_1$, $g(x) = -(1+\varepsilon/2)$ if $x \in K_{-1}$ and $g(x) = 0$ if $x \in K_0$. In fact,

$$x \mapsto \left(1+\frac{\varepsilon}{2}\right)\left(\frac{d(x, K_{-1} \cup K_0)}{d(x, K_{-1} \cup K_0) + d(x, K_1)} - \frac{d(x, K_1 \cup K_0)}{d(x, K_1 \cup K_0) + d(x, K_{-1})}\right)$$

is such a function. ∎

The next lemma estimates the probabilities of the "representative" training sets constructed in the proof of Theorem 1:

LEMMA 3. *Using the notations of the proof of Theorem* 1 *we have*

$$P^n(F_n) \geqslant 1 - 2Me^{-2(\tau^6/M^2)n}.$$

Proof. Let us recall Hoeffding's inequality (cf. [6, Theorem 8.1]), which in particular states that for all i.i.d. random variables $z_i : (\Omega, \mathscr{A}, Q) \to \{0,1\}$ and all $\delta \in (0,1)$, $n \geqslant 1$ we have

$$Q^n\left(\sum_{i=1}^n z_i \leqslant (1-\delta)qn\right) \leqslant e^{-2(\delta q)^2 n},$$

where $q := Q(z_i = 1) = \mathbb{E}z_i$. Thus for $i = 0, \ldots, 2^{m-1} - 2$, $j \in \{-1, 1\}$ and $A \in \mathscr{A}_i^j$ we get

$$\begin{aligned} P^n(F_{n,A}^+) = P^n\Bigg(\Bigg\{&((x_1, y_1), \ldots, (x_n, y_n)) : |\{l : x_l \in A, y_l = j\}| \\ &\geqslant (1-\tau)\left(1 - \frac{i+1}{2^m}\right) P_X(A) n\Bigg\}\Bigg) \\ \geqslant 1 - P^n\Bigg(\Bigg\{&((x_1, y_1), \ldots, (x_n, y_n)) : |\{l : x_l \in A, y_l = j\}| \\ &\leqslant (1-\tau)\int_A (1-s)\, dP_X n\Bigg\}\Bigg) \\ \geqslant 1 - e^{-2(\tau^4/M^2)n}. \end{aligned}$$

This yields $P^n(Z^n \backslash F_{n,A}^+) \leqslant e^{-2(\tau^4/M^2)n} \leqslant e^{-2(\tau^6/M^2)n}$, where $Z := X \times Y$. Analogously, we obtain $P^n(Z^n \backslash F_{n,A}^{\pm}) \leqslant e^{-2(\tau^6/M^2)n}$ for all $A \in \mathscr{A}_{2^{m-1}-1}^j$, $j \in \{-1, 1\}$. Moreover, $P^n(Z^n \backslash F_{n,A}^-) \leqslant e^{-2(\tau^6/M^2)n}$ is trivial for $j \in \{-1, 1\}$ and $A \in \mathscr{A}_0^j$.

Finally, for $i = 1, \ldots, 2^{m-1} - 2$, $j \in \{-1, 1\}$, $A \in \mathscr{A}_i^j$ and $q := \int_A s \, dP_X$ we find

$$\begin{aligned} P^n(F_{n,A}^+) &\geqslant 1 - P^n\left(\left\{((x_1, y_1), \ldots, (x_n, y_n)) : |\{l : x_l \in A, y_l = j\}| \right.\right. \\ &\qquad \left.\left. \leqslant (1 - \tau) \int_A s \, dP_X n \right\}\right) \\ &\geqslant 1 - e^{-2\tau^2 q^2 n} \\ &\geqslant 1 - e^{-2\tau^2 (\tau^2/M)^2 n}, \end{aligned}$$

i.e., $P^n(Z^n \backslash F_{n,A}^+) \leqslant e^{-2(\tau^6/M^2)n}$. The definition of F_n thus yields

$$\begin{aligned} P^n(F^n) &= P^n\left(\bigcap_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-1} \bigcap_{A\in\mathscr{A}_i^j} \left(F_{n,A}^+ \cap F_{n,A}^-\right)\right) \\ &= 1 - P^n\left(\bigcup_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-1} \bigcup_{A\in\mathscr{A}_i^j} \left(\left(Z^n \backslash F_{n,A}^+\right) \cup \left(Z^n \backslash F_{n,A}^-\right)\right)\right) \\ &\geqslant 1 - \sum_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-1} \sum_{A\in\mathscr{A}_i^j} P^n\left(Z^n \backslash F_{n,A}^+\right) - \sum_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-1} \sum_{A\in\mathscr{A}_i^j} P^n\left(Z^n \backslash F_{n,A}^-\right) \\ &\geqslant 1 - 2Me^{-2(\tau^6/M^2)n}. \quad \blacksquare \end{aligned}$$

