Model Selection for Nonparametric Regression

Yuhong Yang
Iowa State University

Abstract: Risk bounds are derived for regression estimation based on model selection over a unrestricted number of models. While a large list of models provides more flexibility, significant selection bias may occur with bias-correction based model selection criteria like AIC. We incorporate a model complexity penalty term in AIC to handle the selection bias. Resulting estimators are shown to achieve a trade-off among approximation error, estimation error and model complexity automatically without prior knowledge about the true regression function. As applications, we demonstrate adaptation property of these estimators over full and sparse approximation function classes with different smoothness. For high-dimensional function estimation by tensor product splines, we show with number of knots and spline order adaptively selected, least squares estimator converges at anticipated rates simultaneously for Sobolev classes with different interaction orders and smoothness parameters.

Keywords and phrases: Adaptive estimation; model complexity; model selection; nonparametric regression; rates of convergence.

Running Head: Model Selection for Regression

Yuhong Yang, 312 Snedecor Hall, Department of Statistics, Iowa State University, Ames, IA 50011-1210

Email: yyang@iastate.edu

Phone: (515) 294-2089

Fax: (515) 294-4040

1. Introduction

Consider the nonparametric regression model

$$Y_i = f(\mathbf{X}_i) + \epsilon_i,\ i = 1, ...n,$$

where $\mathbf{X}_i = (X_{i1}, ..., X_{id}) \in \mathcal{X} \subset R^d$ and the errors are i.i.d. with mean 0 and variance σ^2. We intend to estimate the underlying regression function f based on the observations $(\mathbf{X}_i, Y_i)_{i=1}^n$. To that end, a list of approximating linear models are considered. For instance, one may use finite dimensional polynomial, trigonometric, spline or wavelet models because of their good approximation capabilities to functions in various nonparametric classes and/or computational advantages. Let I be the model index and let Γ be the collection of the indices of the approximating models being considered. Unless stated otherwise, the model list Γ is fixed and does not depend on the sample size n.

For an estimator $\hat{f}$ of f, the discrepancy is measured by the average square error $ASE(\hat{f}) = \frac{1}{n}\sum_{i=1}^n \left(f(\mathbf{X}_i) - \hat{f}(\mathbf{X}_i)\right)^2$. This loss function measures how far away the function estimate is different from the true one on average at the design points $\mathbf{X}_i$, $i = 1, ...n$. The finite-dimensional family of regression functions in model I is denoted by $f_I(\mathbf{x}, \theta)$, $\theta \in R^{m_I}$ with m_I being the model dimension (for simplicity, we use θ instead of θ_I to denote the parameters in model I). In general, no relationship between the parameters in two different models in Γ is assumed.

For each model, we consider the least square estimator for the parameters. Let $M_I = M_{I,n}$ be the projection matrix corresponding to the design matrix of model I.

Example: $\mathcal{X} = [0, 1]^d$. Let $\varphi_1(\mathbf{x})$, $\varphi_2(\mathbf{x})$, ... be a collection of basis functions. Examples are tensor product splines basis of different orders and varying number and locations of knots, or wavelet basis with different resolutions. Let Γ be a collection of some finite subsets of $\mathcal{N} = \{1, 2, ...\}$. For $I = \{i_1, i_2, ...i_m\} \in \Gamma$, let the corresponding approximating model be

$$Y = \sum_{j=1}^m \theta_{i_j}\varphi_{i_j}(\mathbf{x}) + \epsilon.$$

Let

$$\Phi_I = \begin{pmatrix} \varphi_{i_1}(\mathbf{x}_1) & ... & \varphi_{i_m}(\mathbf{x}_1) \\ . & ... & . \\ \varphi_{i_1}(\mathbf{x}_n) & & \varphi_{i_m}(\mathbf{x}_n) \end{pmatrix}$$

be the design matrix, then the projection matrix is $M_I = \Phi_I \left(\Phi_I' \Phi_I\right)^{-1} \Phi_I'$ (here $(\cdot)^{-1}$ denotes the generalized inverse when the matrix is not invertible). The estimator of f based on model I is $\hat{f}_I(\mathbf{x}) = \sum_{j=1}^{m} \hat{\theta}_{i_j} \varphi_{i_j}(\mathbf{x})$, where $\hat{\theta}_{i_j}, 1 \leq j \leq m$ are the least squares estimators. Including more basis functions generally reduces the approximation error but increases the variability of the estimator due to estimating more parameters. A good trade-off between the approximation error and the number of parameters is desired. Another important issue is the choice of Γ. The more models we include in Γ, the better the model with the best trade-off between the approximation error and estimation error among all models being considered, but the harder it is in general to identify it (or other good ones) based on a finite sample.

AIC (Akaike (1972)) is a popular model selection criterion widely used in practice. The *AIC* type criteria (including *AIC*, C_p (Mallows (1973)), *FPE* (Akaike (1970))) add a bias-correction term (penalty) to the residual sum of squares. Shibata (1983) and Li (1987) showed that these criteria share an asymptotic optimality property which says roughly that the accuracy of the estimator based on the selected model is asymptotically the same as that based on the best model in the list (which is, of course, unknown). Similar conclusions were also obtained by Polyak and Tsybakov (1990). These results require certain constraints on number of models to be considered. In fact, to satisfy a summability condition for Theorem 2.1 in Shibata (1983) and condition A.3 in Li (1987), the number of models with dimension m can only be allowed to increase at a polynomial rate in m. When there are exponentially many or more models in competition as in the subset selection case, the chance of selecting a bad model can be substantial using a bias-correction based criteria. The probability that the empirical behavior of at least one model in the list is very much different from expected may be large because of the addition of small errors over a large number of models. The consequence is that comparison of the criterion values might be dramatically different from comparison of risks of the models, based on which the *AIC* type of criteria are derived. Significant selection bias then occurs. In this work, we provide criteria which select good models even when there are exponentially many or more models.

Advantages of consideration of a large number of models have been shown in statistical theory. Research results in that direction include, for example, function estimation by minimum

description length criterion by Barron and Cover (1991), wavelet estimation for Besov classes by Donoho, Johnstone et al (e.g., 1994, 1995), neural network estimation by Barron (1993, 1994), penalized likelihood estimation by Yang and Barron (1998). A general theory on model selection for function estimation was proposed by Barron, Birgé, and Massart (1997).

In this paper, we propose model selection criteria related to AIC and derive general performance bounds. We show the resulting estimators automatically achieve a trade-off among approximation error, estimation error and model complexity. Consequences on adaptive estimation are provided as applications. It is demonstrated that with suitable subset selection, a rich collection of functions characterized by sparse approximation can be adaptively estimated near optimally, yet essentially no price is paid in terms of risk for the much more extensive search when there is in fact no need to do subset selection. It is also shown that when spline orders, numbers of knots, and explanatory variables are selected, a tensor product spline estimator converges at anticipated rate without knowing the interaction order and smoothness of the true regression function.

The paper is organized as follows. In Section 2, model complexity is defined, followed by examples in Section 3. Model selection criteria are proposed in Section 4 and main results of performance bounds are presented in Section 5. Some implications on adaptive estimation are studied in Section 6. Proofs of the results are deferred to an appendix.

2. Model Complexity

The criteria we consider incorporate a model complexity with the residual sum of squares and a bias-correction term. The addition of the model complexity term was suggested by the author's thesis advisor Andrew Barron at Yale University. The ideas of using complexity in statistical estimation have been explored in Rissanen (e.g., 1983, 1984, 1987), Barron (e.g., 1985, 1994), Wallace and Boulton (1987), Hall and Hannan (1988), Barron and Cover (1991), Yu and Speed (1992), and others. The model complexity here does not directly refer to the complication of a specific model (e.g., the number of parameters in the model, which is already accounted for with the bias-correction term), it rather characterizes our view on the models. As seen in the proof of the results, addition of the complexity penalty term regulates the competition among the models to ensure good behavior of the selected model.

Definition: The complexity of a model I in a list Γ is a positive number C_I satisfying the condition:

$$\sum_{I \in \Gamma} 2^{-C_I} \leq 1. \tag{1}$$

There are two interpretations of the model complexity. Based on information theory, C_I can be viewed as the codelength of a prefix-free code to describe the model index. The code can be naturally designed to reflect organization of the models. In the definition, we do not require C_I to be integers. If they are, then there exists a uniquely decodable code with C_I as the codelength (see e.g., Cover and Thomas (1991, Chapter 5)). We may also interpret 2^{-C_I} as a prior probability on model I. However, the criteria we give will not in general be a Bayes procedure (in particular, there is no averaging with respect to distributions for the parameters).

3. Examples of Models and Complexity Assignments

We consider two types of models in this section for examples, which will be further studied later in Section 6 for demonstration of our main results.

Usually models of interest can be indexed in terms of a few non-negative integers. Then description of the model index boils down to description of integers. To describe an integer with a known upper bound L, we may use $\log_2 L$ bits (ignoring rounding up). If there is no known upper bound, it is sufficient to use $\log^*(m) = \log_2(m) + 2\log_2\log_2(m+1)$ bits to describe integer m for this case (cf. Elias (1975) and Rissanen (1983)).

Let $\Phi = \{\phi_1, ..., \phi_k, ...\}$ be a fixed choice of basis functions in $L^2[0,1]^d$.

I). *Nested models (complete models).* Only one model is considered for each dimension. For $m \geq 1$, the family is

$$f_m(x, \theta) = \sum_{i=1}^{m} \theta_i \phi_i(x), \theta = (\theta_1, ..., \theta_m) \in R^m.$$

For this case, the model dimension is naturally an index for the model. Thus one may use $\log^*(m)$ bits to describe integer m, leading to the choice of $C_m = \log^*(m)$. Note that for this case, the model complexity C_m is asymptotically negligible compared to model dimension m as m increases.

