1998

Minimax Nonparametric Classification—Part II: Model Selection for Adaptation

Yuhong Yang
Iowa State University

Minimax Nonparametric Classification—Part II: Model Selection for Adaptation

Abstract
We study nonparametric estimation of a conditional probability for classification based on a collection of finitedimensional models. For the sake of flexibility, different types of models, linear or nonlinear, are allowed as long as each satisfies a dimensionality assumption. We show that with a suitable model selection criterion, the penalized maximum-likelihood estimator has risk bounded by an index of resolvability expressing a good tradeoff among approximation error, estimation error, and model complexity. The bound does not require any assumption on the target conditional probability and can be used to demonstrate the adaptivity of estimators based on model selection. Examples are given with both splines and neural nets, and problems of highdimensional estimation are considered. The resulting adaptive estimator is shown to behave optimally or near optimally over Sobolev classes (with unknown orders of interaction and smoothness) and classes of integrable Fourier transform of gradient. In terms of rates of convergence, performance is the same as if one knew which of them contains the true conditional probability in advance. The corresponding classifier also converges optimally or nearly optimally simultaneously over these classes.

Keywords
minimax adaptive estimation, minimax rates of convergence, model selection, nonparametric classification, neural networks, resolvability, sparse approximation

Disciplines
Statistics and Probability

Comments
This preprint was published as Yuhong Yang, "Minimax Nonparametric Classification-Part II: Model Selection for Adaptation" *IEEE Transactions on Information Theory* (1999): 2285-2292, doi: 10.1109/18.796369.

Minimax Nonparametric Classification—Part II: Model Selection for Adaptation

Yuhong Yang
Department of Statistics
Iowa State University

Abstract—We study nonparametric estimation of a conditional probability for classification based on a collection of finite-dimensional models. For the sake of flexibility, different types of models, linear or nonlinear, are allowed as long as each satisfies a dimensionality assumption. We show that with a suitable model selection criterion, the penalized maximum likelihood estimator has risk bounded by an index of resolvability expressing a good trade-off among approximation error, estimation error, and model complexity. The bound does not require any assumption on the target conditional probability and can be used to demonstrate the adaptivity of estimators based on model selection. Examples are given with both splines and neural nets, and problems of high-dimensional estimation are considered. The resulting adaptive estimator is shown to behave optimally or near optimally over Sobolev classes (with unknown orders of interaction and smoothness) and classes of integrable Fourier transform of gradient. In terms of rates of convergence, performance is the same as if one knew which of them contains the true conditional probability in advance. The corresponding classifier also converges optimally or nearly optimally simultaneously over these classes.

Index Terms—Minimax adaptive estimation, minimax rates of convergence, model selection, nonparametric classification, neural networks, resolvability, sparse approximation, wavelets.

I. Introduction

A. Minimax Adaptive Estimation for Classification

Let $Z_i = (X_i, Y_i), i = 1, ..., n$ be observed independent copies of the random pair $Z = (X, Y)$. Here class label Y takes values in $\{0, 1\}$, and the feature variable X takes values in a space $\mathcal{X}$, which could be of high dimension. Let $f(x) = P\{Y = 1|X = x)$ be the conditional probability of Y taking label 1 given $X = x$. If f is known to be in a nonparametric function class $\mathcal{F}$, convergence of an estimator uniformly over $\mathcal{F}$ is of interest. Minimax risks characterize the

best one can possibly do in that regard. In [22], minimax rates of convergence for estimating $f \in \mathcal{F}$ and also for classification are identified for a general nonparametric class $\mathcal{F}$.

For minimax function estimation, from a practical point of view, an immediate concern is on the choice of the function class. A pessimistic person might insist that the unknown function f can be any conditional probability. Then no rate of convergence is possible at all (see, e.g., [13]). On the other hand, most people are likely to be skeptical about using an estimator optimal for a function class chosen by an overly confident person based on a pure belief. This suggests that for minimax estimation, it is desirable to consider multiple function classes for more flexibility rather than a fixed choice. For example, some times, it is reasonable to assume that f is smooth. As shown in [22], roughly speaking, the smoother a class is, the faster rate of convergence. For instance, for a one-dimensional Sobolev class with square integrable α-th derivative, the rate of convergence for estimating f under square L_2 loss is $n^{-2\alpha/(2\alpha+1)}$, which converges faster for larger α. But in practice, one never knows how large α is. A pessimistic choice of $\alpha = 1$ is too conservative if f happens to be very smooth with a large value of α. Thus one may wish the estimator to automatically achieve the right rate $n^{-2\alpha/(2\alpha+1)}$ without knowing α in advance. Such adaptation with respect to unknown smoothness parameters is the typical concern of adaptive function estimation in the literature (see [20] for some references).

There are many different ways to characterize functions, e.g., in terms of different norms, or different approximating systems. When the feature dimension is high, to overcome the curse of dimensionality, function classes with reduced dimension (such as additive functions, low-interaction-order functions, and neural network classes) are of interest. Not knowing which class (and the associated smoothness parameters, if any) is the best for the underlying function, we wish to have an estimator that performs automatically optimally no matter which one contains the true function. The goal here is more ambitious than adaptation only with respect to smoothness parameters. A pioneering work in this direction is in Barron and Cover [7] for general density estimation using minimum description length criterion. In this paper, we present such adaptation results by model selection for the estimation of the conditional probability and classification based on results on adaptive density estimation in [20]. We demonstrate that with a suitable choice of models and a good model selection criterion, the penalized maximum likelihood estimator of the conditional probability and the corresponding plug-in classifier converge optimally or near optimally simultaneously over multiple, possibly very different, target function classes.