The following lemma estimates the risk of a decision function from above:

LEMMA 4. *With the notations of the proof of Theorem* 1 *we have*

$$\begin{aligned} \mathscr{R}_P\left(f_T^{k,c/n}\right) \leqslant \mathscr{R}_P &+ 2^{-m+1} P_X\left(E_{2^{m-1}-1}^1 \cup E_{2^{m-1}-1}^{-1}\right) \\ &+ \sum_{i=0}^{2^{m-1}-2} \left(1 - \frac{i}{2^{m-1}}\right) P_X\left(E_i^1 \cup E_i^{-1}\right). \end{aligned}$$

Proof. Firstly, with $E_1 := \bigcup_{i=0}^{2^{m-1}-1} E_i^1$ we observe that

$$\begin{aligned} &\int_{B_1(P)} \mathbf{1}_{\left\{f_T^{k,c/n}(x) \neq y\right\}} P(dx, dy) \\ &= \int_{B_1(P)} \left(\mathbf{1}_{\left\{f_T^{k,c/n}(x) \neq -1\right\}} P(y = -1|x) + \mathbf{1}_{\left\{f_T^{k,c/n}(x) \neq 1\right\}} P(y = 1|x)\right) P_X(dx) \end{aligned}$$

$$
\begin{aligned}
&= \int_{B_1(P)\setminus E_1} P(y=-1|x)P_X(dx) + \int_{E_1} P(y=1|x)P_X(dx) \\
&= \int_{B_1(P)} P(y=-1|x)P_X(dx) + \int_{E_1} (1-2P(y=-1|x))P_X(dx) \\
&\leqslant \int_{B_1(P)} P(y=-1|x)P_X(dx) + \sum_{i=0}^{2^{m-1}-1} \left(1-\frac{i}{2^{m-1}}\right) P_X\left(E_i^1\right)
\end{aligned}
$$

holds. Analogously, we obtain

$$
\begin{aligned}
&\int_{B_{-1}(P)} \mathbf{1}_{\left\{f_T^{k,c/n}(x)\neq y\right\}} P(dx,dy) \\
&\quad \leqslant \int_{B_{-1}(P)} P(y=1|x)P_X(dx) + \sum_{i=0}^{2^{m-1}-1} \left(1-\frac{i}{2^{m-1}}\right) P_X\left(E_i^{-1}\right).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\mathscr{R}_P\left(f_T^{k,c/n}\right) &= \int_{B_1(P)} \mathbf{1}_{\left\{f_T^{k,c/n}(x)\neq y\right\}} P(dx,dy) \\
&\quad + \int_{B_{-1}(P)} \mathbf{1}_{\left\{f_T^{k,c/n}(x)\neq y\right\}} P(dx,dy) + \frac{1}{2} P_X(B_0(P)) \\
&\leqslant \mathscr{R}_P + \sum_{i=0}^{2^{m-1}-1} \left(1-\frac{i}{2^{m-1}}\right) P_X\left(E_i^1 \cup E_i^{-1}\right) \\
&= \mathscr{R}_P + 2^{-m+1} P_X\left(E_{2^{m-1}-1}^1 \cup E_{2^{m-1}-1}^{-1}\right) \\
&\quad + \sum_{i=0}^{2^{m-1}-2} \left(1-\frac{i}{2^{m-1}}\right) P_X\left(E_i^1 \cup E_i^{-1}\right). \quad \blacksquare
\end{aligned}
$$

The next lemma provides an estimate for the value of the optimization problem (5) from above:

LEMMA 5. *With the notations of the proof of Theorem* 1 *we have*

$$
\left\langle w_T^{\Phi,c/n}, w_T^{\Phi,c/n} \right\rangle + \frac{c}{n} \sum_{l=1}^{n} \xi_l \leqslant \langle w^*, w^* \rangle + 2c(1-\tau)(\mathscr{R}_P + 9\tau).
$$

