Data Mining and Knowledge Discovery, 6, 259–275, 2002

Support Vector Machines and the Bayes Rule in Classification

YI LIN yilin@stat.wisc.edu
Department of Statistics, University of Wisconsin, Madison, 1210 West Dayton Street, Madison, WI 53706-1685, USA

Editors: Fayyad, Mannila, Ramakrishnan

Received December 30, 1999; Revised May 1, 2001

Abstract. The Bayes rule is the optimal classification rule if the underlying distribution of the data is known. In practice we do not know the underlying distribution, and need to "learn" classification rules from the data. One way to derive classification rules in practice is to implement the Bayes rule approximately by estimating an appropriate classification function. Traditional statistical methods use estimated log odds ratio as the classification function. Support vector machines (SVMs) are one type of large margin classifier, and the relationship between SVMs and the Bayes rule was not clear. In this paper, it is shown that the asymptotic target of SVMs are some interesting classification functions that are directly related to the Bayes rule. The rate of convergence of the solutions of SVMs to their corresponding target functions is explicitly established in the case of SVMs with quadratic or higher order loss functions and spline kernels. Simulations are given to illustrate the relation between SVMs and the Bayes rule in other cases. This helps understand the success of SVMs in many classification studies, and makes it easier to compare SVMs and traditional statistical methods.

Keywords: support vector machine, classification, the Bayes rule, reproducing kernel, reproducing kernel Hilbert space, regularization methods

1. Introduction

Support vector machines (SVMs) have proved highly successful in a number of classification studies. In the classification problems, we are given a training data set of n subjects, and for each subject i, $i = 1, 2, \ldots, n$ in the training data set, we observe an explanatory vector $\mathbf{x}_i \in R^d$, and a label y_i indicating one of several given classes to which the subject belongs. The observations in the training set are assumed to be iid from an unknown probability distribution $P(\mathbf{x}, y)$, or equivalently, they are independent random realizations of the random pair $(\mathbf{X}, Y)$ that has cumulative probability distribution $P(\mathbf{x}, y)$. The task of classification is to derive from the training set a good classification rule, so that once we are given the $\mathbf{x}$ value of a new subject, we can assign a class label to the subject. One possible criterion for judging the quality of a classification rule is the expected misclassification rate, but in practice it is also possible that some other loss function is more appropriate. If we knew the underlying probability distribution $P(\mathbf{x}, y)$, we could derive the optimal classification rule with respect to any given loss function. This optimal rule is usually called the Bayes rule for classification.

In the following we will concentrate on the case where there are only two classes, and where the expected misclassification rate is used as the criterion. This is the case in which SVMs are best developed. In this situation the label y is either 1 or -1. A classification rule is a mapping from R^d to $\{-1, 1\}$. It is easy to see that the expected misclassification rate R of any classification rule $\eta : R^d \to \{-1, 1\}$ can be written as

$$R = E[|Y - \eta(\mathbf{X})|/2] = E[1 - Y\eta(\mathbf{X})]_+/2. \tag{1}$$

Here $(\cdot)_+$ is a function such that τ_+ is τ, if $\tau > 0$; and is 0, otherwise. This function is not really needed in (1), since $[1 - Y\eta(\mathbf{X})]$ will be nonnegative for any classification rule η. For a general real function $f : R^d \to R$, we call $E[1 - Yf(\mathbf{X})]_+$ the generalized comparative Kullback Leibler (GCKL) measure. See Wahba et al. (2000).

Let

$$p(\mathbf{x}) = \Pr\{Y = 1 \mid \mathbf{X} = \mathbf{x}\}$$

Then the (Bayes) rule that minimizes the expected misclassification rate is $\eta^*(\mathbf{x}) = \text{sign}[p(\mathbf{x}) - 1/2]$, or equivalently, $\text{sign}[g(\mathbf{x})]$, where $g(\mathbf{x})$ is the log odds ratio $\log[p(\mathbf{x})/(1 - p(\mathbf{x}))]$.

Since we do not know $P(\mathbf{x}, y)$ in practice, but are only given a sample from it, we can not obtain this Bayes rule exactly. So the question is often how to find a classification rule whose performance is close to that of the Bayes rule, or how to approximate the Bayes rule. Traditional statistical methods try to estimate $[p(\mathbf{x}) - 1/2]$ (or the log odds ratio $g(\mathbf{x})$) from the training data, and then approximate the Bayes rule with $\text{sign}[\hat{p}(\mathbf{x}) - 1/2]$ (or $\text{sign}[\hat{g}(\mathbf{x})]$). Here $\hat{p}(\mathbf{x})$ and $\hat{g}(\mathbf{x})$ are the estimates of $p(\mathbf{x})$ and $g(\mathbf{x})$, respectively. Friedman (1997) discussed how the bias and variance components of the estimation error affects classification error when the estimate is used in a classification rule.

The support vector machine methodology was introduced in Boser et al. (1992). See also Cortes and Vapnik (1995) and Vapnik (1995). Support vector machines have proved highly successful in a number of classification studies. The linear SVMs are motivated by the geometric interpretation of maximizing the margin, and the nonlinear SVMs are characterized by the use of reproducing kernels. (The reproducing kernel is sometimes called kernel in the SVM literature, not to be confused with the kernel estimators in the nonparametric statistics literature). For a tutorial on SVMs for classification, see Burges (1998). Here we give a brief summary of support vector machines for classification, starting from the simple linear support vector machines and moving on to the nonlinear support vector machines.

We start with the simplest case: linear support vector machine trained on separable data. Suppose the two classes of points in the training set can be separated by a linear hyperplane $(\mathbf{x} \cdot \mathbf{w}) + b = 0$, where $\mathbf{w}$ is the normal to the hyperplane. Let d_+ and d_- be the shortest distance from the separating hyperplane to the closest positive and negative examples, respectively. Define the margin of the separating hyperplane to be $(d_+ + d_-)$. It is natural to look for the separating hyperplane with the largest margin. This amounts to the hard margin linear support vector machine: Find $\mathbf{w} \in R^d$, $b \in R$, to minimize $\|\mathbf{w}\|^2$, subject to

$$(\mathbf{x}_i \cdot \mathbf{w}) + b \geq +1 \quad \text{for } y_i = +1; \tag{2}$$

$$(\mathbf{x}_i \cdot \mathbf{w}) + b \leq -1 \quad \text{for } y_i = -1; \tag{3}$$

Once such $\mathbf{w}$ and b are found, we classify according to the sign of $[(\mathbf{w} \cdot \mathbf{x}) + b]$.

When the points in the training data set are not linearly separable, constraints (2) and (3) can not be satisfied simultaneously. We can introduce nonnegative slack variables ξ's to overcome this difficulty, and this results in the soft margin linear support vector machine: Find $\mathbf{w} \in R^d$, $b \in R$, and ξ_i, $i = 1, 2, \ldots, n$, to minimize $(1/n)(\sum_{i=1}^n \xi_i)^q + \lambda \|\mathbf{w}\|^2$, under the constraints

$$(\mathbf{x}_i \cdot \mathbf{w}) + b \geq +1 - \xi_i \quad \text{for } y_i = +1; \tag{4}$$

$$(\mathbf{x}_i \cdot \mathbf{w}) + b \leq -1 + \xi_i \quad \text{for } y_i = -1; \tag{5}$$

$$\xi_i \geq 0, \ \forall i.$$

Here λ is a parameter to be chosen by the user, and q is a positive integer. This is a convex programming problem for any positive integer q; for $q = 2$ and $q = 1$, it is also a quadratic programming problem. The choice $q = 1$ has the further advantage that the Wolfe dual problem has a particularly simple form, and this is the most common choice.

