1998

Combining Different Procedures for Adaptive Regression

Yuhong Yang
Iowa State University

Combining Different Procedures for Adaptive Regression

Abstract
Given any countable collection of regression procedures (e.g., kernel, spline, wavelet, local polynomial, neural nets, etc.), we show that a single adaptive procedure can be constructed to share their advantages to a great extent in terms of global squared L_2 risk. The combined procedure basically pays a price only of order 1/n for adaptation over the collection. An interesting consequence is that for a countable collection of classes of regression functions (possibly of completely different characteristics), a minimax-rate adaptive estimator can be constructed such that it automatically converges at the right rate for each of the classes being considered. A demonstration is given for high-dimensional regression, for which case, to overcome the well-known curse of dimensionality in accuracy, it is advantageous to seek different ways of characterizing a high-dimensional function (e.g., using neural nets or additive modelings) to reduce the influence of input dimension in the traditional theory of approximation (e.g., in terms of series expansion). However, in general, it is difficult to assess which characterization works well for the unknown regression function. Thus adaptation over different modelings is desired. For example, we show by combining various regression procedures that a single estimator can be constructed to be minimax-rate adaptive over Besov classes of unknown smoothness and interaction order, to converge at rate $o(n^{-1/2})$ when the regression function has a neural net representation, and at the same time to be consistent over all bounded regression functions

Keywords
adaptive estimation, combined procedures, minimax rate, nonparametric regression

Disciplines
Statistics and Probability

Comments
This preprint was published as Yuhang Yang, "Combining Different Procedures for Adaptive Regression", *Journal of Multivariate Analysis* (2000): 135-161, doi: 10.1006/jmva.1999.1884.

Combining Different Procedures for Adaptive Regression

Yuhong Yang
Department of Statistics
Iowa State University
yyang@iastate.edu

Abstract

Given any countable collection of regression procedures (e.g., kernel, spline, wavelet, local polynomial, neural nets, etc), we show that a single adaptive procedure can be constructed to share the advantages of them to a great extent in terms of global squared L_2 risk. The combined procedure basically pays a price only of order $1/n$ for adaptation over the collection. An interesting consequence is that for a countable collection of classes of regression functions (possibly of completely different characteristics), a minimax-rate adaptive estimator can be constructed such that it automatically converges at the right rate for each of the classes being considered.

A demonstration is given for high dimensional regression, for which case, to overcome the well-known curse of dimensionality in accuracy, it is advantageous to seek different ways of characterizing a high-dimensional function (e.g., using neural nets or additive modelings) to reduce the influence of input dimension in the traditional theory of approximation (e.g., in terms of series expansion). However, in general, it is difficult to assess which characterization works well for the unknown regression function. Thus adaptation over different modelings is desired. For example, we show by combining various regression procedures, a single estimator can be constructed to be minimax-rate adaptive over Besov classes of unknown smoothness and interaction order, to converge at rate $o(n^{-1/2})$ when the regression function has a neural net representation, and at the same time to be consistent over all bounded regression functions.

Keywords: Adaptive estimation, combined procedures, minimax rate, nonparametric regression.
AMS 1991 subject classifications. Primary 62G07; secondary 62B10, 62C20, 94A29.

1. Introduction. A lot of procedures have been proposed and commonly used for estimating a regression function. Parametric approaches include linear and nonlinear regressions assuming the true regression function is contained in (at least) one of a few finite-dimensional models being considered. Nonparametric procedures include familiar kernel regression (see, e.g., Wand and Jones (1995)), smoothing splines (e.g., Wahba (1990)), local polynomial regression (e.g., Fan and Gijbels (1996))), wavelet estimation (e.g., Donoho et al (1995)) and other methods using a list of approximating models (here the models are rather for operating, i.e., the true regression function may not be in any of them). For high-dimensional function estimation, as it is well-known, one often faces the problem of curse of dimensionality in accuracy. To overcome the curse, various parsimonious models such as projection pursuit (Friedman and Stuetzle (1981)), CART (Breiman et al (1984)), neural networks (e.g., Barron and Barron (1988)), additive models (Buja, Hastie and Tibshirani (1989)), MARS (Friedman (1991)), slicing regression (e.g., Duan and Li (1991)), tensor-product polynomial splines (e.g., Stone et al (1997)) have been proposed. These methods have been demonstrated or proved to work well for functions of different characteristics.

With a lot of regression procedures available (and more to come), from a practical point of view, it is often hard to choose a good one in terms of accuracy due to the difficulty in assessing which scenario describes the current data most appropriately. While nonparametric approaches are more flexible and less restrictive, if one could correctly identify a reasonably simple parametric model, a much more accurate estimator could be obtained. Within a collection of nonparametric approaches, it is also difficult to judge, for instance, if the true regression function is better described by a neural network model or by an additive spline model. Based on these considerations, one wishes to have a single estimation procedure that shares the advantages of candidate procedures automatically.

In this paper, we give a positive result in that direction from one perspective. We show that given any countable collection of estimation procedures for regression, a single procedure can be constructed to behave as well as (or nearly as well as) any procedure in the list in terms of a statistical risk (rate). By mixing a list of procedures, the advantages of them in terms of the risk are combined.

Results on combining an unrestrictive list of statistical estimation procedures are initially given in Yang (1996) for density estimation developed based on earlier work of Barron and his coauthors (e.g., Barron (1987), Clark and Barron (1990), and Barron and Cover (1991)). The results are further developed in Yang (1997) for both density estimation and regression. Later a similar result on density estimation is obtained independently by Catoni (1997). The present paper is developed from the regres-

sion part of Yang (1997) (the part on density estimation will be published in a separate paper (Yang (1999b)).

The paper is organized as follows. In Section 2, some setups are given. A general adaptive scheme and the corresponding adaptation risk bound are presented in Section 3. Section 4 concerns minimax-rate adaptation with respect to a collection of classes of regression functions. A demonstration of the main results is given in Section 5. Section 6 deals with adaptation when variance estimators are available. A discussion follows in Section 7. The proofs of the results are in Section 8.

2. Some setups. We consider the regression model

$$Y_i = u(X_i) + \varepsilon_i,\ i = 1, ...n,$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d. observations from the distribution of (X, Y). The explanatory variable X has an unknown distribution P_X. Given $X_i = x$, the error ε_i is assumed to be normally distributed with unknown mean $u(x)$ and unknown variance $\sigma^2(x)$. The goal is to estimate the regression function u based on $Z^n = (X_i, Y_i)_{i=1}^n$.

Let $\| u - v \| = \left(\int |u(x) - v(x)|^2 dP_X\right)^{1/2}$ be the L_2 distance weighted by the distribution of X. We consider the loss $\| u - \hat{u} \|^2$ as a measure of performance in this work. Another two quantities involved in our analysis are Kullback-Leibler (K-L) divergence and square Hellinger distance, defined as $D(f \parallel g) = \int f \log(f/g)\, d\nu$ and $d_H^2(f, g) = \int \left(\sqrt{f} - \sqrt{g}\right)^2 d\nu$ respectively between two densities f and g with respect to a measure ν.

In this paper, a regression estimation procedure (or strategy), say, δ, refers to a sequence of estimators $\hat{u}_{\delta,1}(x; Z_1)$, ..., $\hat{u}_{\delta,n-1}(x; Z^{n-1})$, ... of the unknown regression function based on observation(s) Z^1, ..., Z^{n-1}, ... respectively. The risk of a procedure δ at sample size n is denoted $R((u, \sigma); n; \delta)$, i.e., $R((u, \sigma); n; \delta) = E \| u - \hat{u}_{\delta,n} \|^2$ with the expectation taken under the regression function u and the variance function σ^2.

Throughout the paper, the regression function u is assumed to be bounded between $-A$ and A with A known.

The joint density of (X, Y) (with respect to the product measure of P_X and the Lebesgue measure) under u and σ^2 is denoted $p_{u,\sigma}(x, y)$, i.e.,

$$p_{u,\sigma}(x, y) = \frac{1}{\sqrt{2\pi}\sigma(x)} e^{\frac{-(y-u(x))^2}{2\sigma^2(x)}}.$$

The symbol "$\asymp$" is used to mean asymptotically of the same order, i.e., $a_n \asymp b_n$ if a_n/b_n is asymptotically upper and lower bounded away from zero.

3. General adaptation risk bounds. Let $\Delta = \{\delta_j, j \geq 1\}$ be a collection of regression estimation procedures with δ_j producing an estimator $\hat{u}_{j,i}$ based on Z^i. The index set $\{j \geq 1\}$ is allowed to degenerate to a finite set. Here no special requirement will be put on the procedures and they could be proposed for different purposes and/or under different assumptions on the regression function (e.g., smoothness, monotonicity, additivity, etc.) resulting in possibly completely different estimators for different j. For instance, procedure δ_1 may be an automated kernel method and procedure δ_2 may be a wavelet method, and δ_3 may be a simple linear regression method and so on. Or the procedures could be of the same type but with different choices of hyper-parameters. For instance, procedure δ_4 may be a method using quadratic splines while procedure δ_5 may be one using cubic splines. Some of the procedures (as δ_1 above) may well be adaptive already in certain scopes.

Now with the chosen countable collection of regression procedures, we ask the question: Can we obtain a single estimation procedure that is adaptive with respect to these procedures in the sense that it works as well as or nearly as well as any procedure in the collection no matter what the true regression function is?

