On the optimal parameter choice for ν-support vector machines

Ingo Steinwart*
Friedrich-Schiller-Universität
07743 Jena, Germany
steinwart@minet.uni-jena.de

April 4, 2002

Abstract

We determine the asymptotically optimal choice of the parameter ν for classifiers of ν-support vector machine (ν-SVM) type which has been introduced by Schölkopf et al.. It turns out that ν should be a close upper estimate of twice the optimal Bayes risk provided that the classifier uses a so-called universal kernel such as the Gaussian RBF kernel. Moreover, several experiments show that this result can be used to implement modified cross validation procedures which both train significantly faster and learn significantly better than standard cross validation techniques.

1 Introduction

One interesting type of support vector machines (SVM's) is the ν-SVM developed in [9]. Unlike classical SVM's it has an adjustable regularization parameter ν which possesses an intuitive meaning: ν is both an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors (cf. [9]). Additionally, both fractions tend almost surely to ν under rather general assumptions on the learning problem and the used kernel. Although this may help to understand the role of ν it is still unclear how to chose ν for a specific learning task. In this work we show that given some a-priori information on the expected optimal Bayes risk—namely, a (close) upper bound $\mathcal{R}$—an asymptotically good estimate of the optimal value of ν is $2\mathcal{R}$.
Before we make this precise let us recall some notions: in the following let X be a compact subset of $\mathbb{R}^d$, $Y := \{-1, 1\}$ and P be a probability measure on $X \times Y$, where X is equipped with the Borel σ-algebra. By disintegration (cf. [5, 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* $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ 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) generated with distribution P independently to T. Here, misclassification means $f_T(x) \neq y$. To make this precise, for a measurable function $f : X \to \{-1, 1\}$ we define the risk of f by

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

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

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

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

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. [4, Thm. 2.1.])

$$\mathcal{R}_P(f_0) = \inf\{\mathcal{R}_P(f) : f : X \to \{-1,1\} \text{ measurable}\} = \int_X s(x)\, P_X(dx) , \tag{1}$$

where the *noise level* $s : X \to \mathbb{R}$ of P 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 $s(x) = 1/2$ 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 $\mathcal{R}_P := \mathcal{R}_P(f_0)$ for the *(optimal) Bayes risk*. 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. Asymptotically, this means that for $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \Pr{}^*\big(\{T \in (X \times Y)^n : \mathcal{R}_P(f_T) \le \mathcal{R}_P + \varepsilon\}\big) = 1 . \tag{2}$$

Here the outer probability measure $\Pr^*$ of P^n is considered in order to avoid the question whether $T \to \mathcal{R}_P(f_T)$ is measurable. Recall, that the algorithm $\mathcal{C}$ is called universally consistent if (2) holds for all P and all $\varepsilon > 0$. Recently, it was shown by the author (cf. [12]) that a certain type of SVM's, namely the 1-norm soft margin classifier, is universally consistent provided that it uses a universal kernel (for a definition see below) and a specific sequence of regularization parameters. Although this result is of great theoretical interest since it shows that in principle SVM's can learn arbitrary classification problems it has the shortcoming that it does not provide a good estimate of the regularization parameter for a given classification task. As indicated above we will show that the ν-SVM approach may help to overcome these problems by incorporating some prior knowledge in terms of the optimal Bayes risk of the considered classification problem.
Before we state the announced results on ν-SVM's let us recall the definition of kernels and ν-SVM's:

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. [3, Ch. 3] and [10]). Obviously, not every kernel is a good kernel, e.g. the kernel that belongs to a constant feature map is not suitable for classification. However, there fortunately exist kernels that fit to all classification problems. To introduce them let $k : X \times X \to \mathbb{R}$ be a kernel and $\Phi : X \to H$ 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 [11, Lem. 2] that this notion is independent of Φ and H. The following definition made in [11] is fundamental for our purposes:

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 $\|f - g\|_\infty \le \varepsilon$.*

In [11, Thm. 9] it was shown that k is universal if $k(x,x) > 0$ for all $x \in X$ and $\mathrm{span}\{\Phi_n : n \geq 1\}$ forms a sub-algebra of $C(X)$ for a suitable feature map $\Phi : X \to \ell_2$ with $\Phi(x) = \big(\Phi_n(x)\big)_{n\geq 1}$. In particular, it turned out that the following kernels were universal (cf. [11, Sect. 3]):

- the Gaussian RBF kernel $\exp(-\sigma\|.-.\|_2^2)$ 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 $k(x,y) := \prod_{i=1}^d \frac{\pi}{2q\sinh(\pi/q)} \cosh(\frac{\pi-|x_i-y_i|}{q})$ for all $0 < q < \infty$ and all compact $X \subset [0, 2\pi)^d$.

Moreover, recall that a continuous kernel is universal if and only if its RKHS is dense in $C(X)$.
Finally, we need a brief description of the ν-SVM algorithm: let k be a kernel on X with feature map $\Phi : X \to H$. Moreover, let $\nu \in (0,1]$ and let $T = ((x_1,y_1),\dots,(x_n,y_n)) \in (X \times Y)^n$ be a training set. Then we denote a solution of the optimization problem

$$\begin{array}{lll} \text{minimize} & F_T(w,b,\rho,\xi) \;:=\; \langle w,w \rangle - \nu\rho + \frac{1}{n}\sum_{i=1}^n \xi_i & \text{for } w,b,\rho,\xi \\ \text{subject to} & y_i\big(\langle w, \Phi(x_i)\rangle + b\big) \geq \rho - \xi_i, & i = 1,\dots,n \\ & \xi_i \geq 0, & i = 1,\dots,n \\ & \rho \geq 0 & \end{array} \tag{3}$$

by $(w_T^{k,\nu}, b_T^{k,\nu}, \rho_T^{k,\nu}, \xi_T^{k,\nu}) \in H \times \mathbb{R} \times [0,\infty) \times \mathbb{R}^n$. An algorithm $\mathcal{C}^{k,\nu}$ that provides the decision function

$$\begin{aligned} f_T^{k,\nu} : X &\to \{-1,1\} \\ x &\mapsto \mathrm{sign}\big(\langle w_T^{k,\nu}, \Phi(x)\rangle + b_T^{k,\nu}\big) \end{aligned}$$

for every training set T is called ν-SVM classifier with kernel k and parameter ν. If k and ν are known from the context we usually omit them in the superscripts.
As usual for SVM's, (3) is solved in practice by the dual problem. Then it turns out that both the dual problem and the decision function are independent of the choice of the feature map of k. However, considering the dual optimization problem also causes a problem: although for the primal problem every $\nu \in [0,1]$ is feasible it was shown in [1] that for the dual problem the set of feasible ν's is smaller in general.
Finally, recall that the decision functions constructed on the basis of (3) coincide with the decision functions of the standard 1-norm soft margin classifier (cf. [9] and [1, Sect. 2]).

