DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1014

November 1, 1999

SUPPORT VECTOR MACHINES AND THE BAYES RULE IN CLASSIFICATION

by

Yi Lin

Support Vector Machines and the Bayes Rule in Classification

Yi Lin
University of Wisconsin, Madison

November 1, 1999

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 SVMs implement the Bayes rule approximately by targeting at some interesting classification functions. This helps understand the success of SVMs in many classification studies, and makes it easier to compare SVMs and traditional statistical 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, $i = 1, 2, ..., n$ in the training data set, we observe an explanatory vector $\underline{x}_i \in R^d$, 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 $P(\underline{x}, y)$, or equivalently, they are independent random realizations of the random pair $(\underline{X}, Y)$ that has cumulative probability distribution $P(\underline{x}, y)$. The task of classification is to derive from the training set a good classification rule, so that once we are given the $\underline{x}$ 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 $P(\underline{x}, y)$, 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 we have 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 η can be written as

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

Here $(\cdot)_+$ is a function such that τ_+ is τ, if $\tau > 0$; and is 0, otherwise. For a general real function $f : R^d \rightarrow R$, we call $E[1 - Yf(\underline{X})]_+$ the generalized comparative Kullback Leibler (GCKL) measure. See Wahba, Lin, and Zhang (1999).

Let

$$p(\underline{x}) = Pr\{Y = 1|\underline{X} = \underline{x}\}$$

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

Since we do not know $P(\underline{x}, y)$ 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 apply the Bayes rule approximately. Traditional statistical methods try to estimate $[p(\underline{x}) - 1/2]$ (or the log odds ratio $g(\underline{x})$) from the training data, and then approximate the Bayes rule with $sign[\hat{p}(\underline{x}) - 1/2]$ (or $sign[\hat{g}(\underline{x})]$). Here $\hat{p}(\underline{x})$ ($\hat{g}(\underline{x})$) is the estimate of $p(\underline{x})$ ($g(\underline{x})$). Friedman (1996) discussed how the bias and variance components of the estimation error affects classification error when the estimate is used in a classification rule.

SVMs are motivated by the geometric interpretation of maximizing the margin, and are characterized by the use of reproducing kernel (called kernel in SVM literature. Not to be confused with the kernel estimators used in nonparametric statistics). For a tutorial on SVMs for classification, see Burges (1998). It has been shown that SVM methodology can be cast as a variational/regularization problem in a reproducing kernel Hilbert space (RKHS). See Wahba (1990), Girosi (1998), Poggio and Girosi (1998). A reproducing kernel over R^d is a positive definite function on $R^d \bigotimes R^d$. Let H_K be a RKHS with reproducing kernel $K(\underline{s}, \underline{t})$, $\underline{s}, \underline{t} \in R^d$. The SVM using reproducing kernel K first minimizes

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

over all the functions of the form $f(\underline{x}) = h(\underline{x}) + const$, and $h \in H_K$. Here q is a positive integer, and $f_i = f(\underline{x}_i)$. Once the minimizer $\tilde{f}$ is found, then the SVM classification rule is $sign[\tilde{f}(\underline{x})]$. A variety of reproducing kernels have been used successfully in practical applications, including polynomial kernels, Gaussian kernels, and Sobolev Hilbert space kernels. The RKHS' for the latter two types of reproducing kernels are of infinite dimension. For a review on the Sobolev Hilbert space kernels, see Wahba (1990). The theory of RKHS ensures that the minimizer of (2) lies in a finite dimensional space, even when the minimization is carried out in an infinite dimensional RKHS. See Wahba (1990). Hence for any positive integer q, the minimization problem (2) becomes a convex programming problem in a finite dimensional space. See Wahba, Lin and Zhang (1999). For $q = 1$ and $q = 2$ it is also a quadratic programming problem. For $q = 1$, the Wolfe dual problem of the minimization is of a particularly simple form, and this is why most SVMs choose to use $q = 1$.

Remark 1.1 *If $\{1\} \subset H_K$, the regularization problem (2) 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 margin achieved on the training set by the classifier. 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 approximately implement the Bayes rule. 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

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - f_i|^q + \lambda ||h||^2_{H_K} \tag{3}$$

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

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

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

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 $q > 1$

Regularization problems similar to (2) and (3) have long been studied in statistics literature, see Wahba (1990) and the reference therein. Cox and O'Sullivan (1990) provided a general framework for studying regularization methods. As in (2) and (3), the method of regularization has two components: a data fit functional component and a regularization penalty component. The data fit component usually approaches a limiting functional as $n \to \infty$. For example, in the SVM situation, this limiting functional is easily seen to be $E[(1-Yf(\underline{X}))_+]^q$. The corresponding limiting functional for (3) is $E|Y-f(\underline{X})|^q$. In general the limiting functional can be used to identify the target function. Under the assumption that the target function is in the RKHS under consideration and certain other general regularity conditions, the solution of the regularization problem approaches the target function as $n \to \infty$. The following lemma identifies the target function for SVM and (3) with $q > 1$ (From now on all proofs are given in the appendix):

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

$$f_q(\underline{x}) = [(p(\underline{x}))^{\frac{1}{q-1}} - (1-p(\underline{x}))^{\frac{1}{q-1}}]/[(p(\underline{x}))^{\frac{1}{q-1}} + (1-p(\underline{x}))^{\frac{1}{q-1}}]$$

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

Thus under certain conditions, the minimizer of (2) $\hat{f}_q$ approaches f_q as $n \to \infty$, and SVM classifier $sign[\hat{f}_q(\underline{x})]$ approximates the Bayes rule $sign[f_q(\underline{x})]$. 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 RKHS, and illustrate how fast $\hat{f}_q$ 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

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

with a norm equivalent to

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

where $M_\nu f = \int_0^1 f^{(\nu)}(u)du$, $\nu = 0, 1, ..., m-1$. It is typical 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 RKHS, and the reproducing kernel of this space is derived in Wahba (1990), chapter 10. 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 $\bigotimes^d H^m$ be the tensor product space of d $H^m([0,1])$ spaces. Then $\bigotimes^d H^m$ is a Hilbert space of functions on $[0,1]^d$, and it can be identified with the Hilbert space of functions

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

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

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

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

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

Now consider the regularization problems (2) and (3) with $H_K = \bigotimes^d H^m$ and $q = 2$. We denote the solution to (2) by f_*, and the solution to (3) by f^*.

