Los Alamos National Laboratory

Stability of Unstable Learning Algorithms

Don Hush, Clint Scovel, and Ingo Steinwart
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545
{dhush,jcs,ingo}@lanl.gov

LANL Technical Report: LA–UR–03–4845

Report Date: July 15, 2003

Abstract

We introduce a formalism called graphical learning algorithms and use it to produce bounds on error deviance for unstable learning algorithms. This formalism suggests a flexible class of extensions of existing algorithms for which risk can be decomposed into *algorithmic* model risk plus estimation error in a way that enables bounds on estimation error and analysis of the algorithmic model risk. For example we obtain error deviance bounds for support vector machines (SVMs) with variable offset parameter and estimation error bounds for variations of SVM where the offset parameter is selected to minimize empirical risk. In addition we prove convergence to the Bayes error for variations of SVM that use a universal kernel and choose the regularization parameter to minimize empirical error. We provide experimental results that suggest that these variations may offer advantages over standard SVMs in both computation and generalization performance.

1 Introduction

Bousquet et al. (Bousquet & Elisseeff, 2002) determine bounds on the deviance between the empirical risk and the risk for stable learning algorithms and Kutin and Niyogi (Kutin & Niyogi, 2002) have extended their work to handle various forms of stability. However many important learning strategies are not stable in any of these senses. For example Vapnik's (Vapnik, 1998) 1-norm soft margin support vector machines (SVMs) which include the offset parameter are not stable. Bousquet et al. show that if the offset parameter is fixed at zero then they are stable and apply their theory to obtain bounds on the risk deviance. However the most commonly used SVMs include the offset parameter as a variable and so the determination of bounds on risk deviance in this case is an important open problem. In this paper we develop a formalism called *graphical learning algorithms* to analyse the statistical stability of unstable learning algorithms and prove a general bound on the risk deviance for them. This result is then applied to SVMs which include the offset parameter to produce bounds on risk deviance which are similar to those of Bousquet et al.

Although risk deviance is extremely important it characterizes only one aspect of an algorithm's performance. Indeed the risk may be large even when the risk deviance is small. To better assess the actual risk we consider the notion of *algorithmic estimation error*. Recall that the smallest possible risk is the so–called Bayes risk. Although there exist algorithms whose risk converges to the Bayes risk (as the number of samples goes to ∞) there may be no algorithm possessing uniform rates of convergence (e.g. see (Devroye, Györfi, & Lugosi, 1996) Chapter 7). To better understand the factors that influence risk much effort has been devoted to the study of the empirical risk minimization algorithm and the decomposition of its risk into two terms; the estimation error which is due to finite samples and the approximation error which is due to the choice of hypothesis class (Devroye et al., 1996; Vapnik, 1998; Vidyasagar, 1997). This decomposition has provided much insight. For example, with mild assumptions on the hypothesis class it is possible to establish rates at which the estimation error converges to zero, independent of the distribution. However, because empirical risk minimization is often computationally intractable there is a gap between this theory and the algorithms that are commonly used in practice. On the other hand Vapnik has introduced the support vector machine algorithm which is computationally tractable and has produced very low risk in a large number of empirical studies. Recently it was proved that SVMs (with universal kernels) converge to the Bayes risk (Steinwart, 2002; Zhang, 2003; Steinwart, 2003a), but so far there are no finite sample bounds on estimation error for SVMs. To better understand the factors that influence the finite sample risk of algorithms other than empirical risk minimization and to help close the gap between theory and practice we consider the following (slightly generalized) notion of estimation error. Let R be the risk of the algorithm applied to the training data. If there exists a value r_* such that both $P(R \geq r_* + \epsilon)$ and $P(R \leq r_* - \epsilon)$ decrease with sample size then we define $R - r_*$ to be the algorithmic estimation error. If this decrease is strong enough then R converges to r_* with probability 1 and r_* is the limit of the risk. Once such an r_* is established the analysis of algorithms decomposes into the analysis of r_* and bounds on estimation error.

The graphical learning algorithm formalism appears to facilitate the development of algorithms for which bounds on estimation error may be obtained. This is accomplished through the introduction of *estimation error models* and is demonstrated to provide bounds on estimation error for variations of the SVM algorithm.

This paper is outlined as follows. In Section 2 we introduce graphical learning algorithms and introduce the fundamental tool in their analysis. In Section 3 we prove general results for classification and apply these results to SVMs which include the variable offset parameter. In Section 4 we introduce estimation error models and describe some general classes of algorithms for which estimation error bounds may be obtained and apply these results to an estimation error version of SVMs. In Section 5 we describe parallel approximations to graphical learning algorithms and provide a result on their performance and computation. In Section 6 we study estimation error model selection. In particular we discuss the implication for important constants obtained in the bounds presented. In addition we use the estimation error model formalism to study variations of SVM which possess improved estimation error bounds and improved computational requirements. Throughout we ignore measurability questions and issues regarding the attainment of infimums. The latter can be handled through approximate minima but complicates the presentation.

2 Graphical learning algorithms

Consider sets X and Y and a probability space $Z = X \times Y$. Let $z = (x, y)$ denote the corresponding random variable with probability measure $\mathcal{P}$ and let $S = supp(z)$ denote the support of z. Let the hypothesis space be decomposed $\mathcal{F} = \mathcal{S} \times \mathcal{U}$ into what we call the stable and unstable components. Let $c : \mathbb{R} \times Y \to \mathbb{R}$ be a cost function and let its associated loss function $l : \mathcal{F} \times Z \to \mathbb{R}$ be defined through $l(f, z) = c(f(x), y)$. The risk associated to the measure $\mathcal{P}$ and the cost function c is defined as

$$R(f) = R_{\mathcal{P}}(f) = E_{\mathcal{P}}(l(f, \cdot)). \tag{1}$$

where we drop the subscript $\mathcal{P}$ when no chance of confusion exists. The empirical risk associated to an n-sample z_n is defined as

$$R_{emp}(f) = E_n(l(f, \cdot)) \tag{2}$$

where E_n is the sample mean operator associated with the n-sample z_n. Since many learning strategies do not have unique solutions, we define a learning strategy to be a set-valued map

$$A : Z^n \to\to \mathcal{F}.$$

Definition 2.1. We say that a learning strategy A is graphical if

$$A_{z_n} = \{(A^u_{z_n}, u) : u \in U_{z_n}\}, \quad \forall z_n \in Z^n$$

for some family of mappings

$$\{A^u : Z^n \to \mathcal{S}, u \in \mathcal{U}\}$$

and for some set-valued map

$$U : Z^n \to\to \mathcal{U}.$$

This definition implies that for each z_n, A_{z_n} is the graph over $\mathcal{U}$ determined the family A^u, $u \in \mathcal{U}$ restricted to a subset U_{z_n} of its domain. In particular the set-valued nature of A is generated by U. As an example of an important class of graphical strategies, consider when A is determined by minimization $A_{z_n} = \arg\min_f J_{z_n}(f)$ of a criterion function. If $A^u_{z_n} = \arg\min_s J_{z_n}(s,u)$ has a unique solution for all u, then the general fact $\min_{s,u} = \min_u \min_s$ implies that A is graphical.

Definition 2.2. Equip $\mathcal{U}$ with a pseudometric d. A graphical learning strategy A (over a pseudometric space) is Lipschitz with respect to the cost function c over a subset $S \subset Z$ if for its associated loss l we have

$$|l\big((A^{u_1}_{z_n}, u_1), z\big) - l\big((A^{u_2}_{z_n}, u_2), z\big)| \le d(u_1, u_2), \quad \forall z \in S, z_n \in S^n.$$

A learning algorithm is defined as a selection $\hat{A}$ from the set-valued map A. That is

$$\hat{A} : Z^n \to \mathcal{F}, \text{ where } \hat{A}_{z_n} \in A_{z_n}, \quad \forall z_n \in Z^n.$$

A selection from a graphical learning strategy has a special form. It is defined through a selection $\hat{u}$ from U as

$$\hat{A}_{z_n} = (A^{\hat{u}_{z_n}}_{z_n}, \hat{u}_{z_n}), \text{ where } \hat{u}_{z_n} \in U_{z_n}, \forall z_n \in Z^n.$$

A Lipschitz graphical learning algorithm is defined as a selection from a Lipschitz graphical learning strategy.

The following simple lemma will be our main tool in analyzing Lipschitz graphical learning strategies.

Lemma 2.1. *Let $\mathcal{X} = (\mathcal{X}_t)_{t\in\mathcal{T}}$ be a real valued stochastic process over a pseudometric space $(\mathcal{T}, d)$. Suppose that $\mathcal{X}$ is Lipschitz*

$$|\mathcal{X}_{t_1} - \mathcal{X}_{t_2}| \le d(t_1, t_2), \quad \forall t_1, t_2 \in \mathcal{T}.$$

Fix $\epsilon > 0$ and consider a cover $\mathcal{O}$ of $\mathcal{T}$ of balls centered at points $t_i, i = 1, .., |\mathcal{O}|$ of radius ϵ. Then

$$\mathcal{P}\Big(\sup_t \mathcal{X}_t > \eta\Big) \le |\mathcal{O}| \sup_{i=1,..,|\mathcal{O}|} \mathcal{P}\Big(\mathcal{X}_{t_i} > \eta - \epsilon\Big)$$

Recall that for a pseudometric space $(\mathcal{T}, d)$ the covering numbers $N(\mathcal{T}, d, \epsilon)$ are the minimum number of closed balls of radius ϵ it takes to cover $\mathcal{T}$. It follows that $|\mathcal{O}| \ge N(\mathcal{T}, d, \epsilon)$ and that there exists a cover such that $|\mathcal{O}| = N(\mathcal{T}, d, \epsilon)$ in the statement of this lemma.

Proof. Let $\mathcal{O}_i$ denote the balls of the cover centered at the points t_i. Since $\mathcal{X}$ is Lipschitz $|\mathcal{X}_t - \mathcal{X}_{t_i}| \le \epsilon$ for all $t \in \mathcal{O}_i$. Consequently

$$\mathcal{P}\Big(\sup_t \mathcal{X}_t > \eta\Big) \le |\mathcal{O}| \sup_i \mathcal{P}\Big(\sup_{t\in\mathcal{O}_i} \mathcal{X}_t > \eta\Big) \le |\mathcal{O}| \sup_i \mathcal{P}\Big(\mathcal{X}_{t_i} + \epsilon > \eta\Big).$$

♦

Lipschitz graphical learning algorithms may not be stable in any of the senses prescribed by Bousquet et al.(Bousquet & Elisseeff, 2002) or Kutin and Niyogi (Kutin & Niyogi, 2002). However, the following application of Lemma 2.1 can be used to apply their results to unstable algorithms.