The nonlinear support vector machine maps the input variable into a high dimensional (often infinite dimensional) feature space, and applies the linear support vector machine in the feature space. Computationally, this can be achieved by the application of a (reproducing) kernel. A reproducing kernel over R^d is a positive definite function on $R^d \otimes R^d$. For an introduction to reproducing kernels and reproducing kernel Hilbert spaces, see Wahba (1990). Let H_K be the reproducing kernel Hilbert space with reproducing kernel $K(\mathbf{s}, \mathbf{t})$, $\mathbf{s}, \mathbf{t} \in R^d$. It has been shown (Wahba, 1999; Evgeniou et al., 1999), that the SVM with kernel K is equivalent to a regularization problem in H_K. The SVM with reproducing kernel K first minimizes

$$\frac{1}{n} \sum_{i=1}^n [(1 - y_i f_i)_+]^q + \lambda \|h\|^2_{H_K} \tag{6}$$

over all the functions of the form $f(\mathbf{x}) = h(\mathbf{x}) + \text{const}$, and $h \in H_K$. Here $f_i = f(\mathbf{x}_i)$. Once the minimizer $\tilde{f}$ is found, then the SVM classification rule is $\text{sign}[\tilde{f}(\mathbf{x})]$. A variety of reproducing kernels have been used successfully in practical applications, including polynomial kernels, Gaussian kernels, and spline kernels (Sobolev Hilbert space kernels). The reproducing kernel Hilbert spaces for the latter two types of reproducing kernels are of infinite dimension. For a review on spline kernels, see Wahba (1990). The theory of reproducing kernel Hilbert spaces ensures that the minimizer of (6) lies in a finite dimensional space, even when the minimization is carried out in an infinite dimensional reproducing kernel Hilbert space. For any positive integer q, the minimization problem (6) becomes a convex programming problem in a n-dimensional space. See Wahba et al. (2000). For $q = 1$ and $q = 2$ it is also a quadratic programming problem.

Remark 1.1. If $\{1\} \subset H_K$, the regularization problem (6) is equivalent to minimizing

$$\frac{1}{n} \sum_{i=1}^n [(1 - y_i f_i)_+]^q + \lambda \|Pf\|^2_{H_K}$$

over H_K, where Pf is the projection of f into the orthogonal complement of $\{1\}$ in H_K.

Several authors have studied the generalization performance of SVMs, See Vapnik (1995), and Shawe-Taylor and Cristianini (1998). These authors established bounds on generalization error based on VC dimension, fat shattering dimension, and the proportion of the training data achieving certain margin. However, SVMs often have very large, even infinite, VC dimension or fat shattering dimension. Hence the bounds established are often very loose, and do not provide a satisfactory explanation as to why SVMs often have good generalization performance. In this paper, we show that SVMs have some interesting asymptotic target functions. Classifying according to the sign of these target functions is equivalent to the Bayes rule. The rate of convergence of the solutions of SVMs to their corresponding target functions is explicitly established in the case of SVMs with quadratic or higher order loss functions and spline kernels. Simulations are given to illustrate the relation between SVMs and the Bayes rule in other cases. This helps explain why SVMs have been successful in practical applications, and facilitates the comparison of SVMs with other traditional statistical methods for classification.

We will also consider the regularization problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - f_i|^q + \lambda\|h\|^2_{H_K} \tag{7}$$

over all the functions of the form $f(\mathbf{x}) = h(\mathbf{x}) + \text{const}$, and $h \in H_K$. With $q = 2$, this is the penalized least square estimation; with $q = 1$ this is the penalized least absolute value estimation. As we will see later, this problem is closely related to the problem of minimizing (6).

Regularization problems similar to (6) and (7) have long been studied in statistics literature, see Wahba (1990) and the reference therein. Examples include penalized least square regression, penalized logistic regression, penalized density estimation, and regularization procedures used in more general nonlinear inverse problems. Cox and O'Sullivan (1990) provided a general framework for studying regularization methods. As in (6) and (7), the method of regularization has two components: a data fit functional component and a regularization penalty component. The data fit functional component dictates that the estimate should follow the pattern in the data, whereas the regularization penalty component imposes smoothness conditions. The data fit component usually approaches a limiting functional as $n \to \infty$. In general the limiting functional can be used to identify the target function: the target function is the minimizer of the limiting functional. Under the assumption that the target function is in the reproducing kernel Hilbert space under consideration and certain other general regularity conditions, the solution of the regularization problem approaches the target function as $n \to \infty$. We give the following simple example for illustration.

Example 1.1. Nonparametric regression. Let $(z_i, \mathbf{w_i})$, $i = 1, 2, \ldots, n$, be an independent random sample of $(Z, \mathbf{W})$. Here Z is a random variable and $\mathbf{W}$ is a random vector. Assume $E(Z \mid \mathbf{W}) = f_0(\mathbf{W})$. The task is to estimate f_0. The data fit component of penalized least square method is

$$\frac{1}{n}\sum_{i=1}^{n}[z_i - f(\mathbf{w}_i)]^2.$$

The limiting functional of this is

$$E[Z - f(\mathbf{W})]^2,$$

which is minimized by f_0.

For more examples in density estimation, hazard regression, and logistic regression, see Cox and O'Sullivan (1990).

Before we proceed further, let us introduce a simple fact:

Lemma 1.1. *For any $a \in [-1, 1]$, and $y \in \{-1, 1\}$, we have $[(1 - ya)_+]^q = |y - a|^q$.*

Proof: For any $a \in [-1, 1]$, and $y \in \{-1, 1\}$, we have

$$|y - a|^q = |y(1 - ya)|^q = |1 - ya|^q = [(1 - ya)_+]^q \qquad \square$$

In the following we first study the cases in which $q > 1$, especially the case when $q = 2$; then we consider the case $q = 1$.

2. SVMs with $q > 1$

In the SVM situation, the limiting functional of the data fit component in (6) is easily seen to be $E[(1 - Yf(\mathbf{X}))_+]^q$. The corresponding limiting functional for (7) is $E|Y - f(\mathbf{X})|^q$. The following lemma identifies the target function for SVM and (7) with $q > 1$ (From now on all proofs are given in the appendix):

Lemma 2.1. *For any $q > 1$, the minimizers of $E[(1 - Yf(\mathbf{X}))_+]^q$ and $E|Y - f(\mathbf{X})|^q$ are the same function given by*

$$f_q(\mathbf{x}) = \left[(p(\mathbf{x}))^{\frac{1}{q-1}} - (1 - p(\mathbf{x}))^{\frac{1}{q-1}}\right] \Big/ \left[(p(\mathbf{x}))^{\frac{1}{q-1}} + (1 - p(\mathbf{x}))^{\frac{1}{q-1}}\right]$$

Also, $\text{sign}[f_q(\mathbf{x})] = \text{sign}[p(\mathbf{x}) - 1/2]$ *for all $q > 1$, and the classification rule* $\text{sign}[f_q(\mathbf{x})]$ *is equivalent to the Bayes rule.*

Thus the asymptotic target of the SVM is f_q, and classifying according to the sign of f_q is equivalent to the Bayes rule. To be specific, let us now specialize to the case $q = 2$. In this case f_q simplifies to $2p - 1$. We will consider a special yet very general reproducing kernel Hilbert space, and illustrate how fast the solution of (6) approaches f_q.