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

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

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

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

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

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

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, Lin and Zhang (1999) considered choosing λ to minimize the estimated generalized comparative Kullback Leibler measure.*

Remark 2.2 *The condition $0 < p(\underline{x}) < 1$ 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 *Under the $||.||_{H^m([0,1])}$ norm, we have*

$$H^m([0, 1]) = H_0^m([0, 1]) \oplus \{1\}$$

where $\{1\}$ is the space of scalars. $H_0^m([0, 1])$ is the subspace (orthogonal to $\{1\}$) consisting of all the functions in $H^m([0, 1])$ satisfying $\int_0^1 f(u)du = 0$. Therefore we have

$$\bigotimes^d H^m = \bigotimes^d [H_0^m([0, 1]) \oplus \{1\}]$$

Identifying the tensor product space of $\{1\}$ with any Hilbert space with that Hilbert space itself, then $\bigotimes^d H^m$ is the direct sum of all the subspaces of the form $H_0^m(x_{j_1}) \otimes H_0^m(x_{j_2}) \otimes \ldots \otimes H_0^m(x_{j_k})$, and the subspaces in this decomposition are all orthogonal to each other. Accordingly, any function in $\bigotimes^d H^m$ can be uniquely decomposed as

$$f(x_1, x_2, ..., x_d) = constant + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + ... + f_{12...d}(x_1, x_2, ..., x_d)$$

with each component in the decomposition in a different subspace. This is the functional ANOVA decomposition, and we usually call the univariate functions in the decomposition the main effects and high dimensional functions the interactions. In practice we can truncate the series by throwing away the higher order interactions to enhance interpretability. Then the RKHS we consider would be a particular subspace of $\bigotimes^d H^m$. If we do that, let v be the highest order of interaction in the model, then similar results to those stated in the theorems above hold with d replaced by v, even though we are still considering space of functions on $[0,1]^d$. For more on the functional ANOVA decomposition, see Wahba (1990) and Lin (1998b).

Remark 2.5 *In some situations we may want to use some RKHS 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 $p(\underline{x})$ is in the assumed RKHS. 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 2.6 *For $q > 2$, similar results can be obtained on how fast the minimizers of (2) and (3) approach f_q by using the framework provided in Cox and O'Sullivan (1990).*

The theorems above show that SVM with $q = 2$ solves a regularization problem to get f_*, which approaches $2p - 1$, then uses $sign[f_*(\underline{x})]$ to approximately implement the Bayes rule $sign[p(\underline{x}) - 1/2]$. Similarly, we can also consider solving (3) 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. See Kaufman (1999). Second, when $p(\underline{x})$ 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 (3) with $q = 2$ can be motivated by the fact that $E(Y|\underline{X} = \underline{x}) = 2p(\underline{x}) - 1$, and we recognize (3) with $q = 2$ as the penalized least square regression method for estimating $E(Y|\underline{X} = \underline{x})$. Intuitively, this method would not be efficient since they do not take into account the fact that $Var(Y|\underline{X} = \underline{x}) = 4p(\underline{x})[1-p(\underline{x})]$ is not a constant and is smaller at places where $p(\underline{x})$ is close to 0 or 1. By proceeding as if the variance is a constant, we are wasting some precision at regions where $p(\underline{x})$ is close to 0 or 1. However, for the purpose of classification, what concerns us most is the region where $p(\underline{x})$ 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 RKHS used in the regularization problem contains $p(\underline{x})$. It conforms to the notion that we should choose reproducing kernel so that $p(\underline{x})$ is in the corresponding RKHS. This condition can be relaxed a little, [see Cox and O'Sullivan (1990),] but $p(\underline{x})$ should at least be close to the RKHS. The space spanned by the linear functions of $\underline{x}$ may not satisfy this. This explains why linear SVM may not perform well in some cases. This is also quite intuitive: linear SVM only linearly divides the whole space into two half spaces, hence in the situations where $sign[p(\underline{x})]$ divides the space into more than two parts, or if the boundary is highly nonlinear, linear SVM can not perform well.

3 SVMs with $q = 1$

This is the most commonly used SVM. In this situation, the limiting functional of the data fit component in (2) is $E[(1 - Yf(X))_+]$. The corresponding limiting functional for (3) is $E|Y - f(\underline{X})|$. The following lemma identifies the target function for SVM and (3) with $q = 1$.

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

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 $sign(p - 1/2)$. However, this target function typically does not lie in the commonly used RKHS, (it is easy to see that $sign(p-1/2)$ is not a smooth function unless $p(\underline{x})$ 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 RHKS' 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 RKHS, 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 $sign(p - 1/2)$].

We can recognize (3) with $q = 1$ as the least absolute value method used in robust regression. In least absolute value regression, the target function is the median $med(Y|\underline{X} = \underline{x})$, and we can see in our case $med(Y|\underline{X} = \underline{x}) = sign[p(\underline{x}) - 1/2]$.

SVM with $q = 1$ is easier to compute than SVM with $q > 1$. It remains effective when $p(\underline{x})$ 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 Theorem 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 $sign(p - 1/2)$ is not in the assumed RKHS. Here we will use a simple simulation to illustrate how, with appropriately chosen tuning parameter λ, SVM with $q = 1$ approaches the target function $sign(p - 1/2)$.

For easy visualization, we will conduct the simulation in one dimension. We take n equidistant points on the interval $[0,1]$. That is, $x_i = i/(n-1)$, $i = 0,1,...,n-1$. Let $p(x) = Pr(Y = 1|X = x) = 1 - |1-2x|$, 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. All figures are given at the end of the paper. It is easy to see that $sign[p(x) - 1/2] = 1$, $x \in (0.25, 0.75)$; -1, otherwise.

We will first consider the RKHS $H^m([0,1])$. The minimizer of (2) 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 $e = (1,...,1)'$, $y = (y_1, y_2, ..., y_n)'$, $c = (c_1, c_2, ..., c_n)'$, and with some abuse of notation, letting $f = (f(x_1), f(x_2), ..., f(x_n))'$ and K now be the $n \times n$ matrix with ijth entry $K(x_i, x_j)$, we have

$$f = Kc + eb$$

and the regularization problem (2) becomes: find (c, b) to minimize

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

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

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

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

$$b = [e'A(I-A)(y - Kc)]/[\alpha'(e - \alpha)].$$

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, Lin and Zhang (1999) proposed finding that value of λ so that the solution f_λ of (2) 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 $sign(p - 1/2)$. Also, heuristically, for such λ that f_λ is close to $sign(p - 1/2)$, 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

$$\frac{1}{n}\sum_{i=1}^{n}[p(x_i)(1-f_\lambda(x_i))_+ + (1-p(x_i))(1+f_\lambda(x_i))_+]. \tag{4}$$

In our simulation here, we can calculate GCKL or (4) 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 (4) directly, but we can always estimate (4) 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 (2) in the case $n = 257$ with smoothing parameters λ such that $n\lambda = 2^{-j}$, $j = 1, 2, ..., 25$. The results are shown in Figure 3 - 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 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 $sign[p(x) - 1/2]$ when GCKL in Figure 3 is close to the minimum.

