DEPARTMENT OF STATISTICS

University of Wisconsin

1210 West Dayton St.

Madison, WI 53706

TECHNICAL REPORT NO. 1029

November 6, 2000

On the Support Vector Machine [1] [2]

by

Yi Lin

[1] **Key words: Support Vector Machine, Bayes Rule, Classification, Sobolev Hilbert Space, Reproducing Kernel, Reproducing Kernel Hilbert Space, Regularization Method.**

[2] Supported by Wisconsin Alumni Research Foundation.

On the Support Vector Machine

Yi Lin

Department of Statistics

University of Wisconsin, Madison

1210 West Dayton Street

Madison, WI 53706-1685

Phone: (608)262-6399

Fax: (608)262-0032

Email: yilin@stat.wisc.edu

November 6, 2000

On the Support Vector Machine

Yi Lin
University of Wisconsin, Madison

Abstract

Classification is a fundamental problem at the intersection of machine learning and statistics. Machine learning methods have enjoyed considerable empirical success. However, they often have an ad hoc quality. It is desirable to have hard theoretical results which might highlight specific quantitative advantages of these methods. The statistical methods often tackle the classification problem through density estimation or regression. Theoretical properties of these statistical methods can be established, but only under the assumption of a fixed order of smoothness. Whether these methods work well when the assumptions are violated is not clear.

The support vector machine (SVM) methodology is a rapidly growing area in machine learning, and is receiving considerable attention in recent years. The SVM has proved highly successful in a number of practical classification studies. In this paper we show that the SVM enjoys excellent theoretical properties which explain the good performance of the SVM. We show that the SVM approaches the the theoretical optimal classification rule (the Bayes rule) in a direct fashion, and its expected misclassification rate quickly converges to that of the Bayes rule. The results are established under very general conditions allowing discontinuity. They testify to the fact that classification is easier than density estimation and regression, and show that the SVM works by taking advantage of this. The results pinpoint the exact mechanism behind the SVM, and clarify the advantage and limitation of the SVM, thus give insights on how the SVM can be extended systematically.

Key Words and Phrases: Support Vector Machine, Bayes Rule, Classification, Sobolev Hilbert Space, Reproducing Kernel, Reproducing Kernel Hilbert Space, Regularization Method.

1 Introduction

In the classification problem, we are given a training data set of n subjects, and for each subject $i \in \{1, 2, ..., n\}$ in the training data set, we observe an explanatory vector $x_i \in R^d$, 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 i.i.d. from an unknown probability distribution $P(x, y)$, or equivalently, they are independent random realizations of the random pair (X, Y) that has cumulative probability distribution $P(x, y)$. The task of classification is to derive from the training set a good classification rule, so that once we are given the x value of a new subject, we can assign a class label to it. One common criterion for accessing the quality of a classification rule is the generalization error rate (expected misclassification rate), though other loss functions are also possible. The situation where there are only two classes and where the misclassification rate is used as the criterion is most commonly encountered in practice. In the following we concentrate on this situation. This binary classification problem (or pattern recognition) has been studied by many authors. See, for example, Devroye, Györfi and Lugosi (1996) and Vapnik (1995) and the references cited therein.

In this paper, the two classes will be called the positive class and the negative class, and will be coded as $+1$ and -1 respectively. Any classification rule η can then be seen as a mapping from R^d to $\{-1, 1\}$. Denote the generalization error rate of a classification rule η as $R(\eta)$. Then $R(\eta) = \int \frac{|y-\eta(x)|}{2} dP$. It is often the case that $\eta(\cdot) = sign[g(\cdot)]$ for some real valued function g. That is, the rule η assigns the subject to the positive class if $g(x)$ is positive, and to the negative class otherwise. In this case, we will use the notations $R(\eta)$ and $R(g)$ interchangeably.

1.1 The Bayes Rule

In the classification problem, if we knew the underlying probability distribution $P(x, y)$, we could derive the optimal classification rule with respect to any given loss function. This optimal rule is usually called the Bayes rule. For the binary classification problem, the Bayes

rule that minimizes the generalization error rate is

$$\eta^*(x) = sign[p(x) - 1/2], \tag{1}$$

where

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

is the conditional probability of the positive class at a given point x. Let $R^* = R(\eta^*)$. Then R^* is the minimal possible value for $R(\cdot)$.

Let $g^+(x)$ be the probability density function of X for the positive population, that is, the conditional density of X given $Y = 1$. Let $g^-(x)$ be the probability density function of X for the negative population. The unconditional ("prior") probabilities of the positive class and negative class in the target population are denoted by π^+ and π^- respectively. Then $p(x)$ can be obtained by the Bayes formula

$$p(x) = \frac{\pi^+ g^+(x)}{\pi^+ g^+(x) + \pi^- g^-(x)} \tag{2}$$

1.2 Common classification methods and the SVM

The statistical approach to classification estimates the conditional class probability $p(x)$ (or equivalently, the log odds $\log[p(x)/(1 - p(x))]$). Once an estimate of $p(x)$ is obtained, we can plug it in (1) to get an approximate Bayes rule. The estimation is often done by logistic regression; or by estimating the densities $g^+(x)$ and $g^-(x)$, and then using (2). It is possible to establish the statistical properties of these methods by making use of the extensive existing results on density estimation and regression. However, these methods posit a fixed order of smoothness assumptions on the conditional probability function $p(x)$ or the densities $g^+(x)$ and $g^-(x)$. This leaves the applicability of these methods in doubt since we never know such smoothness to be the case in practice.

The machine learning community places great emphasis on algorithms and handling large data sets. Many machine learning methods, such as the neural network, the classification tree, and recently the support vector machine, have enjoyed remarkable empirical success, and have attracted tremendous interest. The machine learning methods are often motivated by their heuristic plausibility, and justified by empirical evidence rather than hard theoretical

results that might demonstrate specific quantitative advantages of such methods. In order to have a clear understanding of where and when these methods work well, it is desirable to have theoretical results that pinpoint the advantages and limitations of these methods.

The support vector machine is a new addition to the machine learning toolbox. It was first proposed in Boser, Guyon and Vapnik (1992), and is going through rapid development. The SVM is best developed in the binary classification situation, even though several studies attempted to use the SVM for classifying multiple classes. Since the statistics community is largely unfamiliar with the SVM, in the following we give a brief description of the derivation of the SVM, starting from the simple linear support vector machine and moving on to the nonlinear support vector machine. For a more detailed tutorial on the support vector machine, see Burges (1998).

The SVM is motivated by the intuitive geometric interpretation of maximizing the margin. When the two classes of points in the training set can be separated by a linear hyperplane, it is natural to use the hyperplane that separates the two groups of points in the training set by the largest margin. This amounts to the hard margin linear support vector machine: Find $w \in R^d, b \in R$, to minimize $\|w\|^2$, subject to

$$x_i \cdot w + b \geq +1 \quad for\ y_i = +1; \tag{3}$$

$$x_i \cdot w + b \leq -1 \quad for\ y_i = -1; \tag{4}$$

Once such w and b are found, the SVM classification rule is $sign(w \cdot x + b)$.