B. Model Selection for Flexibility

In light of approximation theory, functions in various infinite-dimensional classes can be well approximated by finite-dimensional families. As showed in [22], with a given approximating system, capability of estimating f or classification is intrinsically related to the capability of approximation to f. As suggested from the minimax results there, a good balance of model dimension (with a suitable choice of terms) and approximation error is likely to result in an optimal estimator in terms of rates of convergence. Not knowing the approximation error and the best terms, a good model selection criterion is desired so that the estimator can automatically achieve the best trade-off.

Different model selection methods can be used for estimating f based on finite-dimensional models. General risk bounds in terms of index of resolvability are obtained in [2] and [7] by complexity regularization and minimum description length (MDL) criterion respectively on suitably chosen nets in the models, and the resolvability bounds are used to demonstrate adaptation in various settings. A similar result is obtained without any discretization using penalized maximum likelihood estimator [20]. The results in this paper are mainly based on work in [20]. For selection of a classifier among candidate classes, general risk bounds are obtained in [15] and [17] based on complexity regularization using empirical complexity and empirical risk minimization, respectively. By using a chaining argument, the risk bounds in [6], [20] and this paper can some times avoid an extra logarithmic factor as appeared in [15] and [17], resulting in optimal rates of convergence. For applications of our risk bounds, we focus on minimax-rate adaptive estimation. For consistency results based on finite-dimensional models, see, e.g., [12] and [16].

II. Method of Estimation

To estimate the conditional probability f, a collection of finite-dimensional families

$$\{f_k(x, \theta^{(k)}), \theta^{(k)} \in \Theta_k\},\ k \in \Gamma,$$

are considered to approximate f. For convenience, we will drop the superscript and use the same symbol θ for the parameters in all models (since no relationship between the parameters in different models will be assumed, this hopefully will not cause any confusion). Here Γ is the collection of indices of the models being considered. The model list is assumed to be fixed and independent of sample size unless otherwise stated. In this paper, we always restrict our

attention to members in each family that are bounded between 0 and 1, as required to be a valid conditional probability function.

A. *Maximum Likelihood Estimation within a Model*

Let $h(x)$ denote the unknown marginal density of the feature variable X with respect to a probability measure μ. For a function g between 0 and 1, let

$$p_g(x,y) = h(x) \cdot g(x)^y (1-g(x))^{1-y}$$

denote the joint density of (X,Y) with conditional probability g with respect to the product measure of μ and the counting measure. For model k, the likelihood function then is

$$\Pi_{i=1}^n p_f(x_i,y_i) = \Pi_{i=1}^n h(x_i) \cdot \Pi_{i=1}^n \left(f_k(x_i,\theta)^{y_i} \left(1 - f_k(x_i,\theta)\right)^{1-y_i} \right).$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ maximizes

$$\Pi_{i=1}^n f_k(X_i,\theta)^{Y_i} \left(1 - f_k(X_i,\theta)\right)^{1-Y_i}$$

over $\theta \in \Theta_k$ (under the restriction that $f_k(\cdot,\theta)$ is bounded between 0 and 1). Note that the MLE does not depend on any knowledge of the distribution of (X,Y).

B. Conditions on the Families

A dimensionality assumption is used for our results. Let $\| \cdot \|_2$ denote the L_2 norm with respect to μ, and let $L_2(h)$ norm denote the L_2 norm weighted by h, i.e., for a function g, $\| g \|_{L_2(h)} = \left(\int g^2(x) h(x) d\mu \right)^{1/2}$. For each $\theta_0 \in \Theta_k$, consider the L_2 balls centered at f_{k,θ_0} in family k defined by

$$B_k(\theta_0, r) = \{ f_{k,\theta} : \theta \in \Theta_k, \| f_{k,\theta_0} - f_{k,\theta} \|_2 \leq r \}.$$

Assumption 1. For each $k \in \Gamma$, there exist a positive constant A_k and an integer $m_k \geq 1$ such that for any $\theta_0 \in \Theta_k$, any $r > 0$ and $\delta \leq 0.0056r$, there exists a δ-net $\mathbf{G}_{k,r,\delta,\theta_0}$ for $B_k(\theta_0, r)$ in the L_∞ norm (i.e., for any $f \in B_k(\theta_0, r)$, there exists $g \in \mathbf{G}_{k,r,\delta,\theta_0}$ such that $\| f - g \|_\infty \leq \delta$) satisfying the following requirement:

$$\mathrm{card}(\mathbf{G}_{k,r,\delta,\theta_0}) \leq \left(\frac{A_k r}{12\sqrt{2}\delta} \right)^{m_k}.$$

Here m_k is called the metric dimension of the model k.

This dimensionality assumption is satisfied by linear approximating families and some non-linear families such as neural networks, and m_k is usually the number of parameters in the family (see Section IV for examples).

C. Model List Γ and Model Complexity

In principle, if an individual family satisfies the above dimensionality assumption, it can be included in the model list Γ as long as the list is countable. Thus different types of families, linear or not, are allowed in the competition for best estimation of f. This provides flexibility that is desirable or even essential for situations such as high-dimensional classification, where one searches for a good parsimonious model to overcome the curse of dimensionality. However, when exponentially many models are considered, significant selection bias might occur with any bias-correction based criteria like AIC [1]. To handle the selection bias, as in [20], we incorporate a model complexity penalty.

For each model in the model list, a complexity C_k is assigned with $L_k = (\log_2 e)C_k$ satisfying the Kraft's inequality: $\sum_k 2^{-L_k} \leq 1$, i.e., $\sum_k e^{-C_k} \leq 1$. See [3] and [7] for similar earlier uses of a model complexity term. The complexity $L_k = C_k \log_2 e$ can be interpreted as the codelength of a uniquely decodable code to describe the models. Complexity assignments can be done naturally according to description or organization of the models. The inclusion of the model complexity term in the criterion in the next subsection regulates the competition among the models to ensure a good risk bound on the selected model.