The same simulation is run with Gaussian kernel:

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

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 Figure 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. Figure 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 $sign[p(x) - 1/2]$ when $(\log_2(n\lambda), \sigma)$ are close to the minimizer of GCKL.

Appendix

Proof of Lemma 2.1: Notice

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

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

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

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

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

$$f_q(\underline{x}) = [(p(\underline{x}))^{\frac{1}{q-1}} - (1-p(\underline{x}))^{\frac{1}{q-1}}]/[(p(\underline{x}))^{\frac{1}{q-1}} + (1-p(\underline{x}))^{\frac{1}{q-1}}]$$

The same line of argument shows that $f_q(\underline{x})$ is also the minimizer of $E|Y - f(\underline{X})|^q$.

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

$$E(Y|\underline{X}) = f_0(\underline{X}), \qquad Var(Y|\underline{X}) = \sigma^2.$$

Lin (1998b) studied the properties of the estimator obtained by minimizing (3) with $q = 2$.

In our present model we have

$$E(Y|\underline{X}) = 2p(\underline{X}) - 1, \qquad Var(Y|\underline{X}) = 4p(\underline{X})(1-p(\underline{X})) \leq 1.$$

Using the same argument as that employed in the proof of Theorem 4.1 in Lin (1998b) with $l_\infty(f) = E[Y - f(\underline{X})]^2$, which is $E[(f(\underline{X}) - (2p(\underline{X})-1))^2 + 4p(\underline{X})(1-p(\underline{X}))]$ in our situation instead of $E[(f(\underline{X}) - (f_0(\underline{X})))^2] + \sigma^2$ as in Lin (1998b), everything goes through exactly as in the proof of Theorem 4.1 in Lin (1998b) with f_* in place of $\hat{f}$, $2p-1$ in place of f_0, and $f_{\underline{X}}$ in place of p in Lin (1998b). So we get that Theorem 4.1 in Lin (1998b) is still valid in our situation. The norm $||.||_a$ is the norm in the space $\bigotimes^d H^{ma}([0,1])$. (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 (1998b). Setting $a = 0$ we get the first expression in our theorem. Setting $a = b$, using Theorem 4.1 and Lemma 2.1 in Lin (1998b), we get the second expression in our theorem.

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

Consider the set $\Omega = \{f \in \bigotimes^d H^m : \quad \sup_{\underline{x}\in[0,1]^d} |f(\underline{x})| < 1\}$. By Lemma 2.1 of Lin (1998b), we see that

$$\sup_{\underline{x}\in[0,1]^d} |f(\underline{x})| \leq C||f||_{\bigotimes^d H^m}$$

for any $f \in \bigotimes^d H^m$. Here C is a constant independent of f. Hence it is easy to check that Ω is an open set in $\bigotimes^d H^m$. We have $f^* \in \Omega$. Since f^* is the minimizer of (3), by Lemma 1.1, we have that f^* is also the minimizer of (2) over Ω. Hence f^* is a local minimum point of (2). Since (2) is a convex functional of f, f^* is also a global minimum point of (2). Hence $f_* = f^*$, and the results now follows from Theorem 2.1.

References

[1] Burges, C.J.C., “A Tutorial on Support Vector Machines for Pattern Recognition”, in *Data Mining and Knowledge Discovery*, Vol. 2, Number 2, p. 121-167, 1998 1998.

[2] Cox, D.D., and O’Sullivan F., “Asymptotic Analysis of Penalized Likelihood and Related Estimates”, *The Annals of Statistics*, Vol. 18, number 4, 1676-1695, 1990.

[3] Girosi, F., “An Equivalence between Sparse Approximation and Support Vector Machines”, *Neural Computation*, Vol. 10(6), 1455-1480, 1998.

[4] Kaufman, L., “Solving the quadratic programming problem arising in support vector classification”, In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, 147-168, 1999.

[5] Lin, Y., *Tensor Product Space ANOVA Models in High Dimensional Function Estimation*, Ph.D. Dissertation, 1998a.

[6] Lin, Y., “Tensor Product Space ANOVA Models”, submitted to *The Annals of Statistics*, 1998b.

[7] Poggio, T., and Girosi, F., “A Sparse Representation for Function Approximation”, *Neural Computation*, Vol. 10, 1445-1454, 1998.

[8] Shawe-Taylor, J., and Cristianini, N., “Robust Bounds on the Generalization from the Margin Distribution”. Neuro COLT Technical Report TR-1998-029, 1998.

[9] Vapnik, V., *The Nature of Statistical Learning Theory*, 1995.

[10] Wahba, G., *Spline Models for Observational Data*, 1990.

[11] Wahba, G., Lin, Y., and Zhang, H., “GACV for Support Vector Machines, or , Another Way to Look at Margin-like Quantities”, to appear in *Advances in Large Margin Classifiers*, 1999.

