Which data-dependent bounds are suitable for SVM's?*

Ingo Steinwart[1]

[1] Friedrich-Schiller-Universität, 07743 Jena, Germany,
steinwart@minet.uni-jena.de

Abstract. Data-dependent, margin-based bounds play an important role in the theoretical analysis of support vector machines (SVM's) for classification. Although existing bounds are usually too large for real-world sample sizes it is often claimed that at least for large sample sizes these bounds can justify the support vector machine approach. The aim of this work is to investigate whether such claims are well-founded. In doing so, it turns out that in the presence of noise many of the known bounds cannot predict well—neither for small nor for large sample sizes. Our analysis is mainly based on two results that describe the margin deviation of SVM decision functions on the training set.
These results are also used for a consideration on sparsity-based bounds. Here we demonstrate that one of the sharpest known bounds cannot be used to explain the generalization performance of SVM's, neither.

1 Introduction

Given a training set T and a decision function f_T constructed by a classifier $\mathcal{C}$ one of the most important questions is how accurately f_T fits to the underlying probability measure P from which T was drawn. Unfortunately, a uniform estimate of the risk of f_T cannot exists in general. To overcome this problem one often tries to collect some additional information during or after the training process about how f_T fits to the *training set* T. Typical examples of such measures are training errors, margin errors, margin deviations, or sparseness of f_T. Once the information is collected it is used in specific estimates on the risk of f_T. If we are lucky in the sense that f_T fits well to T with respect to our measure the used estimates reflect this by predicting a small risk of f_T. The underlying hope of this approach—the so-called luckiness framework (cf. [12])—is that most of the considered classification problems have a benign distribution and this can be observed by the behaviour of f_T on T for a suitable measure, the so-called luckiness function.

Among the most successful algorithms in recent years are support vector machines (SVM's) which are based on a (regularized) maximization of margins in a Hilbert space. Thus, it seems natural to use the margin distribution of f_T on T in order to bound the risk of f_T. Typical estimates of this type can be found e.g. in [11] and [6, Ch. 4]. Because of the intimate relation between the optimization problem on the one hand and the structure of these estimates on the other hand these results have conversely been used to explain the generalization ability of SVM's (cf. [6, Ch. 4]). Using recent results on the asymptotic behavior of the regularized risks optimized by SVM's as well as some simple but important examples we show that this explanation cannot work in almost any interesting case. One of the reasons of this phenomenon is that the known bounds do not reflect the margin distribution which SVM solutions tend to have. Moreover, we also consider a bound which is based on the fraction of γ-margin errors. Here we show that this bound cannot

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

justify SVM's which approximately try to achieve a target margin for a large subset of samples, neither. This consideration is based on a deep result on the asymptotic behaviour of SVM decision functions (cf. Theorem 1) which is also interesting in its own.

Since the decision function of an SVM only depends on the support vectors, SVM's can also be interpreted as a compression scheme. This has been used to bound the expected error of SVM decision functions in terms of their sparseness (cf. e.g. [6, Ch. 4] and [15]). Using the asymptotic results we mentioned above we show that recently found bounds always tend to predict badly for SVM's and noisy classification problems.

2 Preliminaries

In the following (X, τ) is a compact topological T_2-space with countable basis. Recall, that closed, bounded subsets of $\mathbb{R}^d$ are typical examples for such spaces. Moreover, let $Y := \{-1, 1\}$. This set is always assumed to be equipped with the discrete topology.

For a (positive definite) kernel $k : X \times X \to \mathbb{R}$ we denote the corresponding RKHS (cf. [4, Ch. 3]) by H_k or simply by H. We write B_H for its closed unit ball. Recall, that the map $\Phi : X \to H$, $x \mapsto k(x, .)$ fulfills $k(.,.) = \langle \Phi(.), \Phi(.) \rangle_H$. We will sometimes use the quantity $K := \sup\{\sqrt{k(x,x)} : x \in X\}$. Note, that KB_H is the smallest ball in H centered at the origin that contains the image of X under Φ. Moreover, k is continuous if and only if Φ is. In this case H can be continuously embedded into the space of all continuous functions $C(X)$ via $Iw := \langle w, \Phi(.) \rangle_H$. Since we only consider continuous kernels we often identify elements of H as continuous functions on X. If the embedding $I : H \to C(X)$ has a dense image we call k a *universal* kernel (cf. [16, Sect. 3]). In order to consider smooth kernels on bounded C^∞-domains $X \subset \mathbb{R}^d$ (cf. [20, Sect. 3.2]) we write $C^{\infty,\infty}(X \times X)$ for the space of all functions $f : X \times X \to \mathbb{R}$ for which

$$\frac{\partial^{|\alpha_1|+|\alpha_2|}}{\partial_{\alpha_1}\partial_{\alpha_2}} f$$

exists and is continuous for all $\alpha_1, \alpha_2 \in \mathbb{N}_0^d$ (cf. [10, p.40]).

For a given Borel probability measure P on $X \times Y$ there exists a map $x \mapsto P(\,.\,|x)$ from X into the set of all probability measures on Y such that P is the joint distribution of $(P(\,.\,|x))_x$ and of the marginal distribution P_X of P on X (cf. [8, Lem. 1.2.1.]). We call $P(\,.\,|.)$, which is in fact a regular conditional probability, the *supervisor*. Moreover, we often need the noise level $s(x) := \min\{P(1|x), P(-1|x)\}$ of P which describes how noisy the output of the supervisor is.

A *classifier* is an algorithm that constructs to every *training set* $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ a *decision function* $f_T : X \to Y$. In our context it is always assumed that T is i.i.d. according to P, which itself is unknown. Then the decision function $f_T : X \to Y$ constructed by the classifier should guarantee a small probability for the misclassification of an example (x, y) randomly generated according to P. Here, misclassification means $f(x) \neq y$. To make this precise, the risk of a measurable function $f : X \to Y$ is defined by

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

If f is a real valued measurable function we write $\mathcal{R}_P(f) := \mathcal{R}_P(\operatorname{sign} \circ f)$ for short. The smallest achievable risk $\mathcal{R}_P := \inf\{\mathcal{R}_P(f) : f : X \to Y \text{ measurable}\}$ is called *(optimal) Bayes risk.*

In order to state bounds on $\mathcal{R}_P(f_T)$ we need some further risk notions: for $\gamma > 0$ and $t \in \mathbb{R}$ the γ-loss function is defined by $L_\gamma(t) := \max\{0, \gamma - t\}$. Given