II). *Subset models.* Alternatively, one can consider sparse subset models for more flexibility. Let N_k $(k \geq 1)$ be an increasing sequence with $N_k/k \to \infty$ as $k \to \infty$. It is used to control sparsity of the subset models. Let $I = (k, \mathbf{l})$, where $\mathbf{l} = (l_1, .., l_k)$ satisfying $1 \leq l_j \leq N_j$ for

$1 \le j \le k$. Given I, the family is

$$f_I(x,\theta) = \sum_{1\le j\le k} \theta_j \varphi_{l_j}(x).$$

Clearly when N_j/j is big, the jth term in the models is chosen with a large freedom (from $\{\varphi_1(x), ..., \varphi_{N_j}(x)\}$). To describe the model index I, we first describe k using $\log^* k$ bits, then describe $\mathbf{l}$ using $\log N_1 + \log N_2 + ... + \log N_k$ bits. Thus we assign complexity $C_I = \log^* k + \sum_{i=1}^k \log N_i$. Using Stirling's formula, with $N_k \asymp k^\tau$ $(\tau > 1)$, C_I is seen to be of order $k \log k$.

4. Model Selection Criteria

Let $Y^n = (Y_1, Y_2, ..., Y_n)^T$, and let $\hat{Y}_I$ be the projection of Y^n into the space spanned by the columns of the design matrix of model I. Let $r_I = r_{I,n}$ be the rank of the design matrix M_I.

I). σ^2 is known. We propose the following criterion ABC:

$$ABC(I) = \|Y^n - \hat{Y}_I\|^2 + 2r_I\sigma^2 + \lambda\sigma^2 C_I, \tag{2}$$

where λ is a positive constant. The difference between ABC and AIC (C_p) is the addition of the model complexity penalty term.

II). When σ^2 is unknown, one may replace σ^2 in ABC by a consistent estimator $\hat{\sigma}^2$ independent of the models in Γ. For instance, σ^2 can be estimated using nearest neighbor method (see, e.g., Stone (1977)).

III). σ^2 is unknown but an upper bound on σ^2 (say σ_0^2) is available. The criterion is

$$ABC'(I) = \left(1 + \frac{2r_I}{n - r_I}\right)\left(\|Y^n - \hat{Y}_I\|^2 + \lambda\sigma_0^2 C_I\right). \tag{3}$$

Note that without the model complexity term $\lambda\sigma_0^2 C_I$, the criteria is FPE (Akaike (1970)), which is derived based on individual estimation of σ^2 in each model.

5. Main Results

In applications, the explanatory variables can be either random or fixed such as equally spaced design. We give results conditioned on the explanatory variables as well as unconditional ones.

We assume that ϵ_i are i.i.d. $\sim N(0, \sigma^2)$, and for the random design case, they are independent of $\{X_i,\ i = 1, ...n\}$. The normality assumption is essential in our analysis to derive risk (or

in probability) bounds valid for every regression function and without any restriction on the list of operating linear models (see the remark after the proof of Theorem 1). Let E_n denote the expectation with respect to the randomness of the error ϵ_i, $i = 1, ...n$ conditioned on the explanatory variables X_i, $i = 1, ...n$.

Let $f_n = (f(X_1), ..., f(X_n))^T$, and let $\bar{f}_I = M_I f_n$ be the projection of f_n into the column space of the design matrix. Let

$$R_n(f; I) = \frac{1}{n}\|f_n - \bar{f}_I\|^2 + \frac{r_I\sigma^2}{n} + \frac{\lambda\sigma^2 C_I}{n},$$

$$R_n^*(f; \Gamma) = \min_{I\in\Gamma} R_n(f; I), \quad I_n^* = \arg\min R_n(f; I).$$

The meanings of these important quantities are as follows. The first term $\|f_n - \bar{f}_I\|^2/n$ in $R_n(f; I)$ is the approximation error of f_n by model I; the second term $r_I\sigma^2/n$ is the estimation error. The sum of these two terms is the overall risk of the estimator based on model I, i.e., $E_n\left(\|f_n - \hat{Y}_I\|^2/n\right) = \|f_n - \bar{f}_I\|^2/n + r_I\sigma^2/n$. Thus $R_n^*(f; \Gamma)$ characterizes the best trade-off among approximation error, estimation error and model complexity over all models in Γ. We call $R_n^*(f, \Gamma)$ index of resolvability of the unknown function f by models in Γ following a terminology of Barron and Cover (1991). Ideally, model I_n^* should be used. But unfortunately, it could be known only if the true function f were known exactly at the sites X_i, $i = 1, 2, ..., n$, hence the need of model selection criteria. We will compare the performance of estimators based on model selection to the index of resolvability.

Let $ASE(I) = \|f_n - \hat{Y}_I\|^2/n$ denote the average square error of the estimator $\hat{f}_I = f_I(\cdot, \hat{\theta})$ from model I.

I). σ^2 is known. Let $\hat{I}_n$ be the model selected by minimizing ABC over I in Γ.

Theorem 1. When $\lambda \geq 5.1$, we have

$$E_n\left(ASE(\hat{I}_n)\right) \leq \xi R_n^*(f; \Gamma), \tag{4}$$

where ξ is a constant depending only on λ. If in addition, $nR_n^*(f; \Gamma) \to \infty$, then with probability tending to 1, we have

$$ASE(\hat{I}_n) \leq B_0 R_n^*(f; \Gamma), \tag{5}$$

for some constant B_0 depending only on λ.

Remarks:

1. There is no restriction on the size of the list Γ and the risk bound is still valid if Γ is chosen to depend on n.

2. For (4), there is no requirement at all on the underlying function f.

3. In fact, a stronger inequality than (4) holds, namely,

$$E_n\left(ASE(\hat{I}_n)+\frac{\lambda\sigma^2 C_{\hat{I}_n}}{n}\right)\leq \xi R_n^*(f;\Gamma). \tag{6}$$

   Thus the complexity of the selected model is also well controlled.

Theorem 1 characterizes the criterion ABC with good performance bounds on ASE of the selected model in terms of the index of resolvability. Under smoothness conditions on the function f, $R_n^*(f;\Gamma)$ can be easily evaluated through approximation theory for various choices of basis functions. Then upper bounds on the convergence rates of ASE are determined.

The condition $nR_n^*(f;\Gamma)\to\infty$ is satisfied for a typical function in a nonparametric class. It can fail only when there exists a subsequence n_j such that $I_{n_j}^*$ stays the same and $\|f_n-\overline{f}_{I_{n_j}^*}\|^2$ stays bounded. Thus $nR_n^*(f;\Gamma)\to\infty$ if for each $I\in\Gamma$, $\|f_n-\overline{f}_I\|^2\to\infty$ (note that there is no division of n for $\|f_n-\overline{f}_I\|^2$).

For the random design case, the accuracy with respect to the randomness from both errors $\{\varepsilon_i\}$ and independent variables $\{X_i\}$ is of interest. Let m_I denote the number of free parameters in model I. Let

$$\mathcal{R}_n^*(f;\Gamma)=\min_{I\in\Gamma}\left(\inf_{\theta\in R^{m_I}}E\left(f(\mathbf{X})-f_{I,\theta}(\mathbf{X})\right)^2+\frac{m_I\sigma^2}{n}+\frac{\lambda\sigma^2 C_I}{n}\right),$$

be an index of (unconditional) resolvability. From now on, λ is assumed to be taken at least 5.1 unless stated otherwise.

Corollary 1: If $\{X_i\}_{i=1}^n$ are i.i.d. then $E\left(ASE(\hat{I}_n)\right)\leq\xi\mathcal{R}_n^*(f;\Gamma)$.

Note again that the (integrated) approximation error $\inf_{\theta\in R^{m_I}}E\left(f(\mathbf{X})-f_{I,\theta}(\mathbf{X})\right)^2$ is known for f in various nonparametric function classes using familiar bases including polynomial, trigonometric, spline and wavelet. Then by balancing the approximation error bound, estimation error $m_I\sigma^2/n$, and model complexity $\lambda\sigma^2 C_I/n$, upper bounds on convergence rates of the mean average squared error are obtained.

If f is actually well approximated by simple models in the model list Γ, we have the following result. For a constant $A > 0$, let $K_{n,A} = \{I : I \in \Gamma, R_n(f; I) \leq AR_n^*(f;\Gamma)\}$ be the collection of all the models that produce index $R_n(f; I)$ within a multiple of the ideal index $R_n^*(f;\Gamma)$.

Theorem 2.

1. If for every $A > 0$, $\lim_{n\to\infty} \sup_{I\in K_{n,A}} C_I/r_I = 0$, then

$$\frac{ASE(\hat{I}_n)}{\frac{1}{n}\|f_n - \overline{f}_{I_n^*}\|^2 + \frac{r_{I_n^*}\sigma^2}{n}} \longrightarrow 1 \qquad \text{in probability.} \tag{7}$$

2. If the model list $\Gamma = \Gamma_n$ depends on n and $\sup_{I\in\Gamma_n} C_I/r_I \to 0$, then

$$\frac{E_n\left(ASE(\hat{I}_n)\right)}{\inf_{I\in\Gamma_n}\left(\frac{1}{n}\|f_n - \overline{f}_I\|^2 + \frac{r_I\sigma^2}{n}\right)} = 1 + o(1). \tag{8}$$

Remark: If for each $I \in \Gamma$, $\|f_n - \overline{f}_I\|^2 \to \infty$, then it suffices to check the condition for (7) with only $A = B_0$ (B_0 is the same constant as in (5)).

This theorem says that if the models with good trade-off between complexity and accuracy have small complexities compared to model dimensions, then with the ABC criterion, we can do asymptotically as well as if we knew I_n^* in advance. Illustration of this result will be given in the next section, where it is seen that when some sparse subset models are considered in addition to complete models, functions characterized by sparse approximation can be much better estimated, and in the mean time, no essential price in terms of risk is paid for selection among the much larger list of models when there happens to be no need for the subset selection.

II. σ^2 is unknown and is replaced by a consistent estimator $\hat{\sigma}^2$ in ABC. Let $\hat{I}_n'$ be the selected model.

Theorem 3. When $\lambda \geq 5.1$, we have

$$ASE(\hat{I}_n') = O_p\left(R_n^*(f;\Gamma)\right)$$

and (7) still holds under the corresponding condition.