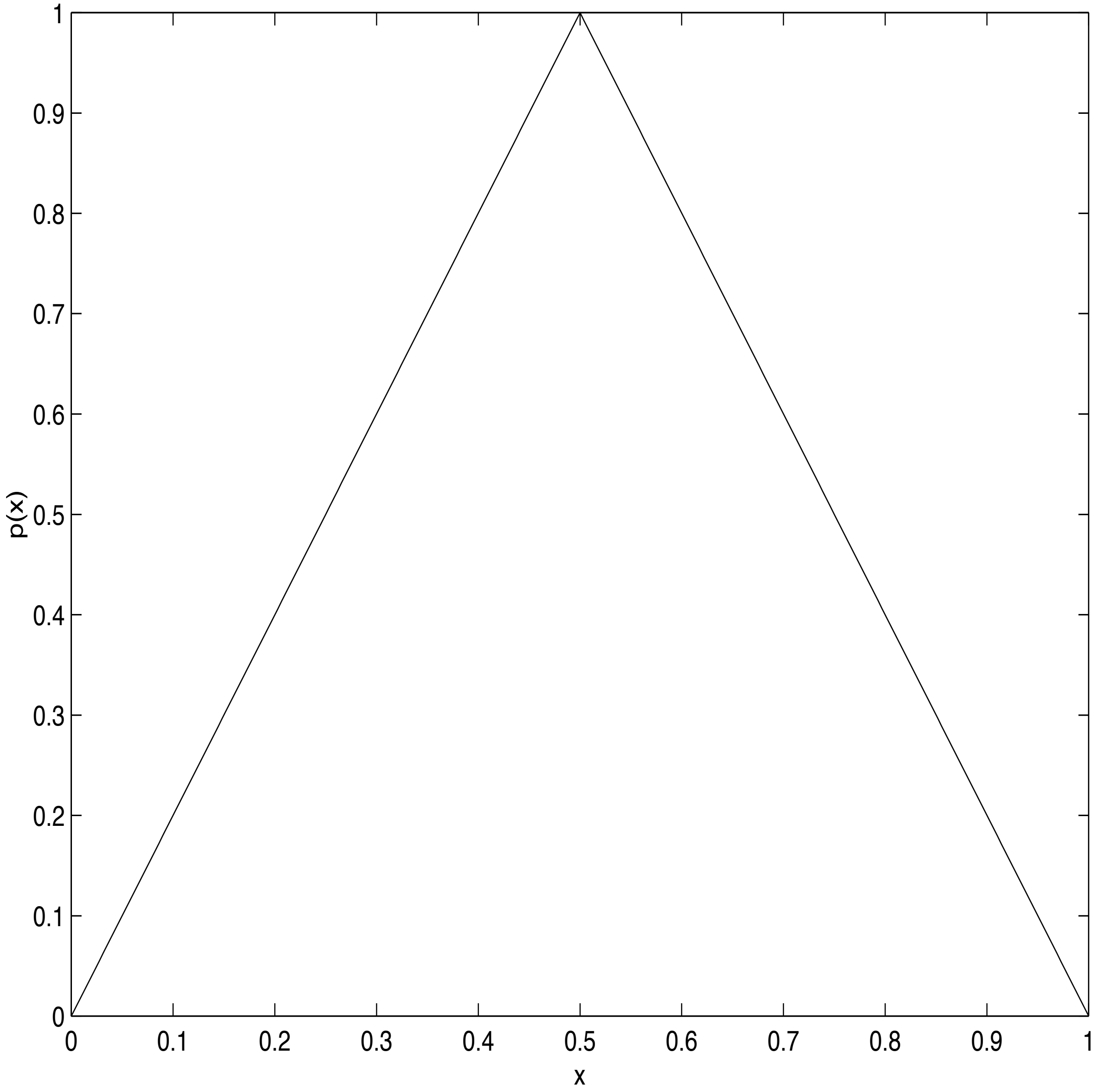

Figure 1: The underlying conditional probability function $p(x) = P\{Y = 1|X = x\}$ in our simulation. The function $sign[p(x) - 1/2]$ is 1, for $0.25 < x < 0.75$; -1, otherwise.

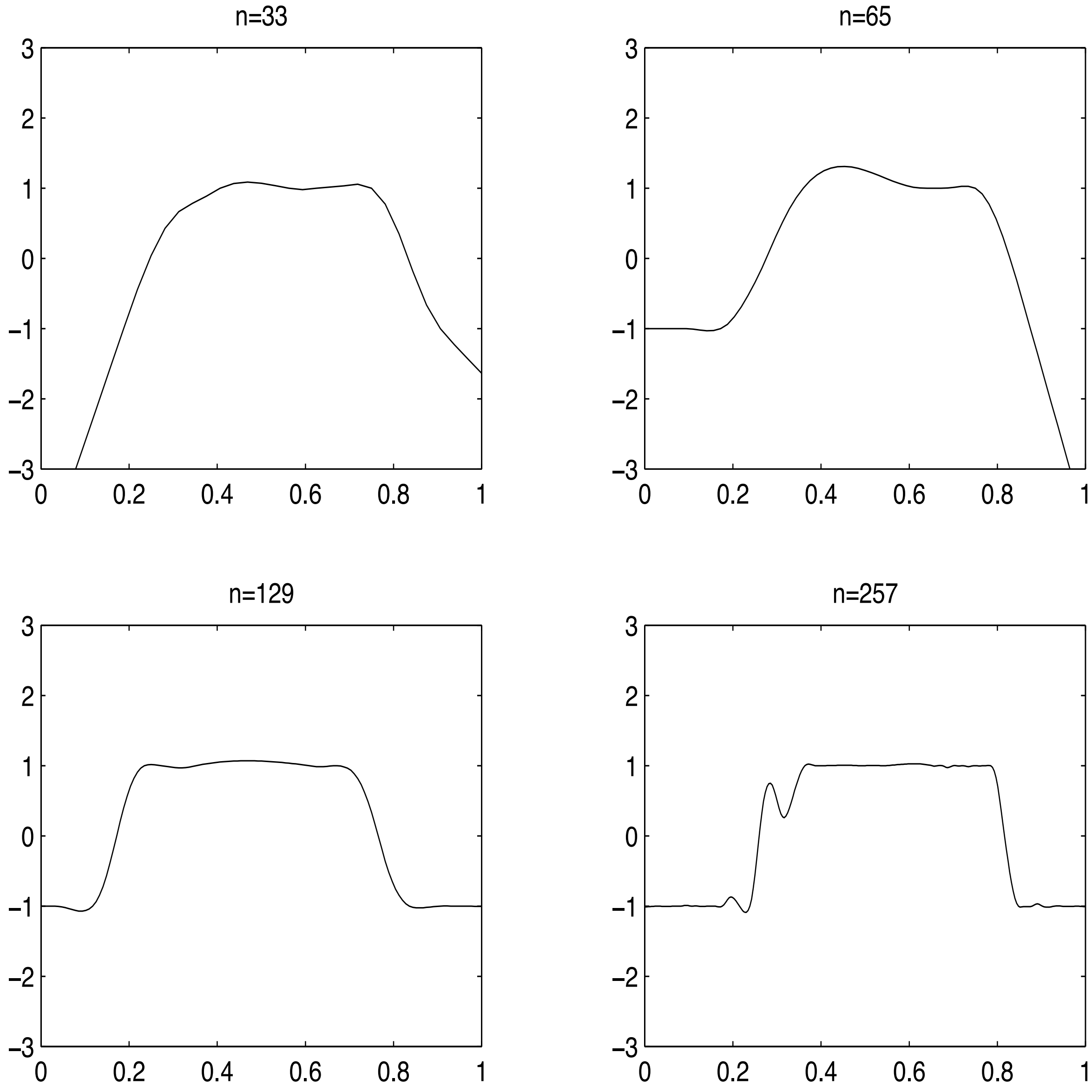

Figure 2: The solutions to the SVM regularization problems with $q = 1$ and the Sobolev Hilbert space kernel for samples of size 33, 65, 129, 257. The tuning parameter λ is chosen to minimize GCKL in each case.