$p \in \{0,1,2\}$ and a Borel probability measure P on $X \times Y$ we define the (p,γ)-risk of a measurable function $f : X \to \mathbb{R}$ by

$$\mathcal{R}_{p,\gamma,P}(f) \;:=\; \mathbb{E}_{(x,y)\sim P} L_\gamma^p(yf(x)) \;,$$

where $0^0 := 0$. The smallest (p,γ)-risk is denoted by $\mathcal{R}_{p,\gamma,P} := \inf\{\mathcal{R}_{p,\gamma,P}(f) \mid f : X \to \mathbb{R} \text{ measurable}\}$. For $p \in \{1,2\}$, $\gamma > 0$ and $\alpha \in [0,1]$ we also define

$$f^*_{p,\gamma}(\alpha) \;:=\; \arg\min_{t\in\mathbb{R}} \alpha L_\gamma^p(1,t) + (1-\alpha) L_\gamma^p(-1,t) \;. \tag{1}$$

Note, that $f^*_{p,\gamma}(\alpha)$ is uniquely determined for all $\alpha \notin \{0, 1/2, 1\}$. In particular we have $f^*_{1,\gamma}(\alpha) = -\gamma$ if $0 < \alpha < 1/2$ and $f^*_{1,\gamma}(\alpha) = \gamma$ if $1/2 < \alpha < 1$ as well as $f^*_{2,\gamma}(\alpha) = \gamma(2\alpha - 1)$ for all $\alpha \in [0,1]$. The main purpose of $f^*_{p,\gamma}$ is that it can be used to construct functions that minimize the (p,γ)-risk. Indeed, we have $\mathcal{R}_{p,\gamma,P} = \mathcal{R}_{p,\gamma,P}\big(f^*_{p,\gamma}(P(1|.))\big)$ for all $p \in \{1,2\}$, $\gamma > 0$ and every Borel probability measure P on $X \times Y$. Given a positive definite and continuous kernel with corresponding RKHS H the *regularized* (p,γ)*-risk* is defined by

$$\mathcal{R}^{reg}_{p,\gamma,P,\lambda}(f) \;:=\; \lambda \|f\|_H^2 + \mathcal{R}_{p,\gamma,P}(f)$$

for all functions $f \in H$ and all $\lambda > 0$. If $\gamma = 1$ we usually omit this index. The *unique* function (cf. [18]) that minimizes $\mathcal{R}^{reg}_{p,\gamma,P,\lambda}(.)$ is denoted by $f_{p,\gamma,P,\lambda}$, or briefly $f_{P,\lambda}$. If P is an empirical measure with respect to $T \in (X \times Y)^n$ we write $\hat{\mathcal{R}}_T(f)$, $\hat{\mathcal{R}}_{p,\gamma,T}(f)$, $\hat{\mathcal{R}}^{reg}_{p,\gamma,T,\lambda}(f)$, $f_{p,\gamma,T,\lambda}$ and $f_{T,\lambda}$, respectively.

In this work we consider SVM's. Recall, that these algorithms minimize $\hat{\mathcal{R}}^{reg}_{p,1,T,\lambda}$, i.e. they solve the optimization problem

$$\min_{f\in H} \; \lambda \|f\|_H^2 + \frac{1}{n}\sum_{i=1}^n L_1^p\big(y_i f(x_i)\big) \;, \tag{2}$$

where $T = \big((x_1,y_1),\dots,(x_n,y_n)\big) \in (X \times Y)^n$ and $p \in \{1,2\}$. The SVM classifier that constructs the decision function $\operatorname{sign} \circ f_{p,1,T,\lambda}$ is called p-norm soft margin classifier (p-SMC).

Finally, for positive sequences (a_n) and (b_n) we write $a_n \preceq b_n$ if there exists a constant $c > 0$ such that $a_n \le c\, b_n$ for all $n \ge 1$. Moreover, $a_n \sim b_n$ means that both $a_n \preceq b_n$ and $b_n \preceq a_n$ hold.

3 Some results on the distribution of the margin deviations

In this section we present two results concerning the distribution of the margin errors for large sample sizes. These results will play the key role in our analysis of existing generalization bounds. For this purpose we write $\hat{X} := \hat{X}_P := \{x \in X : P(1|x) \ne 1/2\}$ for a given Borel probability measure P on $X \times Y$. Moreover, let

$$|y - y'|_\gamma := \begin{cases} |y - y'| & \text{if } -\gamma < y' < \gamma \\ \gamma - y & \text{if } y' \ge \gamma \\ y + \gamma & \text{if } y' \le -\gamma \end{cases}$$

for $y, y' \in \mathbb{R}$ and $\gamma > 0$. Finally for $f : X \to \mathbb{R}$ and $\varepsilon > 0$, $p \in \{1,2\}$, $\gamma > 0$ we define

$$E_\varepsilon(f) \;:=\; E_{\varepsilon,p,\gamma,P}(f) \;:=\; \Big\{x \in \hat{X} \;:\; \big|f(x) - f^*_{p,\gamma}(P(1|x))\big|_\gamma \ge \varepsilon\Big\}$$

Now, our first result which is the main tool in our further considerations reads as follows:

Theorem 1. *Let P be a Borel probability measure on $X \times Y$ and k a universal kernel on X. Moreover, let $\varepsilon, \delta, \gamma > 0$, $p \in \{1,2\}$ and (λ_n) be a positive sequence with $\lambda_n \to 0$ and $\lambda_n^{p+1} n \to \infty$. Then for $n \to \infty$ we have*

$$P^n\Big(T = ((x_1,y_1),\ldots,(x_n,y_n)) : \big|\{i : x_i \in E_{\varepsilon,p,\gamma,P}(f_{p,\gamma,T,\lambda_n})\}\big| \le \delta n\Big) \;\to\; 1.$$

Roughly speaking, this theorem states that for large samples sizes $f_{p,\gamma,T,\lambda_n}(x_i)$ typically equals $f^*_{p,\gamma}(P(1|x_i))$ up to ε for the overwhelming majority of samples. In particular, we obtain information about the margin distribution using the specific form of $f^*_{p,\gamma}(P(1|.))$ for $p = 1, 2$. Since the proof is rather technical and uses some aspects from functional analysis it is worked out in the last section.

Our second results concerns the distribution of $\hat{\mathcal{R}}_{p,\gamma,T}(f_{p,\gamma,T,\lambda})$:

Theorem 2. *Let P be a Borel probability measure on $X \times Y$ and k a continuous kernel on X. Moreover, let $\varepsilon, \gamma > 0$, $p \in \{1,2\}$ and (λ_n) be a positive sequence with $\lambda_n \to 0$ and $\lambda_n^{2p} n \to \infty$. Then for $n \to \infty$ we have*

$$P^n\big(T \in (X \times Y)^n : \hat{\mathcal{R}}_{p,\gamma,T}(f_{p,\gamma,T,\lambda_n}) \ge \mathcal{R}_{p,\gamma,P} - \varepsilon\big) \to 1.$$

The condition $\lambda_n^{2p} n \to \infty$ can be weakened if the kernel k fulfills suitable smoothness assumptions. E.g., for C^∞-kernels we only need $\lambda_n^{p+\delta} n \to \infty$ for an arbitrary small $\delta > 0$. For details we refer to the proof in the last section and [17].

