Statistica Sinica **10**(2000), 1069-1089

ADAPTIVE ESTIMATION IN PATTERN RECOGNITION BY COMBINING DIFFERENT PROCEDURES

Yuhong Yang

Iowa State University

Abstract: We study a problem of adaptive estimation of a conditional probability function in a pattern recognition setting. In many applications, for more flexibility, one may want to consider various estimation procedures targeted at different scenarios and/or under different assumptions. For example, when the feature dimension is high, to overcome the familiar curse of dimensionality one may seek a good parsimonious model among a number of candidates such as CART, neural nets and additive models. For such a situation, one wishes to have an automated final procedure that performs as well as the best candidate.

In this work, we propose a method to combine a countable collection of procedures for estimating the conditional probability. We show that the combined procedure has a property that its statistical risk is bounded above by that of any of the procedure being considered plus a small penalty. Thus asymptotically, the strengths of the different estimation procedures are shared by the combined procedure. A simulation study shows the potential advantage of combining models compared with model selection.

Key words and phrases: Adaptive estimation, conditional probability, logistic regression, minimax-rate adaptation, nonparametric classification.

1. Introduction

In this paper, we study adaptive estimation of a conditional probability function in a pattern recognition setting. For simplicity, consider the two-class case with class labels $Y \in \{0, 1\}$.

Suppose one observes $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, independent copies of a random pair $Z = (X, Y)$. Let $f(x) = P\{Y = 1|X = x)$ be the conditional probability of Y taking label 1 given the feature variable $X = x \in \mathcal{X}$. Here the feature space $\mathcal{X}$ could be of high dimension. We are interested in estimating f.

Besides parametric modelings (e.g., logistic regression), nonparametric methods have been considered. When the dimension of the feature space $\mathcal{X}$ is high, traditional methods (e.g., maximum likelihood estimation based on order selection in a series expansion) often have difficulty estimating exponentially many parameters based on a sample of moderate size, resulting in unsatisfactory performance. Various methods have been proposed, including projection pursuit

(Friedman and Stuetzle (1981)), CART (Breiman, Friedman, Olshen and Stone (1984)), neural networks (e.g., Barron and Barron (1988)), additive models (e.g., Stone (1985), Buja, Hastie and Tibshirani (1989)), tensor-product splines with various interaction orders (e.g., Stone, Hansen, Kooperberg and Truong (1997)) and more. Estimation procedures are developed under different characterizations of the target function.

When a number of estimation procedures are available, how should one be chosen for the data at hand? A solution lies in a good model selection criterion. General results in this direction for estimating the conditional probability include Barron and Cover (1991) using a minimum description length criterion, Barron (1994) using a complexity-penalized criterion, Lugosi and Nobel (1999) using an empirical-complexity based penalization criterion, and Yang (1999c) using a penalized maximum likelihood estimation criterion. These results show that the final estimator automatically adapts to various different characterizations of f.

In this work, we propose a method that allows one to combine a countable collection of procedures for estimating f and derive its adaptation property. We obtain a risk bound for the combined procedure and show that it combines the strengths of the original procedures in the sense that when compared with any of the original procedures, its risk is no more than a small penalty away. Thus the combined procedure works asymptotically as well as the best being considered without knowing which one it is.

The method of combining different procedures here has a connection with data compression in information theory (see, e.g., Barron (1987), Clark and Barron (1990), and Yang and Barron (1999)). Similar adaptation results have been obtained for density estimation and nonparametric regression (with Gaussian errors) in Yang (1996, 2000a, b).

Combining different procedures has been studied in computational learning theory (see, Vovk (1990), Littlestone and Warmuth (1994), Cesa-Bianchi, Freund, Haussler, Shapire and Warmuth (1997), Cesa-Bianchi and Lugosi (1999), and others). The focus is on the worst-case cumulative performance (relative to the best procedure) over all possible sequences of observations without assumptions on how the data are generated. The loss studied in the context of classification is typically $|Y - \hat{f}|$, which is a suitable measure for prediction. In this paper, our interest is in the estimation of the conditional probability under a probabilistic assumption (i.i.d.) on the data. Accordingly the squared L_2 loss $\| \hat{f} - f \|^2$ for estimating f is used. The risk bound, as is well-known, implies a performance bound on the mean error probability for classifying the next Y, which is commonly considered in statistical classification (see Section 5). Compared with the methods mentioned above, for good performance at each sample size (rather than a cumulative fashion, which is suitable for on-line learning), our method involves an averaging over different sample sizes.

The paper is organized as follows. In Section 2, the method of combining a list of procedures is proposed. The adaptation property of the combined procedure is derived in Section 3. In Section 4, we consider minimax-rate adaptation over different classes of conditional probability functions. In Section 5, we give an implication for adaptive classification. The main result is illustrated using an example in Section 6. A computationally feasible algorithm is proposed in Section 7 and a simulation study is presented. The proofs of the results are deferred to Section 8.

2. Method of Adaptation

A procedure for estimating f refers to a sequence of estimators based on observation(s) $Z^1,\ldots, Z^{n-1},\ldots$ respectively.

Let $\Delta = \{\delta_j, j \geq 1\}$ be a collection of estimation procedures for f with δ_j producing an estimator $\hat{f}_{j,i}(x; Z^i)$ based on Z^i. The index set $\{j \geq 1\}$ is allowed to degenerate to a finite set. No restrictive requirement other than a boundedness condition will be placed on the procedures to be combined. For instance, procedure δ_1 may be an automated kernel method, δ_2 may be a logistic regression method, δ_3 may be a neural net method, and so on. Then too, procedure δ_4 may be a method using quadratic splines while procedure δ_5 may be one using cubic splines. Some of the procedures (as δ_1 above) may already be adaptive, in which case the final procedure can provide further adaptation capability.

Let $\underline{\lambda} = \{\lambda_j, j \geq 1\}$ be a set of positive numbers satisfying $\sum_{j\geq 1} \lambda_j = 1$. Here $\underline{\lambda}$ may be viewed as weights (or prior probabilities) of the procedures in Δ. The choice of $\underline{\lambda}$ will be discussed in Section 3.B.

For a given $i_0 \geq 2$, let

$$\beta_{j,i} = \frac{\lambda_j \Pi_{i_0 \leq m \leq i} \hat{f}_{j,m-1}(X_m)^{Y_m} \left(1 - \hat{f}_{j,m-1}(X_m)\right)^{1-Y_m}}{\sum_{l\geq 1} \lambda_l \Pi_{i_0 \leq m \leq i} \hat{f}_{l,m-1}(X_m)^{Y_m} \left(1 - \hat{f}_{l,m-1}(X_m)\right)^{1-Y_m}}, \quad i \geq i_0. \tag{1}$$

Note that $\sum_{j\geq 1} \beta_{j,i} = 1$ and that the $\beta_{j,i}$'s are random, bounded between 0 and 1 depending on Δ and the data Z^i. Then for $i \geq i_0$, let $\tilde{f}_i(x) = \sum_{j\geq 1} \beta_{j,i} \hat{f}_{j,i}(x)$, and for $i = i_0 - 1$, let $\tilde{f}_i(x) = \sum_{j\geq 1} \lambda_j \hat{f}_{j,i}(x)$. These estimators are convex combinations of the original estimators produced by the procedures in Δ, with weights depending on Δ and Z^i. Let δ_* denote this estimation procedure (producing $\{\tilde{f}_n, n \geq 1\}$) with $i_0 = 2$.

