Journal of Machine Learning Research 8 (2007) 775-790 Submitted 6/06; Published 4/07

Sparseness vs Estimating Conditional Probabilities: Some Asymptotic Results

Peter L. Bartlett BARTLETT@CS.BERKELEY.EDU
Division of Computer Science and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Ambuj Tewari AMBUJ@CS.BERKELEY.EDU
Division of Computer Science
University of California
Berkeley, CA 94720-1776, USA

Editor: Gábor Lugosi

Abstract

One of the nice properties of kernel classifiers such as SVMs is that they often produce sparse solutions. However, the decision functions of these classifiers cannot always be used to estimate the conditional probability of the class label. We investigate the relationship between these two properties and show that these are intimately related: sparseness does not occur when the conditional probabilities can be unambiguously estimated. We consider a family of convex loss functions and derive sharp asymptotic results for the fraction of data that becomes support vectors. This enables us to characterize the exact trade-off between sparseness and the ability to estimate conditional probabilities for these loss functions.

Keywords: kernel methods, support vector machines, sparseness, estimating conditional probabilities

1. Introduction

Consider the following familiar setting of a binary classification problem. A sequence $T = ((x_1,y_1), \ldots, (x_n,y_n))$ of i.i.d. pairs is drawn from a probability distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ is the set of labels (which we assume is $\{+1,-1\}$ for convenience). The goal is to use the training set T to predict the label of a new observation $x \in \mathcal{X}$. A common way to approach the problem is to use the training set to construct a decision function $f_T : \mathcal{X} \to \mathbb{R}$ and output $\operatorname{sign}(f_T(x))$ as the predicted label of x.

In this paper, we consider classifiers based on an optimization problem of the form:

$$f_{T,\lambda} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n}\sum_{i=1}^{n} \phi(y_i f(x_i)). \tag{1}$$

Here, H is a reproducing kernel Hilbert space (RKHS) of some kernel k, $\lambda > 0$ is a regularization parameter and $\phi : \mathbb{R} \to [0,\infty)$ is a convex loss function. Since optimization problems based on the non-convex function 0-1 loss $t \mapsto I_{(t \le 0)}$ (where $I_{(\cdot)}$ is the indicator function) are computationally intractable, use of convex loss functions is often seen as using upper bounds on the 0-1 loss to make the problem computationally easier. Although computational tractability is one of the goals we have in mind while designing classifiers, it is not the only one. We would like to compare different convex loss functions based on their statistical and other useful properties. Conditions ensuring Bayes-risk consistency of classifiers using convex loss functions have already been established (Bartlett et al., 2004; Lugosi and Vayatis, 2004; Steinwart, 2005; Zhang, 2004). It has been observed that different cost functions have different properties and it is important to choose a loss function judiciously (e.g., see Wahba, 2002). In order to understand the relative merits of different loss functions, it is important to consider these properties and investigate the extent to which different loss functions exhibit them. It may turn out (as it does below) that different properties are in conflict with each other. In that case, knowing the trade-off allows one to make an informed choice while choosing a loss function for the classification task at hand.

One of the properties we focus on is the ability to estimate the conditional probability of the class label $\eta(x) = P(Y = +1|X = x)$. Under some conditions on the loss function and the sequence of regularization parameters λ_n, the solutions of (1) converge (in probability) to a function $F^*_\phi(\eta(x))$ which is set valued in general (Steinwart, 2003). As long as we can uniquely identify $\eta(x)$ based on a value in $F^*_\phi(\eta(x))$, we can hope to estimate conditional probabilities using $f_{T,\lambda_n}(x)$, at least asymptotically. Choice of the loss function is crucial to this property. For example, the L2-SVM (which uses the loss function $t \mapsto (\max\{0, 1-t\})^2$) is much better than L1-SVM (which uses $t \mapsto \max\{0, 1-t\}$) in terms of asymptotically estimating conditional probabilities.

Another criterion is the sparseness of solutions of (1). It is well known that any solution $f_{T,\lambda}$ of (1) can be represented as

$$f_{T,\lambda}(x) = \sum_{i=1}^{n} \alpha_i^* k(x, x_i) \; . \tag{2}$$

The observations x_i for which the coefficients α_i^* are non-zero are called support vectors. The rest of the observations have no effect on the value of the decision function. Having fewer support vectors leads to faster evaluation of the decision function. Bounds on the number of support vectors are therefore useful to know. Steinwart's recent work (Steinwart, 2004) has shown that for the L1-SVM and a suitable kernel, the asymptotic fraction of support vectors is twice the Bayes-risk. Thus, L1-SVMs can be expected to produce sparse solutions. It was also shown that L2-SVMs will typically not produce sparse solutions.

We are interested in how sparseness relates to the ability to estimate conditional probabilities. What we mentioned about L1 and L2-SVMs leads to several questions. Do we always lose sparseness by being able to estimate conditional probabilities? Is it possible to characterize the exact trade-off between the asymptotic fraction of support vectors and the ability to estimate conditional probabilities? If sparseness is indeed lost when we are able to fully estimate conditional probabilities, we may want to estimate conditional probabilities only in an interval, say $(0.05, 0.95)$, if that helps recover sparseness. Estimating η for x's that have $\eta(x) \geq 0.95$ may not be too crucial for our prediction task. How can we design loss functions which enable us to estimate probabilities in sub-intervals of $[0, 1]$ while preserving as much sparseness as possible?

This paper attempts to answer these questions. We show that if one wants to estimate conditional probabilities in an interval $(1-\gamma_0, \gamma_0)$ for some $\gamma_0 \in (1/2, 1)$, then sparseness is lost on that interval in the sense that the asymptotic fraction of data that become support vectors is lower bounded by $\mathbb{E}_x G(\eta(x))$ where $G(\eta) = 1$ throughout the interval $(1-\gamma_0, \gamma_0)$. Moreover, one cannot recover sparseness by giving up the ability to estimate conditional probabilities in some sub-interval of $(1-\gamma_0, \gamma_0)$. The only way to do that is to decrease γ_0 thereby shortening the interval $(1-\gamma_0, \gamma_0)$. We also derive sharp bounds on the asymptotic number of support vectors for a family of loss functions

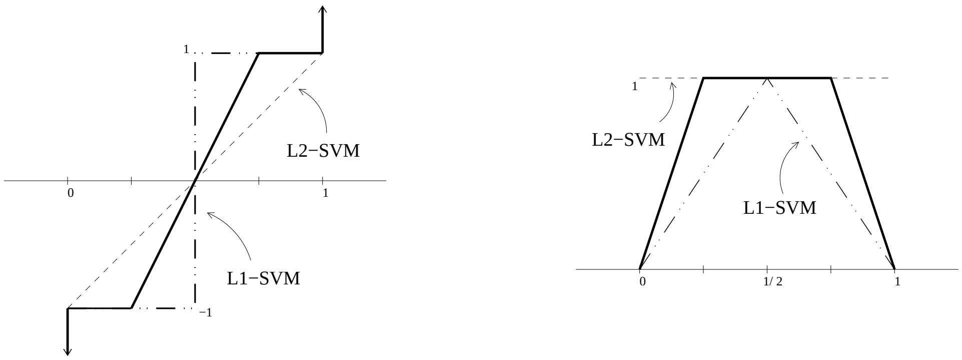

Figure 1: Plots of F_ϕ^* (left) and G (right) for a loss function which is a convex combination of the L1 and L2-SVM loss functions. Dashed lines represent the corresponding plots for the original loss functions.