4 Asymptotical treatment of margin-based bounds

One of the first ideas to explain the generalization performance of SVM's was to use the results of Vapnik and Chervonenkis which estimate $\mathcal{R}_P(f_T)$ in terms of $\hat{\mathcal{R}}_T(f_T)$ and the VC-dimension of the class of functions f_T has been chosen from (usually $\frac{1}{\sqrt{\lambda}} B_H$). Apart from the structural problems concerning the used inductive principle (cf. [2] and [12]) this approach cannot work in general since function classes induced by good, i.e. universal, kernels have infinite VC-dimension whenever X is infinite. One typical approach to avoid this problem is to introduce a target margin $\gamma > 0$. Then some results estimate $\mathcal{R}_P(f_T)$ under the condition $\hat{\mathcal{R}}_{0,\gamma,T}(f_T) = 0$ (cf. [12] and [6, Ch. 4]). Since this is not a realistic assumption for p-SMC's and noisy problems we do not consider these bounds (cf. also [3]). In order to treat outliers other estimates bound $\mathcal{R}_P(f_T)$ by $\hat{\mathcal{R}}_{0,\gamma,T}(f_T)$ and a probabilistic term that depends on quantities like the fat-shattering dimension. Such a result which is often referred to is the following due Bartlett (cf. [1]):

Theorem 3. *Let k be a kernel with RKHS H, $0 < \delta < 1/2$, $0 < \gamma < 1$ and $\lambda > 0$. Then for all Borel probability measures P on $X \times Y$ and all $n \ge 1$ we have*

$$P^n\Big(T : \mathcal{R}_P(f) \le \hat{\mathcal{R}}_{0,\gamma,T}(f) + \varepsilon_n(\delta,\lambda,\gamma,K) \;\; \textit{for all } f \in \tfrac{1}{\sqrt{\lambda}} B_H\Big) \;\ge\; 1 - \delta,$$

where $\varepsilon_n(\delta,\lambda,\gamma,K) := \sqrt{\frac{512K^2}{\lambda\gamma^2 n} \ln\Big(\frac{17e\lambda\gamma^2 n}{128K^2}\Big) \log_2(578n) + \frac{2}{n}\ln\frac{4}{\delta}}$.

It is possible to replace $\varepsilon_n(\delta,\lambda,\gamma,K)$ by an (often) smaller quantity that also depends on the observation in terms of the empirical fat-shattering dimension (cf. [9]). However, in this work we are mainly interested in the prediction term $\hat{\mathcal{R}}_{0,\gamma,T}(f)$. If λ tends slowly to 0 for increasing n, say $\lambda_n \succeq n^{-1/5}$ the term $\varepsilon_n(\delta,\lambda_n,\gamma,K)$ also tends to 0. Moreover, for universal kernels the p-SMC, $p \in \{1,2\}$ is universally consistent in this case (cf. [17]), i.e. we have $\mathcal{R}_P(f_{p,1,T,\lambda_{|T|}}) \to \mathcal{R}_P$ in probability for every P. Thus, in order to use Theorem 3 we should be sure that $\hat{\mathcal{R}}_{0,\gamma,T}(f_T)$

also approximates $\mathcal{R}_P$ since otherwise the estimate becomes useless. Unfortunately, it turns out that such an approximation does not take place in general.

We begin with a result that states a lower (asymptotic) bound on the fraction of γ-margin errors for 2-SMC decision functions. In particular it yields that for all $\gamma > 0$ we may construct sufficiently noisy distributions such that Theorem 3 only gives useless bounds for the 2-SMC.

Theorem 4. *Let $0 < \gamma \leq 1$, $\varepsilon > 0$ and P be a Borel probability measure on $X \times Y$. We define $A_\gamma := \{x \in X : s(x) > (1-\gamma)/2\}$, where s denotes the noise level of P. Then for all universal kernels k on X and all positive sequences (λ_n) with $\lambda \to 0$ and $\lambda_n^3 n \to \infty$ we have:*

$$\lim_{n\to\infty} P^n\Big(T \in (X\times Y)^n : \hat{\mathcal{R}}_{0,\gamma,T}(f_{2,1,T,\lambda_n}) \geq \int_{A_\gamma} (1-s)\, dP_X + \mathcal{R}_P - \varepsilon\Big) \;=\; 1 \;.$$

Proof. Due to space limitations we only sketch the proof. We define $B_\gamma := \{x \in X : s(x) \leq (1-\gamma)/2\}$ and fix a δ with $0 < \delta < \gamma$ and $P_X(A_{\gamma-\delta}) \geq P_X(A_\gamma) - \varepsilon$. Now, we only consider training sets $T = ((x_1, y_1), \ldots, (x_n, y_n))$ such that there are at least $n(P_X(A_{\gamma-\delta}) - \varepsilon)$ samples in $A_{\gamma-\delta}$ and at least $n\big(\int_{B_\gamma} s\, dP_X - \varepsilon\big)$ samples $x_i \in B_\gamma$ with "wrong" label y_i. Here, "wrong" means that $y_i = 1$ if $P(1|x_i) < 1/2$ or $y_i = -1$ if $P(-1|x_i) < 1/2$, respectively. Clearly, the probability of such training sets converges to 1. Moreover, let us additionally assume that our considered training set T fulfills $\big|\{i : x_i \in E_{\delta/2,2,1,P}(f_{T,\lambda_n})\}\big| \leq \varepsilon n$. Theorem 1 guarantees, that the probability of such training sets still converges to 1. Our assumptions on T yields that there are at least $n(P_X(A_{\gamma-\delta}) - 2\varepsilon)$ samples in $A_{\gamma-\delta}$ with $\big||f_{T,\lambda_n}(x_i)| - (1-2s(x_i))\big| < \delta/2$. Obviously, all these samples cause γ-margin errors. Moreover, T has also at least $n\big(\int_{B_\gamma} s\, dP_X - 2\varepsilon\big)$ samples $x_i \in B_\gamma$ with "wrong" label y_i and $\big|f_{T,\lambda_n}(x_i) - (1-2s(x_i))\big| < \delta/2$. Again, these samples cause γ-margin errors. Summing up the considered samples yields

$$\hat{\mathcal{R}}_{0,\gamma,T}(f_T) \;\geq\; P_X(A_{\gamma-\delta}) + \int_{B_\gamma} s\, dP_X - 4\varepsilon \;=\; \int_{A_\gamma} (1-s)\, dP_X + \mathcal{R}_P - 5\varepsilon \;. \quad \square$$

We like to point out, that the probabilities of the above proposition usually tend significantly faster to 1 than the probabilities of Theorem 3. Hence, for sufficiently noisy problems either the probabilistic term of Theorem 3 (for small sample sizes) or the predictive term $\hat{\mathcal{R}}_{0,\gamma,T}(f_T)$ (for large sample sizes) is is too large for meaningful bounds. Note, that this type of argument is also used in the following.