Proof. We will compare the value of optimization problem (5) with the value of the objective function in $(w^*, 0)$. Thus, firstly have to construct an

admissible slack variable ξ^* corresponding to $(w^*, 0)$. For this let (x_l, y_l) be a sample of T. If $x_l \in K_i^1$ for some $i = 0, \ldots, 2^{m-1} - 2$ and x_l is correctly labeled, i.e., $y_l = 1$, we have $y_l \langle w^*, \Phi(x_l)\rangle \geq 1$. Hence, let $\xi_l^* := 0$. Analogously, we define $\xi_l^* := 0$ if $x_l \in K_i^{-1}$ for some $i = 0, \ldots, 2^{m-1} - 2$ and $y_l = -1$. Conversely, if $x_l \in K_i^1$ for some $i = 0, \ldots, 2^{m-1} - 2$ and x_l is not correctly labeled, i.e., $y_l = -1$, we have $y_l \langle w^*, \Phi(x_l)\rangle \geq -(1 + \tau)$ by the definition of w^*. Thus, let $\xi_l^* := 2 + \tau$. Again, if $x_l \in K_i^{-1}$ for some $i = 0, \ldots, 2^{m-1} - 2$ and $y_l = 1$ we analogously define $\xi_l^* := 2 + \tau$. If $x_l \in K_{2^{m-1}-1}$ is positively labeled, i.e., $y_l = 1$, we get $y_l \langle w^*, \Phi(x_l)\rangle \geq -\tau$. Hence, let $\xi_l^* := 1 + \tau$. This may also be done if $x_l \in K_{2^{m-1}-1}$ and $y_l = -1$. Finally, if x_l is neither an element of any K_i^j, $i = 0, \ldots, 2^{m-1} - 2$, $j \in \{-1, 1\}$ nor an element of $K_{2^{m-1}-1}$ we obtain $|\langle w^*, \Phi(x_l)\rangle| \leq 1 + \tau$. Thus let $\xi_l^* := 2 + \tau$ in this case. For brevity's sake, we now define

$$a_1 := \left|\left\{ l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^1, y_l = 1 \right\}\right| + \left|\left\{ l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^{-1}, y_l = -1 \right\}\right|,$$

$$a_2 := \left|\left\{ l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^1, y_l = -1 \right\}\right| + \left|\left\{ l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^{-1}, y_l = 1 \right\}\right|,$$

$$a_3 := |\{ l : x_l \in K_{2^{m-1}-1} \}|,$$

$$a_4 := \left|\left\{ l : x_l \notin K_{2^{m-1}-1} \cup \bigcup_{i=0}^{2^{m-1}-2} \left(K_i^1 \cup K_i^{-1}\right) \right\}\right|.$$

Since the training set T has length n we obviously have $a_1 + a_2 + a_3 + a_4 = n$. Moreover, the above considerations on ξ^* yield

$$\begin{aligned}
\left\langle w_T^{\Phi, c/n}, w_T^{\Phi, c/n} \right\rangle + \frac{c}{n} \sum_{i=1}^{n} \xi_l &\leq \langle w^*, w^* \rangle + \frac{c}{n} \sum_{l=1}^{n} \xi_l^* \\
&\leq \langle w^*, w^* \rangle + \frac{c}{n} ((2 + \tau) a_2 + (1 + \tau) a_3 + (2 + \tau) a_4) \\
&= \langle w^*, w^* \rangle + \frac{c}{n} ((2 + \tau)(n - a_1) - a_3) \qquad (12)
\end{aligned}$$

since $\left(w_T^{\Phi,c/n}, b_T^{\Phi,c/n}\right)$ together with the corresponding slack variable ξ is a solution of problem (5). Furthermore, the construction of F_n implies