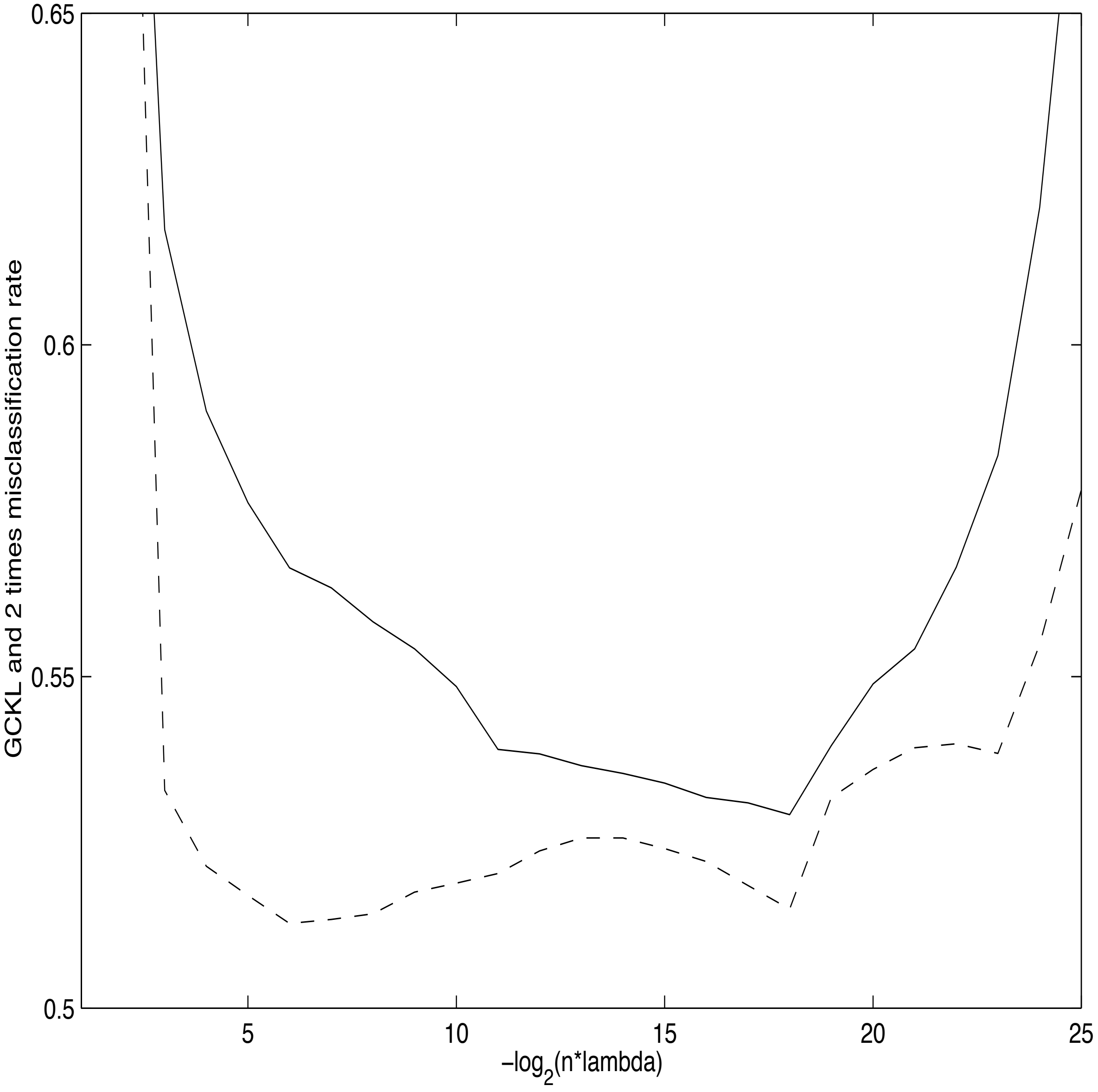

Figure 3: GCKL (solid line) and two times misclassification rate (dashed line) of f_λ with varying λ for a fixed sample with $n = 257$, where f_λ is the solution to the SVM regularization problem with $q = 1$ and the Sobolev Hilbert space kernel. Notice the x-axis is $-\log_2(n\lambda)$. (Larger values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