The next proposition shows that the predictions for the 2-SMC made by Theorem 3 can be trivial for very simple distributions and for all sequences (λ_n). The proof can be found in the last section.

Proposition 1. *Let $0 < \gamma < 1$, $X := \{-1, 1\}$ and P the probability measure on $X \times Y$ defined by $P_X(1) = P_X(-1) = 1/2$ and $P(1|1) = P(-1|-1) = p$ for some fixed p with $1/2 < p < (1+\gamma)/2$. Moreover, let $k(x, x') := x \cdot x'$. Then for all positive sequences (λ_n) we have*

$$P^n\big(T \in (X\times Y)^n : \hat{\mathcal{R}}_{0,\gamma,T}(f_{2,1,T,\lambda_n}) = 1\big) \;\to\; 1 \;.$$

The importance of the above example lies in the fact that it is one of the easiest noisy classification problems. Thus, a good margin bound on the generalization error should certainly produce estimates that are sufficiently close to the Bayes risk, at least for specific choices of λ_n.

The arguments of the proof of Theorem 4 can also be used to consider the 1-SMC in view of the asymptotic behaviour of $\hat{\mathcal{R}}_{0,\gamma,T}(f_{1,1,T,\lambda_n})$. In doing so, it turns out that $\hat{\mathcal{R}}_{0,\gamma,T}(f_{1,1,T,\lambda_n})$ tends to be in $[\mathcal{R}_P, \mathcal{R}_P + \frac{1}{2}P_X(X \setminus \hat{X})]$. The following proposition shows that this is optimal, i.e. there actually exist distributions for which the worst case $\hat{\mathcal{R}}_{0,\gamma,T}(f_T) \to \mathcal{R}_P + \frac{1}{2}P_X(X \setminus \hat{X})$ holds. Again, the proof can be found in the last section.

Proposition 2. *Let $X := \{0,1\}$ and P be a probability measure on $X \times Y$ with $P(1|0) = P(-1|0) =: p_0^+$ and $p_1^+ := P(1|1) > P(-1|1) := p_1^-$. Moreover, let k be the kernel on X defined by $k(x,x') := 1$ if $x = x'$ and $k(x,x') := 0$ otherwise. Finally, let (λ_n) be a positive sequence with $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$. Then for all $0 < \gamma < 1$ and all $\varepsilon > 0$ we have*

$$P^n\Big(T \in (X \times Y)^n : \hat{\mathcal{R}}_{0,\gamma,T}(f_{1,1,T,\lambda_n}) \geq \mathcal{R}_P + \frac{1}{2}P_X(X \setminus \hat{X}) - \varepsilon\Big) \to 1 \ .$$

The above results show that in general the fraction of γ-margin errors is not suitable to bound the risk of SVM decision functions: for the 2-SMC it turned out that regions with large noise level cause bad predictions. Therefore, results like Theorem 3 cannot explain the generalization ability of the 2-SMC. For the 1-SMC the predictions made by $\hat{\mathcal{R}}_{0,\gamma,T}(f_{1,1,T,\lambda_n})$ are optimal up to $\frac{1}{2}P_X(X \setminus \hat{X})$. Although the latter quantity may be large for general distributions we suppose that for most of the real-world problems it is small. Therefore, the main problem for explaining the generalization ability of the 1-SMC by Theorem 3 is certainly the bad behaviour of the probabilistic term $\varepsilon_n(\delta, \lambda, \gamma, K)$.

Another approach in order to justify SVM's is measuring the average deviations from the target margin instead of counting margin errors. Typical examples of such estimates bound $\mathcal{R}_P(f_T)$ in terms of $\hat{\mathcal{R}}^{reg}_{p,\gamma,T}(f_T)$ or $\hat{\mathcal{R}}^{reg}_{p,\gamma,T}(f_T)$, $p \in \{1,2\}$. We will consider a representative result of [11] (cf. also [6, Ch. 4]). In order to state it for a given kernel k let

$$F(T,f,\lambda) \ := \ \frac{64.5(1 + \frac{K^2}{n\lambda})\hat{\mathcal{R}}^{reg}_{2,T,\lambda}(f)}{8e} \ ,$$

where $T \in (X \times Y)^n$ is a training set, f is an element of the RKHS H associated to k and $\lambda > 0$. Moreover, we define

$$H(t) \ := \ t \log_2(1/t)\mathbf{1}_{[0,1/8]}(t) + \frac{3}{8}\mathbf{1}_{(1/8,\infty)}(t) \ .$$

Note, that $H : [0,\infty) \to \mathbb{R}^+$ is an increasing and uniformly continuous function. Now the result of [11] reads as follows:

Theorem 5. *Let (λ_n) be a positive sequence and P be a Borel probability measure on $X \times Y$. Then for all sufficiently large n we have*

$$P^n\big(T : \mathcal{R}_P(f) \leq 16eH(F(T,f,\lambda_n))\log_2(32n) + \varepsilon \text{ for all } f \in H\big) \ \geq \ 1 - 8n2^{-\frac{\varepsilon n}{2}}.$$

Note, that in order to have a simple result we only consider the theorem of [11] for $\gamma = 1$. However, all our results also hold for the general case $\gamma > 0$ by a simple rescaling argument. Furthermore, there exists a similar result bounding $\mathcal{R}_P(f)$ in terms of $\hat{\mathcal{R}}^{reg}_{1,\gamma,T}(f)$ (cf. [6, Thm. 4.24]) for which our analysis also holds.

Since $H(F(T,f,\lambda))$ is minimal if f is the function $f_{2,1,T,\lambda_n}$ constructed by the 2-SMC the above theorem seems to justify this algorithm. As already mentioned above it was shown in [17] that the 2-SMC is universally consistent provided that a universal kernel is used and λ_n tends slowly to 0. The next proposition shows that for noisy problems $16eH(F(T,f_T,\lambda_n))\log_2(32n) \to \infty$ in probability if such a slowly decreasing parameter sequence is used. In other words the prediction for $\mathcal{R}_P(f_T)$ made by Theorem 5 becomes trivial for large sample sizes.

Proposition 3. *Let k be a continuous kernel on X and (λ_n) be a positive null sequence with $\lambda_n^4 n \to \infty$. Then we have*

$$P^n\Big(T \in (X \times Y)^n : F(T, f_{2,1,T,\lambda_n}, \lambda_n) \geq \frac{64.5}{8e}\mathcal{R}_{2,P} - \varepsilon\Big) \;\to\; 1$$

for all Borel probability measures P on $X \times Y$ and all $\varepsilon > 0$. If $k \in C^{\infty,\infty}(X \times X)$ this even holds if we only have $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$.