For a nonnegative integer m, the Sobolev Hilbert space with order m of univariate functions on domain [0, 1], denoted by $H^m([0, 1])$, is defined by

$$H^m([0, 1]) = \left\{ f \;\middle|\; f^{(\nu)} \text{ abs. cont.}, \nu = 0, 1, \ldots, m - 1; f^{(m)} \in L_2 \right\}$$

with a norm equivalent to

$$\|f\|^2_{H^m([0,1])} = \sum_{v=0}^{m-1} (M_v f)^2 + \int_0^1 \left(f^{(m)}(u)\right)^2 du$$

where $M_v f = \int_0^1 f^{(v)}(u)\, du$, $v = 0, 1, \ldots, m-1$. The superscript on f refers to a derivative. It is typical in statistics to impose the mth order smoothness condition on a univariate function by assuming it is in $H^m([0, 1])$. For any positive integer m, the Sobolev Hilbert space $H^m([0, 1])$ is a reproducing kernel Hilbert space. The reproducing kernel of this space is derived in Wahba (1990), chapter 10, and is known as the spline kernel. For example, when $m = 2$, the reproducing kernel is

$$r(s, t) = 1 + k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s - t|),$$

where $k_1(\cdot) = \cdot - 0.5$, $k_2 = (k_1^2 - 1/12)/2$, and $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$.

Let $\otimes^d H^m$ be the tensor product space of d $H^m([0, 1])$ spaces. Then $\otimes^d H^m$ is a Hilbert space of functions on $[0, 1]^d$, and it can be identified with the Hilbert space of functions

$$\Omega_m = \left\{ f : \frac{\partial^{\|\alpha\|_1} f(\mathbf{x})}{\partial \mathbf{x}^\alpha} \in L_2([0, 1]^d),\ \forall \alpha \text{ such that } \|\alpha\|_\infty \le m \right\}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, $\alpha_i \ge 0$, $\alpha_i =$ integer, and $\|\alpha\|_1 \equiv \sum_{i=1}^d \alpha_i$, $\|\alpha\|_\infty \equiv \max\{\alpha_1, \alpha_2, \ldots, \alpha_d\}$. See Lin (1998). The space $\otimes^d H^m$ is also a reproducing kernel Hilbert space, and the reproducing kernel of this space is

$$R(\mathbf{s}, \mathbf{t}) = \prod_{j=1}^d r(s_j, t_j)$$

where $\mathbf{s} = (s_1, \ldots, s_d)$, $\mathbf{t} = (t_1, \ldots, t_d)$.

Recall that $p(\mathbf{x}) = \Pr\{Y = 1 \mid \mathbf{X} = \mathbf{x}\}$. Let the marginal density of $\mathbf{X}$ be denoted by $f_{\mathbf{X}}$. Without loss of generality, assume that $\mathbf{X}$ takes values only in the unit cube $[0, 1]^d$. Also assume that the marginal density of $\mathbf{X}$ is bounded away from 0 and ∞ in the unit cube, i.e., $0 < C_1 \le f_{\mathbf{X}}(\mathbf{x}) \le C_2 < \infty$ for some constants C_1 and C_2.

Now consider the regularization problems (6) and (7) with $H_K = \otimes^d H^m$ and $q = 2$. We denote the solution to (6) by f_*, and the solution to (7) by f^*.

Theorem 2.1. *Assume that $p(\mathbf{x})$ is in $\otimes^d H^m$, then if $\lambda \to 0$, and $n^{-1}\lambda^{-(\frac{3}{2m}+\epsilon)} \to 0$ for some $\epsilon > 0$. Then*

$$\int_{[0,1]^d} [f^* - (2p - 1)]^2 = O(\lambda) + O_p\left[n^{-1}\lambda^{-\frac{1}{2m}}\left(\log\frac{1}{\lambda}\right)^{d-1}\right].$$

$$\sup_{\mathbf{x}\in[0,1]^d} |f^* - (2p - 1)| = O\left(\lambda^{\frac{1}{2}-\frac{1}{4m}-\frac{\epsilon}{4}}\right) + O_p\left[n^{-\frac{1}{2}}\lambda^{-(\frac{1}{2m}+\frac{\epsilon}{4})}\left(\log\frac{1}{\lambda}\right)^{d-1}\right]$$

Theorem 2.2. *Assume that* $p(\mathbf{x})$ *is in* $\otimes^d H^m$, *and* $0 < p(\mathbf{x}) < 1$, $\forall \mathbf{x} \in [0,1]^d$. *Then if* $\lambda \to 0$, *and* $n^{-1}\lambda^{-(\frac{3}{2m}+\epsilon)} \to 0$ *for some* $\epsilon > 0$. *Then*

$$\int_{[0,1]^d} [f_* - (2p-1)]^2 = O_p\left[\lambda + n^{-1}\lambda^{-\frac{1}{2m}}\left(\log\frac{1}{\lambda}\right)^{d-1}\right].$$

$$\sup_{\mathbf{x}\in[0,1]^d} |f_* - (2p-1)| = O_p\left[\lambda^{\frac{1}{2}-\frac{1}{4m}-\frac{\epsilon}{4}} + n^{-\frac{1}{2}}\lambda^{-(\frac{1}{2m}+\frac{\epsilon}{4})}\left(\log\frac{1}{\lambda}\right)^{d-1}\right]$$

Remark 2.1. In the two theorems above, the smoothing parameter λ changes with n. It is actually a sequence $\lambda(n)$. How to choose the smoothing parameter is an important practical problem, and several methods have been proposed in the literature. For example, Wahba et al. (2000) considered choosing λ to minimize the estimated generalized comparative Kullback Leibler measure.

Remark 2.2. The condition $0 < p(\mathbf{x}) < 1$ in Theorem 2.2 is technical, and we believe the result should still be valid without this condition.

Remark 2.3. We believe the results stated for the *sup* norm is not the best possible. There should be room for deriving sharper bounds.

Remark 2.4. In some situations we may want to use some reproducing kernel Hilbert space other than the one considered above. For example, we may want to use the Gaussian kernel. Results similar to those stated in the theorems above should also be obtainable, given that $p(\mathbf{x})$ is in the assumed reproducing kernel Hilbert space. The bounds would usually be different, though. The order of the bounds typically depends on the rate of decay of the eigenvalues of the reproducing kernel. See Cox and O'Sullivan (1990).

Remark 2.5. For $q > 2$, similar results can be obtained on how fast the minimizers of (6) and (7) approach f_q by using the framework provided in Cox and O'Sullivan (1990).

The theorems above show that SVMs with $q = 2$ and the spline kernel solve a regularization problem to get f_*, which approaches $2p - 1$ asymptotically, then uses the sign of $f_*(\mathbf{x})$ to approximately implement the Bayes rule $\text{sign}[p(\mathbf{x}) - 1/2]$. Similarly, we can also consider solving (7) to approximate $2p - 1$.