Theorem 2.1. *Consider a graphical learning algorithm $\hat{A}$ over a pseudometric space $(\mathcal{U}, d)$ which is Lipschitz with respect to the cost function c. Fix $\epsilon > 0$ and $n \geq 1$ and consider a minimal $\mathcal{U}$-cover of balls centered at points $u_i, i = 1,..,N(\mathcal{U}, d, \epsilon)$ of radius ϵ. Then for the risk (1), the empirical risk (2) and any η we have*

$$\mathcal{P}_n\Big(R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n}) > \eta\Big) \leq N(\mathcal{U}, d, \epsilon) \sup_{i=1,..,N(\mathcal{U},d,\epsilon)} \mathcal{P}_n\Big(R(A^{u_i}_{z_n}, u_i) - R_{emp}(A^{u_i}_{z_n}, u_i) > \eta - 2\epsilon\Big)$$

and

$$\mathcal{P}_n\Big(|R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n})| > \eta\Big) \leq N(\mathcal{U}, d, \epsilon) \sup_{i=1,..,N(\mathcal{U},d,\epsilon)} \mathcal{P}_n\Big(|R(A^{u_i}_{z_n}, u_i) - R_{emp}(A^{u_i}_{z_n}, u_i)| > \eta - 2\epsilon\Big)$$

Proof. From Definition 2.1 of a graphical learning strategy

$$\begin{aligned} R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n}) \quad &\leq \sup\nolimits_{f \in A_{z_n}} \left(R(f) - R_{emp}(f)\right) \quad = \sup\nolimits_{u \in U_{z_n}} \left(R(A^u_{z_n}, u) - R_{emp}(A^u_{z_n}, u)\right) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\quad \leq \sup\nolimits_{u \in \mathcal{U}} \left(R(A^u_{z_n}, u) - R_{emp}(A^u_{z_n}, u)\right) \end{aligned} \tag{3}$$

so that

$$\mathcal{P}_n\Big(R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n}) > \eta\Big) \leq \mathcal{P}_n\Big(\sup_{u \in \mathcal{U}} \left(R(A^u_{z_n}, u) - R_{emp}(A^u_{z_n}, u)\right) > \eta\Big). \tag{4}$$

For the righthand side consider the stochastic process

$$(R(A^u_{z_n}, u) - R_{emp}(A^u_{z_n}, u))_{u \in \mathcal{U}}.$$

Since

$$|l\big((A^s_{z_n}, s), z\big) - l\big((A^t_{z_n}, t), z\big)| \leq d(s,t)$$

it follows that

$$|R(A^s_{z_n}, s) - R_{emp}(A^s_{z_n}, s) - R(A^t_{z_n}, t) + R_{emp}(A^t_{z_n}, t)| \leq 2d(s,t)$$

and application of Lemma 2.1 with $\mathcal{T} = \mathcal{U}$ and pseudometric $2d$ at scale 2ϵ obtains the first result. The result for the absolute value follows in a similar way. ♦

Note that Theorem 2.1 is general and characterizes the performance of the learning algorithm in terms of its performance for fixed values of the parameters u and the pseudometric d so all that is needed is to bound the latter. When the algorithm is stable for fixed values of the parameters u, the latter may be analyzed through stability arguments. See Kutin and Niyogi (Kutin & Niyogi, 2002) for a thorough investigation into forms of stability which provide good bounds.

3 Classification

Although Theorem 2.1 can be applied to all of the examples in Bousquet et al. (Bousquet & Elisseeff, 2002), we only show how to apply this framework to classification. Let $Y = \{-1, 1\}$ and define the γ-clipped cost functions $l_\gamma(f, z) = c_\gamma(f(x), y)$ where for $\gamma > 0$

$$c_\gamma(\acute{y}, y) = \begin{cases} 1 & y\acute{y} \leq 0 \\ 1 - \frac{y\acute{y}}{\gamma} & 0 \leq y\acute{y} < \gamma \\ 0 & y\acute{y} \geq \gamma \ . \end{cases}$$

According to (Bousquet & Elisseeff, 2002) a real valued classification algorithm $\hat{A}$ has classification stability β if

$$||\hat{A}_{z_n} - \hat{A}_{z_n^{-i}}||_\infty \leq \beta, \quad \forall i, z_n \in S^n$$

where z_n^{-i} denotes the n-sample z_n minus the i-th point $z_n(i)$. It is said to have uniform stability with respect to the loss function l if

$$||l(\hat{A}_{z_n}, \cdot) - l(\hat{A}_{z_n^{-i}}, \cdot)||_\infty \leq \beta, \quad \forall i, z_n \in S^n.$$

Theorem 3.1. *Consider a graphical learning algorithm $\hat{A}$ over a pseudometric space $(\mathcal{U}, d)$ which is Lipschitz with respect to the cost function c_γ. Let R denote the risk function associated with c_γ. Suppose that for each u, $(A^u, u) : z_n \mapsto (A^u_{z_n}, u)$ has classification stability β. Then for any $\epsilon \geq 0$,*

$$\mathcal{P}_n\Big(|R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n})| > \epsilon + 2\frac{\beta}{\gamma}\Big) \leq N(\mathcal{U}, d, \epsilon/4)e^{-\frac{n\epsilon^2}{2(\frac{4n\beta}{\gamma}+1)^2}}$$

Proof. It is easy to see that classification stability β implies uniform stability β/γ for l_γ. Theorem 12 of Bousquet et al. (Bousquet & Elisseeff, 2002) extends identically to the absolute value of the risk deviance for real classification algorithms with uniform stability β/γ. That is, since $|l_\gamma| \leq 1$, for each $u \in \mathcal{U}$ we have

$$\mathcal{P}_n\Big(|R(A^u_{z_n}, u) - R_{emp}(A^u_{z_n}, u)| > \acute{\epsilon} + 2\frac{\beta}{\gamma}\Big) \leq e^{-\frac{2n\acute{\epsilon}^2}{(\frac{4n\beta}{\gamma}+1)^2}} \tag{5}$$

We apply Theorem 2.1 with $\eta = \hat{\epsilon} + 2\frac{\beta}{\gamma}$ and $\epsilon = \hat{\epsilon}/4$ to (5) with $\acute{\epsilon} = \hat{\epsilon}/2$, followed by $\hat{\epsilon} \mapsto \epsilon$. ♦

What is left is to characterize Lipschitz graphical classification algorithms. Instead of proceeding generally we specialize to soft margin support vector machines. Let H denote a Hilbert space and let $\mathcal{S} = H^* = H$ be the bounded linear functions on H. Let $\mathcal{U} = \mathbb{R}$ be the constant functions on H. Let the model space $\mathcal{F} = \mathcal{S} \times \mathcal{U}$ be the affine functions on $X \subset H$. That is, the point $f = (\psi, b) \in \mathcal{F}$ corresponds to the function $f(x) = \psi \cdot x + b$. On the product space $Z = X \times Y$ consider the penalty function

$$\eta_f(z) = \max(0, 1 - y(\psi \cdot x + b)).$$

Recall that for an n-sample z_n we let E_n denote the sample mean operator. In general we use the subscript n to denote a shorthand for dependence on the n-sample z_n. With this notation let

$$J_n(\psi, b) = |\psi|^2 + CE_n\eta_{(\psi,b)} \tag{6}$$

denote the soft margin criterion with regularization parameter $C > 0$. We define the soft margin classification strategy SVM to be the solutions

$$(\psi, b)_n = \arg\min_{\psi,b} J_n(\psi, b) \tag{7}$$

to the soft margin problem. We note the identification

$$SVM_{z_n} = (\psi, b)_n$$

Theorem 3.2. *Suppose that $|supp(x)| \leq K$ and consider the γ-clipped cost c_γ and its corresponding risk function R. Consider an algorithm $\hat{A}$ which selects a solution to SVM (7) so that when all the data are one class y^* the solution $(\psi, b) = (0, y^*)$ is obtained. Then for any $\gamma > 0$ and $\epsilon \geq 0$,*

$$\mathcal{P}_n\Big(|R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n})| > \epsilon + \frac{CK^2}{\gamma n}\Big) \leq \Big(\frac{64(\sqrt{C}K + \sqrt{2})^3}{\gamma^2\epsilon^2} + 1\Big)e^{-\frac{n\epsilon^2}{2(2\frac{CK^2}{\gamma}+1)^2}} \tag{8}$$

Proof. Howse et al.(Howse, Hush, & Scovel, 2001) show that any selection from (7) which satisfies the assumption of the theorem satisfies $P_{\mathcal{U}}((\psi, b)_n) \in [-(1+\sqrt{C}K), 1+\sqrt{C}K]$. Therefore we define $B = 1 + \sqrt{C}K$ and redefine

$$\mathcal{U} = \mathcal{B} = \{b : |b| \leq B\}. \tag{9}$$

Let

$$\psi_n(b) = \arg\min_{\psi} J_n(\psi, b) \tag{10}$$

denote the solution to the fixed b soft margin problem. It is well known that this solution is unique and so defines a parameterized family of mappings

$$A^b_{z_n} = \psi_n(b).$$

Since

$$\min_{(\psi,b)} = \min_b \min_\psi$$

we obtain that

$$(\psi, b)_n = \{(\psi_n(b), b) : b \in P_{\mathcal{U}}((\psi, b)_n)\}. \tag{11}$$

where $P_{\mathcal{U}}$ is the projection from subsets of $\mathcal{F}$ to subsets of $\mathcal{U}$.

Consequently, SVM is graphical with $U_{z_n} = P_{\mathcal{U}}((\psi, b)_n)$. Bousquet et al. (Bousquet & Elisseeff, 2002) show that if $|supp(x)| \leq K$ then the algorithm A^0 determined by optimizing the $b = 0$ soft margin criterion 10 has classification stability

$$\beta = \frac{CK^2}{2n}. \tag{12}$$

The same proof technique can be used to show that the algorithm A^b determined by optimizing the fixed b optimization problem (10) also has the same classification stability (12) for any b.

We now proceed towards proving that SVM is Lipschitz. To that end we prove the following lemma which is valid for more general measures than the empirical distribution corresponding to an n-sample z_n.