Proof. The assertion is a direct consequence of Theorem 2 and the definition of $F(T, f_T, \lambda_n)$. □

What does happen for other regularization sequences? For example, it is known (cf. [16]) that the 2-SMC with universal kernel and $\lambda_n \to 0$ is also consistent for all noiseless classification problems that ensure a strictly positive margin. If we additionally know $\lambda_n(\log n)^2 \to 0$ then Theorem 5 also implies this result. However, the following examples demonstrate that even for very simple distributions which do not guarantee these restrictive assumptions the estimate of Theorem 5 does not provide any information:

Proposition 4. *Let $X := \{-1, 1\}$ and P the probability measure on $X \times Y$ defined by $P_X(1) = P_X(-1) = 1/2$ and $P(1|1) = P(-1|-1) = p$ for some fixed $p \in (1/2, 1)$. Moreover, let $k(x, x') := x \cdot x'$. Then there exists a constant $c > 0$ such that for all positive sequences (λ_n) we have*

$$P^n\big(T \in (X \times Y)^n : F(T, f_{2,1,T,\lambda_n}, \lambda_n) \geq c\big) \;\to\; 1\ .$$

Note, that due to the symmetry of P the assertion of the above example also holds for universal kernels. Again, the importance of the above example lies in the fact that it is one of the easiest noisy classification problems. Thus, a good margin bound on the generalization error should certainly produce estimates that are sufficiently close to the Bayes risk, at least for specific choices of λ_n. The next example shows that Theorem 5 does not provide any information for noiseless distributions which are concentrated around the decision boundary, either:

Proposition 5. *Let $X := [-1, 1]$ and P be a probability measure on $X \times Y$ such that $h(t) := P_X([-t, 0]) = P_X([0, t])$ fulfills*

$$h\left(\sqrt{\sqrt{\ln t}/t}\right) \sim 1/\sqrt{\ln t}$$

for $t \to \infty$. Furthermore, assume that $P(-1|x) = P(1|-x) = 0$ for all $x \in [0, 1]$ and let $k(x, x') := x \cdot x'$. Then for all positive sequences (λ_n) there exists a constant $c > 0$ such that we have

$$P^n\left(T \in (X \times Y)^n : F(T, f_{2,1,T,\lambda_n}, \lambda_n) \geq \frac{c}{\sqrt{\ln n}}\right) \;\to\; 1\ .$$

Proposition 3 and the above examples may suggest that both the large constants and the logarithmic terms in Theorem 5 cause its bad prediction performance. In other words one might expect that an estimate of the form

$$P^n\Big(T : \mathcal{R}_P(f_{2,1,T,\lambda}) \leq c\,\hat{\mathcal{R}}_{2,T}(f_{2,1,T,\lambda}) + \varepsilon_n(\lambda, \delta)\Big) \geq 1 - \delta\ , \tag{3}$$

where $c > 0$ is a small universal constant would predict well. However, an easy argument shows that we necessarily have $c \geq 1$. Choosing a universal kernel and a regularization sequence (λ_n) with $\lambda_n \to 0$ and $\lambda_n^4 n \to \infty$ the results in [17] yield

$$\hat{\mathcal{R}}_{2,T}(f_{2,1,T,\lambda_n}) \;\to\; \mathcal{R}_{2,P} \;=\; 4\int_X s(1-s)\,dP_X$$

in probability. Moreover, we have $\mathcal{R}_P = \int_X s\,dP_X$ and thus $\hat{\mathcal{R}}_{2,T}(f_{2,1,T,\lambda_n})$ tends to be in $[2\mathcal{R}_P, 4\mathcal{R}_P]$ for large sample sizes. In other words, an estimate of the form (3) for the 2-SMC can only yield good predictions for almost noiseless distributions. Even worse, replacing $\hat{\mathcal{R}}_{2,T}(f_{2,1,T,\lambda_n})$ by another risk functional of $f_{2,1,T,\lambda_n}$ does not solve the problem: since $f_{2,1,T,\lambda_n}$ tends to $2P(1|.) - 1$ the only suitable risk functional is $\hat{\mathcal{R}}_T(f) := \sum_{i=1}^{|T|} \mathbf{1}_{\{y_i f(x_i) \leq 0\}}$. Unfortunately, in order to show that $|\hat{\mathcal{R}}_T(f_{2,1,T,\lambda}) - \mathcal{R}_P(f_{2,1,T,\lambda})| \to 0$ holds *uniformly* and *distribution independent* a finite VC-dimension of $\mathcal{F} := \{f_{2,1,T,\lambda} : T \in (X \times Y)^n, n \in \mathbb{N}\}$ is necessary by the classical VC-theory (cf. [21]) but $\mathcal{F}$ has infinite VC-dimension for universal kernels on infinite sets X (cf. the construction in [16, Thm. 18]).

What does happen for the 1-SMC? Since $f_{1,1,T,\lambda_n}$ tends to $f^*_{1,1}(P(1|.))$ many of the above problems does not occur. Actually, there exists good bounds on $\mathcal{R}_P(f_{1,1,T,\lambda_n})$ with a small probabilistic term which are sharp up to $\frac{1}{2}P_X(\hat{X})$. Due to space limitations we do not go into details but refer to [5] for the most recent results.

5 Asymptotical treatment of sparsity-based bounds

Another kind of data-dependent bounds on the generalization performance of SVM's are based on sparsity properties. Roughly speaking, the idea of these bounds is that the decision functions of SVM's only depend on the support vectors and thus SVM's are a kind of compression scheme. This is exploited in order to prove bounds which seem to have small probabilistic terms. In this section we show that the typical compression rate for SVM's on noisy distributions behaves like $c\,n$, where the factor c depends on the noise of the problem. We will see that for the known bounds this is too bad in order to guarantee good estimates.

One of the sharpest sparsity-based bounds have been proved very recently in [15]. In order to formulate it let $S_{p,\lambda}(T)$ denote the number of support vectors of the function $f_{p,1,T,\lambda}$ for $p = 1, 2$ and $\lambda > 0$. Now, the result reads as follows:

Theorem 6. *Let P be a Borel probability measure on $X \times Y$, $p \in \{1, 2\}$ and $\delta > 0$. Then for all $n \geq 1$ and all $\lambda > 0$ we have*

$$P^n\left(T : \mathcal{R}_P(f_{p,1,T,\lambda}) < 1 - \left(\binom{n}{S_{p,\lambda}(T)} \cdot \frac{n(n+1)}{2\delta}\right)^{-\frac{1}{n - S_{p,\lambda}(T)}}\right) \geq 1 - \delta\ .$$

In order to investigate whether this bound yields sharp results it is necessary to know how much support vectors $f_{p,1,T,\lambda}$ typically have. Recall that every sample for which the decision function makes a 1-margin error is a support vector. Therefore we immediately find by Theorem 4 the following important result which states that the 2-SMC is not sparse classifier for a large class of distributions:

Theorem 7. *Let P be a Borel probability measure on $X \times Y$. We write $R := P_X\big(x : s(x) > 0\big)$, where s denotes the noise level of P. Moreover, let k be a universal kernel and (λ_n) be a regularization sequence with $\lambda_n \to 0$ and $\lambda_n^3 n \to \infty$. Then for all $\varepsilon > 0$ we have*

$$P^n\Big(T \in (X \times Y)^n : S_{2,\lambda_n}(T) \geq n(R - \varepsilon)\Big) \to 1\ .$$

With the help of this theorem we can now estimate the probabilistic term of Theorem 6. Indeed, let n and k be natural numbers with $k > n$. We define $\rho := k/n$. Recall, that Stirling's formula yields

$$\binom{n}{k} = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\frac{n^n}{k^k(n-k)^{n-k}}\exp\left(\frac{\theta(n)}{12n} - \frac{\theta(k)}{12k} - \frac{\theta(n-k)}{12(n-k)}\right)\ ,$$

where θ is a suitable function with $0 < \theta(m) < 1$ for all $m \in \mathbb{N}$. Using $k = \rho n$ we find

$$\binom{n}{k} \geq \frac{\rho^{-\rho n}(1-\rho)^{-(1-\rho)n}}{4\sqrt{\rho(1-\rho)n}}$$

and this yields

$$\begin{aligned} 1 - \left(\binom{n}{k} \cdot \frac{n(n+1)}{2\delta}\right)^{-\frac{1}{n-k}} &\geq 1 - \rho^{\frac{\rho}{1-\rho}}(1-\rho)\left(\frac{8\delta\sqrt{\rho(1-\rho)}}{\sqrt{n}(n+1)}\right)^{\frac{1}{(1-\rho)n}} \\ &\geq 1 - \rho^{\frac{\rho}{1-\rho}}(1-\rho) \end{aligned}$$

for all $\delta \in (0,1]$ and all $n \geq 2$. Since $\rho \mapsto 1-\rho^{\frac{\rho}{1-\rho}}(1-\rho)$ is increasing and continuous on $[0,1]$ we obtain by Theoren 7:

Proposition 6. *Let P be a Borel probability measure on $X \times Y$. We write $R := P_X\big(x : s(x) > 0\big)$, where s denotes the noise level of P. Moreover, let k be a universal kernel and (λ_n) be a regularization sequence with $\lambda_n \to 0$ and $\lambda_n^3 n \to \infty$. Then for all $\varepsilon > 0$ and all $\delta \in (0,1]$ we have*

$$P^n\left(T : 1 - \left(\binom{n}{S_{2,\lambda_n}(T)} \cdot \frac{n(n+1)}{2\delta}\right)^{-\frac{1}{n-S_{2,\lambda_n}(T)}} \geq 1 - R^{\frac{R}{1-R}}(1-R) - \varepsilon\right) \to 1.$$

Since $2\mathcal{R}_P \leq R$ and $1-(2\rho)^{\frac{2\rho}{1-2\rho}}(1-2\rho) \geq 2\rho$ for all $\rho \in [0, 1/2]$ we find by Proposition 6 that for large sample sizes the prediction of Theorem 6 is typically not smaller than $2\mathcal{R}_P$. If the Bayes risk is small or $R/\mathcal{R}_P$ is large its prediction is even significantly worse (cf. Table 1).

By Theorem 3 we obtain that for large samples sizes the fraction of support vectors of the 1-SMC is typically not smaller than the Bayes risk. Using the above considerations we hence find:

Proposition 7. *Let P be a Borel probability measure on $X \times Y$, k be a universal kernel on X and (λ_n) be a positive sequence with $\lambda_n \to 0$ and $\frac{\lambda_n n}{\log(\lambda_n n)\log n} \to \infty$. Then for all $\varepsilon > 0$, $\delta \in (0,1]$ we have*

$$P^n\left(T : 1 - \left(\binom{n}{S_{1,\lambda_n}(T)} \cdot \frac{n(n+1)}{2\delta}\right)^{\frac{-1}{n-S_{1,\lambda_n}(T)}} \geq 1 - \mathcal{R}_P^{\frac{\mathcal{R}_P}{1-\mathcal{R}_P}}(1-\mathcal{R}_P) - \varepsilon\right) \to 1.$$

An easy computation shows $1-\rho^{\frac{\rho}{1-\rho}}(1-\rho) \geq 1.5\rho$ for all $0 \leq \rho \leq 1/2$ and thus the predictions made by Theorem 6 are typically not smaller than $1.5\mathcal{R}_P$. Again, if the Bayes risk is small the factor is significantly larger (cf. Table 1).

Finally, we like to mention that we suppose, that the fraction of support vectors tends to be not smaller than $2\mathcal{R}_P$ (at least for a large class of distributions): recall that the 1-SMC and the ν-SVM are equivalent, i.e. they produce exactly the same set of decision functions. Moreover, under some regularity conditions on P and k the parameter ν is the limit of the fraction support vectors (cf. [14] and [13]). Finally, it was shown in [19] that a good choice of ν is close to $2\mathcal{R}_P$. Unfortunately, it is unknown whether smaller values of ν also produce asymptotically almost optimal results.

6 Proofs

Proof of Theorem 1. By [17, Prop. 3.2] we have $\mathcal{R}^{reg}_{p,\gamma,P,\lambda_n}(f_{P,\lambda_n}) \to \mathcal{R}_{p,\gamma,P}$ for $n \to \infty$ and thus

$$\lim_{n\to\infty} \mathcal{R}_{p,\gamma,P}(f_{P,\lambda_n}) = \mathcal{R}_{p,\gamma,P} . \tag{4}$$

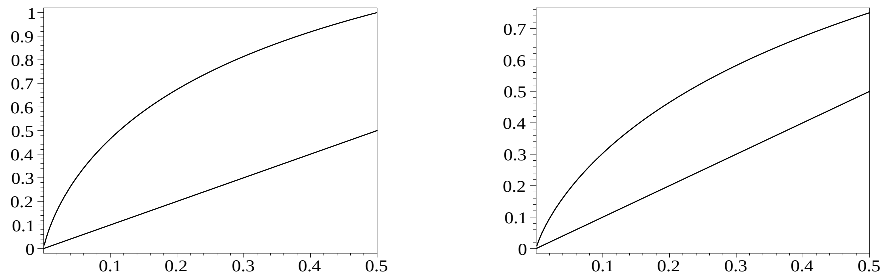

Table 1. Left: Behaviour of $\rho \mapsto 1-(2\rho)^{\frac{2\rho}{1-2\rho}}(1-2\rho)$ compared to $\rho \mapsto \rho$. Right: Behaviour of $\rho \mapsto 1-\rho^{\frac{\rho}{1-\rho}}(1-\rho)$ compared to $\rho \mapsto \rho$.

