

Sparseness vs Estimating Conditional Probabilities: Some Asymptotic Results

Peter L. Bartlett¹ and Ambuj Tewari²

¹ Division of Computer Science and Department of Statistics
University of California, Berkeley
`bartlett@cs.berkeley.edu`

² Division of Computer Science
University of California, Berkeley
`ambuj@cs.berkeley.edu`

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 bounds for the number of support vectors. This enables us to characterize the exact trade-off between sparseness and the ability to estimate conditional probabilities for these loss functions.

1 Introduction

Consider the following familiar setting of a binary classification problem. A sequence $T = ((x_1, y_1), \dots, (x_n, y_n))$ 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 $f_T : \mathcal{X} \rightarrow \mathbb{R}$ and output $\text{sign}(f_T(x))$ as the predicted label of x .

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

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

Here, H is a reproducing kernel Hilbert space (RKHS) of some kernel k , $\lambda > 0$ is a regularization parameter and $\phi : \mathbb{R} \rightarrow [0, \infty)$ is a convex loss function. Since optimization problems based on the non-convex function 0-1 loss $t \mapsto I_{(t \leq 0)}$ (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 [2, 4, 9, 12]. It has been observed that different cost functions have different properties and it is important to choose a loss function judiciously (see, for example, [10]). 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 [7]. 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

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

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 [8] 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 $(\gamma, 1 - \gamma)$ for some $\gamma \in (0, 1/2)$, 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 $(\gamma, 1 - \gamma)$. Moreover, one cannot recover sparseness by giving up the ability to estimate conditional probabilities in some sub-interval of $(\gamma, 1 - \gamma)$. The only way to do that is to increase γ thereby shortening the interval $(\gamma, 1 - \gamma)$. We also derive sharp bounds on the asymptotic number of support vectors for a family of loss functions of the form:

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

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 $(\gamma, 1 - \gamma)$ for some value of γ . The asymptotic fraction of support vectors is then $\mathbb{E}_x G(\eta(x))$, where $G(\eta)$ is a function that increases linearly from 0 to 1 as η goes from 0 to γ . 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).

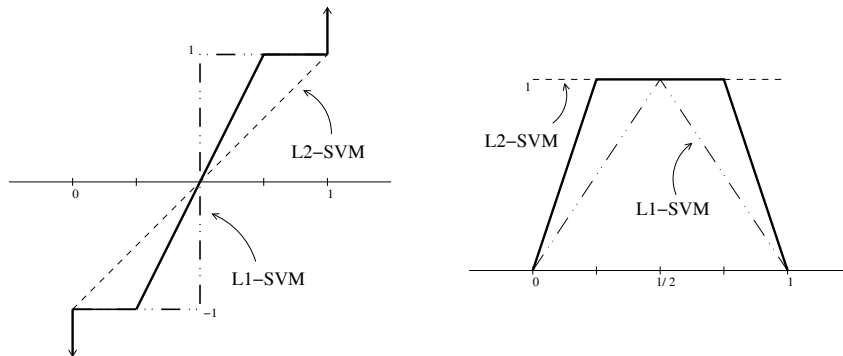


Fig. 1. Plots of $F_\phi^*(\eta)$ (left) and $G(\eta)$ (right) for a loss function which is a convex combination of the L1 and L2-SVM loss functions. Dashed lines represent the corresponding plots for the original loss functions.

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 $f : \mathcal{X} \rightarrow \mathbb{R}$, 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 $F_\phi^*(\eta)$ 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 [7] 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 [2]. If ϕ is classification calibrated then it is guaranteed that for any sequence f_n such that $R_{\phi,P}(f_n) \rightarrow R_{\phi,P}$, we have $R_P(f_n) \rightarrow R_P$. 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