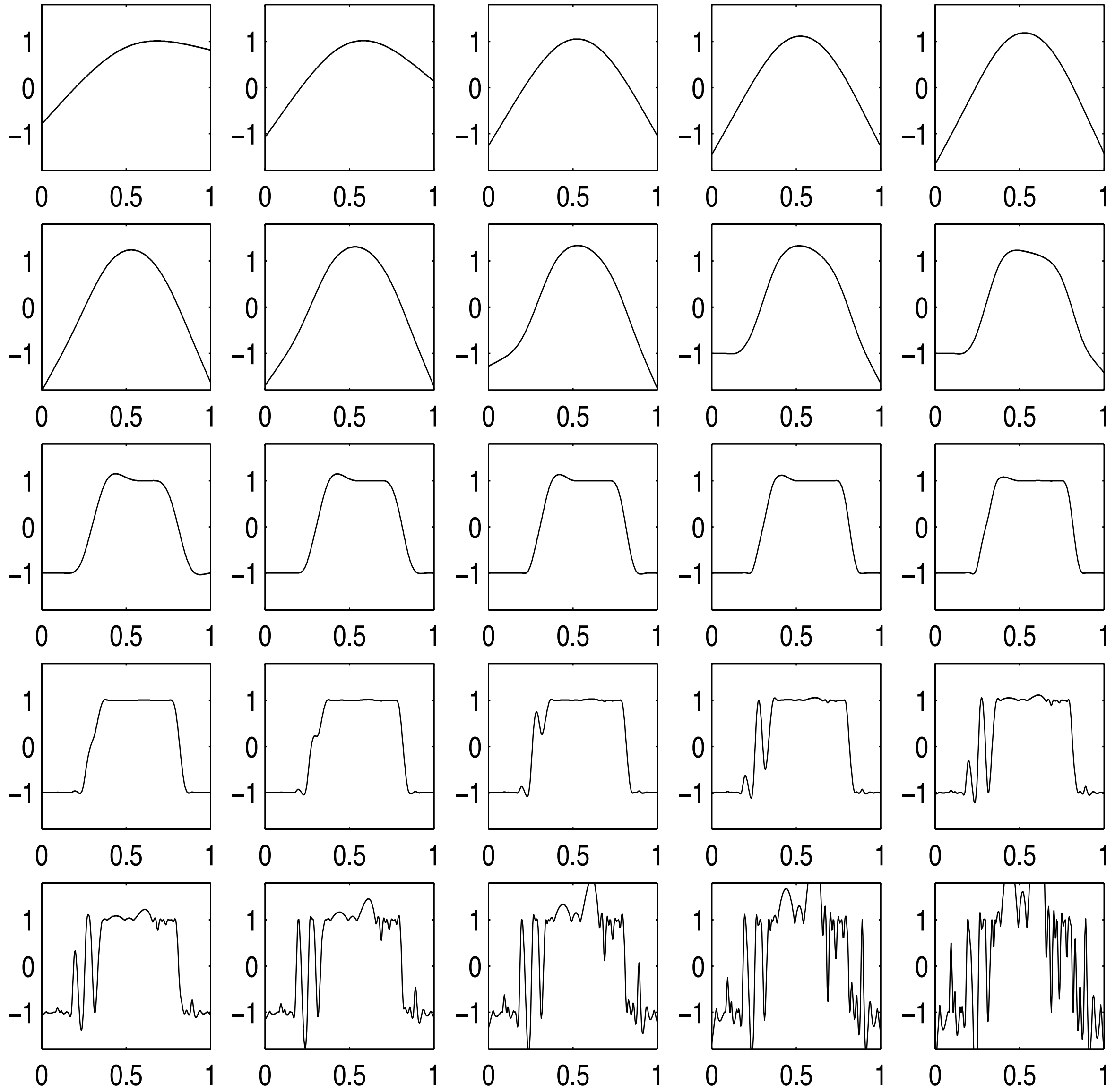

Figure 4: For the same sample as in Figure 3, the solutions to the SVM regularization problem with $q = 1$ and the Sobolev Hilbert space kernel for $n\lambda = 2^{-1}, 2^{-2}, ..., 2^{-25}$. We see the solution is close to $sign[p(x) - 1/2]$ when GCKL in Figure 3 is close to the minimum.

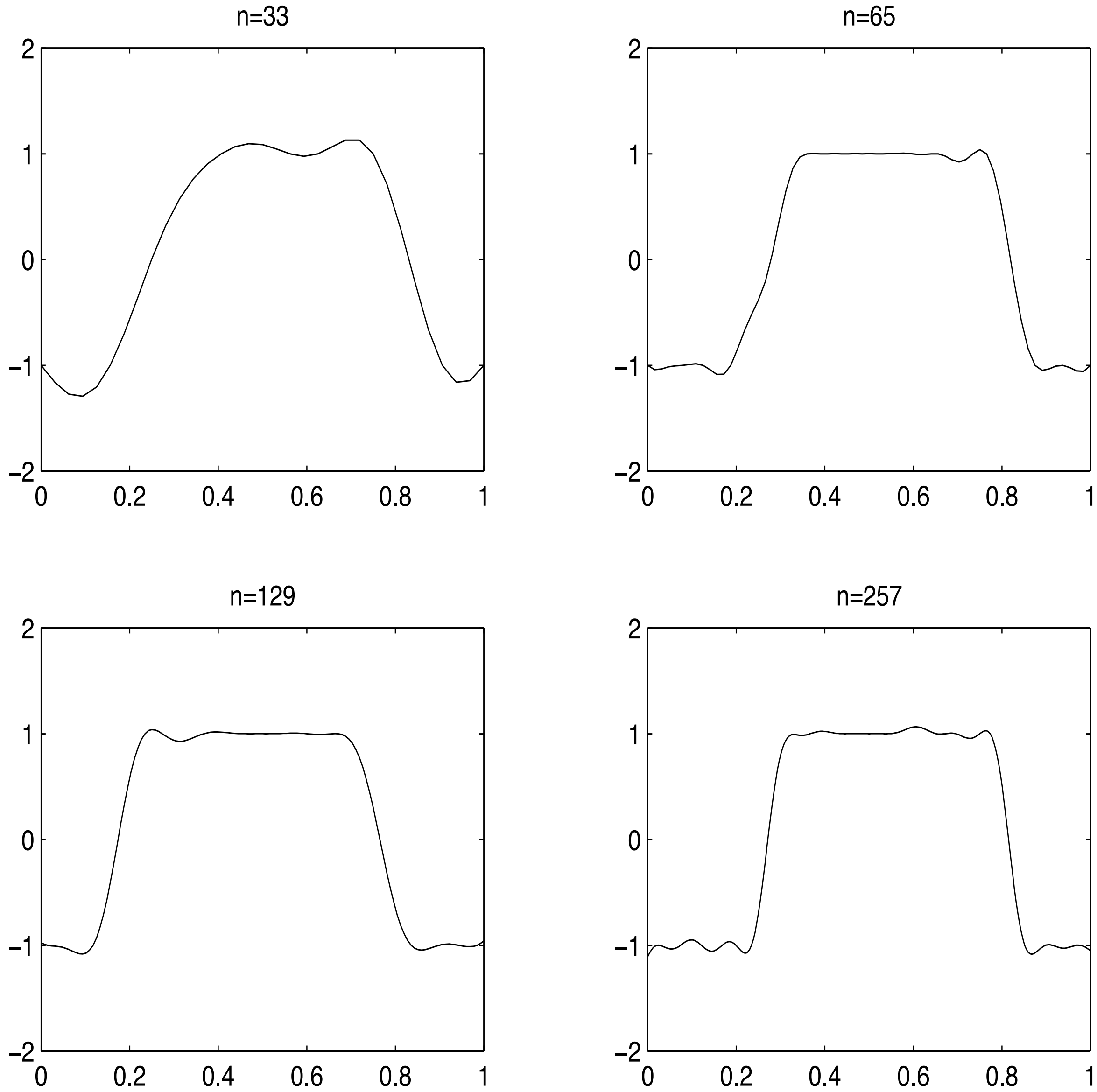

Figure 5: The solutions to the SVM regularization problems with $q = 1$ and the Gaussian kernel for samples of size 33, 65, 129, 257. The tuning parameter λ and σ are chosen to minimize GCKL in each case.