Compared with the traditional statistical method of estimating the log odds ratio and using the sign of the estimate to approximate the Bayes rule, SVM enjoys two advantages. First, the computation load of SVM is not so heavy as that of the methods of estimating log odds ratio. Second, when $p(\mathbf{x})$ is (or is close to) 0 or 1, the log odds ratio is (or is close to) $-\infty$ or ∞, and the method of estimating log odds ratio is ineffective and computationally unstable. SVM is more suitable for this situation.

The method of minimizing (7) with $q = 2$ can be motivated by the fact that $E(Y \mid \mathbf{X} = \mathbf{x}) = 2p(\mathbf{x}) - 1$, and we recognize (7) with $q = 2$ as the penalized least square regression method for estimating $E(Y \mid \mathbf{X} = \mathbf{x})$. Intuitively, this method would not be efficient since they do not take into account the fact that $\text{Var}(Y \mid \mathbf{X} = \mathbf{x}) = 4p(\mathbf{x})[1 - p(\mathbf{x})]$ is not a

constant and is smaller at places where $p(\mathbf{x})$ is close to 0 or 1. By proceeding as if the variance is a constant, we are wasting some precision at regions where $p(\mathbf{x})$ is close to 0 or 1. However, for the purpose of classification, what concerns us most is the region where $p(\mathbf{x})$ is not too far away from 1/2, and the efficiency lost there for estimation is small.

One of the conditions of the theorems is that the reproducing kernel Hilbert space used in the regularization problem contains $p(\mathbf{x})$. It conforms to the notion that we should choose reproducing kernel so that $p(\mathbf{x})$ is in the corresponding reproducing kernel Hilbert space. This condition can be relaxed a little, (see Cox and O'Sullivan (1990)) but $p(\mathbf{x})$ should at least be close to the reproducing kernel Hilbert space.

3. SVMs with $q = 1$

This is the most commonly used SVM. In this situation, the limiting functional of the data fit component in (6) is $E[(1 - Yf(\mathbf{X}))_+]$. The corresponding limiting functional for (7) is $E|Y - f(\mathbf{X})|$. The following lemma identifies the target function for SVM and (7) with $q = 1$.

Lemma 3.1. *The minimizer of* $E[(1 - Yf(\mathbf{X}))_+]$ *and* $E|Y - f(\mathbf{X})|$ *are both* $\operatorname{sign}(p-1/2)$.

Thus instead of targeting at $(p - 1/2)$, and then using the sign of the estimate to approximate the Bayes rule, SVM with $q = 1$ takes aim directly at $\operatorname{sign}(p - 1/2)$. However, this target function typically does not lie in the commonly used reproducing kernel Hilbert space, (it is easy to see that $\operatorname{sign}(p-1/2)$ is not a smooth function unless $p(\mathbf{x})$ is always larger than 1/2 or always smaller than 1/2.) though it can be approximated arbitrarily closely in the L_2 norm by the functions in the reproducing kernel Hilbert spaces such as the tensor product Sobolev Hilbert space and the one induced by the Gaussian kernel. Since we are solving the regularization problem in the assumed reproducing kernel Hilbert space, we encounter the Gibbs phenomenon. That is, the solution may behave erratically at the discontinuous point. This in general is not a serious problem for classification, since we are mainly concerned with the location of the classification boundary [consisting of the discontinuous points of $\operatorname{sign}(p - 1/2)$].

We can recognize (7) with $q = 1$ as the least absolute value method used in robust regression. In least absolute value regression, the target function is the median $\operatorname{med}(Y \mid \mathbf{X} = \mathbf{x})$, and we can see in our case $\operatorname{med}(Y \mid \mathbf{X} = \mathbf{x}) = \operatorname{sign}[p(\mathbf{x}) - 1/2]$.

The computational complexity of SVM with $q = 1$ is less than those of SVM with $q > 1$ and the traditional nonparametric logistic regression. See Kaufman (1999). It remains effective when $p(\mathbf{x})$ is close or equal to 0 or 1. One special property of SVM with $q = 1$ is that it magnifies the contrast between the two sides of the classification boundary: on one side, the value of the classification function is close to 1; on the other side, it is close to -1. This is different from the SVM with $q = 2$, for which the value of f_q is very close on the two sides of the boundary.

It is much harder to derive theoretic results similar to Theorems 2.1 and 2.2 for SVMs with $q = 1$. One reason is that $(1 - yf)_+$ is not differentiable. The other reason is that the target function $\operatorname{sign}(p - 1/2)$ is not in the assumed reproducing kernel Hilbert space.

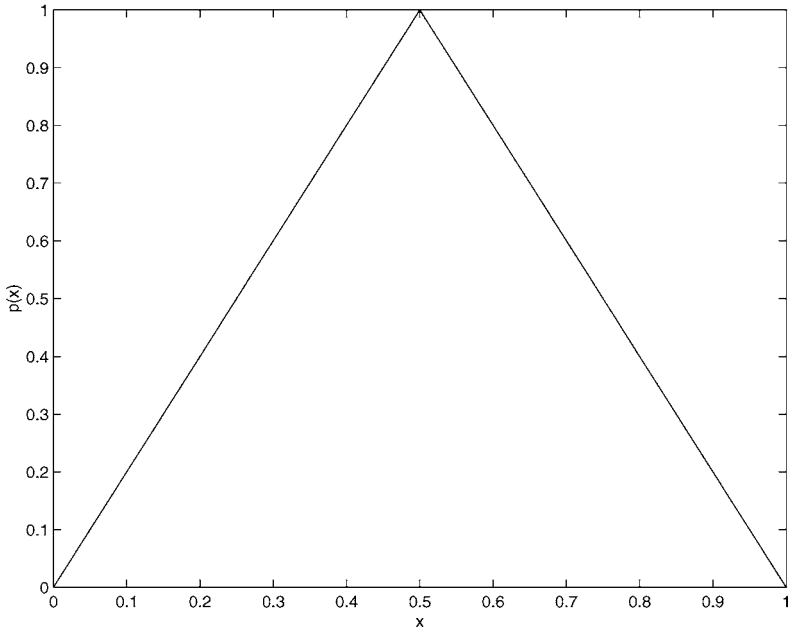

Figure 1. The underlying conditional probability function $p(x) = P\{Y = 1 \mid X = x\}$ in our simulation. The function $\text{sign}[p(x) - 1/2]$ is 1, for $0.25 < x < 0.75$; -1, otherwise.

Here we will use a simple simulation to illustrate how, with appropriately chosen tuning parameter λ, SVM with $q = 1$ approaches the target function $\text{sign}(p - 1/2)$.

For easy visualization, we will conduct the simulation in one dimension. We take n equidistant points on the interval $[0, 1]$. That is, $x_i = i/(n-1)$, $i = 0, 1, \ldots, n-1$. Let $p(x) = \Pr(Y = 1 \mid X = x) = 1 - |1 - 2x|$, and randomly generate y_i to be 1 or -1 with probability $p(x_i)$ and $1 - p(x_i)$. The picture of $p(x)$ is given in figure 1. It is easy to see that $\text{sign}[p(x) - 1/2] = 1$, $x \in (0.25, 0.75)$; -1, otherwise.

