Mach Learn (2007) 67: 197–206
DOI 10.1007/s10994-007-5004-z

Stability of Unstable Learning Algorithms

Don Hush · Clint Scovel · Ingo Steinwart

Received: 3 July 2003 / Revised: 16 October 2006 /
Accepted: 10 January 2007 / Published online: 30 March 2007

Abstract We introduce graphical learning algorithms and use them to produce bounds on error deviance for unstable learning algorithms which possess a partial form of stability. As an application we obtain error deviance bounds for support vector machines (SVMs) with variable offset parameter.

Keywords Learning · Stability · Generalization

1 Introduction

Bousquet and Elisseeff (2002) determine bounds on the deviance between the empirical and true risk for stable learning algorithms and Kutin and 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 (Boser et al. 1992) 1-norm soft margin support vector machine (SVM) which includes the offset parameter is not stable. In this paper we develop *graphical learning algorithms* to analyse the statistical stability of unstable learning algorithms that possess a partial form of stability 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 Bousquet and Elisseeff obtained without the offset parameter. The results are presented in Sect. 2 and the proofs are contained in Sect. 3.

Editor: Avrim Blum

D. Hush · C. Scovel (✉) · I. Steinwart
CCS-3, Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: jcs@lanl.gov

D. Hush
e-mail: dhush@lanl.gov

I. Steinwart
e-mail: ingo@lanl.gov

2 Results

Consider sets X and Y and let P be a probability measure on $Z := X \times Y$ with corresponding random variable $z = (x, y)$. Let the model space be a nonempty set $\mathcal{F}$ and consider a loss function $L : \mathcal{F} \times Z \to \mathbb{R}$. The risk associated to the measure P and the loss function is defined as

$$\mathcal{R}_{L,P}(f) := \mathbb{E}_{z\sim P}(L(f, z)). \tag{1}$$

Note that if we identify a training set $T = (z_1, \ldots, z_n) \in Z^n$ with its empirical measure, then

$$\mathcal{R}_{L,T}(f) := \frac{1}{n}\sum_{i=1}^{n} L(f, z_i)$$

is the empirical risk of f. The goal of the learning problem is to find an f such that $\mathcal{R}_{L,P}(f)$ is minimal where P is unknown and T is an i.i.d. sample from P. A typical strategy is to minimize $\mathcal{R}_{L,T}$ or a regularized version. However, often solutions to these optimization problems are not unique so we define a learning strategy to be a set-valued map

$$A : Z^n \to 2^{\mathcal{F}},$$

where $2^{\mathcal{F}}$ denotes the power set of $\mathcal{F}$. Let us denote the value of the map A at the sample $T \in Z^n$ by the subset $A_T \subset \mathcal{F}$. Then a learning algorithm is defined as a selection $\hat{A}$ from the set-valued map A. That is,

$$\hat{A} : Z^n \to \mathcal{F}, \quad \text{where } \hat{A}_T \in A_T,\ \forall T \in Z^n.$$

We now define graphical learning strategies which are the key notion of this work.

Definition 2.1 Let the hypothesis space be decomposed into $\mathcal{F} = \mathcal{U} \times \mathcal{S}$. We say that a learning strategy A is graphical with respect to this decomposition if there exists a family $(A^u)_{u\in\mathcal{U}}$ of mappings

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

and a set-valued map

$$U : Z^n \to 2^{\mathcal{U}}$$

such that

$$A_T = \{(u, A^u_T) : u \in U_T\}, \quad \forall T \in Z^n.$$

This definition implies that for each $T \in Z^n$, A_T is the graph over $\mathcal{U}$ determined by the family $(A^u)_{u\in\mathcal{U}}$ restricted to the subset $U_T \subset \mathcal{U}$. In particular the set-valued nature of A is generated by U. In the following we will refer to $\mathcal{S}$ as the stable space and $\mathcal{U}$ as the unstable space.

We define a graphical learning algorithm to be a selection from a graphical learning strategy. It is easy to show that a graphical learning algorithm $\hat{A}$ corresponding to a graphical learning strategy $A_T = \{(u, A^u_T) : u \in U_T\}$ is determined by a selection $\hat{u}$ from U through

$$\hat{A}_T = (\hat{u}_T, A^{\hat{u}_T}_T), \quad \text{where } \hat{u}_T \in U_T, \forall T \in Z^n.$$

The following illustrates an important example of a graphical learning strategy.