of the form:

$$\phi(t) = h((t_0 - t)_+),\ t_0 > 0$$

where t_+ denotes $\max\{0,t\}$ and h is a continuously differentiable convex function such that $h'(0) \geq 0$. Each loss function in the family allows one to estimate probabilities in the interval $(1-\gamma_0, \gamma_0)$ for some value of γ_0. The asymptotic fraction of support vectors is then $\mathbb{E}_x G(\eta(x))$, where $\eta \mapsto G(\eta)$ is a function that increases linearly from 0 to 1 as η goes from 0 to $1-\gamma_0$. For example, if $\phi(t) = \frac{1}{3}((1-t)_+)^2 + \frac{2}{3}(1-t)_+$ then conditional probabilities can be estimated in $(1/4, 3/4)$ and $G(\eta) = 1$ for $\eta \in (1/4, 3/4)$ (see Fig. 1).

2. Notation and Known Results

Let P be the probability distribution over $\mathcal{X} \times \mathcal{Y}$ and let $T \in (\mathcal{X} \times \mathcal{Y})^n$ be a training set. Let $\mathbb{E}_P(\cdot)$ denote expectations taken with respect to the distribution P. Similarly, let $\mathbb{E}_x(\cdot)$ denote expectations taken with respect to the marginal distribution on $\mathcal{X}$. Let $\eta(x)$ be $P(Y = +1|X = x)$. For a decision function $f : \mathcal{X} \to \mathbb{R}$, define its risk as

$$R_P(f) = \mathbb{E}_P I_{(yf(x)\leq 0)}\ .$$

The Bayes-risk $R_P = \inf\{R_P(f) : f \text{ measurable}\}$ is the least possible risk. Given a loss function ϕ, define the ϕ-risk of f by

$$R_{\phi,P}(f) = \mathbb{E}_P \phi(yf(x))\ .$$

The optimal ϕ-risk $R_{\phi,P} = \inf\{R_{\phi,P}(f) : f \text{ measurable}\}$ is the least achievable ϕ-risk. When the expectations in the definitions of $R_P(f)$ and $R_{\phi,P}(f)$ are taken with respect to the empirical measure corresponding to T, we get the empirical risk $R_T(f)$ and the empirical ϕ-risk $R_{\phi,T}(f)$ respectively. Conditioning on x, we can write the ϕ-risk as

$$\begin{aligned} R_{\phi,P}(f) &= E_x[E(\phi(yf(x)|x)] \\ &= E_x[\eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x))] \\ &= E_x[C(\eta(x), f(x))]\ . \end{aligned}$$

Here, we have defined $C(\eta,t) = \eta\phi(t) + (1-\eta)\phi(-t)$. To minimize the ϕ-risk, we have to minimize $C(\eta,\cdot)$ for each $\eta \in [0,1]$. So, define the set valued function F^*_ϕ by

$$F^*_\phi(\eta) = \{t : C(\eta,t) = \min_{s\in\bar{\mathbb{R}}} C(\eta,s)\}$$

where $\bar{\mathbb{R}}$ is the set of extended reals $\mathbb{R} \cup \{-\infty,\infty\}$. Any measurable selection f^* of F^*_ϕ actually minimizes the ϕ-risk. The function F^*_ϕ is plotted for three choices of ϕ in Fig. 1. From the definitions of $C(\eta,t)$ and $F^*_\phi(\eta)$, it is easy to see that $F^*_\phi(\eta) = -F^*_\phi(1-\eta)$. Steinwart (2003) also proves that $\eta \mapsto F^*_\phi(\eta)$ is a monotone operator. This means that if $\eta_1 > \eta_2$, $t_1 \in F^*_\phi(\eta_1)$ and $t_2 \in F^*_\phi(\eta_2)$ then $t_1 \geq t_2$.

A convex loss function is called classification calibrated if the following two conditions hold:

$$\eta < \frac{1}{2} \Rightarrow F^*_\phi(\eta) \subset [-\infty,0) \text{ and } \eta > \frac{1}{2} \Rightarrow F^*_\phi(\eta) \subset (0,+\infty] \ .$$

A necessary and sufficient condition for a convex ϕ to be classification calibrated is that $\phi'(0)$ exists and is negative (Bartlett et al., 2004). If ϕ is classification calibrated then it is guaranteed that for any sequence f_n such that $R_{\phi,P}(f_n) \to R_{\phi,P}$, we have $R_P(f_n) \to R_P$. Thus, classification calibrated loss functions are good in the sense that minimizing the ϕ-risk leads to classifiers that have risks approaching the Bayes-risk. Note, however, that in the optimization problem (1), we are minimizing the regularized ϕ-risk

$$R^{reg}_{\phi,T,\lambda} = \lambda\|f\|^2_H + R_{\phi,T} \ .$$

Steinwart (2005) has shown that if one uses a classification calibrated convex loss function, a universal kernel (one whose RKHS is dense in the space of continuous functions over $\mathcal{X}$) and a sequence of regularization parameters such that $\lambda_n \to 0$ sufficiently slowly, then $R_{\phi,P}(f_{T,\lambda_n}) \to R_{\phi,P}$. In another paper (Steinwart, 2003), he proves that this is sufficient to ensure the convergence in probability of f_{T,λ_n} to $F^*_\phi(\eta(\cdot))$. That is, for all $\varepsilon > 0$

$$P_x(\{x \in \mathcal{X} : \rho(f_{T,\lambda_n}(x), F^*_\phi(\eta(x))) \geq \varepsilon\}) \to 0. \tag{3}$$

The function $\rho(t,B)$ is just the distance from t to the point in B which is closest to t. The definition given by Steinwart (2003) is more complicated because one has to handle the case when $B \cap \mathbb{R} = \emptyset$. We will ensure in our proofs that F^*_ϕ is not a singleton set just containing $+\infty$ or $-\infty$.