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

Steinwart [9] has shown that if one uses an 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 $\lambda_n \rightarrow 0$ sufficiently slowly, then $R_{\phi,P}(f_{T,\lambda_n}) \rightarrow R_{\phi,P}$. In another paper [7], 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 $\epsilon > 0$

$$P_x(\{x \in \mathcal{X} : \rho(f_{T,\lambda_n}(x), F_\phi^*(\eta(x))) \geq \epsilon\}) \rightarrow 0 \tag{3}$$

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 [7] 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 [7] 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

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

For the last step, we simply defined

$$G(\eta) = \eta I_{(0 \notin \partial\phi(F_\phi^*(\eta)))} + (1 - \eta) I_{(0 \notin \partial\phi(-F_\phi^*(\eta)))} . \quad (4)$$

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 $G(\eta) = 1$ on $[0, 1]$.*

Proof. $t_0 = \infty$ implies that for all t , $0 \notin \partial\phi(t)$. Using (4), we get $G(\eta) = \eta \cdot 1 + (1 - \eta) \cdot 1 = 1$. \square

Therefore, let us assume that $t_0 < \infty$. The next lemma tell us about the signs of $\phi'_-(t_0)$ and $\phi'_+(t_0)$.

Lemma 2. *If $t_0 < \infty$, then $\phi'_-(t_0) \leq 0$ and $\phi'_+(t_0) \geq 0$.*

Proof. Suppose $\phi'_-(t_0) > 0$. This implies $\partial\phi(t_0) > 0$. Since subdifferential is a monotone operator, we have $\partial\phi(t) > 0$ for all $t > t_0$. By definition of t_0 , $0 \notin \partial\phi(t)$ for $t < t_0$. Thus, $\{t : 0 \in \partial\phi(t)\} = \emptyset$, which contradicts the fact that $t < \infty$. Now, suppose that $\phi'_+(t_0) = -\epsilon$, such that $\epsilon > 0$. Since $\lim_{t' \downarrow t} \phi'_-(t') = \phi'_+(t_0)$ (see [6], Theorem 24.1), we can find a $t' > t_0$ sufficiently close to t_0 such that $\phi'_-(t') \leq -\epsilon/2$. Therefore, by monotonicity of the subdifferential, $\partial\phi(t) < 0$, for all $t < t'$. This implies $t' \leq \inf\{t : 0 \in \partial\phi(t)\}$, which is a contradiction since $t' > t_0$. \square

The following lemma describes the function $F_\phi^*(\eta)$ near 0 and 1. Note that we have $\phi'_-(-t_0) \leq \phi'_+(-t_0) \leq \phi'(0) < 0$. Also $\phi'(0) \leq \phi'_-(t_0) \leq 0$.

Lemma 3. $t_0 \in F_\phi^*(\eta)$ iff $\eta \in [1 - \gamma, 1]$, where γ is defined as

$$\gamma = \frac{\phi'_-(t_0)}{\phi'_-(t_0) + \phi'_+(-t_0)}.$$

Moreover, $F_\phi^*(\eta)$ is the singleton set $\{t_0\}$ for $\eta \in (1 - \gamma, 1)$.

Proof. $t_0 \in F_\phi^*(\eta) \Leftrightarrow t_0$ minimizes $C(\eta, \cdot) \Leftrightarrow 0 \in \partial_2 C(\eta, t_0)$, where ∂_2 denotes that the subdifferential is with respect to the second variable. This is because $C(\eta, \cdot)$, being a linear combination of convex functions, is convex. Thus, a necessary and sufficient condition for a point to be a minimum is that the subdifferential there should contain zero. Now, using the linearity of the subdifferential operator and the chain rule, we get

$$\begin{aligned} \partial_2 C(\eta, t_0) &= \eta \partial\phi(t_0) - (1 - \eta) \partial\phi(-t_0) \\ &= [\eta \phi'_-(t_0) - (1 - \eta) \phi'_+(-t_0), \eta \phi'_+(t_0) - (1 - \eta) \phi'_-(-t_0)]. \end{aligned}$$

Hence, $0 \in \partial_2 C(\eta, t_0)$ iff the following two conditions hold.