Example 2.2 Consider an optimization criterion function J_T and suppose the learning strategy is determined by minimization: $A_T := \arg\min_f J_T(f)$. Then we show that if the minimization problem determined by fixing the unstable variable $A_T^u := \arg\min_s J_T(u, s)$ has a unique solution for all u, then A is graphical. For the proof let us define $U_T := \arg\min_u J_T(u, A_T^u)$. Our goal is to prove that $A_T = \{(u, A_T^u) : u \in U_T\}$ which is clearly graphical. To that end suppose that $(u', s') \in A_T$. Then $J_T(u', s') \leq J_T(u', s)$ for all $s \in \mathcal{S}$ implies that $s' = A_T^{u'}$. Moreover, since $J_T(u', s') \leq J_T(u, s)$ for all (u, s) it follows that $J_T(u', A_T^{u'}) \leq J_T(u, A_T^u)$ for all u and so we conclude that $u' \in U_T$. Consequently we obtain that $A_T \subset \{(u, A_T^u) : u \in U_T\}$. Conversely, consider a point $(u', A_T^{u'})$ with $u' \in U_T$. It follows that $J_T(u', A_T^{u'}) \leq J_T(u, A_T^u)$ and from the definition of A_T^u it follows that $J_T(u, A_T^u) \leq J_T(u, s)$ for all $s \in \mathcal{S}$. Consequently we obtain $J_T(u', A_T^{u'}) \leq J_T(u, s)$ for all (u, s) and so conclude that $\{(u, A_T^u) : u \in U_T\} \subset A_T$.

We now define graphical strategies which are Lipschitz continuous when the unstable parameter is varied. To do so, recall that a pseudometric is a nonnegative bivariate function which is symmetric and satisfies the triangle inequality.

Definition 2.3 Let $(\mathcal{U}, d)$ be a pseudometric space. A graphical learning strategy A with respect to $\mathcal{U} \times \mathcal{S}$ is Lipschitz continuous with respect to the loss function L over a subset $Z_0 \subset Z$ if we have

$$\left|L((u_1, A_T^{u_1}), z) - L((u_2, A_T^{u_2}), z)\right| \leq d(u_1, u_2), \quad \forall z \in Z_0,\ T \in Z_0^n,\ u_1, u_2 \in \mathcal{U}.$$

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

The following theorem provides performance guarantees for a Lipschitz continuous graphical learning algorithm in terms of its performance for fixed values of the unstable parameters and the covering numbers of the unstable space. Recall that for a pseudometric space $(\mathcal{U}, d)$ the covering numbers $N(\mathcal{U}, d, \epsilon)$ are defined by

$$N(\mathcal{U}, d, \epsilon) := \min\left\{n \geq 1 : \exists u_1, \ldots, u_n \in \mathcal{U} \text{ with } \mathcal{U} \subset \bigcup_{i=1}^n B_\varepsilon(u_i)\right\}, \quad \varepsilon > 0,$$

where $B_\varepsilon(u_i)$ denotes the set of points u such that $d(u, u_i) \leq \varepsilon$.

Theorem 2.4 *Consider a pseudometric space $(\mathcal{U}, d)$ and a graphical learning algorithm $\hat{A}$ with respect to $\mathcal{U} \times \mathcal{S}$ which is Lipschitz continuous with respect to the loss function L over Z. Then for any $\delta > 0, \epsilon > 0$ we have*

$$\begin{aligned} &P^n\left(T \in Z^n : \mathcal{R}_{L,P}(\hat{A}_T) - \mathcal{R}_{L,T}(\hat{A}_T) > \delta\right) \\ &\quad \leq N(\mathcal{U}, d, \epsilon) \sup_{u \in \mathcal{U}} P^n\left(T \in Z^n : \mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u) > \delta - 2\epsilon\right) \end{aligned}$$

and

$$\begin{aligned} &P^n\left(T \in Z^n : |\mathcal{R}_{L,P}(\hat{A}_T) - \mathcal{R}_{L,T}(\hat{A}_T)| > \delta\right) \\ &\quad \leq N(\mathcal{U}, d, \epsilon) \sup_{u \in \mathcal{U}} P^n\left(T \in Z^n : |\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)| > \delta - 2\epsilon\right). \end{aligned}$$

Remark 2.5 A common method for defining learning strategies on product spaces $\mathcal{U} \times \mathcal{S}$ can be described as follows. Let $\acute{A} : Z^n \to \mathcal{S}$ denote any $\mathcal{S}$ valued learning algorithm. Then the learning strategy $A_T = \{(u, \acute{A}_T) : u \in U_T\}$ where $U : Z^n \to 2^{\mathcal{U}}$ is any set valued map is obviously graphical and Lipschitz continuous with the trivial pseudometric and therefore satisfies the assumptions of Theorem 2.4. For example, consider SVMs which include an offset parameter (see the discussion at the end of this section for a description of SVMs). It is well known that all solutions to the standard SVM learning strategy have the same Hilbert space component A_T and vary only in their offset parameter. If instead of choosing one of these offsets one chooses the offset parameter by minimizing the empirical classification error, then this theorem applies.