We will first consider the reproducing kernel Hilbert space $H^m([0, 1])$. The minimizer of (6) with $q = 1$ is known to have the form

$$f(\cdot) = \sum_{i=1}^{n} c_i K(\cdot, x_i) + b$$

where K is the reproducing kernel of $H_0^m([0, 1])$:

$$K(s, t) = k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s - t|),$$

where $k_1(\cdot) = \cdot - 0.5$, $k_2 = (k_1^2 - 1/12)/2$, and $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$.

Letting $\mathbf{e} = (1, \ldots, 1)'$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$, $\mathbf{c} = (c_1, c_2, \ldots, c_n)'$, and with some abuse of notation, letting $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))'$ and K now be the $n \times n$ matrix

with ijth entry $K(x_i, x_j)$, we have

$$\mathbf{f} = K\mathbf{c} + \mathbf{e}b$$

and the regularization problem (6) becomes: find $(\mathbf{c}, b)$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + \lambda \mathbf{c}' K\mathbf{c}.$$

We solve the above problem by considering its dual problem. Let Y be the $n \times n$ diagonal matrix with y_i in the iith position, and let $H = \frac{1}{2n\lambda} YKY$. The dual problem has the form

$$\max L = -\frac{1}{2}\alpha' H\alpha + \mathbf{e}'\alpha$$

subject to $0 \le \alpha_i \le 1$, $i = 1, 2, \ldots, n$, and $\mathbf{y}'\alpha = 0$. Here $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)'$. Once we get the α's, we get $\mathbf{c}$'s by $\mathbf{c} = \frac{1}{2n\lambda} Y\alpha$, and b can be computed robustly by

$$b = [\mathbf{e}' A(I - A)(\mathbf{y} - K\mathbf{c})]/[\alpha'(\mathbf{e} - \alpha)].$$

as long as there exists an i for which $0 < \alpha_i < 1$. Here A is the $n \times n$ diagonal matrix with α_i in the iith position.

The choice of the smoothing parameter λ is important. Wahba et al. (2000) proposed finding that value of λ so that the solution f_λ of (6) minimizes GCKL $E[(1 - Yf_\lambda(X))_+]$. By Lemma 3.1, using the λ that minimizes GCKL of f_λ in a sense reassures that the chosen f_λ is close to $\text{sign}(p - 1/2)$. Also, heuristically, for such λ that f_λ is close to $\text{sign}(p - 1/2)$, we can see from (1) that GCKL of f_λ is close to two times the expected misclassification rate of f_λ, therefore the λ that minimizes the GCKL of f_λ should be close to a local minimum point of the expected misclassification rate, though this local minimum may not be the global minimum.

An approximant of the GCKL for f_λ is

$$\frac{1}{n}\sum_{i=1}^{n}[p(x_i)(1 - f_\lambda(x_i))_+ + (1 - p(x_i))(1 + f_\lambda(x_i))_+]. \tag{8}$$

In our simulation here, we can calculate GCKL or (8) directly for any f_λ, since we know what $p(x)$ is. In reality, we do not know the true $p(x)$, hence we can not calculate (8) directly, but we can always estimate (8) with a test data set.

We run the simulation for $n = 33, 65, 129, 257$. In each case the smoothing parameter λ is chosen so that GCKL for f_λ is minimized. The result is shown in figure 2.

To illustrate how the smoothing parameter influences the solution, we give the solutions to (6) in the case $n = 257$ with smoothing parameters λ such that $n\lambda = 2^{-j}$, $j = 1, 2, \ldots, 25$. The results are shown in figures 3 and 4. We can see in figure 3 that the minimizer of GCKL coincides with a local minimum point of the expected misclassification rate. This local

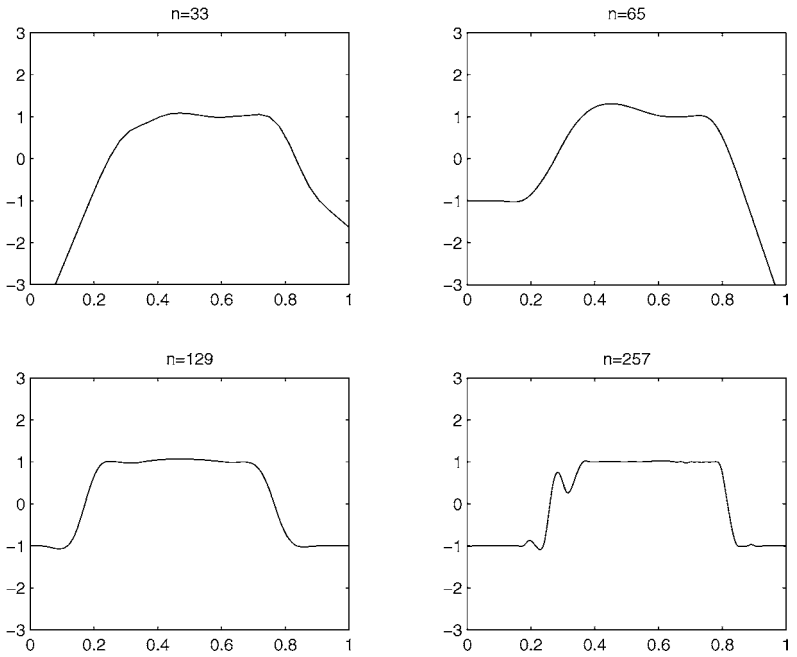

Figure 2. The solutions to the SVM regularization problems with $q = 1$ and the Sobolev Hilbert space kernel for samples of size 33, 65, 129, 257. The tuning parameter λ is chosen to minimize GCKL in each case.

minimum of the expected misclassification rate is not the global minimum, but the value of the local minimum is close to the value of the global minimum. It is often the case in our simulations that the expected misclassification rate fluctuates much more than GCKL. This is easy to understand since the expected misclassification rate depends only on the the points where the estimate crosses the x-axis, which is usually just a few points, whereas GCKL depends on almost the whole function estimate. We see in figure 4 the solution to the SVM regularization problem with $q = 1$ is close to $\text{sign}[p(x) - 1/2]$ when GCKL in figure 3 is close to the minimum.

The same simulation is run with Gaussian kernel:

$$K(s,t) = \exp\left[-\frac{(s-t)^2}{2\sigma^2}\right].$$

For Gaussian kernel, there is an additional tuning parameter σ. We use GCKL to find a good choice of λ and σ jointly. The minimum of GCKL is searched on a mesh of $(\log_2(n\lambda), \sigma)$. The relevant results are shown in figures 5–8. Figure 5 shows, in the cases when the sample size is $n = 33, 65, 129, 257$, the solutions to the regularization problem when $(\log_2(n\lambda), \sigma)$ are chosen to minimize GCKL. Figures 6–8 are for the case $n = 257$. For this sample the minimum of GCKL is found at $\log_2(n\lambda) = -9$ and $\sigma = 0.09$. Again we see the solution to the SVM regularization problem is close to $\text{sign}[p(x) - 1/2]$ when $(\log_2(n\lambda), \sigma)$ are close to the minimizer of GCKL.