2 Results

Our first result estimates the risk of the decision functions $f_T^{k,\nu}$ provided that ν is chosen to be larger than twice the optimal Bayes risk:

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 $X \times Y$, all $\nu > 2\mathcal{R}_P$ and all $\varepsilon > 0$ there exists a constant $c > 0$ such that for all $n \geq 1$ we have*

$$\Pr{}^*\big(\{T \in (X \times Y)^n : \ \mathcal{R}_P(f_T^{k,\nu}) \leq \nu - \mathcal{R}_P + \varepsilon\}\big) \ \geq \ 1 - e^{-cn} \ .$$

The following corollary shows that given an upper bound $\mathcal{R}$ on $\mathcal{R}_P$ the decision function with respect to $\nu = 2\mathcal{R}$ almost surely achieves a risk not larger than $\mathcal{R}_P + 2(\mathcal{R} - \mathcal{R}_P)$ in the limit. In other words, if we know $\mathcal{R}_P$ up to $\delta > 0$ we obtain a risk not larger that $\mathcal{R}_P + 2\delta$:

Corollary 1 *Let $X \subset \mathbb{R}^d$ be compact, k be a universal kernel on X and P be a Borel probability measure on $X \times Y$. Suppose we know an upper bound $\mathcal{R} \in (0, 1/2]$ on $\mathcal{R}_P$, i.e. $\mathcal{R} > \mathcal{R}_P$. Then for $\nu := 2\mathcal{R}$ we have*

$$\Pr{}^*\Big((x_i, y_i) \in (X \times Y)^\infty : \limsup_{n\to\infty} \mathcal{R}_P\big(f^{k,\nu}_{((x_1,y_1),\dots,(x_n,y_n))}\big) \leq \mathcal{R}_P + 2(\mathcal{R} - \mathcal{R}_P)\Big) \ = \ 1 \ .$$

Unfortunately, the dual formulation of (3) has some linear constraints which are difficult to treat algorithmically (cf. [1] and [10, Ch. 7.5]). In particular, standard decomposition methods are not directly applicable. One way to avoid these difficulties is to consider the following, modified optimization problem, instead:

$$\begin{array}{lll} \text{minimize} & \langle w, w\rangle - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i & \text{for } w, \rho, \xi_i \\ \text{subject to} & y_i\langle w, \Phi(x_i)\rangle \geq \rho - \xi_i, & i = 1, \dots, n \\ & \xi_i \geq 0, & i = 1, \dots, n \\ & \rho \geq 0 \ . & \end{array} \tag{4}$$

The advantage of this modification is that its dual problem can be treated easilier since one of the critical constraints vanishes. In particular it is possible to modify existing decomposition techniques for standard SVM's in order to solve the dual problem of (4) efficiently. Unlike the dual problem of standard ν-SVM's the dual problem of (4) also enjoys the property that every $\nu \in [0, 1]$ is feasible. Furthermore, the classifier corresponding to this modified optimization problem have all the above properties of the original algorithm:

Theorem 2 *For the classifier based on the modified problem (4) Theorem 1 and Corollary 1 also hold.*

3 Experiments

The theoretical results of the previous section are of asymptotical nature. Therefore, it is an almost natural question whether they can be applied in practice. In order to answer this question we made several experiments using the datasets of Rätsch (cf. [6]). The training sets of these classification problems are relatively small, i.e. they range from 140 to 1300 samples (cf. Table 1). Thus, they are certainly a real challenge for our theoretical results. Moreover, several excellent classifiers including RBF neural networks, SVM's, kernel Fisher discriminant (KFD) and variants of Adaboost based on RBF neural networks have already been tested on these datasets. Therefore, we have both estimates on the Bayes risks and a comparison to these classifiers.
For simplicity, in our experiments we only considered ν-SVM's with Gaussian RBF kernels $k(x, x') = \exp(-\sigma\|x - x'\|_2^2)$. To solve the corresponding dual problem we used the ν-SVM routine from Lin's and Chang's LIBSVM 2.33 (cf. [2]). We first trained a ν-SVM using cross validation in order to determine

	THYROID	TITANIC	HEART	CANCER	BANANA	RINGNORM	TWONORM	WAVEFORM	DIABETIS	SOLAR	GERMAN	SPLICE	IMAGE
training patterns	140	150	170	200	400	400	400	400	468	666	700	1000	1300
test patterns	75	2051	100	77	900	7000	7000	4600	300	400	300	2175	1010
input dimensions	5	3	13	9	2	20	20	21	8	9	20	60	18

Table 1: Sample sizes of the classification problems

	RBF-Net	AdaBoost	LP-AdaB.	QP-AdaB.	AdaB.-Reg	SVM	KFD	ν-SVM c.v.	ν-SVM ±25%	ν-SVM ±10%	ν-SVM ±0%
THYROID	4.52 ±2.12	4.40 ±2.18	4.59 ±2.22	4.35 ±2.18	4.55 ±2.19	4.80 ±2.19	**4.20** ±2.07	4.56 ±1.87	4.71 ±1.88	4.73 ±2.40	4.61 ±2.40
TITANIC	23.26 ±1.34	22.58 ±1.18	23.98 ±4.38	22.71 ±1.05	22.64 ±1.20	**22.42** ±1.02	23.25 ±2.05	24.17 ±5.05	24.62 ±6.83	23.83 ±3.49	26.05 ±8.55
HEART	17.55 ±3.25	20.29 ±3.44	17.49 ±3.53	17.17 ±3.44	16.47 ±3.51	15.95 ±3.26	16.14 ±3.39	**15.50** ±3.12	15.87 ±3.13	16.16 ±3.04	17.59 ±2.60
CANCER	27.64 ±4.71	30.36 ±4.73	26.79 ±6.08	25.91 ±4.61	26.51 ±4.47	26.04 ±4.74	**24.77** ±4.63	26.27 ±4.56	26.75 ±4.03	26.25 ±4.07	26.01 ±4.26
BANANA	10.76 ±0.42	12.26 ±0.67	10.73 ±0.43	10.90 ±0.46	10.85 ±0.42	11.53 ±0.66	10.75 ±0.45	10.53 ±0.48	**10.48** ±0.47	10.49 ±0.50	10.66 ±0.52
RINGNORM	1.70 ±0.21	1.93 ±0.24	2.24 ±0.46	1.86 ±0.22	1.58 ±0.12	1.66 ±0.12	**1.49** ±0.12	1.68 ±0.15	1.68 ±0.12	1.79 ±0.15	1.76 ±0.13
TWONORM	2.85 ±0.28	3.03 ±0.28	3.17 ±0.43	2.97 ±0.26	2.70 ±0.24	2.96 ±0.23	2.61 ±0.15	**2.46** ±0.14	2.74 ±0.21	2.74 ±0.23	3.23 ±0.30
WAVEFORM	10.66 ±1.08	10.84 ±0.58	10.53 ±1.02	10.07 ±0.51	**9.79** ±0.81	9.88 ±0.43	9.86 ±0.44	10.04 ±0.51	10.26 ±0.40	10.30 ±0.43	10.50 ±0.56
DIABETIS	24.29 ±1.88	26.47 ±2.29	24.11 ±1.90	25.39 ±2.20	23.79 ±1.80	23.53 ±1.73	**23.21** ±1.63	23.43 ±1.60	**23.21** ±1.57	23.28 ±1.81	23.98 ±1.54
SOLAR	34.37 ±1.95	35.7 ±1.79	34.74 ±2.00	36.22 ±1.80	34.2 ±2.18	32.43 ±1.82	33.16 ±1.72	32.34 ±1.79	**32.33** ±1.81	32.34 ±1.80	33.68 ±1.64
GERMAN	24.71 ±2.38	27.45 ±2.50	24.79 ±2.22	25.25 ±2.14	24.34 ±2.08	**23.61** ±2.07	23.71 ±2.20	24.68 ±4.96	23.64 ±2.19	23.68 ±2.13	23.85 ±2.08
SPLICE	9.95 ±0.78	10.14 ±0.51	10.22 ±1.59	10.11 ±0.52	**9.50** ±0.65	10.88 ±0.66	10.52 ±0.64	11.11 ±0.69	11.12 ±0.72	11.21 ±0.56	11.05 ±0.64
IMAGE	3.32 ±0.65	2.73 ±0.66	2.76 ±0.61	**2.67** ±0.63	**2.67** ±0.61	2.96 ±0.60	4.76 ±0.58	3.07 ±0.70	3.13 ±0.63	3.13 ±0.54	3.09 ±0.62