Lemma 3.1. *Let Q be a probability measure on Z with bounded support. Define the criterion*

$$J_Q(\psi, b) = |\psi|^2 + CE_Q\eta_{(\psi,b)} \tag{13}$$

and let

$$\psi_Q(b) = \arg\min_{\psi} J_Q(\psi, b) \tag{14}$$

denote the unique (Zhang, 2001; Steinwart, 2003b) solution to the fixed b soft margin problem at Q. Then

$$|\psi_Q(b_1) - \psi_Q(b_2)|^2 \leq C|b_1 - b_2|. \tag{15}$$

Proof. Consider a function $F(h) = |h|^2 + L(h)$ where $L : H \to \mathbb{R}$ is convex and finite. Then according to Barbu et. al.(Barbu & Precupanu, 1978) the subdifferentials add

$$\partial_h F = 2h + \partial_h L.$$

Barbu et al. credit Rockafellar with this infinite dimensional extension of his finite dimensional result (Rockafellar, 1970). Let h^* be a minimizer of F. It follows that $0 \in \partial_{h^*} F = 2h^* + \partial_{h^*} L$ so that $-2h^* \in \partial_{h^*} L$. However, by the definition of subdifferential,

$$L(h) - L(h^*) \geq \langle -2h^*, h - h^* \rangle, \ \forall h.$$

Therefore

$$F(h) - F(h^*) = |h|^2 - |h^*|^2 + L(h) - L(h^*) \geq |h|^2 - |h^*|^2 - < 2h^*, h - h^* >= |h - h^*|^2.$$

and consequently we obtain the inequality of Bousquet et al. (Bousquet & Elisseeff, 2002)

$$F(h) - F(h^*) \geq |h - h^*|^2 \tag{16}$$

which they derived under the additional assumption of differentiability of L. Since Q has bounded support $CE_Q\eta_{(\psi,b)}$ is finite so we can apply (16) to $J_Q(\psi, b)$ with b fixed to obtain

$$|\psi_Q(b_1) - \psi_Q(b_2)|^2 \leq J_Q(\psi_Q(b_2), b_1) - J_Q(\psi_Q(b_1), b_1)$$

and

$$|\psi_Q(b_1) - \psi_Q(b_2)|^2 \leq J_Q(\psi_Q(b_1), b_2) - J_Q(\psi_Q(b_2), b_2).$$

Adding the two we obtain

$$2|\psi_Q(b_1) - \psi_Q(b_2)|^2 \leq J_Q(\psi_Q(b_2), b_1) - J_Q(\psi_Q(b_1), b_1) + J_Q(\psi_Q(b_1), b_2) - J_Q(\psi_Q(b_2), b_2)$$

$$= C\Big(E_Q\eta_{(\psi_Q(b_2),b_1)} - E_Q\eta_{(\psi_Q(b_1),b_1)} + E_Q\eta_{(\psi_Q(b_1),b_2)} - E_Q\eta_{(\psi_Q(b_2),b_2)}\Big)$$

$$= CE_Q\Big(\eta_{(\psi_Q(b_2),b_1)} - \eta_{(\psi_Q(b_1),b_1)} + \eta_{(\psi_Q(b_1),b_2)} - \eta_{(\psi_Q(b_2),b_2)}\Big).$$

Since $||\eta_{(\psi_Q(b_2),b_1)} - \eta_{(\psi_Q(b_2),b_2)}||_\infty \leq |b_1 - b_2|$ and $||\eta_{(\psi_Q(b_1),b_2)} - \eta_{(\psi_Q(b_1),b_1)}||_\infty \leq |b_1 - b_2|$ the proof is finished. ♦

We continue. The definition 2.1 of a graphical strategy extends to set-valued maps $\mathcal{A} : \mathfrak{P} \to \mathcal{F}$ where $\mathfrak{P}$ is a set of probability measures. Indeed, in Section 4 such maps are used to analyze estimation error.

Lemma 3.2. *Let Q be a probability measure on Z with supp$(Q_X) \leq K$ where Q_X is the X-marginal of Q, and let the set-valued map $\mathcal{A}$ be graphical (Def.2.1) with components*

$$\mathcal{A}^b_Q = \psi_Q(b)$$

the fixed b soft margin solutions at Q (14). Then $\mathcal{A}$ is Lipschitz with respect to l_γ for the metric

$$d_\gamma(b_1, b_2) = \frac{1}{\gamma}(\sqrt{C}K\sqrt{|b_1 - b_2|} + |b_1 - b_2|).$$

Proof.

$$|l_\gamma\big((\mathcal{A}^{b_1}_Q, b_1), z\big) - l_\gamma\big((\mathcal{A}^{b_2}_Q, b_2), z\big)| = |l_\gamma\big((\psi_Q(b_1), b_1), z\big) - l_\gamma\big((\psi_Q(b_2), b_2), z\big)|$$

$$= |c_\gamma\big((\psi_Q(b_1), b_1)(x), y\big) - c_\gamma\big((\psi_Q(b_2), b_2)(x), y\big)| \leq \frac{1}{\gamma}|(\psi_Q(b_1), b_1)(x) - (\psi_Q(b_2), b_2)(x)|$$

$$= \frac{1}{\gamma}|\psi_Q(b_1)\cdot x + b_1 - \psi_Q(b_2)\cdot x - b_2| \leq \frac{1}{\gamma}\big(|\psi_Q(b_1) - \psi_Q(b_2)|K + |b_1 - b_2|\big).$$

Application of Lemma 3.1 finishes the proof. ♦

The fact that SVM is Lipschitz with respect to l_γ for the metric d_γ now follows from Lemma 3.2 by letting Q be the empirical distribution of the n-sample z_n.

To apply Theorem 3.1 we need to bound the covering numbers $N(\mathcal{B}, d_\gamma, \epsilon/4)$. Because these covering numbers are small compared with exponential decay of the probability bounds we bound them crudely.

Lemma 3.3.

$$N(\mathcal{B}, d_\gamma, \epsilon) \leq \frac{4(\sqrt{C}K + \sqrt{2})^3}{\gamma^2\epsilon^2} + 1.$$

Proof. Let $d(b_1,b_2) = \sqrt{C}K\sqrt{|b_1-b_2|} + |b_1-b_2|$ so that $d_\gamma = \frac{1}{\gamma}d$. Since $N(\mathcal{B},d_\gamma,\epsilon) = N(\mathcal{B},d,\gamma\epsilon)$ we bound $N(\mathcal{B},d,\epsilon)$. Since $|b| \leq B$ it follows that $|b_1-b_2| \leq \sqrt{|b_1-b_2|}\sqrt{2B}$ so that

$$d(b_1,b_2) = \sqrt{C}K\sqrt{|b_1-b_2|} + |b_1-b_2| \leq (\sqrt{C}K+\sqrt{2B})\sqrt{|b_1-b_2|}$$

so that if we let $\alpha = \sqrt{C}K+\sqrt{2B}$ then $d(b_1,b_2) \leq \alpha\sqrt{|b_1-b_2|}$ and we obtain

$$N(\mathcal{B},d,\epsilon) \leq N(\mathcal{B},\alpha\sqrt{|\ |},\epsilon) = N(\mathcal{B},|\ |,\frac{\epsilon^2}{\alpha^2}).$$

Since the latter is bounded by

$$\frac{2B\alpha^2}{\epsilon^2}+1$$

and we can bound

$$B\alpha^2 = (\sqrt{C}K+1)\Big(\sqrt{C}K+\sqrt{2(\sqrt{C}K+1)}\Big)^2 \leq 2(\sqrt{C}K+1)(CK^2+2(\sqrt{C}K+1))$$

$$\leq 2(\sqrt{C}K+1)(\sqrt{C}K+\sqrt{2})^2 \leq 2(\sqrt{C}K+\sqrt{2})^3$$

the proof is finished. ♦

We now proceed to finish the proof of Theorem 3.2. The classification stability bound (12) combined with the fact that Lemma 3.2 implies that SVM is Lipschitz with respect to l_γ for the metric d_γ and the bound on the covering numbers of Lemma 3.3 applied to Theorem 3.1 finishes the proof. ♦

Instead of using the stability result (12) of (Bousquet & Elisseeff, 2002) we can alternatively apply Theorem 4.2. Indeed, assuming in (32) that Q is the empirical measure with respect to z_n^{-i} we find

$$\begin{aligned}
||\psi_{z_n}(b) - \psi_{z_n^{-i}}(b)||_\infty &\leq C\Big|\frac{1}{n}\sum_{j=1}^{n} z_n(j)h(z_n(j)) - \frac{1}{n-1}\sum_{j\neq i} z_n(j)h(z_n(j))\Big| \\
&= C\Big|\frac{1}{n}z_n(i)h(z_n(i)) - \frac{1}{n(n-1)}\sum_{j\neq i} z_n(j)h(z_n(j))\Big| \\
&\leq \frac{CK}{n} + \frac{CK}{n-1}.
\end{aligned}$$

Hence this path yields that SVM has classification stability $\beta = \frac{2CK}{n-1}$. Note, that this is slightly worse than (12). Furthermore, instead of using Theorem 12 of (Bousquet & Elisseeff, 2002) to establish (5) and Theorem 3.2 one can also apply (33). In this case the term $2\beta/\gamma$ in (5) vanishes and hence $\frac{CK^2}{\gamma n}$ in (8) disappears. However, the arising constants in the exponential term on the right sides of (5) and (8) are slightly larger. Therefore we do not go into details of this route.

Although this path from Theorem 3.1 to a result like Theorem 3.2 is general, we note that for SVM we can actually do a little better because it has some special structure. Indeed, it is well known that it is only b which can vary on the solutions. Namely

$$(\psi, b)_n = \{(\psi_n, b); b \in P_{\mathcal{U}}((\psi, b)_n)\}$$

where ψ_n does not depend on b. If one considers the proof of Theorem 2.1, in the first line one can assert that $\psi_n(b) = \psi_n$ is constant in b for optimal b before expanding the supremum to all of the unstable set. Then one needs only use a pseudometric d such that

$$|l\big((\psi_n, b), z\big) - l\big((\psi_n, b_i), z\big)| \leq d(b, b_i)$$

for the constant ψ_n. Then the proof of Lemma 3.2 shows that we can choose

$$d_\gamma(b_1, b_2) = \frac{1}{\gamma}|b_1 - b_2|.$$

Consequently one obtains Theorem 3.2 with covering numbers linear in $\frac{1}{\gamma\epsilon}$

$$N(\mathcal{B}, d_\gamma, \epsilon/4) \leq \frac{8(\sqrt{C}K + 1)}{\gamma\epsilon} + 1$$

instead of the quadratic $\frac{64(\sqrt{C}K+\sqrt{2})^3}{\gamma^2\epsilon^2} + 1$. That is

Theorem 3.3. *Suppose that $|supp(x)| \leq K$ and consider the γ-clipped cost c_γ and its corresponding risk function R. Consider an algorithm $\hat{A}$ which selects a solution to SVM (7) so that when all the data are one class y^* the solution $(\psi, b) = (0, y^*)$ is obtained. Then for any $\gamma > 0$ and $\epsilon \geq 0$,*

$$\mathcal{P}_n\Big(|R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n})| > \epsilon + \frac{CK^2}{\gamma n}\Big) \leq \Big(\frac{8(\sqrt{C}K + 1)}{\gamma\epsilon} + 1\Big)e^{-\frac{n\epsilon^2}{2(2\frac{CK^2}{\gamma}+1)^2}}.$$

We can state this bound in the alternative form. If $\delta \leq e^{-1}$ and $n \geq \frac{(2CK^2+\gamma)^2}{32(\sqrt{C}K+1)^2}\ln\frac{1}{\delta}$, then with probability greater than $1-\delta$ we have

$$|R(\hat{A}_{z_n}) - R_{emp}(\hat{A}_{z_n})| \leq \frac{CK^2}{\gamma n} + \frac{\sqrt{2}(\frac{2CK^2}{\gamma} + 1)}{\sqrt{n}}\sqrt{\ln\Big(12\sqrt{n}\frac{\sqrt{C}K + 1}{2CK^2 + \gamma}\Big) + \ln\frac{1}{\delta}}. \tag{17}$$

To prove the bound (17) observe that if $\delta(x) = (\frac{a}{\sqrt{x}} + 1)e^{-bx}$ we can bound the inverse $x(\delta)$ as follows: $\delta = (\frac{a}{\sqrt{x}} + 1)e^{-bx}$ implies that $x = \frac{1}{b}(\ln(\frac{a}{\sqrt{x}} + 1) + \ln\frac{1}{\delta})$ which implies that $x \geq \frac{1}{b}\ln\frac{1}{\delta}$ so that $x \leq \frac{1}{b}(\ln(\frac{a\sqrt{b}}{\sqrt{\ln\frac{1}{\delta}}} + 1) + \ln\frac{1}{\delta})$ and if $b \geq \frac{1}{a^2}\ln\frac{1}{\delta}$ and $\delta \leq e^{-1}$ then it follows that $x \leq \frac{1}{b}(\ln 2a\sqrt{b} + \ln\frac{1}{\delta})$. Then apply with $x = \epsilon^2$, $a = \frac{8(\sqrt{C}K+1)}{\gamma}$, and $b = \frac{n}{2(\frac{2CK^2}{\gamma}+1)^2}$ and use $8\sqrt{2} \leq 12$.