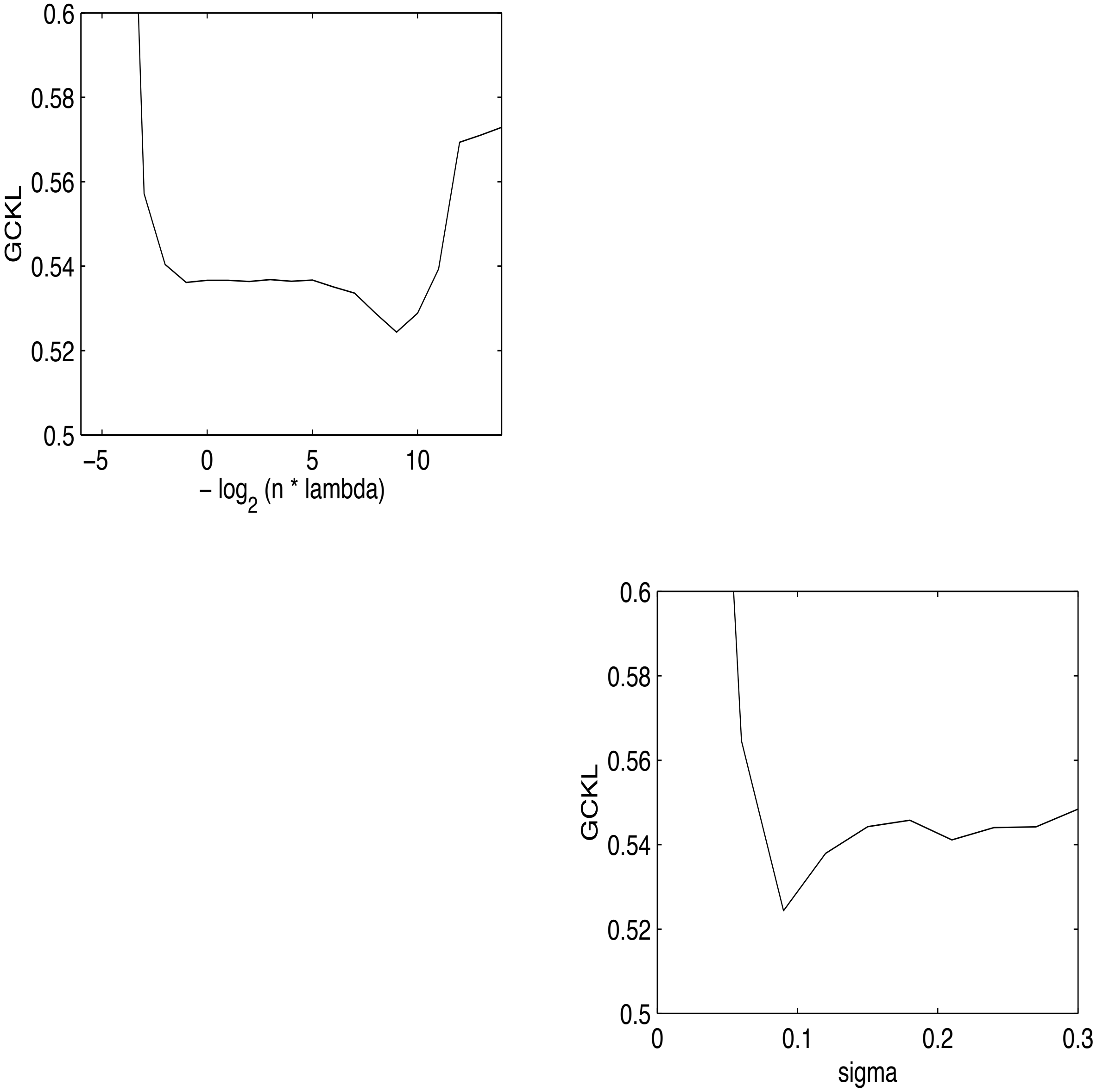

Figure 6: GCKL of $f_{\sigma,\lambda}$ with varying σ and λ for a fixed sample with $n = 257$, where $f_{\sigma,\lambda}$ is the solution to the SVM regularization problem with $q = 1$ and the Gaussian kernel $\exp[-\frac{(s-t)^2}{2\sigma^2}]$. Upper left: GCKL of $f_{\sigma,\lambda}$ with σ fixed to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