Table 2: Test errors of the classifiers

the free parameters ν and σ. To make our results comparable with those of [6] we followed the cross validation procedure of [7] (also communicated by [8]) which we briefly recall: each of the first five training sets was divided into ten parts. For a parameter pair (ν, σ) the ν-SVM was trained on the union of nine parts and the resulting decision function was tested on the remaining part. This was done for every part and then the average classification error was computed. In order to find the pair (ν, σ) with the best average error a two step strategy was chosen:

- A coarse search, where all pairs of *feasible* values (σ, ν) for $\nu = 0.1 \cdot k - 0.05$, $k = 1, \ldots, 10$ and $\sigma = 5 \cdot 10^k$, $k = -4, \ldots, 2$ were considered.
- A finer search, where all pairs of *feasible* values (σ, ν) for $\nu = \nu_1 + 0.01 \cdot k$, $k = -5, \ldots, 5$ and $\sigma = 0.2 \cdot \sigma_1 \cdot k$, $k = 1, \ldots, 10$ were considered. Here, ν_1 and σ_1 denote the values determined in the first step.

This procedure was done for the first five training sets. Finally the medians of the resulting values $\nu^{(1)}, \ldots, \nu^{(5)}$ and $\sigma^{(1)}, \ldots, \sigma^{(5)}$ were chosen, respectively. This approach is denoted by "ν-SVM c.v." in Tables 2, 3 and 4.
For the following training approaches we assumed that we had some information on the Bayes risk $\mathcal{R}_P$ of the problem: for each classification problem we picked the error rate $\mathcal{R}$ achieved by the respectively best classifiers considered in [6]. Our second method was based on the idea that we only had a vague knowledge on $\mathcal{R}_P$. This was modeled by a coarse search for ν, where only the (feasible) values $2\mathcal{R} + 0.1 \cdot k$, $k = -2, \ldots, 2$ were considered and a fine search for ν, where the (feasible) values $2\mathcal{R} + 0.025 \cdot k$, $k = -2, \ldots, 2$ were considered. The parameter σ were determined analogously to the above description.

	RBF-Net	AdaBoost	LP-AdaB.	QP-AdaB.	AdaB.-Reg	SVM	KFD	ν-SVM c.v.	ν-SVM ±25%	ν-SVM ±10%	ν-SVM ±0%
All	1.01	1.98	1.05	1.01	0.55	0.48	0.46	0.57	0.62	0.58	1.05
Samples > 200	0.79	1.67	0.86	1.10	0.43	0.43	0.50	0.42	0.34	0.38	0.70

Table 3: Average test errors of the classifiers

		Thyroid	Titanic	Heart	Cancer	Banana	Ringnorm	Twonorm	Waveform	Diabetis	Solar	German	Splice	Image
ν	c. v.	0.340	0.5000	0.530	0.4900	0.2300	0.2100	0.3600	0.3500	0.6000	0.7500	0.5700	0.26	0.0700
	$\pm 25\%$	0.134	0.5484	0.469	0.4954	0.2396	0.1298	0.1522	0.2708	0.5892	0.7986	0.5222	0.24	0.0534
	$\pm 10\%$	0.084	0.4984	0.419	0.4954	0.2646	0.0798	0.1522	0.2458	0.5142	0.7486	0.5222	0.14	0.0534
	$\pm 0\%$	0.084	0.4484	0.319	0.4954	0.2146	0.0298	0.0522	0.1958	0.4642	0.6486	0.4722	0.19	0.0534
σ	c. v.	0.70	0.7	0.0005	0.02	0.6	0.06	0.0002	0.02	0.0005	0.0070	0.006	0.009	0.03
	$\pm 25\%$	0.90	0.6	0.0010	0.10	0.6	0.08	0.0080	0.01	0.0060	0.0010	0.005	0.008	0.09
	$\pm 10\%$	0.06	20.0	0.0008	0.07	0.5	0.07	0.0060	0.01	0.0006	0.0100	0.020	0.008	0.60
	$\pm 0\%$	0.10	0.1	0.0010	0.04	0.9	0.08	0.0500	0.04	0.0008	0.0009	0.060	0.009	0.03

Table 4: Used parameters of the different approaches

This approach is denoted by "ν-SVM $\pm 25\%$" in Tables 2, 3 and 4. In our third training approach we assumed that we had a rather precise knowledge on $\mathcal{R}_P$: we only performed a coarse search for ν, where the (feasible) values $2\mathcal{R} + 0.05 \cdot k$, $k = -2, \ldots, 2$ were considered. The fine search for ν was omitted. Again, σ were determined analogously to the above description. This approach is denoted by "ν-SVM $\pm 10\%$" in Tables 2, 3 and 4. In our forth training method we trusted our estimate $\mathcal{R}$ of $\mathcal{R}_P$, i.e. we defined $\nu := 2\mathcal{R}$. The free parameter σ was determined using cross validation analogously to the above description. This approach is denoted by "ν-SVM $\pm 0\%$" in Tables 2, 3 and 4.

Table 2 shows the obtained error rates of [6] and ours. In order to compare the algorithms we picked the respectively best error (bold faced) for every classification problem. Then we computed for each classifier and all problems the difference of the best error rate to its respective error rate. The average differences for each classifier are listed in Table 3. Since it turned out that the ν-SVM's behaved better on the larger training sets (cf. Table 2) we also computed the average differences for the training sets with more than 200 samples (cf. Table 3, again). Finally, the used parameter pairs of the different approaches are listed in Table 4.