$$\begin{aligned}
\frac{2+\tau}{n}(n-a_1) &\leqslant (2+\tau)\left(1 - \sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2} \sum_{A\in\mathscr{A}_i^j} (1-\tau)\left(1-\frac{i+1}{2^m}\right) P_X(A)\right)\\
&\leqslant 2 - 2(1-\tau)\sum_{i=0}^{2^{m-1}-2}\left(1-\frac{i+1}{2^m}\right) P_X\left(\tilde{\boldsymbol{K}}_i^1 \cup \tilde{\boldsymbol{K}}_i^{-1}\right) + 5\tau\\
&= 2(1-\tau)\left(1 - \sum_{i=0}^{2^{m-1}-2}\left(1-\frac{i+1}{2^m}\right) P_X\left(\tilde{\boldsymbol{K}}_i^1 \cup \tilde{\boldsymbol{K}}_i^{-1}\right)\right) + 7\tau\\
&\leqslant 2(1-\tau)\left(1 - \sum_{i=0}^{2^{m-1}-2} P_X\left(\tilde{\boldsymbol{K}}_i^1 \cup \tilde{\boldsymbol{K}}_i^{-1}\right)\right.\\
&\qquad \left. + \sum_{i=0}^{2^{m-1}-2} \frac{i}{2^m} P_X\left(\tilde{\boldsymbol{K}}_i^1 \cup \tilde{\boldsymbol{K}}_i^{-1}\right)\right) + 9\tau.
\end{aligned}$$

Considering $F_{n,A}^+$ and $F_{n,A}^-$ for all $A \in \mathscr{A}^j_{2^{m-1}-1}$, $j \in \{-1,1\}$ we also get

$$\begin{aligned}
\frac{a_3}{n} &\geqslant 2 \sum_{\substack{A\in\mathscr{A}^j_{2^{m-1}-1}\\ j\in\{-1,1\}}} (1-\tau)\left(\frac{1}{2}-\frac{1}{2^m}\right) P_X(A)\\
&\geqslant 2(1-\tau)\left(\frac{1}{2}-\frac{1}{2^m}\right)\left(P_X\left(\tilde{\boldsymbol{K}}_{2^{m-1}-1}\right) - 2\tau\right)\\
&\geqslant 2(1-\tau)\left(P_X\left(\tilde{\boldsymbol{K}}_{2^{m-1}-1}\right) - \left(\frac{1}{2}-\frac{1}{2^m}\right) P_X\left(\tilde{\boldsymbol{K}}_{2^{m-1}-1}\right)\right) - 6\tau.
\end{aligned}$$

If we combine these estimates with inequality (8) we thus obtain

$$\begin{aligned}
&\frac{1}{n}\left((2+\tau)(n-a_1) - a_3\right)\\
&\quad\leqslant 2(1-\tau)\left(1 - \sum_{i=0}^{2^{m-1}-1} P_X\left(\tilde{\boldsymbol{K}}_i^1 \cup \tilde{\boldsymbol{K}}_i^{-1}\right) + \sum_{i=0}^{2^{m-1}-1} \frac{i}{2^m} P_X\left(\tilde{\boldsymbol{K}}_i^1 \cup \tilde{\boldsymbol{K}}_i^{-1}\right)\right) + 15\tau\\
&\quad\leqslant 2(1-\tau)\left(\tau + \sum_{i=0}^{2^{m-1}-1} \frac{i}{2^m} P_X(X_i)\right) + 15\tau
\end{aligned}$$

$$\leqslant 2(1-\tau)(\tau + \mathscr{R}_P) + 15\tau$$

$$\leqslant 2(1-\tau)(\mathscr{R}_P + 9\tau).$$

The assertion now follows with estimate (12). ∎

The last lemma estimates the value of optimization problem (5) from below:

LEMMA 6. *With the notations of the proof of Theorem* 1 *we have*

$$\frac{c}{n}\sum_{l=1}^{n}\xi_l \geqslant (1-\tau)^2 c\left(2\mathscr{R}_P + \sum_{i=0}^{2^{m-1}-2}\left(1-\frac{i}{2^{m-1}}\right)P_X\left(E_i^1 \cup E_i^{-1}\right) - 9\tau\right).$$

Proof. For $i = 0, \ldots, 2^{m-1}-2$ and $j \in \{-1, 1\}$ we define

$$I_i^j := \bigcup_{\substack{A\in\mathscr{A}_i^j \\ A\cap E_i^j \neq \emptyset}} \{l : x_l \in A\},$$

$$J_i^j := \bigcup_{\substack{A\in\mathscr{A}_i^j \\ A\cap E_i^j \neq \emptyset}} \{l : x_l \in A\}.$$