Theorem 2.4 is independent of any notion of stability. However, if a Lipschitz continuous graphical learning algorithms is stable for fixed values of the unstable parameter, Theorem 2.4 may be used to provide performance guarantees through stability arguments. For example, although many important algorithms may not be stable in any of the senses prescribed by Bousquet and Elisseeff (2002) or Kutin and Niyogi (2002), many of them possess this partial form of stability. As an illustration we now demonstrate how Theorem 2.4 can be applied to classification. Here $\mathcal{S}$ and $\mathcal{U}$ are spaces of functions from X to $\mathbb{R}$, $\mathcal{F} = \mathcal{U} + \mathcal{S}$ and $Y = \{-1, 1\}$. However instead of the usual classification loss function we consider a continuous approximation. Namely, for $\gamma \geq 0$ define the γ-clipped loss function L_γ by

$$L_\gamma(f, z) = \begin{cases} 1, & yf(x) \leq 0, \\ 1 - \dfrac{yf(x)}{\gamma}, & 0 < yf(x) \leq \gamma, \\ 0, & yf(x) > \gamma. \end{cases}$$

Note that L_0 is the standard 0–1 classification loss. According to Bousquet and Elisseeff (2002) a real valued classification algorithm $\hat{A}$ has classification stability $\beta \geq 0$ if

$$\|\hat{A}_T - \hat{A}_{T^{-i}}\|_\infty \leq \beta, \quad \forall i, T \in Z_0^n,$$

where $T^{-i} := (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ denotes the $n-1$-sample which one obtains by removing the i-th point from the n-sample $T = (z_1, \ldots, z_n)$. It is said to have uniform stability with respect to the loss function L if

$$\|L(\hat{A}_T, \cdot) - L(\hat{A}_{T^{-i}}, \cdot)\|_\infty \leq \beta, \quad \forall i, T \in Z_0^n.$$

We can now prove performance guarantees for Lipschitz continuous graphical learning algorithms which are stable for fixed values of the unstable parameter.

Theorem 2.6 *Consider a pseudometric space $(\mathcal{U}, d)$ and a graphical learning algorithm $\hat{A}$ with respect to $\mathcal{U} \times \mathcal{S}$ which is Lipschitz continuous with respect to the loss function L_γ over Z. Suppose that for each u, $(u, A^u) : T \mapsto u + A_T^u$ has classification stability β. Then for any $\epsilon > 0$,*

$$P^n\left(T \in Z^n : |\mathcal{R}_{L_\gamma, P}(\hat{A}_T) - \mathcal{R}_{L_\gamma, T}(\hat{A}_T)| > \epsilon + 2\beta\gamma^{-1}\right) \leq N(\mathcal{U}, d, \epsilon/4) e^{-\frac{n\epsilon^2}{2(4n\beta\gamma^{-1}+1)^2}}.$$

Let us now apply Theorem 2.6 to support vector machines *with* offset parameter. To that end, let H denote a reproducing kernel Hilbert space of functions on X with kernel k such

that $K := \sup_{x\in X} \sqrt{k(x,x)} < \infty$. For $\lambda > 0$ we define the unstable space by

$$\mathcal{U}_\lambda := \{b \in \mathbb{R} : |b| \le 1 + \lambda^{-1/2} K\}. \tag{2}$$

Define the model space $\mathcal{F} := \mathcal{U}_\lambda \times H$ and consider the loss function L_γ. To define the SVM, we consider the hinge loss function $L_{\text{hinge}}(f,z) := \max\{0, 1 - yf(x)\}$ and define the SVM criterion function at $f = (b,h) \in \mathcal{U}_\lambda \times H$ to be

$$\mathcal{R}_{T,\lambda}(f) := \lambda \|h\|_H^2 + \mathcal{R}_{L_{\text{hinge}},T}(b+h), \tag{3}$$

where $\lambda > 0$. Consider the fixed-b learning strategy A^b_λ defined by

$$A^b_{T,\lambda} := \arg\min_{h\in H} \mathcal{R}_{T,\lambda}(b,h). \tag{4}$$

It is well known that solutions to (4) are unique. In the following theorem we will consider a general class of SVM learning strategies which are defined by

$$A_\lambda := \{(b, A^b_{T,\lambda}) : b \in U_T\},$$

where $U : Z^n \to 2^{\mathcal{U}_\lambda}$ is some set valued map.