Furthermore, for $E := E_{\varepsilon,p,\gamma,P}(f_{P,\lambda_n})$ we find

$$
\begin{aligned}
&\mathcal{R}_{p,\gamma,P}(f_{P,\lambda_n}) \\
&\geq \int\limits_{X\setminus E}\int\limits_{Y} L^p_\gamma\big(y, f^*_{p,\gamma}(P(1|x))\big)P(dy|x)P_X(dx) + \int\limits_{E}\int\limits_{Y} L^p_\gamma(y, f(x))P(dy|x)P_X(dx) \\
&= \mathcal{R}_{p,\gamma,P} + \int\limits_{E}\int\limits_{Y}\Big(L^p_\gamma(y, f(x)) - L^p_\gamma\big(y, f^*_{p,\gamma}(P(1|x))\big)\Big)P(dy|x)P_X(dx) \\
&\geq \mathcal{R}_{p,\gamma,P} + \int_E \Delta(x)P_X(dx)\ ,
\end{aligned}
$$

where we have defined

$$
\begin{aligned}
\Delta(x) := \min&\left\{\int\limits_{Y} L^p_\gamma(y,y')P(dy|x) : y' \in \mathbb{R} \text{ with } \big|y' - f^*_{p,\gamma}(P(1|x))\big|_\gamma \geq \varepsilon\right\} \\
&- \int\limits_{Y} L^p_\gamma\big(y, f^*_{p,\gamma}(P(1|x))\big)P(dy|x)\ .
\end{aligned}
$$

By the definition of $|.-.|_\gamma$ we get $\Delta(x) > 0$ for all $x \in \hat{X}$. Hence the measures P_X and ΔdP_X both restricted to $\hat{X}$ are absolutely continuous to each other. This together with (4) yields

$$\lim_{n\to\infty} P_X\big(E_\varepsilon(f_{P,\lambda_n})\big) \ = \ 0$$

for all $\varepsilon > 0$. In particular, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $P_X\big(E_{\varepsilon/2}(f_{P,\lambda_n})\big) \leq \delta/2$. This implies

$$P^n\Big(T = ((x_1,y_1),\ldots,(x_n,y_n)) : \big|\{i : x_i \in E_{\varepsilon/2}(f_{P,\lambda_n})\}\big| \leq \delta n\Big) \to 1 \qquad (5)$$

for $n \to \infty$ by Hoeffding's inequality. An easy calculation shows that $E_\varepsilon(f_{T,\lambda_n}) \subset E_{\varepsilon/2}(f_{P,\lambda_n})$ whenever $\|f_{T,\lambda_n} - f_{P,\lambda_n}\| \leq \varepsilon/2$. The latter typically holds for large sample sizes: indeed, in [18, Thm. 1.2] it was shown that there exists a measurable function $h : X \times Y \to \mathbb{R}$ with $\|h\|_\infty \leq |L^p_{\gamma|[-a,a]}|_1$ and

$$\|f_{P,\lambda_n} - f_{T,\lambda_n}\| \leq \frac{1}{\lambda_n}\,\|\mathbb{E}_P h\Phi - \mathbb{E}_T h\Phi\|\ . \qquad (6)$$

Here, $|L^p_{\gamma|[-a,a]}|_1$ denotes the Lipschitz-constant of L^p_γ restricted to $[-a,a]$, $a := K/\sqrt{\lambda_n}$ and $\mathbb{E}_P h\Phi$ is the H-valued *Bochner-integral* of $h\Phi$ with respect to P. Now, [22, Thm. 3.3.4] yields

$$P^n\Big(T \in (X\times Y)^n : \|\mathbb{E}_P h\Phi - \mathbb{E}_T h\Phi\| \geq \frac{\varepsilon\lambda_n}{2}\Big) \leq 2\exp\Big(-\frac{\varepsilon\lambda_n^2 n}{64c_n^2 + 8\lambda_n\varepsilon c_n}\Big), \quad (7)$$

where $c_n := pK(\gamma + K/\sqrt{\lambda_n})^{p-1}$. Combining estimates (6) with (7) shows that $P^n(T : \|f_{T,\lambda_n} - f_{P,\lambda_n}\| \le \varepsilon/2) \to 1$ for $n \to \infty$. Therefore, we can replace $E_{\varepsilon/2}(f_{P,\lambda_n})$ by $E_\varepsilon(f_{T,\lambda_n})$ in (5). □

Proof of Theorem 2. Using Lemma 5.2 and Lemma 5.3 in [17] we find that $\left|\hat{\mathcal{R}}_{p,\gamma,T}(f_{T,\lambda_n}) - \mathcal{R}_{p,\gamma,P}(f_{T,\lambda_n})\right| \to 0$ holds in probability under the assumptions of Theorem 2. Now, the assertion follows by the trivial inequality $\mathcal{R}_{p,\gamma,P}(f_{T,\lambda_n}) \ge \mathcal{R}_{p,\gamma,P}$. □

Proof of Proposition 1. We first fix an $\varepsilon > 0$ with $\varepsilon \le \min\{(1+\gamma)/2 - p, p - 1/2\}$. Let us consider a training set $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ with

$$\frac{p-\varepsilon}{2} n \;\le\; \left|\{(x_i, y_i) \in T : x_i = y_i = j\}\right| \;\le\; \frac{p+\varepsilon}{2} n \tag{8}$$

for $j = \pm 1$. Hoeffding's inequality ensures that the probability of the occurrence of such a training set converges exponentially fast to 1. Identifying the RKHS of k with $\mathbb{R}$ we have to minimize

$$\lambda_n w^2 + a\big(\max\{0, 1-w\}\big)^2 + (1-a)\big(\max\{0, 1+w\}\big)^2 \tag{9}$$

for $w \in \mathbb{R}$ in order to find the functional constructed by the 2-SMC. Here, a is the fraction of correctly labeled samples of T. An easy calculation shows that (9) is minimized for

$$w^* \;:=\; \frac{2a-1}{1+\lambda} \;<\; 2a - 1 \;\le\; 2(p+\varepsilon) - 1 \;\le\; \gamma\ .$$

Note, that (8) yields $w^* > 0$. Since w^* and $-w^*$ are the margins of the correctly labeled and incorrectly labeled samples, respectively, the assertion follows. □

Proof of Proposition 2. Let $p_0^- := p_0^+$. Given a training set T we denote the fraction of samples in $x \in \{0,1\}$ with positive label and negative label by a_x^+ and a_x^-, respectively. By Hoeffding's inequality the probability of $|a_x^+ - p_x^+| \le \gamma\lambda_n/2$ and $|a_x^- - p_x^-| \le \gamma\lambda_n/2$ tends to 1. Since for large n, i.e. small λ_n, the regularized risk

$$\lambda_n\langle w, w\rangle + \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i\langle w, \Phi(x_i)\rangle\}$$

is minimized in (w_1, w_2) with $w_1 := (a_0^+ - a_0^-)/(2\lambda)$ and $w_2 := 1$ we observe that the margins for samples x_i with $x_i = 0$ are in $[-\gamma/2, \gamma/2]$ while the margins are 1 and -1 for correctly labeled and incorrectly labeled samples x_i with $x_i = 1$, respectively. Now, the assertion follows since $p_0^+ + p_0^- + p_1^- = \mathcal{R}_P + \frac{1}{2}P_X(X \setminus \hat{X})$. □