Table 3 shows that the ν-SVM's performed worse than the best classifiers SVM and KFD if we consider all training sets. However, considering only the training sets with more than 200 samples the ranking changes: while the ν-SVM trained by cross validation performed like the standard SVM our second and third training approach which incorporated knowledge on the Bayes risk behaved significantly better. Moreover, recall that our second training method was approximately twice faster than standard cross validation and our third method was even approximately four-times faster than standard cross validation. In other words, incorporating prior knowledge on the Bayes risk yielded both better performance and shorter training times. The fourth approach which trusted our estimate on the Bayes risk still worked slightly worse than the standard cross validation ansatz. However, its training time was approximately *twenty-times* shorter than the training time of the standard approach. Therefore, the fourth approach may be interesting if short training times are important. In particular, for large training sets this approach may be a good choice since in this case our theoretical results may have a larger impact on the generalization performance.

4 Conclusions

In this paper we have shown that an asymptotical optimal choice of the regularization parameter ν for ν-SVM's is an arbitrary close upper bound of twice the Bayes risk of the considered optimization problem. Therefore, prior knowledge on this Bayes risk can be used for an effective parameter search. Moreover, we have demonstrated in several experiments that this approach is not only of theoretical nature but also of practical interest. In particular it has turned out that a restricted parameter search for ν which has been based on our theoretical results has performed significantly better and faster than standard cross validation techniques.

On the theoretical side it would be interesting whether our almost sure upper bound on the risk achieved by ν-SVM's (cf. Cor. 1) is also a *lower* bound. This together with the ν-property of [9] would yield a new insight into the sparseness of SVM decision functions.

5 Proofs of the theorems

Before we prove Theorem 1 let us recall that the covering numbers of a metric space (X, d) are defined by

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

for all $\varepsilon > 0$. The space (X, d) is precompact if and only if $\mathcal{N}((X,d),\varepsilon)$ is finite for all $\varepsilon > 0$. We also need the following result which has been proved in [11, Lem. 3]:

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 id $: (X, |.|) \to (X, d_k)$ *is continuous. In particular, $N\big((X,d_k),\varepsilon\big)$ is finite for all $\varepsilon > 0$.*

Proof of Theorem 1: Trivially, we may assume without loss of generality, that $\varepsilon \in (0, 1]$. We define $\delta := \nu - 2\mathcal{R}_P$ and $\tau := \frac{\min\{\delta,\varepsilon\}}{18}$. Furthermore, we fix an integer m with $\frac{1}{2^m} \le \tau \le \frac{1}{2^{m-1}}$. Let

$$\begin{aligned} X_i \ &:= \ \Big\{x \in X \ : \ \frac{i}{2^m} \le s(x) < \frac{i+1}{2^m}\Big\} \ , \qquad\qquad i = 0,\dots,2^{m-1}-2 \\ X_{2^{m-1}-1} \ &:= \ \Big\{x \in X \ : \ \frac{1}{2} - \frac{1}{2^m} \le s(x) \le \frac{1}{2}\Big\} \ , \end{aligned}$$

where $s : X \to \mathbb{R}$ denotes the noise level of P. Recalling equation (1), this definition immediately yields

$$\sum_{i=0}^{2^{m-1}-1} \frac{i}{2^m} P_X(X_i) \ \le \ \mathcal{R}_P \ \le \ \sum_{i=0}^{2^{m-1}-1} \frac{i}{2^m} P_X(X_i) + \tau \ . \tag{5}$$

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

$$\begin{aligned} P_X(X_i^j \setminus \tilde{K}_i^j) \ &\le \ \frac{\tau}{2^m} \ , \qquad\qquad i = 0,\dots,2^{m-1}-2 \ , \ j \in \{-1,1\} \\ P_X(X_{2^{m-1}-1} \setminus \tilde{K}_{2^{m-1}-1}) \ &\le \ \frac{\tau}{2^m} \ . \end{aligned}$$

For later purpose, we write $\tilde{K}^1_{2^{m-1}-1} := \tilde{K}_{2^{m-1}-1} \cap \big(B_1(P) \cup B_0(P)\big)$ and $\tilde{K}^{-1}_{2^{m-1}-1} := \tilde{K}_{2^{m-1}-1} \cap B_{-1}(P)$. 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

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

hold and $\langle w^*, \Phi(.)\rangle$ only takes values between $-(1+\tau)$ and $1+\tau$. We define

$$\sigma := \frac{\tau(\nu - 2\mathcal{R}_P - 17\tau)}{2\|w^*\|_H^2} .$$

Then for every $i = 0, \dots, 2^{m-1}-1$ and $j \in \{-1,1\}$ there exists a finite partition $\tilde{\mathcal{A}}_i^j$ of $\tilde{K}_i^j$ such that each $A \in \tilde{\mathcal{A}}_i^j$ 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 $|\tilde{\mathcal{A}}_i^j| \leq \mathcal{N}\big((X, d_k), \sigma\big)$. We define $M := \frac{2}{\tau}\mathcal{N}\big((X,d_k),\sigma\big)$ and

$$\mathcal{A}_i^j := \Big\{ A \in \tilde{\mathcal{A}}_i^j \ : \ P_X(A) \geq \frac{2\tau}{M} \Big\} .$$

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

$$\begin{aligned}
F_{n,A}^+ &:= \Big\{\big((x_1,y_1),\dots,(x_n,y_n)\big) : \big|\{l : x_l \in A, y_l = j\}\big| \geq (1-\tau)\Big(1 - \frac{i+1}{2^m}\Big) P_X(A)\, n\Big\} \\
F_{n,A}^- &:= \Big\{ \big((x_1,y_1),\dots,(x_n,y_n)\big) : \big|\{l : x_l \in A, y_l \neq j\}\big| \geq (1-\tau)\frac{i}{2^m} P_X(A)\, n\Big\} ,
\end{aligned}$$

where $n \geq 1$, $i = 0, \dots, 2^{m-1}-2$, $j \in \{-1,1\}$ and $A \in \mathcal{A}_i^j$. Moreover, for $A \in \mathcal{A}_{2^{m-1}-1}^j$, $j \in \{-1,1\}$ let

$$\begin{aligned}
F_{n,A}^+ &:= \Big\{ \big((x_1,y_1),\dots,(x_n,y_n)\big) : \big|\{l : x_l \in A, y_l = 1\}\big| \geq (1-\tau)\Big(\frac{1}{2} - \frac{1}{2^m}\Big) P_X(A)\, n\Big\} \\
F_{n,A}^- &:= \Big\{ \big((x_1,y_1),\dots,(x_n,y_n)\big) : \big|\{l : x_l \in A, y_l = -1\}\big| \geq (1-\tau)\Big(\frac{1}{2} - \frac{1}{2^m}\Big) P_X(A)\, n\Big\}.
\end{aligned}$$