Remark 2.7 The standard SVM

$$A_{T,\lambda} := \arg\min_{f\in\mathbb{R}\times H} \mathcal{R}_{T,\lambda}(f) \tag{5}$$

is such a strategy if we make the stipulation that if all the data have the class label y^* then we choose the solution $(b,h) = (y^*, 0)$. To see this observe that Howse et al. (2002) (see Steinwart and Scovel 2007 for a published and more general result) show that any such selection $\hat{A}_\lambda$ from (5) satisfies $b \in \mathcal{U}_\lambda$ so the a priori restriction from $\mathbb{R}$ to $\mathcal{U}_\lambda$ changes nothing. Moreover, more general SVMs, for which the offset parameter b is determined differently than the standard SVM, are also such strategies. For example consider when $A_{T,\lambda} := \{(b, A^b_{T,\lambda}) : b \in U_T\}$ for $A^b_{T,\lambda}$ defined in (4) and b is defined by minimizing the empirical $\hat{\gamma}$-clipped loss:

$$U_T := \arg\min_{b\in\mathcal{U}_\lambda} \mathcal{R}_{L_{\hat{\gamma}},T}(b + A^b_{T,\lambda})$$

for some $\hat{\gamma} \ge 0$.

Theorem 2.8 *Consider a graphical learning algorithm with respect to $\mathcal{U}_\lambda \times H$ defined by $\hat{A}_\lambda := \{(b, A^b_{T,\lambda}) : b \in U_T\}$ such that $A^b_{T,\lambda}$ satisfies the fixed-b SVM (4) and $U : Z^n \to 2^{\mathcal{U}_\lambda}$ is any set valued map. Then for any $\gamma > 0$ and $\epsilon > 0$ we have*

$$P^n\left(T \in Z^n : |\mathcal{R}_{L_\gamma,P}(\hat{A}_{T,\lambda}) - \mathcal{R}_{L_\gamma,T}(\hat{A}_{T,\lambda})| > \epsilon + \frac{K^2}{\lambda\gamma n}\right)$$
$$\le \left(\frac{64(\lambda^{-\frac{1}{2}}K + \sqrt{2})^3}{\gamma^2\epsilon^2} + 1\right) e^{-\frac{n\epsilon^2}{2(2\frac{K^2}{\lambda\gamma}+1)^2}}. \tag{6}$$

Remark 2.9 Let us state the bound (6) in the alternative form to compare it with the result of Example 2 of Bousquet and Elisseeff (2002) for soft margin SVMs with $b = 0$. In Sect. 3 we prove that if $\delta \leq e^{-1}$ then with probability greater than $1 - \delta$ we have

$$|\mathcal{R}_{L_\gamma,P}(\hat{A}_{T,\lambda}) - \mathcal{R}_{L_\gamma,T}(\hat{A}_{T,\lambda})| \leq \frac{K^2}{\lambda\gamma n} + \frac{\sqrt{2}(\frac{2K^2}{\lambda\gamma}+1)}{\sqrt{n}}\sqrt{\ln\left(32n\frac{(\lambda^{-1/2}K+\sqrt{2})^3}{(2\lambda^{-1}K^2+\gamma)^2}+1\right)+\ln\frac{1}{\delta}}. \tag{7}$$

The bound (7) can be compared with the result of Bousquet and Elisseeff 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 2.6 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 logarithm term. Consequently the stability analysis of Bousquet and Elisseeff (2002), which only worked for $b = 0$, can be carried over to algorithms which choose b in some other way.