When the points in the training data set are not linearly separable, constraints (3) and (4) can not be satisfied simultaneously. We can introduce nonnegative slack variables ξ_i's to overcome this difficulty. This results in the soft margin linear support vector machine: Find $w \in R^d, b \in R$, and ξ_i, $i = 1, 2, ..., n$, to minimize $1/n \sum_i \xi_i + \lambda \|w\|^2$, under the constraints

$$x_i \cdot w + b \geq +1 - \xi_i \quad for\ y_i = +1; \tag{5}$$

$$x_i \cdot w + b \leq -1 + \xi_i \quad for\ y_i = -1; \tag{6}$$

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

Here λ is a control parameter to be chosen by the user. Notice (5) and (6) can be combined as

$$\xi_i \geq 1 - y_i(x_i \cdot w + b).$$

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. It turns out that the computation of this linear SVM in the feature space can be carried out in the original space through a (reproducing) kernel trick. Therefore we do not really need to know the feature space and the transformation to the feature space. The nonlinear support vector machine with kernel K is equivalent to a regularization problem in the reproducing kernel Hilbert space (RKHS) H_K: Find $f(x) = h(x) + b$ with $h \in H_K$, $b \in R$, and ξ_i, $i = 1, 2, ..., n$, to minimize

$$\frac{1}{n}(\sum_i \xi_i) + \lambda \|h\|^2_{H_K}, \tag{7}$$

under the constraints

$$\xi_i \geq 1 - y_i f(x_i), \tag{8}$$

$$\xi_i \geq 0, \ \forall i. \tag{9}$$

Once the solution $\hat{f}$ is found, the SVM classification rule is $sign(\hat{f})$.

Commonly used kernels include Gaussian kernels, spline kernels, and polynomial kernels. Wahba (1990) contains some detailed introduction to reproducing kernels and reproducing kernel Hilbert spaces.

The theory of RKHS ensures that the solution to (7), (8), and (9) lies in a finite dimensional space, even when the RKHS H_K is of infinite dimension. See Wahba (1990). Hence the SVM problem (7), (8), and (9) becomes a mathematical programming problem in a finite dimensional space. See Wahba, Lin and Zhang (1999). The computation of the SVM is often done with the dual formulation of this mathematical programming problem. This dual formulation is a quadratic programming problem with a simple form. It turns out that the SVM solution enjoys certain sparsity: usually the final solution depends only on a small proportion of the data points. These points are called support vectors. This sparsity can be exploited for fast computation, and the SVM has been applied to very large datasets. See,

for example, Vapnik (1979), Osuna, Freund and Girosi (1997), Platt (1999), for some basic ideas of fast computation of the SVM.

Denote $l_n(f) = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i))]_+$. Here $a_+ = a$ if $a \geq 0$, and $a_+ = 0$ if $a < 0$. The limit functional of $l_n(f)$ is $l(f) = \int [1 - yf(x)]_+ dP$. We can see that (7), (8), (9) is equivalent to minimizing

$$l_n(f) + \lambda \|h\|^2_{H_K}. \tag{10}$$

Several authors have studied the generalization error rate of the SVM, 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, the VC dimension or the fat shattering dimension of the nonlinear SVM is often very large, even infinite. Hence the bounds established are often very loose, or even trivial (larger than 1), and do not provide a satisfactory explanation as to why the SVM often has good generalization performance.

Due to the heuristic fashion in which the SVM is derived, it has not been clear how the SVM is related to Bayes rule, and how the generalization error rate of the SVM compares with the minimal possible value R^*. Some confusions exist in practice on what to do with the SVM when the appropriate measure of risk is not the expected misclassification rate, and how the SVM can be used for multi-class classification. In this paper, we clarify matters by pinpointing the exact mechanism behind the SVM. This will enable us to extend the SVM methodology and develop new algorithms based on the basic ideas of the SVM.

2 Statements of Our Results

From (10) we see the nonlinear SVM is another example of the penalized method very often used in statistics. Lin (1999) showed that the minimizer of $l(f)$ is $sign[p(x) - 1/2]$, which is exactly the Bayes rule. This strongly suggests that the SVM solution is aiming at approaching the Bayes rule. Lin (1999) demonstrated with simulations that with Gaussian kernel and spline kernel the solution to (10) approaches to the function $sign[p(x) - 1/2]$. One point worth mentioning is that the function $sign[p(x) - 1/2]$ is usually discontinuous

and does not belong to any RKHS commonly used in practice, while the solution to (10) is in the RKHS H_K. This is different from the situation of many penalized methods.

In this paper we consider the first order spline kernel in the situation $d = 1$. The simple reproducing kernel under consideration facilitates the proofs. However, in principle the same line of argument can be applied to the SVM with other commonly used reproducing kernels. We show that under very general conditions without any smoothness assumption, the solution to (10) converges to $sign[p(x) - 1/2]$. We further show that under very mild boundary conditions on $p(x)$, the generalization error rate of the SVM converges to R^* at a certain rate. These conditions are much weaker than the usual smoothness conditions imposed in regression and density estimation, and can easily be satisfied by nonsmooth, even discontinuous functions. Also the implementation of the SVM does not require any a priori information of the conditions.

Assumption 1 *The density $d(x)$ of X is supported on $[-1, 1]$, and it is bounded away from zero and infinity in this interval. That is, there exists constants $D_2 > D_1 > 0$, such that $D_1 \leq d(x) \leq D_2$ for all $x \in [-1, 1]$.*

The first order Sobolev Hilbert space of functions on any interval $[b_1, b_2]$, denoted by $H^1[b_1, b_2]$, is defined by

$$H^1[b_1, b_2] = \{f | f \text{ abs. cont.}; \ \ f' \in L_2[b_1, b_2]\},$$

with the Sobolev Hilbert norm

$$\|f\|^2 = \int_{b_1}^{b_2} f^2 dx + \int_{b_1}^{b_2} (f')^2 dx$$

In this paper we will write $H^1[-1, 1]$ as simply H^1. It is well known that this is a RKHS and the corresponding RK is called the first order spline kernel. With this RK, (10) becomes

$$\min_{f \in H^1} l_n(f) + \lambda \int_{-1}^{1} (f')^2 dx, \tag{11}$$

or equivalently,

$$\min_{f \in H^1} l_n(f) \quad \text{subject to} \quad \int_{-1}^{1} (f')^2 dx \leq M. \tag{12}$$

Here λ or M is the smoothing parameter. Let the solution to (11) be denoted by $\hat{f}$. Let the solution to (12) be denoted by $\hat{f}_M$. We will allow the smoothing parameters to vary with the sample size n. In this paper, any function with a hat, such as $\hat{f}$ and $\hat{f}_M$, are random (depending on the training sample). We use the notation E_c for the expectation conditional on the training sample (X_i, Y_i), $i = 1, 2, ..., n$. Then $E_c[\hat{g}(X)] = \int \hat{g}(x)dP$ for any random function $\hat{g}$ depending on the training sample.

Theorem 1 *Under Assumption 1, suppose $p(x)$ is bounded away from $1/2$ from below by some positive constant D_3 in an interval $[x_0 - \delta, x_0 + \delta]$. Then there exists a positive number Λ depending only on D_3 and δ, such that for any fixed $\lambda < \Lambda$, or for any fixed sequence $\lambda_{(n)}$ going to zero, we have*

$$|sign[p(x_0) - 1/2] - \hat{f}(x_0)| = O_p(n^{-1/3}\lambda^{-2/3}).$$

The same result is valid for $p(x)$ bounded away from $1/2$ from above.

The result is uniform over all the functions $p(x)$ and points x_0 satisfying that $p(x)$ is bounded from below by $1/2 + D_3$ (or from above by $1/2 - D_3$) in the interval $[x_0 - \delta, x_0 + \delta]$. Theorem 1 shows that the SVM solution converges to the Bayes rule $sign[p(x) - 1/2]$. This uncovers the mechanism by which the SVM works. Notice that $\hat{f}$ is absolutely continuous, whereas $sign[p(x) - 1/2]$ is usually discontinuous.

To investigate the global performance of the SVM, we consider the generalization error rate. For any classification rule η, it is natural to access its quality by looking at how fast $R(\eta)$ converges to the minimal possible value R^*. The convergence $R(\eta) \to R^*$ was proved for various classification rules (not including the SVM though). Furthermore, certain bounds on the difference $E(R(\eta) - R^*)$ are known for finite sample sizes. See, for example, Devroye, Györfi and Lugosi (1996) and the references therein.