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

$$F_n := \bigcap_{\substack{i=0 \\ j\in\{-1,1\}}}^{2^{m-1}-1} \ \bigcap_{A \in \mathcal{A}_i^j} \big(F_{n,A}^+ \cap F_{n,A}^-\big) .$$

By Lemma 3 there exists a constant $c > 0$ with $P^n(F_n) \geq 1 - e^{-cn}$ for all $n \geq 1$. Therefore it suffices to show that $\mathcal{R}_P(f_T) \leq \nu - \mathcal{R}_P + \varepsilon$ holds for all $T \in F_n$. Let us assume the converse, i.e. there exists a training set $T = ((x_1,y_1),\dots,(x_n,y_n)) \in F_n$ with

$$\mathcal{R}_P(f_T) \ > \ \nu - \mathcal{R}_P + \varepsilon . \tag{6}$$

Then for $i = 0, \dots, 2^{m-1}-2$ and $j \in \{-1,1\}$ we denote the set of misclassified points (compared with the optimal Bayes decision rule) in X_i^j by E_i^j, i.e.

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

Analogously, let

$$E_{2^{m-1}-1}^j := \big\{ x \in X_{2^{m-1}-1} \cap B_j(P) \ : \ f_T(x) \neq j \big\}$$

and $E := \bigcup_{i=0}^{2^{m-1}-1} (E_i^{-1} \cup E_i^1)$. Since we know by Lemma 4 that

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

holds, our assumption (6) yields

$$\int_E (1-2s)\, dP_X \;>\; \nu - 2\mathcal{R}_P + \varepsilon \; . \tag{7}$$

Now let us denote the slack variables, which correspond to a fixed solution (w_T, b_T, ρ_T) of our optimization problem (3), by $\xi_1, \ldots, \xi_n$. Then Lemma 5 yields

$$\langle w_T, w_T\rangle - \nu\rho_T + \frac{1}{n}\sum_{l=1}^{n} \xi_l \;<\; 0 \; . \tag{8}$$

On the other hand, by Lemma 7 and inequality (7) we obtain

$$\frac{1}{n}\sum_{l=1}^{n} \xi_l \;\geq\; \rho_T \left(2\,\mathcal{R}_P + \int_E (1-2s) dP_X - 15\tau\right) \;>\; \rho_T\;(\nu + \varepsilon - 15\tau)\; .$$

Therefore our assumption (6) must be false since inequality (8) yields

$$0 \;>\; \langle w_T, w_T\rangle - \nu\rho_T + \frac{1}{n}\sum_{l=1}^{n} \xi_l \;\geq\; -\nu\rho_T + \rho_T\;(\nu + \varepsilon - 15\tau) \;\geq\; 0. \;◄$$

Proof of Corollary 1: The assertion is a direct consequence of the tail bound of Theorem 1.◄

Proof of Theorem 2: Note, that only in Lemmas 5, 6 and 7 the function F_T is considered. Since in all cases we compare the solution of (3) with a vector with $b = 0$ the assertion of Theorem 1 remains true for the modified classifier. ◄

6 Proofs of the lemmas

In this section we show the lemmas used in the proof of Theorem 1. Readers familiar with the techniques of [12] may skip most of the lemmas. In particular this holds for Lemmas 2, 3, 4 and 7 as well as for almost all of the proof of Lemma 5.
We begin with the following result which is needed to construct an almost optimal decision function. It is a direct consequence of Urysohn's Lemma and the definition of universal kernels. For a proof we refer to [11, Prop. 5] and [12, Lem. 2]:

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*

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

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 there exists a constant $c > 0$ independent of n with*

$$P^n(F_n) \geq 1 - e^{-cn}\; .$$

Proof: Analogously to the proof of Lemma 3 in [12] we easily find by Hoeffding's inequality (cf. [4, Th. 8.1]) that

$$P^n(F^{\pm}_{n,A}) \;\geq\; 1 - e^{-2\frac{\tau^6}{M^2}n}$$

holds for all $n \geq 1$ and $A \in \mathcal{A}^j_i$, $i = 0, \ldots, 2^{m-1}-1$, $j \in \{-1, 1\}$. Since M and τ do not depend on n the assertion follows. ◀

In the following lemma the risk of a decision function is computed:

Lemma 4 *With the notations of the proof of Theorem 1 we have*

$$\mathcal{R}_P(f_T) \;=\; \mathcal{R}_P + \int_E (1-2s)dP_X \ .$$

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

$$\begin{aligned}
\int_{B_1(P)} \mathbf{1}_{\{f_T(x)\neq y\}}\ P(dx,dy) &= \int_{B_1(P)} \Big(\mathbf{1}_{\{f_T(x)\neq -1\}} P(y=-1|x) + \mathbf{1}_{\{f_T(x)\neq 1\}} P(y=1|x)\Big) P_X(dx) \\
&= \int_{B_1(P)\backslash E_1} P(y=-1|x)\, P_X(dx) + \int_{E_1} P(y=1|x)\, P_X(dx) \\
&= \int_{B_1(P)} s(x)\, P_X(dx) + \int_{E_1} (1-2s(x))\, P_X(dx)
\end{aligned}$$

holds. Since we obtain an analogous result for $B_{-1}(P)$ the assertion follows.◀

The next lemma provides an estimate for the value of the optimization problem (3) from above:

Lemma 5 *With the notations of the proof of Theorem 1 we have*

$$\langle w_T, w_T\rangle - \nu\rho_T + \frac{1}{n}\sum_{l=1}^n \xi_l \;\leq\; -\frac{(\nu - 2\mathcal{R}_P - 17\tau)^2}{4\langle w^*, w^*\rangle} \ .$$

Proof: Firstly, we will compare the value of the optimization problem (3) with the value of the objective function in $(\rho w^*, 0, \rho, \xi^*)$. Thus we have to construct an admissible slack variable ξ^* corresponding to ρw^*, 0 and $\rho > 0$. For this let (x_l, y_l) be a sample of T. If $x_l \in K^1_i$ 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 $x_l \in K^{-1}_i$ for some $i = 0, \ldots, 2^{m-1}-2$ and $y_l = -1$. Conversely, if $x_l \in K^1_i$ 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 $\xi^*_l := \rho(2+\tau)$. Again, if $x_l \in K^{-1}_i$ for some $i = 0, \ldots, 2^{m-1}-2$ and $y_l = 1$ we analogously define $\xi^*_l := \rho(2+\tau)$. 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 $\xi^*_l := \rho(1+\tau)$. 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^j_i, $i = 0, \ldots, 2^{m-1}-2$, $j \in \{-1, 1\}$ nor an element of $K_{2^{m-1}-1}$

we obtain $|\langle w^*, \Phi(x_l)\rangle| \leq 1+\tau$. Thus let $\xi_l^* := \rho(2+\tau)$ in this case. For brevity's sake we now define

$$\begin{aligned}
a_1 &:= \Big|\Big\{l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^1,\ y_l = 1\Big\}\Big| + \Big|\Big\{l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^{-1},\ y_l = -1\Big\}\Big| \\
a_2 &:= \Big|\Big\{l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^1,\ y_l = -1\Big\}\Big| + \Big|\Big\{l : x_l \in \bigcup_{i=0}^{2^{m-1}-2} K_i^{-1},\ y_l = 1\Big\}\Big| \\
a_3 &:= \big|\{l : x_l \in K_{2^{m-1}-1}\}\big| \\
a_4 &:= \Big|\Big\{l : x_l \notin K_{2^{m-1}-1} \cup \bigcup_{i=0}^{2^{m-1}-2} (K_i^1 \cup K_i^{-1})\Big\}\Big| \ .
\end{aligned}$$