D. Model Selection Criterion

Choose $\hat{k}$ to minimize the penalized log-likelihood

$$-\sum_{i=1}^{n} \Big(Y_i \log f_k(X_i, \hat{\theta}) + (1 - Y_i) \log \Big(1 - f_k(X_i, \hat{\theta}) \Big) \Big) + \lambda_k m_k + 9.49 C_k, \tag{1}$$

where $\lambda_k > 0$ is a constant to be determined. It is the criterion in [20] specialized for the classification problem. Differently from AIC, a model complexity is added in (1) and the dimensionality penalty coefficient λ_k is allowed to depend on the model in general. The final estimator from model selection then is

$$\hat{f}(x) = f_{\hat{k}}(x, \hat{\theta}),$$

i.e., the MLE in the selected model.

E. Index of Resolvability

Borrowing a terminology from Barron and Cover [7], we define the index of resolvability (relative to the choice of models in Γ) as follows:

$$R_n(f) = \inf_{k \in \Gamma} \{ \inf_{\theta \in \Theta_k} \|f - f_{k,\theta}\|^2_{L_2(h)} + \frac{\lambda_k m_k}{n} + \frac{9.49 C_k}{n} \}. \tag{2}$$

It provides a trade-off among the approximation error, estimation error and the model complexity relative to sample size. As will be seen in the next section, under some conditions, it provides an upper bound on the risk of the estimator $\hat{f}$. Then for examining the performance of $\hat{f}$ for a target class, one only needs to evaluate the index of resolvability for the class using the chosen approximation system.

III. Main results

We consider the square $L_2(h)$ loss for the estimation of f. For classification, a Bayes decision g^* with knowledge of f is to classify Y as class 1 for all x that $f(x) \geq 1/2$ and classify Y as class 0 otherwise, i.e., $g^*(x) = 1$ if $f(x) \geq 1/2$ and $g^*(x) = 0$ if $f(x) < 1/2$. For a classifier $\delta = \delta(x; Z^n)$ based on $Z^n = (X_i, Y_i)_{i=1}^n$, the loss we consider is the difference between the error probability under δ and the Bayesian error probability. Then the risk of a classifier δ is

$$r(f; \delta; n) = EP(Y \neq \delta(X; Z^n)|Z^n) - P(Y \neq g^*(X)).$$

We call it mean error probability regret (MEPR). 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 classifies Y as class 0 otherwise. As is well-known (see [13], p. 95), a plug-in classifier has a risk bound

$$r(f; \delta;) \leq 2 \left(E\|f - \hat{f}\|^2_{L_2(h)} \right)^{1/2}.$$

It is shown in [22] that for many familiar function classes (e.g., bounded variation and Besov), if $\hat{f}$ is minimax-rate optimal for f in the class, the upper bound obtained this way is also optimal for classification.

A. Risk Bounds in Terms of Resolvability

Our risk bounds require an additional assumption on the families. Assume that each family in Γ is uniformly bounded away from 0 and 1, i.e., there exists a constant $0 < \alpha_k \leq 1/2$ such that

$$\alpha_k \leq f_k(x, \theta) \leq 1 - \alpha_k \text{ for all } \theta \in \Theta_k. \tag{3}$$

The natural approximating families usually do not satisfy this condition. Then one can consider an increasing sequence of sub-families, e.g., $\{f_{k,\theta} : 1/j \leq f_k(x, \theta) \leq 1 - 1/j\}$, $j \geq 1$ and treat each one as a model (see [20] for a similar treatment). The complication can be avoided by a modification of the data as will be discussed in subsection C.

Define

$$\Lambda(A) = 4.75 \log A + 27.93. \tag{4}$$

Assume that the (unknown) feature density h is uniformly lower bounded away from zero, i.e., there exists a known constant $\underline{a} > 0$ such that $h \geq \underline{a}$. In the model selection criterion (1), choose λ_k sufficiently large:

$$\lambda_k \geq \Lambda\left(A_k/(4\underline{a}\alpha_k)\right).$$

Theorem 1. Assume Assumption 1 and the boundness assumption in (3) are satisfied. Then the estimator $f_{\hat{k},\hat{\theta}}$ has risk bounded by

$$E\|f - f_{\hat{k},\hat{\theta}}\|^2_{L_2(h)} \leq \frac{5314}{\alpha_k^2} R_n(f).$$

Let $\delta_{f_{\hat{k},\hat{\theta}}}$ be the plug-in classifier. Then the MEPR is bounded by

$$r(f; \delta_{f_{\hat{k},\hat{\theta}}}; n) \leq \frac{72.9}{\alpha_k}\sqrt{R_n(f)}.$$

Note that the above conclusion does not depend on any condition on f and holds for every sample size. This property is essential for obtaining minimax adaptivity based on model selection over different function classes, see Sections VI and V for examples.

Remarks:

1) Since the resolvability bound in the theorem is valid for any sample size, the model list Γ is allowed to change according to the sample size.

2) From [20], the requirement in Assumption 1 only needs to be checked for $r \geq 7.8\sqrt{\frac{m_k}{n}}$ for the conclusion of Theorem 1 to hold. If this weaker requirement is satisfied with $A_{k,n}$ (depending also on the sample size) instead of A_k, and we use $\lambda_{k,n} \geq \Lambda\left(A_{k,n}/\left(4\underline{a}\alpha_k\right)\right)$ in the criterion, the conclusion of Theorem 1 is still valid.