For the rate of such convergence, the only studies in the literature that we know of are that of Marron (1983) and Mammen and Tsybakov (1999). Under smoothness assumptions on the densities $g^+(x)$ and $g^-(x)$, Marron (1983) proved that the optimal rates of convergence are the same as those of the mean integrated squared error in density estimation. He also showed that under these assumptions the density estimation approach to the classification

problem is asymptotically optimal. The error criterion he used is the integrated (over all prior probabilities q from 0 to 1) difference $R_q(\eta) - R_q^*$, where $R_q(\cdot)$ and R_q^* are the generalization error rate when $\pi^+ = q$. Mammen and Tsybakov (1999) imposes conditions on the decision region and assumes that the decision region belongs to a known class $\mathcal{G}$ of possible "candidate" regions. The δ-entropy with bracketing of the class $\mathcal{G}$ is assumed to be finite and varies with δ at a certain rate. They studied the asymptotic properties of direct minimum contrast estimators and sieve estimators, and found the optimal rate of $R(\eta) - R^*$ for classes of boundary fragments. The rates they obtained are faster than those in Marron (1983). They concluded that direct estimation procedures such as the empirical risk minimization can achieve better performance in terms of the generalization error rate than the density estimation based method. However, the direct minimum contrast estimators and sieve estimators are hard to implement and need a priori knowledge of the class $\mathcal{G}$. Our second result (to be stated) is in spirit closer to the results in Mammen and Tsybakov (1999), but we study the asymptotic properties of the SVM, which do not assume a priori knowledge of a candidate class with finite δ-entropy (with bracket).

For any function g, if we classify according to the sign of $g(x)$, then it is easy to see the misclassification rate $R(g)$ is equal to $l[sign(g)]/2$. By Theorem 1 we see that $sign(\hat{f}) \approx \hat{f}$. So $l(\hat{f}) \approx 2R(\hat{f})$. Therefore it is also reasonable to use $l(\cdot)$ to assess the performance of the SVM. In fact, $l(\cdot)$ is called the GCKL in Wahba, Lin, and Zhang (1999), and was used to adaptively tune the smoothing parameter for the SVM. It might be advantageous in many situations to consider $l(\cdot)$ rather the $R(\cdot)$, since $l(\cdot)$ is continuous and convex, whereas $R(\cdot)$ is discontinuous and not convex. It is shown in Lin (1999) that the Bayes rule η^* is the minimizer of $l(f)$ over all function f. In this paper we also consider how fast the GCKL of the SVM converges to $l(\eta^*)$.

Before we can state our second result, we need to characterize the behavior of $p(x)$ at its cross points with 1/2. We say a point r is a positive cross point if there exists a positive number $a > 0$, such that $p(x) > 0$ in $(r, r + a]$ and $p(x) < 0$ in $[r - a, r)$. Negative cross points are defined likewise.

Assumption 2 *The function $p(x)$ crosses 1/2 finite many (k) times, that is, $sign[p(x)-1/2]$*

has finite many pieces; and there exists $\zeta > 0$ *and* $D_4 > 0$ *such that for any cross point* r_j, $j = 1, 2, ..., k$, *there exists* $\alpha_j \geq 0$, *and* $D_{6j} > D_{5j} > 0$, *satisfying*

$$D_{5j}|x - r_j|^{\alpha_j} \leq |p(x) - 1/2| \leq D_{6j}|x - r_j|^{\alpha_j}, \quad \forall x \in (r_j - \zeta, r_j + \zeta), \tag{13}$$

and $p(x)$ *is bounded away from* $1/2$ *by* D_4 *when* x *is more than* ζ *away from all the cross points. Denote* $\max_j D_{6j} = \bar{D}$, $\min_j D_{5j} = \underline{D}$, $\max_j \alpha_j = \bar{\alpha}$, *and* $\min_j \alpha_j = \underline{\alpha}$.

It falls right out from this assumption that $k \leq 2/\zeta$. Assumption 2 is related to the condition (4) in Mammen and Tsybakov (1999). Assumption 2 could, in particular, be satisfied by nonsmooth, even discontinuous functions. Notice also that we allow the possibility that some α_j is zero. This represents possible discontinuity at the cross points. Notice the implementation of the SVM does not require any a priori information in Assumption 2.

We will consider the setup (12) and its solution $\hat{f}_M$ in our second result for technical convenience. For any $\theta > 0$, denote $\rho(\theta) = \min(\underline{\alpha} + 1 - \theta, \theta/\bar{\alpha}, (\underline{\alpha} + 2)/(\bar{\alpha} + 2))$. ($\theta/\bar{\alpha} = +\infty$ if $\bar{\alpha} = 0$.)

Theorem 2 *Under Assumption 1 and Assumption 2, for any fixed* $\theta > 0$, *suppose* $M_{(n)} \sim n^t$ *for some* $0 < t \leq 2/[3(1 + \rho(\theta))]$, *then for any fixed* $s > 0$, *there exists finite constant* $D(s)$ *depending on* s, *and* $N > 0$, *such that for any* $n > N$,

$$n^{\gamma s} E[l(\hat{f}_M) - l(\eta^*)]^s \leq D(s); \tag{14}$$

$$n^{\gamma s} E[R(\hat{f}_M) - R(\eta^*)]^s \leq D(s); \tag{15}$$

where $\gamma = \min\{[t(\rho(\theta) + \theta)], 2/3 - [t(1 + \theta)]/3\}$. *The constants* $D(s)$ *and* N *depend on* $p(x)$ *only through* ζ, $\bar{D}$, $\underline{D}$, D_4, *and* $\bar{\alpha}$.

For example, if $\bar{\alpha} = 0$, $\theta = 1/2$, then $\rho(\theta) = 1/2$, and $\gamma = 4/9$ with $t = 4/9$; if $\bar{\alpha} = \underline{\alpha} = 2$, $\theta = 2$, then $\rho(\theta) = 1$, and $\gamma = 1/2$ with $t = 1/6$.

The proof of Theorem 2 uses general results from empirical process theory, and follows an argument employed in the proof of Theorem 1 of Mammen and Tsybakov (1999). One complication is that the L_2 norm is not readily bounded by difference measured by $l(\cdot)$. We derive (29) to overcome this difficulty.

3 Discussion

The SVM makes no a priori assumption of a fixed order of smoothness or a fixed class of possible "candidate" decision regions. It easily accommodates discontinuity. Its generalization error rate goes to that of the Bayes rate with a fast rate of convergence.

The reason why we can obtain results under very general conditions is that the function $sign[p(x) - 1/2]$, though may be discontinuous, is often simpler than the function $p(x)$ or $g^+(x)$ and $g^-(x)$. The SVM takes advantage of this fact by aiming directly at the simpler function $sign[p(x) - 1/2]$ which is more directly related to the decision rule. Several authors have observed that classification is easier than regression and density estimation. See Devroye, Györfi and Lugosi (1996) and Mammen and Tsybakov (1999). Our results further confirm this.

It is important to understand the mechanism behind the SVM. The SVM implement the Bayes rule in an interesting way: Instead of estimating $p(x)$, it estimates $sign[p(x) - 1/2]$. This has advantages when our goal is binary classification with minimal expected misclassification rate. However, this also means that in some other situations the SVM needs to be modified, and should not be used as is.

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 nonstandard situations the Bayes rule that minimizes the expected misclassification cost can be expressed as $sign[p(x) - c]$, where $c \in (0, 1)$ is not equal to $1/2$. Hence the SVM as is will not perform optimally in this situation, and there is no direct way of getting $sign[p(x) - c]$ from $sign[p(x) - 1/2]$. Lin, Lee, and Wahba (2000) contains some extension of the SVM to such nonstandard situations.

Multi-class (N-class) classification problem arises naturally in practice. The Bayes rule in this case assigns the class label corresponding to the largest conditional class probability. Some authors suggested training N one-versus-rest SVMs and taking the class for a test subject to be that corresponding to the largest value of the classification functions. Our results show that this approach should work well when one of the conditional class probabilities is

larger than 1/2, (there is a majority class), but will not approach the Bayes rule when there is no majority class.