$$\eta \phi'_-(t_0) - (1 - \eta) \phi'_+(-t_0) \leq 0 \tag{5}$$

$$\eta \phi'_+(t_0) - (1 - \eta) \phi'_-(-t_0) \geq 0 \tag{6}$$

The inequality (6) holds for all $\eta \in [0, 1]$ since $\phi'_+(t_0) \geq 0$ and $\phi'_-(-t_0) < 0$. The other inequality is equivalent to

$$\eta \geq \frac{-\phi'_+(-t_0)}{-\phi'_-(t_0) - \phi'_+(-t_0)}.$$

Moreover, the inequalities are strict when $\eta \in (1 - \gamma, 1)$. Therefore, t_0 is the unique minimizer of $C(\eta, \cdot)$ for these values of η . \square

Corollary 4. $-t_0 \in F_\phi^*(\eta)$ iff $\eta \in [0, \gamma]$. Moreover, $F_\phi^*(\eta)$ is the singleton set $\{-t_0\}$ for $\eta \in (0, \gamma)$.

Proof. Straightforward once we observe that $F_\phi^*(1 - \eta) = -F_\phi^*(\eta)$. \square

The next lemma states that if $F_\phi^*(\eta_1)$ and $F_\phi^*(\eta_2)$ intersect for $\eta_1 \neq \eta_2$ then ϕ must have points of non-differentiability. This means that differentiability of the loss function ensures that one can uniquely identify η via any element in $F_\phi^*(\eta)$.

Lemma 5. Suppose $\eta_1 \neq \eta_2$ and $\eta_1, \eta_2 \in (\gamma, 1 - \gamma)$. Then $F_\phi^*(\eta_1) \cap F_\phi^*(\eta_2) \neq \emptyset$ implies that

- $F_\phi^*(\eta_1) \cap F_\phi^*(\eta_2)$ is a singleton set ($= \{t\}$ say).
- ϕ is not differentiable at one of the points $t, -t$.

Proof. Without loss of generality assume $\eta_1 > \eta_2$. Suppose $t > t'$ and $t, t' \in F_\phi^*(\eta_1) \cap F_\phi^*(\eta_2)$. This contradicts the fact that F_ϕ^* is monotonic since $t' \in F_\phi^*(\eta_1)$, $t \in F_\phi^*(\eta_2)$ and $t' < t$. This establishes the first claim. To prove the second claim, suppose $F_\phi^*(\eta_1) \cap F_\phi^*(\eta_2) = \{t\}$ and assume, for sake of contradiction, that ϕ is differentiable at t and $-t$. Since $\eta_1, \eta_2 \in (\gamma, 1 - \gamma)$, Lemma 3 and Corollary 4 imply that $t \neq \pm t_0$. Therefore, $t \in (-t_0, t_0)$ and $\phi'(t), \phi'(-t) > 0$. Also, $t \in F_\phi^*(\eta_1) \cap F_\phi^*(\eta_2)$ implies that

$$\begin{aligned}\eta_1 \phi'(t) - (1 - \eta_1) \phi'(-t) &= 0 \\ \eta_2 \phi'(t) - (1 - \eta_2) \phi'(-t) &= 0.\end{aligned}$$

Subtracting and rearranging, we get

$$(\phi'(t) + \phi'(-t))(\eta_1 - \eta_2) = 0$$

which is absurd since $\eta_1 > \eta_2$ and $\phi'(t), \phi'(-t) > 0$. \square

Theorem 6. Let ϕ be an classification calibrated convex loss function such that $t_0 = \inf\{t : 0 \in \partial\phi(t)\} < \infty$. Then, for $G(\eta)$ as defined in (4), we have

$$G(\eta) = \begin{cases} 1 & \eta \in (\gamma, 1 - \gamma) \\ \min\{\eta, 1 - \eta\} & \eta \in [0, \gamma] \cup [1 - \gamma, 1] \end{cases} \quad (7)$$