The bound (17) can be compared with the result of Example 2 of Bousquet et al (Bousquet & Elisseeff, 2002) for soft margin SVMs with $b = 0$ as follows. Setting $\gamma = 1$, the $\frac{1}{n}$ term is identical, the coefficient in front of the large square root is larger by a factor of two and inside the large square root is an additional $\ln n$ and constant term. The factor of 2 can mostly be removed by modifying Theorem 3.1 so that $\epsilon/4$ in the covering numbers becomes more like $\epsilon/10$ with a better constant in the exponential. This however increases the coefficient in the ln term. We do not provide the details.

4 Estimation Error

The bounds of Theorem 3.1 are bounds on risk deviance. They control how different the risk $R(\hat{A})$ is from the empirical risk $R_{emp}(\hat{A})$ but they do not say anything about the size of the risk which could be large. To assess the size of the risk we consider the notion of algorithmic estimation error $R(\hat{A}) - r_*$ presented in the introduction. We seek a risk value r_* and bounds on concentration of the risk $R(\hat{A})$ about r_*. To accomplish this we introduce *learning models* as a natural extension of the notion of *learning strategies* and show how the analysis of estimation error can be accomplished for graphical Lipschitz learning models through the analysis of estimation error models.

A learning model should be thought of as the learning strategy applied to an infinite number of samples. Specifically, let $\mathfrak{P}$ denote a set of probability measures on Z. We define a learning model to be a set-valued map

$$\mathcal{A} : \mathfrak{P} \to\to \mathcal{F}$$

where we note the similarity with the definition of a learning strategy. This similarity is fundamental in what follows. For a fixed cost function c, we recall from (1) that the risk function R depends on the measure $\mathcal{P}$ and so extend it to depend on an arbitrary measure $Q \in \mathfrak{P}$ and denote this extension R_Q. For specific cost function c we define *the model risk* at $\mathcal{P}$ to be

$$r_{\mathcal{P}}(\mathcal{A}) = \inf_{f \in \mathcal{A}_{\mathcal{P}}} R_{\mathcal{P}}(f) \tag{18}$$

where we will consider the value $r_* = r_{\mathcal{P}}(\mathcal{A})$ in defining estimation error. Let $\mathcal{I} : Z^n \to \mathfrak{P}$ denote the map from n-samples z_n to their corresponding empirical distributions

$$\mathcal{I} z_n = \frac{1}{n} \sum_{i=1,n} \delta_{z_n(i)}.$$

The learning strategy A canonically induced by a learning model $\mathcal{A}$ and the map $\mathcal{I}$ is

$$A = \mathcal{A}\mathcal{I}.$$

This learning strategy is obtained by simply applying the learning model to the empirical distribution associated with the sample data. Such learning strategies are symmetric under permutation of the indices labeling the n-sample. We use the shorthand $R_{emp} = R_{\mathcal{I}z_n}$. The definitions of graphical and Lipschitz *models* are the same as for learning *strategies* in Definitions 2.1 and 2.2 and the canonical strategy inherits these properties.

Now that we have defined the model risk value (18) we seek models which possess bounds on estimation error. To that end, consider a family of mappings $\mathcal{A}^u : \mathfrak{P} \to \mathcal{F}, u \in \mathcal{U}$ and define a graphical learning model $\mathcal{A}$ by choosing the optimal u values through risk minimization $U_Q = \arg\min_{u\in\mathcal{U}} R_Q(\mathcal{A}^u_Q, u)$. That is, an *estimation error model* is defined as

$$\mathcal{A}_Q = \{(\mathcal{A}^u_Q, u) : u \in \arg\min_{u'\in\mathcal{U}} R_Q(\mathcal{A}^{u'}_Q, u')\}. \tag{19}$$

and its model risk at $\mathcal{P}$ (18) can be written

$$r_\mathcal{P}(\mathcal{A}) = \inf_{u\in\mathcal{U}} R_\mathcal{P}(\mathcal{A}^u_\mathcal{P}, u). \tag{20}$$

It is Lipschitz with respect to c when the family $\mathcal{A}^u, u \in \mathcal{U}$ satisfies Definition 3.2. Its canonical learning strategy ($A = \mathcal{A}\mathcal{I}$) is

$$A_{z_n} = \{(A^u_{z_n}, u) : u \in \arg\min_{u'\in\mathcal{U}} R_{emp}(A^{u'}_{z_n}, u')\}. \tag{21}$$

where $A^u_{z_n} = \mathcal{A}^u_{\mathcal{I}z_n}$.

By choosing $Q = \mathcal{I}z_n$ we can use the following lemma to bound the estimation error.

Lemma 4.1. *Consider a family $\mathcal{A}^u : \mathfrak{P} \to \mathcal{F}, u \in \mathcal{U}$, a cost function c, and its corresponding family of risk operators $R_Q, Q \in \mathfrak{P}$. Consider the estimation error model (19) and its model risk $r_\mathcal{P}(\mathcal{A})$ at $\mathcal{P}$ (20). Let $\hat{A}$ be a selection from $\mathcal{A}$. Then for every $Q \in \mathfrak{P}$*

$$R_\mathcal{P}(\hat{A}_Q) - r_\mathcal{P}(\mathcal{A}) \leq \sup_u \left(R_\mathcal{P}(\mathcal{A}^u_Q, u) - R_\mathcal{P}(\mathcal{A}^u_\mathcal{P}, u)\right) + 2\sup_u |R_Q(\mathcal{A}^u_Q, u) - R_\mathcal{P}(\mathcal{A}^u_Q, u)|$$

and

$$r_\mathcal{P}(\mathcal{A}) - R_\mathcal{P}(\hat{A}_Q) \leq \sup_u \left(R_\mathcal{P}(\mathcal{A}^u_\mathcal{P}, u) - R_\mathcal{P}(\mathcal{A}^u_Q, u)\right)$$

where the supremums are taken over all of $\mathcal{U}$.

Proof. Since $\mathcal{A}$ is graphical, $R_\mathcal{P}(\hat{A}_Q) = R_\mathcal{P}(\mathcal{A}^{u_Q}_Q, u_Q)$ for some u_Q. Let $u_\mathcal{P}$ be a point where the infimum (20) $r_\mathcal{P}(\mathcal{A}) = R_\mathcal{P}(\mathcal{A}^{u_\mathcal{P}}_\mathcal{P}, u_\mathcal{P})$ is attained. It follows from (19) that $R_Q(\mathcal{A}^{u_Q}_Q, u_Q) \leq R_Q(\mathcal{A}^u_Q, u)\ \forall u \in \mathcal{U}$ and from (20) that $R_\mathcal{P}(\mathcal{A}^{u_\mathcal{P}}_\mathcal{P}, u_\mathcal{P}) \leq R_\mathcal{P}(\mathcal{A}^u_\mathcal{P}, u)\ \forall u \in \mathcal{U}$. Therefore,

$$\begin{aligned} R_\mathcal{P}(\mathcal{A}^{u_Q}_Q, u_Q) &= R_Q(\mathcal{A}^{u_Q}_Q, u_Q) + R_\mathcal{P}(\mathcal{A}^{u_Q}_Q, u_Q) - R_Q(\mathcal{A}^{u_Q}_Q, u_Q) \\ &\leq R_Q(\mathcal{A}^{u_\mathcal{P}}_Q, u_\mathcal{P}) + \sup_u |R_Q(\mathcal{A}^u_Q, u) - R_\mathcal{P}(\mathcal{A}^u_Q, u)|. \end{aligned} \tag{22}$$

but since

$$\begin{aligned} R_Q(\mathcal{A}^{u_\mathcal{P}}_Q, u_\mathcal{P}) &= R_\mathcal{P}(\mathcal{A}^{u_\mathcal{P}}_Q, u_\mathcal{P}) + R_Q(\mathcal{A}^{u_\mathcal{P}}_Q, u_\mathcal{P}) - R_\mathcal{P}(\mathcal{A}^{u_\mathcal{P}}_Q, u_\mathcal{P}) \\ &\leq R_\mathcal{P}(\mathcal{A}^{u_\mathcal{P}}_Q, u_\mathcal{P}) + \sup_u |R_Q(\mathcal{A}^u_Q, u) - R_\mathcal{P}(\mathcal{A}^u_Q, u)| \end{aligned} \tag{23}$$

we have

$$R_{\mathcal{P}}(\hat{A}_Q) = R_{\mathcal{P}}(\mathcal{A}_Q^{u_Q}, u_Q) \le R_{\mathcal{P}}(\mathcal{A}_Q^{u_{\mathcal{P}}}, u_{\mathcal{P}}) + 2\sup_u |R_Q(\mathcal{A}_Q^u, u) - R_{\mathcal{P}}(\mathcal{A}_Q^u, u)|.$$

Consequently,

$$R_{\mathcal{P}}(\hat{A}_Q) - r_{\mathcal{P}}(\mathcal{A}) \le R_{\mathcal{P}}(\mathcal{A}_Q^{u_{\mathcal{P}}}, u_{\mathcal{P}}) - R_{\mathcal{P}}(\mathcal{A}_{\mathcal{P}}^{u_{\mathcal{P}}}, u_{\mathcal{P}}) + 2\sup_u |R_Q(\mathcal{A}_Q^u, u) - R_{\mathcal{P}}(\mathcal{A}_Q^u, u)|$$

and the first result follows. For the second result we observe

$$r_{\mathcal{P}}(\mathcal{A}) - R_{\mathcal{P}}(\hat{A}_Q) = R_{\mathcal{P}}(\mathcal{A}_{\mathcal{P}}^{u_{\mathcal{P}}}, u_{\mathcal{P}}) - R_{\mathcal{P}}(\mathcal{A}_Q^{u_Q}, u_Q) \le R_{\mathcal{P}}(\mathcal{A}_{\mathcal{P}}^{u_Q}, u_Q) - R_{\mathcal{P}}(\mathcal{A}_Q^{u_Q}, u_Q)$$

and the proof is finished.