III. σ^2 is unknown but an upper bound σ_0^2 on σ^2 is known.

Theorem 4: Let $\hat{I}_n$ be the model selected by minimizing ABC' over Γ. Assume for each $I \in \Gamma$, $\|f_n - \overline{f}_I\|^2 \to \infty$ and $R_n^*(f;\Gamma) \to 0$. Then when $\lambda \geq 40$, with probability tending to 1, we have

$$ASE(\hat{I}_n) \leq \left(B_1 \sigma_0^2/\sigma^2\right) R_n^*(f;\Gamma),$$

where B_1 is a constant depending only on λ.

A result similar to Theorem 2 also holds for ABC'. As mentioned before, the condition $\|f_n - \overline{f}_I\|^2 \to \infty$ is satisfied for a typical nonparametric function. The condition $R_n^*(f;\Gamma) \to 0$ holds if the true function f can be well approximated by some models in Γ with not too large model complexities.

6. Application on Adaptive Estimations

In recent years, adaptive estimation becomes a main topic on nonparametric curve estimation (see, e.g., Efroimovich and Pinsker (1984), Efroimovich (1985), Härdle and Marron (1985), Barron and Cover (1991), Lepskii (e.g., 1991), Donoho, Johnstone et al (1994, 1996), Birgé and Massart (1996), Brown and Low (1995), Yang and Barron (1998), and Yang (1997)). Model selection provides a practical mean to obtain estimators that are adaptive to many possible different characteristics of the unknown underlying function (see, e.g., Barron and Cover (1991), Barron, Birgè and Massart (1997)), and Yang and Barron (1998)). The results in the proceeding section provide risk bounds in terms of approximation capability, model dimension, and model complexity of the operating models. In Theorem 1, no condition on the function to be estimated is even required for (4), hence the risk bounds hold for all regression functions. To provide estimators that are adaptive over multiple target functions classes, one can construct different kinds of approximating models each suitable for one or more classes and then use the model selection criterion to choose a good one based on data. To prove adaptation, one just need to examine the index of resolvability for each class and show it converges at an appropriate rate.

In this section, we demonstrate the above point by showing adaptation with respect to interaction order and smoothness, adaptations with respect to both full and sparse approximation sets of functions. We concentrate on criterion ABC with σ^2 known. When σ^2 is unknown, with the criterion in Section 4.II or 4.III, based on Theorems 3 and 4, analogous results hold on the average squared error of the selected model. But the conclusions are weaker, in terms

of convergence in probability instead of convergence in risk. The explanatory variables $\{X_i\}_{i=1}^n$ are assumed to be independent and uniformly distributed on $[0,1]^d$ in this section.

6.1. Adaptation with respect to interaction order and smoothness

For high dimensional function estimation with series expansion methods, complete models with terms up to certain orders in the expansion usually do not produce satisfactory estimators due to curse of dimensionality. In contrast, parsimonious subset models may significantly increase estimation accuracy. In this subsection, we consider adaptation by model selection over Sobolev classes with different interaction orders and smoothness.

For $r \geq 1$, let $\mathbf{z}_r = (z_1, ..., z_r) \in [0,1]^r$. For $\mathbf{k} = (k_1, ..., k_r)$ with nonnegative integer components k_i, define $|\mathbf{k}| = \sum_{i=1}^r k_i$. Let $D^{\mathbf{k}}$ denote the differentiation operator $D^{\mathbf{k}} = \partial^{|\mathbf{k}|}/\partial z_1^{k_1} \cdot \cdot \cdot \partial z_r^{k_r}$. For an integer α, define 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 $\| g \|_{W_2^{\alpha,r}} \leq C$. It is a r-dimensional Sobolev class. Now consider the following function classes on $[0,1]^d$ of different interaction orders and smoothness:

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

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

$$\cdots$$

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

with $\alpha \geq 1$ and $C > 0$. The simplest class $S_1(\alpha; C)$ contains additive functions (no interaction), 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. Then by results of Yang and Barron (1997, Theorem 10), the minimax rate of convergence under square L_2 loss for estimating a regression function in $S_r(\alpha; C)$ is $n^{-2\alpha/(2\alpha+r)}$ for $1 \leq r \leq d$ (the result is suggested by the heuristic dimensionality reduction principle of Stone (1985)). Note that the convergence rate does not depend on the input dimension d. Thus low order interaction classes are worth exploring for better accuracy.

Suppose the true regression function is in one of the classes in $\{S_r(\alpha; C) : 1 \leq r \leq d, \alpha \geq 1, C > 0\}$. Stone (1994) proposed tensor product spline models for estimating such a high-dimensional function (or more generally its components of different interaction orders in a

functional ANOVA decomposition) in a general context including density estimation, nonparametric regression, and conditional density estimation. He demonstrated that suitable splines models can result in estimators converging at expected rates. However, his results require knowledge of smoothness parameters and interaction order, therefore are not adaptive. For regression, a similar non-adaptive result was obtained earlier by Chen (1991) under more restrictive condition with fixed balanced design. We here show that by model selection, an adaptive estimator can be obtained based only on data in the regression setting. Previously, adaptation results with respect to smoothness in the context of generalized additive modeling was obtained by Burman (1990) using cross-validation.

Let $\varphi_{m,q,1}(x), \varphi_{m,q,2}(x), ...,\varphi_{m,q,m}(x)$ be the B-spline basis of order q (piecewise polynomial of order less than q) on $[0,1]$ with $m-q+2$ $(m \geq q)$ equally spaced knots. For $1 \leq r \leq d$, 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 subverter 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 $1 \leq i_l \leq m_l, 1 \leq l \leq r$. Then given the spline order $\mathbf{q}_r$ and $\mathbf{m}_r$, the tensor products

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

have interaction order r.

Given $r, \mathbf{q}_r, \mathbf{m}_r$, consider the linear combinations of the functions in (9). Let $I = (r, \mathbf{q}_r, \mathbf{m}_r)$ be the model index. When $r < d$, the functions in (9) are not all linearly independent, with model dimension of order $\Pi_i^r m_i$ in any case. Then we choose a subset of independent basis and the coefficients are estimated by least squares method based on the observations $(\mathbf{X}_i, Y_i)_{i=1}^n$. Let $\hat{f}_I = \hat{f}_{(r,\mathbf{q}_r,\mathbf{m}_r)}$ denote the corresponding function estimator.

To adaptively select $I = (r, \mathbf{q}_r, \mathbf{m}_r)$ by criterion ABC, we need to assign model complexity. To describe the model index, 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. Thus we assign model complexity $C_I = \log_2 d + \sum_{j=1}^r \log^* q_j + \sum_{j=1}^r \log^* m_j$. Let $\hat{I}_n$ be the model selected by ABC over all valid choices of $(r, \mathbf{q}_r, \mathbf{m}_r)$. We have the following result.

Theorem 5: The estimator $\hat{f}_{\hat{I}_n}$ has mean average square error for the enlarged Sobolev classes bounded as follows

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

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

It is anticipated that the minimax rate of convergence of mean ASE is of the same order as the minimax risk under L_2 loss (cf. Chen (1990, pp. 1863-1864)). If that is confirmed, then without knowing the true interaction order r and the smoothness parameter α, the estimator based on ABC is minimax-rate adaptive over these enlarged Sobolev classes.

6.2. Adaptation with respect to full approximation sets of functions

Let $\Phi = \{\phi_1, ..., \phi_k, ...\}$ be a fixed choice of fundamental sequence in $L^2[0,1]^d$ (that is, linear combinations are dense in $L^2[0,1]^d$). Let $\Upsilon = \{\gamma_0, ..., \gamma_k, ...\}$ for which $\gamma_k \downarrow 0$ as $k \to \infty$. Let $\eta_0(f) = \| f \|_2$ and $\eta_k(f) = \min_{\{a_i\}} \| f - \sum_{i=1}^k a_i \phi_i \|_2$ for $k \geq 1$ be the k-th degree of approximation of $f \in L^2[0,1]^d$ by the system Φ. Let $\mathcal{F}(\Upsilon, \Phi)$ be all functions in $L_2[0,1]^d$ with the approximation errors bounded by Υ, i.e.,

$$\mathcal{F}(\Upsilon, \Phi) = \{f \in L_2[0,1]^d : \eta_k(f) \leq \gamma_k, k = 0, 1, ...\}.$$

They are called full approximation sets of functions (Lorentz (1966)). Some familiar function classes, e.g., Sobolev classes are essentially full approximation sets (in the sense that each is contained between two full approximation sets of essentially the same size).

From the defining property, to estimate a function in full approximation sets, it is natural to consider the finite dimensional families

$$f_m(x, \theta) = \sum_{i=1}^m \theta_i \phi_i, \theta = (\theta_1, ..., \theta_m) \in R^m$$

for $m \geq 1$. Intuitively, a good choice of m should balance the approximation error and estimation error, resulting in an optimal estimator. Indeed, Yang and Barron (1997) showed that for full approximation sets of functions, the minimax square L_2 risk is of order $\frac{m}{n}$ (estimation error) or γ_m^2 (approximation error) when they are balanced (i.e., $\frac{m}{n} \asymp \gamma_m^2$) under a general condition on the approximation error sequence Υ, namely, there exist $0 < c' < c < 1$ such that

$$c' \gamma_m \leq \gamma_{2m} \leq c \gamma_m. \tag{10}$$

It is true for $\gamma_m \sim m^{-\alpha}$ and also for $\gamma_m \sim m^{-\alpha}(\log m)^{\beta}$, $\alpha > 0$, $\beta \in R$ (which covers classical classes such as Sobolev). Let $m_n(\Upsilon)$ be a sequence of model dimensions that attains the balance of γ_m^2 and m/n. Then $m_n(\Upsilon)/n$ is the minimax rate of convergence for the class $\mathcal{F}(\Upsilon, \Phi)$. For instance, the minimax rate is $n^{-2\alpha/(1+2\alpha)}(\log n)^{-2\beta/(1+2\alpha)}$ if $\gamma_m \sim m^{-\alpha}(\log m)^{\beta}$. In applications, of course, we do not know the order of approximation error γ_m, therefore can not choose $m_n(\Upsilon)$ directly. To use ABC, we assign model complexity $C_m = \log^* m$ as in Section 3. Applying Corollary 1, together with that C_m is negligible compared to m, we have the following conclusion.