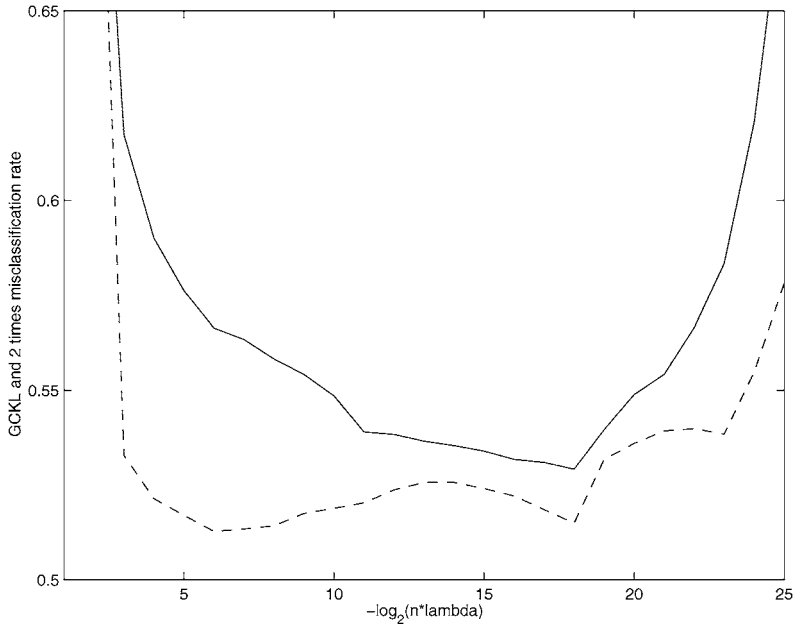

Figure 3. GCKL (solid line) and two times misclassification rate (dashed line) of f_λ with varying λ for a fixed sample with $n = 257$, where f_λ is the solution to the SVM regularization problem with $q = 1$ and the Sobolev Hilbert space kernel. Notice the x-axis is $-\log_2(n\lambda)$. (Larger values of λ correspond to the points on the left.)

4. Conclusion

We studied the relation between SVMs and the Bayes rule of classification. Lemmas 2.1 and 3.1 hold for any kernel, and they identify the asymptotic target of SVMs. It is shown that classifying with the sign of these target functions is equivalent to the Bayes rule. The question of how well SVMs approach their corresponding asymptotic targets is studied for SVMs with quadratic loss (or higher order loss) and spline kernels. Theorem 2.2 gives the rate of convergence of such SVMs approaching their corresponding target functions. Intuitively, the theory of regularization methods and Lemmas 2.1 and 3.1 suggests that SVMs with $q = 1$, and SVMs with other kernels, should also approach their target functions. However, results similar to Theorem 2.2 can be very hard to obtain for SVMs with $q = 1$. We used some simulations to demonstrate how SVMs with $q = 1$ approaches the Bayes rule. The simulation is based on a very simple problem, and only provide some illustration. Further study is definitely needed. Lin (2000b) derived some theoretical results for SVMs with $q = 1$ and the first order spline kernel.

The insight we obtain from the relation between SVMs and the Bayes rule has practical importance. For example, in practice it is often the case that the costs of false positive and false negative are different. It is also possible that the fraction of members of the classes in the training set is different than those in the general population (sampling bias). In such situations the Bayes rule that minimizes the expected misclassification cost can be

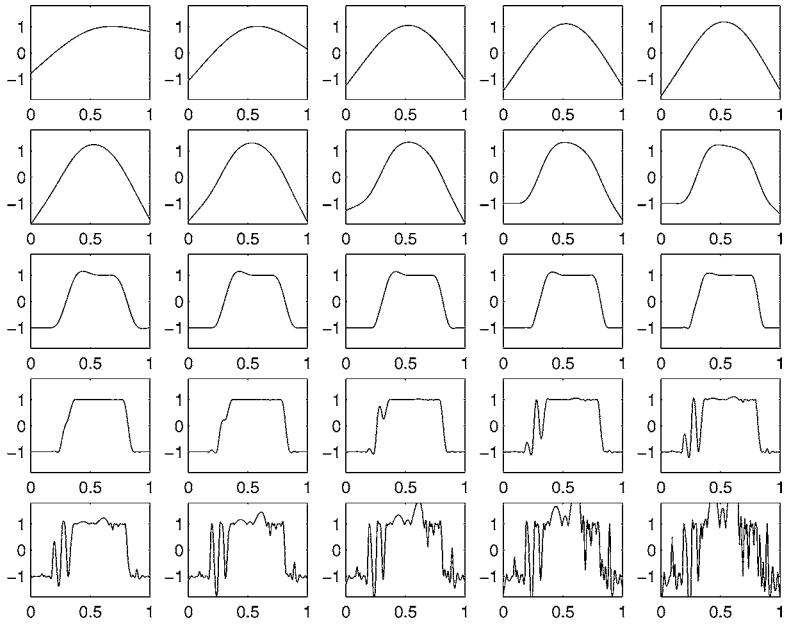

Figure 4. For the same sample as in figure 3, the solutions to the SVM regularization problem with $q = 1$ and the Sobolev Hilbert space kernel for $n\lambda = 2^{-1}, 2^{-2}, \ldots, 2^{-25}$. The first row corresponds to $n\lambda = 2^{-1}, 2^{-2}, \ldots, 2^{-5}$, from left to right. The second row corresponds to $n\lambda = 2^{-6}, 2^{-7}, \ldots, 2^{-10}$; the third row corresponds to $n\lambda = 2^{-11}, 2^{-12}, \ldots, 2^{-15}$; and so on. We see the solution is close to $\text{sign}[p(x) - 1/2]$ when GCKL in figure 3 is close to the minimum.

expressed as $\text{sign}[p(\mathbf{x}) - c]$, where $c \in (0, 1)$ is not equal to $1/2$. Hence we need to modify the formulation of SVMs accordingly for them to perform optimally in this situation. Lin et al. (2002) contains some extension of the SVM to such nonstandard situations.

Appendix

Proof of Lemma 2.1: Notice

$$E[(1 - Yf(\mathbf{X}))_+]^q = E\{E\{[(1 - Yf(\mathbf{X}))_+]^q \mid \mathbf{X}\}\}$$

We can minimize $E[(1 - Yf(\mathbf{X}))_+]^q$ by minimizing $E\{[(1 - Yf(\mathbf{X}))_+]^q \mid \mathbf{X} = \mathbf{x}\}$ for every fixed $\mathbf{x}$.

For any fixed $\mathbf{x}$, we have $E\{[(1 - Yf(\mathbf{X}))_+]^q \mid \mathbf{X} = \mathbf{x}\} = p(\mathbf{x})[(1 - f(\mathbf{x}))_+]^q + (1 - p(\mathbf{x}))[(1 + f(\mathbf{x}))_+]^q$. Let us search for $\bar{w}$ that minimizes $A(w) = p(\mathbf{x})[(1 - w)_+]^q + (1 - p(\mathbf{x}))[(1 + w)_+]^q$.

First notice that the minimizer of $A(w)$ must be in $[-1, 1]$. For any w outside $[-1, 1]$, let $w' = \text{sign}(w)$, then w' is in $[-1, 1]$ and it is easy to check $A(w') < A(w)$. So we can restrict our search in $[-1, 1]$.