♦

Now consider the SVM model

$$\mathcal{SVM}_Q = \arg\min_{\psi, b} (|\psi|^2 + CE_Q\eta_{(\psi,b)}). \tag{24}$$

As mentioned before

$$\mathcal{SVM}_Q = \{(\psi_Q, b) : b \in P_{\mathcal{B}}(\mathcal{SVM}_Q)\}$$

is graphical with the choice $U_Q = P_{\mathcal{B}}(\mathcal{SVM}_Q)$ and is Lipschitz. However, because this model is unstable and not directly related to risk optimization, bounds on its estimation error appear difficult to obtain. It turns out that making the selection process in the unstable parameters based on risk minimization is sufficient to remedy this situation. That is, let $\mathcal{U} = \mathcal{B}$ and let $\mathcal{S}\dot{\mathcal{V}}\mathcal{M}$ denote the estimation error model (19) defined with the unique solution

$$\mathcal{A}_Q^b = \psi_Q(b) = \arg\min_{\psi} (|\psi|^2 + CE_Q\eta_{(\psi,b)}) \tag{25}$$

to the fixed b soft margin problem determined by Q. Since $P_{\mathcal{B}}(\mathcal{SVM}_Q) \subset \mathcal{B}$ it follows from the definition of model risk (20) that $\mathcal{S}\dot{\mathcal{V}}\mathcal{M}$ provides no degradation in model risk

$$r_Q(\mathcal{S}\dot{\mathcal{V}}\mathcal{M}) \le r_Q(\mathcal{SVM}). \tag{26}$$

We can now bound the estimation error for $\mathcal{S}\dot{\mathcal{V}}\mathcal{M}$.

Theorem 4.1. *Suppose that $|supp(x)| \le K$ and consider the γ-clipped cost function and its associated family of risk operators R_Q. Let $\mathcal{B} = \{b : |b| \le 1 + \sqrt{C}K\}$ and define*

$$r_* = \inf_b R_{\mathcal{P}}(\psi_{\mathcal{P}}(b), b). \tag{27}$$

where $\psi_{\mathcal{P}}(b)$ is the solution to the infinite sample fixed b soft margin problem (25) at $\mathcal{P}$. Consider an algorithm $\hat{A}$ which minimize the empirical risk R_{emp} over all $(\psi_n(b), b), b \in \mathcal{B}$ where $\psi_n(b)$ is the solution to the fixed b soft margin problem

$$\psi_n(b) = \arg\min_{\psi} (|\psi|^2 + CE_n\eta_{(\psi,b)}). \tag{28}$$

Then for any $0 < \epsilon \leq 1$

$$\mathcal{P}_n\Big(R_{\mathcal{P}}(\hat{A}_{z_n}) > r_* + \frac{2CK^2}{\gamma n} + \epsilon\Big) \leq (\frac{1536(\sqrt{C}K + \sqrt{2})^3}{\gamma^2\epsilon^2} + 3)e^{-\frac{n\epsilon^2}{32(2\frac{CK^2}{\gamma}+1)^2}}$$

and

$$\mathcal{P}_n\Big(R_{\mathcal{P}}(\hat{A}_{z_n}) < r_* - \epsilon\Big) \leq 2(\frac{64(\sqrt{C}K + \sqrt{2})^3}{\gamma^2\epsilon^2} + 1)e^{-\frac{n\epsilon^2}{8(2\frac{CK^2}{\gamma}+1)^2}}.$$

Proof. Let $R_{emp} = R_{\mathcal{I}z_n}$ denote the empirical risk for the γ-clipped cost function. Since $\hat{A}$ is a selection from the canonical strategy of $\mathcal{S}\dot{\mathcal{V}}\mathcal{M}$ and $r_* = r_{\mathcal{P}}(\mathcal{S}\dot{\mathcal{V}}\mathcal{M})$ we can apply Lemma 4.1 with $Q = \mathcal{I}z_n$ to obtain

$$\begin{aligned}&\mathcal{P}_n\Big(R_{\mathcal{P}}(\hat{A}_{z_n}) > r_* + \tfrac{2CK^2}{\gamma n} + \epsilon\Big) \\ &\leq \mathcal{P}_n(\sup_b |R_{\mathcal{P}}(\psi_n(b), b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b), b)| > \tfrac{\epsilon}{2}) + \mathcal{P}_n(\sup_b |R_{emp}(\psi_n(b), b) - R_{\mathcal{P}}(\psi_n(b), b)| > \tfrac{CK^2}{\gamma n} + \tfrac{\epsilon}{4})\end{aligned} \tag{29}$$

and

$$\mathcal{P}_n\Big(R_{\mathcal{P}}(\hat{A}_{z_n}) < r_* - \epsilon\Big) \leq \mathcal{P}_n(\sup_b |R_{\mathcal{P}}(\psi_n(b), b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b), b)| > \epsilon) \tag{30}$$

We first address the $\mathcal{P}_n(\sup_b |R_{\mathcal{P}}(\psi_n(b), b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b), b)| > \xi)$ term which appears in inequality (29) with $\xi = \epsilon/2$ and (30) with $\xi = \epsilon$. To do so observe that Lemma 3.2 implies that $\mathcal{S}\dot{\mathcal{V}}\mathcal{M}$ is Lipschitz with respect to the metric d_γ in Lemma 3.2. Consequently we can apply Lemma 2.1 with pseudometric $2d_\gamma$ to the stochastic process $(|R_{\mathcal{P}}(\psi_n(b), b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b), b))|)_{b\in\mathcal{B}}$ to obtain

$$\begin{aligned}&\mathcal{P}_n\Big(\sup_b |R_{\mathcal{P}}(\psi_n(b), b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b), b)| > \xi\Big) \\ &\leq N(\mathcal{B}, d_\gamma, \xi/4) \sup_{i=1,..,N(\mathcal{B},d_\gamma,\alpha)} \mathcal{P}_n\Big(|R_{\mathcal{P}}(\psi_n(b_i), b_i) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b_i), b_i)| > \xi/2\Big)\end{aligned} \tag{31}$$

We now appeal to a corollary of Steinwart's (Steinwart, 2003b) extension to infinite dimensions of a result of Zhang (Zhang, 2001).

Theorem 4.2. *(Zhang, Steinwart) Let $\psi_n(b)$ and $\psi_{\mathcal{P}}(b)$ be as above. Then there exists a function $h : Z \to [-1, 1]$ such that for all probability measures Q we have*

$$|\psi_n(b) - \psi_Q(b)| \leq C\left|\frac{1}{n}\sum_{j=1}^{n} z_n(j)h(z_n(j)) - E_{z\sim Q}zh(z)\right| \tag{32}$$

Furthermore, for every $\acute{\xi} \geq 0$

$$\mathcal{P}_n(|\psi_n(b) - \psi_{\mathcal{P}}(b)| > \acute{\xi}) \leq 2e^{-\frac{n\acute{\xi}^2}{8C^2K^2+2\acute{\xi}CK}}. \tag{33}$$

Consequently we obtain

$$\mathcal{P}_n(|R_{\mathcal{P}}(\psi_n(b),b)-R_{\mathcal{P}}(\psi_{\mathcal{P}}(b),b))| > \xi/2) \leq \mathcal{P}_n(\frac{1}{\gamma}K|\psi_n(b)-\psi_{\mathcal{P}}(b)| > \xi/2) \leq 2e^{-\frac{n\xi^2}{32\frac{C^2K^4}{\gamma^2}+\frac{4\xi CK^2}{\gamma}}}. \tag{34}$$

and using (31) and Lemma 3.3 we obtain

$$\mathcal{P}_n(\sup_b |R_{\mathcal{P}}(\psi_n(b),b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b),b)| > \xi) \leq 2(\frac{64(\sqrt{C}K+\sqrt{2})^3}{\gamma^2\xi^2}+1)e^{-\frac{n\xi^2}{32\frac{C^2K^4}{\gamma^2}+\frac{4\xi CK^2}{\gamma}}}.$$

When $0 \leq \xi \leq 1$ this easily simplifies to

$$\mathcal{P}_n(\sup_b |R_{\mathcal{P}}(\psi_n(b),b) - R_{\mathcal{P}}(\psi_{\mathcal{P}}(b),b)| > \xi) \leq 2(\frac{64(\sqrt{C}K+\sqrt{2})^3}{\gamma^2\xi^2}+1)e^{-\frac{n\xi^2}{8(2\frac{CK^2}{\gamma}+1)^2}}. \tag{35}$$

Setting $\xi = \epsilon$ and utilizing (30) obtains the second result of the theorem.

We now consider the second term in the inequality (29). As mentioned in the proof of Theorem 2.1 the stochastic process $|R_{emp}(\psi_n(b),b) - R_{\mathcal{P}}(\psi_n(b),b)|_{b\in\mathcal{B}}$ is Lipschitz with respect to the pseudometric $2d_\gamma$ so we can apply Lemma 2.1 to obtain

$$\begin{aligned}&\mathcal{P}_n\Big(\sup_b |R_{\mathcal{P}}(\psi_n(b),b) - R_{emp}(\psi_n(b),b)| > \tfrac{\epsilon}{4} + \tfrac{CK^2}{\gamma n}\Big)\\ &\leq N(\mathcal{B},d_\gamma,\epsilon/16)\sup_b \mathcal{P}_n\Big(|R_{\mathcal{P}}(\psi_n(b),b) - R_{emp}(\psi_n(b),b)| > \epsilon/8 + \tfrac{CK^2}{\gamma n}\Big).\end{aligned} \tag{36}$$

Utilizing the result of Bousquet et al (5) and Lemma 3.3 for the righthand side we obtain

$$\mathcal{P}_n\Big(\sup_b |R_{\mathcal{P}}(\psi_n(b),b) - R_{emp}(\psi_n(b),b)| > \frac{\epsilon}{4}+\frac{CK^2}{\gamma n}\Big) \leq \Big(\frac{1024(\sqrt{C}K+\sqrt{2})^3}{\gamma^2\epsilon^2}+1\Big)e^{-\frac{n\epsilon^2}{32(2\frac{CK^2}{\gamma}+1)^2}} \tag{37}$$

Combining this result with (35) with $\xi = \epsilon/2$ in (29) obtains the second result of Theorem 4.1. ♦

We note that as a consequence of the proof we can conclude bounds on estimation error for the $b = 0$ soft margin classifier by setting $\mathcal{B} = \{0\}$. Then the complexity due to the covering numbers disappears and the bounds improve in concentration.