Theorem 6: Let $\hat{I}_n$ be selected using criterion ABC with $\lambda \geq 5.1$. Then the estimator $\hat{f}_{\hat{I}_n}$ satisfies

$$\sup_{f \in \mathcal{F}(\Upsilon,\Phi)} E\left(ASE(\hat{I}_n)\right) = O\left(m_n(\Upsilon)/n\right)$$

simultaneously for all choices of Υ satisfying (10).

This result shows that the estimator based on order selection is adaptive among the full approximation sets of functions without knowledge of the approximation error.

6.3. Subset selection for sparse approximation sets of functions

Instead of full approximation, one can also consider sparse approximation. Let Φ and Υ be as in the previous subsection with Υ satisfying the condition in (10). As in Section 3, Let $N_k > k$ $(k \geq 1)$ be a chosen increasing sequence of integers satisfying $\liminf N_k/k = \infty$ and let $\mathcal{N} = \{N_1, N_2, ...\}$. For simplicity, take N_k of order k^{τ} for some $\tau > 1$. Let $\widetilde{\eta}_k(g) = \min_{l_1 \leq N_1, l_2 \leq N_2 ..., l_k \leq N_k} \min_{\{a_i\}} \parallel g - \sum_{i=1}^{k} a_i \phi_{l_i} \parallel_2$ be called the k-th degree of sparse approximation of $g \in L^2[0,1]^d$ by the system Φ under the chosen sparsity constraint $\mathcal{N}$. Here for $k = 0$, there is no approximation and $\widetilde{\eta}_0(g) = \parallel g \parallel_2$. The k-th term used to approximate g is selected from N_k basis functions. Let $\mathcal{S}(\Upsilon, \Phi, \mathcal{N})$ be all functions in $L_2[0,1]^d$ with the sparse approximation errors bounded by Υ, i.e.,

$$\mathcal{S}(\Upsilon, \Phi, \mathcal{N}) = \{g \in L_2[0,1]^d : \widetilde{\eta}_k(g) \leq \gamma_k, k = 0, 1, ...\}.$$

Sparse approximation provides much more freedom of approximation yet can still enjoy simplicity of linearity to a great extent. Minimax bounds are given for sparse approximation sets in Yang and Barron (1997). For instance, if $\gamma_k \sim k^{-\alpha}$, $\alpha > 0$, then the minimax rate is

between $\left(n \log^{2\alpha} n\right)^{-2\alpha/(1+2\alpha)}$ and $(n/\log n)^{-2\alpha/(1+2\alpha)}$.

Corresponding to sparse approximation, consider subset models:

$$f_I(x,\theta) = \sum_{1 \le j \le k} \theta_j \varphi_{l_j}(x),$$

where $I = (k, \mathbf{l})$ with $\mathbf{l} = (l_1, ..., l_k)$ satisfying $1 \le l_j \le N_j$. The model complexity C_I can be assigned as in Section 3. From Corollary 1, we have the following result.

Theorem 7: Let $\hat{I}_n$ be selected among the subset models using criterion ABC with $\lambda \ge 5.1$. Then the risk is bounded by

$$\sup_{f \in \mathcal{S}(\Upsilon,\Phi,\mathcal{N})} E\left(ASE(\hat{I}_n)\right) = O\left(m_n(\Upsilon) \log n/n\right)$$

simultaneously for all choices of Υ satisfying (10).

Remark: For the above result, Φ and $\mathcal{N}$ are held fixed. Generally, one is allowed to consider different choices of Φ and $\mathcal{N}$. A similar result still holds with suitable assignment of model complexity.

Since $\mathcal{S}(\Upsilon, \Phi, \mathcal{N})$ contains $\mathcal{F}(\Upsilon, \Phi)$, the minimax square L_2 risk of $\mathcal{S}(\Upsilon, \Phi, \mathcal{N})$ is lower bounded by order $m_n(\Upsilon)/n$. Thus from Theorem 7, by the subset selection, the mean ASE is within at most a logarithmic factor of the anticipated rates of convergence automatically over the different sparse approximation sets. In contrast, if only complete models are considered (i.e., considering only full approximation), the rate in general could be much worse as seen in the following example.

Example: Sparse Fourier Series. Consider Fourier basis on $[0,1]$: $\phi_1(x) = 1, \phi_2(x) = \sin(2\pi x), \phi_3(x) = \cos(2\pi x), \cdots$. With γ_k of order $k^{-\alpha}$ $(\alpha \ge 1)$, the full approximation set $\mathcal{F}(\Upsilon, \Phi)$ is essentially a periodic Sobolev class with smoothness parameter α (i.e., $\mathcal{F}(\Upsilon, \Phi)$ is contained between two Sobolev balls of different radii, see, e.g., Lorentz (1966)). From Theorem 6, by considering the complete models, the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)}$ is achieved using the selected model. However, if the true regression function is in $\mathcal{S}(\Upsilon, \Phi, \mathcal{N})$ with $N_k = k^2$, then the rate of convergence of the estimator for the larger class $\mathcal{S}(\Upsilon, \Phi, \mathcal{N})$ deteriorates to at least $n^{-\alpha/(\alpha+1)}$ (see Yang and Barron (1997)). When the sparse subset models are considered, from Theorem 7, the rate improves to $n^{-2\alpha/(2\alpha+1)} \log n$.

We suspect a logarithmic factor in the risk bound in Theorem 7 is needed for the sparse approximation sets. However, it still appears even if f is in a full approximation set, which we know from Theorem 6 is not necessary when full approximation sets alone are considered. We next show by considering both types of models, we get the advantages of both of them as if we knew which type to consider in advance. To that end, let the model index be $I = (\delta, H)$, where $\delta = 0$ indicates it is a complete model for which case, $H = m$, and $\delta = 1$ indicates it is a sparse subset model for which case, $H = (k, \mathbf{l})$. To describe the new model index, we just need to add an extra bit describing the value of δ. Then H can be described as before. This gives us $C_{(0,m)} = 1 + \log^* m$ for a complete model and $C_{(1,(k,\mathbf{l}))} = 1 + \log^* k + \sum_{i=1}^{k} \log N_i$ for a subset model. Let Γ_1 and Γ_2 denote the lists of complete models and sparse subset models respectively. Then it can be easily verified that

$$R_n(f; \Gamma_1 \cup \Gamma_2) = \lambda \log 2/n + \min\left(R_n(f; \Gamma_1), R_n(f; \Gamma_2)\right).$$

That is, the index of resolvability when both types of models are considered is the minimum of the indices of resolvability when they are considered separately plus $\lambda \log 2/n$, which is negligible for nonparametric estimation. Let $\hat{I}_n$ be the selected model. Then we have the following result.

Corollary 2: The estimator $\hat{f}_{\hat{I}_n}$ satisfies

$$\sup_{f \in \mathcal{F}(\Upsilon, \Phi)} E\left(ASE(\hat{I}_n)\right) = O\left(m_n(\Upsilon)/n\right).$$

and

$$\sup_{f \in \mathcal{S}(\Upsilon, \Phi, \mathcal{N})} E\left(ASE(\hat{I}_n)\right) = O\left(m_n(\Upsilon)\log(n)/n\right)$$

simultaneously for all choices of Υ satisfying (10). In addition, if f has approximation error $c_1 k^{-\alpha} \leq \eta_k(f) \leq c_2 k^{-\alpha}$ for some constants c_1 and c_2 and $\alpha > 0$, then

$$\frac{ASE(\hat{I}_n)}{\inf_{I \in \Gamma_1 \cup \Gamma_2} ASE(I)} \to 1 \text{ in probability.}$$

This corollary shows the advantages of enlarging the list of models to be selected from. The resulting estimator converges at least near optimally for sparse approximation sets; stay at the anticipated rates for full approximation sets; and furthermore, for a function with approximation error regularly decreasing, it has ASE asymptotically equivalent in probability to the smallest

value one can get with knowledge of which model is the best in advance. The last statement in Corollary 2 follows directly from Theorem 2.

Acknowledgments

The author is very grateful to Andrew Barron for his invaluable insights, suggestions, and guidance during this work. Discussions with David Olive were very helpful. He also wants to thank the referees for their comments on an earlier draft of this paper.

Appendix: Proofs of the Results

For the proofs of Theorems 1 and 2, without lose of generality, assume $\sigma^2 = 1$.

Proof of Theorem 1: Let $e_n = (\epsilon_1, ..., \epsilon_n)^T$. Using $\hat{Y}_I = M_I Y_n$, $Y_n = f_n + e_n$, and expanding the square, one obtains

$$\begin{aligned} & \|Y_n - \hat{Y}_I\|^2 + 2r_I + \lambda C_I \\ = \; & \|f_n - \hat{Y}_I\|^2 + e_n^{'} e_n + 2e_n^{'}(f_n - M_I f_n) + 2(r_I - e_n^{'} M_I e_n) + \lambda C_I. \end{aligned}$$

Let $\mathrm{rem}_1(I) = e_n^{'}(f_n - M_I f_n)$ and $\mathrm{rem}_2(I) = (r_I - e_n^{'} M_I e_n)$. For simplicity, denote I_n^*, $R_n(f;I)$, and $R_n^*(f;\Gamma)$ by I_n, $R_n(I)$ and R_n^* respectively. By definitions of $\hat{I}$ and I_n, and using projection property of M_I, we have

$$\begin{aligned} & \|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}} \\ = \; & ABC(\hat{I}) - e_n{}' e_n - 2\mathrm{rem}_1(\hat{I}) - 2\mathrm{rem}_2(\hat{I}) \\ \leq \; & ABC(I_n) - e_n{}' e_n - 2\mathrm{rem}_1(\hat{I}) - 2\mathrm{rem}_2(\hat{I}) \\ = \; & \|f_n - \hat{Y}_{I_n}\|^2 + \lambda C_{I_n} + 2\mathrm{rem}_1(I_n) + 2\mathrm{rem}_2(I_n) - 2\mathrm{rem}_1(\hat{I}) - 2\mathrm{rem}_2(\hat{I}) \\ = \; & \|f_n - M_I f_n\|^2 + r_{I_n} \; + \lambda C_{I_n} + 2\mathrm{rem}_1(I_n) + \mathrm{rem}_2(I_n) - 2\mathrm{rem}_1(\hat{I}) - 2\mathrm{rem}_2(\hat{I}). \end{aligned}$$