A closely related procedure will also be used. For each n, choose an integer N_n with $2 \leq N_n \leq n$ (the role of N_n will be explained in Section 3.B; unless stated otherwise, N_n is of order n). Taking $i_0 = n - N_n + 2$ in the definition of $\beta_{j,i}$'s, we define $\hat{f}_n^*(x) = (1/N_n) \sum_{i=n-N_n+1}^{n} \tilde{f}_i(x)$, our final adaptive estimator of

f at sample size n. Let δ^* denote this procedure (which produces the sequence of estimators $\{\hat{f}_n^*,\ n \geq 1\}$). By construction, it is a convex combination of the original estimators at various sample sizes up to n, with adaptive weights depending on the list Δ and the realization of the training sample.

Finally we point out that the adaptive weights for the combined procedures δ^* and δ_* are based on a connection with universal coding in information theory (see, e.g., Barron (1987), Barron and Cover (1991), and Yang (1996, Lemma 2.6).

3. Adaptation Risk Bound

A. Risk of interest. Let μ denote the (unknown) distribution of the feature random variable X. We measure the loss of an estimator of f in terms of a squared norm. Let $\|\cdot\|_2$ denote the L_2 norm with respect to μ and consider two risks: the average cumulative risk and the individual risk. For a procedure δ producing estimators $\hat{f}_1,\ \hat{f}_2, \ldots$ based on $Z^1, Z^2, \ldots$ respectively, the *individual risk* at sample size n is $R(f;n;\delta) = E \parallel f - \hat{f}_n \parallel_2^2$, where the expectation is taken with respect to Z^n, f being the true conditional probability function. Since f is between 0 and 1, this is well defined. Alternatively, for $n \geq 1$, one can consider the *average cumulative risk* up to sample size n, $R_{seq}(f;n;\delta) = n^{-1}\sum_{i=1}^n E \parallel f - \hat{f}_i \parallel_2^2$.

B. Risk bound for the combined procedure. A technical condition will be used for our first theoretical result. We assume that, for each procedure δ_j in the list Δ, the estimators are uniformly bounded away from 0 and 1 (at least for large samples), i.e., there exists a constant $0 < A_j < 1/2$ such that $\hat{f}_{j,i}(x)$ is always between A_j and $1 - A_j$ for all x and all sample size $i \geq 1$. The constants $\{A_j, j \geq 1\}$ are not required to be uniformly bounded away from zero. A data modification to relax the requirement will be discussed in the next subsection.

Theorem 1. *For a collection of procedures $\Delta = \{\delta_j, j \geq 1\}$ satisfying the above condition, and for any weight $\underline{\lambda}$, we can construct estimation procedures δ_* and δ^* as in Section 2 such that for any underlying conditional probability function f,*

$$R_{seq}(f;n;\delta_*) \leq 2 \inf_j \left(\frac{1}{n} \log \frac{1}{\lambda_j} + A_j^{-2} R_{seq}(f;n;\delta_j) \right), \tag{2}$$

$$R(f;n;\delta^*) \leq 2 \inf_j \left(\frac{1}{N_n} \log \frac{1}{\lambda_j} + A_j^{-2} \sum_{l=n-N_n+1}^{n} R(f;l;\delta_j)/N_n \right). \tag{3}$$

Remark. Note that the adaptation method given in Section 2 does not require the boundness assumption on the procedures. Though risk bounds blow up without such an assumption, the combined procedures may not be bad in practice.

From (2), the average cumulative risk of the adaptive procedure δ_* is automatically bounded by a multiple of the average cumulative risk of each procedure δ_j plus a small penalty. Note that $(1/N_n)\sum_{l=n-N_n+1}^{n} R(f;l;\delta_j)$ is the average accuracy of δ_j between sample sizes $n - N_n + 1$ and n. Ideally one would like to replace it by $R(f;n;\delta_j)$ (the risk of δ_j at sample size n) in the individual risk bound (3), but we suspect that it is not true in general. For most interesting applications, with a proper choice of N_n, $(1/N_n)\sum_{l=n-N_n+1}^{n} R(f;l;\delta_j)$ is bounded like $R(f;n;\delta_j)$ for a reasonable estimation procedure (see, e.g., the proof of Theorem 3). For instance, for a decreasing risk around a polynomial order $n^{-r}\eta(n)$ for some $0 < r \le 1$ and $\eta(n)$ (e.g., $\log n$) being a slowly changing function, $(1/N_n)\sum_{l=n-N_n+1}^{n} R(f;l;\delta_j)$ is indeed of the same order as $n^{-r}\eta(n)$ for any choice of $N_n \le \tau n$ with $0 < \tau < 1$ (the choice of $N_n = n$ results in an extra logarithmic factor for a parametric rate with $r = 1$). If this is the case for good procedures in the collection Δ (which have small risks but not too small weights), one has

$$R(f;n;\delta^*) = O\left(\inf_j \left(\frac{1}{n}\log\frac{1}{\lambda_j} + A_j^{-2} R(f;n;\delta_j)\right)\right). \tag{4}$$

Based on the above discussion, we later informally interpret the risk bound in (3) as if (4) held.

From the above, for individual and average cumulative risks, without knowing which procedures are good for the underlying f, we pay the price of a penalty $(1/n)\log(1/\lambda_j)$ (of order $1/n$) for adaptation over the estimation procedures in Δ.

Now we briefly discuss the roles of N_n and $\underline{\lambda}$. If for a given j, $R(f;l;\delta_j)$ is decreasing in l (as expected for a good estimation procedure), then clearly a larger N_n decreases the penalty term in the risk bound in (3) involving the weight, but increases the main term involving the risk of the procedure. For the familiar cases with $R(f;n;\delta_j)$ decreasing at a polynomial order, any choice of $N_n \sim \tau n$ with $0 < \tau < 1$ yields the right rate of convergence. From a computational point of view, a larger N_n increases the computation time polynomially. Due to averaging over different sample sizes in the construction of δ^*, we expect that the estimator produced with larger N_n to be more robust to the presence of outliers, but this remains to be verified.

For the weight assignment $\underline{\lambda}$, if Δ is a small finite set with a few procedures of interest, a natural choice is a uniform weight on the procedures. For example, if both a parametric model and a nonparametric one are plausible, one can combine them with $\lambda_1 = \lambda_2 = 1/2$. When there are countably many procedures to be combined, one may assign the weights according to a natural or reasonable way to describe the index of the procedures (see, e.g., Barron and Cover (1991), Hall

and Hannan (1988), Rissanen, Speed and Yu (1992), Yang and Barron (1998) for demonstrations of such assignments in related work). Of course the assignment is always subjective to some degree, but when the weights are chosen reasonably it usually does not affect the rate of convergence. In practice, we may assign smaller weights λ_j for more complex estimation procedures. Then the risk bound in (3) is a trade-off between accuracy and complexity.

In principle, one can also use the method in Section 2 to deal with adaptation over estimation procedures indexed by continuous hyper-parameters (e.g., bandwidths for kernels). With the hyper-parameters properly discretized, an adaptive procedure is obtained and the penalty term in the risk bound (3) usually does not affect the rate of convergence for nonparametric estimation.

C. A modification to improve the adaptation risk bound. In Theorem 1, the original procedures are assumed to be bounded away from zero and one. This restriction can be weakened by a technique used earlier in Yang and Barron (1998, 1999). The idea is to modify the data so that the conditional probability is bounded away from 0 and 1 (to avoid a technical difficulty that arises in relating Kullback-Leibler divergence and the chi-square distance).