Now, our first goal is to show that

$$\frac{1}{n}\sum_{l\in I_i^j}\xi_l \geqslant (1-\tau)^2\left(1-\frac{1}{2^m}\right)\sum_{\substack{A\in\mathscr{A}_i^j \\ A\cap E_i^j \neq \emptyset}} P_X(A), \tag{13}$$

$$\frac{1}{n}\sum_{l\in J_i^j}\xi_l \geqslant (1-\tau)^2\frac{i}{2^{m-1}}\sum_{\substack{A\in\mathscr{A}_i^j \\ A\cap E_i^j = \emptyset}} P_X(A), \tag{14}$$

$$\frac{1}{n}\sum_{\substack{A\in\mathscr{A}_{2^{m-1}-1}^{\pm 1} \\ x_l\in A}}\xi_l \geqslant (1-\tau)^2\left(1-\frac{1}{2^{m-1}}\right)\left(P_X\left(\tilde{\boldsymbol{K}}_{2^{m-1}-1}\right) - 2\tau\right) \tag{15}$$

hold for all $i = 0, \ldots, 2^{m-1}-2$ and $j \in \{-1, 1\}$. For this, we firstly compare the value of optimization problem (5) with the value of the objective function in $(0, 0)$ and obtain

$$\left\langle w_T^{\Phi,c/n}, w_T^{\Phi,c/n}\right\rangle \leqslant \left\langle w_T^{\Phi,c/n}, w_T^{\Phi,c/n}\right\rangle + \frac{c}{n}\sum_{l=1}^{n}\xi_l \leqslant c,$$

i.e., $\|w_T^{\Phi,c/n}\| \leqslant \sqrt{c}$. To show inequality (13) let $A \in \mathscr{A}_i^j$ with $A \cap E_i^j \neq \emptyset$. Then for fixed $z \in A \cap E_i^j$ we define $a := -(\langle w_T^{\Phi,c/n}, \Phi(z)\rangle + b_T^{\Phi,c/n})$. Without loss of generality we may assume $j = 1$, i.e., $a \geqslant 0$. Now, for an index l with $x_l \in A$ and $y_l = 1$ we have $\|\Phi(x_l) - \Phi(z)\| = d_k(x_l, z) \leqslant \sigma = \frac{\tau}{\sqrt{c}}$ and this yields

$$\begin{aligned} 1 - \xi_l &\leqslant \left\langle w_T^{\Phi,c/n}, \Phi(x_l)\right\rangle + b_T^{\Phi,c/n} \\ &= \left\langle w_T^{\Phi,c/n}, \Phi(x_l) - \Phi(z)\right\rangle + \left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle + b_T^{\Phi,c/n} \\ &\leqslant \|w_T^{\Phi,c/n}\| \cdot \|\Phi(x_l) - \Phi(z)\| - a \\ &\leqslant \tau - a, \end{aligned}$$

i.e., $\xi_l \geqslant 1 - \tau + a > 0$. Analogously, for an index l with $x_l \in A$ and $y_l = -1$ we obtain

$$\begin{aligned} 1 - \xi_l &\leqslant -\left(\left\langle w_T^{\Phi,c/n}, \Phi(x_l)\right\rangle_T^{\Phi,c/n}\right) \\ &= -\left\langle w_T^{\Phi,c/n}, \Phi(x_l) - \Phi(z)\right\rangle - \left(\left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle + b_T^{\Phi,c/n}\right) \\ &\leqslant \tau + a, \end{aligned}$$

i.e., $\xi_l \geqslant \max\{0, 1 - \tau - a\}$. Let us suppose that $1 - \tau - a \geqslant 0$. Then by the definition of F_n we get

$$\begin{aligned} \frac{1}{n}\sum_{x_l \in A} \xi_l &\geqslant (1 - \tau + a)(1 - \tau)\left(1 - \frac{i+1}{2^m}\right)P_X(A) + (1 - \tau - a)(1 - \tau)\frac{i}{2^m}P_X(A) \\ &= (1 - \tau)^2\left(1 - \frac{1}{2^m}\right)P_X(A) + a(1 - \tau)\left(1 - \frac{2i+1}{2^m}\right)P_X(A) \\ &\geqslant (1 - \tau)^2\left(1 - \frac{1}{2^m}\right)P_X(A) \end{aligned}$$