4 The Proof of Our Results

The notation $a_n \sim b_n$ means $c_1 a_n \leq b_n \leq c_2 a_n$ for all n, and some positive constants c_1 and c_2. Any constants here and later in the proofs are generic positive constants not depending on n, λ, M, or the sample, and may depend on $p(x)$ and $d(x)$ only through D_1, D_2, D_3, δ, ζ, D_4, $\bar{D}$, $\underline{D}$, and $\bar{\alpha}$. Consecutive appearances of c without subscript may stand for different positive constants.

Lemma 4.1 *Under Assumption 1, suppose $p(x)$ be bounded away from 1/2 from below by some positive constant D_3 in $[x_0 - \delta, x_0]$. For any fixed number $a \in [-1, 1]$, let f_a be the solution to the variational problem:*

$$\min_{\substack{f \in H^1[x_0-\delta, x_0] \\ f(x_0-\delta)=a}} E[(1 - Yf(X))_+ 1_{\{x_0-\delta \leq X \leq x_0\}}] + \lambda \int_{x_0-\delta}^{x_0} (f')^2 dx, \tag{16}$$

then when λ is small enough, there exists $\epsilon \in [0, \delta)$, such that $\epsilon \sim \lambda^{1/2}(1-a)^{1/2}$, and $f_a = 1$ for $x \in [x_0 - \delta + \epsilon, x_0]$. Also,

$$\int_{x_0-\delta}^{x_0-\delta+\epsilon} (f_a')^2 dx \sim \lambda^{-1/2}(1-a)^{3/2}$$

$$\int_{x_0-\delta}^{x_0-\delta+\epsilon} (1 - f_a) dx \sim \lambda^{1/2}(1-a)^{3/2}$$

$$E[(1 - Yf_a(X))_+ 1_{\{x_0-\delta \leq X \leq x_0\}}] + \lambda \int_{x_0-\delta}^{x_0} (f_a')^2 dx - E[(1-Y)_+ 1_{\{x_0-\delta \leq X \leq x_0\}}] \sim \lambda^{1/2}(1-a)^{3/2}$$

Proof: Without loss of generality, we assume $x_0 = 0$.

It is easy to see that $f_a(x) \in [-1, 1]$, $\forall x \in [-\delta, 0]$. Otherwise the truncation of f_a into $[-1, 1]$ would still be in $H^1[-\delta, 0]$, and has a smaller value for (16). Now let us restrict our attention to functions satisfying $|f(x)| \leq 1$, $\forall x \in [-\delta, 0]$. Under this constraint we have

$$\begin{aligned} & E\{[1 - Yf(X)]_+ 1_{\{-\delta \leq X \leq 0\}}\} + \lambda \textstyle\int_{-\delta}^{0} (f')^2 dx \\ = \;& E[(1 - Yf(X)) 1_{\{-\delta \leq X \leq 0\}}] + \lambda \textstyle\int_{-\delta}^{0} (f')^2 dx \\ = \;& \textstyle\int_{-\delta}^{0} d(x) dx - \int_{-\delta}^{0} g(x) f(x) dx + \lambda \int_{-\delta}^{0} (f')^2 dx, \end{aligned} \tag{17}$$

where $g(x) = d(x)[2p(x) - 1]$.

From (17) we can see that f_a must be monotone increasing. Otherwise suppose $f_a(x_1) > f_a(x_2)$, $-\delta \le x_1 < x_2 \le 0$. Consider the function defined as

$$\tilde{f}_a(x) = \begin{cases} f_a(x) & : \quad x \in [-\delta, x_1] \\ \max\{f_a(x_1), f_a(x)\} & : \quad x \in [-x_1, 0]. \end{cases}$$

Then $\tilde{f}_a \in H^1[-\delta, 0]$, $\tilde{f}_a(-\delta) = a$, $|\tilde{f}_a| \le 1$ for any $x \in [-\delta, 0]$, and $\tilde{f}_a$ gives a smaller value of (17).

Let $G(x) = \int_0^x g(t)dt$. Then $G(0) = 0$, $G(x)$ is continuous and strictly monotone increasing in $[-\delta, 0]$. Integrating by parts, we have (17) is the same as

$$\begin{aligned} & \int_{-\delta}^0 d(x)dx + aG(-\delta) + \int_{-\delta}^0 G(x)f'(x)dx + \lambda \int_{-\delta}^0 (f')^2 dx \\ = \quad & \lambda \int_{-\delta}^0 [f' + G/(2\lambda)]^2 - \int_{-\delta}^0 G^2/(4\lambda)dx + \int_{-\delta}^0 d(x)dx + aG(-\delta) \end{aligned}$$

From the above we know $h_a = f_a'$ solves the problem

$$\min_{\substack{h \in L^2[-\delta,0], \\ h \ge 0}} \lambda \int_{-\delta}^0 [h + G/(2\lambda)]^2,$$

subject to the constraint

$$1 - a - \int_{-\delta}^0 h(x) \ge 0. \tag{18}$$

Since $p(x) \ge 1/2 + D_3$, and $D_1 \le d(x) \le D_2$, by the definition of $G(\cdot)$, there exists a positive constant Λ, such that for any $\lambda \le \Lambda$, we have

$$\int_{-\delta}^0 -G/(2\lambda)dx > 2 \ge 1 - a. \tag{19}$$

So the constraint (18) is not trivial. Introducing Lagrange multiplier $\mu > 0$ for the constraint (18), we get

$$\begin{aligned} & \lambda \int_{-\delta}^0 [h + G/(2\lambda)]^2 dx - \mu[1 - a - \int_{-\delta}^0 h(x)dx] \\ = \quad & \lambda \int_{-\delta}^0 (f' + (G + \mu)/(2\lambda))^2 dx - \mu(1 - a) - \int_{-\delta}^0 (\mu^2 + 2G\mu)/(4\bar{\alpha}). \end{aligned}$$

So $h_a = [-(G + \mu)/(2\lambda)]_+$. We also have $1 - a - \int_{-\delta}^0 h_a(x) = 0$, which means $f_a(0) = 1$.

Recalling that $-G$ is continuous, strictly decreasing to 0 on $[-\delta, 0]$. Let $-G$ crosses μ at $-\delta+\epsilon$, where $-\delta < \epsilon < 0$. Then

$$\begin{aligned} f_a(x) &= 1 \quad x \in [-\delta+\epsilon, 0] \\ f_a(x) & \qquad \text{strictly increases from } \ a \ \text{ to } \ 1 \ \text{ in } \ [-\delta, -\delta+\epsilon]. \end{aligned} \tag{20}$$

By the definition of $g(x)$ and $G(x)$, there exists positive constants c_1 and c_2, such that $c_1 < g(x) = G'(x) < c_2$, $\forall x \in [-\delta, 0]$. It is easy to see that $2\lambda(1-a) = \int_{-\delta}^{0} 2\lambda h_a = \int_{-\delta}^{-\delta+\epsilon}(-G-\mu)dx$ is in between $1/2c_1\epsilon^2$ and $1/2c_2\epsilon^2$. Therefore ϵ is in between $2c_2^{-1/2}\lambda^{1/2}(1-a)^{1/2}$ and $2c_1^{-1/2}\lambda 1/2(1-a)^{1/2}$.