In addition to the observed i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, let W_i be an independently generated Bernoulli random variable with success probability $1/2$, and let $\widetilde{Y}_i$ be Y_i or W_i with probability $(\rho, 1-\rho)$ for some $1/2 \le \rho < 1$, independently for $i = 1, \ldots, n$. The conditional probability of $\widetilde{Y}_i$ taking value 1 given $X = x$ is $g(x) = \rho f(x) + (1-\rho)/2$. The new function g is bounded below by $(1-\rho)/2 > 0$ and above by $\rho + (1-\rho)/2 < 1$. Applying the procedures in Δ based on the modified data $\widetilde{Z}^n = (X_i, \widetilde{Y}_i)_{i=1}^n$, we have estimators $\hat{g}_{j,i}(x)$ of g. In accordance with the bounds on g, one can project the estimators to that range, i.e., let $\widetilde{g}_{j,i}(x)$ be the minimizer of $\| s - \hat{g}_{j,i} \|_2$ over all function s with $(1-\rho)/2 \le s(x) \le \rho + (1-\rho)/2$. By convexity, the risk of $\widetilde{g}_{j,i}$ is no greater than $\hat{g}_{j,i}$ (see e.g., Yang and Barron (1999)). Then, applying the adaptation recipe in Section 2 to the modified estimators, from Theorem 1 we have an adaptive procedure producing estimator $\hat{g}_n^*$ of g based on $\widetilde{Z}^n$, with $E_{\widetilde{Z}^n}\|\hat{g}_n^* - g\|_2^2$ bounded above in terms of the risks of the procedures in estimating g based on $\widetilde{Z}^n$. Since $f(x) = g(x)/\rho - (1-\rho)/(2\rho)$, let $\hat{f}_{rand}(x) = \hat{g}_n^*(x)/\rho - (1-\rho)/(2\rho)$. Then $\hat{f}_{rand}(x)$ is nonnegative and integrates to one (with respect to x). It is a randomized estimator of f and, due to convexity of the loss, one may take conditional expectation with respect to the randomness in the generated random variables to obtain a nonrandomized estimator with no greater risk. The squared L_2 risk of $\hat{f}_{rand}$ is related to the risk of $\hat{g}_n^*$:

$$E\|f - \hat{f}_{rand}\|_2^2 = \rho^{-2} E_{\widetilde{Z}^n}\|\hat{g}_n^* - g\|_2^2, \tag{5}$$

where the first expectation is taken with respect to both the original randomness in Z^n and that from the generated random variables. Let $\delta_\dagger$ and $\delta^\dagger$ be the

combined procedures based on the modified data corresponding to δ_* and δ^* given in Theorem 1 respectively. We have the following conclusion.

Theorem 2. *For any given $\Delta = \{\delta_j, j \geq 1\}$ and $\underline{\lambda}$, we can construct estimation procedures $\delta_\dagger$ and $\delta^\dagger$ such that for any underlying conditional probability function f, risks are bounded in terms of the risks of the original procedures at $g = \rho f + (1-\rho)/2$:*

$$R_{seq}(f; n; \delta_\dagger) \leq 2\rho^{-2} \inf_j \left(\frac{1}{n} \log \frac{1}{\lambda_j} + \left(\frac{2}{1-\rho} \right)^2 R_{seq}(g; n; \delta_j) \right), \tag{6}$$

$$R\left(f; n; \delta^\dagger\right) \leq 2\rho^{-2} \inf_j \left(\frac{1}{N_n} \log \frac{1}{\lambda_j} + \left(\frac{2}{1-\rho} \right)^2 \sum_{l=n-N+1}^{n} R(g; l; \delta_j)/N_n \right). \tag{7}$$

Remarks. 1. The bigger is ρ, the more we make use of the original data. From a practical point of view, ρ should be taken to be close to 1, but not too close to blow up the risk bounds.

2. The technical difficulty that arises when f is close to 0 or 1, in our derivation of the adaptation risk bound for estimating f, seems not to be a problem when classification alone is concerned, because being close to zero or one are easy cases for classification.

3. In data modification, we generate the W_i's using constant conditional probability 1/2. Alternatively one may use any function uniformly bounded away from zero and one as the conditional probability to generate W_i's. A similar result follows.

For nonparametric estimators, the risks at $g(x) = \rho f(x) + (1-\rho)/2$ and f are usually bounded at the same rate. If this is the case for the procedures in Δ, then (6) and (7) yield the desired adaptive rate of convergence. See the proof of Theorem 3 for an application of this result.

4. Minimax-Rate Adaptation

In addition to the convergence property of a procedure at an individual f, uniform convergence over a class of conditional probability functions, say, $\mathcal{F}$, is also of interest. To that end, the worst-case risk of a procedure over $\mathcal{F}$ can be compared to the minimax risk to assess its performance. The minimax risk also provides an answer to the question of how large the sample size should be to guarantee a certain accuracy for every member in $\mathcal{F}$. In this section, we apply the adaptation risk bounds in Section 3 to obtain minimax-rate adaptive procedures of f over different classes of conditional probability functions.

The minimax risk under the squared L_2 loss for estimating a conditional probability in $\mathcal{F}$ is $R(\mathcal{F}; n) = \min_{\hat{f}} \max_{f \in \mathcal{F}} E \parallel f - \hat{f} \parallel_2^2$, where $\hat{f}$ is over all valid estimators based on Z^n.

The rate of convergence of the above minimax risk (as well as a minimax risk for classification) is studied in Yang (1999b). It is shown that for estimating f, in general, the minimax rate is determined by the massiveness of $\mathcal{F}$ as measured by metric entropy.

A procedure that produces a sequence of estimators of f (at different sample sizes) achieving the minimax rate of convergence is said to be minimax-rate optimal. Let $\{\mathcal{F}_j, j \geq 1\}$ be a collection of classes of conditional probability functions. If a procedure is simultaneously minimax-rate optimal for every $\mathcal{F}_j$, we say it is minimax-rate adaptive over the collection. An important question concerning adaptation then is: Is it possible to construct a minimax-rate adaptive estimator for a general collection of classes of conditional probability functions? We have the following result in the positive direction.

Assume that each class $\mathcal{F}_j$ is convex, containing at least one common member uniformly bounded away from 0 and 1. Otherwise, the function classes could be completely different, e.g., $\mathcal{F}_1$ may consist of all nondecreasing functions between 0 and 1, $\mathcal{F}_2$ may be a neural network class, and some other classes may be Sobolev with various interaction order and smoothness (see Section 6).

For a given class $\mathcal{F}$, if $R(\mathcal{F}; \lfloor n/2 \rfloor)$ and $R(\mathcal{F}; n)$ converge at the same order, we say that the minimax risk of the class $\mathcal{F}$ is *rate-regular*. The familiar rates of convergence $n^{-\alpha}(\log n)^{\beta}$ for some $0 \leq \alpha \leq 1$ and $\beta \in R$ are rate-regular. We assume that each class is rate-regular.

Theorem 3. *Under the above assumptions, we can construct a minimax-rate adaptive procedure for the classes* $\{\mathcal{F}_j, j \geq 1\}$.

Thus under a mild condition, a single procedure can be constructed to automatically perform optimally in terms of rate of convergence for a general collection $\{\mathcal{F}_j, j \geq 1\}$ without knowing which one contains the true conditional probability function.

5. Implication for Classification

For classification, one needs to predict the label of Y according to the feature variable $X = x$. Formally, a classifier κ based on the training data Z^n is a mapping from $\mathcal{X} \times \{\mathcal{X} \times \{0,1\}\}^n$ to $\{0,1\}$. For a given classifier $\kappa = \kappa(x; Z^n)$ based on Z^n, the mean error probability is $EP(Y \neq \kappa(X; Z^n)|Z^n)$. If f were known, this error probability is minimized over all choices of classifiers by a Bayes decision κ^* which predicts Y as class 1 if $f(x) \geq 1/2$ and class 0 otherwise. A risk of interest for studying a classifier is

$$r(f; n; \kappa) = EP\{Y \neq \kappa(X; Z^n)|Z^n)\} - P\{Y \neq \kappa^*(X)\}. \tag{8}$$

It is a measure of performance of κ relative to the Bayes rule.