since $(2i + 1)2^{-m} < 1$ and $a \geqslant 0$. On the other hand, if $1 - \tau - a < 0$ we have $1 - \tau + a > 2 - 2\tau$ and this, together with $(2i + 1)2^{-m} < 1$, implies

$$\begin{aligned} \frac{1}{n}\sum_{x_l \in A} \xi_l &\geqslant (1 - \tau + a)(1 - \tau)\left(1 - \frac{i+1}{2^m}\right)P_X(A) \\ &> (1 - \tau)^2\left(2 - \frac{2i+2}{2^m}\right)P_X(A) \\ &> (1 - \tau)^2\left(1 - \frac{1}{2^m}\right)P_X(A). \end{aligned}$$

Thus, we finally obtain

$$\frac{1}{n}\sum_{l\in I_i^j}\xi_l = \frac{1}{n}\sum_{\substack{A\in\mathscr{A}_i^j\\ A\cap E_i^j\neq\emptyset}}\sum_{x_l\in A}\xi_l \geqslant (1-\tau)^2\left(1-\frac{1}{2^m}\right)\sum_{\substack{A\in\mathscr{A}_i^j\\ A\cap E_i^j\neq\emptyset}} P_X(A).$$

Now we prove inequality (14). For this let $A\in\mathscr{A}_i^j$ with $A\cap E_i^j=\emptyset$. Then for fixed $z\in A\cap\left(X\backslash E_i^j\right)$ we define $a := -\left(\left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle + b_T^{\Phi,c/n}\right)$. Without loss of generality we may assume $j=-1$, i.e., $a\geqslant 0$. For an index l with $x_l\in A$ and $y_l=-1$ we thus obtain

$$\begin{aligned}1-\xi_l &\leqslant -\left(\left\langle w_T^{\Phi,c/n}, \Phi(x_l)\right\rangle + b_T^{\Phi,c/n}\right)\\ &= -\left\langle w_T^{\Phi,c/n}, \Phi(x_l)-\Phi(z)\right\rangle - \left(\left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle + b_T^{\Phi,c/n}\right)\\ &\leqslant \tau + a,\end{aligned}$$

i.e., $\xi_l\geqslant\max\{0, 1-\tau-a\}$. Analogously, for an index l with $x_l\in A$ and $y_l=1$ we obtain

$$\begin{aligned}1-\xi_l &\leqslant \left\langle w_T^{\Phi,c/n}, \Phi(x_l)\right\rangle + b_T^{\Phi,c/n}\\ &= \left\langle w_T^{\Phi,c/n}, \Phi(x_l)-\Phi(z)\right\rangle + \left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle + b_T^{\Phi,c/n}\\ &\leqslant \tau - a,\end{aligned}$$

i.e., $\xi_l\geqslant 1-\tau+a>0$. Let us suppose that $1-\tau-a\geqslant 0$. From the definition of F_n we get

$$\begin{aligned}\frac{1}{n}\sum_{x_l\in A}\xi_l &\geqslant (1-\tau-a)(1-\tau)\left(1-\frac{i+1}{2^m}\right)P_X(A) + (1-\tau+a)(1-\tau)\frac{i}{2^m}P_X(A)\\ &= (1-\tau)^2\left(1-\frac{1}{2^m}\right)P_X(A) - a(1-\tau)\left(1-\frac{2i+1}{2^m}\right)P_X(A)\\ &\geqslant (1-\tau)^2\frac{i}{2^{m-1}}P_X(A)\end{aligned}$$

since $(2i+1)2^{-m}<1$ and $a\leqslant 1-\tau$. On the other hand, if $1-\tau-a<0$ we have $1-\tau+a>2-2\tau$ and this implies

$$\frac{1}{n}\sum_{x_l\in A}\xi_l \geqslant (1-\tau+a)(1-\tau)\frac{i}{2^m}P_X(A)\geqslant (1-\tau)^2\frac{i}{2^{m-1}}P_X(A).$$

Therefore, we obtain

$$\frac{1}{n}\sum_{l\in I_i^j}\xi_l = \frac{1}{n}\sum_{\substack{A\in\mathscr{A}_i^j\\ A\cap E_i^j=\emptyset}}\sum_{x_l\in A}\xi_l \geqslant (1-\tau)^2\frac{i}{2^{m-1}}\sum_{\substack{A\in\mathscr{A}_i^j\\ A\cap E_i^j=\emptyset}} P_X(A).$$