3.1. An adaptation recipe. The following is a recipe to get an adaptive procedure by mixing appropriately the proposed ones in Δ. Unless stated otherwise, we assume σ is upper and lower bounded by known constants $\overline{\sigma} > 1$ and $\underline{\sigma} = \overline{\sigma}^{-1}$. Since σ may not be known, we consider a list of variance functions $\Xi = \{\sigma_k^2(x) : k \geq 1\}$ bounded accordingly, hoping that one of them is suitably close to the true one. Adaptation schemes utilizing estimators of σ^2 will be given in Section 6. Let $\underline{\pi} = \{\pi_j, j \geq 1\}$ and $\underline{\omega} = \{\omega_k, k \geq 1\}$ be two sets of positive numbers satisfying $\sum_{j\geq 1} \pi_j = 1$ and $\sum_{k\geq 1} \omega_k = 1$. Here $\underline{\pi}$ may be viewed as weights or prior probabilities of the procedures in Δ and similarly $\underline{\omega}$ as weights for Ξ.

For each n, choose an integer N with $1 \leq N_n \leq n$. The role of N_n as used earlier in Catoni (1997) will be discussed later. Unless stated otherwise, N_n is always chosen to be of order n. Define

$$q_{n-N+1}(x, y; z^{n-N+1}) = \sum_{j\geq 1, k\geq 1} \pi_j \omega_k p_{\hat{u}_{j,n-N+1}, \sigma_k}(x, y)$$

$$q_{n-N+2}(x, y; z^{n-N+2}) = \frac{\sum_{j\geq 1, k\geq 1} \pi_j \omega_k p_{\hat{u}_{j,n-N+1}, \sigma_k}(x_{n-N+2}, y_{n-N+2}) p_{\hat{u}_{j,n-N+2}, \sigma_k}(x, y)}{\sum_{j\geq 1, k\geq 1} \pi_j \omega_k p_{\hat{u}_{j,n-N+1}, \sigma_k}(x_{n-N+2}, y_{n-N+2})}$$

$$\cdots$$

$$q_i(x, y; z^i) = \frac{\sum_{j\geq 1, k\geq 1} \pi_j \omega_k \left(\prod_{l=n-N+1}^{i-1} p_{\hat{u}_{j,l}, \sigma_k}(x_{l+1}, y_{l+1})\right) p_{\hat{u}_{j,i}, \sigma_k}(x, y)}{\sum_{j\geq 1, k\geq 1} \pi_j \omega_k \prod_{l=n-N+1}^{i-1} p_{\hat{u}_{j,l}, \sigma_k}(x_{l+1}, y_{l+1})}$$

$$\cdots$$

$$q_n(x, y; z^n) = \frac{\sum_{j\geq 1, k\geq 1} \pi_j \omega_k \left(\prod_{l=n-N+1}^{n-1} p_{\hat{u}_{j,l}, \sigma_k}(x_{l+1}, y_{l+1})\right) p_{\hat{u}_{j,n}, \sigma_k}(x, y)}{\sum_{j\geq 1, k\geq 1} \pi_j \omega_k \prod_{l=n-N+1}^{n-1} p_{\hat{u}_{j,l}, \sigma_k}(x_{l+1}, y_{l+1})}.$$

Note that the dependence of $q_i(x, y; z^i)$ on z^i is through the estimators $\hat{u}_{j,l}$ based on $Z^l = z^l$ for $j \geq 1$

and $n - N + 1 \le l \le i$. Let

$$\hat{g}_n(y|x) = \frac{1}{N} \sum_{i=n-N+1}^{n} q_i(x, y; Z^i).$$

Given x, it is a convex combination of Gaussian densities with random weights (which does not depend on knowledge of P_X or u) and it is an estimator of the conditional density of Y given $X = x$. For a given x, let $\hat{u}_n(x)$ and $\hat{\sigma}(x)$ be the minimizer of the Hellinger distance $d_H\left(\hat{g}_n(\cdot|x), \phi_{t,s}\right)$ between $\hat{g}_n(y \mid x)$ and the normal density $\phi_{t,s}(y)$ with mean t and variance s^2 over choices $|t| \le A$ (recall that A is the known bound on the regression function) and $\underline{\sigma} < s \le \overline{\sigma}$. We use $\hat{u}_n$ as our final adaptive estimator at sample size n. Let δ^* denote this procedure producing $\{\hat{u}_n,\ n \ge 1\}$.

3.2. Risk bound.

THEOREM 1: *For any given countable collection of estimation procedures $\Delta = \{\delta_j, j \ge 1\}$, a list of variance functions $\Xi = \{\sigma_k^2(x) : k \ge 1\}$, and weights $\underline{\pi}$ and $\underline{\omega}$, we can construct a single estimation procedure δ^* as given above such that for any underlying regression function u with $\|u\|_\infty \le A$ and $\sigma \le \overline{\sigma}$, we have*

$$R\left((u,\sigma); n; \delta^*\right) \le C_{A,\overline{\sigma}} \left\{ \inf_k \left(\frac{1}{N} \log \frac{1}{\omega_k} + \|\sigma^2 - \sigma_k^2\|^2 \right) + \inf_j \left(\frac{1}{N} \log \frac{1}{\pi_j} + \frac{1}{N} \sum_{l=n-N+1}^{n} R\left((u,\sigma); l; \delta_j\right) \right) \right\}. \tag{1}$$

REMARKS:

1. The risk bound is still valid if Δ, Ξ and the weights $\underline{\pi}$ and $\underline{\omega}$ are chosen to depend on the sample size n.

2. The adaptation recipe above also produces an estimator $\hat{\sigma}^2$ of σ^2. It has risk $E\left(\|\sigma - \hat{\sigma}\|^2\right)$ bounded by a similar quantity as the above upper bound. See the remark to the proof of Theorem 1.

3. The lower bound assumption on σ is not essential. One can always make the condition satisfied by adding independently generated (small) noise to the response.

4. For an explicit expression of $C_{A,\overline{\sigma}}$, see (5) in the proof of Theorem 1.

To understand Theorem 1, we first talk about the term $(1/N)\sum_{l=n-N+1}^{n} R\left((u,\sigma); l; \delta_j\right)$. Ideally one would like to replace it with $R\left((u,\sigma); n; \delta_j\right)$ (the risk of δ_j at sample size n), which we suspect to be incorrect in general, but have not come up with a counter-example. While this remains to be proven or disproved, an application of the risk bound is generally not affected by the gap as we explain next. For an unknown regression function, the risk of a good procedure should decrease as the sample size increases. For a decreasing risk, the influence of N_n is clear: larger N_n decreases the two penalty terms in the risk bound in (1) involving the weights but increases the main term involving the risk of the

procedure. For a risk decreasing around a polynomial order $n^{-r}\eta(n)$ for some $0 < r \leq 1$ and $\eta(n)$ (e.g., $\log n$) being a slowly changing function (as is usually the case for both parametric and nonparametric estimations), $(1/N)\sum_{l=n-N+1}^{n} R((u,\sigma); l; \delta_j)$ is of the same order as $n^{-r}\eta(n)$ for any choice of $N_n \leq \tau n$ for some $0 < \tau < 1$ (the choice of $N_n = n$ results in an extra logarithmic factor for a parametric rate with $r = 1$). For such a choice of $N_n = \lfloor \tau n \rfloor$, if one uses $\hat{u}_{j,n-N+1}$ instead of $\hat{u}_{j,i}$ for all i between $n - N + 2$ and n in the construction of the adaptive estimator, one gets a most likely rougher (due to ignoring the observations $Z_{n-N+2}, ..., Z_n$) but simpler bound

$$R\left((u,\sigma); n; \delta^*\right) \leq C'_{A,\overline{\sigma}} \left\{ \inf_k \left(\frac{1}{N} \log \frac{1}{\omega_k} + \|\sigma^2 - \sigma_k^2\|^2 \right) + \inf_j \left(\frac{1}{N} \log \frac{1}{\pi_j} + R((u,\sigma); \lfloor \tau n \rfloor; \delta_j) \right) \right\}.$$

When N_n is chosen of order n, $R((u,\sigma); \lfloor \tau n \rfloor; \delta)$ is within a multiple of $R((u,\sigma); n; \delta)$ for probably almost all of the interesting applications. If that is the case (only needed for good procedures in the list which have both small risks and not too small weights), then the difference between $R((u,\sigma); n; \delta_j)$ (the ideal one) and $(1/N)\sum_{l=n-N+1}^{n} R((u,\sigma); l; \delta_j)$ in the risk bound in Theorem 1 does not have any effect on the risk bound beyond a constant factor.

Now we discuss the result of Theorem 1. The term $\inf_k \left((1/N)\log(1/\omega_k) + \|\sigma^2 - \sigma_k^2\|^2 \right)$ is a trade-off between the distance between σ^2 and σ_k^2 in Ξ and the weight on it relative to N (of order n). It is a penalty for not knowing σ^2. Theorem 1 implies that in addition to this penalty for unknown σ, we pay the price of a penalty $(1/N)\log(1/\pi_j)$ of order $1/n$ for adaptation over the estimation procedures in Δ. As will be seen, the penalties are negligible for many interesting applications.

In practice, we may assign smaller weights π_j for more complex estimation procedures. Then the risk bound in (1) is a trade-off between accuracy and complexity plus the penalty for not knowing σ. For a complex procedure (with a small prior probability), its role in the risk bound becomes significant only when the sample size becomes large.

Next we give some direct implications of Theorem 1 for several cases.