Given an estimator $\hat{f}$ of f, the plug-in classifier classifies Y as class 1 for x with $\hat{f}(x) \geq 1/2$ and class 0 otherwise. It is well-known (see e.g. Devroye, Györfi, and Lugosi (1996, p.95)) that a plug-in classifier has a risk bound $r(f; n; \kappa) \leq 2(E\|f - \hat{f}\|_2^2)^{1/2}$. It is shown in Yang (1999b) that for most of the familiar function classes (e.g., bounded variation, Sobolev or Besov), if $\hat{f}_n$ is minimax-rate optimal for $\mathcal{F}$, so is the plug-in classifier. For a collection of such classes of conditional probability functions, from Theorems 1 and 2, we have the conclusion that the plug-in classifier based on the combined procedure for estimating f is minimax-rate adaptive for classification.

For the pure purpose of classification, one does not have to estimate the conditional probability function f. Devroye (1988) derived nice risk bounds and convergence properties for an adaptive classifier obtained by selecting among a candidate set based on empirical risk minimization.

6. An Example of High-Dimensional Estimation

When the dimension of the feature variable X is high, one faces the familiar curse of dimensionality: the accuracy of the traditional estimators are often not satisfactory even with a moderate sample size. To overcome the problem, different parsimonious models have been suggested. In applications, not knowing which model is good, adaptivity over a collection of plausible ones is desired. For a demonstration, we consider three types of procedures: data-dependent histograms, neural nets, and splines with various interaction orders and smoothness.

Assume that $\mathcal{X} = [0, 1]^d$ and that the distribution μ of X is dominated by the Lebesgue measure with a density bounded away from zero and infinity.

A. Procedures to be combined.

(i). Histograms using data-dependent partitionings. Let $\Pi_n(Z^n)$ be a data dependent partition of R^d. For a given x, let $\Pi_n[x]$ denote the cell containing x. Then the histogram estimate of the conditional probability function f is defined as

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in \Pi_n[x]\}}}{\sum_{i=1}^n I_{\{X_i \in \Pi_n[x]\}}}.$$

(When the denominator is zero, define $\hat{f}_n(x) = 0$.) Consistency and almost sure convergence results have been established under conditions on the partitioning (see, Stone (1974), Devroye and Györfi (1983), Gordon and Olshen (1984), Breiman, Friedman, Olshen and Stone (1984), Nobel (1996) and others). A tree-structured partitioning results in a regression tree. When the partition is properly carried out, the procedure is consistent for every conditional probability function f. Let δ_1 denote such a histogram procedure.

(ii). Neural nets. Consider feedforward neural network models with one layer of sigmoidal nonlinearities, which have the form $f_k(x, \theta) = \sum_{i=1}^k c_i \sigma(v_i \cdot x + b_i) + c_0$.

The function is parametrized by θ, consisting of $v_i \in R^d, b_i, c_i \in R$ with the restriction that $0 \le f_k(x,\theta) \le 1$. Here σ is a given sigmoidal function with $\| \sigma \|_\infty \le 1$, $\lim_{z\to\infty}\sigma(z) = 1$ and $\lim_{z\to-\infty}\sigma(z) = 0$. A model selection criterion can be used to select k. Consider, for example, the penalized maximum likelihood estimation using a criterion studied in Yang (1999c). Let δ_2 denote this neural net procedure.

(iii). Tensor-product spline models. Let $\varphi_{m,q,1}(x), \ldots, \varphi_{m,q,m}(x)$ be the B-spline basis on [0,1]. For $1 \le r \le d$, let $J_r = (j_1, \ldots, j_r)$ $(j_1 < j_2 < \cdots < j_r)$ be an ordered vector of elements from $\{1, \ldots, d\}$, and let $\mathcal{J}_r$ denote the set of all possible such choices. Let $\mathbf{x}_{J_r} = (x_{j_1}, \ldots, x_{j_r})$ be the subvecter of $\mathbf{x}$ with subscripts in J_r. Let $\mathbf{m}_r = (m_1, \ldots, m_r)$ and $\mathbf{q}_r = (q_1, \ldots, q_r)$ be vectors of integers. Let $\mathbf{i}_r = (i_1, \ldots, i_r)$, with $1 \le i_l \le m_l, 1 \le l \le r$. Then given the spline order $\mathbf{q}_r$ and $\mathbf{m}_r$, the tensor products

$$\{\varphi_{\mathbf{i}_r}(\mathbf{x}_{J_r}) = \Pi_{l=1}^r \varphi_{m_l,q_l,i_l}(x_{j_l}) : J_r \in \mathcal{J}_r; 1 \le i_l \le m_j \text{ for } 1 \le l \le r\} \qquad (9)$$

have interaction order $r-1$.

For each choice of $I = (r, \mathbf{q}_r, \mathbf{m}_r)$, consider the family of linear combinations of the splines in (9). Based on a model selection criterion, a penalized maximum likelihood estimation procedure is shown in Yang (1999c) to be minimax-rate adaptive over Sobolev classes (see below). Let δ_3 denote this spline procedure.

B. Related function classes.

(i). Neural network classes. Let $N(C)$ be the closure in $L_2[0,1]^d$ of the set of all functions $g : R^d \to [0,1]$ of the form $g(x) = c_0 + \sum_{i\ge1} c_i\sigma(v_i \cdot x + b_i)$, with $|c_0| + \sum_{i\ge1}|c_i| \le C$, and $\|v_i\| = 1$, where σ is a sigmoidal function. The minimax rate of $N(C)$ under squared L_2 loss is shown in Yang (1999b) to be slightly better than $n^{-1/2}$ for large d. Neural net estimators, including δ_3 with the number of nodes (k) selected by a criterion, are shown to converge at the rate $(\log n/n)^{1/2}$ (see, e.g., Barron (1994), McCaffrey and Gallant (1994), Barron, Birgé and Massart (1999), and Yang (1999c)). This rate is independent of d and is close to the optimal rate of convergence when d is large.

(ii). Sobolev classes with various interaction-order and smoothness. For $r \ge 1$, let $\mathbf{z}_r = (z_1, \ldots, z_r) \in [0,1]^r$. For $\mathbf{k} = (k_1, \ldots, k_r)$ with nonnegative integer components k_i, define $|\mathbf{k}| = \sum_{i=1}^r k_i$. Let $D^{\mathbf{k}}$ denote the differentiation operator $D^{\mathbf{k}} = \partial^{|\mathbf{k}|}/\partial z_1^{k_1} \cdots \partial z_r^{k_r}$. For an integer α, define the Sobolev norm $\| g \|_{W_2^{\alpha,r}} = \| g \|_2 + \sum_{|\mathbf{k}|=\alpha} \int_{[0,1]^r} |D^{\mathbf{k}} g|^2 d\mathbf{z}_r$. Let $W_2^{\alpha,r}(C)$ denote the set of all functions g on $[0,1]^r$ with $\| g \|_{W_2^{\alpha,r}} \le C$. Consider the following function classes of different interaction orders and smoothness:

$$S_1(\alpha;C) = \{\textstyle\sum_{i=1}^d g_i(x_i) : g_i \in W_2^{\alpha,1}(C), 1 \le i \le d\},$$
$$S_2(\alpha;C) = \{\textstyle\sum_{1\le i<j\le d} g_{i,j}(x_i, x_j) : g_{i,j} \in W_2^{\alpha,2}(C), 1 \le i < j \le d\},$$
$$\cdots$$
$$S_d(\alpha;C) = W_2^{\alpha,d}(C),$$

with $\alpha \geq 1$ and $C > 0$. (Since f is bounded between zero and one, we restrict attention to members in that range accordingly.) The simplest class $S_1(\alpha; C)$ contains additive functions (no interaction), and with larger r, functions in $S_r(\alpha; C)$ have higher order interactions. The minimax rate of convergence under squared L_2 loss for estimating $f \in S_r(\alpha; C)$ is shown in Yang (1999c) to be $n^{-2\alpha/(2\alpha+r)}$ for $1 \leq r \leq d$ (cf. Stone (1994) and Nicoleris and Yatracos (1997)). Note that the convergence rate depends on the interaction order but not on the input dimension d, as suggested by the heuristic dimensionality reduction principle of Stone (1985). Thus low order interaction classes are worth exploring for better accuracy. Stone et al (1997) propose a general adaptive spline methodology by model selection for function estimation including estimating the conditional probability. However, it remains to be seen in theory if the methods have the intended adaptation capability.