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

Proof. Using Lemmas 2 and 3, we have $0 \in \partial\phi(F_\phi^*(\eta))$ for $\eta \in [1 - \gamma, 1]$. If $\eta < 1 - \gamma$, Lemma 3 tells us that $t_0 \notin F_\phi^*(\eta)$. Since F_ϕ^* is monotonic, $F_\phi^*(\eta) < t_0$. Since $t_0 = \inf\{t : 0 \in \partial\phi(t)\}$, $0 \notin \partial\phi(F_\phi^*(\eta))$ for $\eta \in [0, 1 - \gamma)$. Thus, we can write $I_{(0 \notin \partial\phi(F_\phi^*(\eta)))}$ as $I_{(\eta \notin [1 - \gamma, 1])}$. Also $I_{(0 \notin \partial\phi(-F_\phi^*(\eta)))} = I_{(0 \notin \partial\phi(F_\phi^*(1 - \eta)))}$. Plugging this in (4), we get

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

Since $\gamma \leq 1/2$, we can write $G(\eta)$ in the form given above. \square

Corollary 7. *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. Lemma 3 and Corollary 4 tell us that $\eta_1 \in (\gamma, 1 - \gamma)$. Rest follows from Theorem 6. \square

The preceding theorem and corollary have important implications. First, we can hope to have sparseness only for values of $\eta \in [0, \gamma] \cup [1 - \gamma, 1]$. 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 is $1/2$ and 0 respectively. For these two classifiers, the lower bounds $\mathbb{E}_x G(\eta(x))$ obtained after plugging in γ in (7) are the ones proved initially [7]. For the L1-SVM, the bound was later significantly improved [8]. 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 sharp improved bounds, that the bound is indeed loose for a family of loss functions.

4 Improved Bounds

We will consider convex loss functions of the form

$$\phi(t) = h((t_0 - t)_+) \tag{8}$$

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. Since we are not interested in everywhere differentiable loss functions we want a strict inequality. In other words the loss function is constant for all $t \geq t_0$ and is continuously differentiable before that. 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:

$$\arg \min_w \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \phi(y_i w^T x_i) . \tag{9}$$

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

$$\arg \min_w \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \phi(t_0 - \xi_i) \tag{10}$$

$$\text{subject to } \xi_i = t_0 - y_i w^T x_i \text{ for all } i . \tag{11}$$

Introducing Lagrange multipliers, we get the Lagrangian:

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

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

$$w = \frac{1}{2\lambda} \sum_{i=1}^n \alpha_i y_i x_i \quad (12)$$

$$\alpha_i \in -\partial\phi(t_0 - \xi_i)/n . \quad (13)$$

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

$$-\partial\phi(t_0 - \xi_i)/n = \begin{cases} \{h'(\xi_i)/n\} & \xi_i > 0 \\ [0, h'(0)/n] & \xi_i = 0 \\ \{0\} & \xi_i < 0 . \end{cases} \quad (14)$$

Let (w^*, ξ_i^*) be a solution of (10). Then we have

$$\begin{aligned} \lambda \|w^*\|^2 &= \lambda (w^*)^T \left(\frac{1}{2\lambda} \sum_{i=1}^n \alpha_i^* y_i x_i \right) \\ &= \frac{1}{2} \sum_{i=1}^n \alpha_i^* y_i (w^*)^T x_i = \frac{1}{2} \sum_{i=1}^n \alpha_i^* (t_0 - \xi_i^*) . \end{aligned} \quad (15)$$

4.2 Asymptotic Fraction of Support Vectors

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 [8], 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 (8). Define $G(\eta)$ as

$$G(\eta) = \begin{cases} \eta/\gamma & 0 \leq \eta \leq \gamma \\ 1 & \gamma < \eta < 1 - \gamma \\ (1 - \eta)/\gamma & 1 - \gamma \leq \eta \leq 1 \end{cases}$$

where $\gamma = h'(0)/(h'(0) + h'(2t_0))$. Since ϕ is differentiable on $(-t_0, t_0)$, Lemma 5 implies that F_ϕ^* is invertible on $(\gamma, 1 - \gamma)$. Thus, one can estimate conditional probabilities in the interval $(\gamma, 1 - \gamma)$. Let $\#SV(f_{T,\lambda})$ denote the number of support vectors in the solution (2):

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

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

Theorem 8. *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 (8), and a regularization sequence which tends to 0 sufficiently slowly, we have*

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

in probability.