By the definition of ϵ, we have $c_1(-\delta+\epsilon-x)/(2\lambda) \le h_a(x) \le c_2(-\delta+\epsilon-x)/(2\lambda)$, $x \in [-\delta, -\delta+\epsilon]$. So

$$\int_{-\delta}^{-\delta+\epsilon} h_a^2 dx \in (c_1^2/12\epsilon^3\lambda^{-2}, c_2^2/12\epsilon^3\lambda^{-2}),$$

$$\int_{-\delta}^{-\delta+\epsilon} g(x)(1-f_a(x))dx \in (c_1^2/12\epsilon^3\lambda^{-1}, c_2^2/12\epsilon^3\lambda^{-1}).$$

$$\begin{aligned} & E[(1-Yf_a(X))_+ 1_{\{-\delta \le X \le 0\}}] + \lambda \int_{-\delta}^{0} (f_a')^2 dx - E[(1-Y)_+ 1_{\{-\delta \le X \le 0\}}] \\ = & \int_{-\delta}^{-\delta+\epsilon} g(x)(1-f_a(x))dx + \lambda \int_{-\delta}^{-\delta+\epsilon} h_a^2 dx \\ \in & (c_1^2\epsilon^3/(6\lambda), c_2^2\epsilon^3/(6\lambda)) \\ \subset & (4/3\lambda^{1/2}(1-a)^{3/2}c_1^2c_2^{-3/2}, 4/3\lambda^{1/2}(1-a)^{3/2}c_2^2c_1^{-3/2}). \end{aligned}$$

$\triangle$.

Proof of Theorem 1: Without loss of generality, assume $x_0 = 0$, and that $p(x)$ is bounded from 1/2 from below. It is easy to see that $|\hat{f}| \le 1$. Denote $a = \hat{f}(-\delta)$, $b = \hat{f}(\delta)$, and $\mathcal{F}_\delta = \{f \in H^1[-\delta, \delta], f(-\delta) = a, f(\delta) = b\}$. Consider problems

$$\min_{f \in \mathcal{F}_\delta} 1/n \sum_{i=1}^{n} [(1 - Y_i f(X_i))_+ 1_{-\delta \le X_i \le \delta}] + \lambda \int_{-\delta}^{\delta} (f')^2 dx \tag{21}$$

$$\min_{f \in \mathcal{F}_\delta} E[(1 - Yf(X))_+ 1_{\{-\delta \le X \le \delta\}}] + \lambda \int_{-\delta}^{\delta} (f')^2 dx \tag{22}$$

Then the restriction of $\hat{f}$ to $[-\delta, \delta]$ is a solution to (21). Let $\bar{f}_\delta$ be the solution to (22), then by Lemma 4.1, for small enough λ, we have $\bar{f}_\delta(x) = 1, \forall x \in (-\delta+\epsilon_1, \delta-\epsilon_2)$; and $\bar{f}_\delta$ strictly

increases from a to 1 in $[-\delta, -\delta+\epsilon_1]$, strictly decreases from 1 to b in $[\delta-\epsilon_2, \delta]$, where $\epsilon_1 < \delta$, and $\epsilon_2 < \delta$.

Denote $1-\hat{f}(0)$ by ω, and $\mathcal{F}_\delta^\omega = \{f \in H^1[-\delta,\delta], f(-\delta)=a, f(\delta)=b, f(0)=1-\omega\}$. Let $\bar{f}_{\delta\omega}$ be the solution to

$$\min_{f\in\mathcal{F}_\delta^\omega} E\{[1-Yf(X)]_+ 1_{\{-\delta\le X\le\delta\}}\} + \lambda\int_{-\delta}^{\delta}(f')^2dx \tag{23}$$

From Lemma 4.1 we can see that for λ small enough, we have $\bar{f}_{\delta\omega}(x)=1, \forall x\in(-\delta+\epsilon_1, -\epsilon_3)\cup(\epsilon_4, \delta-\epsilon_2)$; and $\bar{f}_{\delta\omega}$ strictly increases from a to 1 in $[-\delta, -\delta+\epsilon_1]$, strictly decreases from 1 to b in $[\delta-\epsilon_2,\delta]$, strictly decreases from 1 to $1-\omega$ in $(-\epsilon_3, 0)$, and strictly increases form $1-\omega$ to 1 in $(0,\epsilon_4)$; and $\bar{f}_{\delta\omega}$ is identical to $\bar{f}_\delta$ other than on $[-\epsilon_3,\epsilon_4]$. And

$$E_c[(1-Y\bar{f}_{\delta\omega}(X))_+1_{\{-\delta\le X\le\delta\}}]+\lambda\int_{-\delta}^{\delta}(\bar{f}'_{\delta\omega})^2dx - E_c[(1-Y\bar{f}_\delta(X))_+1_{\{-\delta\le X\le\delta\}}]-\lambda\int_{-\delta}^{\delta}(\bar{f}'_\delta)^2dx \ge c_3\lambda^{1/2}\omega^{3/2}$$

Since $\bar{f}_{\delta\omega}$ is the solution to (23), we get

$$E_c[(1-Y\hat{f}(X))_+1_{\{-\delta\le X\le\delta\}}]+\lambda\int_{-\delta}^{\delta}(\hat{f}')^2dx - E_c[(1-Y\bar{f}_\delta(X))_+1_{\{-\delta\le X\le\delta\}}]-\lambda\int_{-\delta}^{\delta}(\bar{f}'_\delta)^2dx \ge c_3\lambda^{1/2}\omega^{3/2}. \tag{24}$$

On the other hand, the left hand side of (24) is equal to

$$\begin{aligned}
&-1/n\textstyle\sum_{i=1}^n[(1-Y_i\hat{f}(X_i))1_{\{-\delta\le X_i\le\delta\}}] + E_c[(1-Y\hat{f}(X))1_{\{-\delta\le X\le\delta\}}]\\
&+1/n\textstyle\sum_{i=1}^n[(1-Y_i\hat{f}(X_i))_+1_{\{-\delta\le X_i\le\delta\}}] + \lambda\int_{-\delta}^{\delta}(\hat{f}')^2dx\\
&-E_c[(1-Y\bar{f}_\delta(X))1_{\{-\delta\le X\le\delta\}}] - \lambda\int_{-\delta}^{\delta}(\bar{f}'_\delta)^2dx\\
\le\quad &-1/n\textstyle\sum_{i=1}^n[(1-Y_i\hat{f}(X_i))1_{\{-\delta\le X_i\le\delta\}}] + E_c[(1-Y\hat{f}(X))1_{\{-\delta\le X\le\delta\}}]\\
&+1/n\textstyle\sum_{i=1}^n[(1-Y_i\bar{f}_\delta(X_i))_+1_{\{-\delta\le X_i\le\delta\}}] + \lambda\int_{-\delta}^{\delta}(\bar{f}'_\delta)^2dx\\
&-E_c[(1-Y\bar{f}_\delta(X))1_{\{-\delta\le X\le\delta\}}] - \lambda\int_{-\delta}^{\delta}(\bar{f}'_\delta)^2dx\\
=\quad &1/n\textstyle\sum_{i=1}^n[(Y_i(\hat{f}-\bar{f}_\delta)(X_i))1_{\{-\delta\le X_i\le\delta\}}] - E_c[(Y(\hat{f}-\bar{f}_\delta)(X))1_{\{-\delta\le X\le\delta\}}]\\
=\quad &1/n\textstyle\sum_{i=1}^n[(Y_iq(X_i))1_{\{-\delta\le X_i\le\delta\}}] - E_c[(Yq(X))1_{\{-\delta\le X\le\delta\}}],
\end{aligned}$$

where $q=\hat{f}-\bar{f}_\delta\in H^1[-1,1]$. ($\bar{f}_\delta$ is extended to the interval $[-1,1]$ continuously. It is constant in $[-1,-\delta]$ or $[\delta,1]$.) The first inequality comes from the fact that $\hat{f}$ solves (21). Therefore we have

$$E\left[1/n\sum_{i=1}^n[(Y_iq(X_i))1_{\{-\delta\le X_i\le\delta\}}] - E_c[(Yq(X))1_{\{-\delta\le X\le\delta\}}]\right]^2 \ge c_3^2E(\lambda^{1/2}\omega^{3/2})^2, \tag{25}$$

Consider an orthonormal basis $\{\phi_j\}$ in $L_2[-1,1]$, such that

$$\langle \phi_j, \phi_k \rangle_{L^2} = \delta_{jk};$$
$$\langle \phi_j, \phi_k \rangle_{H^1} = \lambda_j \delta_{jk}$$

then $\lambda_j \sim j^2$. See Silverman (1982), or Cox and O'Sullivan (1990), or Lin (2000).