5 Parallelization

Highly parallelizable approximations to Lipschitz graphical learning models appear easy to construct. For example consider the estimation error model $\mathcal{A}$ from line (19) and substitute a parallel approximation A to it in the following way. Let $\tilde{u}_i$ be the centroids of a cover $\mathcal{O}$ of $\mathcal{U}$ by balls of radius α and let $\tilde{\mathcal{U}} = \{\tilde{u}_i\}$ be the set of centroids. Define the α-approximate model

$$\mathsf{A}_Q = \{(A_Q^u, u) : u \in \arg\min_{u'\in\tilde{\mathcal{U}}} R_Q(A_Q^{u'}, u')\} \tag{38}$$

This model consists of $|\mathcal{O}| \geq N(\mathcal{U}, d, \alpha)$ parallel evaluations of A^u_Q for $u \in \tilde{\mathcal{U}}$ followed by a minimization of $R_Q(A^u_Q, u)$ over the set $\tilde{\mathcal{U}}$ of size $|\mathcal{O}|$. A fully parallel algorithm for the canonical strategy follows directly. Although in theory one can choose a cover such that $|\mathcal{O}| = N(\mathcal{U}, d, \alpha)$ it is not essential. As far as generalization performance is concerned we can prove the following.

Lemma 5.1. *Consider a family $\mathcal{A}^u : \mathfrak{P} \to \mathcal{F}, u \in \mathcal{U}$, a cost function c, and its corresponding family of risk operators $R_Q, Q \in \mathfrak{P}$. Let $r_{\mathcal{P}}(\mathcal{A})$ denote the model risk (20) for the estimation error model $\mathcal{A}$ (19) at $\mathcal{P}$. Let $Lip(c)$ be the Lipschitz norm of the cost function c in its first argument. Let $\hat{\mathsf{A}}$ be a selection from the parallel approximate learning model A (38). Then for every $Q \in \mathfrak{P}$ we have*

$$R_{\mathcal{P}}(\hat{\mathsf{A}}_Q) - r_{\mathcal{P}}(\mathcal{A}) \leq \alpha Lip(c) + \sup_u (R_{\mathcal{P}}(\mathcal{A}^u_Q, u) - R_{\mathcal{P}}(\mathcal{A}^u_{\mathcal{P}}, u)) + 2\sup_u |R_Q(\mathcal{A}^u_Q, u)) - R_{\mathcal{P}}(\mathcal{A}^u_Q, u)|$$

and

$$r_{\mathcal{P}}(\mathcal{A}) - R_{\mathcal{P}}(\hat{\mathsf{A}}_Q) \leq \sup_u (R_{\mathcal{P}}(\mathcal{A}^u_{\mathcal{P}}, u) - R_{\mathcal{P}}(\mathcal{A}^u_Q, u)).$$

Proof. Since A is the graphical model (19) over $\tilde{\mathcal{U}}$ we obtain the inequalities of Lemma 4.1 in terms of $r_{\mathcal{P}}(\mathsf{A}) = \min_{u \in \tilde{\mathcal{U}}} R_{\mathcal{P}}(\mathcal{A}^u_{\mathcal{P}}, u)$. Since $r_{\mathcal{P}}(\mathsf{A}) \geq r_{\mathcal{P}}(\mathcal{A})$ and $r_{\mathcal{P}}(\mathsf{A}) \leq r_{\mathcal{P}}(\mathcal{A}) + \alpha Lip(c)$ the proof is finished. ♦

Lemma 5.1 can be use to prove bounds on estimation error with respect to $r_{\mathcal{P}}(\mathcal{A})$ in the same way we used Lemma 4.1 in the proof of Theorem 4.1 at lines (29) and (30). The bound for above $r_{\mathcal{P}}(\mathcal{A})$(as in Theorem 4.1) will contain the penalty $\alpha Lip(c)$ indicating the performance price paid for using the α-approximate model (38).

6 Model Selection

The VC theorem allows the study of the empirical error minimization model to be analysed in terms of the approximation error $e_{\mathcal{H}} = \inf_{h \in \mathcal{H}} e(h)$ and the VC dimension of the hypothesis class $\mathcal{H}$. In a similar way estimation error models may be analysed in terms of their model risk r and their estimation error bounds such as in Theorem 4.1. Therefore, we can begin the study of estimation error model selection.

We begin by investigating what assumptions concerning $\mathfrak{P}$ can say about model selection. In particular, we discuss some observations made in (Howse et al., 2001) concerning the term CK^2 appearing in the results of Theorems 3.2,3.3, and 4.1. Let us distinguish between the assumption $|supp(x)| \leq K$ and the assumption "it is known that $|supp(x)| \leq K$". In the latter case we choose C to depend upon K. Let C_K note such a dependence. Let $\psi = K^{-1}\acute{\psi}$ and write $\acute{Q}$ for the transformation of Q induced by scaling $x = K\acute{x}$ such that $|supp(\acute{Q}_x)| \leq 1$. Then it is easy to see that if we write the criterion with dependence on C that

$$J^C_Q(\psi, b) = K^{-2} J^{CK^2}_{\acute{Q}}(\acute{\psi}, b)$$

so that

$$C_K K^2 = C_1$$

implies scaling covariance. Therefore when $|supp(x)| \leq K$ is known we would choose the constant $C = \frac{C_1}{K^2}$. Then the bounds of Theorems 3.2, 3.3, and 4.1, can be written in terms of $C_K K^2 = C_1$ and the dependence of the bounds on K disappears.

We change tact. Suppose we consider that the estimation error model $\mathcal{S}\dot{\mathcal{V}}\mathcal{M}$ is an appealing alternative to $\mathcal{SVM}$ since $r_{\mathcal{P}}(\mathcal{S}\dot{\mathcal{V}}\mathcal{M}) \leq r_{\mathcal{P}}(\mathcal{SVM})$ (26) and it possesses the estimation error bounds of Theorem 4.1. However, $\mathcal{SVM}$ has special structure that allows an alternative estimation error model with desirable properties. As mentioned in the proof of Theorem 3.2 the H component of $\mathcal{SVM}_Q$ is a constant ψ_Q. Therefore if we define

$$\mathcal{S}\ddot{\mathcal{V}}\mathcal{M}_Q = \{(\psi_Q, b) : b \in \arg\min_{b \in \mathcal{B}} R_Q(\psi_Q, b)\} \tag{39}$$

where

$$\psi_Q = P_H(\mathcal{SVM}_Q)$$

is this constant (in b), with model risk at $\mathcal{P}$

$$r_{\mathcal{P}}(\mathcal{S}\ddot{\mathcal{V}}\mathcal{M}) = \inf_{b \in \mathcal{B}} R_{\mathcal{P}}(\psi_{\mathcal{P}}, b), \tag{40}$$

it follows that

$$r_{\mathcal{P}}(\mathcal{S}\ddot{\mathcal{V}}\mathcal{M}) \leq r_{\mathcal{P}}(\mathcal{SVM}).$$

In addition since ψ_Q is constant in b there is no need for parallelization and the covering number complexity can be reduced as in going from Theorem 3.2 to Theorem 3.3 and in the proof of a version of Theorem 4.1. Namely we can obtain that for any selection from the canonical strategy of $\mathcal{S}\ddot{\mathcal{V}}\mathcal{M}$, if $n \geq \frac{(2CK^2+\gamma)^2}{32(\sqrt{C}K+1)^2} \ln \frac{4}{\delta}$, then with probability greater than $1 - \delta$ we have

$$r_{\mathcal{P}}(\mathcal{S}\ddot{\mathcal{V}}\mathcal{M}) - \epsilon(\delta) \leq R_{\mathcal{P}} \leq r_{\mathcal{P}}(\mathcal{S}\ddot{\mathcal{V}}\mathcal{M}) + \frac{2CK^2}{\gamma n} + \epsilon(\delta)$$

where

$$\epsilon(\delta) = \frac{4\sqrt{2}(\frac{2CK^2}{\gamma} + 1)}{\sqrt{n}} \sqrt{\ln\Big(8\sqrt{n}\frac{\sqrt{C}K + 1}{2CK^2 + \gamma}\Big) + \ln\frac{4}{\delta}}.$$

Now consider that the choice of regularization constant C in the criterion $J_Q(\psi, b) = |\psi|^2 + CE_Q\eta_{(\psi,b)}$ defining $\mathcal{SVM}$ (24) is an important open question in the design of support vector machines. Let us treat the following model which gives a solution to this problem: assume

$$\mathcal{F} = \mathcal{S} \times \mathcal{B} \times \mathcal{C}$$

where $\mathcal{C} \subset \mathbb{R}_+$ has no effect on the functions:

$$(\psi, b, C)(x) = \psi \cdot x + b.$$

Define the family

$$\mathcal{A}_Q^C = \big\{(\psi_Q^C, b_Q^C) : (\psi_Q^C, b_Q^C) \in \arg\min_{(\psi,b)} (|\psi|^2 + CE_Q\eta_{(\psi,b)})\big\}$$

to be the solutions of the soft margin criterion for fixed C and Q. For a given cost function consider the estimation error model (19) defined by

$$\overline{\mathcal{SVM}}_Q = \{(\psi_Q^C, b_Q^C) : C \in \arg\min_{C' \in \mathcal{C}} R_Q(\psi_Q^{C'}, b_Q^{C'})\}. \tag{41}$$

with model risk at $\mathcal{P}$

$$r_{\mathcal{P}}(\overline{\mathcal{SVM}}) = \inf_{C \in \mathcal{C}} R_{\mathcal{P}}(\psi_{\mathcal{P}}^C, b_{\mathcal{P}}^C). \tag{42}$$

In almost any application SVMs use a kernel $k : X \times X \to \mathbb{R}$ which maps the data from the input space X to the *reproducing kernel Hilbert space* H of k. In particular, the optimization problem (7) is then solved in H. Furthermore, recall that a kernel on a compactum X is said to be *universal* if its reproducing kernel Hilbert space is dense in $C(X)$, (cf. (Steinwart, 2001)). The best known universal kernel is the Gaussian RBF kernel $k(x, x') = \exp(-\sigma^2 |x - x'|^2)$.