In Yang (1999c), it is shown that squared L_2 risk of the procedure δ_3 (the penalized maximum likelihood estimation based on model selection) is automatically bounded above by the right order $n^{-2\alpha/(2\alpha+r)}$ for each class $S_r(\alpha; C)$ without knowing α and r.

C. Adaptation property of the combined procedure. By combining the three procedures δ_1, δ_2, and δ_3 with a uniform weight ($\lambda_1 = \lambda_2 = \lambda_3 = 1/3$) and with a data modification as used for Theorem 2, we have a single procedure, say δ^*, with the following adaptation property.

Theorem 4. *The procedure δ^* has risk bounded simultaneously as follows:* $\sup_{f \in S_r(\alpha;C)} R(f; n; \delta^*) = O(n^{-2\alpha/(2\alpha+r)})$ *for all* $1 \leq r \leq d$, $\alpha \geq 1$*;* $\sup_{f \in N(C)} R(f; n; \delta^*) = O(\frac{\log n}{n})^{\frac{1}{2}}$. *In addition, δ^* is consistent for every conditional probability function f.*

From this result, the combined procedure is minimax-rate adaptive over the Sobolev classes, it achieves a rate close to the optimal (when d is large) for the neural network class, and it is consistent for every conditional probability function.

7. A Computationally Feasible Algorithm

A. ACM algorithm. So far, we have concentrated on theoretical development of adaptive pattern recognition. In the formulation of the adaptation method in Section 2, we allow countably many procedures to be combined. This is clearly impossible to implement in practice within a finite time. If the complexities of the procedures in a list Δ are well understood, one could use an appropriate rule to exclude procedures that are too complex at the given sample size. For instance, if one considers various finite-dimensional models based on approximations of the conditional probability function f with n observations, one could not fit models

with more than n parameters and they can be ruled out. In general, if we have a reasonable way of assigning the weights λ_j for the procedures, then a sensible way to avoid infinite computation time is to rule out procedures with weights sufficiently small relative to the sample size (e.g., less than $n^{-\rho}$ for some $\rho > 0$). Now with n given, assume that we have a finite number of procedures to be combined.

Notice that the adaptive estimator given in Section 2 depends on the order of observations. Since the observations are assumed to be i.i.d., the order does not contain useful information for estimating f. Under the (convex) squared L_2 loss, the estimator can be improved by taking the conditional expectation given the observations (ignoring the order). That is, one should permute the order of the observations in all possible ways and compute the average of the resulting estimators. This, however, is computationally prohibitive due to the large number of permutations, even for a small sample size. We next propose an algorithm compromising appropriately between the theoretical risk bound and computational feasibility.

First, instead of permuting the order of observations in all possible ways, we can randomly permute a manageably large number of times and compute the average of the corresponding estimators. Secondly, to save computation time, we do not update the estimators sequentially after each new observation (which is used in constructing $\hat{f}_n^*$ in Section 2). In addition, instead of averaging over different sample sizes as used for computing $\hat{f}_n^*$, we take the last estimator. We call the algorithm ACM (**A**daptive **C**lassification by **M**ixing). For simplicity, assume n is even and there are not too many procedures to be combined. We simply assign a uniform weight to the procedures. Let $\Delta = \{\delta_1, \ldots, \delta_K\}$ denote the procedures.

Algorithm ACM

- *Step 1.* Randomly permute the order of observations. Split the data into two parts $Z^{(1)} = (X_i, Y_i)_{i=1}^{n/2}$ and $Z^{(2)} = (X_i, Y_i)_{i=n/2+1}^{n}$.
- *Step 2.* Estimate f by $\hat{f}_{k,n/2}$ based on $Z^{(1)}$ for each procedure $\delta_k \in \Delta$.
- *Step 3.* Assign the weights.

$$W_k = \frac{\Pi_{n/2+1\le m\le n}\hat{f}_{k,n/2}(X_m)^{Y_m}\left(1-\hat{f}_{k,n/2}(X_m)\right)^{1-Y_m}}{\sum_{l\ge 1}\Pi_{n/2+1\le m\le n}\hat{f}_{l,n/2}(X_m)^{Y_m}\left(1-\hat{f}_{l,n/2}(X_m)\right)^{1-Y_m}}$$

- *Step 4.* Repeat the above steps $(M-1)$ times. Average the weights over the M permutations. Denote the weights by $\hat{W}_k$, $1 \le k \le K$.

- *Step 5.* Compute the convex combination of the estimators $\hat{f}_{k,n}$ for $1 \le k \le K$: $\hat{f}_n(x) = \sum_{k=1}^{K} \hat{W}_k \hat{f}_{k,n}(x)$.

When the number of permutations M is suitably large, the final estimator $\hat{f}_n$ is stable. One could use an objective rule to decide how large M should be. For instance, stop permuting when the estimator meets a convergence criterion. Notice that in the assignment of the weights W_k in Step 3, the same estimator $\hat{f}_{k,n/2}$ (instead of $\hat{f}_{k,i}$ for $n/2+1 \le i \le n$ as in Section 2) is used for each k without updating. This reduces the computation dramatically. In the meantime, as we show next, we do not sacrifice much in terms of the risk bound.

For $n/2+1 \le i < n$, define $\hat{f}_i(x) = \sum_{k=1}^{K} \hat{W}_{k,i} \hat{f}_{k,n/2}(x)$, where $\hat{W}_{k,i}$ is the average of $W_{k,i}$ over the M permutations with

$$W_{k,i} = \frac{\Pi_{n/2+1 \le m \le i} \hat{f}_{k,n/2}(X_m)^{Y_m} \left(1 - \hat{f}_{k,n/2}(X_m)\right)^{1-Y_m}}{\sum_{l \ge 1} \Pi_{n/2+1 \le m \le i} \hat{f}_{l,n/2}(X_m)^{Y_m} \left(1 - \hat{f}_{l,n/2}(X_m)\right)^{1-Y_m}}.$$

Corollary 1. *Assume that the boundness condition used for Theorem* 1 *is satisfied. Then the average cumulative risk of the estimators* $\hat{f}_i$ *satisfies*

$$\frac{1}{n/2} \sum_{i=n/2+1}^{n} E \parallel f - \hat{f}_i \parallel^2$$
$$\le 2 \inf_{1 \le k \le K} \left(\frac{1}{n/2} \log K + A_k^{-2} \left(\frac{n-2}{n} E \parallel f - \hat{f}_{k,n/2} \parallel^2 + \frac{2}{n} E \parallel f - \hat{f}_{k,n} \parallel^2 \right) \right)$$

Remarks. 1. Note that for a good procedure δ_k, the risk $E \parallel f - \hat{f}_{k,l} \parallel^2$ decreases as the sample size l increases. Then the quantity $\frac{n-2}{n} E \parallel f - \hat{f}_{k,n/2} \parallel^2 + \frac{2}{n} E \parallel f - \hat{f}_{k,n} \parallel^2$ is worse than $\sum_{l=n/2+1}^{n} E \parallel f - \hat{f}_{k,l} \parallel^2$ (as would appear if one uses the method in Section 2). If $\hat{f}_{k,n/2}$ and $\hat{f}_{k,n}$ converge at the same rate, as is usually the case for both parametric and nonparametric estimations, then the simplified algorithm does not affect the rate of convergence while it dramatically reduces the computation time.