Case 1. σ^2 is known, say $\sigma^2(x) = \sigma_0^2(x)$ for some known $\sigma_0^2(x)$. We then take one element (i.e., σ_0^2) in Ξ. The risk bound becomes

$$R\left((u,\sigma_0); n; \delta^*\right) \leq C_{A,\overline{\sigma}} \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{1}{N} \sum_{l=n-N+1}^{n} R\left((u,\sigma_0); l; \delta_j\right) \right\}.$$

Case 2. $\sigma^2(x)$ is an unknown constant, i.e., $\sigma^2(x) \equiv v^2$. Accordingly we can discretize v^2 at accuracy $\sqrt{\log n}/n^{1/2}$ to get the adaptation risk bound

$$R\left((u,v); n; \delta^*\right) \leq \tilde{C}_{A,\overline{\sigma}} \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{\log n}{N} + \frac{1}{N} \sum_{l=n-N+1}^{n} R\left((u,v); l; \delta_j\right) \right\}.$$

For the above two cases, the price for adaptation is basically of order $1/n$ and $\log n/n$ respectively, which is negligible for nonparametric rates. Thus for any bounded regression function, the combined procedure performs asymptotically as well as any procedure in the list Δ (or nearly so, possibly losing a logarithmic factor for parametric rates) .

Theorem 1 implies a simple consequence on consistency. A procedure δ is said to be consistent for u if $R((u,\sigma);n;\delta) \to 0$ as $n \to \infty$.

COROLLARY 1: For Cases 1 and 2, the combined procedure δ^* is consistent whenever any of the procedures in the list Δ is so.

PROOF: Suppose δ_{j^*} is consistent. Since $R\left((u,\sigma);n;\delta_{j^*}\right) \to 0$, we have $(1/N)\sum_{l=n-N+1}^{n} R((u,\sigma);l;\delta_{j^*}) \to 0$ as $n \to \infty$. Under a choice of N_n of order n, $(1/N)\log(1/\pi_j) + \log n/N \to 0$. The consistency of δ^* follows from the risk bounds for the above two cases. This completes the proof of Corollary 1.

As will be illustrated later in Section 5, using a combined adaptive procedure, we can seek opportunity of various convergence rates without sacrificing consistency.

Case 3. $\sigma^2(x)$ is known to be in a class of variance functions. More generally than the above two cases, let Σ be a class of variance functions bounded above and below. Let $M(\epsilon)$ be the $L_2(P_X)$ metric entropy of Σ, i.e., $M(\epsilon)$ is the logarithm of the smallest size of an ϵ-cover of Σ under the $L_2(P_X)$ distance. Let ϵ_n be determined by $M(\epsilon_n)/N = \epsilon_n^2$. Take $\underline{\omega}$ to be a uniform weight on a covering set, i.e., $\omega_k = e^{-M(\epsilon_n)}$. This choice of ϵ_n gives a good trade-off between $(1/N)\log(1/\omega_j)$ and $\|\sigma^2 - \sigma_k^2\|^2$. We have the following corollary.

COROLLARY 2: *For any given Δ, a class of variance function Σ and weight $\underline{\pi}$, we can construct a single estimation procedure δ^* such that for any underlying regression function u with $\|u\|_\infty \le A$ and $\sigma^2 \in \Sigma$, we have*

$$R((u,\sigma);n;\delta^*) \le C_{A,\overline{\sigma}} \left\{ \epsilon_n^2 + \inf_j \left(\frac{1}{N}\log\frac{1}{\pi_j} + \frac{1}{N}\sum_{l=n-N+1}^{n} R((u,\sigma);l;\delta_j) \right) \right\}. \tag{2}$$

REMARK: If σ is known to be in any one in a collection of families of variance functions $\{\Sigma_l, l \ge 1\}$, then a similar adaptation risk bound holds with an additional penalty term $(1/N)\log(1/\varsigma_l)$, where $\{\varsigma_l : l \ge 1\}$ is the weight on $\{\Sigma_l, l \ge 1\}$.

If Σ is a parametric family (i.e., of finite-dimension) as in Case 2, then $M(\epsilon)$ is usually of order $\log(1/\epsilon)$, resulting in ϵ_n^2 of order $\log n/n$, which is negligible for a nonparametric rate of convergence. Thus for nonparametric estimation, we lose little not knowing the variance function in a parametric

class. In contrast, when Σ is a nonparametric class, we may pay a rather high price. For instance, if Σ consists of all nondecreasing functions, then $M(\epsilon)$ is of order $1/\epsilon$ and ϵ_n^2 determined accordingly is of order $n^{-2/3}$, which is damaging for a fast nonparametric rate, e.g., $n^{-4/5}$.

From Theorem 1, by mixing different procedures, we have a single procedure that shares the advantages of them automatically in terms of the risk. Besides the L_2 risk as we consider, other performance measures (e.g., L_∞ risk or local risks) are useful from both theoretical and practical point of views. It then is of interest to investigate if similar adaptation procedures exist for other loss functions and if not, what are the prices one needs to pay for adaptation.

The result of Theorem 1 can go somewhat beyond what we have said so far (i.e., roughly *if one procedure in the list works well, so does the combined adaptive procedure*). Even if none of the procedures in Δ works optimally for an unknown regression function u in some sense (e.g., in terms of minimax rate of convergence as will be studied in Section 4), the combined strategy can still be optimal. The hope is that there exists a sequence of procedures in the list approaching an optimal one and further more, the weights (prior probabilities) on these procedures do not decrease too fast. For such a case, a genuine trade-off between accuracy $\frac{1}{N}\sum_{l=n-N+1}^{n} R\left((u,\sigma);l;\delta_j\right)$ and weight $\frac{1}{N}\log\frac{1}{\pi_j}$ is necessary. More discussion will be given in Section 4 on adaptation over an uncountable collection of classes of regression functions.

3.3. A practical consideration. The adaptation procedure in Section 3.1 is hard to implement in practice since the estimator $\hat{u}_n$ depends on a minimization step involving the Hellinger distance. An alternative approach to construct a regression estimator based on the conditional density estimator $\hat{g}_n$ is simply taking the mean value of the density $\hat{g}_n$ at each x. Because $\hat{g}_n$ is a mixture of Gaussians, the mean is easily computed. Intuitively the modified estimator should behave similarly as $\hat{u}_n$ because closeness of $\hat{g}_n(y|x)$ to $p_{\hat{u}_n,\hat{\sigma}}$ at each x under the Hellinger distance suggests that the mean of $\hat{g}_n$ at each x should be close to $\hat{u}_n(x)$ (at least under some conditions). The computation then lies mainly in computing the estimators produced by the procedures in the collection Δ at various sample sizes (e.g., from $n/2$ to n). If the number of procedures being considered depends on the sample size n and it increases polynomially in n, and when the computation times for the original procedures are all of polynomial orders uniformly, then the computation time for the combined estimator is also polynomial in n.

Another practical consideration is on σ^2. Discretization as used in Section 3.1 substantially increase computation. When good estimators of σ^2 are available, simpler adaptive procedures can be constructed. See Section 6 for details.

After the submission of this work, further developments toward applications have been made by the author. Built on the work here, a practically feasible algorithm ARM is proposed. Simulation results support the theoretical findings. Some of these results are reported in Yang (1999c).

4. Adaptation with respect to function classes. In light of Theorem 1, the adaptation recipe can be used to derive minimax-rate adaptive estimators over function classes. For simplicity, in this section, we assume $\sigma^2(x)$ is a unknown constant bounded above by $\overline{\sigma}^2$.

4.1. *Minimax-rate adaptation.* Let $\mathcal{U}$ be a class of regression functions. Consider the minimax risk for estimating a regression in $\mathcal{U}$:

$$R(\mathcal{U}; n) = \min_{\hat{u}_n} \max_{u \in \mathcal{U}, \sigma \le \overline{\sigma}} E \parallel u - \hat{u}_n \parallel^2,$$

where $\hat{u}_n$ is over all estimators based on $Z^n = (X_i, Y_i)_{i=1}^n$ and the expectation is taken under u and σ. The minimax risk measures how well one can estimate u uniformly over the class $\mathcal{U}$. Let $\{\mathcal{U}_j,\ j \ge 1\}$ be a collection of classes of regression functions uniformly bounded between $-A$ and A. Assume the true function u is in (at least) one of the classes, i.e., $u \in \cup_{j \ge 1} \mathcal{U}_j$. The question we want to address is: Without knowing which class contains u, can we have a single estimator such that it converges automatically at the minimax optimal rate of the class that contains u? If such an estimator exists, we call it a minimax-rate adaptive estimator with respect to the classes $\{\mathcal{U}_j,\ j \ge 1\}$.

A lot of results have been obtained on minimax-rate adaptation (or even with the right constant for some cases) for specific functions classes such as Sobolev and Besov under various performance measures, including Efroimovich and Pinsker (1984), Efroimovich (1985), Härdle and Marron (1985), Lepski (1991), Golubev and Nussbaum (1992), Donoho and Johnstone (e.g., 1998), Delyon and Juditsky (1994), Mammen and van de Geer (1997), Tsybakov (1995), Goldenshluger and Nemirovski (1997), Lepski et al (1997), Devroye and Lugosi (1997), and others. A method is proposed by Juditsky and Nemirovski (1996) to aggregate estimators to adapt to within order $n^{-1/2}$ in risk. General schemes have also been proposed for the construction of adaptive estimators in Barron and Cover (1991) based on minimum description length (MDL) criterion using ϵ-nets. Other adaptation schemes and adaptation bounds by model selection have been developed later including very general penalized contrast criteria in Birgé and Massart (1996), and Barron, Birgé and Massart (1999) with many interesting applications on adaptive estimation; penalized maximum likelihood or least squares criteria in Yang and Barron (1998) and Yang (1999a); and complexity penalized criteria based on V-C theory (e.g., Lugosi and Nobel (1996)). In the opposite direction, for the case of estimating a regression function at a point,

negative results have been obtained by Lepski (1991) and Brown and Low (1996)).