Proof. Let us fix an $\epsilon > 0$. 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 $\lambda_n \rightarrow 0$ slowly. Since $f_{P,\lambda}$ is an analytic function, for any constant c , we have

$$P_x(\{x \in X : f_{P,\lambda}(x) = c\}) > 0 \Rightarrow f(x) = c \text{ } P_x\text{-a.s.} \quad (16)$$

Assume that $P_x(\{x \in \mathcal{X} : f_{P,\lambda}(x) = t_0\}) > 0$. By (16), 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 $R_{\phi,P}(f_{P,\lambda}) \rightarrow R_{\phi,P}$ and $R(t_0) \neq R_{\phi,P}$ 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

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

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

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

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

Let $m(\lambda) = \inf\{\delta(\lambda') : \lambda' \geq \lambda\}$ be a decreasing version of $\delta(\lambda)$. Using Proposition 33 from [7] with $\epsilon = m(\lambda_n)$, we conclude that for a sequence $\lambda_n \rightarrow 0$ sufficiently slowly, the probability of a training set T such that

$$\|f_{T,\lambda_n} - f_{P,\lambda_n}\| < m(\lambda_n)/K \quad (17)$$

converges to 1 as $n \rightarrow \infty$. 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 [8]). We now relate the 2-norm of an f to its ∞ -norm.

$$\begin{aligned} f(x) &= \langle k(x, \cdot), f(\cdot) \rangle \leq \|k(x, \cdot)\| \|f\| \\ &= \sqrt{\langle k(x, \cdot), k(x, \cdot) \rangle} \|f\| \\ &= k(x, x) \|f\| \leq K \|f\| \end{aligned} \quad (18)$$

Thus, (17) gives us

$$\|f_{T,\lambda_n} - f_{P,\lambda_n}\|_\infty < m(\lambda_n) . \quad (19)$$

Step 2. In the second step, we relate the fraction of support vectors to an empirical average. Suppose that, in addition to (19), our training set T satisfies

$$\lambda_n \|f_{T,\lambda_n}\|^2 + R_{\phi,P}(f_{T,\lambda_n}) \leq R_{\phi,P} + \epsilon \quad (20)$$

$$|\{i : x_i \in A_{\delta(\lambda_n)}\}| \leq 2\epsilon n . \quad (21)$$

The probability of such a T also converges to 1. For (20), see the proof of Theorem III.6 in [9]. Since $P_x(A_{\delta(\lambda_n)}) \leq \epsilon$, (21) 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, (20) gives us $\lambda_n \|f_{T,\lambda_n}\|^2 \leq \epsilon$. Now we use (15) to get

$$\left| \sum_{i=1}^n \alpha_i^* t_0 - \sum_{i=1}^n \alpha_i^* \xi_i^* \right| \leq 2\epsilon . \quad (22)$$

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 (19), $f_{T,\lambda_n}(x_i) \neq \pm t_0$. Thus we can have at most $2\epsilon n$ elements in the set B by (21). Equation (14) gives us a bound on α_i^* for $i \in B$ and therefore

$$\left| \sum_{i \in B} \alpha_i^* t_0 \right| \leq 2\epsilon n \times h'(0)t_0/n = 2h'(0)t_0\epsilon . \quad (23)$$

Using (14), we get $\alpha_i = 0$ for $i \in A$. By definition of B , $\xi_i^* = 0$ for $i \in B$. Therefore, (22) and (23) give us

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

where $c_1 = 2(1 + h'(0)t_0)$ is just a constant. We use (14) once again to write α_i^* as $h'(\xi_i^*)/n$ for $i \in C$:

$$\left| \frac{1}{n} \sum_{i \in C} h'(\xi_i^*)t_0 - \frac{1}{n} \sum_{i \in C} h'(\xi_i^*)\xi_i^* \right| < c_1 \epsilon . \quad (24)$$

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 $N_B \leq 2\epsilon n$ and therefore

$$\frac{N_C}{n} \leq \frac{\#SV(f_{T,\lambda_n})}{n} \leq \frac{N_C}{n} + 2\epsilon . \quad (25)$$

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 (24) to the whole training set.