Since f_{T,λ_n} converges to $F^*_\phi(\eta(\cdot))$, the plots in Fig. 1 suggest that the L2-SVM decision function can be used to estimate conditional probabilities in the whole range $[0,1]$ while it not possible to use the L1-SVM decision function to estimate conditional probabilities in any interval. However, the L1-SVM is better if one considers the asymptotic fraction of support vectors. Under some conditions on the kernel and the regularization sequence, Steinwart proved that the fraction is $\mathbb{E}_x[2\min(\eta(x), 1-\eta(x))]$, which also happens to be the optimal ϕ-risk for the hinge loss function. For L2-SVM, he showed that the asymptotic fraction is $P_x(\{x \in \mathcal{X} : 0 < \eta(x) < 1\})$, which is the probability of the set where noise occurs. Observe that we can write the fraction of support vectors as $\mathbb{E}_x[G(\eta(x))]$ where $G(\eta) = 2\min\{\eta, 1-\eta\}$ for the hinge loss and $G(\eta) = I_{(\eta \notin \{0,1\})}$ for the squared hinge loss. We will see below that these two are extreme cases. In general, there are loss functions which allow one to estimate probabilities in an interval centered at 1/2 and for which $G(\eta) = 1$ only on that interval.

Steinwart (2003) also derived a general lower bound on the asymptotic number of support vectors in terms of the probability of the set

$$S = \{(x,y) \in \mathcal{X}_{cont} \times \mathcal{Y} : 0 \notin \partial\phi(yF_\phi^*(\eta(x)))\} \ .$$

Here, $\mathcal{X}_{cont} = \{x \in \mathcal{X} : P_x(\{x\}) = 0\}$ and $\partial\phi$ denotes the subdifferential of ϕ. In the simple case of a function of one variable $\partial\phi(x) = [\phi'_-(x), \phi'_+(x)]$, where ϕ'_- and ϕ'_+ are the left and right hand derivatives of ϕ (which always exist for convex functions). If $\mathcal{X}_{cont} = \mathcal{X}$, one can write $P(S)$ as

$$\begin{aligned} P(S) &= \mathbb{E}_P[I_{(0 \notin \partial\phi(yF_\phi^*(\eta(x))))}] \\ &= \mathbb{E}_x[\eta(x) I_{(0\notin\partial\phi(F_\phi^*(\eta(x))))} + (1-\eta(x)) I_{(0\notin\partial\phi(-F_\phi^*(\eta(x))))}] \\ &= \mathbb{E}_x G(\eta(x)) \ . \end{aligned} \tag{4}$$

For the last step, we simply defined

$$G(\eta) = \eta I_{(0\notin\partial\phi(F_\phi^*(\eta)))} + (1-\eta) I_{(0\notin\partial\phi(-F_\phi^*(\eta)))} \ . \tag{5}$$

3. Preliminary Results

We will consider only classification calibrated convex loss functions. Since ϕ is classification calibrated we know that $\phi'(0) < 0$. Define t_0 as

$$t_0 = \inf\{t : 0 \in \partial\phi(t)\}$$

with the convention that $\inf \emptyset = \infty$. Because $\phi'(0) < 0$ and subdifferentials of a convex function are monotonically decreasing, we must have $t_0 > 0$. However, it may be that $t_0 = \infty$. The following lemma says that sparse solutions cannot be expected if that is the case.

Lemma 1 *If $t_0 = \infty$, then $G(\eta) = 1$ for $\eta \in [0,1]$.*

Proof $t_0 = \infty$ implies that for all t, $0 \notin \partial\phi(t)$. Using (5), we get $G(\eta) = \eta.1 + (1-\eta).1 = 1$. ■

Thus, for losses like the exponential loss $t \mapsto \exp(-t)$, the bound (4) says that the fraction of support vectors approaches 1 asymptotically. Since we are interested in loss functions that lead to sparse solutions, let us assume that

$$\text{(A1)} \qquad t_0 < \infty \ .$$

Although not immediate from the definition, a continuity argument involving the one-sided derivatives gives us $0 \in \partial\phi(t_0)$.

We now proceed to investigate the general form of the function $\eta \mapsto F_\phi^*(\eta)$. Before we begin, we make another assumption about the number of points in the interval $(-t_0, t_0)$ where the function ϕ fails to be differentiable. Define the set

$$D = \{t \in (0, t_0) : \text{ either } \phi'(t) \text{ or } \phi'(-t) \text{ does not exist } \} \ .$$

We assume that

$$\text{(A2)} \qquad |D| < \infty \ .$$

Note that, since ϕ is convex, we know that D can be at most countably infinite. Let $t_1 > t_2 > \ldots > t_L$ be the set D sorted in decreasing order. We have $t_0 > t_1$ and $t_L > 0$. Set $\beta_0 = 1$ and define

$$\beta_l = \frac{1}{1 + \frac{\phi'_+(t_l)}{\phi'_-(-t_l)}} \ , \ \ 1 \le l \le L \ ,$$

$$\gamma_l = \frac{1}{1 + \frac{\phi'_-(t_l)}{\phi'_+(-t_l)}} \ , \ \ 0 \le l \le L \ .$$

With the possible exception of $\phi'_-(t_0)$ (which can be zero), all one-sided derivatives appearing in the definitions above are negative. For $s < t$, we have $\phi'_-(s) \le \phi'_+(s) \le \phi'_-(t) \le \phi'_+(t)$. Using this fact, we get

$$\frac{1}{2} \le \gamma_L \le \beta_L \le \ldots \le \gamma_1 \le \beta_1 \le \gamma_0 \le \beta_0 = 1 \ .$$

Since $t_l \in D$ for $1 \le l \le L$, at least one of the inequalities, $\phi'_-(t_l) < \phi'_+(t_l)$ or $\phi'_-(-t_l) < \phi'_+(-t_l)$, is strict. So $\gamma_l < \beta_l$ for $1 \le l \le L$ and we get

$$\frac{1}{2} \le \gamma_L < \beta_L \le \ldots \le \gamma_1 < \beta_1 \le \gamma_0 \le \beta_0 = 1 \ .$$

Using the observation that $F^*_\phi(1-\eta) = -F^*_\phi(\eta)$, we restrict ourselves to examining the behavior of $F^*_\phi(\eta)$ on an interval containing $[1/2, 1]$.

Theorem 2 *Let ϕ be a classification calibrated convex loss function satisfying assumptions (A1) and (A2). Then the following statements hold true.*

1. *For $0 \le l \le L$ and for $\eta \in (\gamma_l, \beta_l)$, $F^*_\phi(\eta) = \{t_l\}$.*
2. *If I is one of the (possibly degenerate) intervals $[\beta_l, \gamma_{l-1}]$, $1 \le l \le L$, then there exists a continuous non-decreasing function g_I mapping $[t_l, t_{l-1}]$ onto I such that, for $\eta \in I$,*

   $$F^*_\phi(\eta) = g_I^{-1}(\eta) \ ,$$

   where $g_I^{-1}(\eta) = \{t : g_I(t) = \eta\}$.
3. *The above also holds for the (possibly degenerate) interval $I = [1-\gamma_L, \gamma_L]$ but here the function g_I maps $[-t_L, t_L]$ onto I.*
4. *$F^*_\phi(1)$ is either $[t_0, t']$, for some $t' \ge t_0$, or $[t_0, \infty)$.*

Proof (Part 1) Denoting the subdifferential with respect to the second argument by ∂_2, we have

$$\begin{aligned} \partial_2 C(\eta, t_l) &= \eta \partial\phi(t_l) - (1-\eta)\partial\phi(-t_l) \\ &= [\eta\phi'_-(t_l) - (1-\eta)\phi'_+(-t_l), \eta\phi'_+(t_l) - (1-\eta)\phi'_-(-t_l)] \ . \end{aligned}$$

For $\gamma_l < \eta < \beta_l$, we have

$$\eta\phi'_-(t_l) - (1-\eta)\phi'_+(-t_l) < 0 < \eta\phi'_+(t_l) - (1-\eta)\phi'_-(-t_l) \ ,$$

and so t_l is the unique minimizer of $C(\eta,\cdot)$.