Using $nR_n(\hat{I}) = \|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}} + \mathrm{rem}_2(\hat{I})$, we obtain

$$\begin{aligned} \frac{R_n(\hat{I})}{R_n^*} & \leq \frac{nR_n^* + 2\mathrm{rem}_1(I_n) + \mathrm{rem}_2(I_n) - 2\mathrm{rem}_1(\hat{I}) - \mathrm{rem}_2(\hat{I})}{nR_n^*} \\ & = 1 + \frac{2\mathrm{rem}_1(I_n) + \mathrm{rem}_2(I_n) - 2\mathrm{rem}_1(\hat{I}) - \mathrm{rem}_2(\hat{I})}{nR_n^*} \; . \end{aligned}$$

If we can show that for any $0 < \delta < 1$, there exist $g_1(\delta) > 0$ and $g_2(\delta) > 0$ such that for each sample size n, with probability no less than $1 - 3\delta$, we have that for all $I \in \Gamma$

$$|\mathrm{rem}_1(I)| \leq \tau_1 \left(nR_n(I) + g_1(\delta) \right). \tag{11}$$

$$|\mathrm{rem}_2(I)| \le \tau_2 \left(nR_n(I) + g_2(\delta)\right), \tag{12}$$

where τ_1 and τ_2 are two constants satisfying $2\tau_1 + \tau_2 < 1$, then with probability no less than $1 - 3\delta$,

$$\frac{R_n(\hat{I})}{R_n^*} \le 1 + (2\tau_1 + \tau_2)\frac{R_n(\hat{I})}{R_n^*} + 2\tau_1\left(1 + \frac{2g_1(\delta)}{nR_n^*}\right) + \tau_2\left(1 + \frac{2g_2(\delta)}{nR_n^*}\right). \tag{13}$$

Using that $nR_n^* \ge r_{I_n} \ge 1$, we know that with probability no less than $1 - 3\delta$,

$$\frac{R_n(\hat{I})}{R_n^*} \le \frac{1 + 2\tau_1(1 + 2g_1(\delta)) + \tau_2(1 + 2g_2(\delta))}{1 - 2\tau_1 - \tau_2}, \tag{14}$$

and

$$\begin{aligned} & \frac{\|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}}}{nR_n^*} \\ \le\ & \frac{nR_n(\hat{I}) + |\mathrm{rem}_2(\hat{I})|}{nR_n^*} \\ \le\ & (1 + \tau_2)\frac{R_n(\hat{I})}{R_n^*} + \tau_2 g_2(\delta) \\ \le\ & \frac{(1 + \tau_2)(1 + 2\tau_1(1 + 2g_1(\delta)) + \tau_2(1 + 2g_2(\delta))}{1 - 2\tau_1 - \tau_2} + \tau_2 g_2(\delta). \end{aligned} \tag{15}$$

Let $Z_n = \left(\|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}}\right)/(nR_n^*)$, and $\xi_1 = (1 + \tau_2)(1 + 2\tau_1 + \tau_2)/(1 - 2\tau_1 - \tau_2)$, $\xi_2 = \overline{\lambda}\left(2(1 + \tau_2)(2\tau_1 + \tau_2)/(1 - 2\tau_1 - \tau_2) + \tau_2\right)$ for some constant $\overline{\lambda}$ (to be given later). If $g_1(\delta) = g_2(\delta) = \overline{\lambda}\log_2(1/\delta)$, then conditioned on X_i, $1 \le i \le n$, $P_n\{Z_n \ge \xi_1 + \xi_2 \log_2(1/\delta)\} \le 3\delta$. Thus

$$E_n\left(\frac{Z_n - \xi_1}{\xi_2}\right)^+ = \int_0^\infty P_n\{\frac{Z_n - \xi_1}{\xi_2} \ge t\}dt \le 3\int_0^\infty 2^{-t}dt = \frac{3}{\ln 2}.$$

Let $\xi = 3\xi_2/\ln 2 + \xi_1$, then we have $E_n\left(\|f_n - \hat{Y}_{\hat{I}}\|^2/n + \lambda C_{\hat{I}}/n\right) \le \xi R_n^*$. Thus, we need only to prove (11) and (12) with $g_1(\delta) = g_2(\delta) = \overline{\lambda}\log_2(1/\delta)$.

The following two facts are useful.

Fact 1. If $Z \sim N(0,1)$, then $P(|z| \ge t) \le e^{-t^2/2}$.

Fact 2. If $Z_m \sim \chi_m^2$, then

$$P(Z_m - m \ge \kappa m) \le e^{-\frac{m}{2}(\kappa - \ln(1+\kappa))}, \quad \kappa > 0$$

$$P(Z_m - m \le -\kappa m) \le e^{-\frac{m}{2}(\ln(\frac{1}{1-\kappa}) + \kappa)}, \quad 0 < \kappa < 1.$$

These familiar large deviation bounds follow by evaluating the Cramer-Chernoff exponents for Normal and Chi-square distributions.

Notice $\mathrm{rem}_1(I) \sim N(0, \|f_n - M_I f_n\|^2)$, by Fact 1, $P\left(|\mathrm{rem}_1(I)|/\|f_n - M_I f_n\| \geq t_I\right) \leq e^{-t_I^2/2}$. Taking $t_I^2 = 2(\ln 2)(C_I - \log_2(\delta)$, we have

$$\begin{array}{ll} P\left(\sup_{I\in\Gamma} \frac{|\mathrm{rem}_1(I)|}{\|f_n - M_I f_n\| t_I} \geq 1\right) & \leq \sum_{I\in\Gamma} P\left(\frac{|\mathrm{rem}_1(I)|}{\|f_n - M_I f_n\| t_I} \geq 1\right) \\ & \leq \sum_{I\in\Gamma} 2^{-(C_I - \log_2(\delta))} \\ & \leq \delta. \end{array}$$

Thus, with probability no less than $1-\delta$, for all $I \in \Gamma$,

$$|\mathrm{rem}_1(I)| \leq \|f_n - M_I f_n\| \left(2(\ln 2)(C_I - \log_2(\delta))\right)^{1/2}.$$

For the other remainder term, because ${e_n}' M_I e_n \sim \chi^2_{r_I}$, by taking $\rho_{1,I}$ such that

$$\frac{r_I}{2}(\rho_{1,I} - \ln(\rho_{1,I} + 1)) = (\ln 2)(C_I - \log_2(\delta)), \tag{16}$$

by Fact 2,

$$P(\mathrm{rem}_2(I) \leq -\rho_{1,I} r_I \ \text{ for some } I \in \Gamma) \leq \sum_{I\in\Gamma} 2^{-(C_I - \log_2(\delta))} \leq \delta. \tag{17}$$

Take $\rho_{2,I}$ such that

$$\frac{r_I}{2}(\rho_{2,I} + \ln(\frac{1}{1-\rho_{2,I}})) = (\ln 2)(C_I - \log_2(\delta)), \tag{18}$$

then

$$P(\mathrm{rem}_2(I) \geq \rho_{2,I} r_I \ \text{ for some } I \in \Gamma) \leq \sum_{I\in\Gamma} 2^{-(C_I - \log_2(\delta))} \leq \delta.$$

If we can choose λ large enough (not depending on δ) so that

$$\|f_n - M_I f_n\| \sqrt{2(\ln 2)(C_I - \log_2 \delta)} \leq \tau_1(n R_n(I) + g_1(\delta)),$$

$\rho_{1,I} r_I \leq \tau_2(n R_n(I) + g_2(\delta))$, and $\rho_{2,I} r_I \leq \tau_2(n R_n(I) + g_2(\delta))$, then (11) and (12) are satisfied. Equivalently we need

$$\lambda + g_1(\delta)/C_I \geq (1/\tau_1)\left(\|f_n - M_I f_n\|/C_I^{1/2}\right)\left(2\ln 2 - 2(\ln 2)\log_2 \delta/C_I\right)^{1/2} - r_I/C_I - \|f_n - M_I f_n\|^2/C_I,$$

$$\lambda + g_2(\delta)/C_I \geq (\rho_{1,I}/\tau_2 - 1) r_I/C_I - \|f_n - M_I f_n\|^2/C_I,$$

$$\lambda + g_2(\delta)/C_I \geq (\rho_{2,I}/\tau_2 - 1) r_I/C_I - \|f_n - M_I f_n\|^2/C_I.$$

Let $s = \|f_n - M_I f_n\|^2 / (C_I - \log_2(\delta))$. Using the relationships in (16) and (18), it suffices to require that for all $s > 0$, $\rho_1 > 0$, and $0 < \rho_2 < 1$,

$$\lambda + g_1(\delta)/C_I \geq (1 - \log_2(\delta)/C_I) \left((2(\ln 2)s)^{1/2} / \tau_1 - s \right),$$

$$\lambda + g_2(\delta)/C_I \geq (1 - \log_2(\delta)/C_I) \left((\rho_1/\tau_2 - 1) 2(\ln 2) / (\rho_1 - \ln(\rho_1 + 1)) \right),$$

$$\lambda + g_2(\delta)/C_I \geq (1 - \log_2(\delta)/C_I) \left((\rho_2/\tau_2 - 1) 2(\ln 2) / (\rho_2 - \ln(1 - \rho_2)) \right).$$

The third requirement is automatically satisfied in the presence of the second one. Thus it suffices to require that

$$\lambda \geq h(\tau_1, \tau_2) = \max \left(\sup_{s \geq 0} \left((2(\ln 2)s)^{1/2} / \tau_1 - s \right), \sup_{\rho > 0} (\rho/\tau_2 - 1) 2(\ln 2) / (\rho - \ln(\rho + 1)) \right),$$

with the choice of $g_1(\delta) = g_2(\delta) = (-\log_2 \delta)\, h(\tau_1, \tau_2)$ (i.e., $\overline{\lambda} = h(\tau_1, \tau_2)$). It is easily seen that for any $\tau_1 > 0$, $\tau_2 > 0$, $h(\tau_1, \tau_2)$ is less than infinity. Let $\lambda_0 = \min_{0 < \tau_2 < 1} h((1 - \tau_2)/2, \tau_2)$. Then if $\lambda > \lambda_0$, by continuity, there exist τ_1 and τ_2 with $2\tau_1 + \tau_2 < 1$ such that the desired requirements are satisfied. With suitably chosen τ_2 at the value of 4.78, λ_0 can be shown to be smaller than 5.1.