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

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

$$\left| \frac{N_C}{n} - \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{h'((\xi_i^*)_+)t_0 - h'((\xi_i^*)_+)(\xi_i^*)_+}{h'(0)t_0} \right) \right| \leq c_2 \epsilon . \quad (26)$$

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 (26) can be written as

$$\left| \frac{N_C}{n} - \mathbb{E}_T g(yf_{T,\lambda_n}(x)) \right| \leq c_2 \epsilon . \quad (27)$$

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,

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

which, together with (18), implies that

$$\|f_{T,\lambda_n}\| \leq \sqrt{\frac{h(t_0)}{\lambda_n}} \quad (28)$$

$$\|f_{T,\lambda_n}\|_\infty \leq K \sqrt{\frac{h(t_0)}{\lambda_n}} . \quad (29)$$

Let \mathcal{F}_{λ_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 [11]):

$$\mathcal{N}_2(\mathcal{F}_{\lambda_n}, \epsilon, n) \leq e^{\frac{\kappa h(t_0)}{\lambda_n \epsilon^2} \log(2n+1)} . \quad (30)$$

Define $L_g(\lambda_n)$ as

$$L_g(\lambda_n) = \sup \left\{ \frac{|g(t) - g(t')|}{|t - t'|} : t, t' \in \left[-K\sqrt{\frac{h(t_0)}{\lambda_n}}, +K\sqrt{\frac{h(t_0)}{\lambda_n}} \right], t \neq t' \right\} \quad (31)$$

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}_{λ_n} (see, for example, Lemma 14.13 on p. 206 in [1]):

$$\mathcal{N}_2(\mathcal{G}_{\lambda_n}, \epsilon, n) \leq \mathcal{N}_2(\mathcal{F}_{\lambda_n}, \epsilon/L_g(\lambda_n), n) . \quad (32)$$

Now, using a result of Pollard (see Section II.6 on p. 30 in [5]) and the fact that 1-norm covering numbers are bounded above by 2-norm covering numbers, we get

$$\begin{aligned} P^n \left(T \in (\mathcal{X} \times \mathcal{Y})^n : \sup_{\tilde{g} \in \mathcal{G}_{\lambda_n}} |\mathbb{E}_T \tilde{g}(x, y) - \mathbb{E}_P \tilde{g}(x, y)| > \epsilon \right) \\ \leq 8\mathcal{N}_2(\mathcal{G}_{\lambda_n}, \epsilon/8, n) e^{-n\epsilon^2 \lambda_n / 512 L_g^2(\lambda_n) K^2 h(t_0)} . \end{aligned} \quad (33)$$

The estimates (30) and (32) 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

$$|\mathbb{E}_T g(yf_{T, \lambda_n}(x)) - \mathbb{E}_P g(yf_{T, \lambda_n}(x))| \leq \epsilon \quad (34)$$

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

$$\eta g(t^*) + (1 - \eta)g(-t^*) = G(\eta) . \quad (35)$$

This is easily verified for $\eta \in [0, \gamma] \cup [1 - \gamma, 1]$ since $g(t) = 0$ for $t \geq t_0$ and $g(-t_0) = 1/\gamma$. For $\eta \in (\gamma, 1 - \gamma)$ we have

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

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 (35) for all $\eta \in [0, 1]$. Define the sets $E_n = \{x \in \mathcal{X} : \rho(f_{T, \lambda_n}(x), F_\phi^*(\eta(x))) \geq \epsilon\}$. We have $P_x(E_n) \rightarrow 0$ by (3). We now bound the difference between the two quantities of interest.