3 Proofs

Proof of Theorem 2.4 From Definition 2.1 of a graphical learning strategy we obtain

$$\begin{aligned}\mathcal{R}_{L,P}(\hat{A}_T) - \mathcal{R}_{L,T}(\hat{A}_T) &\leq \sup_{f\in A_T}(\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,T}(f)) \\ &= \sup_{u\in U_T}(\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) \\ &\leq \sup_{u\in\mathcal{U}}(\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)).\end{aligned}$$

Moreover since

$$\left|L((u_1, A_T^{u_1}), z) - L((u_2, A_T^{u_2}), z)\right| \leq d(u_1, u_2), \quad u_1, u_2 \in \mathcal{U}$$

it follows that

$$|\mathcal{R}_{L,P}(u_1, A_T^{u_1}) - \mathcal{R}_{L,T}(u_1, A_T^{u_1}) - \mathcal{R}_{L,P}(u_2, A_T^{u_2}) + \mathcal{R}_{L,T}(u_2, A_T^{u_2})| \leq 2d(u_1, u_2),$$
$$u_1, u_2 \in \mathcal{U}.$$

Let $\mathcal{O} \subset \mathcal{U}$ denote an ϵ-net of size $N(\mathcal{U}, d, \epsilon)$. We conclude that

$$\sup_{u\in\mathcal{U}}(\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) \leq \sup_{u\in\mathcal{O}}(\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) + 2\epsilon.$$

It then follows that

$$\begin{aligned}&P^n\left(T \in Z^n : \sup_{u\in\mathcal{U}}(\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) > \delta\right) \\ &\leq P^n\left(T \in Z^n : \sup_{u\in\mathcal{O}}(\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) > \delta - 2\epsilon\right)\end{aligned}$$

$$\leq N(\mathcal{U}, d, \epsilon) \sup_{u \in \mathcal{O}} P^n\big(T \in Z^n : (\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) > \delta - 2\epsilon\big)$$

$$\leq N(\mathcal{U}, d, \epsilon) \sup_{u \in \mathcal{U}} P^n\big(T \in Z^n : (\mathcal{R}_{L,P}(u, A_T^u) - \mathcal{R}_{L,T}(u, A_T^u)) > \delta - 2\epsilon\big).$$

The result for the absolute value follows in a similar way. □

Proof of Theorem 2.6 It is easy to see that classification stability β implies uniform stability β/γ for L_γ. Furthermore note that the proof of Theorem 12 of Bousquet and Elisseeff (2002) can be slightly modified to provide the same result for the absolute value of the risk deviance for real classification algorithms with uniform stability β/γ. That is, since $|L_\gamma(f, z)| \leq 1$, $f \in \mathcal{F}$, $z \in Z$, we have

$$P^n\left(T \in Z^n : |\mathcal{R}_{L_\gamma,P}(u + A_T^u) - \mathcal{R}_{L_\gamma,T}(u + A_T^u)| > \frac{\epsilon}{2} + 2\beta\gamma^{-1}\right) \leq e^{-\frac{n\epsilon^2}{2(4n\beta\gamma^{-1}+1)^2}} \tag{8}$$

for all $u \in \mathcal{U}$. Consequently Theorem 2.4 gives

$$P^n\big(T \in Z^n : |\mathcal{R}_{L_\gamma,P}(\hat{A}_T) - \mathcal{R}_{L_\gamma,T}(\hat{A}_T)| > \delta\big)$$

$$\leq N(\mathcal{U}, d, \epsilon/4) \sup_{u \in \mathcal{U}} P^n\left(T \in Z^n : |\mathcal{R}_{L_\gamma,P}(u + A_T^u) - \mathcal{R}_{L_\gamma,T}(u + A_T^u)| > \delta - \frac{\epsilon}{2}\right)$$

and if we set $\delta := \epsilon + 2\beta\gamma^{-1}$ and utilize (8) we obtain the result. □

We now prove a lemma controlling how the fixed-b SVM defined in (5) can vary with b.

Lemma 3.1 *Let Q be a probability measure on Z and H be a reproducing kernel Hilbert space of functions on X with kernel k such that $\sup_{x \in X} \sqrt{k(x, x)} < \infty$. Define the regularized risk at $f = (b, h)$ to be*

$$\mathcal{R}_{Q,\lambda}(f) := \lambda \|h\|_H^2 + \mathcal{R}_{L_{\text{hinge}},Q}(b + h) \tag{9}$$

and let

$$A_{Q,\lambda}^b := \arg\min_{h \in H} \mathcal{R}_{Q,\lambda}(b, h) \tag{10}$$

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

$$\left\| A_{Q,\lambda}^{b_1} - A_{Q,\lambda}^{b_2} \right\|_H^2 \leq \lambda^{-1}|b_1 - b_2|. \tag{11}$$