(Part 2) Let $I = [\beta_l, \gamma_{l-1}]$. Note that for $t \in (t_l, t_{l-1})$, both $\phi'(t)$ and $\phi'(-t)$ are defined. Being derivatives of a convex function, they are continuous. So we define, for $t \in (t_l, t_{l-1})$,

$$g_I(t) = \frac{1}{1+\frac{\phi'(t)}{\phi'(-t)}} \ .$$

Convexity also implies that $g_I(t)$ is monotonically increasing. In fact, it will be strictly increasing if one of $\phi(t)$ and $\phi(-t)$ is strictly convex. However, if both $\phi(t)$ and $\phi(-t)$ are linear on some open interval, $g_I(t)$ will be constant over that interval. Define $g_I(t_l) = \beta_l$ and $g_I(t_{l-1}) = \gamma_{l-1}$. Since $\lim_{t\downarrow t_l} \phi'(t) = \phi'_+(t_l)$ and $\lim_{t\downarrow t_l} \phi'(-t) = \phi'_-(-t_l)$ (e.g., see Rockafellar, 1970, Chap. 24), we have $\lim_{t\downarrow t_l} g_I(t) = \beta_l = g_I(t_l)$. Similarly, $\lim_{t\uparrow t_{l-1}} g_I(t) = \gamma_{l-1} = g_I(t_{l-1})$. So, we have defined a continuous monotonically increasing function on $[\beta_l, \gamma_{l-1}]$ onto $[t_l, t_{l-1}]$. If $t \in (t_l, t_{l-1})$ then, for any η,

$$\begin{aligned} t \in F^*_\phi(\eta) &\text{ iff } \eta\phi'(t) + (1-\eta)\phi'(-t) = 0 \\ &\text{iff } \frac{1}{1+\frac{\phi'(t)}{\phi'(-t)}} = \eta \\ &\text{iff } g_I(t) = \eta \\ &\text{iff } t \in g_I^{-1}(\eta) \ . \end{aligned}$$

It remains to handle t_l and t_{l-1}. For $\eta \in I$, $t_l \in F^*_\phi(\eta)$ iff $\eta = \beta_l$. Since $g_I(t_l) = \beta_l$, we also have $t_l \in g_I^{-1}(\eta)$ iff $\eta = \beta_l$. Arguing similarly for t_{l-1}, for $\eta \in I$, $t_l \in F^*_\phi(\eta)$ iff $\eta = \gamma_{l-1}$. Since $g_I(t_{l-1}) = \gamma_{l-1}$, we also have $t_{l-1} \in g_I^{-1}(\eta)$ iff $\eta = \gamma_{l-1}$.

(Part 3) Once we observe that both $\phi'(t)$ and $\phi'(-t)$ exist for $t \in (-t_L, t_L)$, the proof proceeds as in part 2.

(Part 4) Let $t' = \sup\{t : 0 \in \partial\phi(t)\}$. If the set in question is unbounded, then $0 \in \partial\phi(t)$ for all $t \geq t_0$ and so $F^*_\phi(1) = [t, \infty)$. If the set is bounded above then the supremum t' is well defined. The t's which minimize $C(1,\cdot) = \phi(t)$ are then precisely those in the interval $[t_0, t']$. ■

The above theorem says a couple of interesting things about the relationship between ϕ and F^*_ϕ. Points of non-differentiability of ϕ lead to the function F^*_ϕ being constant on certain intervals. On an interval I where F^*_ϕ is not constant, there exists a function g_I such that given $t \in F^*_\phi(\eta)$, one can recover η by the relation $\eta = g_I(t)$. We can express this by saying that $F^*_\phi(\eta)$ is invertible on intervals that correspond to differentiable portions of $\phi(t)$ and $\phi(-t)$ (see Fig. 2 for an example).

Theorem 3 *Let ϕ be an classification calibrated convex loss function satisfying assumptions (A1) and (A2). Then, for $G(\eta)$ as defined in* (5), *we have*

$$G(\eta) = \begin{cases} 1 & \eta \in (1-\gamma_0, \gamma_0)\ , \\ \min\{\eta, 1-\eta\} & \eta \in [0, 1-\gamma_0] \cup [\gamma_0, 1]\ , \end{cases} \tag{6}$$

where $\gamma_0 = \phi'_+(-t_0)/(\phi'_+(-t_0) + \phi'_-(t_0))$.

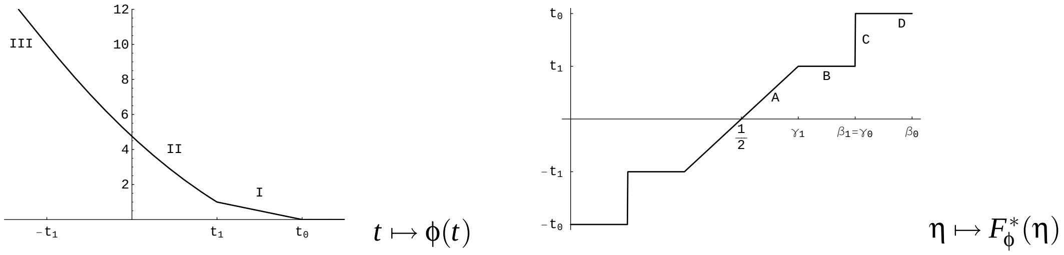

Figure 2: The loss function (left) is composed of two linear parts (I & III) and a quadratic part (II). The function F^*_ϕ (right) is constant on the regions marked B and D correponding to the 2 points of non-differentiability, viz. t_0, t_1. The vertical part C is due to ϕ being linear on the intervals $[-t_0, -t_1]$ and $[t_1, t_0]$. Region A arises due to the quadratic part of the loss function.