We give below a general result on minimax-rate adaptation without requiring any special property on the classes over which adaptation is desired. Some regularity conditions will be used for our results.

Definition: If the minimax risk sequence satisfies

$$R(\mathcal{U}; \lfloor n/2 \rfloor) \asymp R(\mathcal{U}; n),$$

we say the minimax risk of the class $\mathcal{U}$ is *rate-regular*.

A rate of convergence is said to be nonparametric if it converges no faster than n^{θ} for some $0 < \theta < 1$.

The familiar rates of convergence $n^{-\alpha}(\log n)^{\beta}$ for some $0 < \alpha \leq 1$ and $\beta \in R$ are rate-regular. For a rate-regular risk, together with that $R(\mathcal{U}; i)$ is nonincreasing in the sample size i, we have that $(1/n)\sum_{\lfloor n/2 \rfloor}^{n} R(\mathcal{U}; i)$ is of order $R(\mathcal{U}; n)$.

4.2. *Result.* We have the following result on minimax-rate adaptation.

THEOREM 2: *Let $\{\mathcal{U}_j,\ j \geq 1\}$ be any collection of uniformly bounded function classes. Assume further that the minimax risk of each of the classes is rate-regular. Then we can construct a minimax-rate adaptive procedure such that it automatically achieves the optimal rate of convergence for any class in the collection with a nonparametric regular-rate and it is within a logarithmic factor of the optimal for a parametric rate.*

REMARK: If σ^2 is known, we do not need to pay any price in terms of rates. That is, a minimax-rate adaptive estimator can be constructed for any countable collection of rate-regular classes (parametric or nonparametric).

Theorem 2 shows the existence of minimax-rate adaptive estimators over a countable collection of function classes. Conclusions can also be made for adaptation over an uncountable collection of classes such as Besov classes as we discuss next.

4.3. Adaptation over an uncountable collection of classes. Consider a collection of function classes $\{\mathcal{U}_\alpha : \alpha \in (0, \infty)\}$ indexed by a continuous hyper-parameter α (e.g., a smoothness parameter). To apply the adaptation recipe, we need a suitable discretization of α and a proper assignment of weight π_j such that for every class $\mathcal{U}_{\alpha_0}$, we can find a sequence of classes $\mathcal{U}_{\alpha_n}$ in the discretization approaching $\mathcal{U}_{\alpha_0}$ suitably quickly with the weight on α_n not decreasing too fast.

Assume $\{\mathcal{U}_\alpha : \alpha \in (0, \infty)\}$ satisfy the conditions in Theorem 2 and has an ordering relationship $\mathcal{U}_{\alpha_1} \subset \mathcal{U}_{\alpha_2}$ for $\alpha_1 > \alpha_2$. Let $M(n; \alpha)$ be an upper bound on the minimax risk $R(\mathcal{U}_\alpha; n)$ of class $\mathcal{U}_\alpha$ of the right order. Assume that there exists a decreasing sequence $b_n \to 0$ such that for each α, $\alpha_n \uparrow \alpha$ no

slower than b_n implies $M(n;\alpha_n)/M(n;\alpha)$ is bounded (the bound is allowed to depend on α).

THEOREM 3: Under the above conditions, we can construct an adaptive strategy δ^* such that for each α,

$$\max_{u\in\mathcal{U}_\alpha,\sigma\le\overline{\sigma}} R((u,\sigma);n;\delta^*) \le C_{A,\alpha,\overline{\sigma}}\left(\frac{\log(1/b_n)}{n}+\frac{\log n}{n}+R(\mathcal{U}_\alpha;n)\right).$$

REMARK: The result can be easily generalized to the case when α is a vector of parameters (e.g., Besov classes).

As long as b_n can be chosen reasonably slow, say, $b_n \asymp n^{-\beta}$ for some $\beta > 0$, the additional penalty $\log(1/b_n)/n$ is basically of order $\log n/n$. The familiar nonparametric rate of convergence is $n^{-2\alpha/(2\alpha+d)}(\log n)^{h(\alpha)}$ for some $\alpha > 0$ (a smoothness parameter), d (dimension) and $h(\alpha)$ (e.g., $h(\alpha)=0$) (see, e.g., Birgé (1986) and Yang and Barron (1999)). Then b_n can be chosen as large as $1/\log n$ and the combined strategy converges at the optimal rate for a nonparametric rate $R(\mathcal{U}_\alpha;n)$. The proposition applies to classical function classes such as ellipsoidal, Hölder and others.

5. An illustration. Consider estimating a regression function on $[0,1]^d$ assumed to be in $\mathcal{H}(C)$ which consists of all functions that are bounded between $-C$ and C for a known positive constant C. Assume that we have a list $\mathcal{L}$ of a few candidate regression procedures obtained under different assumptions. Not knowing if any of these procedures works well for the unknown regression function u, it is hypothesized that u might be in one of a collection of functions classes denoted by $\mathcal{T}$ (which are not handled well by the procedures in $\mathcal{L}$). Some concerns here are: (a). parametric versus nonparametric: we want the flexibility of nonparametric approaches in $\mathcal{L}$ but do not want to lose too much accuracy when some simple parametric model in $\mathcal{L}$ happens to work well; (b). adaptation with respect to different nonparametric procedures in $\mathcal{L}$; (c). consistency versus rates of convergence: we want the estimator to be consistent for every regression function in $\mathcal{H}(C)$ yet it permits fast rates of convergence when the hypothesis that the true regression function u is in a class in $\mathcal{T}$ happens to be right.

The collection $\mathcal{L}$ may be chosen to include a few familiar parametric procedures (e.g., simple linear or generalized linear models) as well as some nonparametric procedures, for instance, kernel estimator with automatically selected bandwidth (e.g., Härdle and Marron (1985) and Devroye and Lugosi (1997)), CART (Breiman et al (1984)) and others as mentioned in the introduction section of this paper.

The following classes will be involved in one choice of $\mathcal{T}$.

1. *Besov classes of different interaction orders.* For $1 \le \sigma, q \le \infty$ and $\alpha > 0$, let $B^{\alpha,r}_{q,\sigma}(C)$ be the collections of all functions $g \in L_q[0,1]^r$ such that the Besov norm satisfies $\| g \|_{B^\alpha_{q,\sigma}} \le C$ (see, e.g., Triebel (1992) and DeVore and Lorentz (1993)). When $\alpha/d > 1/q$, the functions in $B^\alpha_{q,\sigma}$ are uniformly bounded.

Consider the following function classes on $[0,1]^d$:

$$S_{q,\sigma}^{\alpha,1}(C) = \{\sum_{i=1}^d g_i(x_i) : g_i \in B_{q,\sigma}^{\alpha,1}(C), 1 \le i \le d\}$$

$$S_{q,\sigma}^{\alpha,2}(C) = \{\sum_{1\le i<j\le d} g_{i,j}(x_i,x_j) : g_{i,j} \in B_{q,\sigma}^{\alpha,2}(C), 1 \le i < j \le d\}$$

$$\ldots$$

$$S_{q,\sigma}^{\alpha,d}(C) = B_{q,\sigma}^{\alpha,d}(C).$$

The simplest function class $S_{q,\sigma}^{\alpha,1}(C)$ contains additive functions (no interaction). These classes have different input dimensions with increasing complexity when r increases. The metric entropies of these classes are of the same orders as $B_{q,\sigma}^{\alpha,1}(C)$, ..., $B_{q,\sigma}^{\alpha,d}(C)$ respectively. By results of Yang and Barron (1999), if the unknown design density with respect to Lebesgue measure is bounded above and away from zero on $[0,1]^d$, the minimax rate of convergence under square L_2 loss for estimating a regression function in $S_{q,\sigma}^{\alpha,r}(C)$ is $n^{-2\alpha/(2\alpha+r)}$ for $1 \le r \le d$, as suggested by the heuristic dimensionality reduction principle of Stone (1985). Rates of convergence and adaptive estimation using wavelets for $B_{q,\sigma}^{\alpha,1}(C)$ with α unknown are studied in Donoho and Johnstone (1998). Results on non-adaptive rates of convergence in more restrictive Sobolev classes are given in Stone (1994) and Nicoleris and Yatracos (1997). Adaptive estimation over Sobolev classes with unknown smoothness and interaction order by model selection is in Yang (1999a).

2. *Neural network classes.* Let $N(C)$ be the closure in $L_2[0,1]^d$ of the set of all functions $g : R^d \to R$ of the form $g(x) = c_0 + \sum_i c_i\sigma(v_i \cdot x + b_i)$, with $|c_0| + \sum_i |c_i| \le C$, and $\|v_i\| = 1$, where σ is the step function $\sigma(t) = 1$ for $t \ge 0$, and $\sigma(t) = 0$ for $t < 0$. The minimax rate under square L_2 loss is shown to be bounded between

$$n^{-(1+2/d)/(2+1/d)}(\log n)^{-(1+1/d)(1+2/d)/(2+1/d)} \text{ and } (n/\log n)^{-(1+1/d)/(2+1/d)} \tag{3}$$

when the design density is bounded above and away from zero on $[0,1]^d$ (see Yang and Barron (1999)). When d is moderately large, the rate is slightly better than $n^{-1/2}$ (independent of d).