3) For selecting a classifier among candidate classes, risk bounds are obtained in [15] and [17] based on empirical complexity regularization and structural risk minimization respectively without any assumption on the distribution of (X, Y). Our risk bound for classification through estimating f is somewhat restrictive because of the assumption on h, but our approach seems easier to compute and is readily applicable for deriving rates of convergence based on approximation theory.

Proof of Theorem 1: We apply Theorem 1 in [20]. Let w_β denote the Bernoulli probability function with success probability β, i.e., $w_\beta(y) = \beta^y(1-\beta)^{1-y}$ for $y = 0$ or 1. It is easy to verify that the square Hellinger distance between w_{β_1} and w_{β_2} satisfies $d_H^2\left(w_{\beta_1}, w_{\beta_2}\right) \geq (\beta_1 - \beta_2)^2/2$. As a consequence, for any two functions f_1 and f_2 between 0 and 1, using that $h \geq \underline{a}$, we have

$$d_H^2(p_{f_1}, p_{f_2}) \quad = \quad \int h(x) d_H^2\left(w_{f_1}, w_{f_2}\right)\mu\left(dx\right) \geq \frac{1}{2}\|f_1 - f_2\|^2_{L_2(h)} \geq \frac{\underline{a}^2}{2}\|f_1 - f_2\|_2^2. \qquad (5)$$

Thus $\{f_{k,\theta} : d_H(p_{f_{k,\theta^0}}, p_{f_{k,\theta}}) \le r\} \subset \{f_{k,\theta} : \| f_{k,\theta_0} - f_{k,\theta} \|_{L_2(h)} \le \sqrt{2}r/\underline{a}\}$. Note that under the boundness assumption on the family, the sup-norm distance between $p_{f_{k,\theta_1}}$ and $p_{f_{k,\theta_2}}$ is upper bounded by $(1/\alpha_k)$ times the sup-norm distance between f_{k,θ_1} and f_{k,θ_2}. From all above, under Assumption 1, Assumption 1 in [20] is satisfied with the constants $A_k/(4\underline{a}\alpha_k)$ and m_k in the cardinality bound. Then by Theorem 1 in [20], we have

$$Ed_H^2(p_f, p_{f_{\hat{k},\hat{\theta}}}) \le 2657 \inf_{k\in\Gamma}\{D(p_f \parallel p_{f_{k,\theta^*}}) + \frac{\lambda_k m_k}{n} + \frac{9.49C_k}{n}\},$$

where

$$D(p_f \parallel p_{f_{k,\theta^*}}) = \inf_{\theta\in\Theta_k} \int h\left(f\log\frac{f}{f_{k,\theta}} + (1-f)\log\frac{1-f}{1-f_{k,\theta}}\right)d\mu$$

is the smallest Kullback-Leibler (K-L) divergence between f and the family. By the familiar chi-square bound on the K-L divergence, together with the boundness assumption, it can be shown that for $\alpha \le \beta_2 \le 1-\alpha$, the K-L divergence between w_{β_1} and w_{β_2} satisfies $D(w_{\beta_1} \parallel w_{\beta_2}) \le (\beta_1-\beta_2)^2/\alpha^2$. As a consequence, for $\theta \in \Theta_k$ with $\alpha_k \le f_{k,\theta} \le 1-\alpha_k$, we have

$$D\left(p_f \parallel p_{f_{k,\theta}}\right) = \int hD\left(w_f \parallel w_{f_{k,\theta}}\right)d\mu \le \frac{\|f-f_{k,\theta}\|_{L_2(h)}^2}{\alpha_k^2}.$$

The conclusion follows together with (5). This completes the proof of Theorem 1.

To apply Theorem 1, if the true conditional probability f is bounded away from 0 and 1 but with the bounds unknown, then as mentioned earlier, we can consider a sequence of increasing subfamilies from each original natural family. With suitable complexity assignment, the estimator based on model selection converges as if we knew the bounds in advance (see [20]).

If f may approach 0 or 1 at some points or even be 0 or 1 on some subset of $\mathcal{X}$, the risk bounds in Theorem 1 can be large due to division of a small number α_k^2 for good models. A natural question is: Do the bounds really characterize the performance of the estimator $\hat{f}$ and the plug-in classifier for this case? We do not have a definite answer, but we tend to think that the estimator should behave reasonably and the largeness of the bounds probably mainly comes from the technical analysis relating different distances. The following subsection is intended to avoid that difficulty. In the mean time, the treatment avoids the consideration of subfamilies for each original natural family.

C. A Modification to Improve the Risk Bound

We apply a technique used in [20], [21] and [22]. The idea is to modify the data so that the conditional probability becomes bounded away from 0 and 1. In addition to the observed

i.i.d. sample $(X_1, Y_1), ..., (X_n, Y_n)$, we generate some random variables. Let W_i, $1 \le i \le n$, be independently generated Bernoulli random variables with success probability 1/2. Let $\widetilde{Y}_i$ be Y_i or W_i with probability $(1/2, 1/2)$ according to the outcome of Bernoulli(1/2) random variables V_i generated independently for $i = 1, ..., n$. The conditional probability of $\widetilde{Y}_i$ taking value 1 given $X = x$ is $g(x) = (f(x) + 1/2)/2$. The new function g is bounded below by 1/4 and above by 3/4. Now applying the model selection criterion in (1) on the new data $\widetilde{Z}^n = (X_i, \widetilde{Y}_i)_{i=1}^n$ using the families $\{f_{k,\theta} : \theta \in \Theta_k \text{ and } 1/4 \le f_{k,\theta} \le 3/4\}$ restricted accordingly, we have an estimator $\hat{g}$ of g based on $\widetilde{Z}^n$. From $f(x) = 2g(x) - 1/2$, let

$$\hat{f}_{rand}(x) = 2\hat{g}(x) - 1/2.$$

Then $\hat{f}_{rand}(x)$ is a valid randomized estimator of f depending on Z^n and the generated random variables. (If one wishes, one can take conditional expectation of $\hat{f}_{rand}$ with respect to the randomness in the generated random variables to get a nonrandomized estimator of f with no larger risk.)