Now if the cost function involved in (41) is the misclassification cost, let $e_Q(f)$ denote the misclassification risk of a function $f : X \to \mathbb{R}$. Then $\overline{\mathcal{SVM}}$ chooses the pairs (ψ_Q^C, b_Q^C), $C \in \mathcal{C}$ with the smallest misclassification risk and the model risk (42) is

$$\mathfrak{e}_{\mathcal{P}}(\overline{\mathcal{SVM}}) = \inf_{C \in \mathcal{C}} e_{\mathcal{P}}(\psi_{\mathcal{P}}^C, b_{\mathcal{P}}^C).$$

Furthermore, if a universal kernel is used and $\mathcal{C}$ is unbounded we have

$$\mathfrak{e}_{\mathcal{P}}(\overline{\mathcal{SVM}}) = e_{Bayes}\,,$$

where e_{Bayes} is the Bayes risk. In other words, the model risk of $\overline{\mathcal{SVM}}$ equals the Bayes risk in this case. The learning strategy of $\overline{\mathcal{SVM}}$ with misclassification cost function implements empirical risk minimization over $\mathcal{C}$ to determine (ψ_n^C, b_n^C, C). However, for empirical data using a universal kernel and an unbounded set $\mathcal{C}$ leads to overfitting since for all training sets without contradicting samples (ψ_n^C, b_n^C) achieves zero empirical misclassification risk whenever C is large enough, (cf. (Steinwart, 2001)). Hence we assume that $\mathcal{C} = \mathcal{C}_n$ is bounded and depends on the size n of the training set. The following theorem shows that under some assumptions all learning algorithms (with the restrictions made in Theorem 3.2) based on the described strategy are universally consistent:

Theorem 6.1. *Let $k : X \times X \to \mathbb{R}$ be a continuous kernel with reproducing kernel Hilbert space H. Assume, that X is compact and k is universal. Furthermore, let $\mathcal{C}_n \subset \mathbb{R}^+$ be such that $\inf \mathcal{C}_n \to \infty$ and $\sqrt{\frac{\log n}{n}} \sup \mathcal{C}_n \to 0$ for $n \to \infty$. For each n-sample z_n choose an arbitrary regularization constant $C_{z_n} \in \mathcal{C}_n$. Then we have*

$$e_{\mathcal{P}}(\psi_n^{C_{z_n}}, b_n^{C_{z_n}}) \;\to\; e_{Bayes}$$

in probability for $n \to \infty$. If k is a Gaussian RBF kernel on $X \subset \mathbb{R}^d$ the condition $\sqrt{\frac{\log n}{n}} \sup \mathcal{C}_n \to 0$ can be replaced by the weaker assumption $\frac{1}{n} \sup \mathcal{C}_n |\log \sup \mathcal{C}_n|^{d+1} \to 0$.

Proof. The proof essentially consists of the following simple observation: for all $C \in \mathcal{C}_n$ we have

$$\begin{aligned} |\psi_Q^C|_H &\le \sqrt{C} &&\le \sqrt{\sup \mathcal{C}_n} =: c_n\,, \\ |b_Q^C|_H &\le 1 + \sqrt{C}K &&\le 1 + \sqrt{\sup \mathcal{C}_n}K\,, \end{aligned}$$

where $K := \sup_{x\in X} k(x,x)$. Like in the proof of Lemma III.10 (and Example III.9 for the Gaussian RBF kernel) in (Steinwart, 2003a) we also find

$$\sup_{\substack{|\psi|\le c_n \\ |b|\le 1+c_n K}} \left|E_n\eta_{(\psi,b)} - E_{\mathcal{P}}\eta_{(\psi,b)}\right| \to 0$$

in probability for $n\to\infty$. In particular this yields

$$\left|E_n\eta_{(\psi_n^{C_{zn}},b_n^{C_{zn}})} - E_{\mathcal{P}}\eta_{(\psi_n^{C_{zn}},b_n^{C_{zn}})}\right| \to 0.$$

The rest of the proof follows the approach discussed in the introduction of (Steinwart, 2003a). ♦

Finally we combine the ideas of the models $\ddot{\mathcal{SVM}}$ and $\overline{\mathcal{SVM}}$: again, let $\mathcal{F} = \mathcal{S}\times\mathcal{B}\times\mathcal{C}$ where $\mathcal{U} = \mathcal{B}\times\mathcal{C}$ is the new unstable component, and—as above— $\mathcal{C}\subset\mathbb{R}_+$ has no effect on the functions, $(\psi,b,C)(x) = \psi\cdot x + b$. Define the family

$$\mathcal{A}_Q^{b,C} = \psi_Q^C = \arg\min_{\psi}\left(|\psi|^2 + CE_Q\eta_{(\psi,b)}\right)$$

to be the ψ component of the fixed C soft margin solution determined by Q. For a given cost function consider the estimation error model (19) defined by

$$\tilde{\mathcal{SVM}}_Q = \{(\psi_Q^{C_Q}, b_Q) : (b_Q, C_Q) \in \arg\min_{(b,C)\in\mathcal{B}\times\mathcal{C}} R_Q(\psi_Q^C, b)\}. \tag{43}$$

with model risk at $\mathcal{P}$

$$r_{\mathcal{P}}(\tilde{\mathcal{SVM}}) = \inf_{(b,C)\in\mathcal{B}\times\mathcal{C}} R_{\mathcal{P}}(\psi_{\mathcal{P}}^C, b). \tag{44}$$

Note, that in contrast to $\ddot{\mathcal{SVM}}$ and $\overline{\mathcal{SVM}}$ the model $\tilde{\mathcal{SVM}}$ determines both parameters b and C by risk minimization. Furthermore, if we denote by $\mathcal{SVM}(C)$ the model $\mathcal{SVM}$'s dependence on C, with a similar meaning for $\ddot{\mathcal{SVM}}(C)$, it follows that

$$r_{\mathcal{P}}(\tilde{\mathcal{SVM}}) = \inf_{C\in\mathcal{C}} r_{\mathcal{P}}(\ddot{\mathcal{SVM}}(C)) \le \inf_{C\in\mathcal{C}} r_{\mathcal{P}}(\mathcal{SVM}(C)). \tag{45}$$

If the chosen cost function is the γ-clipped cost function the next theorem shows that under some assumptions all learning algorithms based on this model are universally consistent:

Theorem 6.2. *Let $k: X\times X\to\mathbb{R}$ be a universal kernel on a compact space X and $\mathcal{P}$ such that $\mathcal{P}\big(x\in X : \mathcal{P}(1|x) = 1/2\big) = 0$. Furthermore, let $\mathcal{C}_n\subset\mathbb{R}^+$ be such that $\inf\mathcal{C}_n\to\infty$ and $\frac{1}{\sqrt{n}}\sup\mathcal{C}_n\to 0$ for $n\to\infty$. For each n-sample z_n choose (C_{z_n}, b_{z_n}) according to the model (43) with respect to the γ-clipped loss function. Then we have*

$$e_{\mathcal{P}}(\psi_n^{C_{zn}}, b_{z_n}) \;\to\; e_{Bayes}$$

in probability for $n\to\infty$.

Proof. By our assumption on $\mathcal{P}$ and Theorem 3.9 in (Steinwart, 2003b) we see that for all sequences (f_n) of functions with $E_{\mathcal{P}}\eta_{(f_n)} \to \min_{f:X\to\mathbb{R}} E_{\mathcal{P}}\eta_{(f)}$ we have

$$R_{\mathcal{P}}(f_n) \to \min_{f:X\to\mathbb{R}} R_{\mathcal{P}}(f) =: R_{\mathcal{P}} ,$$

where the risk is with respect to the γ-clipped loss function. In particular we find

$$R_{\mathcal{P}} \leq r_{\mathcal{P}}(\tilde{\mathcal{SVM}}) \leq \inf_{C\in\mathcal{C}_n} r_{\mathcal{P}}(\mathcal{SVM}(C)) \to R_{\mathcal{P}}$$

by (45). By our assumption on $\sup\mathcal{C}_n$ Theorem 4.2 guarantees

$$||\psi_n^{C_{\mathcal{P}}} - \psi_{\mathcal{P}}^{C_{\mathcal{P}}}||_\infty \to 0$$

in probability. Furthermore, using the idea of the proof of Theorem 6.1 the results of (Steinwart, 2003a) yields

$$\left|R_{emp}(\psi_n^{C_{z_n}}, b_{z_n}) - R_{\mathcal{P}}(\psi_n^{C_{z_n}}, b_{z_n})\right| + \left|R_{emp}(\psi_n^{C_{\mathcal{P}}}, b_{\mathcal{P}}) - R_{\mathcal{P}}(\psi_n^{C_{\mathcal{P}}}, b_{\mathcal{P}})\right| \to 0$$

in probability. Hence for all $\varepsilon > 0$ and sufficiently large n the following estimate is true with high probability

$$\begin{aligned}
R_{\mathcal{P}}(\psi_n^{C_{z_n}}, b_{z_n}) &\leq R_{emp}(\psi_n^{C_{z_n}}, b_{z_n}) + \varepsilon \\
&\leq R_{emp}(\psi_n^{C_{\mathcal{P}}}, b_{\mathcal{P}}) + \varepsilon \\
&\leq R_{emp}(\psi_{\mathcal{P}}^{C_{\mathcal{P}}}, b_{\mathcal{P}}) + 2\varepsilon \\
&\leq R_{\mathcal{P}}(\psi_{\mathcal{P}}^{C_{\mathcal{P}}}, b_{\mathcal{P}}) + 3\varepsilon \\
&\leq R_{\mathcal{P}} + 4\varepsilon .
\end{aligned}$$

Now the assertion follows since c_γ is admissible in the sense of (Steinwart, 2003a). ♦

7 Experiments

This section describes experimental results for the $\mathcal{SVM}$, $\ddot{\mathcal{SVM}}$, $\overline{\mathcal{SVM}}$ and $\tilde{\mathcal{SVM}}$ canonical learning strategies applied to three different data sets. The first two data sets are synthetically generated according to Fukunaga's so-called I-$4I$ and I-Λ distributions (Fukunaga, 1990), and the third is the **Spambase** data set from the UCI repository (Blake & Merz, 1998).

For the synthetic data sets we set $d = 8$ and generate samples from $\mathbb{R}^d \times \{-1, 1\}$ according to the I-$4I$ and I-Λ distributions. For both distributions the class marginals are $\mathcal{P}(y = -1) = \mathcal{P}(y = 1) = 0.5$ and the class conditional distributions are Gaussian. For the I-$4I$ distribution the class conditional means and covariances are

$$\begin{aligned}
&\mu_{-1} = \mu_1 = 0 \\
&\Sigma_{-1} = I, \quad \Sigma_1 = 4I,
\end{aligned}$$

and for the I-Λ distribution they are

$$\begin{aligned}
&\mu_{-1} = 0, \quad \mu_1 = (3.86, 3.10, 0.84, 0.84, 1.64, 1.08, 0.26, 0.01) \\
&\Sigma_{-1} = I, \quad \Sigma_1 = diag(8.41, 12.06, 0.12, 0.22, 1.49, 1.77, 0.35, 2.73).
\end{aligned}$$

For each distribution we generate a training set with $n = 200$ samples and a test set with $100,000$ samples.

The **Spambase** data set contains 4601 samples from $\mathbb{R}^d_+ \times \{-1, 1\}$ where $d = 57$. This data set contains 1813 samples with $y = -1$ and 2788 samples with $y = 1$. We perform random sampling (without replacement) to split this data into a training set with $n = 3601$ samples and a test set with 1000 samples. We also normalize the data as follows. We compute (univariate) sample means and standard deviations for each of the 57 input dimensions over the 3601 training samples, and then normalize all 4601 samples by subtracting the means and dividing by the standard deviations.