Let $\mathcal{T}$ consist of all the Besov classes of different interaction order and smoothness parameters, and the neural network class. Note that for some of the Besov classes, from Donoho and Johnstone (1998), linear procedures such as kernel estimators can not be optimal, suggesting the need of other estimation procedures. The desire here is that the to-be-constructed adaptive procedure automatically adapts to the interaction order and smoothness over the Besov classes, and retains a good rate $o(n^{-1/2})$ if unfortunately both the interaction-order is high and the smoothness parameter α is small relative to d (i.e., curse of dimensionality, as well-known) but fortunately it has the neural net representation. More

function classes with different characterizations can be considered here as well to increase the chance to capture the underlying regression function to overcome the curse of dimensionality.

METHOD OF ADAPTATION. To achieve our goals, it suffices to construct a consistent estimator for $\mathcal{H}(C)$ and optimal-rate estimators for the Besov classes and the neural net work class separately and then combine them appropriately together with the procedures in $\mathcal{L}$.

For regression estimation, universally consistent estimators have been derived under L_q loss without any assumption on the joint distribution of (X, Y) other than the necessary existence of the corresponding moment of Y (see, e.g., Stone (1977) and Devroye and Wagner (1980)). Thus we have a consistent estimator for $\mathcal{H}(C)$ under $L_2(P_X)$ loss. By Theorem 3, it can be shown that, in principle, for each choice of $1 \leq r < d$, one can construct an adaptive estimator for the classes $S^{\alpha,r}_{q,\sigma}(C)$ with α and q satisfying $\alpha > 1/q$ by a suitable discretization of α and q using the adaptation recipe. Then one can combine the d adaptive estimators to obtain further adaptivity in terms of the interaction order as well. For the neural network class $N(C)$, from (3) an estimator can be obtained at a rate $o(n^{-1/2})$. Estimators at rate $O(\log n/n^{1/2})$ using finite-dimensional neural network models are in e.g., Barron (1994).

Finally, we combine the above three procedures together with the ones in $\mathcal{L}$ (e.g., with equal weights). Then the combined procedure has the desired properties, i.e., it is consistent for $\mathcal{H}(C)$, adapts with respect to the procedures in $\mathcal{L}$ (possibly losing a logarithmic factor for parametric rates here, but it can be avoided if one uses good estimators of σ^2 instead of discretization as will be discussed in Section 6), adapts to smoothness and interaction order of the Besov classes, and converge at a rate $o(n^{-1/2})$ if the true regression function has a neural net representation. In addition, if a procedure in $\mathcal{L}$ converges at a good rate for a nonparametric class, so does the combined procedure. Recently, Donoho (1997) shows that a dyadic CART is nearly minimax-rate adaptive (within a logarithmic factor) to unknown anisotropic smoothness for the case $d = 2$ with equally spaced fixed design. Assuming a similar result holds for general d with a random design, the above final adaptive procedure shares that property as well.

6. Adaptation utilizing estimators of σ^2. In the construction of an adaptive estimator in Section 3, a discretization is used for σ^2. When good estimators of σ^2 are available, simpler and better adaptive procedures can be constructed.

6.1. Adaptation using an independent estimator of σ^2. Sometimes, it is possible to have a good estimator of σ^2, e.g., by nearest neighbor method (e.g., Stone (1977)) or based on additional information. Then a different adaptation recipe can be used. Let $\hat{\sigma}$ be an estimator of σ independent of the sample

$Z_1, ..., Z_n$ (or one could set aside a portion of data for the estimation of σ). Then we use this estimator in the definition of q_l's (instead of mixing over Ξ) in Section 3.1, i.e., $q_{n-N+1}(x,y)$ is redefined as $\sum_{j\geq 1}\pi_j p_{\hat{u}_{j,n-N+1},\hat{\sigma}(x)}(x,y)$ and we make similar modifications for others. Proceed as before and let $\tilde{u}_n(x)$ be the minimizer of the Hellinger distance $d_H\left(\hat{g}_n(\cdot|x), \phi_{t,\hat{\sigma}(x)}\right)$ between $\hat{g}_n(y \mid x)$ and the normal density $\phi_{t,\hat{\sigma}(x)}(y)$ with mean t and variance $\hat{\sigma}^2(x)$ over choices of t with $|t| \leq A$. Take $\tilde{u}_n$ as our final adaptive estimator and call this estimation procedure δ_*. Let $\underline{s}$ and $\overline{s}$ denote the minimum and maximum value of $\hat{\sigma}(x)$ over x respectively. Similarly, let $\underline{\sigma}$ and $\overline{\sigma}$ denote the minimum and maximum of $\sigma(x)$ respectively.

THEOREM 4: *For the combined procedure, for any regression function* u with $\| u \|_\infty \leq A$, *we have*

$$R((u,\sigma);n;\delta_*) \leq \frac{8A^2}{1-Ee^{-A^2/(\overline{\sigma}^2+\overline{s}^2)}} \cdot \left(E\left(\left(2+\log\left(1+\frac{\overline{s}^2}{\underline{\sigma}^2}\right)\right)\|\frac{\sigma^2-\hat{\sigma}^2}{\hat{\sigma}^2}\|^2\right)\right.$$
$$\left. + \inf_j\left\{\frac{2}{N}\log\frac{1}{\pi_j} + E\left(\frac{1}{\underline{s}^2}\right)\frac{1}{N}\sum_{l=n-N+1}^{n} R((u,\sigma);l;\delta_j)\right\}\right).$$

REMARKS:

1. The above bound is not very useful when $Ee^{-A^2/(\overline{\sigma}^2+\overline{s}^2)}$ is close to 1 or $E\left(\underline{s}^{-2}\right)$ is large. If $\sigma(x)$ is uniformly upper bounded, a good estimator should have $Ee^{-A^2/(\overline{\sigma}^2+\overline{s}^2)}$ bounded away from 1. The term $E\left(\underline{s}^{-2}\right)$ is likely to be large when $\sigma(x)$ may be close to zero. To get it controlled mathematically, one can always add independently generated noise to the responses and restrict attention accordingly to $\underline{s}$ bounded away from zero. The increase of noise level usually do not change the risk beyond a constant factor.

2. If there are several plausible estimators of $\sigma(x)$ available independent of Z^n, one can mix over them as well to obtain a similar result.

From the above theorem, with σ^2 estimated, the adaptive procedure basically pays the price of the discrepancy of the variance estimator and $(1/N)\log(1/\pi_j)$ (of order $1/n$). Note that we do not require a known upper bound on σ above. If an estimator $\hat{\sigma}^2$ converges at rate $1/n$ in mean square error, then unlike the adaptation procedure by discretizing σ^2, the above adaptation risk bound does not lose a logarithmic factor for parametric cases.

6.2. Adaptation with variance estimators from regression procedures. Many regression procedures provide estimates of both the regression function and the variance function $\sigma(x)$ (some assuming the variance function is constant). One can construct an adaptive procedure accordingly. Assume that $\sigma(x)$ is upper bounded by a known constant $\overline{\sigma}$ and lower bounded by $\overline{\sigma}^{-1}$. Let $\Delta = \{\delta_j : j \geq 1\}$ be a collection of regression procedures with δ_j producing estimator $\hat{u}_{j,l}$ and $\hat{\sigma}_{j,l}$ based on Z^l for $l \geq 1$ (the variance

estimators are assumed to take values in the known range). Redefine $q_{n-N+1}(x,y)$ in Section 3.1 as $\sum_{j\ge 1}\pi_j p_{\hat{u}_{j,n-N+1},\hat{\sigma}_{j,n-N+1}(x)}(x,y)$ and make similar modifications for others. Proceed as before and let $\overline{u}_n(x)$ and $\hat{\sigma}_n(x)$ be the minimizer of the Hellinger distance $d_H\left(\hat{g}_n(\cdot|x),\phi_{t,s}\right)$ between $\hat{g}_n(y\mid x)$ and the normal density $\phi_{t,s}(y)$ with mean t and variance s^2 over choices of t with $|t|\le A$ and $s\le\overline{\sigma}$. Take $\overline{u}_n$ as our final adaptive estimator and call this estimation procedure $\delta_\dagger$.

THEOREM 5: *The combined procedure satisfies that for any regression function u with $|u|\le A$, we have*

$$R((u,\sigma);n;\delta_\dagger)\le C_{A,\overline{\sigma}}\inf_j\left\{\frac{1}{N}\log\frac{1}{\pi_j}+\frac{1}{N}\sum_{l=n-N+1}^{n}E\|\frac{\sigma^2-\hat{\sigma}^2_{j,l}}{\hat{\sigma}^2_{j,l}}\|^2+\frac{1}{N}\sum_{l=n-N+1}^{n}E\|\frac{u-\hat{u}_{j,l}}{\hat{\sigma}_{j,l}}\|^2\right\}.$$

REMARKS:

1. If some regression procedures in the collection Δ do not provide estimators of the variance function, for the construction of an adaptive estimator, one can use an independent variance estimator (if available), or borrow a variance estimator from another procedure, or discretize σ^2 as in Section 3 for these procedures to get a similar result.

2. As discussed before, with $N\sim\tau n$ for some $0<\tau<1$, if good procedures have decreasing risks, and if the variance estimators are bounded away from zero, then the above upper bound is basically of order $\inf_j\{(1/N)\log(1/\pi_j)+E\|\sigma^2-\hat{\sigma}^2_{j,n}\|^2+E\|u-\hat{u}_{j,n}\|^2\}$. If a parametric procedure with a variance estimator of order $1/n$ in risk in the collection Δ happens to be optimal, the above adaptive procedure avoids a possible extra logarithmic factor compared to that by discretizing σ^2 in Case 2 in Section 3.2.