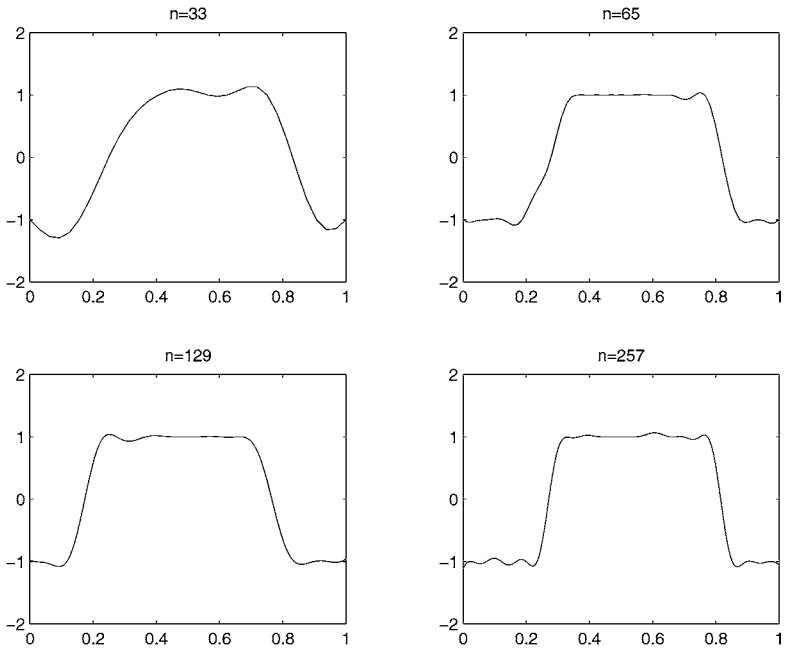

Figure 5. The solutions to the SVM regularization problems with $q = 1$ and the Gaussian kernel for samples of size 33, 65, 129, 257. The tuning parameter λ and σ are chosen to minimize GCKL in each case.

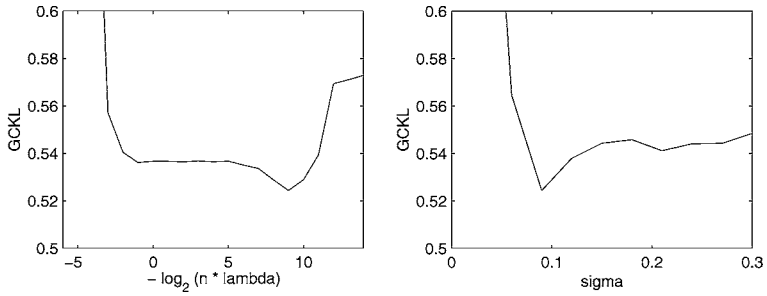

Figure 6. GCKL of $f_{\sigma,\lambda}$ with varying σ and λ for a fixed sample with $n = 257$, where $f_{\sigma,\lambda}$ is the solution to the SVM regularization problem with $q = 1$ and the Gaussian kernel $\exp[-\frac{(s-t)^2}{2\sigma^2}]$. Upper left: GCKL of $f_{\sigma,\lambda}$ with σ fixed at 0.09. Notice the x-axis is $-\log_2(n\lambda)$. Lower right: GCKL of $f_{\sigma,\lambda}$ with $n\lambda$ fixed at 2^{-9}.

For $w \in [-1, 1]$, $A(w) = p(\mathbf{x})(1-w)^q + [1-p(\mathbf{x})](1+w)^q$. By taking derivative with respect to w, we get $\bar{w} = [(p(\mathbf{x}))^{\frac{1}{q-1}} - (1-p(\mathbf{x}))^{\frac{1}{q-1}}]/[(p(\mathbf{x}))^{\frac{1}{q-1}} + (1-p(\mathbf{x}))^{\frac{1}{q-1}}]$. Therefore the minimizer of $E[(1-Yf(\mathbf{X}))_+]^q$ is

$$f_q(\mathbf{x}) = \left[(p(\mathbf{x}))^{\frac{1}{q-1}} - (1-p(\mathbf{x}))^{\frac{1}{q-1}}\right] \Big/ \left[(p(\mathbf{x}))^{\frac{1}{q-1}} + (1-p(\mathbf{x}))^{\frac{1}{q-1}}\right]$$

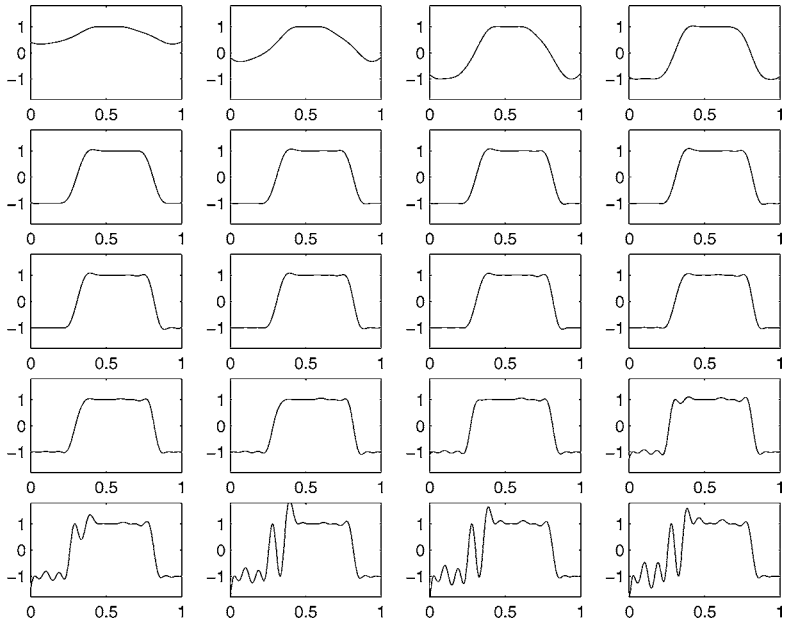

Figure 7. For the same sample as in figure 6, with σ fixed at 0.09, the solutions to the SVM regularization problem with $q = 1$ and Gaussian kernel for $n\lambda = 2^5, 2^4, \ldots, 2^{-14}$. The first row corresponds to $n\lambda = 2^5, 2^4, 2^3, 2^2$; the second row corresponds to $n\lambda = 2^1, 2^0, 2^{-1}, 2^{-2}$; and so on. We see the solution is close to $\text{sign}[p(x) - 1/2]$ when GCKL in figure 6 (upper left picture) is close to the minimum.

The same line of argument shows that $f_q(\mathbf{x})$ is also the minimizer of $E|Y - f(\mathbf{X})|^q$. The proof of the rest of Lemma 2.1 is straight forward. □

Proof of Lemma 3.1: Follow the same line of proof as that of Lemma 2.1. □

Proof of Theorem 2.1: Consider the problem of estimating f_0 with iid sample from the model

$$E(Y \mid \mathbf{X}) = f_0(\mathbf{X}), \quad \text{Var}(Y \mid \mathbf{X}) = \sigma^2.$$

Lin (2000a) studied the properties of the estimator obtained by minimizing (7) with $q = 2$.