We perform the following experiments. For the I-$4I$ and I-Λ data we employ the Gaussian RBF kernel $k(x, x') = \exp(-|x - x'|^2/d)$. For the higher dimensional **Spambase** data no kernel is used. For each of the three data sets and for each value of C from the list

$$.01,\ .05,\ .1,\ .5,\ 1,\ 5,\ 10,\ 50,\ 100,\ 500,\ 1000,\ 5000,\ 10000$$

we apply the canonical learning strategy for $\mathcal{SVM}$ to the training set to obtain a classifier (ψ_n, b_n) that minimizes the soft margin criterion, and we apply the canonical learning strategy for $\mathcal{S\ddot{V}M}$ to the training set to obtain a classifier $(\ddot{\psi}_n, \ddot{b}_n)$ where $\ddot{\psi}_n = \psi_n$ minimizes the soft margin criterion and $\ddot{b}_n$ minimizes the γ–clipped risk $R_{emp}(\ddot{\psi}_n, \cdot)$ with $\gamma = 0.01$. We then use the test sets to compute estimates of the γ–clipped risks $R_{\mathcal{P}}$ for these classifiers.

Results are shown in Figure 1 where we provide two plots for each of the three data sets; the first plot shows the clipped risks $R_{\mathcal{P}}$ (solid lines) and R_{emp} (dashed lines) with $\gamma = 0.01$ for the $\mathcal{SVM}$ (denoted by the $\square$ symbol) and $\mathcal{S\ddot{V}M}$ (denoted by $\times$ symbol) classifiers, and the second plot shows the classification error $e_{\mathcal{P}}$ (i.e. $R_{\mathcal{P}}$ with $\gamma = 0$) for these same classifiers. The second plots also show the Bayes classification error e_{Bayes} for the I-$4I$ and I-Λ distributions, and the classification error $e_{\mathcal{P}}$-*doc* reported in the documentation for the **Spambase** data.

These results exhibit several noteworthy characteristics. First, the risk deviances $|R_{\mathcal{P}} - R_{emp}|$ are small, especially for the **Spambase** data where a larger training set is used. Second, both the $\mathcal{SVM}$ and $\mathcal{S\ddot{V}M}$ strategies achieve nearly the same smallest risk value with the appropriate choices of C. In addition these smallest risk values are quite good in that they are close to the Bayes risk for the I-$4I$ and I-Λ distributions and close to the documented risk value for the **Spambase** data. Of the two strategies, $\mathcal{S\ddot{V}M}$ seems more attractive. In all three cases it achieves near–optimal risk values over a larger range of C suggesting that it is more robust to the choice of C. In addition, since soft margin algorithms are typically faster for smaller C, and $\mathcal{S\ddot{V}M}$ achieves smaller risks at smaller values of C, learning may be computationally more efficient for $\mathcal{S\ddot{V}M}$.

Finally, although we did not implement them directly we can infer results for $\overline{\mathcal{SVM}}$ and $\mathcal{S\tilde{V}M}$ as follows. A brute force implementation of the canonical strategy for $\overline{\mathcal{SVM}}$ can be obtained by employing the $\mathcal{SVM}$ strategy at all values of $C \in \mathcal{C}$ to produce a set of classifiers, and then choosing a classifier from this set that minimizes the empirical risk. A brute force implementation of the canonical stratgy for $\mathcal{S\tilde{V}M}$ can be obtained by employing a similar procedure with the $\mathcal{S\ddot{V}M}$ strategy. Thus, our experiments allow us to infer results for $\overline{\mathcal{SVM}}$ and $\mathcal{S\tilde{V}M}$ where $\mathcal{C}$ is the finite set of points listed above. The fact that the empirical risk will be zero for sufficiently large C when universal kernels are used (cf. (Steinwart, 2001)) is verified by our results for the I–I and I–Λ data. In addition larger values of C can lead to larger risk deviance, as predicted by the theorems in previous sections and validated by our experiments (e.g. this is most apparent in the results for the I-$4I$ data set in Figure 1(a)). This partially explains why the values of C that minimize empirical risk are larger than those that minimize risk, and why the $\overline{\mathcal{SVM}}$ and $\mathcal{S}\tilde{\mathcal{V}}\mathcal{M}$ strategies may overfit when $\mathcal{C}$ contains sufficiently large values. Nevertheless, the excess risk that results from these strategies is small in our experiments, suggesting that these simple stratgegies may provide a useful method for choosing C in practice.

References

Barbu, V., & Precupanu, T. (1978). *Convexity and optimization in banach spaces.* The Netherlands: Sijthoff and Noordhoff Publishers.

Blake, C., & Merz, C. (1998). *UCI repository of machine learning databases.* http://www.ics.uci.edu/∼mlearn/MLRepository.html: University of California, Irvine, Dept. of Information and Computer Sciences.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition.* New York, NY: Springer.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). San Diego, CA: Academic Press.

Howse, J., Hush, D., & Scovel, C. (2001). Linking learning strategies and performance for support vector machines. *unpublished.*

Kutin, S., & Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. *University of Chicago Report, TR-2002-03.*

Rockafellar, R. (1970). *Convex analysis.* Princeton: Princeton University Press.

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

Steinwart, I. (2002). Support vector machines are universally consistent. *J. Complexity*, *18*, 768–791.

Steinwart, I. (2003a). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory, accepted with revisions.*

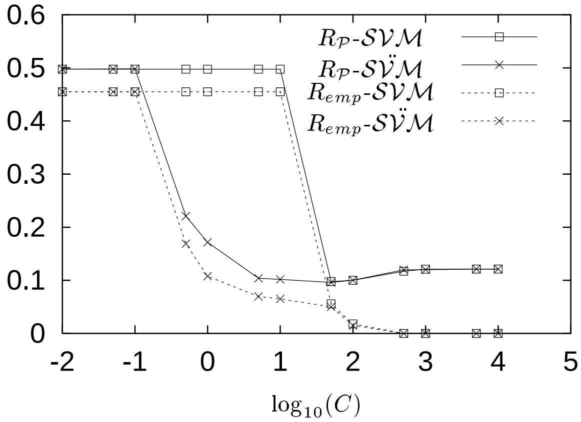
Steinwart, I. (2003b). Sparseness of support vector machines. *Journal of Machine Learning Research, accepted with revisions.*

Vapnik, V. N. (1998). *Statistical learning theory.* New York: John Wiley and Sons, Inc.

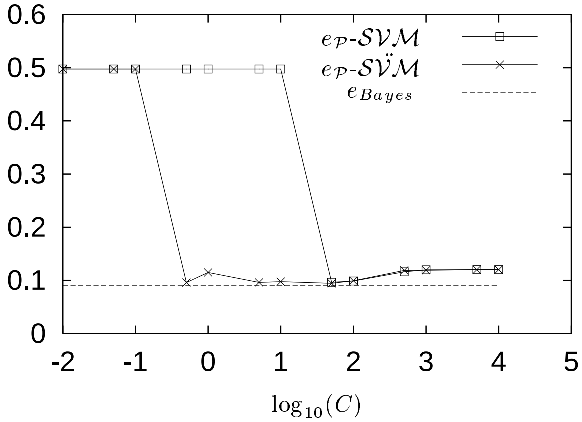
Vidyasagar, M. (1997). *A theory of learning and generalization.* Berlin: Springer-Verlag.

Zhang, T. (2001). Convergence of large margin separable linear classification. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* **13** (pp. 357–363). MIT Press.

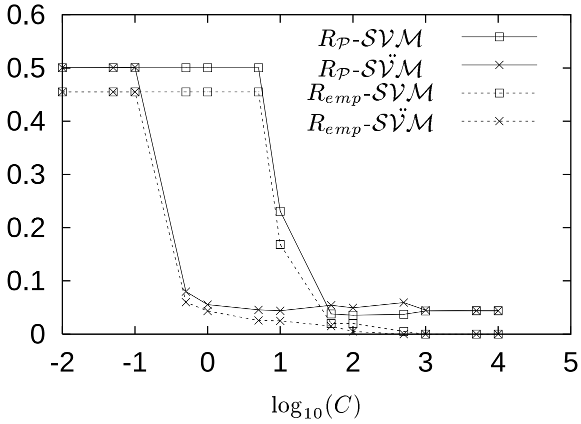
Zhang, T. (2003). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, *to appear.*



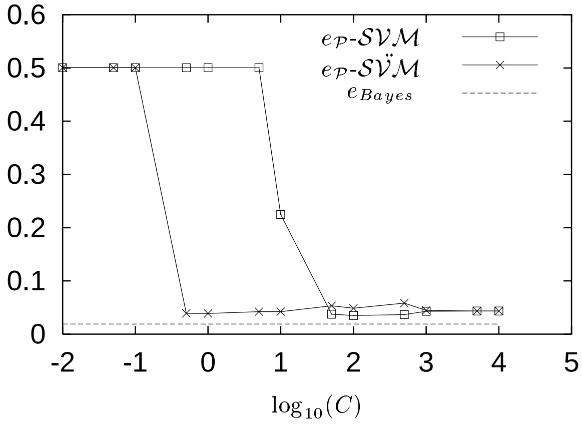
(a) *I*-4*I*



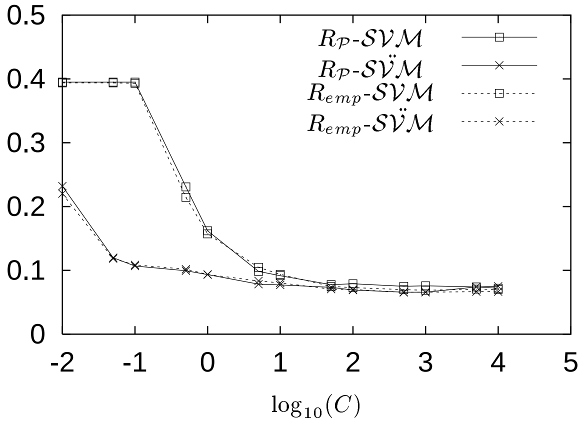
(b) *I*-4*I*



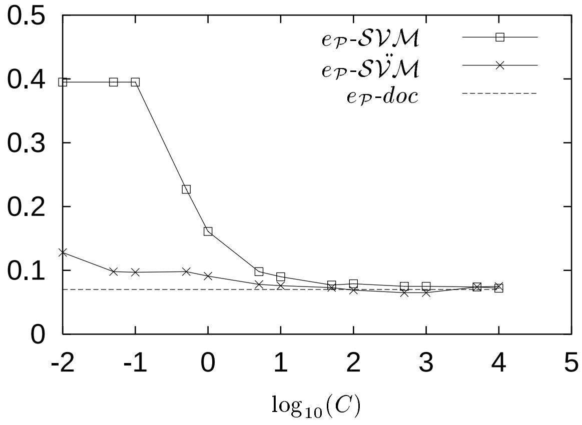
(c) *I*-Λ



(d) *I*-Λ



(e) **Spambase**



(f) **Spambase**

Figure 1: Plots of risk verses $\log_{10}(C)$ for $\mathcal{SVM}$ and $\mathcal{S\ddot{V}M}$ learning strategies applied to the *I*-4*I*, *I*-Λ and **Spambase** data sets.