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

$$\begin{aligned}
\langle w_T, w_T\rangle - \nu\rho_T + \frac{1}{n}\sum_{l=1}^n \xi_l &\leq \rho^2\langle w^*, w^*\rangle - \nu\rho + \frac{1}{n}\sum_{l=1}^n \xi_l^* \\
&\leq \rho^2\langle w^*, w^*\rangle - \nu\rho + \frac{\rho}{n}\Big((2+\tau)a_2 + (1+\tau)a_3 + (2+\tau)a_4\Big) \\
&= \langle \rho^2 w^*, w^*\rangle - \nu\rho + \frac{\rho}{n}\Big((2+\tau)(n-a_1) - a_3\Big) \qquad (9)
\end{aligned}$$

since (w_T, b_T, ρ_T) together with the corresponding slack variables $\xi_1,\ldots,\xi_n$ is a solution of problem (3). Furthermore, the construction of F_n implies

$$\begin{aligned}
\frac{2+\tau}{n}\big(n-a_1\big) &\leq (2+\tau)\Bigg(1 - \sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2} \sum_{A\in\mathcal{A}_i^j} (1-\tau)\Big(1-\frac{i+1}{2^m}\Big)P_X(A)\Bigg) \\
&\leq 2 - 2\,(1-\tau)\sum_{i=0}^{2^{m-1}-2}\Big(1-\frac{i+1}{2^m}\Big)P_X\big(\tilde K_i^1 \cup \tilde K_i^{-1}\big) + 5\tau \\
&= 2(1-\tau)\Bigg(1 - \sum_{i=0}^{2^{m-1}-2}\Big(1-\frac{i+1}{2^m}\Big)P_X\big(\tilde K_i^1 \cup \tilde K_i^{-1}\big)\Bigg) + 7\tau \\
&\leq 2(1-\tau)\Bigg(1 - \sum_{i=0}^{2^{m-1}-2} P_X\big(\tilde K_i^1 \cup \tilde K_i^{-1}\big) + \sum_{i=0}^{2^{m-1}-2}\frac{i}{2^m}P_X\big(\tilde K_i^1 \cup \tilde K_i^{-1}\big)\Bigg) + 9\tau \ .
\end{aligned}$$

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

$$\begin{aligned}
\frac{a_3}{n} &\geq 2\sum_{\substack{A\in\mathcal{A}_{2^{m-1}-1}^j\\ j\in\{-1,1\}}} (1-\tau)\Big(\frac{1}{2}-\frac{1}{2^m}\Big)P_X(A) \\
&\geq 2\,(1-\tau)\Big(\frac{1}{2}-\frac{1}{2^m}\Big)\big(P_X(\tilde K_{2^{m-1}-1}) - 2\tau\big) \\
&\geq 2\,(1-\tau)\Big(P_X(\tilde K_{2^{m-1}-1}) - \Big(\frac{1}{2}-\frac{1}{2^m}\Big)P_X(\tilde K_{2^{m-1}-1})\Big) - 6\tau \ .
\end{aligned}$$

If we combine these estimates with inequality (5) we thus obtain

$$\begin{aligned}
\frac{1}{n}\Big((2+\tau)(n-a_1)-a_3\Big) &\le 2(1-\tau)\Big(1-\sum_{i=0}^{2^{m-1}-1} P_X\big(\tilde K_i^1\cup\tilde K_i^{-1}\big)+\sum_{i=0}^{2^{m-1}-1}\frac{i}{2^m}P_X\big(\tilde K_i^1\cup\tilde K_i^{-1}\big)\Big)+15\tau \\
&\le 2(1-\tau)\Big(\tau+\sum_{i=0}^{2^{m-1}-1}\frac{i}{2^m}P_X(X_i)\Big)+15\tau \\
&\le 2(1-\tau)(\tau+\mathcal{R}_P)+15\tau \\
&\le 2\mathcal{R}_P+17\tau\ .
\end{aligned}$$

Thus we may continue estimate (9) to

$$\langle w_T,w_T\rangle-\nu\rho_T+\frac{1}{n}\sum_{l=1}^n\xi_l\ \le\ \rho^2\langle w^*,w^*\rangle-\nu\rho+2\rho\mathcal{R}_P+17\rho\tau\ .$$

Since this holds for all $\rho>0$ we may minimize the right side with respect to ρ. An easy computation shows that this yields the assertion. ◀

Lemma 6 *With the notations of the proof of Theorem 1 we have*

$$\|w_T\|\sigma\ \le\ \rho_T\tau\ .$$

Proof: By Lemma 5 we know that

$$-\nu\rho_T\ \le\ F_T(w_T,b_T,\rho_T,\xi_T)\ \le\ -\frac{(\nu-2\mathcal{R}_P-17\tau)^2}{4\langle w^*,w^*\rangle}$$

holds and thus we find

$$\rho_T\ \ge\ \nu\rho_T\ \ge\ \frac{(\nu-2\mathcal{R}_P-17\tau)^2}{4\langle w^*,w^*\rangle}\ . \tag{10}$$

Moreover, comparing $F_T(w_T,b_T,\rho_T,\xi_T)$ with $F_T\big(0,0,\rho_T,(\rho_T,\dots,\rho_T)\big)$ yields

$$\langle w_T,w_T\rangle-\nu\rho_T\ \le\ \langle w_T,w_T\rangle-\nu\rho_T+\frac{1}{n}\sum_{l=1}^n\xi_l\ \le\ -\nu\rho_T+\rho_T$$

and thus we also have $\langle w_T,w_T\rangle\le\rho_T$. This together with inequality (10) implies

$$\sigma\ =\ \frac{\tau(\nu-2\mathcal{R}_P-17\tau)}{2\|w^*\|}\ \le\ \tau\sqrt{\rho_T}\ \le\ \tau\frac{\rho_T}{\|w_T\|}$$

by the definition of σ and therefore the assertion follows.◀

The last lemma estimates the value of the optimization problem (3) from below:

Lemma 7 *With the notations of the proof of Theorem 1 we have*

$$\frac{1}{n}\sum_{l=1}^n\xi_l\ \ge\ \rho_T\ \Big(2\,\mathcal{R}_P+\int_E(1-2s)dP_X-15\tau\Big)\ .$$

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

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

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

Now, our first goal is to show that

$$\frac{1}{n} \sum_{l \in I_i^j} \xi_l \geq \rho_T (1-\tau)^2 \left(1 - \frac{1}{2^m}\right) \sum_{\substack{A \in \mathcal{A}_i^j \\ A \cap E_i^j \neq \emptyset}} P_X(A) \tag{11}$$

$$\frac{1}{n} \sum_{l \in J_i^j} \xi_l \geq \rho_T (1-\tau)^2 \, \frac{i}{2^{m-1}} \sum_{\substack{A \in \mathcal{A}_i^j \\ A \cap E_i^j = \emptyset}} P_X(A) \tag{12}$$

$$\frac{1}{n} \sum_{\substack{A \in \mathcal{A}^{\pm 1}_{2^{m-1}-1} \\ x_l \in A}} \xi_l \geq \rho_T (1-\tau)^2 \left(1 - \frac{1}{2^{m-1}}\right) \big(P_X(\tilde{K}_{2^{m-1}-1}) - 2\tau\big) \tag{13}$$

hold for all $i = 0, \dots, 2^{m-1} - 2$ and $j \in \{-1, 1\}$. To show inequality (11) let $A \in \mathcal{A}_i^j$ with $A \cap E_i^j \neq \emptyset$. Then for fixed $z \in A \cap E_i^j$ we define $a := -\big(\langle w_T, \Phi(z) \rangle + b_T\big)$. Without loss of generality we may assume $j = 1$, i.e. $a \geq 0$. Now, for an index l with $x_l \in A$ and $y_l = 1$ we have $\|\Phi(x_l) - \Phi(z)\| = d_k(x_l, z) \leq \sigma$. By Lemma 6 this yields

$$\begin{aligned}
\rho_T - \xi_l &\leq \langle w_T, \Phi(x_l) \rangle + b_T \\
&= \big\langle w_T, \Phi(x_l) - \Phi(z) \big\rangle + \langle w_T, \Phi(z) \rangle + b_T \\
&\leq \|w_T\| \cdot \|\Phi(x_l) - \Phi(z)\| - a \\
&\leq \rho_T \tau - a \,,
\end{aligned}$$

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

$$\rho_T - \xi_l \leq -\big(\langle w_T, \Phi(x_l) \rangle + b_T\big) = -\big\langle w_T, \Phi(x_l) - \Phi(z) \big\rangle - \big(\langle w_T, \Phi(z) \rangle + b_T\big) \leq \rho_T \tau + a \,,$$

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

$$\begin{aligned}
\frac{1}{n} \sum_{x_l \in A} \xi_l &\geq \big(\rho_T(1-\tau) + a\big)(1-\tau)\Big(1 - \frac{i+1}{2^m}\Big) P_X(A) + \big(\rho_T(1-\tau) - a\big)(1-\tau) \, \frac{i}{2^m} \, P_X(A) \\
&= \rho_T(1-\tau)^2 \Big(1 - \frac{1}{2^m}\Big) P_X(A) + a \, (1-\tau) \Big(1 - \frac{2i+1}{2^m}\Big) P_X(A) \\
&\geq \rho_T(1-\tau)^2 \Big(1 - \frac{1}{2^m}\Big) P_X(A)
\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{n} \sum_{x_l \in A} \xi_l &\geq \big(\rho_T(1-\tau) + a\big)(1-\tau)\Big(1 - \frac{i+1}{2^m}\Big) P_X(A) \\
&> \rho_T(1-\tau)^2 \Big(2 - \frac{2i+2}{2^m}\Big) P_X(A) \\
&> \rho_T(1-\tau)^2 \Big(1 - \frac{1}{2^m}\Big) P_X(A) \,.
\end{aligned}$$

Thus we finally obtain

$$\frac{1}{n}\sum_{l\in I_i^j}\xi_l \;=\; \frac{1}{n}\sum_{\substack{A\in\mathcal{A}_i^j\\ A\cap E_i^j\neq\emptyset}}\;\sum_{x_l\in A}\xi_l \;\geq\; \rho_T(1-\tau)^2\Bigl(1-\frac{1}{2^m}\Bigr)\sum_{\substack{A\in\mathcal{A}_i^j\\ A\cap E_i^j\neq\emptyset}} P_X(A)\ .$$

Now we prove inequality (12). For this let $A\in\mathcal{A}_i^j$ with $A\cap E_i^j=\emptyset$. Then for fixed $z\in A\cap(X\setminus E_i^j)$ we define $a:=-\bigl(\langle w_T,\Phi(z)\rangle+b_T\bigr)$. Without loss of generality we may assume $j=-1$, i.e. $a\geq 0$. For an index l with $x_l\in A$ and $y_l=-1$ we thus obtain

$$\rho_T-\xi_l\;\leq\;-\bigl(\langle w_T,\Phi(x_l)\rangle+b_T\bigr)\;=\;-\bigl\langle w_T,\Phi(x_l)-\Phi(z)\bigr\rangle-\bigl(\langle w_T,\Phi(z)\rangle+b_T\bigr)\;\leq\;\rho_T\tau+a\ ,$$

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

$$\rho_T-\xi_l\;\leq\;\langle w_T,\Phi(x_l)\rangle+b_T\;=\;\bigl\langle w_T,\Phi(x_l)-\Phi(z)\bigr\rangle+\langle w_T,\Phi(z)\rangle+b_T\;\leq\;\rho_T\tau-a\ ,$$

i.e. $\xi_l\geq\rho_T(1-\tau)+a>0$. Let us suppose that $\rho_T(1-\tau)-a\geq 0$. Then by the definition of F_n we get

$$\begin{aligned}
\frac{1}{n}\sum_{x_l\in A}\xi_l \;&\geq\; \bigl(\rho_T(1-\tau)-a\bigr)(1-\tau)\Bigl(1-\frac{i+1}{2^m}\Bigr)P_X(A)+\bigl(\rho_T(1-\tau)+a\bigr)(1-\tau)\;\frac{i}{2^m}P_X(A)\\
&=\; \rho_T(1-\tau)^2\Bigl(1-\frac{1}{2^m}\Bigr)P_X(A)-a\;(1-\tau)\Bigl(1-\frac{2i+1}{2^m}\Bigr)P_X(A)\\
&\geq\; \rho_T(1-\tau)^2\;\frac{i}{2^{m-1}}\;P_X(A)
\end{aligned}$$