3. Throughout the paper, the errors are assumed to be normally distributed. This assumption is not essential for the main results as long as the shape of the error distribution is known. In a later work by the author (Yang (1999c)), similar adaptation methods are proposed for a general error distribution (e.g., double exponential).

4. Only random designs are studied in this work. It is not clear to us if similar results hold for fixed designs.

7. Discussion. Adaptive function estimation has attracted a lot of attention in recent years. Many adaptive estimators have been proposed for smoothness function classes. Adaptivity of these estimators with respect to smoothness parameters basically comes from a certain automatic selection of a tuning parameter associated with a general procedure (e.g., bandwidth for a kernel or local polynomial procedure, a smoothing parameter for smoothing splines, order of approximation for series expansion estimators, or a subset of a wavelet expansion). In terms of global risks, general model selection theories

developed in pioneering work of Barron and Cover (1991) and subsequent papers (e.g., Barron, Birgé and Massart (1999), and Yang (1999a)) provide more flexibility by allowing models of many different basis (e.g., wavelets and neural nets) to be considered at the same time and as a consequence, the estimators can adapt to different types of characteristics. For the theories obtained in that direction, the models can be quite general in terms of approximation of the true regression function, but are still restricted to be of similar nature (e.g., of finite metric dimension) with similar estimation methods (e.g., by minimum contrasts).

The adaptation schemes given in this paper allow one to combine advantages of any countable collection of regression procedures (in terms of L_2 risk) without requiring any restrictive properties on the procedures. This provides more flexibility in estimating a regression function. Thus estimation procedures (including adaptive ones as mentioned above) designed under various (possibly completely different) assumptions can be combined at the same time, significantly increasing the chance of capturing the true characteristics of the unknown regression function.

8. Proofs of the results.

PROOF OF THEOREM 1: We first construct an adaptive estimator of the conditional density of Y given X and then derive an adaptive regression estimator based on it. Given the conditional variance $\sigma^2(x)$, the estimators of the regression function naturally give estimators of the conditional density of Y given $X = x$ by $\hat{p}^{\sigma}_{j,i}(y|x; Z^i) = p_{\hat{u}_{j,i},\sigma}(x, y)$ for $i \geq 1, j \geq 1$. For simplicity, let $i_0 = n - N + 1$ and denote $(x_l, y_l)_{l=i_0+1}^{n+1}$ by $z_{i_0+1}^{n+1}$. Let

$$g_j^{\sigma}\left(z_{i_0+1}^{n+1}\right) = \Pi_{l=i_0}^{n} \hat{p}^{\sigma}_{j,l}(y_{l+1}|x_{l+1}; z^l).$$

Conditioned on z^{i_0} and $x_{i_0+1}, ..., x_{n+1}$, it is a density in $y_{i_0+1}, ..., y_{n+1}$. Now mix these densities over different procedures (j) and different variance functions σ_k to get

$$g^{(n)}\left(z_{i_0+1}^{n+1}\right) = \sum_{j\geq 1, k\geq 1} \pi_j \omega_k g_j^{\sigma_k}\left(z_{i_0+1}^{n+1}\right).$$

Ignoring that $P = P_X$ is unknown, $\hat{f}_l(x, y) = q_l(x, y; Z^l)$ can be viewed as an estimator of the joint density of (X, Y) (or conditional density of Y given X) with respect to the product measure of P and Lebesgue (or Lebesgue). The cumulative risk of $\hat{f}_l(x, y)$ based on Z^l, $i_0 \leq l \leq n$ satisfy

$$\begin{aligned}
&\textstyle\sum_{l=i_0}^{n} E D(p_{u,\sigma} \parallel \hat{f}_l) = \sum_{l=i_0}^{n} E \int p_{u,\sigma}(x, y) \log \frac{p_{u,\sigma}(x,y)}{\hat{f}_l(x,y)} dy P(dx) \\
&= \textstyle\sum_{l=i_0}^{n} E \int p_{u,\sigma}(x_{l+1}, y_{l+1}) \log \frac{p_{u,\sigma}(x_{l+1},y_{l+1})}{\hat{f}_l(x_{l+1},y_{l+1})} dy_{l+1} P(dx_{l+1}) \\
&= E \int \Pi_{l=i_0}^{n} p_{u,\sigma}(x_{l+1}, y_{l+1}) \left(\textstyle\sum_{l=i_0}^{n} \log \frac{p_{u,\sigma}(x_{l+1},y_{l+1})}{q_l(x_{l+1},y_{l+1};Z_1,...,Z_{i_0},z_{i_0+1},...,z_{l+1})} \right) dy_{i_0+1} \cdots dy_{n+1} P(dx_{i_0+1}) \cdots P(dx_{n+1}) \\
&= E \int \Pi_{l=i_0}^{n} p_{u,\sigma}(x_{l+1}, y_{l+1}) \left(\log \frac{\Pi_{l=i_0}^{n} p_{u,\sigma}(x_{l+1},y_{l+1})}{g^{(n)}\left(z_{i_0+1}^{n+1}\right)} \right) dy_{i_0+1} \cdots dy_{n+1} P(dx_{i_0+1}) \cdots P(dx_{n+1})
\end{aligned}$$

For any u, σ and any $j \geq 1$, $k \geq 1$, since $\log(x)$ is an increasing function, we have

$$\begin{aligned}
& \int \Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1}) \left(\log \frac{\Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1})}{g^{(n)}\left(z_{i_0+1}^{n+1}\right)} \right) dy_{i_0+1} \cdots dy_{n+1} P(dx_{i_0+1}) \cdots P(dx_{n+1}) \\
\leq & \int \Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1}) \left(\log \frac{\Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1})}{\pi_j \omega_k g_j^{\sigma_k}\left(z_{i_0+1}^{n+1}\right)} \right) dy_{i_0+1} \cdots dy_{n+1} P(dx_{i_0+1}) \cdots P(dx_{n+1}) \\
= & \log \frac{1}{\pi_j} + \log \frac{1}{\omega_k} + \int \Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1}) \log \frac{\Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1})}{g_j^{\sigma_k}(z_{i_0+1}^{n+1})} dy_{i_0+1} \cdots dy_{n+1} P(dx_{i_0+1}) \cdots P(dx_{n+1}).
\end{aligned}$$

The last term above can be bounded in terms of risks of the estimators produced by strategy δ_j. Indeed, as earlier but going backwards together with Lemma 1, we have

$$\begin{aligned}
& E \int \Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1}) \log \frac{\Pi_{l=i_0}^n p_{u,\sigma}(x_{l+1}, y_{l+1})}{g_j^{\sigma_k}\left(z_{i_0+1}^{n+1}\right)} dy_{i_0+1} \cdots dy_{n+1} P(dx_{i_0+1}) \cdots P(dx_{n+1}) \\
= & \sum_{l=i_0}^n E D(p_{u,\sigma} \parallel \hat{p}_{j,l}^{\sigma_k}) \\
= & \sum_{l=i_0}^n \left(E\left(\frac{\sigma^2(X) - \sigma_k^2(X)}{2\sigma_k^2(X)} \right) - \frac{1}{2} E \log \left(1 + \frac{\sigma^2(X) - \sigma_k^2(X)}{\sigma_k^2(X)} \right) + E\left(\frac{1}{2\sigma_k^2(X)} (u(X) - \hat{u}_{j,l}(X))^2 \right) \right) \\
\leq & \sum_{l=i_0}^n \left(C_{\overline{\sigma}} E\left(\sigma^2(X) - \sigma_k^2(X)\right)^2 + \frac{\overline{\sigma}^2}{2} E\left(\| u - \hat{u}_{j,l} \|^2 \right) \right),
\end{aligned}$$

where for the last step, we use Lemma 1 and the assumption that $\sigma_k(x)$ is lower bounded away from zero by $\overline{\sigma}^{-1}$, and the constant $C_{\overline{\sigma}}$ equals $\overline{\sigma}^2((\overline{\sigma}^{-4} - 1) - \log(\overline{\sigma}^{-4}))/\left(2(\overline{\sigma}^{-4} - 1)^2\right)$. Thus we have

$$\sum_{l=i_0}^n E D(p_{u,\sigma} \parallel \hat{f}_l) \leq \log \frac{1}{\pi_j} + \log \frac{1}{\omega_k} + \sum_{l=i_0}^n \left(C_{\overline{\sigma}} \| \sigma^2 - \sigma_k^2 \|^2 + \frac{\overline{\sigma}^2}{2} E\left(\| u - \hat{u}_{j,l} \|^2 \right) \right). \tag{4}$$

For $\hat{g}_n(y|x) = (1/N) \sum_{i=i_0}^n \hat{f}_i(x, y)$, by convexity, we have

$$E D(p_{u,\sigma} \parallel \hat{g}_n) \leq \frac{1}{N} \sum_{l=i_0}^n E D(p_{u,\sigma} \parallel \hat{f}_l).$$

Since the above inequality (4) holds for all j and k, minimizing over j and k, we have

$$E D(p_{u,\sigma} \parallel \hat{g}_n) \leq \inf_k \left\{ \frac{1}{N} \log \frac{1}{\omega_k} + C_{\overline{\sigma}} \| \sigma^2 - \sigma_k^2 \|^2 \right\} + \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{\overline{\sigma}^2}{2N} \sum_{l=i_0}^n R((u, \sigma); l; \delta_j) \right\}.$$

Since the square Hellinger distance is upper bounded by the K-L divergence, the above risk bound also upper bounds $E d_H^2(p_{u,\sigma}, \hat{g}_n)$.