Proof If $\eta \in [\gamma_0, 1]$, Theorem 2 tells us that $t_0 \in F^*_\phi(\eta)$ and hence $0 \in \partial\phi(F^*_\phi(\eta))$. If $\eta < 1-\gamma_0$, monotonicity of $F^*_\phi(\eta)$ implies that $F^*_\phi(\eta) \subseteq (-\infty, t_0)$. Since $t_0 = \inf\{t : 0 \in \partial\phi(t)\}$, $0 \notin \partial\phi(F^*_\phi(\eta))$ for $\eta \in [0, \gamma_0)$. Thus, we can write $I_{(0\notin\partial\phi(F^*_\phi(\eta)))}$ as $I_{(\eta\notin[\gamma_0,1])}$. Also $I_{(0\notin\partial\phi(-F^*_\phi(\eta))} = I_{(0\notin\partial\phi(F^*_\phi(1-\eta))}$. Plugging this in (5), we get

$$\begin{aligned} G(\eta) &= \eta I_{(\eta\notin[\gamma_0,1])} + (1-\eta) I_{(1-\eta\notin[\gamma_0,1])} \\ &= \eta I_{(\eta\notin[\gamma_0,1])} + (1-\eta) I_{(\eta\notin[0,1-\gamma_0])} \ . \end{aligned}$$

Since $\gamma_0 \geq 1/2$, we can write $G(\eta)$ in the form given above. ■

Corollary 4 *If* $\eta_1 \in (0,1)$ *is such that* $F^*_\phi(\eta_1) \cap F^*_\phi(\eta) = \emptyset$ *for* $\eta \neq \eta_1$, *then* $G(\eta) = 1$ *on* $[\min\{\eta_1, 1-\eta_1\}, \max\{\eta_1, 1-\eta_1\}]$.

Proof Theorem 1, part 1 tells us that such an η_1 cannot lie in $[0, 1-\gamma_0] \cup [\gamma_0, 1]$. Rest follows from Theorem 3. ■

The preceding theorem and corollary have important implications. First, we can hope to have sparseness only for values of $\eta \in [0, 1-\gamma_0] \cup [\gamma_0, 1]$. Second, we cannot estimate conditional probabilities in these two intervals because $F^*_\phi(\cdot)$ is not invertible there. Third, any loss function for which $F^*_\phi(\cdot)$ is invertible, say at $\eta_1 < 1/2$, will necessarily not have sparseness on the interval $[\eta_1, 1-\eta_1]$.

Note that for the case of L1 and L2-SVM, γ_0 is $1/2$ and 1 respectively. For these two classifiers, the lower bounds $\mathbb{E}_x G(\eta(x))$ obtained after plugging in γ_0 in (6) are the ones proved initially (Steinwart, 2003). For the L1-SVM, the bound was later significantly improved (Steinwart, 2004). This suggests that $\mathbb{E}_x G(\eta(x))$ might be a loose lower bound in general. In the next section we will show, by deriving a sharp asymptotic result, that the bound is indeed loose for a family of loss functions.

4. Asymptotic Fraction of Support Vectors

We will consider convex loss functions of the form

$$\phi(t) = h((t_0 - t)_+) \ . \tag{7}$$

The function h is assumed to be continuously differentiable and convex. We also assume $h'(0) > 0$. The convexity of ϕ requires that $h'(0)$ be non-negative. The reason we rule out $h'(0) = 0$ is that $\gamma_0 = 1$ in that case and Theorem 3 already tells us we will not have any sparseness. So, in this section we are not interested in loss functions that are differentiable everywhere. In other words, our assumption is that the loss function is constant for all $t \geq t_0$ and is continuously differentiable to the left of t_0. Further, the only discontinuity in the derivative is at t_0. Without loss of generality, we may assume that $h(0) = 0$ because the solutions to (1) do not change if we add or subtract a constant from ϕ. Note that we obtain the hinge loss if we set $h(t) = t$. We now derive the dual of (1) for our choice of the loss function.

4.1 Dual Formulation

For a convex loss function $\phi(t) = h((t_0 - t)_+)$, consider the optimization problem:

$$\arg\min_w \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \phi(y_i w^T x_i) \ .$$

Make the substitution $\xi_i = t_0 - y_i w^T x_i$ to get

$$\begin{gathered} \arg\min_w \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \phi(t_0 - \xi_i) \\ \text{subject to } \xi_i = t_0 - y_i w^T x_i \text{ for all } i \ . \end{gathered} \tag{8}$$

Introducing Lagrange multipliers, we get the Lagrangian:

$$\mathscr{L}(w,\xi,\alpha) = \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \phi(t_0 - \xi_i) + \sum_{i=1}^n \alpha_i(t_0 - y_i w^T x_i - \xi_i) \ .$$

Minimizing this with respect to the primal variables w and ξ_i's, gives us

$$\begin{aligned} w &= \frac{1}{2\lambda}\sum_{i=1}^n \alpha_i y_i x_i \ , \\ \alpha_i &\in -\partial\phi(t_0 - \xi_i)/n \ . \end{aligned}$$

For the specific form of ϕ that we are working with, we have

$$-\partial\phi(t_0 - \xi_i)/n = \begin{cases} \{h'(\xi_i)/n\} & \xi_i > 0 \ , \\ [0, h'(0)/n] & \xi_i = 0 \ , \\ \{0\} & \xi_i < 0 \ . \end{cases} \tag{9}$$

Let (w^*, ξ_i^*) be a solution of (8). Then we have

$$\begin{aligned} \lambda\|w^*\|^2 &= \lambda(w^*)^T\left(\frac{1}{2\lambda}\sum_{i=1}^n \alpha_i^* y_i x_i\right) \\ &= \frac{1}{2}\sum_{i=1}^n \alpha_i^* y_i (w^*)^T x_i = \frac{1}{2}\sum_{i=1}^n \alpha_i^*(t_0 - \xi_i^*) \ . \end{aligned} \tag{10}$$

4.2 Result About Asymptotic Fraction of Support Vectors and Its Proof

Recall that a kernel is called universal if its RKHS is dense in the space of continuous functions over $\mathcal{X}$. Suppose the kernel k is universal and analytic. This ensures that any function in the RKHS H of k is analytic. Following Steinwart (2004), we call a probability distribution P non-trivial (with respect to ϕ) if

$$R_{\phi,P} < \inf_{b\in\mathbb{R}} R_{\phi,P}(b) \ .$$

We also define the P-version of the optimization problem (1):

$$f_{P,\lambda} = \arg\min_{f\in H} \lambda \|f\|_H^2 + E_P \phi(yf(x)) \ .$$

Further, suppose that $K = \sup\{\sqrt{k(x,x)} : x \in \mathcal{X}\}$ is finite. Fix a loss function of the form (7). Let G be the function defined as

$$G(\eta) = \begin{cases} \eta/(1-\gamma_0) & 0 \le \eta \le 1-\gamma_0 \ , \\ 1 & 1-\gamma_0 < \eta < \gamma_0 \ , \\ (1-\eta)/(1-\gamma_0) & \gamma_0 \le \eta \le 1 \ , \end{cases}$$

where $\gamma_0 = h'(2t_0)/(h'(0)+h'(2t_0))$. Note that this definition is different from the one given in (5).