Proof Consider the more general situation where we have a function $F : H \to \mathbb{R}$ defined by $F(h) := \lambda\|h\|^2 + G(h)$ where $G : H \to \mathbb{R}$ is convex and finite. Then according to Barbu et al. (Barbu and Precupanu 1986) the subdifferentials add

$$\partial F(h) = 2\lambda h + \partial G(h).$$

Let h^* be a minimizer of F. It follows that $0 \in \partial F(h^*) = 2\lambda h^* + \partial G(h^*)$ so that $-2\lambda h^* \in \partial G(h^*)$. However, by the definition of subdifferential,

$$G(h) - G(h^*) \geq \langle -2\lambda h^*, h - h^* \rangle, \quad \forall h.$$

Therefore

$$F(h) - F(h^*) = \lambda\|h\|^2 - \lambda\|h^*\|^2 + G(h) - G(h^*) \geq \lambda\|h\|^2 - \lambda\|h^*\|^2 - \langle 2\lambda h^*, h - h^*\rangle$$

and consequently we obtain the inequality of Bousquet and Elisseeff (2002)

$$\lambda\|h - h^*\|^2 \leq F(h) - F(h^*) \tag{12}$$

which they derived under the additional assumption of differentiability of G. To apply (12) observe that since $\sup_{x\in X}\sqrt{k(x,x)} < \infty$ it follows that $\mathcal{R}_{L_{\text{hinge}},Q}(b + \cdot)$ is a finite convex function for all b. Therefore by setting $G(h) := \mathcal{R}_{L_{\text{hinge}},Q}(b_1 + h)$ and $F(h) := \mathcal{R}_{Q,\lambda}(b_1, h)$ we can apply (12) to obtain

$$\lambda\|A^{b_2}_{Q,\lambda} - A^{b_1}_{Q,\lambda}\|^2_H \leq \mathcal{R}_{Q,\lambda}(b_1, A^{b_2}_{Q,\lambda}) - \mathcal{R}_{Q,\lambda}(b_1, A^{b_1}_{Q,\lambda})$$

and by setting $G(h) := \mathcal{R}_{L_{\text{hinge}},Q}(b_2 + h)$ and $F(h) := \mathcal{R}_{Q,\lambda}(b_2, h)$ we obtain

$$\lambda\|A^{b_1}_{Q,\lambda} - A^{b_2}_{Q,\lambda}\|^2_H \leq \mathcal{R}_{Q,\lambda}(b_2, A^{b_1}_{Q,\lambda}) - \mathcal{R}_{Q,\lambda}(b_2, A^{b_2}_{Q,\lambda}).$$

Adding the two we obtain

$$\begin{aligned}
&2\lambda\|A^{b_1}_{Q,\lambda} - A^{b_2}_{Q,\lambda}\|^2_H \\
&\quad\leq \mathcal{R}_{Q,\lambda}(b_1, A^{b_2}_{Q,\lambda}) - \mathcal{R}_{Q,\lambda}(b_1, A^{b_1}_{Q,\lambda}) + \mathcal{R}_{Q,\lambda}(b_2, A^{b_1}_{Q,\lambda}) - \mathcal{R}_{Q,\lambda}(b_2, A^{b_2}_{Q,\lambda}) \\
&\quad= \mathcal{R}_{L_{\text{hinge}},Q}(b_1 + A^{b_2}_{Q,\lambda}) - \mathcal{R}_{L_{\text{hinge}},Q}(b_1 + A^{b_1}_{Q,\lambda}) + \mathcal{R}_{L_{\text{hinge}},Q}(b_2 + A^{b_1}_{Q,\lambda}) \\
&\qquad - \mathcal{R}_{L_{\text{hinge}},Q}(b_2 + A^{b_2}_{Q,\lambda}) \\
&\quad= \mathbb{E}_{z\sim Q}\big(L_{\text{hinge}}(b_1 + A^{b_2}_{Q,\lambda}, z) - L_{\text{hinge}}(b_1 + A^{b_1}_{Q,\lambda}, z) + L_{\text{hinge}}(b_2 + A^{b_1}_{Q,\lambda}, z) \\
&\qquad - L_{\text{hinge}}(b_2 + A^{b_2}_{Q,\lambda}, z)\big).
\end{aligned}$$

Since $\|L_{\text{hinge}}(b_1 + A^{b_2}_{Q,\lambda}, \cdot) - L_{\text{hinge}}(b_2 + A^{b_2}_{Q,\lambda}, \cdot)\|_\infty \leq |b_1 - b_2|$ and $\|L_{\text{hinge}}(b_2 + A^{b_1}_{Q,\lambda}, \cdot) - L_{\text{hinge}}(b_1 + A^{b_1}_{Q,\lambda}, \cdot)\|_\infty \leq |b_1 - b_2|$ the result follows. □

Proof of Theorem 2.8 To apply Theorem 2.6 we show that the assumptions imply that A_λ is Lipschitz continuous and that A^b_λ is classification stable for each value of b with the same stability. We then bound the covering numbers.