Now let us derive an estimator of u based on $\hat{g}_n$. Note that $\hat{g}_n(y|x)$ is a mixture of Gaussians depending on the procedures in the collection, variance functions in the list Ξ and the weights $\underline{\pi}$ and $\underline{\omega}$ (but not on P or u). For a given x, $\hat{u}_n(x)$ and $\hat{\sigma}(x)$ minimize the Hellinger distance $d_H\left(\hat{g}_n(\cdot|x), \phi_{t,s}\right)$

between $\hat{g}_n(y \mid x)$ and the normal density $\phi_{t,s}(y)$ with mean t and variance s^2 over choices $|t| \leq A$, $\overline{\sigma}^{-1} \leq s \leq \overline{\sigma}$. By triangle inequality, given x,

$$\begin{aligned} d_H\left(\phi_{u(x),\sigma(x)}, \phi_{\hat{u}_n(x),\hat{\sigma}(x)}\right) &\leq d_H\left(\phi_{u(x),\sigma(x)}, \hat{g}_n(\cdot \mid x)\right) + d_H\left(\phi_{\hat{u}_n(x),\hat{\sigma}(x)}, \hat{g}_n(\cdot \mid x)\right) \\ &\leq 2 d_H\left(\phi_{u(x),\sigma(x)}, \hat{g}_n(\cdot \mid x)\right). \end{aligned}$$

As a consequence,

$$\begin{aligned} d_H^2(p_{u,\sigma}, p_{\hat{u}_n,\hat{\sigma}}) &= \int d_H^2\left(\phi_{u(x),\sigma(x)}, \phi_{\hat{u}_n(x),\hat{\sigma}(x)}\right) P(dx) \\ &\leq 4 \int d_H^2\left(\phi_{u(x),\sigma(x)}, \hat{g}_n(\cdot \mid x)\right) P(dx) = 4 d_H^2(p_{u,\sigma}, \hat{g}_n). \end{aligned}$$

From Lemma 1, we have that

$$d_H^2\left(\phi_{u(x),\sigma(x)}, \phi_{\hat{u}_n(x),\hat{\sigma}(x)}\right) \geq 2\left(1 - e^{-(u(x)-\hat{u}_n(x))^2/\left(4\sigma(x)^2+4\hat{\sigma}(x)^2\right)}\right) \geq 2\left(1 - e^{-(u(x)-\hat{u}_n(x))^2/\left(8\overline{\sigma}^2\right)}\right)$$

The concave function $(1 - e^{-v})$ is above the chord $(v/B)(1 - e^{-B})$ for $0 \leq v \leq B$. Thus using $v = (u(x) - \hat{u}_n(x))^2/\left(8\overline{\sigma}^2\right)$ and $B = A^2/2$, we obtain

$$\int (u(x) - \hat{u}_n(x))^2 P(dx) \leq \frac{2A^2}{\left(1 - e^{-A^2/\left(2\overline{\sigma}^2\right)}\right)} d_H^2(p_{u,\sigma}, p_{\hat{u}_n,\hat{\sigma}}).$$

From all above, we have

$$R\left((u,\sigma); n; \delta^*\right) \leq$$

$$\frac{8A^2}{1 - e^{-A^2/2\overline{\sigma}^2}} \left(\inf_k \left\{ \frac{1}{N} \log \frac{1}{\omega_k} + C_{\overline{\sigma}} \| \sigma^2 - \sigma_k^2 \|^2 \right\} + \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{\overline{\sigma}^2}{2N} \sum_{l=i_0}^{n} R((u,\sigma); l; \delta_j) \right\} \right). \quad (5)$$

This completes the proof of Theorem 1.

REMARK: From Lemma 1, we have

$$d_H^2(p_{u,\sigma}, p_{\hat{u}_n,\hat{\sigma}}) \geq 2\left(1 - \sqrt{\frac{2\sigma\hat{\sigma}}{\hat{\sigma}^2 + \sigma^2}}\right) \geq 2\underline{C}(\sigma - \hat{\sigma})^2,$$

where $\underline{C}$ depends only on $\overline{\sigma}$. As a consequence, we have that $E\|\sigma - \hat{\sigma}\|^2$ is upper bounded similarly as in (5).

PROOF OF THEOREM 2: For each class $\mathcal{U}$, let δ_j be a minimax-rate optimal procedure producing estimators $\hat{u}_{j,i}$, $i \geq 0$. That is, there exists a constant C_j such that

$$R\left((u,\sigma); l; \delta_j\right) \leq C_j \cdot R(\mathcal{U}_j; l)$$

for $u \in \mathcal{U}_j, \sigma \le \overline{\sigma}$ and all $l \ge 0$. (We may take C_j to be any number bigger than 1 independent of j, but it is not necessary for the result.) Now combining the procedures as for Case 2 in Section 3.2 with $N_n = \lfloor n/2 \rfloor$, we have that for each j^0,

$$\begin{aligned}
\max_{u \in \mathcal{U}_{j^0}, \sigma \le \overline{\sigma}} R\left((u,\sigma); n; \delta^*\right) &\le C\left(\frac{1}{n}\log\frac{1}{\pi_{j^0}} + \frac{\log n}{n} + \max_{u \in \mathcal{U}_{j^0}, \sigma \le \overline{\sigma}}\left(\frac{1}{N}\sum_{l=n-N+1}^{n} R((u,\sigma); l; \delta_{j^0})\right)\right) \\
&\le C\left(\frac{1}{n}\log\frac{1}{\pi_{j^0}} + \frac{\log n}{n} + \frac{1}{N}\sum_{l=n-N+1}^{n}\left(\max_{u \in \mathcal{U}_{j^0}, \sigma \le \overline{\sigma}} R((u,\sigma); l; \delta_{j^0})\right)\right) \\
&\le C\left(\frac{1}{n}\log\frac{1}{\pi_{j^0}} + \frac{\log n}{n} + \frac{1}{N}\sum_{l=n-N+1}^{n} C_{j^0} R(\mathcal{U}_{j^0}; l)\right) \\
&\le C\left(\frac{1}{n}\log\frac{1}{\pi_{j^0}} + \frac{\log n}{n} + C_{j^0} R(\mathcal{U}_{j^0}; n-N+1)\right) \\
&\le C\left(\frac{1}{n}\log\frac{1}{\pi_{j^0}} + \frac{\log n}{n} + \widetilde{C}_{j^0} R(\mathcal{U}_{j^0}; n)\right),
\end{aligned}$$

where C is a constant depending only on A and $\overline{\sigma}$, for the 4th inequality, we use the monotonicity of $R(\mathcal{U}; l)$ in l, and for the last inequality, we use the assumption that each class has a rate-regular risk. When $R(\mathcal{U}_{j^0}; n)$ is at a nonparametric rate, the two penalty terms are negligible. This completes the proof of Theorem 2.

Proof of Theorem 3: We first assume that $\alpha \in [a, b]$ with $0 < a < b < \infty$. Consider a discretization of α at accuracy of order b_n. Let Υ_n be the set of the discretized values. A uniform weight on Υ_n gives π_j of order $\log(1/b_n)$. For each $\alpha \in \Upsilon_n$, let δ_α be a minimax-rate optimal procedure with

$$\max_{u \in \mathcal{U}_\alpha, \sigma \le \overline{\sigma}} R\left((u,\sigma); n; \delta_\alpha\right) \le 2R(\mathcal{U}_\alpha; n)$$

for all $n \ge 1$. Combining the procedures $\delta_\alpha, \alpha \in \Upsilon_n$ using the adaptation recipe with $N_n \sim n/2$ and σ^2 discretized as in Case 2 in Section 3.2, we have an adaptive procedure δ^* with

$$R\left((u,\sigma); n; \delta^*\right) \le C_{A,\overline{\sigma}} \inf_{\alpha \in \Upsilon_n}\left\{\frac{\log\left(1/b_n\right)}{n} + \frac{\log n}{n} + \frac{1}{n}\sum_{l=\lfloor n/2 \rfloor}^{n} R((u,\sigma); l; \delta_\alpha)\right\}.$$