Now suppose $nR_n^* \to \infty$ as $n \to \infty$. From (13), using $\|f_n - \hat{Y}_I\|^2 = nR_n(I) - \mathrm{rem}_2(I)$ and (12), we have that with exception probability no bigger than 3δ,

$$\frac{\|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}}}{nR_n^*} \leq \frac{(1 + \tau_2)(1 + 2\tau_1(1 + 2g_1(\delta)/nR_n^*) + \tau_2(1 + 2g_2(\delta)/nR_n^*)}{(1 - 2\tau_1 - \tau_2)} + \frac{\tau_2 g_2(\delta)}{nR_n^*}.$$

Because for any fixed δ, $g_1(\delta)$ and $g_2(\delta)$ are asymptotically negligible compared to nR_n^*, the second conclusion of Theorem 1 follows. This complete the proof of Theorem 1.

Remark: The essential ingredients in the above analysis are the exponential inequalities in Facts 1 and 2. They are used to bound the magnitude of the two remainder terms $\mathrm{rem}_1(I)$ and $\mathrm{rem}_2(I)$. Normality makes the quadratic remainder term $\mathrm{rem}_2(I)$ chi-square (centered) distributed and therefore easy to handle. In general, the distribution of this term depends on both the distribution of the errors and the design matrix of the model. Thus without normality, it seem unlikely that one can obtain such a general risk bound as in Theorem 1 without any conditions on the unknown regression function and the operating linear models.

Proof of Corollary 1: Let $\widetilde{I}_n^*$ be the model achieving $\mathcal{R}_n^*$. Observing that $\|f_n - \overline{f}_I\|^2 \leq \|f_n - f_{I,\theta^*}\|^2$, where θ^* is the minimizer of $E\left(f(\mathbf{X}) - f_{I,\theta}(\mathbf{X})\right)^2$ over $\theta \in R^{m_I}$ and $f_{I,\theta^*} = (f_I(X_1,\theta^*), ...f_I(X_n,\theta^*))^T$, we have

$$\begin{aligned}\|f_n - \overline{f}_{I_n^*}\|^2/n + r_{I_n^*}\sigma^2/n + \lambda\sigma^2 C_{I_n^*}/n &\leq \inf_{I\in\Gamma}\{\|f_n - f_{I,\theta^*}\|^2/n + r_I\sigma^2/n + \lambda\sigma^2 C_I/n\} \\ &\leq \|f_n - f_{\widetilde{I}_n^*,\theta^*}\|^2/n + r_{\widetilde{I}_n^*}\sigma^2/n + \lambda\sigma^2 C_{\widetilde{I}_n^*}/n \\ &\leq \|f_n - f_{\widetilde{I}_n^*,\theta^*}\|^2/n + m_{\widetilde{I}_n^*}\sigma^2/n + \lambda\sigma^2 C_{\widetilde{I}_n^*}/n.\end{aligned}$$

Together with (4), by taking expectation further with respect to $\{X_i\}_{i=1}^n$, and using that $E\left(\|f_n - f_{\widetilde{I}_n^*,\theta^*}\|^2/n\right) = E\left(f(\mathbf{X}) - f_{\widetilde{I}_n^*,\theta^*}(\mathbf{X})\right)^2$, we conclude that the expected value of the sum of ASE and the model complexity (over n) of the selected model is bounded above by a multiple of $\mathcal{R}_n^*$. This completes the proof of Corollary 1.

Proof of Theorem 2. From (13) with $g_1(\delta) = g_2(\delta) = \overline{\lambda}\log_2(1/\delta)$, for any fixed $0 < \delta < 1$, there exists τ_1^0 and τ_2^0 and $A = A_\delta = \left(1 + (2\tau_1^0 + \tau_2^0)(1 - 2\overline{\lambda}\log_2(\delta)/nR_n^*)\right)/(1 - 2\tau_1^0 - \tau_2^0)$ such that $P(\hat{I} \in K_{n,A}) \geq 1 - 3\delta$. If we can show that for any τ_1, τ_2 with $2\tau_1 + \tau_2 < 1$, when n is large enough,

$$P\left(|\mathrm{rem}_1(I)| \geq \tau_1 nR_n(I) \ \text{ for some } I \in K_{n,A}\right) \leq \delta \tag{19}$$

$$P\left(|\mathrm{rem}_2(I)| \geq \tau_2 nR_n(I) \ \text{ for some } \ I \in K_{n,A}\right) \leq 2\delta, \tag{20}$$

then similarly to the analysis in the proof of Theorem 1 (see (15)), but considering only models in $K_{n,A}$ instead of Γ, when $\hat{I} \in K_{n,A}$ (which implies that $\hat{I}$ minimizes ABC over $K_{n,A}$), one can show that with probability no less than $1 - (3\delta + 3\delta)$,

$$\frac{R_n(\hat{I})(1-\tau_2)}{R_n^*} \leq \frac{\|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}}}{nR_n^*} \leq \frac{(1+\tau_2)(1+2\tau_1+\tau_2)}{1-2\tau_1-\tau_2}.$$

Note the r.h.s of the above inequality goes to 1 when $\tau_1 \to 0$, $\tau_2 \to 0$ and that $R_n(\hat{I})/R_n^* \geq 1$. As a result, $\left(\|f_n - \hat{Y}_{\hat{I}}\|^2 + \lambda C_{\hat{I}}\right)/(nR_n^*) \longrightarrow 1$ in probability. Using $\sup_{I\in K_{n,A}} C_I/r_I \to 0$ together with (20), it is seen that for $I \in K_{n,A}$, λC_I is negligible in probability compared to $\|f_n - \hat{Y}_I\|^2 = \|f_n - M_I f_n\|^2 + r_I - \mathrm{rem}_2(I)$. Also λC_{I_n} is negligible compared to $\|f_n - M_{I_n} f_n\|^2 + r_{I_n}$. As a consequence, we have both $\|f_n - \hat{Y}_{\hat{I}}\|^2/nR_n^* \to 1$ and $\|f_n - \hat{Y}_{\hat{I}}\|^2/\left(\|f_n - M_{I_n} f_n\|^2 + r_{I_n}\right) \to 1$ in probability. Now let's prove (19) and (20). In the proof of Theorem 1, we have shown that with probability no less than $1 - \delta$, for all $I \in \Gamma$,

$$|\mathrm{rem}_1(I)| \leq \|f_n - M_I f_n\|\left(2(\ln 2)(C_I - \log_2(\delta))\right)^{1/2}. \tag{21}$$

Because $\|f_n - M_I f_n\|^2 + r_I + \lambda C_I \geq 2\|f_n - M_I f_n\|\sqrt{r_I}$ and $\sup_{I \in K_{n,A}} C_I/r_I \to 0$, we have that for any fixed $0 < \delta < 1$, $|\mathrm{rem}_1(I)| \leq \tau_1 R_n(I)$ holds with probability no less than $1-\delta$ for all $I \in K_{n,A}$ and any $\tau_1 > 0$ when n is large enough. For the other remainder term, from the proof of Theorem 1, we know that with probability no less than $1-\delta$, $\mathrm{rem}_2(I) \leq \rho_{2,I} r_I$ for all $I \in \Gamma$, where $\rho_{2,I}$ satisfies (18). So with probability no less than $1-\delta$,

$$\mathrm{rem}_2(I) \leq 2(\ln 2)(C_I - \log_2(\delta)) \text{ for all } I \in \Gamma. \tag{22}$$

Similarly, with probability no less than $1-\delta$, $\mathrm{rem}_2(I) \geq -\rho_{1,I} r_I$ for all $I \in \Gamma$, where $\rho_{1,I}$ satisfies (16). From $\sup_{I \in K_{n,A}} C_I/r_I \to 0$, we have $\sup_{I \in K_{n,A}} \rho_{1,I} \to 0$. Therefore for n large enough, $\rho_{1,I} - \ln(1+\rho_{1,I}) \geq (\rho_{1,I})^2/4$ for all $I \in K_{n,A}$. Then $\mathrm{rem}_2(I) \geq -\left(8(\ln 2)(C_I - \log_2(\delta)) r_I\right)^{1/2}$ holds with probability no less than $1-\delta$ for all $I \in K_{n,A}$. For any fixed $0 < \delta < 1$, again from the assumption $\sup_{I \in K_{n,A}} C_I/r_I \to 0$, it follows that $\sup_{I \in K_{n,A}} (C_I - \log_2(\delta)) / (nR_n(I)) \to 0$ and $\sup_{I \in K_{n,A}} \left(8(C_I - \log_2(\delta)) r_I\right)^{1/2} / (nR_n(I)) \to 0$. So with probability no less than $1-2\delta$, $|\mathrm{rem}_2(I)| \leq \tau_2 n R_n(I)$ for all $I \in K_{n,A}$ and for every τ_2 when n is large enough.