2. As in Theorem 1, one could average the estimators $\hat{f}_i$ for $n/2+1 \le i < n$ to get an estimator with guaranteed good individual risk bound, but that significantly increases the computation time.

3. Using a treatment similar to that in Section 3.C, the boundness condition on the procedures can be relaxed.

The algorithm ACM can be computationally intensive if M is large and/or there are a large number of procedures to be combined.

B. A simulation study. The following is a simulation study intended to provide some understanding of the actual performance of ACM in a simple setting.

We consider logistic regression with a number of feature variables. We compare ACM with familiar model selection criteria AIC, BIC and cross-validation.

Let $X = (X_1, \ldots, X_8)$ be a feature vector with independent components all uniformly distributed on $[-1, 1]$. Consider nested logistic regression families

$$f_k(x, \theta) = \frac{e^{\theta_0 + \sum_{i=1}^{k} \theta_i x_i}}{1 + e^{\theta_0 + \sum_{i=1}^{k} \theta_i x_i}}$$

for $1 \le k \le 8$. It is assumed that the true conditional probability function f is in one of these families. The criteria AIC (Akaike (1973)) and BIC (Schwartz (1978)) select the model that minimizes $-\text{loglikelihood} + k$ and $-\text{loglikelihood} + k/2 \log n$ respectively. For the leave-one-out cross-validation (see, e.g., Stone (1974)), for a given family, each observation is removed once and the estimator of f based on the remaining data is used to classify the case being removed while the error rate of misclassification is recorded. The family that minimizes this test error rate is selected. Differently from these model selection criteria, ACM combines the families rather than selecting one. The number of random permutations M is chosen to be 100 for ACM.

Regarding the difference between selecting and combining (or mixing) models, it seems intuitively clear that when the models are hard to distinguish (e.g., due to the small sample size relative to the number of models), selection causes a larger variability compared with combining the models appropriately. For such a case, ACM is likely to perform better than the model selection criteria. This has been demonstrated in the context of parametric regression (see Yang (1999d)). In the study here, the sample size is chosen to be $n = 100$, which is not very large relative to the number of models (i.e., 8) being considered. Two cases are chosen with different numbers of parameters in the true models. For the first one, $f(x) = (e^{1+0.8x_1+0.5x_2})/(1 + e^{1+0.8x_1+0.5x_2})$, and for the second, $f(x) = (e^{1+0.8x_1+0.5x_2+0.9x_3+0.4x_4+0.2x_5})/(1 + e^{1+0.8x_1+0.5x_2+0.9x_3+0.4x_4+0.2x_5})$.

Both squared L_2 loss for estimating f and the probability of misclassifying a future case (error probability (EP)) are considered as performance measures in the simulation study. In addition, L_1 loss and logarithmic loss are considered for comparison. Here the risk under logarithmic loss is $E \int (f(x) \log(f(x)/\hat{f}(x)) + (1 - f(x)) \log((1 - f(x))/(1 - \hat{f}(x))))\mu(dx)$. Five thousand additional observations are drawn from the true underlying distribution to simulate these quantities, and one hundred replications are used to simulate the corresponding risks. The numbers in the parentheses in Table 1 are the corresponding standard errors.

From Table 1, it is clear that ACM outperforms the model selection criteria for the estimation of f under all loss functions. For the two cases, compared with the model selection criteria, the risk of ACM is reduced respectively by at least

15% and 32% under squared L_2 loss, by at least 9% and 17% under L_1 loss, and by at least 17% and 39% under logarithmic loss. For classification risk, ACM and BIC perform similarly and significantly better than AIC and Cross-validation for both cases.

Table 1. Comparing ACM with AIC, BIC and cross-validation in logistic regression.

	Case 1				Case 2			
	L_2 Sq	L_1	Log-Loss	EP	L_2 Sq	L_1	Log-Loss	EP
AIC	0.0109 (0.0009)	0.0786 (0.0034)	0.0322 (0.0032)	0.3171 (0.0041)	0.0143 (0.0007)	0.0921 (0.0026)	0.0410 (0.0025)	0.3467 (0.0038)
BIC	0.0073 (0.0005)	0.0675 (0.0019)	0.0193 (0.0013)	0.3061 (0.0038)	0.0169 (0.0006)	0.1025 (0.0019)	0.0452 (0.0017)	0.3305 (0.0045)
CV	0.0103 (0.0009)	0.0762 (0.0031)	0.0286 (0.0027)	0.3170 (0.0040)	0.0157 (0.0007)	0.0974 (0.0022)	0.0425 (0.0019)	0.3371 (0.0044)
ACM	0.0062 (0.0005)	0.0612 (0.0021)	0.0161 (0.0012)	0.3092 (0.0036)	0.0097 (0.0005)	0.0760 (0.0020)	0.0251 (0.0013)	0.3302 (0.0037)

ACM and cross-validation are more computer intensive compared with AIC and BIC. With $M = n = 100$, ACM and cross-validation take about the same computing time. Note that the standard errors associated with ACM ($M = 100$) are all smaller compared with the model selection criteria.

8. Proof of the Results

Proof of Theorem 1. Let $K_f(x, y) = f(x)^y(1 - f(x))^{1-y}$ denote the joint density of (X, Y) with conditional probability function f with respect to the product measure $\mu \otimes \nu$, where ν is the counting measure on {0,1}. For $y = 0$ and 1, we have $K_{\widetilde{f}_i}(x, y) = \sum_{j\geq 1} \beta_{j,i} K_{\hat{f}_{j,i}}(x, y)$. Thus for $i \geq i_0$,

$$K_{\widetilde{f}_i}(x_{i+1}, y_{i+1})$$
$$= \frac{\sum_{j\geq 1} \lambda_j \left(\Pi_{i_0 \leq m \leq i} \hat{f}_{j,m-1}(x_m)^{y_m} \left(1-\hat{f}_{j,m-1}(x_m)\right)^{1-y_m} \right) \hat{f}_{j,i}(x_{i+1})^{y_{i+1}} \left(1-\hat{f}_{j,i}(x_{i+1})\right)^{1-y_{i+1}}}{\sum_{l\geq 1} \lambda_l \Pi_{i_0 \leq m \leq i} \hat{f}_{l,m-1}(x_m)^{y_m} \left(1-\hat{f}_{l,m-1}(x_m)\right)^{1-y_m}}.$$

As a consequence, together with the definition of $\tilde{f}_{i_0-1}$, we have

$$\Pi_{i=i_0-1}^{n} K_{\widetilde{f}_i}(x_{i+1}, y_{i+1}) = \sum_{j\geq 1} \lambda_j \Pi_{i=i_0-1}^{n} \hat{f}_{j,i}(x_{i+1})^{y_{i+1}} \left(1 - \hat{f}_{j,i}(x_{i+1})\right)^{1-y_{i+1}}. \tag{10}$$