Let q_j be the coefficients of q with respect to $\{\phi_j\}$. Then

$$\begin{aligned}
& 1/n \sum_{i=1}^n [(Y_i q(X_i)) 1_{\{-\delta \le X_i \le \delta\}}] - E_c[(Yq(X)) 1_{\{-\delta \le X \le \delta\}}] \\
= & \sum_j \left\{ q_j \left[1/n \sum_{i=1}^n [(Y_i \phi_j(X_i)) 1_{\{-\delta \le X_i \le \delta\}}] - E[(Y\phi_j(X)) 1_{\{-\delta \le X \le \delta\}}] \right] \right\} \\
\le & \left[\sum_j \lambda_j q_j^2 \right]^{1/2} \left\{ \sum_j \lambda_j^{-1} \left[1/n \sum_{i=1}^n [(Y_i \phi_j(X_i)) 1_{\{-\delta \le X_i \le \delta\}}] - E[(Y\phi_j(X)) 1_{\{-\delta \le X \le \delta\}}] \right]^2 \right\}^{1/2} \\
= & \|q\|_{H^1} \left\{ \sum_j \lambda_j^{-1} \left[1/n \sum_{i=1}^n [(Y_i \phi_j(X_i)) 1_{\{-\delta \le X_i \le \delta\}}] - E[(Y\phi_j(X)) 1_{\{-\delta \le X \le \delta\}}] \right]^2 \right\}^{1/2}
\end{aligned}$$

But we have

$$\begin{aligned}
& E \left[1/n \sum_{i=1}^n [(Y_i \phi_j(X_i)) 1_{\{-\delta \le X_i \le \delta\}}] - E[(Y\phi_j(X)) 1_{\{-\delta \le X \le \delta\}}] \right]^2 \\
\le & 1/n E \left[Y \phi_j(X) 1_{\{-\delta \le X \le \delta\}} \right]^2 \\
\le & 1/n E \left[\phi_j(X) \right]^2 \\
= & 1/n \int_{-1}^1 \phi_j^2(x) d(x) dx \\
\le & D_2/n.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& E \left\{ \sum_j \lambda_j^{-1} \left[1/n \sum_{i=1}^n [(Y_i \phi_j(X_i)) 1_{\{-\delta \le X_i \le \delta\}}] - E[(Y\phi_j(X)) 1_{\{-\delta \le X \le \delta\}}] \right]^2 \right\} \\
\le & D_2/n \sum_j \lambda_j^{-1} \sim 1/n \sum_j j^{-2} \sim 1/n.
\end{aligned}$$

Since $\hat{f}$ is the solution to (11), comparing with zero function, we get $\lambda \int_{-1}^1 (\hat{f}')^2 \le 1$. Since $|\hat{f}| \le 1$, we have $\|\hat{f}\|_{H^1}^2 \le 2 + 1/\lambda$. Also we have shown that

$$\int_{-1}^1 (\bar{f}_\delta')^2 = \int_{-\delta}^{\delta} (\bar{f}_\delta')^2 \le c_4 \lambda^{-1/2}.$$

So we can see $\|q\|_{H^1}^2 \leq 4 + 1/\lambda + c_4\lambda^{-1/2}$. Therefore we get

$$E\left[1/n\sum_{i=1}^{n}[(Y_i q(X_i))1_{\{-\delta\leq X_i\leq\delta\}}] - E_c[(Yq(X))1_{\{-\delta\leq X\leq\delta\}}]\right]^2 \leq c_5\lambda^{-1}/n,$$

Combining the last inequality with (25), we get $E\omega^3 \leq c_6 n^{-1}\lambda^{-2}$. This is a little stronger than the conclusion of Theorem 1. $\triangle$.

Proof of Theorem 2: Let $\bar{f}_M$ be the solution to

$$\min_{\int_{-1}^{1}(f')^2\leq M} E[1 - Yf(X)]_+. \tag{26}$$

We will need the following lemma in the proof.

Lemma 4.2 *For sufficiently large M, we have*

$$\int_{-1}^{1} |sign(p - 1/2) - \bar{f}_M|dx \leq c_7 M^{-(\underline{\alpha}+2)/(\bar{\alpha}+2)} \tag{27}$$

$$l(\bar{f}_M) - l(\eta^*) = \int_{-1}^{1} |sign(p - 1/2) - \bar{f}_M]||2p - 1|d(x)dx \leq c_8 M^{-(\underline{\alpha}+1)}. \tag{28}$$

Furthermore, for any $f \in H^1$ satisfying $|f(x)| \leq 1$, $\forall x \in [-1, 1]$, and fixed $\theta > 0$, we have

$$\int_{-1}^{1}(f - \bar{f}_M)^2 d(x)dx \leq c_9\left\{[l(f) - l(\bar{f}_M)]M^{\theta} + M^{-\rho(\theta)}\right\}. \tag{29}$$

Proof of Lemma 4.2: (26) is equivalent to

$$\min_{f\in H^1} E[1 - Yf(X)]_+ + \lambda\int_{-1}^{1}(f')^2 dx \tag{30}$$

for some $\lambda_{(M)}$ depending on M. It is easy to see $|\bar{f}_M| \leq 1$.

Let us first concentrate on one interval $[r_j, r_{j+1}]$ for some fixed j. Without loss of generality, assume $p(x) > 1/2$ in (r_j, r_{j+1}), and $r_j = -\delta$, $r_{j+1} = \delta$ for some $\delta > \zeta/2$.

Denote $a = \bar{f}_M(-\delta)$, $b = \bar{f}_M(\delta)$, and $\mathcal{F}_\delta = \{f \in H^1[-\delta, \delta], f(-\delta) = a, f(\delta) = b\}$. Then the restriction of $\bar{f}_M$ to the interval $[-\delta, \delta]$ is the solution to the following variational problem:

$$\min_{f\in\mathcal{F}_\delta} E[(1 - Yf(X))_+ 1_{\{-\delta\leq X\leq\delta\}}] + \lambda\int_{-\delta}^{\delta}(f')^2 dx.$$

Now we can follow a proof that is similar to the proof of Lemma 4.1, [the proof is identical to the proof of Lemma 4.1 up to (18). After that use the boundary condition (13)

at r_j and r_{j+1}]. Some tedious by straightforward calculation yields, for M large enough, there exists $\epsilon_1 \in [0,\zeta)$, $\epsilon_2 \in [0,\zeta)$, such that $\epsilon_1 \sim [\lambda(1-a)]^{1/(\alpha_j+2)}$, $\epsilon_2 \sim [\lambda(1-b)]^{1/(\alpha_{j+1}+2)}$; and $\bar{f}_M(x) = 1, \forall x \in (-\delta+\epsilon_1, \delta-\epsilon_2)$; $\bar{f}_\delta$ strictly increases from a to 1 in $[-\delta, -\delta+\epsilon_1]$, strictly decreases from 1 to b in $[\delta-\epsilon_2, \delta]$; and

$$\int_{-\delta}^{-\delta+\epsilon_1} (\bar{f}'_M)^2 dx \sim \lambda^{-1/(\alpha_j+2)}(1-a)^{(2\alpha_j+3)/(\alpha_j+2)} \tag{31}$$

$$\int_{-\delta}^{-\delta+\epsilon_1} (1-\bar{f}_M) dx \sim \lambda^{1/(\alpha_j+2)}(1-a)^{(\alpha_j+3)/(\alpha_j+2)} \tag{32}$$

$$\int_{-\delta}^{-\delta+\epsilon_1} (1-\bar{f}_M)(2p-1) \sim \lambda^{(\alpha_j+1)/(\alpha_j+2)}(1-a)^{(2\alpha_j+3)/(\alpha_j+2)} \tag{33}$$

$$E[(1-Y\bar{f}_M(X))_+ 1_{\{-\delta\le X\le 0\}}] + \lambda\int_{-\delta}^{0}(\bar{f}'_M)^2 dx - E[(1-Y)_+ 1_{\{-\delta\le X\le 0\}}] \sim \lambda^{(\alpha_j+1)/(\alpha_j+2)}(1-a)^{(2\alpha_j+3)/(\alpha_j+2)}$$

Consider the two sides to the cross point r_j, since $\bar{f}_M$ solves (30), we can see that $1-a \sim 1+a \sim 1$. Summing up over all the intervals, we get from (31),

$$\int_{-1}^{1} (\bar{f}'_M)^2 dx \sim \lambda^{-1/(\underline{\alpha}+2)},$$

but the left hand side must be equal to M. Therefore we get $M \sim \lambda^{-1/(\underline{\alpha}+2)}$.