For the proof of the second result, we need to treat the two remainder terms more carefully. Let

$$\tau_1(\delta) = \sup_{I \in \Gamma_n} \frac{\|f_n - M_I f_n\| \left(2(\ln 2)(C_I - \log_2(\delta))\right)^{1/2}}{\|f_n - M_I f_n\|^2 + r_I + \lambda C_I + \log^2(\delta)}.$$

Then using $a + b \geq 2\sqrt{ab}$ for $a, b > 0$, we have

$$\tau_1(\delta) \leq \sup_{I \in \Gamma_n} \left(\frac{(\ln 2)(C_I - \log_2(\delta))}{2\left(r_I + \lambda C_I + \log_2^2 \delta\right)} \right)^{1/2} \leq \left(\frac{\ln 2}{2}\right)^{1/2} \left(\sup_{I \in \Gamma_n} \left(\frac{C_I}{r_I} + \frac{1}{2\sqrt{r_I}} \right) \right)^{1/2}.$$

Let $\tau_{1,0}$ denote the right hand side of the second inequality above. Then together with (21), we have that with probability no less than $1-\delta$, for all $I \in \Gamma_n$, $|\mathrm{rem}_1(I)| \leq \tau_{1,0}(nR_n(I) + \log_2^2 \delta)$. From (22), with probability no less than $1-\delta$, $\mathrm{rem}_2(I) \leq 2(\ln 2)(C_I - \log_2(\delta))$ for all $I \in \Gamma_n$. From (17), with probability no less than $1-\delta$, $\mathrm{rem}_2(I) \geq -\rho_{1,I} r_I$ for all $I \in \Gamma_n$, where $\frac{r_I}{2}(\rho_{1,I} - \ln(1+\rho_{1,I})) = (\ln 2)(C_I - \log_2(\delta))$. For $\rho_{1,I} \leq 1$, $\rho_{1,I} - \ln(1+\rho_{1,I}) \geq \rho_{1,I}^2/4$. Then $\mathrm{rem}_2(I) \geq -\, r_I^{1/2} \left(r_I \rho_{1,I}^2\right)^{1/2} \geq -\left(8(\ln 2)(C_I - \log_2(\delta)) r_I\right)^{1/2}$ holds with probability no less than $1-\delta$. For $\rho_{1,I} \geq 1$, because $\rho_{1,I} - \ln(1+\rho_{1,I}) \geq \rho_{1,I}/2$, we have that $r_I \rho_{1,I}/2 \leq 2\ln 2(C_I - \log_2 \delta)$. So $\mathrm{rem}_2(I) \geq -4(\ln 2)(C_I - \log_2 \delta)$ with probability no less than $1-\delta$.

Similarly to bounding $\tau_1(\delta)$, we have

$$\begin{aligned}
&\sup_{I\in\Gamma_n} \frac{\max\left((8\ln 2(C_I-\log_2(\delta))r_I)^{1/2}, 4\ln 2(C_I-\log_2\delta)\right)}{nR_n(I)+\log_2^2\delta} \\
\leq\ & \max\left(\left(8\ln 2\sup_{I\in\Gamma_n}\left(C_I/r_I+1/\left(2r_I^{1/2}\right)\right)\right)^{1/2}, 4\ln 2\sup_{I\in\Gamma_n}\left(C_I/r_I+1/\left(2r_I^{1/2}\right)\right)\right).
\end{aligned}$$

Let $\tau_{2,0}$ denote the right hand side of the above inequality. It follows that with probability no less than $1-2\delta$, $|\mathrm{rem}_2(I)| \leq \tau_{2,0}(nR_n(I)+\log_2^2\delta)$ holds for all $I\in\Gamma_n$. All together and from (13), we have that with probability no less than $1-6\delta$,

$$W_n = \frac{R_n(\hat{I})}{R_n^*} \leq \frac{1+(2\tau_{1,0}+\tau_{2,0})(1+2\log_2^2\delta/\underline{r}_n)}{1-2\tau_{1,0}-\tau_{2,0}},$$

where $\underline{r}_n = \min_{I\in\Gamma_n} r_I$. Let

$$\widetilde{W} = \left(W_n - \frac{1+2\tau_{1,0}+\tau_{2,0}}{1-2\tau_{1,0}-\tau_{2,0}}\right) / \frac{2\left(2\tau_{1,0}+\tau_{2,0}\right)}{\underline{r}_n\left(1-2\tau_{1,0}-\tau_{2,0}\right)}.$$

Then $P_n\{\widetilde{W}\geq\log_2^2\delta\}\leq 6\delta$ for $0<\delta<1$. It follows that $E_n\tilde{W}^+ = \int_0^\infty P_n\{\ \tilde{W}\geq t\}dt \leq 6\int_0^\infty 2^{-\sqrt{t}}dt = 12/\left(\ln 2\right)^2$. Thus

$$E_n\left(W_n\right) \leq \frac{1+2\tau_{1,0}+\tau_{2,0}}{1-2\tau_{1,0}-\tau_{2,0}} + \frac{24\left(2\tau_{1,0}+\tau_{2,0}\right)}{\left(\ln 2\right)^2\underline{r}_n\left(1-2\tau_{1,0}-\tau_{2,0}\right)}.$$

Because $\sup_{I\in\Gamma_n} C_I/r_I \to 0$, hence $\tau_{1,0}$ and $\tau_{2,0}$ tend to 0. Then the r.h.s. of the above inequality tends to 1 as $n\to\infty$. Together with $W_n\geq 1$, we conclude that $\lim_{n\to\infty} E_n\left(W_n\right)=1$. Because C_I is uniformly negligible compared to r_I, we have

$$\frac{\inf_{I\in\Gamma}\left(\|f_n-\overline{f}_I\|^2/n + r_I\sigma^2/n\right)}{\inf_{I\in\Gamma}\left(\|f_n-\overline{f}_I\|^2/n + r_I\sigma^2/n + \lambda\sigma_0^2C_I/n\right)} \to 1.$$

Then (8) follows. This completes the proof of Theorem 2.

Proof of Theorem 3: Expanding squares as in the proof of Theorem 1, we have

$$\begin{aligned}
&\|Y_n-\hat{Y}_I\|^2 + 2r_I\hat{\sigma}^2 + \lambda C_I\hat{\sigma}^2 \\
=\ & e_n'e_n + \left(\|f_n-M_If_n\|^2 + r_I\sigma^2 + \lambda\sigma^2C_I\right) + 2e_n'(f_n-M_If_n) + (r_I\sigma^2 - e_n'M_Ie_n) + \\
& (2r_I+\lambda C_I)\left(\hat{\sigma}^2-\sigma^2\right).
\end{aligned}$$

From the proof of Theorem 1, for $\lambda\geq 5.1$, there exist τ_1 and τ_2 with $2\tau_1+\tau_2\leq 1-\gamma$ for some small $\gamma>0$, such that for any $0<\delta<1$, there exist $g_1(\delta)$ and $g_2(\delta)$ such that (11) and (12)

are satisfied with probability no less than $1-3\delta$. Let $\beta_n = P\left(|\hat{\sigma}^2 - \sigma^2| > \gamma\sigma^2/2\right)$, then because $\hat{\sigma}^2$ is consistent, $\beta_n \to 0$. Let $\mathrm{rem}_3(I) = (2r_I + \lambda C_I)\left(\hat{\sigma}^2 - \sigma^2\right)$. From above, with probability no less than $1-3\delta-\beta_n$, the three remainder terms are well controlled as in (11), (12), and $|\mathrm{rem}_3(I)| \le (2r_I+\lambda C_I)\gamma\sigma^2/2 \le (\gamma/2)nR_n(I)$. Proceed as in the proof of Theorem 1, we obtain an inequality similar to (15) with a different upper bound in terms of δ. The proof of the second assertion in Theorem 3 can be handled similarly using the argument in the proof of Theorem 2. This completes the proof of Theorem 3.

Proof of Theorem 4: We prove the slightly stronger conclusion that with probability tending to 1, $ASE(\hat{I}) \le B_1 \inf_{I\in\Gamma}\left(\left\|f_n - \overline{f}_I\right\|^2 + r_I\sigma^2 + \lambda\sigma_0^2 C_I\right)$. Let $A_I = I_{m_I\times m_I} - M_I$, where $I_{m_I\times m_I}$ is the $m_I \times m_I$ identity matrix. By removing a common term $(e_n^{'}e_n)$ for all models, the criterion is theoretically equivalent to

$$\begin{aligned}
\mathrm{crit}(I) &= \|A_I f_n\|^2 - r_I\sigma^2 + (r_I\sigma^2 - e_n^{'}M_I e_n) + 2e_n^{'}A_I f_n \\
&\quad + \frac{2r_I}{n-r_I}\left(\|Y_n - \hat{Y}_I\|^2 + \lambda\sigma_0^2 C_I\right) + \lambda\sigma_0^2 C_I \\
&= \|A_I f_n\|^2 + r_I\left(\frac{2}{n-r_I}\left(\|Y_n - \hat{Y}_I\|^2 + \lambda\sigma_0^2 C_I\right) - \sigma^2\right) \\
&\quad + \lambda\sigma_0^2 C_I + 2\mathrm{rem}_1(I) + \mathrm{rem}_2(I),
\end{aligned}$$

where $\mathrm{rem}_1(I)$ and $\mathrm{rem}_2(I)$ are defined in the proof of Theorem 1. Note also that

$$\|Y_n - \hat{Y}_I\|^2 + \lambda\sigma_0^2 C_I = \|A_I f_n\|^2 + (n-r_I)\sigma^2 + \left(e_n^{'}A_I e_n - (n-r_I)\sigma^2\right) + 2e_n^{'}A_I f_n + \lambda\sigma_0^2 C_I.$$

Let $T(I) = \|A_I f_n\|^2 + (n-r_I)\sigma^2 + \lambda\sigma_0^2 C_I$. Redefine $R_n(I) = \|A_I f_n\|^2 + r_I\sigma^2 + \lambda\sigma_0^2 C_I$ (note that $\|A_I f_n\|^2 = \left\|f_n - \overline{f}_I\right\|^2$). Similarly to the proof of Theorem 1, one can show that if $\lambda > h(\tau_1,\tau_2)$, then there exist two constants τ_1 and τ_2 with $2\tau_1+\tau_2<1$ such that for any $\delta>0$, with probability no less than $1-5\delta$, $|\mathrm{rem}_1(I)| \le \tau_1\left(nR_n(I) + g_1(\delta)\right)$, $|\mathrm{rem}_2(I)| \le \tau_2\left(nR_n(I)+g_2(\delta)\right)$, and $|e_n^{'}A_I e_n - (n-r_I)\sigma^2| \le \tau_2\left(T(I)+g_2(\delta)\right)$, where $g_1(\delta) = g_2(\delta) = \overline{\lambda}\log_2(1/\delta)$. Under the condition that for each $I\in\Gamma$, $\|f_n - \overline{f}_I\|^2 \to \infty$, for any $\epsilon>0$, when n is large enough, $nR_n(I) + g_1(\delta) \le (1+\epsilon)nR_n(I)$, $nR_n(I)+g_2(\delta) \le (1+\epsilon)nR_n(I)$, and $(T(I)+g_2(\delta)) \le (1+\epsilon)T(I)$. Then with probability no less than $1-5\delta$, we have

$$\mathrm{crit}(I) \ge \|A_I f_n\|^2 + r_I\left(\frac{2\left(1-(1+\epsilon)(2\tau_1+\tau_2)\right)T(I)}{n-r_I} - \sigma^2\right)$$

$$
\begin{aligned}
& -(1+\epsilon)(2\tau_1+\tau_2)R_n(I)+\lambda\sigma_0^2 C_I \\
\geq\ & \|A_I f_n\|^2 + r_I\left(1-(1+\epsilon)(4\tau_1+2\tau_2)\right)\sigma^2 - (1+\epsilon)(2\tau_1+\tau_2)R_n(I)+\lambda\sigma_0^2 C_I \\
\geq\ & \left(1-(1+\epsilon)(6\tau_1+3\tau_2)\right)R_n(I)
\end{aligned}
$$