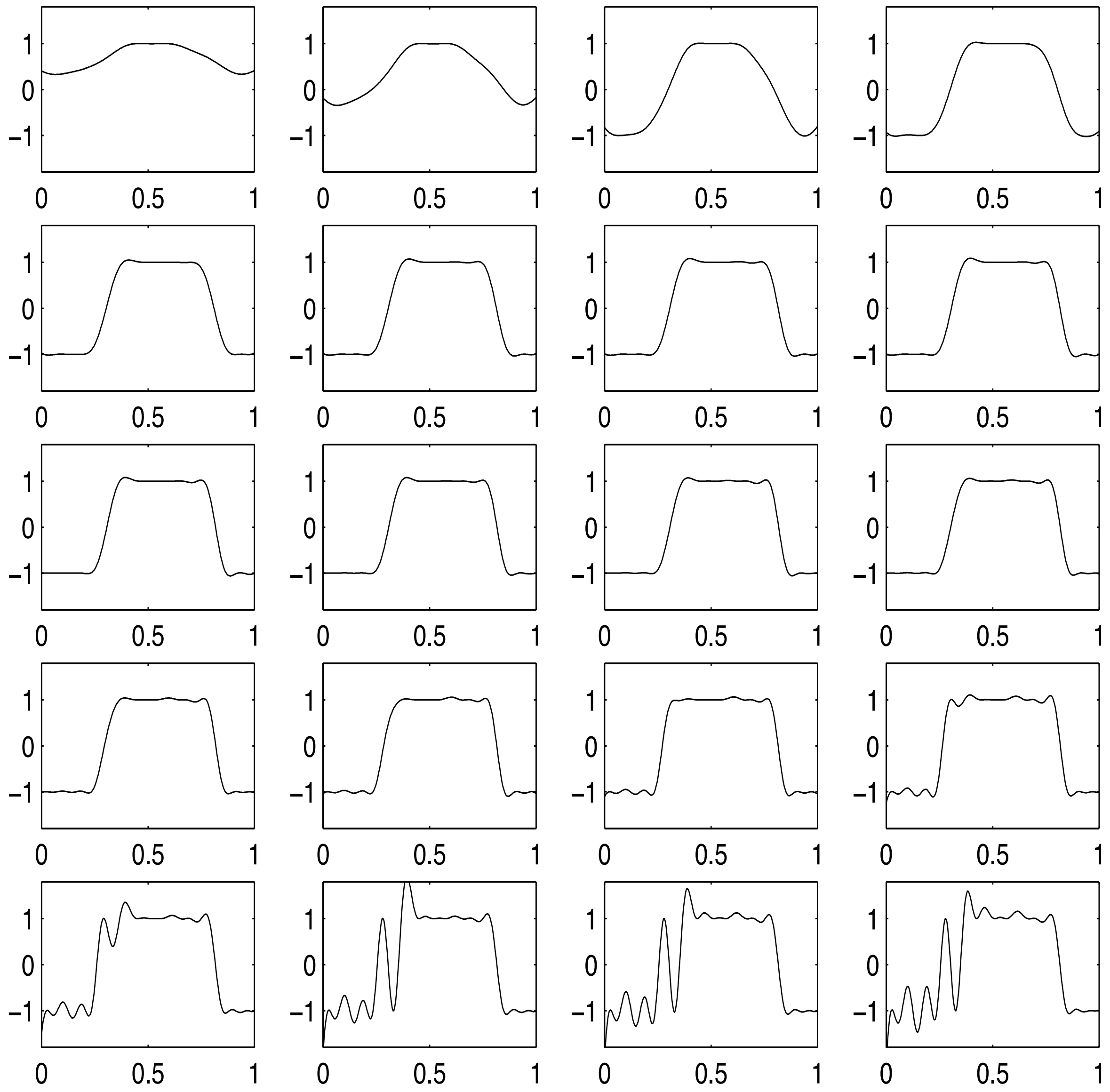

Figure 7: For the same sample as in Figure 6, with σ fixed at 0.09, the solutions to the SVM regularization problem with $q = 1$ and Gaussian kernel for $n\lambda = 2^5, 2^4, ..., 2^{-14}$. We see the solution is close to $sign[p(x) - 1/2]$ when GCKL in Figure 6 (upper left picture) is close to the minimum.

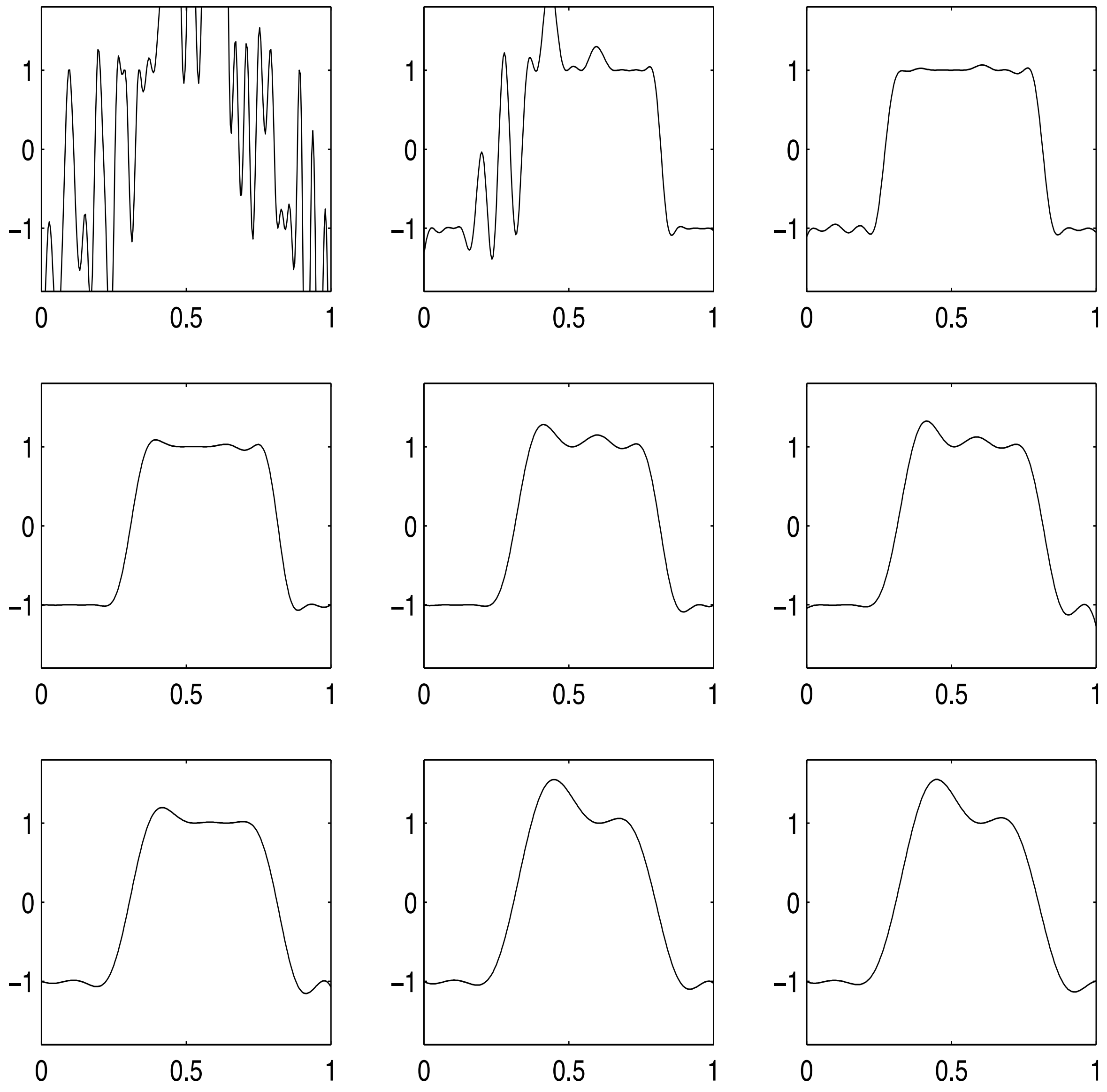

Figure 8: For the same sample as in Figure 6, with $n\lambda$ fixed at 2^{-9}, the solutions to the SVM regularization problem with $q = 1$ and Gaussian kernel for $\sigma = 0.03, 0.06, ..., 0.27$. We see the solution is close to $sign[p(x) - 1/2]$ when GCKL in Figure 6 (lower right picture) is close to the minimum.