Now we treat inequality (15). For this let $A\in\mathscr{A}^{\pm 1}_{2^{m-1}-1}$ and fix $z\in A$. Moreover, we define $a := -\left(\left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle b_T^{\Phi,c/n}\right)$. Suppose that we have an index l with $x_l\in A$ and $y_l=-1$. Then we obtain

$$\begin{aligned}1-\xi_l &\leqslant -\left(\left\langle w_T^{\Phi,c/n}, \Phi(x_l)\right\rangle + b_T^{\Phi,c/n}\right)\\ &= -\left\langle w_T^{\Phi,c/n}, \Phi(x_l)-\Phi(z)\right\rangle - \left(\left\langle w_T^{\Phi,c/n}, \Phi(z)\right\rangle + b_T^{\Phi,c/n}\right)\\ &\leqslant \tau - a,\end{aligned}$$

i.e., $\xi_l\geqslant\max\{0, 1-\tau-a\}$. Analogously, we check that $y_l=1$ implies $\xi_l\geqslant \max\{0, 1-\tau+a\}$ for all l with $x_l\in A$. If $a\in[-(1-\tau), 1-\tau]$ we thus obtain

$$\begin{aligned}\frac{1}{n}\sum_{\substack{A\in\mathscr{A}^{\pm 1}_{2^{m-1}-1}\\ x_l\in A}}\xi_l &\geqslant \left((1-\tau-a)(1-\tau)\left(\frac{1}{2}-\frac{1}{2^m}\right)\right.\\ &\qquad \left.+(1-\tau+a)(1-\tau)\left(\frac{1}{2}-\frac{1}{2^m}\right)\right)P_X(K_{2^{m-1}-1})\\ &\geqslant (1-\tau)^2\left(1-\frac{1}{2^{m-1}}\right)(P_X(\tilde{\boldsymbol{K}}_{2^{m-1}-1})-2\tau).\end{aligned}$$

On the other hand, if $a>1-\tau$ we have $1-\tau+a>2-2\tau$ and therefore inequality (15) also follows. Finally, for $a<-(1-\tau)$ we get $1-\tau-a> 2-2\tau$ and thus we obtain inequality (15) in this case, too. Having proved (13)–(15) we may now estimate

$$\begin{aligned}\frac{1}{n}\sum_{l=1}^n\xi_l \geqslant (1-\tau)^2\Bigg(&\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\Bigg(\left(1-\frac{1}{2^m}\right)\sum_{\substack{A\in\mathscr{A}_i^j\\ A\cap E_i^j\neq\emptyset}}P_X(A)\\ &+\frac{2i}{2^m}\sum_{\substack{A\in\mathscr{A}_i^j\\ A\cap E_i^j=\emptyset}}P_X(A)\Bigg)+\left(1-\frac{2}{2^m}\right)P_X(\tilde{\boldsymbol{K}}_{2^{m-1}-1})-2\tau\Bigg)\end{aligned}$$

$$= (1-\tau)^2 \left(\sum_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-2} \left(\left(1 - \frac{2i+1}{2^m}\right) \sum_{\substack{A\in\mathscr{A}_i^j \\ A\cap E_i^j \neq \emptyset}} P_X(A). \right.\right.$$

$$\left.\left. + \frac{2i}{2^m} \sum_{A\in\mathscr{A}_i^j} P_X(A) \right) + \left(1 - \frac{2}{2^m}\right) P_X(\tilde{\boldsymbol{K}}_{2^{m-1}-1}) - 2\tau \right)$$

$$\geqslant (1-\tau)^2 \left(\sum_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-2} \left(\left(1 - \frac{2i+1}{2^m}\right) \left(P_X\left(E_i^j\right) - \frac{\tau}{2^m}\right) + \frac{2i}{2^m} P_X(\tilde{\boldsymbol{K}}_i^j) \right) \right.$$

$$\left. + \left(1 - \frac{2}{2^m}\right) P_X(\tilde{\boldsymbol{K}}_{2^{m-1}-1}) - 4\tau \right).$$