Assumption 2: Each family in Γ satisfies the condition that if s is in $\{f_{k,\theta} : \theta \in \Theta_k\}$, so is $(s + 1/2)/2$.

The condition holds if a family allows shift and scale change as is the case with any linear family including the constant term, and some nonlinear families such as the neural networks in Section IV.

Theorem 2. Assume Assumptions 1 and 2 are satisfied. Take $\lambda_k \ge \Lambda(A_k/\underline{a})$ in the model selection criterion in (1) applied to the modified data $\widetilde{Z}^n$. Then for the estimator $\hat{f}_{rand}$, we have

$$E\|f - \hat{f}_{rand}\|^2_{L_2(h)} \le MR_n(f),$$

where M is an absolute constant. The plug-in classifier $\delta_{\hat{f}_{rand}}$ has risk bounded by

$$r(f; \delta_{\hat{f}_{rand}}; n) \le \sqrt{MR_n(f)}.$$

Proof of Theorem 2: Applying Theorem 1, we have $E_{\widetilde{Z}^n}\|\hat{g} - g\|^2_{L_2(h)} \le M' R_n(g)$, where $M' = 5314 \cdot 16$. As a consequence, the risk of $\hat{f}_{rand}$ is bounded as follows:

$$E\|f - \hat{f}_{rand}\|^2_{L_2(h)} = 4E_{\widetilde{Z}^n}\|\hat{g} - g\|^2_{L_2(h)} \le 4M' R_n(g), \tag{6}$$

where the first expectation is taken with respect to both the original randomness in Z^n and that from the generated random variables. Under Assumption 2, it can be easily verified that

$R_n(g) \le R_n(f)$ (see [23] for detail). The result of Theorem 2 follows. This completes the proof of Theorem 2.

D. Resolvability and Adaptation

Through the risk bounds in terms of resolvability, adaptation results can be readily obtained. Given a class of conditional probability, by evaluating $R_n(f)$ for f in the class, a uniform upper bound on the convergence rate of the estimator based on model selection is obtained. If f is believed to be in one of a collection of possibly very different function classes, we can consider the union of approximating families that are suitable for each one (or more) of the classes. Based on the resolvability bounds and the relationship between approximation and minimax rates in [22], with a suitable model complexity assignment, the estimator often converges optimally or near optimally simultaneously over the considered classes. See Section IV for a demonstration. This provides a tool to show minimax adaptivity of the estimator based on model selection. Even when f is not in any of the considered classes based on which the approximating models are chosen, the resolvability bound is still meaningful and expresses a good trade-off among the approximation error, estimation error and model complexity.

IV. Some Applications

In this section, we apply the model selection results for two examples. More applications including general linear and sparse approximations are given in [23].

Assume the feature variable X is supported in $[0,1]^d$ for some $d \ge 1$ and has Lebesgue density h bounded above and below:

$$0 < \underline{a} \le h \le \overline{a} < \infty,$$

for some known constants $\underline{a}$ and $\overline{a}$. Some of the results given below are sketched and detailed treatments can be done as in [20]. Let $\hat{f}$ denote the estimator based on model selection using the modified data as in Theorem 2.

A. Univariate log-spline models

For $m \ge q$, let $\varphi_{m,q,1}(x), \varphi_{m,q,2}(x), ...,\varphi_{m,q,m}(x)$ be the B-spline basis (some piecewise polynomials of order less than q) with $m-q+2$ equally spaced knots (see, e.g., [10]). Let

$$f_{m,q}(x,\theta) = \sum_{i=1}^{m} \theta_i \varphi_{m,q,i}(x)$$

be the corresponding approximating family. Let us index our models by $k = (m,q)$. They satisfy Assumption 1 for some constants $A_k = A_q$ depending only on the spline order q and

$m_k = m$ (for detail, see [23]). We specify the model complexity in a natural way to describe the index as follows:

1) describe q using $\log_2^* q$ bits

2) describe m using $\log_2^* m$ bits,

where the function $\log^*$ is defined by $\log^* i = \log(i+1) + 2\log\log(i+1)$ for $i \geq 1$. This leads to a natural choice of $C_{(m,q)} = \log^* q + \log^* m$.

Assume f belongs to $W_2^{s^*}(U^*)$ for some $s^* \geq 1$ and $U^* > 0$, where $W_2^s(U)$ is a Sobolev class which includes all the functions with squared integral of the s-th derivative bounded by U. The hyper-parameters s^* and U^* are not known. To investigate the performance of the estimator based on model selection, in light of Theorem 2, we only need to evaluate the index of resolvability for f in the Sobolev classes with the spline models.