Summing up over all the intervals, we obtain (27) and (28) from (32) and (33).

For (29), we have

$$\begin{aligned}
&\int_{-\delta}^{\delta} (f-\bar{f}_M)^2 d(x) dx \\
=\ & \int_{\substack{[-\delta,\delta] \\ f>\bar{f}_M}} (f-\bar{f}_M)^2 d(x) dx + \int_{\substack{[-\delta,\delta] \\ f<\bar{f}_M}} (f-\bar{f}_M)^2 d(x) dx \\
\le\ & c\left[\int_{\substack{[-\delta,\delta] \\ f>\bar{f}_M}} (1-\bar{f}_M) dx + \int_{\substack{[-\delta,\delta] \\ f<\bar{f}_M}} (\bar{f}_M - f) d(x) dx\right] \\
\le\ & c\left[\int_{-\delta}^{\delta} (1-\bar{f}_M) dx + \int_{\substack{[-\delta,\delta] \\ f<\bar{f}_M}} (\bar{f}_M - f) d(x) dx\right] \\
\le\ & c\left[M^{-(\underline{\alpha}+2)/(\bar{\alpha}+2)} + \int_{\substack{[-\delta,\delta], f<\bar{f}_M \\ p-1/2\le M^{-\theta}}} (\bar{f}_M - f) d(x) dx + \int_{\substack{[-\delta,\delta], f<\bar{f}_M \\ p-1/2> M^{-\theta}}} (\bar{f}_M - f) d(x) dx\right] \\
\le\ & c\left[M^{-\rho(\theta)} + M^{-\theta/\bar{\alpha}} + M^{\theta}\int_{\substack{[-\delta,\delta], f<\bar{f}_M \\ p-1/2> M^{-\theta}}} (\bar{f}_M - f)[2p(x)-1] d(x) dx\right] \qquad (34) \\
\le\ & c\left[M^{-\rho(\theta)} + M^{\theta}[\int_{[-\delta,\delta]} (\bar{f}_M - f)[2p(x)-1] d(x) dx - \int_{\substack{[-\delta,\delta] \\ f>\bar{f}_M}} (\bar{f}_M - f)[2p(x)-1] d(x) dx]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq c\left[M^{-\rho(\theta)} + M^{\theta}[\int_{[-\delta,\delta]}(\bar{f}_M - f)[2p(x)-1]d(x)dx + \int_{[-\delta,\delta]}(1-\bar{f}_M)(2p-1)dx]\right] \\
&\leq c\left[M^{-\rho(\theta)} + M^{\theta}\int_{[-\delta,\delta]}(\bar{f}_M - f)[2p(x)-1]d(x)dx + M^{\theta-(\underline{\alpha}+1)}\right] \\
&\leq c\left[M^{-\rho(\theta)} + M^{\theta}\int_{[-\delta,\delta]}(\bar{f}_M - f)[2p(x)-1]d(x)dx\right].
\end{aligned}
$$

Here (34) follows from the boundary condition (13) in Assumption 2.

Summing up over all the intervals, we get (29). $\triangle$.

We now prove Theorem 2.

In Lemma 1 of Mammen and Tsybakov (1999), put in $Z = (X, Y)$, $z = (x, y)$, where $y \in \{-1, 1\}$, $x \in [-1, 1]$. Let $\mathcal{H} = \{h(z) = -M^{-1/2} y f(x) : f \in H^1, \int_{-1}^{1}(f')^2 dx \leq M; |f(x)| \leq 1, \forall x \in [-1, 1]\}$.

Let $H_B(\delta, \mathcal{H}, P)$ be the δ-entropy with bracketing of $\mathcal{H}$. Let $H_\infty(\delta, \mathcal{H})$ be the δ-entropy of $\mathcal{H}$ for the supremum norm, and $\bar{H}_\infty(\delta, \mathcal{H})$ be the δ-entropy of $\mathcal{H}$ for the supremum norm requiring the centers of the covering balls be in $\mathcal{H}$. For a definition of these concepts, see van de Geer (1999). Define $\mathcal{F} = \{f \in H^1 : \int_{-1}^{1}(f')^2 dx \leq 1; |f(x)| \leq M^{-1/2}, \forall x \in [-1, 1]\}$. Then for any $\delta > 0$, we have

$$
\begin{aligned}
H_B(\delta, \mathcal{H}, P) &\leq H_\infty(\delta/2, \mathcal{H}) && (35)\\
&\leq \bar{H}_\infty(\delta/2, \mathcal{H}) && (36)\\
&= \bar{H}_\infty(\delta/2, \mathcal{F}) && (37)\\
&\leq H_\infty(\delta/4, \mathcal{F}) && (38)\\
&\leq c\delta^{-1}, && (39)
\end{aligned}
$$

where (35) follows from Lemma 2.1 of van de Geer (1999). (36) is by definition. For (37), notice that any function h in $\mathcal{H}$ can be written as $-yf(x)$ with $f \in \mathcal{F}$, and vice versa, and that for $f_1, f_2 \in \mathcal{F}$, we have $|[-yf_1(x)] - [-yf_2(x)]| = |f_1(x) - f_2(x)|$, for any $x \in [-1, 1]$, $y \in \{-1, 1\}$. (38) is easy to check, and (39) is well known. See, for example, Theorem 2.4 of van de Geer (1999).

Write $h_0(z) = -M^{-1/2} y \bar{f}_M(x)$. Then by Lemma 1 of Mammen and Tsybakov (1999), there exists constants $c_{10} > 0$, $c_{11} > 0$ such that

$$Pr\left\{\sup_{h\in\mathcal{H}} \frac{|n^{-1/2}\sum_{i=1}^n\{(h-h_0)(Z_i) - E(h-h_0)(Z_i)\}|}{\{\|h-h_0\|_{L_2(P)} \bigvee n^{-1/3}\}^{1/2}} > c_{10}\nu\right\} \le c_{11}e^{-\nu}$$

for $\nu \ge 1$. Here $a \bigvee b = \max(a,b)$. This is equivalent to

$$Pr\left\{\sup_{\substack{|f|\le 1\\ \int_{-1}^1 (f')^2 dx \le M}} \frac{|n^{-1/2}\sum_{i=1}^n\{Y_i(\bar{f}_M - f)(X_i) - EY_i(\bar{f}_M - f)(X_i)\}|}{\{M^{1/2}\|f-\bar{f}_M\|_{L_2(P)} \bigvee Mn^{-1/3}\}^{1/2}} > c_{10}\nu\right\} \le c_{11}e^{-\nu} \quad (40)$$

for $\nu \ge 1$.

Now define $V_n = n^{1/2}M^{-(1+\theta)/4}\left\{[l(\hat{f}_M) - l(\bar{f}_M)] - [l_n(\hat{f}_M) - l_n(\bar{f}_M)]\right\} / \left\{l(\hat{f}_M) - l(\bar{f}_M)\right\}^{1/4}$.