Moreover, since inequality (8) we have

$$\sum_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-2} \frac{2i}{2^m} P_X\left(\tilde{\boldsymbol{K}}_i^j\right) + \left(1 - \frac{2}{2^m}\right) P_X\left(\tilde{\boldsymbol{K}}_{2^{m-1}-1}\right)$$

$$\geqslant \sum_{i=0}^{2^{m-1}-2} \frac{2i}{2^m} \left(P_X(X_i) - \frac{2\tau}{2^m}\right) + \left(1 - \frac{2}{2^m}\right) \left(P_X(X_{2^{m-1}-1}) - \frac{2\tau}{2^m}\right)$$

$$= \sum_{i=0}^{2^{m-1}-1} \frac{2i}{2^m} P_X(X_i) - \sum_{i=0}^{2^{m-1}-1} \frac{2i}{2^m} \frac{2\tau}{2^m}$$

$$\geqslant 2\mathscr{R}_P - 3\tau$$

and thus we may continue the above estimate to

$$\frac{1}{n} \sum_{l=1}^{n} \xi_l$$

$$\geqslant (1-\tau)^2 \left(\sum_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-2} \left(1 - \frac{2i+1}{2^m}\right) \left(P_X\left(E_i^j\right) - \frac{\tau}{2^m}\right) + 2\mathscr{R}_P - 7\tau \right). \qquad (16)$$

Furthermore, we also get

$$\begin{aligned}
&\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2} \left(1-\frac{2i+1}{2^m}\right)\left(P_X\left(E_i^j\right)-\frac{\tau}{2^m}\right)\\
&\quad= \sum_{i=0}^{2^{m-1}-2}\left(1-\frac{2i}{2^m}\right)P_X\left(E_i^1\cup E_i^{-1}\right)-\sum_{i=0}^{2^{m-1}-2}\left(1-\frac{2i}{2^m}\right)\frac{2\tau}{2^m}\\
&\qquad\quad-\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\frac{1}{2^m}P_X\left(E_i^j\right)+\sum_{i=0}^{2^{m-1}-2}\frac{2\tau}{2^{2m}}\\
&\quad\geqslant \sum_{i=0}^{2^{m-1}-2}\left(1-\frac{2i}{2^m}\right)P_X\left(E_i^1\cup E_i^{-1}\right)-2\tau
\end{aligned}$$

and therefore inequality (16) yields

$$\frac{c}{n}\sum_{l=1}^{n}\xi_l\geqslant(1-\tau)^2c\left(2\mathscr{R}_P+\sum_{i=0}^{2^{m-1}-2}\left(1-\frac{i}{2^{m-1}}\right)P_X\left(E_i^1\cup E_i^{-1}\right)-9\tau\right). \quad \blacksquare$$

REFERENCES

1. B. E. Boser, I. M. Guyon, and V. N. Vapnik, A training algorithm for optimal margin classifiers, *in* "Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory," (D. Haussler, Ed.), pp. 144–152, ACM Press, New York, 1992.
2. C. J. C. Burges and D. J. Crisp, Uniqueness of the SVM solution, *in* "Advances in Neural Information Processing," (M. I. Jordan, T. K. Leen, and K.-R. Müller, Ed.), Vol. 2, pp. 223–229, MIT Press, Cambridge, MA, 2000.
3. B. Carl and I. Stephani, "Entropy, Compactness and the Approximation of Operators," Cambridge Univ. Press, Cambridge, UK, 1990.
4. C. Cortes and V. N. Vapnik, Support vector networks, *Mach. Learning* **20** (1995), 273–297.
5. N. Cristianini and J. Shawe-Taylor, "Support Vector Machines and other Kernel-Based Learning Methods," Cambridge Univ. Press, Cambridge, UK, 2000.
6. L. Devroye, L. Györfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," Springer, New York, 1997.
7. R. M. Dudley, A course on empirical processes, *Lecture Notes Math.* **1097** (1984), 1–142.
8. B. Schölkopf and A. J. Smola, "Learning with Kernels," MIT Press, Cambridge, MA, 2002.
9. I. Steinwart, On the influence of the kernel on the consistency of support vector machines, *J. Mach. Learning Res.* **2** (2001), 67–93.
10. V. N. Vapnik, "Statistical Learning Theory," Wiley, New York, 1998.
11. V. N. Vapnik and A. Lerner, Pattern recognition using generalized portrait method, *Automat. Remote Control* **24** (1963), 774–780.