For the approximation error of $f \in W_2^{s^*}(U^*)$ with the splines of the right order $q = s^*$, by [11], together with the boundness assumption on h, we have that for each $m \geq s^*$,

$$\inf_\theta \|f - f_{k,\theta}\|^2_{L_2(h)} \leq \overline{a}^2 \inf_\theta \|f - f_{k,\theta}\|_2^2 \leq \frac{\overline{a}^2 K}{(m - s^* + 2)^{s^*}} \|f^{(s^*)}\|_2^2 \leq \frac{\overline{a}^2 K U^*}{(m - s^* + 2)^{s^*}}$$

for some absolute constant K. As a consequence, by restricting attention to the splines of the right order s^*, taking $\lambda_k = \Lambda(A_q/\underline{a})$, the index of resolvability is upper bounded by

$$\begin{aligned} R_n(f) &\leq \inf_{m \geq s^*} \left(\inf_\theta \|f - f_{(m,s^*),\theta}\|^2_{L_2(h)} + \frac{\lambda_{(m,s^*)} m}{n} + \frac{9.49 C_{(m,s^*)}}{n} \} \right) \\ &\leq \inf_{m \geq s^*} \left(\frac{\overline{a}^2 K U^*}{(m - s^* + 2)^{s^*}} + \frac{\Lambda(A_{s^*}/\underline{a}) m}{n} + \frac{9.49 C_{(m,s^*)}}{n} \} \right) \\ &\leq M n^{-\frac{2s^*}{2s^*+1}}, \end{aligned}$$

where for the last inequality, we take m of order $n^{1/(2s^*+1)}$ and the constant M depends on $\underline{a}$, $\overline{a}$, s^* and U^*. As a consequence, for any $f \in W_2^{s^*}(U^*)$,

$$E \parallel f - \hat{f} \parallel^2_{L_2(h)} \leq M' \cdot n^{-\frac{2s^*}{2s^*+1}},$$

where the constant M' depends only on $\underline{a}$, $\overline{a}$, s^* and U^*.

Note that the rate $n^{-\frac{2s^*}{2s^*+1}}$ is the minimax optimal rate of convergence for estimating f in $W_2^{s^*}(U^*)$ (which is a special case of Besov classes) and $n^{-\frac{s^*}{2s^*+1}}$ is the minimax optimal rate for classification [22]. Thus the model selection guarantees the optimal rate of convergence for the Sobolev classes without knowledge of U^* and s^*.

B. Neural Network Models

Consider feedforward neural network models with one layer of sigmoidal nonlinearities, which have the following form:

$$f_l(x,\theta) = \sum_{j=1}^{l} \eta_j \phi(a_j^T x + b_j) + \eta_0.$$

The function is parametrized by θ, consisting of η_0, $a_j \in R^d$, $b_j, \eta_j \in R$, for $j = 1, 2, ..., l$. Here ϕ is a given Lipschitz sigmoidal function with $\| \phi \|_\infty \leq 1$, $\lim_{z\to\infty} \phi(z) = 1$ and $\lim_{z\to-\infty} \phi(z) = 0$.

The target function $f(x)$ is assumed to have a Fourier representation of the form $f(x) = \int_{R^d} e^{i\omega^T x} \widetilde{f}(\omega) d\omega$. Let $\varsigma_g = \int |\omega|_1 |\widetilde{f}(\omega)| d\omega$, where $|\omega|_1 = \sum_{j=1}^d |\omega_j|$ is the l_1 norm of ω in R^d. Consider the class $\mathcal{F}(\varsigma) = \{g : \varsigma_g \leq \varsigma\}$, which was introduced and extensively studied in [4]. Consider the parameter space

$$\Theta_{l,\tau_l,\varsigma} = \{\theta : \max_{1\leq j\leq l} |a_j|_1 \leq \tau_l, \max_{1\leq j\leq l} |b_j| \leq \tau_l, \sum_{j=1}^{l} |\eta_j| \leq 2\varsigma\}.$$

Since ς is not known, we consider all positive integer values. Then the model index is $k = (l, \varsigma)$ with $l \geq 1$ and $\varsigma \geq 1$. When τ_l is chosen appropriately as in [5], under a mild condition on ϕ, Assumption 1 is satisfied with $A_{k,n} \leq \text{const} \times l^{\beta_2 - \frac{1}{2}} \sqrt{n}$ and $m_k = ld + 2l + 1$. We can use $\log_2^* l + \log_2^* \varsigma$ bits to describe the model index k. As a consequence, the model complexity is assigned as $C_k = \log^* l + \log^* \varsigma$. Applying Theorem 2, and examining the index of resolvability, we have

$$\sup_{f\in\mathcal{F}(\varsigma)} E \| f - \hat{f} \|^2 \leq M \left(\frac{d \log n}{n} \right)^{\frac{1}{2}},$$

where the constant M depends only on ς, ϕ, $\underline{a}$, and $\overline{a}$.

Note that the feature dimension affects the speed of convergence polynomially (instead of exponentially) as in [3] and [5] but without discretization of the parameter spaces. The corresponding plug-in classifier converges at rate $O\left(\frac{d \log n}{n}\right)^{\frac{1}{4}}$. Similarly to the class $N(C)$ in [22], these rates are close to the minimax optimal ones when d is large. The same rates of convergence are obtained recently in [14] relaxing the Lipschitz requirement on ϕ to bounded variation.

V. A Demonstration of High-Dimensional Estimation

When the dimension of the feature variable X is high, estimation of f by traditional methods (e.g., histogram, kernel, complete models based on series expansion) faces the curse of dimensionality. Alternative parsimonious models can some times improve estimation accuracy significantly. For such a situation, it is desirable to have a lot of flexibility to capture the

unknown characteristics of the underlying function. To that end, different types of models are considered. This is a step further from the consideration of adaptation with respect to unknown smoothness parameters of the same type of function classes.

Assume f is supported in $[0,1]^d$ with d large and h is bounded above and below as in Section IV. We consider two methods of dimensionality reduction: low-order tensor-product splines, and neural nets.

A. Two Types of Models

1) Tensor-Product Splines: Tensor-product spline models have been proposed and studied in [19] for general function estimation. It was shown that suitable splines models can result in estimators converging at optimal rates in probability. However, the method there requires knowledge of smoothness parameters and interaction order, and therefore is not adaptive.