In our present model we have

$$E(Y \mid \mathbf{X}) = 2p(\mathbf{X}) - 1, \quad \text{Var}(Y \mid \mathbf{X}) = 4p(\mathbf{X})(1 - p(\mathbf{X})) \leq 1.$$

Using the same argument as that employed in the proof of Theorem 4.1 in Lin (2000a) with $l_\infty(f) = E[Y - f(\mathbf{X})]^2$, which is $E[(f(\mathbf{X}) - (2p(\mathbf{X}) - 1))^2 + 4p(\mathbf{X})(1 - p(\mathbf{X}))]$ in our situation instead of $E[(f(\mathbf{X}) - (f_0(\mathbf{X})))^2] + \sigma^2$ as in Lin (2000a), everything goes through exactly as in the proof of Theorem 4.1 in Lin (2000a) with f_* in place of $\hat{f}$, $2p-1$ in

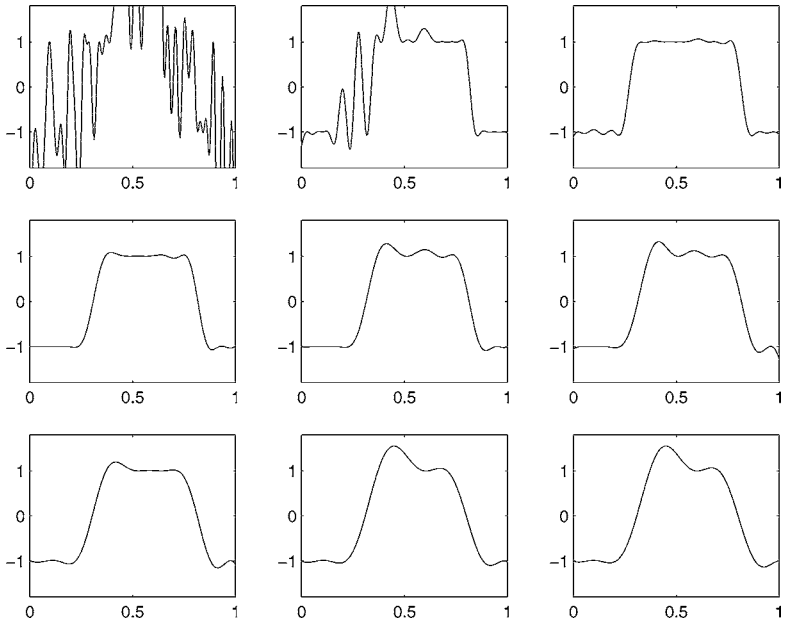

Figure 8. For the same sample as in figure 6, with $n\lambda$ fixed at 2^{-9}, the solutions to the SVM regularization problem with $q = 1$ and Gaussian kernel for $\sigma = 0.03, 0.06, \ldots, 0.27$. The first row corresponds to $\sigma = 0.03, 0.06, 0.09$; the second row corresponds to $\sigma = 0.12, 0.15, 0.18$; and so on. We see the solution is close to $\text{sign}[p(x) - 1/2]$ when GCKL in figure 6 (lower right picture) is close to the minimum.

place of f_0, and $f_{\mathbf{X}}$ in place of p in Lin (2000a). So we get that Theorem 4.1 in Lin (2000a) is still valid in our situation. The norm $\|\cdot\|_a$ is the norm in the space $\otimes^d H^{ma}([0, 1])$. (If ma is not an integer, then $H^{ma}([0, 1])$ is a fractional order Sobolev space.)

Now set $b = \frac{1}{2m} + \frac{\epsilon}{2}$ in Theorem 4.1 of Lin (2000a). Setting $a = 0$ we get the first expression in our theorem. Setting $a = b$, using Theorem 4.1 and Lemma 2.1 in Lin (2000a), we get the second expression in our theorem. □

Proof of Theorem 2.2: Since $0 < p(\mathbf{x}) < 1$, $\forall \mathbf{x} \in [0, 1]^d$, and $p(\mathbf{x})$ is continuous, we have that $\sup_{\mathbf{x}\in[0,1]^d} |2p - 1| < 1$. Also, by Theorem 2.1, under our condition, we have $\sup_{\mathbf{x}\in[0,1]^d} |f^* - (2p - 1)| = o_p(1)$. Hence we can take n large enough so that the event $\sup_{\mathbf{x}\in[0,1]^d} |f^*| < 1$ occurs with probability arbitrarily close to one. For the remainder of the proof we restrict attention to this event.

Consider the set $\Omega = \{f \in \otimes^d H^m : \sup_{\mathbf{x}\in[0,1]^d} |f(\mathbf{x})| < 1\}$. By Lemma 2.1 of Lin (2000a), we see that

$$\sup_{\mathbf{x}\in[0,1]^d} |f(\mathbf{x})| \le C\|f\|_{\otimes^d H^m}$$

for any $f \in \otimes^d H^m$. Here C is a constant independent of f. Hence it is easy to check that Ω is an open set in $\otimes^d H^m$. We have $f^* \in \Omega$. Since f^* is the minimizer of (7), by Lemma 1.1,

we have that f^* is also the minimizer of (6) over Ω. Hence f^* is a local minimum point of (6). Since (6) is a convex functional of f, f^* is also a global minimum point of (6). Hence $f_* = f^*$, and the results now follows from Theorem 2.1. □

Acknowledgment

This work was partly supported by Wisconsin Alumni Research Foundation. The author wishes to thank one referee for the careful review and helpful comments.

References

Boser, B.E., Guyon, I.M., and Vapnik, V.N. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, D. Haussler (Ed.). Pittsburgh, PA: ACM Press.

Burges, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121–167.

Cortes, C. and Vapnik, V.N. 1995. Support vector networks. Machine Learning, 20:273–297.

Cox, D.D. and O'Sullivan, F. 1990. Asymptotic analysis of penalized likelihood and related estimates. The Annals of Statistics, 18(4):1676–1695.

Evgeniou, T., Pontil, M., and Poggio, T. 1999. A unified framework for regularization networks and support vector machines. Technical Report, M.I.T. Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences.

Friedman, J.H. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery, 1(1):55–77.

Kaufman, L. 1999. Solving the quadratic programming problem arising in support vector classification. In Advances in Kernel Methods—Support Vector Learning, B. Schölkopf, C.J.C. Burges, and A.J. Smola (Eds.). Cambridge, MA: MIT Press, pp. 147–168.

Lin, Y. 1998. Tensor product space ANOVA models in high dimensional function estimation. Ph.D. Dissertation, University of Pennsylvania.

Lin, Y. 2000a. Tensor product space ANOVA models. The Annals of Statistics, 28(3):734–755.

Lin, Y. 2000b. On the support vector machine. Technical Report 1029, Department of Statistics, University of Wisconsin, Madison.

Lin, Y., Lee, Y., and Wahba, G. 2002. Support vector machines for classification in nonstandard situations. Machine Learning, 46:191–202.

Shawe-Taylor, J. and Cristianini, N. 1998. Robust bounds on the generalization from the margin distribution. Neuro COLT Technical Report TR-1998-029.

Vapnik, V.N. 1995. The Nature of Statistical Learning Theory. New York: Springer Verlag.

Wahba, G. 1990. Spline Models for Observational Data. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Wahba, G. 1999. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In Advances in Kernel Methods—Support Vector Learning, B. Scholkopf, C.J.C. Burges, and A.J. Smola (Eds.). Cambridge, MA: MIT Press.

Wahba, G., Lin, Y., and Zhang, H. 2000. *GACV* for support vector machines, or, another way to look at margin-like quantities. In Advances in Large Margin Classifiers, A.J. Smola, P. Bartlett, B. Scholkopf, and D. Schurmans (Eds.). Cambridge, MA and London, England: MIT Press.