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

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

Therefore, we obtain

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

Now we treat inequality (13). For this let $A\in\mathcal{A}^{\pm1}_{2^{m-1}-1}$ and fix $z\in A$. Moreover, we define $a:=-\bigl(\langle w_T,\Phi(z)\rangle+b_T\bigr)$. Suppose that we have an index l with $x_l\in A$ and $y_l=-1$. Then we obtain

$$\rho_T-\xi_l\;\leq\;-\bigl(\langle w_T,\Phi(x_l)\rangle+b_T\bigr)\;=\;-\bigl\langle w_T,\Phi(x_l)-\Phi(z)\bigr\rangle-\bigl(\langle w_T,\Phi(z)\rangle+b_T\bigr)\;\leq\;\rho_T\tau+a\ ,$$

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

$$\begin{aligned}
\frac{1}{n}\sum_{\substack{A\in\mathcal{A}^{\pm1}_{2^{m-1}-1}\\ x_l\in A}}\xi_l \;&\geq\; \Bigl(\bigl(\rho_T(1-\tau)-a\bigr)(1-\tau)\Bigl(\frac{1}{2}-\frac{1}{2^m}\Bigr)+\bigl(\rho_T(1-\tau)+a\bigr)(1-\tau)\Bigl(\frac{1}{2}-\frac{1}{2^m}\Bigr)\Bigr)\;P_X(K_{2^{m-1}-1})\\
&\geq\; \rho_T(1-\tau)^2\Bigl(1-\frac{1}{2^{m-1}}\Bigr)\bigl(P_X(\tilde{K}_{2^{m-1}-1})-2\tau\bigr)\ .
\end{aligned}$$

On the other hand, if $a > \rho_T(1-\tau)$ we have $\rho_T(1-\tau)+a > \rho_T(2-2\tau)$ and therefore inequality (13) also follows. Finally, for $a < -\rho_T(1-\tau)$ we get $\rho_T(1-\tau)-a > \rho_T(2-2\tau)$ and thus we obtain inequality (13) in this case, too. Having proved (11), (12) and (13) we may now estimate

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{n}\xi_l &\geq \rho_T(1-\tau)^2\Bigg(\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\bigg(\Big(1-\frac{1}{2^m}\Big)\sum_{\substack{A\in\mathcal{A}_i^j\\ A\cap E_i^j\neq\emptyset}}P_X(A)+\frac{2i}{2^m}\sum_{\substack{A\in\mathcal{A}_i^j\\ A\cap E_i^j=\emptyset}}P_X(A)\bigg)+\Big(1-\frac{2}{2^m}\Big)P_X(\tilde{K}_{2^{m-1}-1})-2\tau\Bigg)\\
&=\rho_T(1-\tau)^2\Bigg(\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\bigg(\Big(1-\frac{2i+1}{2^m}\Big)\sum_{\substack{A\in\mathcal{A}_i^j\\ A\cap E_i^j\neq\emptyset}}P_X(A)+\frac{2i}{2^m}\sum_{A\in\mathcal{A}_i^j}P_X(A)\bigg)+\Big(1-\frac{2}{2^m}\Big)P_X(\tilde{K}_{2^{m-1}-1})-2\tau\Bigg)\\
&\geq \rho_T(1-\tau)^2\Bigg(\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\bigg(\Big(1-\frac{2i+1}{2^m}\Big)\Big(P_X(E_i^j)-\frac{\tau}{2^m}\Big)+\frac{2i}{2^m}P_X(\tilde{K}_i^j)\bigg)+\Big(1-\frac{2}{2^m}\Big)P_X(\tilde{K}_{2^{m-1}-1})-4\tau\Bigg).
\end{aligned}$$

Moreover, since inequality (5) we have

$$\begin{aligned}
&\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\frac{2i}{2^m}P_X(\tilde{K}_i^j)+\Big(1-\frac{2}{2^m}\Big)P_X(\tilde{K}_{2^{m-1}-1})\\
\geq\ &\sum_{i=0}^{2^{m-1}-2}\frac{2i}{2^m}\Big(P_X(X_i)-\frac{2\tau}{2^m}\Big)+\Big(1-\frac{2}{2^m}\Big)\Big(P_X(X_{2^{m-1}-1})-\frac{2\tau}{2^m}\Big)\\
=\ &\sum_{i=0}^{2^{m-1}-1}\frac{2i}{2^m}P_X(X_i)-\sum_{i=0}^{2^{m-1}-1}\frac{2i}{2^m}\frac{2\tau}{2^m}\\
\geq\ & 2\,\mathcal{R}_P-3\tau
\end{aligned}$$

and thus we may continue the above estimate to

$$\frac{1}{n}\sum_{l=1}^{n}\xi_l \ \geq\ \rho_T(1-\tau)^2\Bigg(\sum_{\substack{i=0\\ j\in\{-1,1\}}}^{2^{m-1}-2}\Big(1-\frac{2i+1}{2^m}\Big)\Big(P_X(E_i^j)-\frac{\tau}{2^m}\Big)+2\,\mathcal{R}_P-7\tau\Bigg). \tag{14}$$

Furthermore, we also get

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

and therefore inequality (14) and the definition of the X_i's yield

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{n}\xi_l &\geq \rho_T(1-\tau)^2\left(2\,\mathcal{R}_P+\sum_{i=0}^{2^{m-1}-2}\Big(1-\frac{i}{2^{m-1}}\Big)P_X\big(E_i^1\cup E_i^{-1}\big)-9\tau\right)\\
&\geq \rho_T(1-\tau)^2\left(2\,\mathcal{R}_P+\sum_{i=0}^{2^{m-1}-1}\Big(1-\frac{i}{2^{m-1}}\Big)P_X\big(E_i^1\cup E_i^{-1}\big)-11\tau\right)\\
&\geq \rho_T(1-2\tau)\left(2\,\mathcal{R}_P+\int_E(1-2s)dP_X-11\tau\right)\\
&\geq \rho_T\left(2\,\mathcal{R}_P+\int_E(1-2s)dP_X-15\tau\right).\blacktriangleleft
\end{aligned}$$

References

[1] C.-C. Chang and C.-J. Lin, Training nu-support vector classifiers: theory and algorithms, *Neural Computation* **13** (2001), 2119-2147

[2] C.-C. Chang and C.-J. Lin, LIBSVM 2.33, `http://www.csie.ntu.edu.tw/~cjlin/`

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

[4] L. Devroye, L. Györfi and G. Lugosi, "A Probabilistic Theory of Pattern Recognition", Springer, New York, 1997

[5] R.M. Dudley, A course on empirical processes, *Lecture Notes in Math.* **1097** (1984), 1-142

[6] G. Rätsch, `http://ida.first.gmd.de/~raetsch/data/benchmarks.htm`

[7] G. Rätsch, T. Onoda and K.-R. Müller, Soft margins for AdaBoost, *Machine Learning* **42** (2001), 287-320

[8] G. Rätsch, Oral communication

[9] B. Schölkopf, A.J. Smola, R.C. Williamson and P.L. Bartlett, New support vector algorithms, *Neural Computation* **12** (2000), 1207-1245

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

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

[12] I. Steinwart, Support vector machines are universally consistent, *J. Complexity*, accepted