Let $\varphi_{m,q,1}(x), ..., \varphi_{m,q,m}(x)$ be the B-spline basis as in Section IV. For $1 \le r \le d$, let $J_r = (j_1, ..., j_r)$ $(j_1 < j_2 < ... < j_r)$ be an ordered vector of elements from $\{1, 2, ..., d\}$ and let $\mathcal{J}_r$ denote the set of all possible such choices. Let $\mathbf{x}_{J_r} = (x_{j_1}, ..., x_{j_r})$ be the subvector of $\mathbf{x}$ with subscript in J_r. Let $\mathbf{m}_r = (m_1, ..., m_r)$ and $\mathbf{q}_r = (q_1, ..., q_r)$ be vectors of integers. Let $\mathbf{i}_r = (i_1, ..., i_r)$ 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

$$\{\varphi_{\mathbf{i}_r}(\mathbf{x}_{J_r}) = \Pi_{l=1}^r \varphi_{m_l,q_l,i_l}(x_{j_l}) : J_r \in \mathcal{J}_r; 1 \le i_l \le m_j \text{ for } 1 \le l \le r\} \tag{7}$$

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 (7). Since the functions in (7) are not all linearly independent when $r < d$, the dimension m_I of the family is then less than $\binom{d}{r} \Pi_{i=1}^r m_i$. But in any case, m_I is of order $\Pi_{i=1}^r m_i$. It can be shown that Assumption 1 is satisfied (see [23]).

2) Neural Network Models: As in Section IV.B.

B. Target Classes

1) Sobolev Classes with Different Orders of Interaction and Smoothness: For $r \ge 1$, $\mathbf{l} = (l_1, ..., l_r)$ with nonnegative integer components l_i, define $|\mathbf{l}| = \sum_{i=1}^r l_i$. Let $\mathbf{z}_r = (z_1, ..., z_r) \in [0,1]^r$. Let $D^{\mathbf{l}}$ denote the differentiation operator $D^{\mathbf{l}} = \partial^{|\mathbf{l}|}/\partial z_1^{l_1} \cdots \partial z_r^{l_r}$. For an integer α, the Sobolev norm is $\| g \|_{W_2^{\alpha,r}} = \| g \|_2 + \sum_{|\mathbf{l}|=\alpha} \int_{[0,1]^r} |D^{\mathbf{l}} g|^2 d\mathbf{z}_r$. Let $W_2^{\alpha,r}(C)$ denote the set of all functions g on $[0,1]^r$ with $\| g \|_{W_2^{\alpha,r}} \le C$. It is a r-dimensional Sobolev class. Now consider the

following function classes on $[0,1]^d$ of different interaction orders and smoothness:

$$S_1(\alpha;C) = \{\textstyle\sum_{i=1}^d g_i(x_i) : g_i \in W_2^{\alpha,1}(C), 1 \le i \le d\}$$

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

$$\ldots$$

$$S_d(\alpha;C) = W_2^{\alpha,d}(C)$$

with $\alpha \ge 1$ and $C > 0$. The simplest class $S_1(\alpha;C)$ contains additive functions (no interaction) with each component in a univariate Sobolev class, and with r increases, functions in $S_r(\alpha;C)$ have higher order interactions. The L_2 metric entropies of these classes are of the same orders as $W_2^{\alpha,1}(C)$, ..., $W_2^{\alpha,d}(C)$ respectively, i.e., the metric entropy of $S_r(\alpha;C)$ is of order $\epsilon^{-r/\alpha}$. Then by Theorem 1 in [22], the minimax rate of convergence under square $L_2(h)$ loss for estimating $f \in S_r(\alpha;C)$ is $n^{-2\alpha/(2\alpha+r)}$ for $1 \le r \le d$. For classification, the minimax rate of the mean error probability regret is $n^{-\alpha/(2\alpha+r)}$ using Theorem 2 in [22]. Note that these convergence rates *do not* depend on the input dimension d, but rather on the true interaction order. When the interaction order is low relative to d, the rate of convergence is much better compared to $n^{-2\alpha/(2\alpha+d)}$, overcoming the curse of dimensionality.

2) Classes with Integrable Fourier Transform of Gradient: $\{\mathcal{F}(\varsigma) : \varsigma > 0\}$ as in Section IV.B.

C. Assignment of Model Complexity

Based on the structure of the models, it is natural to first describe type using $\log 2$ bits, and then within the same type describe the other involved hyper-parameters. The description of the neural network models is already mentioned in Section IV. So we now describe the tensor-product models. Let $(r, \mathbf{q}_r, \mathbf{m}_r)$ be the hyper-parameters defining a spline model. To describe $(r, \mathbf{q}_r, \mathbf{m}_r)$ (after specifying that the type is spline), we just need to describe a few integers. Since r is between 1 and d, we only need $\log_2 d$ bits to describe r. To describe $\mathbf{q}_r$ and $\mathbf{m}_r$, we use $\sum_{j=1}^r \log^* q_j$ and $\sum_{j=1}^r \log^* m_j$ bits respectively. Together with the description of the model type, we have the following assignment of the overall complexity.

1) Spline Models: $\log 2 + \log d + \sum_{j=1}^r \log^* q_j + \sum_{j=1}^r \log^* m_j$

2) Neural Network Models: $\log 2 + \log^* l + \log^* \varsigma$.

D. Adaptive Rate of Convergence

Let $\hat{f}$ be the estimator as in Theorem 2 based on the above models.