Since ϕ is differentiable on $(-t_0,t_0)$, Theorem 2, part 2 implies that F_ϕ^* is invertible on $(1-\gamma_0,\gamma_0)$. Thus, one can estimate conditional probabilities in the interval $(1-\gamma_0,\gamma_0)$. Let $\#SV(f_{T,\lambda})$ denote the number of support vectors in the solution (2):

$$\#SV(f_{T,\lambda}) = |\{i : \alpha_i^* \ne 0\}| \ .$$

The next theorem says that the fraction of support vectors converges to the expectation $\mathbb{E}_x G(\eta(x))$ in probability.

Theorem 5 *Let H be the RKHS of an analytic and universal kernel on $\mathbb{R}^d$. Further, let $\mathcal{X} \subset \mathbb{R}^d$ be a closed ball and P be a probability measure on $\mathcal{X} \times \{\pm 1\}$ such that P_x has a density with respect to the Lebesgue measure on X and P is non-trivial. Suppose* $\sup\{\sqrt{k(x,x)} : x \in \mathcal{X}\} < \infty$*. Then for a classifier based on* (1)*, which uses a loss function of the form* (7)*, and a regularization sequence which tends to* 0 *sufficiently slowly, we have*

$$\frac{\#SV(f_{T,\lambda_n})}{n} \to \mathbb{E}_x G(\eta(x))$$

in probability.

Proof Let us fix an $\varepsilon > 0$. The proof will proceed in four steps of which the last two simply involve relating empirical averages to expectations.

Step 1. In this step we show that $f_{P,\lambda_n}(x)$ is not too close to $\pm t_0$ for most values of x. We also ensure that $f_{T,\lambda_n}(x)$ is sufficiently close to $f_{P,\lambda_n}(x)$ provided $\lambda_n \to 0$ slowly. Since $f_{P,\lambda}$ is an analytic function, for any constant c, we have

$$P_x(\{x \in X : f_{P,\lambda}(x) = c\}) > 0 \Rightarrow f(x) = c \ P_x\text{-a.s.} \tag{11}$$

Assume that $P_x(\{x \in \mathcal{X} : f_{P,\lambda}(x) = t_0\}) > 0$. By (11), we get $P_x(\{x \in \mathcal{X} : f_{P,\lambda}(x) = t_0\}) = 1$. But for small enough λ, $f_{P,\lambda} \neq t_0$ since $R_{\phi,P}(f_{P,\lambda}) \to R_{\phi,P}$ and $R_{\phi,P}(t_0) \neq R_{\phi,P}$ by the non-triviality of P. Therefore, assume that for all sufficiently large n, we have

$$P_x(\{x \in \mathcal{X} : f_{P,\lambda_n}(x) = t_0\}) = 0 \ .$$

Repeating the reasoning for $-t_0$ gives us

$$P_x(\{x \in \mathcal{X} : \left|f_{P,\lambda_n}(x) - t_0\right| \leq \delta\}) \downarrow 0 \text{ as } \delta \downarrow 0 \ ,$$

$$P_x(\{x \in \mathcal{X} : \left|f_{P,\lambda_n}(x) + t_0\right| \leq \delta\}) \downarrow 0 \text{ as } \delta \downarrow 0 \ .$$

Define the set $A_\delta(\lambda) = \{x \in \mathcal{X} : \left|f_{P,\lambda}(x) - t_0\right| \leq \delta \text{ or } \left|f_{P,\lambda}(x) + t_0\right| \leq \delta\}$. For small enough λ and for all $\varepsilon' > 0$, there exists $\delta > 0$ such that $P_x(A_\delta(\lambda)) \leq \varepsilon'$. Therefore, we can define

$$\delta(\lambda) = \frac{1}{2}\sup\{\delta > 0 : P_x(A_\delta(\lambda)) \leq \varepsilon\} \ .$$

Let $m(\lambda) = \inf\{\delta(\lambda') : \lambda' \geq \lambda\}$ be a decreasing version of $\delta(\lambda)$. Using Proposition 33 from Steinwart (2003) with $\varepsilon = m(\lambda_n)/K$, we conclude that for a sequence $\lambda_n \to 0$ sufficiently slowly, the probability of a training set T such that

$$\| f_{T,\lambda_n} - f_{P,\lambda_n} \| < m(\lambda_n)/K \tag{12}$$

converges to 1 as $n \to \infty$. It is important to note that we can draw this conclusion because $m(\lambda) > 0$ for $\lambda > 0$ (See proof of Theorem 3.5 in Steinwart, 2004). We now relate the ∞-norm of an f to its 2-norm.

$$\begin{aligned} f(x) &= \langle k(x,\cdot), f(\cdot)\rangle \leq \|k(x,\cdot)\| \, \|f\| \\ &= \sqrt{\langle k(x,\cdot), k(x,\cdot)\rangle}\|f\| \\ &= \sqrt{k(x,x)}\|f\| \leq K\|f\| \ . \end{aligned} \tag{13}$$

Thus, (12) gives us

$$\| f_{T,\lambda_n} - f_{P,\lambda_n} \|_\infty < m(\lambda_n) \ . \tag{14}$$

Step 2. In the second step, we relate the fraction of support vectors to an empirical average. Suppose that, in addition to (14), our training set T satisfies

$$\lambda_n \| f_{T,\lambda_n} \|^2 + R_{\phi,P}(f_{T,\lambda_n}) \leq R_{\phi,P} + \varepsilon \ , \tag{15}$$

$$\left|\{i : x_i \in A_{\delta(\lambda_n)}\}\right| \leq 2\varepsilon n \ . \tag{16}$$

The probability of such a T also converges to 1. For (15), see the proof of Theorem 3.5 in Steinwart (2005). Since $P_x(A_{\delta(\lambda_n)}) \leq \varepsilon$, (16) follows from Hoeffding's inequality. By definition of $R_{\phi,P}$, we have $R_{\phi,P} \leq R_{\phi,P}(f_{T,\lambda_n})$. Thus, (15) gives us $\lambda_n \| f_{T,\lambda_n} \|^2 \leq \varepsilon$. Now we use (10) to get

$$\left| \sum_{i=1}^{n} \alpha_i^* t_0 - \sum_{i=1}^{n} \alpha_i^* \xi_i^* \right| \leq 2\varepsilon \ . \tag{17}$$