Since $\hat{f}_M$ solves (12), we have

$$V_n \ge n^{1/2}M^{-(1+\theta)/4}[l(\hat{f}_M) - l(\bar{f}_M)]^{3/4}. \quad (41)$$

Now consider the event $A = \{l(\hat{f}_M) - l(\bar{f}_M) > M^{-(\rho(\theta)+\theta)}\}$. If A holds, then

$$\begin{aligned}
V_n &\le \sup_{\substack{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1;\\ l(f)-l(\bar{f}_M)>M^{-(\rho(\theta)+\theta)}}} n^{1/2}M^{-(1+\theta)/4}\frac{[l(f)-l(\bar{f}_M)] - [l_n(f)-l_n(\bar{f}_M)]}{[l(f)-l(\bar{f}_M)]^{1/4}} \\
&\le \sup_{\substack{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1;\\ l(f)-l(\bar{f}_M)>M^{-(\rho(\theta)+\theta)}}} n^{1/2}M^{-(1+\theta)/4}\frac{|1/n\sum_{i=1}^n\{Y_i(f-\bar{f}_M)(X_i) - EY_i(f-\bar{f}_M)(X_i)\}|}{[l(f)-l(\bar{f}_M)]^{1/4}} \\
&\le \sup_{\substack{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1;\\ l(f)-l(\bar{f}_M)>M^{-(\rho(\theta)+\theta)}}} \frac{|n^{-1/2}\sum_{i=1}^n\{Y_i(f-\bar{f}_M)(X_i) - EY_i(f-\bar{f}_M)(X_i)\}|}{[M^{1+\theta}(l(f)-l(\bar{f}_M))]^{1/4}} \\
&\le \sup_{\substack{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1;\\ l(f)-l(\bar{f}_M)>M^{-(\rho(\theta)+\theta)}}} \frac{|n^{-1/2}\sum_{i=1}^n\{Y_i(f-\bar{f}_M)(X_i) - EY_i(f-\bar{f}_M)(X_i)\}|}{\{M^{1+\theta}[(cM^{-\theta}\|f-\bar{f}_M\|^2_{L_2(P)} - M^{-(\rho(\theta)+\theta)}) \bigvee M^{-(\rho(\theta)+\theta)}]\}^{1/4}} \quad (42) \\
&\le \sup_{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1} \frac{|n^{-1/2}\sum_{i=1}^n\{Y_i(f-\bar{f}_M)(X_i) - EY_i(f-\bar{f}_M)(X_i)\}|}{\{M^{1+\theta}[(cM^{-\theta}\|f-\bar{f}_M\|^2_{L_2(P)}) \bigvee M^{-(\rho(\theta)+\theta)}]\}^{1/4}} \quad (43) \\
&= c \sup_{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1} \frac{|n^{-1/2}\sum_{i=1}^n\{Y_i(f-\bar{f}_M)(X_i) - EY_i(f-\bar{f}_M)(X_i)\}|}{[(M^{1/2}\|f-\bar{f}_M\|_{L_2(P)}) \bigvee M^{(1-\rho(\theta))/2}]^{1/2}} \\
&\le c \sup_{\int_{-1}^1 (f')^2dx\le M; |f(x)|\le 1} \frac{|n^{-1/2}\sum_{i=1}^n\{Y_i(f-\bar{f}_M)(X_i) - EY_i(f-\bar{f}_M)(X_i)\}|}{[(M^{1/2}\|f-\bar{f}_M\|_{L_2(P)}) \bigvee Mn^{-1/3}]^{1/2}}
\end{aligned}$$

where (42) follows from (29), and (43) follows from the fact that $a \bigvee b \geq (a-b) \bigvee b \geq (a \bigvee b)/2$ for any $a > 0$, $b > 0$. The last step follows from that $M_{(n)} \sim n^t$ for some $0 < t \leq 2/[3(1+\rho(\theta))]$.

Therefore it follows from (40) that

$$\limsup_{n\to\infty} E[V_n^s 1_A] \leq C(s), \tag{44}$$

for all $s > 0$ and finite constants $C(s)$ depending on s.

By (41) and (44), we have

$$E\{[l(\hat{f}_M) - l(\bar{f}_M)]^s 1_A\} \leq C(4s/3)n^{-\gamma s} \tag{45}$$

On A^c, we have $l(\hat{f}_M) - l(\bar{f}_M) \leq M^{-(\rho(\theta)+\theta)}$, so

$$E\{[l(\hat{f}_M) - l(\bar{f}_M)]^s 1_{A^c}\} \leq M^{-\gamma s} \tag{46}$$

By the definition of $\rho(\theta)$, we have $\underline{\alpha} + 1 \geq \rho(\theta) + \theta$. Noticing $(a+b)^s \leq 2^s(a^s + b^s)$ for any $a > 0, b > 0$, we see that (45) and (46) combined with (28) gives (14). Then (15) follows directly from the following lemma.

Lemma 4.3 *$R(g) - R(\eta^*) \leq l(g) - l(\eta^*)$ for any function g satisfying $|g(x)| \leq 1$, $\forall x \in [-1, 1]$.*

Proof of Lemma 4.3: we have

$$\begin{aligned} R(g) - R(\eta^*) &= 1/2\{l[sign(g)] - l[sign(2p-1)]\} \\ &= \int_{-1}^{1} 1/2[sign(2p-1) - sign(g)][2p(x) - 1]d(x)dx \\ &\leq \int_{-1}^{1} [sign(2p-1) - g][2p(x) - 1]d(x)dx \\ &= l(g) - l[sign(2p-1)] = l(g) - l(\eta^*). \end{aligned}$$

$\triangle$.

Acknowledgements: This work was partly supported by Wisconsin Alumni Research Foundation. The author would like to thank Professor Grace Wahba for introducing him to the field of the SVM, and for stimulating discussions.

References

[1] Boser, B., Guyon, I. and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Fifth Annual Conference on Computational Learning Theory, Pittsburgh ACM, pp, 142-152.

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

[3] Cox, D.D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics* **18** 1676-1695.

[4] Devroye, L., Györfi, L., and Lugosi, G. (1996). A probabilistic theory of pattern recognition. Springer, New York.

[5] Lin, Y. (1999). Support vector machines and the Bayes rule in classification. University of Wisconsin - Madison technical report 1014. Submitted.

[6] Lin, Y. (2000). Tensor product space ANOVA models. To appear in *Ann. Statist.* **27**.

[7] Lin, Y., Lee, Y. and Wahba, G. (2000). Support vector machines for classification in nonstandard situations. Technical Report 1016. Department of Statistics, University of Wisconsin, Madison. Submitted.

[8] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808-1829.

[9] Marron, J. S. (1983). Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.* **11** 1142-1155.

[10] Osuna, E., Freund, R. and Girosi, F. (1997). An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural networks for signal processing VII — Proceedings of the 1997 IEEE workshop,* pages 276 - 285, New York. IEEE.

[11] Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in kernel methods — Support vector learning*, pages 185 - 208, Cambridge, MA, MIT Press.

[12] Shawe-Taylor, J. and Cristianini, N. (1998). "Robust Bounds on the Generalization from the Margin Distribution". Neuro COLT Technical Report TR-1998-029.

[13] Silverman, B.W. (1982). On the estimation of a probability density function by the maximum penalised likelihood method. *The Annals of Statistics* **10** 795-810.

[14] van de Geer, S. (1999). Empirical processes in M-estimation. Cambridge university press.

[15] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York.

[16] Wahba, G. (1990). Spline Models for Observational Data. Philadelphia, PA: Society for Industrial and Applied Mathematics.

[17] Wahba, G., Lin, Y. and Zhang, H. (1999). *GACV* for support vector machines, or , another way to look at margin-like quantities. To appear in A. J. Smola, P. Bartlett, B. Scholkopf & D. Schurmans (Eds.), Advances in Large Margin Classifiers. Cambridge, MA & London, England: MIT Press.