Let us first show that A_λ is Lipschitz continuous. To that end we prove the following lemma which we note is valid for more general measures than empirical measures.

Lemma 3.2 *Let H be a reproducing kernel Hilbert space of functions on X with kernel k such that $K := \sup_{x\in X}\sqrt{k(x,x)} < \infty$ and let $\mathcal{U} \subset \mathbb{R}$. Consider a graphical learning algorithm with respect to $\mathcal{U} \times H$ defined by $\hat{A}_{T,\lambda} := \{(b, A^b_{T,\lambda}) : b \in U_T\}$ such that $A^b_{T,\lambda}$ satisfies the fixed-b SVM (4) and $U : Z^n \to 2^{\mathcal{U}}$ is any set valued map. Then A_λ is Lipschitz continuous with respect to L_γ for the pseudometric*

$$d_{\gamma,\lambda}(b_1, b_2) := \frac{1}{\gamma}\big(\lambda^{-1/2}K\sqrt{|b_1 - b_2|} + |b_1 - b_2|\big).$$

Proof We first note that since $|a| + |b| \le \left(\sqrt{|a|} + \sqrt{|b|}\right)^2$ that $d_{\gamma,\lambda}$ is a pseudometric. Since L_γ is Lipschitz continuous in its first variable with constant γ^{-1}, for all values of the second variable $z = (x, y)$ we obtain that

$$\begin{aligned}\left|L_\gamma((b_1, A^{b_1}_{T,\lambda}), z) - L_\gamma((b_2, A^{b_2}_{T,\lambda}), z)\right| &\le \frac{1}{\gamma}\left|A^{b_1}_{T,\lambda}(x) + b_1 - A^{b_2}_{T,\lambda}(x) - b_2\right| \\ &\le \frac{1}{\gamma}\left(|b_1 - b_2| + |A^{b_1}_{T,\lambda}(x) - A^{b_2}_{T,\lambda}(x)|\right) \\ &\le \frac{1}{\gamma}\left(|b_1 - b_2| + K\|A^{b_1}_{T,\lambda} - A^{b_2}_{T,\lambda}\|_H\right).\end{aligned}$$

Application of Lemma 3.1 then finishes the proof. □

Bousquet and Elisseeff (2002) show that the soft margin solution without offset $A^0_{T,\lambda}$ defined in (4) has classification stability

$$\beta = \frac{K^2}{2\lambda n}. \tag{13}$$

The same proof technique can be used to show that $A^b_{T,\lambda}$ has the same classification stability (13) for any b.

We now bound the covering numbers $N(\mathcal{U}_\lambda, d_{\gamma,\lambda}, \epsilon/4)$. Because these covering numbers are small compared with the exponential decay of the probability bounds we bound them crudely.

Lemma 3.3 *Consider $\mathcal{U}_\lambda$ defined in (2) and $d_{\gamma,\lambda}$ defined in (3.2). Then for all $\epsilon > 0$ we have*

$$N(\mathcal{U}_\lambda, d_{\gamma,\lambda}, \epsilon) \le \frac{4(\lambda^{-1/2}K + \sqrt{2})^3}{\gamma^2\epsilon^2} + 1.$$

Proof Let $d_\lambda(b_1, b_2) := \lambda^{-1/2}K\sqrt{|b_1 - b_2|} + |b_1 - b_2|$ so that $d_{\gamma,\lambda} = \frac{1}{\gamma}d_\lambda$ and $N(\mathcal{U}_\lambda, d_{\gamma,\lambda}, \epsilon) = N(\mathcal{U}_\lambda, d_\lambda, \gamma\epsilon)$. Define $B := 1 + \lambda^{-1/2}K$. For $b_1, b_2 \in \mathcal{U}_\lambda = [-B, B]$ we have $|b_1 - b_2| \le 2B$ from which we conclude that $|b_1 - b_2| \le \sqrt{|b_1 - b_2|}\sqrt{2B}$ so that

$$d_\lambda(b_1, b_2) = \lambda^{-1/2}K\sqrt{|b_1 - b_2|} + |b_1 - b_2| \le \left(\lambda^{-1/2}K + \sqrt{2B}\right)\sqrt{|b_1 - b_2|}.$$