Proof of Proposition 4. Let $T = ((x_1, y_1), \ldots, (x_n, y_n))$ be a training set with

$$\left|\{(x_i, y_i) \in T : x_i = y_i = j\}\right| \ge \frac{1}{4}n(3p-1)$$
$$intertextand\left|\{(x_i, y_i) \in T : x_i = j,\ y_i \ne j\}\right| \ge \frac{1}{4}n(1-p)$$

for $j = \pm 1$. Hoeffding's inequality ensures that the probability of the occurrence of such a training set converges to 1. Identifying the RKHS of k with $\mathbb{R}$ we write $\xi_i := \max\{0, 1 - y_i x_i w\}$ for $w \in \mathbb{R}$. In order to minimize $\lambda_n w^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^2$ it suffices to consider $w \in [-1, 1]$ in our situation. Then we obtain

$$\lambda_n w^2 + \frac{1}{n}\sum_{i=1}^{n} \xi^2 \;\ge\; \lambda_n w^2 + \frac{1}{2}(3p-1)(1-w)^2 + \frac{1}{2}(1-p)(1+w)^2$$

An easy calculation shows that the right hand side is minimal for $w = (2p-1)/(\lambda_n+p)$ and we obtain

$$\lambda_n w^2 + \frac{1}{n}\sum_{i=1}^{n}\xi^2 \;\geq\; \frac{\lambda_n p - 3p^2 + 4p - 1}{\lambda_n + p} \;\geq\; 4p - 3p^2 - 1 > 0\ . \quad \square$$

Proof of Proposition 5. For brevity's sake let us fix sequences (t_n) and (a_n) with $t_n := \sqrt{\sqrt{\ln n}/n}$ and $a_n := nh(t_n) - \sqrt{n\ln(n)}$. Hoeffding's inequality yields

$$P^n\Big((x_1,y_1),\dots,(x_n,y_n)) \in (X\times Y)^n : \big|\{i : x_i \in [0,t_n]\}\big| \geq a_n\Big) \;\geq\; 1 - 1/n\ .$$

Now let us assume that we have a training set T of length n which has at least a_n samples in $[-t_n, 0]$ and $[0, t_n]$, respectively. Considering the optimization problem for the 2-SMC and $w \in [-1/t_n, 1/t_n]$ we find

$$\begin{aligned}
\lambda_n w^2 + \frac{1}{n}\sum_{i=1}^{n}(1-y_i w x_i)^2 &\geq \lambda_n w^2 + \frac{1}{n}\sum_{x_i\in[-t_n,t_n]}(1-y_i w x_i)^2 \\
&\geq \lambda_n w^2 + \frac{a_n}{n}(1-wt_n)^2 \\
&\geq \frac{\lambda_n a_n}{\lambda_n n + a_n t_n^2}\ .
\end{aligned}$$

Since for $w \notin [-1/t_n, 1/t_n]$ the regularized risk minimized by the 2-SMC is always larger than for $w \in [-1/t_n, 1/t_n]$ we thus obtain

$$F(T, w, \lambda_n) \;\geq\; \frac{64.5}{8e}\Big(1 + \frac{1}{\lambda_n n}\Big)\frac{\lambda_n a_n}{\lambda_n n + a_n t_n^2} \;\succeq\; \frac{1}{\sqrt{\ln n}}\ . \quad \square$$

References

1. P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, *IEEE Transaction on Information Theory* **44** (1998), 525-536
2. P. Bartlett and J. Shawe-Taylor, Generalization performance of support vector machines and other pattern classifiers, in B. Schölkopf, C.J.C. Burges, and A.J. Smola editors, "Advances in Kernel Methods - Support Vector Learning", MIT Press, 1999, 43-54
3. S. Ben-David, N. Eiron and U. Simon, Limitations of learning via embeddings in Euclidean half-spaces, in *Proceedings of the 14th Annual Conference on Computational Learning Theory*, LNAI **2111**, 385-401, 2001
4. C. Berg, J.P.R. Christensen and P. Ressel, Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions, Springer, New York, 1984
5. O. Bousquet and A. Elisseeff, Stability and generalization, *Journal of Machine Learning Research* **2** (2002), 499-526
6. N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, 2000
7. L. Devroye, L. Györfi and G. Lugosi, "A Probabilistic Theory of Pattern Recognition", Springer, New York, 1997
8. R.M. Dudley, A Course on Empirical Processes, *Lecture Notes in Math.* **1097**, 1-142, 1984
9. B. Kégl, T. Linder and G. Lugosi, Data-dependent margin-based generalization bounds for classification, in *Proceedings of the 14th Annual Conference on Computational Learning Theory*, LNAI **2111**, 368-384, 2001
10. K. Ritter, Average-Case Analysis of Numerical Problems, *Lecture Notes in Math.* **1733** (2000)

11. J. Shawe-Taylor and N. Cristianini, Margin Distribution Bounds on Generalization, in *Proceedings of the European Conference on Computational Learning Theory, EuroCOLT 99*, LNAI **1572**, 263-273, 1999
12. J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson and M. Anthony, Structural risk minimization over data-dependent hierarchies, *IEEE Trans. Inf. Theory* **44** (1998), 1926-1940
13. B. Schölkopf and A.J. Smola, "Learning with Kernels", MIT Press, 2002
14. B. Schölkopf, A.J. Smola, R.C. Williamson and P.L. Bartlett, New support vector algorithms, *Neural Computation* **12** (2000), 1207-1245
15. M. Seeger, PAC-Bayesian generalization error bounds for Gaussian process classification, preprint, `http://www.dai.ed.ac.uk/~seeger/papers/gpmcall-tr.ps.gz`
16. I. Steinwart, On the influence of the kernel on the consistency of support vector machines, *Journal of Machine Learning Research* **2** (2001), 67-93
17. I. Steinwart, Consistency of SVM's and other regularized risk minimizers, preprint, `http://www.minet.uni-jena.de/~ingo/consist.ps`
18. I. Steinwart, On the stability of support vector machines, preprint, `http://www.minet.uni-jena.de/~ingo/stability.ps`
19. I. Steinwart, On the optimal parameter choice for ν-support vector machines, submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*
20. H. Triebel, Theory of Function Spaces, Akademische Verlagsgesellschaft Geest & Portig, Leipzig, 1983
21. V.N. Vapnik, "Statistical Learning Theory", Wiley, 1998
22. V. Yurinsky, Sums and Gaussian Vectors, *Lecture Notes in Math.* **1617**, 1995