Define three disjoint sets: $A = \{i : \xi_i^* < 0\}, B = \{i : \xi_i^* = 0\}$ and $C = \{i : \xi_i^* > 0\}$. We now show that B contains few elements. If x_i is such that $i \in B$ then $\xi_i^* = 0$ and we have $y_i f_{T,\lambda_n}(x_i) = t_0 \Rightarrow f_{T,\lambda_n}(x_i) = \pm t_0$. On the other hand, if $x_i \notin A_{\delta(\lambda_n)}$ then $\min\{|f_{P,\lambda_n}(x_i) - t_0|, |f_{P,\lambda_n}(x_i) + t_0|\} > \delta(\lambda_n) \geq m(\lambda_n)$, and hence, by (14), $f_{T,\lambda_n}(x_i) \neq \pm t_0$. Thus we can have at most $2\varepsilon n$ elements in the set B by (16). Equation (9) gives us a bound on α_i^* for $i \in B$ and therefore

$$\left| \sum_{i \in B} \alpha_i^* t_0 \right| \leq 2\varepsilon n \times h'(0) t_0 / n = 2h'(0) t_0 \varepsilon \; . \tag{18}$$

Using (9), we get $\alpha_i = 0$ for $i \in A$. By definition of B, $\xi_i^* = 0$ for $i \in B$. Therefore, (17) and (18) give us

$$\left| \sum_{i \in C} \alpha_i^* t_0 - \sum_{i \in C} \alpha_i^* \xi_i^* \right| \leq 2(1 + h'(0) t_0)\varepsilon = c_1 \varepsilon \; .$$

where $c_1 = 2(1 + h'(0)t_0)$ is just a constant. We use (9) once again to write α_i^* as $h'(\xi_i^*)/n$ for $i \in C$:

$$\left| \frac{1}{n} \sum_{i \in C} h'(\xi_i^*) t_0 - \frac{1}{n} \sum_{i \in C} h'(\xi_i^*) \xi_i^* \right| < c_1 \varepsilon \; . \tag{19}$$

Denote the cardinality of the sets B and C by N_B and N_C respectively. Then we have $N_C \leq \#SV(f_{T,\lambda_n}) \leq N_C + N_B$. But we showed that $N_B \leq 2\varepsilon n$ and therefore

$$\frac{N_C}{n} \leq \frac{\#SV(f_{T,\lambda_n})}{n} \leq \frac{N_C}{n} + 2\varepsilon \; . \tag{20}$$

Observe that $(\xi_i^*)_+ = 0$ for $i \in A \cup B$ and $(\xi_i^*)_+ = \xi_i^*$ for $i \in C$. Thus, we can extend the sums in (19) to the whole training set.

$$\left| \frac{1}{n} \sum_{i=1}^{n} h'((\xi_i^*)_+) t_0 - (n - N_C) \frac{h'(0)t_0}{n} - \frac{1}{n} \sum_{i=1}^{n} h'((\xi_i^*)_+)(\xi_i^*)_+ \right| < c_1 \varepsilon \; .$$

Now let $c_2 = c_1 / h'(0)t_0$ and rearrange the above sum to get

$$\left| \frac{N_C}{n} - \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{h'((\xi_i^*)_+)t_0 - h'((\xi_i^*)_+)(\xi_i^*)_+}{h'(0)t_0} \right) \right| \leq c_2 \varepsilon \; . \tag{21}$$

Define $g(t)$ as

$$g(t) = 1 - \frac{h'((t_0 - t)_+)t_0 - h'((t_0 - t)_+)(t_0 - t)_+}{h'(0)t_0} \; .$$

Now (21) can be written as

$$\left| \frac{N_C}{n} - \mathbb{E}_T g(y f_{T,\lambda_n}(x)) \right| \leq c_2 \varepsilon \; . \tag{22}$$

Step 3. We will now show that the empirical average of $g(yf_{T,\lambda_n}(x))$ is close to its expectation. We can bound the norm of f_{T,λ_n} as follows. The optimum value for the objective function in (1) is upper bounded by the value it attains at $f = 0$. Therefore,

$$\lambda_n \| f_{T,\lambda_n} \|^2 + R_{\phi,T}(f_{T,\lambda_n}) \leq \lambda_n . 0^2 + R_{\phi,T}(0) = \phi(0) = h(t_0)$$

which, together with (13), implies that

$$\begin{aligned} \| f_{T,\lambda_n} \| &\leq \sqrt{\frac{h(t_0)}{\lambda_n}} , \\ \| f_{T,\lambda_n} \|_\infty &\leq K\sqrt{\frac{h(t_0)}{\lambda_n}} . \end{aligned} \tag{23}$$

Let $\mathcal{F}_{\lambda_n}$ be the class of functions with norm bounded by $\sqrt{h(t_0)/\lambda_n}$. The covering number in 2-norm of the class satisfies (see, for example, Definition 1 and Corollary 3 in Zhang (2002)):

$$\mathcal{N}_2(\mathcal{F}_{\lambda_n}, \varepsilon, n) \leq e^{\frac{Kh(t_0)}{\lambda_n \varepsilon^2} \log(2n+1)} . \tag{24}$$

Define $L_g(\lambda_n)$ as

$$L_g(\lambda_n) = \sup \left\{ \frac{|g(t) - g(t')|}{|t - t'|} : t, t' \in \left[-K\sqrt{\frac{h(t_0)}{\lambda_n}}, +K\sqrt{\frac{h(t_0)}{\lambda_n}} \right], t \neq t' \right\} . \tag{25}$$

Let $\mathcal{G}_{\lambda_n} = \{(x,y) \mapsto g(yf(x)) : f \in \mathcal{F}_{\lambda_n}\}$. We can express the covering numbers of this class in terms of those of $\mathcal{F}_{\lambda_n}$ (see, for example, Lemma 14.13 on p. 206 in Anthony and Bartlett (1999)):

$$\mathcal{N}_2(\mathcal{G}_{\lambda_n}, \varepsilon, n) \leq \mathcal{N}_2(\mathcal{F}_{\lambda_n}, \varepsilon / L_g(\lambda_n), n) . \tag{26}$$

Now, using a result of Pollard (see Pollard, 1984, Section II.6, p. 30) and the fact that 1-norm covering numbers are bounded above by 2-norm covering numbers, we get

$$P^n \left(T \in (\mathcal{X} \times \mathcal{Y})^n : \sup_{\tilde{g} \in \mathcal{G}_{\lambda_n}} |\mathbb{E}_T \tilde{g}(x,y) - \mathbb{E}_P \tilde{g}(x,y)| > \varepsilon \right) \\ \leq 8\mathcal{N}_2(\mathcal{G}_{\lambda_n}, \varepsilon/8, n) e^{-n\varepsilon^2 \lambda_n / 512 L_g^2(\lambda_n) K^2 h(t_0)} .$$

The estimates (24) and (26) imply that if

$$\frac{n\lambda_n^2}{L_g^4(\lambda_n) \log(2n+1)} \rightarrow \infty \text{ as } n \rightarrow \infty$$

then the probability of a training set which satisfies

$$\left| \mathbb{E}_T g(yf_{T,\lambda_n}(x)) - \mathbb{E}_P g(yf_{T,\lambda_n}(x)) \right| \leq \varepsilon \tag{27}$$

tends to 1 as $n \rightarrow \infty$.