Theorem 3: The estimator $\hat{f}$ has risk bounded simultaneously for the target classes as

follows

$$\sup_{f \in S_r(\alpha;C)} E \parallel f - \hat{f} \parallel^2 = O\left(n^{-2\alpha/(2\alpha+r)}\right)$$

for all $1 \leq r \leq d$, $\alpha \geq 1$ and $C > 0$ and

$$\sup_{f \in \mathcal{F}(\varsigma)} E \parallel f - \hat{f} \parallel^2 = O\left(\frac{d \log n}{n}\right)^{\frac{1}{2}}$$

for all $\varsigma > 0$. The plug-in classifier based on $\hat{f}$ has risk uniformly bounded for the above classes by orders of the square root of the above rates respectively.

The rates for estimating f and for classification are optimal for the Sobolev classes. As mentioned in Section IV, the above rate for $\mathcal{F}(\varsigma)$ is close to the optimal rate when d is large. Note that the convergence rates for $\mathcal{F}(\varsigma)$ and the Sobolev classes with r much smaller than d are fast. Thus the curse of dimensionality is automatically avoided if the unknown conditional probability f is well approximated by the neural nets or lower-order tensor-product splines. In practice, other parsimonious models can be considered at the same time for even more flexibility.

Proof of Theorem 3: Based on Theorem 2, we only need to examine the index of resolvabilty for the function classes. Observing that the index of resolvability is increased by only $9.49 \log 2/n$ for $\mathcal{F}(\varsigma)$ compared to that when only the neural nets are considered as in Section IV, the rate for $\mathcal{F}(\varsigma)$ follows. It remains to derive the rates for the Sobolev classes. To that end, the main task is to upper bound the approximation error for these classes by the tensor-product splines. For $f \in W_2^{\alpha,r}(C)$, from [18] (Theorem 12.8 and Equation 13.69) as used in [19], with $\mathbf{q}_r^*$ satisfying $|\mathbf{q}_r^*| = \alpha$ and $\mathbf{m}_r^* = (m, m, ..., m)$, for the spline model $I = (r, \mathbf{q}_r^*, \mathbf{m}_r^*)$, the approximation error $\inf_\theta \parallel f - f_{I,\theta} \parallel^2$ is upper bounded by $Mm^{-2\alpha}$, where the constant M depends only on r, $\mathbf{q}_r^*$ and C. As a consequence, the approximation error of class $S_r(\alpha; C)$ by model I is upper bounded by order $\binom{d}{r} Mm^{-2\alpha}$. The model dimension of I is of order m^r. Note that for given r and $\mathbf{q}_r^*$, the model complexity $\log 2 + \log d + \sum_{j=1}^r \log^* q_j + r \log^* m$ is asymptotically negligible compared to m^r. From all above, by optimizing m, the index of resolvability for class $S_r(\alpha; C)$ is seen to be of order $n^{-2\alpha/(2\alpha+r)}$. This completes the proof of Theorem 3.

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Info. Theory*, pp. 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest, 1973.

[2] A.R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas editor, Kluwer, Dordrecht, Netherlands, 1991, pp. 561-576.

[3] A.R. Barron, "Neural net approximation," in *Proc. Yale Workshop Adaptive Learning Syst.*, K. Narendra, Ed., Yale University, May 1992.

[4] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory* vol. 39, pp. 930-945, 1993.

[5] A.R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning* vol. 14, 115-133, 1994.

[6] A.R. Barron, L. Birgé and P. Massart, "Risk bounds for model selection via penalization," To appear in *Probability Theory and Related Fields.* 1996.

[7] A.R. Barron, and T.M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory* vol. 37, 1034-1054, 1991.

[8] A.R. Barron, L. Birgé and P. Massart, "Risk bounds for model selection via penalization," To appear in *Probability Theory and Related Fields,* 1996.

[9] L. Birgé and P. Massart, "From model selection to adaptive estimation," *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam*, (D. Pollard, E. Torgersen and G. Yang, eds.), Springer, New York, 1996.

[10] C. de Boor, *A practical Guide to Splines*, Springer-Verlag New York, 1978.

[11] C. de Boor and G.J. Fix, "Spline approximation by quasiinterpolents," *J. Approx. Theory*, vol. 8, pp. 19-45, 1973.

[12] L. Devroye, "Automatic Pattern Recognition: a Study of the Probability of Error," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 10, 530-543, 1988.

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

[14] A. Krzyżak and T. Linder, "Radial basis function networks and complexity regularization in function learning," *IEEE Trans. Neural Networks*, vol. 9, pp. 247-256, 1998.

[15] G. Lugosi and A. Nobel, "Adaptive model selection using empirical complexities," submitted to *Ann. Statistics*, 1996.

[16] G. Lugosi and K. Zeger, "Nonparametric Estimation via Empirical Risk Minimization," *IEEE Trans. on Information Theory*, vol. 41, pp. 677-687, 1995.

[17] G. Lugosi and K. Zeger, "Concept learning using complexity regularization," *IEEE Trans. on Information Theory*, vol. 42, pp. 48-54, 1996.

[18] L.L. Schumaker, *Spline functions: Basic theory*, Wiley-Interscience, New York, 1981.

[19] C.J. Stone, "The use of polynomial splines and their tensor products in multivariate function estimation," *Ann. Statistics*, vol. 22, pp. 118-184, 1994.

[20] Y. Yang and A.R. Barron, "An asymptotic property of model selection criteria, " *IEEE Trans. Inform. Theory,* vol. 44, pp. 95-116, 1998.

[21] Y. Yang and A.R. Barron, "Information-theoretic determination of minimax rates of convergence, " conditionally accepted by *Ann. Statistics*, 1999.

[22] Y. Yang, "Minimax nonparametric classification—part I: rates of convergence," accepted by *IEEE Trans. Inform. Theory,* 1998.

[23] Y. Yang, "Minimax nonparametric classification—part II: model selection for adaptation," Preprint #8, Department of Statistics, Iowa State University, 1998.