For the above inequalities to be useful, we need $6\tau_1+3\tau_2<1$. Let I_n be the model minimizing $R_n(I)$ among the candidate models. From above, with exception probability less than 5δ, for I_n,

$$
\begin{aligned}
\mathrm{crit}(I_n) \ \leq\ & \|A_{I_n} f\|^2 + r_I\left(\frac{2(1+(1+\epsilon)(2\tau_1+\tau_2))T(I_n)}{n-r_{I_n}}-\sigma^2\right) \\
& +(1+\epsilon)(2\tau_1+\tau_2)R_n(I_n)+\lambda\sigma_0^2 C_I.
\end{aligned}
$$

Under the assumption that $R_n^*(f;\Gamma)\to 0$, and since $R_n(I_n)\leq(\sigma_0^2/\sigma^2)\,R_n^*(f;\Gamma)$, we have that $r_{I_n}/n\to 0$ and $\left(\|A_{I_n}f_n\|^2+\lambda\sigma_0^2 C_{I_n}\right)/(n-r_{I_n})\to 0$. So when the sample size is large enough, $T(I_n)/(n-r_I)\to 1$ and $\mathrm{crit}(I_n)\leq(1+\epsilon)\left(1+(1+\epsilon)(6\tau_1+3\tau_2)\right)R_n(I_n)$. Thus for any $\delta>0$, when the sample size is large enough, we have that with probability no less than $1-5\delta$,

$$
\begin{aligned}
R_n(\hat{I}) \ &\leq\ \mathrm{crit}(\hat{I})/\left(1-(1+\epsilon)(6\tau_1+3\tau_2)\right) \\
&\leq\ \mathrm{crit}(I_n)/\left(1-(1+\epsilon)(6\tau_1+3\tau_2)\right) \\
&\leq\ (1+\epsilon)\left(1+(1+\epsilon)(6\tau_1+3\tau_2)\right)R_n(I_n)/\left(1-(1+\epsilon)(6\tau_1+3\tau_2)\right).
\end{aligned}
$$

That is

$$
R_n(\hat{I})/R_n(I_n)\leq(1+\epsilon)\left(1+(1+\epsilon)(6\tau_1+3\tau_2)\right)/\left(1-(1+\epsilon)(6\tau_1+3\tau_2)\right)
$$

with exception probability less than 5δ. Proceed as in the proof of Theorem 1, we know with probability tending to 1, $\left(\|f_n-\hat{Y}_{\hat{I}}\|^2+\lambda C_{\hat{I}}\right)/R_n(I_n)\leq B_1$, where the constant B_1 depends on ϵ, τ_1 and τ_2, provided $(1+\epsilon)(6\tau_1+3\tau_2)<1$. Minimizing $h(\tau_1,\tau_2)$ over τ_1 and τ_2 in the region $6\tau_1+3\tau_2<1$, the minimum value is less than 40. Thus the asymptotic results hold when $\lambda\geq 40$. This completes the proof of Theorem 4.

Proof of Theorem 5: From Corollary 1, we only need to examine the index of resolvability $\mathcal{R}_n^*$ for the enlarged Sobolev classes $S_r(\alpha;C)$. To that end, the main task is to upper bound the approximation error for these classes by the tensor-product splines. For $g\in W_2^{\alpha,r}(C)$, from Schumaker (1981, Theorem 12.8 and Equation 13.69) as used in Stone (1994), 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 $\int_0^1 (g(\mathbf{z}_r) - g_{I,\theta^*}(\mathbf{z}_r))^2 \, d\mathbf{z}_r$ is upper bounded by $Am^{-2\alpha}$, where the constant A 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 $d!/(r!(d-r)!)Am^{-2\alpha}$. The model dimension m_I is of order m^r. Note that for given r and $\mathbf{q}_r^*$, the model complexity $\log_2 d + \sum_{j=1}^r \log^* q_j + r \log^* m$ is asymptotically negligible compared to m_I. From all above,

$$\begin{aligned}
\sup_{f \in S_r(\alpha;C)} \mathcal{R}_n^*(f;\Gamma) &\leq \inf_m \left(\sup_{f \in S_r(\alpha;C)} \int \left(f(\mathbf{x}) - f_{(r,\mathbf{q}_r^*,\mathbf{m}_r^*),\theta^*}(\mathbf{x}) \right)^2 d\mathbf{x} + \frac{m_{(r,\mathbf{q}_r^*,\mathbf{m}_r^*)}\sigma^2}{n} + \frac{\lambda\sigma^2 C_{(r,\mathbf{q}_r^*,\mathbf{m}_r^*)}}{n} \right) \\
&= O\left(\inf_m \left(m^{-2\alpha} + m^r/n \right) \right) \\
&= O\left(n^{-2\alpha/(2\alpha+r)} \right),
\end{aligned}$$

where for the last step, m is taken of order $n^{1/(2\alpha+r)}$. This completes the proof of Theorem 5.

Proof of Theorem 6: By Corollary 1, we only need to show $\sup_{f \in \mathcal{F}(\Upsilon,\Phi)} \mathcal{R}_n^*(f;\Gamma_1) = O\left(m_n(\Upsilon)/n\right)$ for each choice of Υ satisfying (10), where Γ_1 is the list of complete models. For $f \in \mathcal{F}(\Upsilon,\Phi)$, by definition, we have $\inf_{\theta \in R^m} E\left(f(\mathbf{X}) - f_{m,\theta}(\mathbf{X})\right)^2 \leq \gamma_m^2$, together with $C_m/m \to 0$, we know $\sup_{f \in \mathcal{F}(\Upsilon,\Phi)} \mathcal{R}_n^*(f;\Gamma_1) \leq A \inf_m \left(\gamma_m^2 + m\sigma^2/n\right) = O\left(m_n(\Upsilon)/n\right)$, where A is a constant not depending on m. The conclusion follows.

Proof of Theorem 7: As mentioned in Section 3, for N_k of order k^τ $(\tau > 1)$, C_I is of order $k \log k$. Let Γ_2 denote the list of the subset models. Then as in the proof of Theorem 6, $\sup_{f \in \mathcal{S}(\Upsilon,\Phi,\mathcal{N})} \mathcal{R}_n^*(f;\Gamma_2) \leq A' \inf_k \left(\gamma_k^2 + k\sigma^2/n + k \log k/n\right) = O\left(m_n(\Upsilon) \log n/n\right)$, where for the last step, we take k of order $m_n(\Upsilon)$. The conclusion then follows from Corollary 1.

References

[1] Akaike, H. (1970). Statistical prediction identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

[2] Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory* 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest.

[3] Barron, A.R. (1985). Logistically smooth density estimation. Ph.D. dissertation, Department of Electrical Engineering, Stanford University.

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

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

[6] Barron, A.R. and Sheu, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19**, 1347-1369.

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

[8] Barron, A.R., Birgé, L. and Massart, P. (1997). Risk bounds for model selection via penalization. To appear in *Probability Theory and Related Fields.*

[9] Birgé, L. and Massart, P. (1996). From model selection to adaptive estimation. *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam* 55-87, (D. Pollard, E. Togersen and G. Yang, eds.), Springer, New York.

[10] Brown, L.D. and Low, M.G. (1995). A constrained risk inequality with applications to nonparametric functional estimation. Technical report, Department of Statistics, University of Pennsylvania.

[11] Burman, P. (1990). Estimation of generalized additive models. *J. Multi. Anal.* **32**, pp. 230-255.

[12] Chen, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19**, 1855-1868.

[13] Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*, Wiley, New York.

[14] Donoho, D. and Johnstone, I. (1994). Minimax Estimation via Wavelet Shrinkage. Technical Report 402, Department of Statistics, Stanford University.

[15] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508-539.

[16] Efroimovich, S.Yu. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory probab. Appl.* **30**, 557-568.

[17] Efroimovich, S.Yu. and Pinsker, M.S. (1984). A self-educating nonparametric filtration algorithm. *Automation and Remote Control* **45**, 58-65.

[18] Elias, P. (1975). Universal codeword sets and representation of integers. *IEEE Trans. Information Theory* **21**, 194-203.

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

[20] Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465-1481.

[21] Lepskii, O.V. (1991). Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory probab. Appl.* **36**, 682-697.

[22] Li, K.C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 958-975.

[23] Lorentz, G.G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soci.* **72**, 903-937.

[24] Mallows, C.L. (1973). Some comments on C_p. *Technometrics***15**, 661-675.

[25] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag New York.

[26] Polyak, B.T. and Tsybakov, A.B. (1990). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Probab. Application* **35**, 293-306.

[27] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416-431.

[28] Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **30**, 629-636.

[29] Rissanen, J. (1987). Stochastic complexity. *J. R. Statist. Soc. B* **49**, 223-239.

[30] Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math.* **35**, 415-423.

[31] Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595-620.

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

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

[34] Wallace, C.S. and Boulton, D.M. (1987). Estimation and inference by compact coding. *J. R. Statist. Soc. B* **49**, 240-265.

[35] Yang, Y. (1997). On adaptive function estimation. Submitted to *Ann. Statist.*.

[36] Yang, Y. and Barron, A.R. (1997). Information-theoretic determination of minimax rates of convergence. Submitted to *Ann. Statist.*.

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

[38] Yu, B. and Speed, T. (1992). Data compression and histogram. *Probability Theory and Related Fields* **92**, 195-229.