Step 4. The last step in the proof is to show that $\mathbb{E}_P g(yf_{T,\lambda_n}(x))$ is close to $E_x G(\eta(x))$ for large enough n. Write $\mathbb{E}_P g(yf_{T,\lambda_n}(x))$ as

$$\mathbb{E}_P g(yf_{T,\lambda_n}(x)) = \mathbb{E}_x[\eta(x)g(f_{T,\lambda_n}(x)) + (1-\eta(x))g(-f_{T,\lambda_n}(x))] \ .$$

Note that if $t^* \in F^*_\phi(\eta)$ then

$$\eta g(t^*) + (1-\eta)g(-t^*) = G(\eta) \ . \tag{28}$$

Let us verify this for three separate cases. First, when $\eta \in (0, 1-\gamma_0] \cup [\gamma_0, 1)$ the only possible values for t^* are t_0 or $-t_0$ (by Theorem 2, Part 1). Since, $g(t_0) = 0$ and $g(-t_0) = 1/(1-\gamma_0)$ the equality holds. Second, when $\eta = 1$ (the argument for $\eta = 0$ is similar), the left hand side is $g(t^*)$. In this case $t^* \geq t_0$, but the definition of g implies that $g(t) = 0$ for all $t \geq t_0$. Since $G(1) = 0$, the equality holds in this case too. Lastly, for $\eta \in (\gamma_0, 1-\gamma_0)$ we have

$$\eta g(t^*) + (1-\eta)g(-t^*) = 1 - \frac{t^*}{t_0 h'(0)}\left(\eta h'(t_0 - t^*) - (1-\eta)h'(t_0 + t^*)\right) \ .$$

Since t^* minimizes $\eta h(t_0 - t) + (1-\eta)h(t_0 + t)$ and h is differentiable, we have $\eta h'(t_0 - t^*) - (1-\eta)h'(t_0 + t^*) = 0$. Thus, we have verified (28) for all $\eta \in [0,1]$.

Define the sets $E_n = \{x \in \mathcal{X} : \rho(f_{T,\lambda_n}(x), F^*_\phi(\eta(x)) \geq \varepsilon\}$. We have $P_x(E_n) \to 0$ by (3). We now bound the difference between the two quantities of interest.

$$\begin{aligned}
&\left|\,\mathbb{E}_P g(yf_{T,\lambda_n}(x)) - \mathbb{E}_x G(\eta(x))\,\right| \\
&= \left|\,\mathbb{E}_x[\eta(x)g(f_{T,\lambda_n}(x)) + (1-\eta(x))g(-f_{T,\lambda_n}(x))] - \mathbb{E}_x G(\eta(x))\,\right| \\
&\leq \mathbb{E}_x \left|\,\eta(x)g(f_{T,\lambda_n}(x)) + (1-\eta(x))g(-f_{T,\lambda_n}(x)) - G(\eta(x))\,\right| \\
&= I_1 + I_2 \leq |I_1| + |I_2|
\end{aligned} \tag{29}$$

where the integrals I_1 and I_2 are

$$\begin{aligned}
I_1 &= \int_{E_n} \eta(x)g(f_{T,\lambda_n}(x)) + (1-\eta(x))g(-f_{T,\lambda_n}(x)) - G(\eta(x))\, dP_x \ , \\
I_2 &= \int_{\mathcal{X}\setminus E_n} \eta(x)g(f_{T,\lambda_n}(x)) + (1-\eta(x))g(-f_{T,\lambda_n}(x)) - G(\eta(x))\, dP_x \ .
\end{aligned}$$

Using (23) and (25) we bound $|g(\pm f_{T,\lambda_n}(x))|$ by $g(0) + L_g(\lambda_n)K\sqrt{h'(t_0)/\lambda_n}$. Since $g(0) = 1$ and $|G(\eta)| \leq 1$, we have

$$|I_1| \leq \left(1 + g(0) + L_g(\lambda_n)K\sqrt{\frac{h'(t_0)}{\lambda_n}}\right) P_x(E_n) \ .$$

If $\lambda_n \to 0$ slowly enough so that $L_g(\lambda_n)P_x(E_n)/\sqrt{\lambda_n} \to 0$, then for large n, $|I_1| \leq \varepsilon$. To bound $|I_2|$, observe that for $x \in \mathcal{X} \setminus E_n$, we can find a $t^* \in F^*_\phi(\eta(x))$, such that $|\,f_{T,\lambda_n}(x) - t^*| \leq \varepsilon$. Therefore,

$$\eta(x)g(f_{T,\lambda_n}(x)) + (1-\eta(x))g(-f_{T,\lambda_n}(x)) = \eta(x)g(t^*) + (1-\eta(x))g(-t^*) + \Delta \ .$$

where $|\Delta| \leq c_3\varepsilon$ and the constant c_3 does not depend on λ_n. Using (28), we can now bound $|I_2|$:

$$|I_2| \leq c_3\varepsilon(1 - P_x(E_n)) \leq c_3\varepsilon \ .$$

We now use (29) to get

$$\left|\,\mathbb{E}_P g(yf_{T,\lambda_n}(x)) - \mathbb{E}_x G(\eta(x))\,\right| \leq (c_3 + 1)\varepsilon \ . \tag{30}$$

Finally, combining (20), (22), (27) and (30) proves the theorem. ■

5. Conclusion

We saw that the decision functions obtained using minimization of regularized empirical ϕ-risk approach $F_\phi^*(\eta(\cdot))$. It is not possible to preserve sparseness on intervals where $F_\phi^*(\cdot)$ is invertible. For the regions outside that interval, sparseness is maintained to some extent. For many convex loss functions, the general lower bounds known previously turned out to be quite loose.

But that leaves open the possibility that the previously known lower bounds are actually achievable by some loss function lying outside the class of loss functions we considered. However, we conjecture that it is not possible. Note that the right hand side of Theorem 5 only depends on the left derivative of the loss function at t_0 and the right derivative at $-t_0$. The derivatives at other points do not affect the asymptotic number of support vectors. This suggests that the assumption of the differentiability of ϕ before the point where it attains its minimum can be relaxed. It may be that results on the continuity of solution sets of convex optimization problems can be applied here (e.g., see Fiacco, 1983).

Acknowledgments

This work was supported in part by ARO under MURI grant DAAD 190210383. Thanks to Grace Wahba and Laurent El Ghaoui for helpful discussions.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Large margin classifiers: Convex loss, low noise and convergence rates. In *Advances in Neural Information Processing Systems* **16**. MIT Press, 2004.

Anthony V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press, New York, 1983.

Gábor Lugosi and Nicolas Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32(1):30–55, 2004.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

Ingo Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.

Ingo Steinwart. Sparseness of support vector machines – some asymptotically sharp bounds. In *Advances in Neural Information Processing Systems* **16**. MIT Press, 2004.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

Grace Wahba. Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences USA*, 99(26):16524–16530, 2002.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.