Consequently, if we let $\alpha := \lambda^{-1/2}K + \sqrt{2B}$ then $d_\lambda(b_1, b_2) \le \alpha\sqrt{|b_1 - b_2|}$ and we obtain

$$N(\mathcal{U}_\lambda, d_\lambda, \epsilon) \le N\left(\mathcal{U}_\lambda, \alpha\sqrt{|\cdot|}, \epsilon\right) = N\left(\mathcal{U}_\lambda, |\cdot|, \frac{\epsilon^2}{\alpha^2}\right) \le \frac{2B\alpha^2}{\epsilon^2} + 1.$$

Finally, the inequality $(a + b)^2 \le 2a^2 + 2b^2$ implies that

$$\begin{aligned}B\alpha^2 &= (\lambda^{-1/2}K + 1)\left(\lambda^{-1/2}K + \sqrt{2(\lambda^{-1/2}K + 1)}\right)^2 \\ &\le 2(\lambda^{-1/2}K + 1)\left(\lambda^{-1}K^2 + 2(\lambda^{-1/2}K + 1)\right) \\ &\le 2(\lambda^{-1/2}K + 1)(\lambda^{-1/2}K + \sqrt{2})^2 \\ &\le 2(\lambda^{-1/2}K + \sqrt{2})^3.\end{aligned}$$

□

We can now conclude the proof of Theorem 2.8. The classification stability bound (13) combined with the fact that Lemma 3.2 implies that A_λ is Lipschitz continuous with respect to L_γ for the pseudometric $d_{\gamma,\lambda}$ and the bound on the covering numbers of Lemma 3.3 applied to Theorem 2.6 finishes the proof. □

Proof of Remark 2.9 To prove the alternative form (7) observe that for $x, a, b > 0$ and $\delta(x) = (\frac{a}{x} + 1)e^{-bx}$ we can bound the inverse $x(\delta)$ as follows: $\delta = (\frac{a}{x} + 1)e^{-bx}$ implies that $x = \frac{1}{b}(\ln(\frac{a}{x} + 1) + \ln\frac{1}{\delta})$ which implies that $x \geq \frac{1}{b}\ln\frac{1}{\delta}$ so that $x \leq \frac{1}{b}(\ln(\frac{ab}{\ln\frac{1}{\delta}} + 1) + \ln\frac{1}{\delta})$. Therefore if $\delta \leq e^{-1}$ it follows that $x \leq \frac{1}{b}(\ln(ab + 1) + \ln\frac{1}{\delta})$. Setting $x = \epsilon^2$, $a = \frac{64(\lambda^{-1/2}K+\sqrt{2})^3}{\gamma^2}$, and $b = \frac{n}{2(\frac{2K^2}{\lambda\gamma}+1)^2}$ then gives the result. □

4 Discussion

Theorem 2.6 is general in that it shows how to express bounds on risk deviance for learning algorithms in terms of bounds on the risk deviance for the algorithm restricted to fixed values of a parameter space. This result is then applied to SVMs which include the offset parameter to produce bounds on risk deviance which are similar to those Bousquet and Elisseeff obtained without the offset parameter. However it is useful to note that this inductive risk deviance theorem requires no notion of stability. This technique is similar to structural risk minimization (Vapnik 1998) where a discrete collection of classes is to be optimized over. The main difference is that in structural risk minimization no continuity assumptions are necessary when moving between the classes since simple union bounds are used. On the other hand, we thank one of the referees for pointing out that this method could be used in structural risk minimization over a continuous family of classes if the simpler notion of continuity is replaced by $|L(u_1, a, z) - L(u_2, a, z)| \leq d(u_1, u_2)$. In Theorem 2.6 the Definition 2.3 needs to be a stronger assumption since the result is in terms of learning algorithms which are more general than those constructed from risk minimization.

References

Barbu, V., & Precupanu, Th. (1986). *Convexity and optimization in Banach spaces*. Dordrecht: Reidel.

Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In *Fifth annual ACM workshop on computational learning theory*. Pittsburgh, 27–29 July.

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

Howse, J., Hush, D., & Scovel, C. (2002). Linking learning strategies and performance for support vector machines. http://www.c3.lanl.gov/ml/pubs_select.shtml.

Kutin, S., & Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. In *Proc. of uncertainty in artificial intelligence*. Edmonton, August.

Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research 4*, 1071–1105.

Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, to appear.

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

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