$$\begin{aligned}
& |\mathbb{E}_P g(y f_{T, \lambda_n}(x)) - \mathbb{E}_x G(\eta(x))| \\
&= |\mathbb{E}_x [\eta(x)g(f_{T, \lambda_n}(x)) + (1 - \eta(x))g(-f_{T, \lambda_n}(x))] - \mathbb{E}_x G(\eta(x))| \\
&\leq \mathbb{E}_x |\eta(x)g(f_{T, \lambda_n}(x)) + (1 - \eta(x))g(-f_{T, \lambda_n}(x)) - G(\eta(x))| \\
&= I_1 + I_2 \leq |I_1| + |I_2|
\end{aligned} \tag{36}$$

where the integrals I_1 and I_2 are

$$I_1 = \int_{E_n} \eta(x)g(f_{T, \lambda_n}(x)) + (1 - \eta(x))g(-f_{T, \lambda_n}(x)) - G(\eta(x)) dP_x \tag{37}$$

$$I_2 = \int_{\mathcal{X} \setminus E_n} \eta(x)g(f_{T, \lambda_n}(x)) + (1 - \eta(x))g(-f_{T, \lambda_n}(x)) - G(\eta(x)) dP_x. \tag{38}$$

Using (29) and (31) 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 $\lambda_n \rightarrow 0$ slowly enough so that $L_g(\lambda_n)P_x(E_n)/\sqrt{\lambda_n} \rightarrow 0$, then for large n , $|I_1| \leq \epsilon$. 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 $|f_{T, \lambda_n}(x) - t^*| \leq \epsilon$. Therefore

$$\begin{aligned}
& \eta(x)g(f_{T, \lambda_n}(x)) + (1 - \eta(x))g(-f_{T, \lambda_n}(x)) \\
&= \eta(x)g(t^*) + (1 - \eta(x))g(-t^*) + \Delta.
\end{aligned} \tag{39}$$

where $|\Delta| \leq c_3\epsilon$ and the constant c_3 does not depend on λ_n . Using (35), we can now bound $|I_2|$:

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

We now use (36) to get

$$|\mathbb{E}_P g(y f_{T, \lambda_n}(x)) - \mathbb{E}_x G(\eta(x))| \leq (c_3 + 1)\epsilon. \tag{40}$$

Finally, combining (25), (27), (34) and (40) proves the theorem. \square

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 bound of Theorem 8 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 (see, for example, [3]).

Acknowledgements

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

1. Anthony, M. and Bartlett, P. L.: *Neural network learning: Theoretical foundations*. Cambridge University Press, Cambridge (1999)
2. Bartlett, P. L., Jordan, M.I. and McAuliffe, J.D.: Large Margin Classifiers: convex loss, low noise and convergence rates. In *Advances in Neural Information Processing Systems* **16**. MIT Press, Cambridge, MA (2004)
3. Fiacco, A. V.: *Introduction to sensitivity and stability analysis in nonlinear programming*. Academic Press, New York (1983)
4. Lugosi, G. and Vayatis, N.: On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics* **32**:1 (2004) 30–55
5. Pollard, D.: *Convergence of stochastic processes*. Springer-Verlag, New York (1984)
6. Rockafellar, R. T.: *Convex analysis*. Princeton University Press, Princeton (1970)
7. Steinwart, I.: Sparseness of support vector machines. *Journal of Machine Learning Research* **4** (2003) 1071–1105
8. Steinwart, I.: Sparseness of support vector machines – some asymptotically sharp bounds. In *Advances in Neural Information Processing Systems* **16**. MIT Press, Cambridge, MA (2004)
9. Steinwart, I.: Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, to appear
10. Wahba, G.: Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences USA* **99**:26 (2002) 16524–16530
11. Zhang, T.: Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research* **2** (2002) 527–550
12. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* **32**:1 (2004) 56–85