For $2 \leq i_0 \leq n$, we have

$$\sum_{i=i_0-1}^{n} E \int K_f(x, y) \log \frac{K_f(x, y)}{K_{\widetilde{f}_i}(x, y)} \mu \otimes \nu(dxdy)$$

$$
\begin{aligned}
&= \sum_{i=i_0-1}^{n} E \int K_f(x_{i+1}, y_{i+1}) \log \frac{K_f(x_{i+1}, y_{i+1})}{K_{\widetilde{f}_i}(x_{i+1}, y_{i+1})} \mu \otimes \nu(dx_{i+1} dy_{i+1}) \\
&= \sum_{i=i_0-1}^{n} E \int \Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1}) \log \frac{K_f(x_{i+1}, y_{i+1})}{K_{\widetilde{f}_i}(x_{i+1}, y_{i+1})} \mu \otimes \nu(dx_{i_0} dy_{i_0}) \\
&\qquad \cdots \mu \otimes \nu(dx_{n+1} dy_{n+1}) \\
&= E \int \Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1}) \log \frac{\Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1})}{\Pi_{i=i_0-1}^{n} K_{\widetilde{f}_i}(x_{i+1}, y_{i+1})} \mu \otimes \nu(dx_{i_0} dy_{i_0}) \\
&\qquad \cdots \mu \otimes \nu(dx_{n+1} dy_{n+1}) \\
&= E \int \Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1}) \log \\
&\qquad \frac{\Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1})}{\sum_{j \geq 1} \lambda_j \Pi_{i=i_0-1}^{n} \hat{f}_{j,i}(x_{i+1})^{y_{i+1}} \left(1 - \hat{f}_{j,i}(x_{i+1})\right)^{1-y_{i+1}}} \\
&\qquad \mu \otimes \nu(dx_{i_0} dy_{i_0}) \cdots \mu \otimes \nu(dx_{n+1} dy_{n+1}) \\
&\leq E \int \Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1}) \log \frac{\Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1})}{\lambda_j \Pi_{i=i_0-1}^{n} \hat{f}_{j,i}(x_{i+1})^{y_{i+1}} \left(1 - \hat{f}_{j,i}(x_{i+1})\right)^{1-y_{i+1}}} \mu \\
&\qquad \otimes \nu(dx_{i_0} dy_{i_0}) \cdots \mu \otimes \nu(dx_{n+1} dy_{n+1}) \\
&\leq \log(1/\lambda_j) + E \int \Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1}) \log \\
&\qquad \frac{\Pi_{i=i_0-1}^{n} K_f(x_{i+1}, y_{i+1})}{\Pi_{i=i_0-1}^{n} \hat{f}_{j,i}(x_{i+1})^{y_{i+1}} \left(1 - \hat{f}_{j,i}(x_{i+1})\right)^{1-y_{i+1}}} \\
&\qquad \mu \otimes \nu(dx_{i_0} dy_{i_0}) \cdots \mu \otimes \nu(dx_{n+1} dy_{n+1}) \\
&= \log(1/\lambda_j) + \sum_{i=i_0-1}^{n} E \int K_f(x, y) \log \frac{K_f(x, y)}{K_{\hat{f}_{j,i}}(x, y)} \mu \otimes \nu(dx dy),
\end{aligned}
$$

where, for the fourth equality, we use equation (10); for the first inequality, we use the fact that $\log(x)$ is an increasing function; and for the last step, we use the relationships for the first four equalities but in the reverse direction. Let $D(p \parallel q) = \int p \log(p/q)$ and $d_H^2(p, q) = \int \left(\sqrt{p} - \sqrt{q}\right)^2$ denote the Kullback-Leibler (K-L) divergence and the squared Hellinger distance between two densities p and q respectively. We now bound the K-L divergence $D(K_f \parallel K_{\hat{f}_{j,i}})$ in terms of the L_2 distance:

$$
\begin{aligned}
&\int K_f(x, y) \log \frac{K_f(x, y)}{K_{\hat{f}_{j,i}}(x, y)} \mu \otimes \nu(dx dy) \\
&= \int \left(f(x) \log \frac{f(x)}{\hat{f}_{j,i}(x)} + (1 - f(x)) \log \frac{1 - f(x)}{1 - \hat{f}_{j,i}(x)} \right) \mu(dx)
\end{aligned}
$$

$$\leq \int \left(\frac{\left(f(x) - \hat{f}_{j,i}(x)\right)^2}{\hat{f}_{j,i}(x)} + \frac{\left(f(x) - \hat{f}_{j,i}(x)\right)^2}{1 - \hat{f}_{j,i}(x)} \right) \mu(dx)$$
$$\leq \frac{1}{A_j^2} \parallel f - \hat{f}_{j,i} \parallel_2^2,$$

where the first inequality follows from the familiar bound on K-L divergence by chi-square distance, i.e., $\int p \log (p/q) \leq \int (p-q)^2 / q$ for densities p and q, and the second inequality follows from the boundness assumption on the procedures. Thus we have

$$\sum_{i=i_0-1}^{n} ED(K_f \parallel K_{\widetilde{f}_i}) \leq \log(1/\lambda_j) + \frac{1}{A_j^2} \sum_{i=i_0-1}^{n} E \parallel f - \hat{f}_{j,i} \parallel_2^2$$

for each j. Since the squared Hellinger distance is always upper bounded by the K-L divergence, we have

$$\sum_{i=i_0-1}^{n} E d_H^2(K_f, K_{\widetilde{f}_i}) \leq \inf_{j \geq 1} \left(\log(1/\lambda_j) + \frac{1}{A_j^2} \sum_{i=i_0-1}^{n} E \parallel f - \hat{f}_{j,i} \parallel_2^2 \right). \tag{11}$$

Note that $d_H^2(K_f, K_g) = \int d_Y^2(f(x), g(x)) \mu(dx)$, where

$$d_Y^2(f(x), g(x)) = \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 + \left(\sqrt{1 - f(x)} - \sqrt{1 - g(x)} \right)^2$$
$$\geq \frac{1}{4}(f(x) - g(x))^2 + \frac{1}{4}(1 - f(x) - (1 - g(x)))^2 = \frac{1}{2}(f(x) - g(x))^2.$$

(For the above inequality, we use the fact that f and g are upper bounded by 1.) As a consequence, we have

$$\sum_{i=i_0-1}^{n} E \parallel f - \tilde{f}_i \parallel_2^2 \leq 2 \inf_{j \geq 1} \left(\log(1/\lambda_j) + \frac{1}{A_j^2} \sum_{i=i_0-1}^{n} E \parallel f - \hat{f}_{j,i} \parallel_2^2 \right).$$

Taking $i_0 = 2$, the above inequality yields the cumulative risk bound (2). For the individual risk bound, taking $i_0 = n - N_n + 2$ and, by convexity of squared L_2 loss, we have

$$E \parallel f - \hat{f}_n^* \parallel_2^2 \leq \frac{1}{N_n} \sum_{i=n-N_n+1}^{n} E \parallel f - \tilde{f}_i \parallel_2^2$$
$$\leq 2 \inf_{j \geq 1} \left(\frac{\log(1/\lambda_j)}{N_n} + \frac{1}{A_j^2 N_n} \sum_{i=n-N_n+1}^{n} E \parallel f - \hat{f}_{j,i} \parallel_2^2 \right).$$

This completes the proof of Theorem 1.

Proof of Theorem 3. Let $\mathcal{R}_j(n)$ denote the minimax squared L_2 risk of the class $\mathcal{F}_j$, i.e., $\mathcal{R}_j(n) = \min_{\hat{f}} \max_{f\in\mathcal{F}_j} E \parallel f - \hat{f} \parallel_2^2$, where the minimum is taken over all estimators based on Z^n. For each class $\mathcal{F}_j$, let δ_j be a minimax-rate optimal procedure, i.e., $\sup_{f\in\mathcal{F}_j} R(f;n;\delta_j) \le C_j\mathcal{R}_j(n)$ for some constant $C_j > 1$. From the assumptions, there exists a function $f_0 \in \mathcal{F}_j$ uniformly bounded away from zero and one. We modify the data as in Theorem 2 with, for instance, $\rho = 1/2$, $N_n = n/2$ (ignore rounding), and $\lambda_j = c/j^2$ ($c = \sum_{j\ge1} j^{-2}$). Instead of generating W_i's using the constant conditional probability $1/2$ in the derivation of Theorem 2, we generate W_i using the conditional probability $f_0(X_i)$ at X_i for $1 \le i \le n$. A result similar to Theorem 2 then holds. For the following proof, for simplicity, we assume $f_0 = 1/2$ is in each $\mathcal{F}_j$. The proof for a general f_0 is similar.

Let $g_f = f/2 + 1/2$. By Theorem 2, we have a combined procedure $\delta^\dagger$ such that for each $j^* \ge 1$ and every $f \in \mathcal{F}_{j^*}$,

$$\begin{aligned} R\left(f;n;\delta^\dagger\right) &\le 8 \inf_j \left(\frac{2}{n}\log\frac{1}{\lambda_j} + 32 \sum_{l=n/2}^{n} R\left(g_f;l;\delta_j\right)/n \right) \\ &\le 8 \left(\frac{2}{n}\log\frac{1}{\lambda_{j^*}} + 32 \sum_{l=n/2}^{n} \sup_{g\in\mathcal{F}_{j^*}} R\left(g;l;\delta_{j^*}\right)/n \right) \\ &\le 8 \left(\frac{2}{n}\log\frac{1}{\lambda_{j^*}} + 32 \sum_{l=n/2}^{n} C_{j^*}\mathcal{R}_{j^*}(l)/n \right) \\ &\le 8 \left(\frac{2}{n}\log\frac{1}{\lambda_{j^*}} + 16 C_{j^*}\mathcal{R}_{j^*}(n/2) \right). \end{aligned}$$

For the second inequality, we use the fact that g_f is in $\mathcal{F}_{j^*}$ under the convexity assumption; for the third inequality, we use the fact that δ_{j^*} is minimax-rate optimal; for the last inequality, we use the fact that the minimax risk $\mathcal{R}_{j^*}(l)$ is nonincreasing in l. As a consequence,

$$\sup_{f\in\mathcal{F}_{j^*}} R\left(f;n;\delta^\dagger\right) \le 8 \left(\frac{2}{n}\log\frac{1}{\lambda_{j^*}} + 16 C_{j^*}\mathcal{R}_{j^*}(n/2) \right).$$

Under the rate-regular assumption on the classes, $\mathcal{R}_{j^*}(n/2)$ is of the same order as $\mathcal{R}_{j^*}(n)$. The penalty $(2/n)\log(1/\lambda_{j^*})$ is of order $1/n$ for each j^* and does not affect the rate of convergence. Thus $\delta^\dagger$ converges at the rate $\mathcal{R}_{j^*}(n)$ uniformly over $\mathcal{F}_{j^*}$ for each j^*. This completes the proof of Theorem 3.

Proof of Theorem 4. Since the constant function $1/2$ is in the Sobolev and neural network classes, by convexity of these classes, $g = f/2+1/2$ is also in these classes. The rate of convergence of the combined procedure then follows directly from Theorem 2. Since δ_3 is consistent for every conditional probability function, we have $R(f;l;\delta_j) \to 0$ as $l \to \infty$. As a consequence, $\sum_{l=n/2}^{n} R(f;l;\delta_j)/n \to 0$ as $n \to \infty$. The consistency of δ^* then follows from (7). This completes the proof of Theorem 4.

Proof of Corollary 1. Since $\| f - \hat{f} \|_2^2$ is a convex loss, averaging over random permutations does not increase the risk. The conclusion of Corollary 1 then follows directly from Theorem 1.

Acknowledgements

The author thanks Hyung-Woo Kim for writing up the Splus program for the simulation study. He is also grateful to the anonymous referees for comments and suggestions on an earlier draft of this paper which led to a significant improvement. This research was partially supported by US National Security Agency Grant MDA9049910060.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory* (edited by B. N. Petrov and F. Csaki), 267-281. Akademia Kiado, Budapest.

Barron, A. R. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (edited by T. M. Cover and B. Gopinath), 85-91. Springer-Verlag, New York.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39**, 930-945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14**, 115-133.

Barron, A. R. and Barron, R. L. (1988). Statistical learning networks: a unifying view. In *Computer Science and Statistics: Proceeding of the 21st Interface*, 192-203. Alexandria, Virginia.

Barron, A. R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301-413.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37**, 1034-1054.

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothing and additive models. *Ann. Statist.* **17**, 453-555.

Cesa-Bianchi, N., Freund, Y., Haussler, D. P., Schapire, R. and Warmuth, M. K. (1997). How to use expert advise? *J. ACM* **44**, 427-485.

Cesa-Bianchi, N. and Lugosi, G. (1999). On prediction of individual sequences. To appear in *Ann. Statist.*.

Clarke, B. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **36**, 453-471.

Devroye, L. (1988). Automatic pattern recognition: a study of the probability of error. *IEEE Trans. Pattern Analysis and Machine Intelligence* **10**, 530-543.

Devroye, L. and Györfi, L. (1983). Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression functions. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics* (edited by F. Konecny, J. Mogyor ódi and W. Wertz), 67-76. Akadémiai Kiadó, Budapest, Hungary.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Friedman, J. and Stuetzel, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817-823.

Gorden, L. and Olshen, L. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **15**, 147-163.

Hall, P. and Hannan, E. J. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika* **75**, 705-714.

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Inform. Comput.* **108**, 212-261.

Lugosi, G. and Nobel, A. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27**, 1830-1864.

McCaffrey, D. F. and Gallant, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks* **7**, 147-158.

Nicoleris, T. and Yatracos, Y. G. (1997). Rate of convergence of estimates, Kolmogrov's entropy and the dimensionality reduction principle in regression. *Ann. Statist.* **25**, 2493-2511.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *Ann. Statist.* **24**, 1084-1105.

Rissanen, J., Speed, T. P. and Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Inform. Theory* **38**, 315-323.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118-184.

Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371-1470.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 111-147.

Vovk, V. G. (1990). Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, 372-383.

Yang, Y. (1996). *Minimax Optimal Density Estimation*. Ph.D. Dissertation, Department of Statistics, Yale University.

Yang, Y. (1999a). Model selection for nonparametric regression. *Statist. Sinica* **9**, 475-499.

Yang, Y. (1999b). Minimax nonparametric classification—part I: rates of convergence. *IEEE Trans. Inform. Theory* **45**, 2271-2284.

Yang, Y. (1999c). Minimax nonparametric classification—part II: model selection for adaptation. *IEEE Trans. Inform. Theory* **45**, 2285-2291.

Yang, Y. (1999d). Regression with multiple candidate models: selecting or mixing? Technical Report #8, Department of Statistics, Iowa State University.

Yang, Y. (2000a). Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74**, 135-161.

Yang, Y. (2000b). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75-87.

Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory* **44**, 95-116.

Yang, Y. and Barron, A. R. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564-1599.

312 Snedecor Hall, Department of Statistics, Iowa State University, Ames, IA 50011-1210, U.S.A.
E-mail: yyang@iastate.edu

(Received October 1998; accepted April 2000)