As a consequence, for each $\beta \in \Upsilon_n$, we have

$$\begin{aligned}
\max_{u \in \mathcal{U}_\beta, \sigma \le \overline{\sigma}} R\left((u,\sigma); n; \delta^*\right) &\le C_{A,\overline{\sigma}} \inf_{\alpha \in \Upsilon_n}\left\{\frac{\log\left(1/b_n\right)}{n} + \frac{\log n}{n} + \frac{1}{n}\sum_{l=\lfloor n/2 \rfloor}^{n} \max_{u \in \mathcal{U}_\beta, \sigma \le \overline{\sigma}} R((u,\sigma); l; \delta_\alpha)\right\} \\
&\le C_{A,\overline{\sigma}}\left(\frac{\log\left(1/b_n\right)}{n} + \frac{\log n}{n} + \frac{1}{n}\sum_{l=\lfloor n/2 \rfloor}^{n} 2R(\mathcal{U}_\beta; l)\right) \\
&\le C'_{A,\overline{\sigma}}\left(\frac{\log\left(1/b_n\right)}{n} + \frac{\log n}{n} + R(\mathcal{U}_\beta; \lfloor n/2 \rfloor)\right),
\end{aligned}$$

where for the last inequality, we use the nonincreasing property of the minimax risk in sample size. Now for any $\alpha_0 \in [a,b]$, there exists a $\beta_n \in \Upsilon_n$ with $\beta_n \uparrow \alpha_0$ at order b_n. Then using the assumptions on the relationship between the function classes and that on the minimax risks, we have

$$\begin{aligned}
\max_{u\in\mathcal{U}_{\alpha_0},\sigma\le\overline{\sigma}} R\left((u,\sigma);n;\delta^*\right) &\le \max_{u\in\mathcal{U}_{\beta_n},\sigma\le\overline{\sigma}} R\left((u,\sigma);n;\delta^*\right) \\
&\le C'_{A,\overline{\sigma}}\left(\frac{\log\left(1/b_n\right)}{n}+\frac{\log n}{n}+R(\mathcal{U}_{\beta_n};\lfloor n/2\rfloor)\right) \\
&\le C'_{A,\overline{\sigma}}\left(\frac{\log\left(1/b_n\right)}{n}+\frac{\log n}{n}+M(\lfloor n/2\rfloor;\beta_n)\right) \\
&\le C'_{A,\overline{\sigma},\alpha_0}\left(\frac{\log\left(1/b_n\right)}{n}+\frac{\log n}{n}+M(n;\alpha_0)\right),
\end{aligned}$$

where for the last inequality, the constant may depend on α_0 since the assumption on the minimax risks does not require uniformity. From above, δ^* is adaptive over the classes $\{\mathcal{U}_\alpha : \alpha \in [a,b]\}$. Now without knowing bounds on the true α, we can consider a sequence of compact intervals, e.g., $[a_k,b_k]$, $k \ge 1$ with $b_k \to \infty$ and $a_k \to 0$. For each k, construct an adaptive procedure as above and then combine these procedures. The final procedure has the desired adaptive property without requiring knowledge of bounds on α. This completes the proof of Theorem 3.

Proof of Theorem 4: For the modified adaptive procedure with an independent estimator $\hat{\sigma}$, again by triangle inequality together with the (new) definition of $\widetilde{u}$, we have

$$\begin{aligned}
d_H\left(\phi_{u(x),\sigma(x)},\phi_{\widetilde{u}(x),\hat{\sigma}(x)}\right) &\le d_H\left(\phi_{u(x),\sigma(x)},\hat{g}_n(\cdot\mid x)\right)+d_H\left(\phi_{\widetilde{u}(x),\hat{\sigma}(x)},\hat{g}_n(\cdot\mid x)\right) \\
&\le d_H\left(\phi_{u(x),\sigma(x)},\hat{g}_n(\cdot\mid x)\right)+d_H\left(\phi_{u(x),\hat{\sigma}(x)},\hat{g}_n(\cdot\mid x)\right) \\
&\le d_H\left(\phi_{u(x),\sigma(x)},\hat{g}_n(\cdot\mid x)\right)+d_H\left(\phi_{u(x),\sigma(x)},\phi_{u(x),\hat{\sigma}(x)}\right)+d_H\left(\phi_{u(x),\sigma(x)},\hat{g}_n(\cdot\mid x)\right) \\
&\le 2d_H\left(\phi_{u(x),\sigma(x)},\hat{g}_n(\cdot\mid x)\right)+d_H\left(\phi_{u(x),\sigma(x)},\phi_{u(x),\hat{\sigma}(x)}\right).
\end{aligned}$$

By Lemma 1, we have

$$d_H^2\left(\phi_{u(x),\sigma(x)},\phi_{u(x),\hat{\sigma}(x)}\right)=2\left(1-\sqrt{\frac{2\sigma(x)\hat{\sigma}(x)}{\sigma^2(x)+\hat{\sigma}^2(x)}}\right)\le\frac{2(\sigma(x)-\hat{\sigma}(x))^2}{\sigma^2(x)+\hat{\sigma}^2(x)}.$$

As a consequence,

$$\begin{aligned}
d_H^2(p_{u,\sigma},p_{\widetilde{u},\hat{\sigma}}) &= \int d_H^2\left(\phi_{u(x),\sigma(x)},\phi_{\widetilde{u}(x),\hat{\sigma}}(x)\right)P(dx) \\
&\le 8\int d_H^2\left(\phi_{u(x),\sigma(x)},\hat{g}_n(\cdot\mid x)\right)P(dx)+4E\left(\frac{(\sigma(X)-\hat{\sigma}(X))^2}{\sigma^2(X)+\hat{\sigma}^2(X)}\right).
\end{aligned}$$

From Lemma 1 again, we have

$$d_H^2\left(\phi_{u(x),\sigma(x)},\phi_{\widetilde{u}(x),\hat{\sigma}(x)}\right)\ge 2\left(1-e^{-(u(x)-\widetilde{u}(x))^2/\left(4\sigma^2(x)+4\hat{\sigma}^2(x)\right)}\right).$$

Then similarly as before in the proof of Theorem 1, we have

$$E\int (u(x)-\widetilde{u}(x))^2 P(dx) \le \frac{2A^2}{1-Ee^{-A^2/\left(\overline{\sigma}^2+\overline{s}^2\right)}} E d_H^2(p_{u,\sigma}, p_{\widetilde{u},\hat{\sigma}}).$$

Note that $Ed_H^2(p_{u,\sigma},\hat{g}_n) \le ED(p_{u,\sigma}\|\hat{g}_n)$, which can be bounded similarly as in the proof of Theorem 1, namely,

$$ED(p_{u,\sigma}\|\hat{g}_n) \le E\left(\left(1+\log\left(1+\frac{\overline{s}^2}{\underline{\sigma}^2}\right)\right)\|\frac{\sigma^2-\hat{\sigma}^2}{\hat{\sigma}^2}\|^2/2\right)+\inf_j\left\{\frac{1}{N}\log\frac{1}{\pi_j}+E\left(\frac{1}{\underline{s}^2}\right)\cdot\frac{1}{2N}\sum_{l=i_0}^{n} R((u,\sigma);l;\delta_j)\right\}.$$

In applying Lemma 1 above, we use an inequality $((x-1)-\log x)/(x-1)^2 \le 1+\log(1+1/x)$ for $x>0$, which can be easily verified. Altogether, we have

$$R((u,\sigma);n;\delta^*) \le \frac{8A^2}{1-Ee^{-A^2/\left(\overline{\sigma}^2+\overline{s}^2\right)}}\cdot\left(E\left(1+\log\left(1+\frac{\overline{s}^2}{\underline{\sigma}^2}\right)\right)\|\frac{\sigma^2-\hat{\sigma}^2}{\hat{\sigma}^2}\|^2\right.$$

$$\left.+\,E\left(\frac{(\sigma(X)-\hat{\sigma}(X))^2}{\sigma^2(X)+\hat{\sigma}^2(X)}\right)+\inf_j\left\{\frac{2}{N}\log\frac{1}{\pi_j}+E\left(\frac{1}{\underline{s}^2}\right)\cdot\frac{1}{N}\sum_{l=n-N+1}^{n} R((u,\sigma);l;\delta_j)\right\}\right).$$

Observing that

$$\frac{(\sigma-\hat{\sigma})^2}{\sigma^2+\hat{\sigma}^2} \le \left(\frac{\sigma^2-\hat{\sigma}^2}{\hat{\sigma}^2}\right)^2,$$

the conclusion then follows. This completes the proof of Theorem 4.

Proof of Theorem 5: The proof follows similarly as those for Theorems 1 and 4.

Let $p_{a,\sigma}$ denote the normal density with mean a and variance σ^2.

Lemma 1: The K-L divergence and Hellinger distance between two normal densities satisfy

$$D(p_{a_1,\sigma_1}\|p_{a_2,\sigma_2}) = \frac{\sigma_1^2-\sigma_2^2}{2\sigma_2^2}-\frac{1}{2}\log\left(1+\frac{\sigma_1^2-\sigma_2^2}{\sigma_2^2}\right)+\frac{(a_1-a_2)^2}{2\sigma_2^2}$$

$$d_H^2(p_{a_1,\sigma_1},p_{a_2,\sigma_2}) = 2\left(1-\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}e^{\frac{-(a_1-a_2)^2}{4(\sigma_1^2+\sigma_2^2)}}\right) \ge 2\left(1-e^{\frac{-(a_1-a_2)^2}{4(\sigma_1^2+\sigma_2^2)}}\right).$$

If $\sigma_1^2/\sigma_2^2 \ge \gamma > 0$, then

$$D(p_{a_1,\sigma_1}\|p_{a_2,\sigma_2}) \le c_\gamma\left(\frac{\sigma_1^2-\sigma_2^2}{\sigma_2^2}\right)^2+\frac{(a_1-a_2)^2}{2\sigma_2^2},$$

where $c_\gamma = ((\gamma-1)-\log\gamma)/\left(2(\gamma-1)^2\right)$.

Proof of Lemma 1: The calculations are straightforward. For the second inequality, we use the fact that for $t>(-1)$, $(t-\log(1+t))/t^2$ is decreasing in t.

Acknowledgment. The anonymous reviewers are thanked for their very helpful comments.

References

[1] A.R. Barron, "Are Bayes rules consistent in information?" *Open Problems in Communication and Computation*, 85-91. T. M. Cover and B. Gopinath editors, Springer-Verlag, 1987.

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

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

[4] A.R. Barron and R.L. Barron, "Statistical learning networks: a unifying view," in *Computer Science and Statistics: Proceeding of the 21st Interface*, 1988.

[5] A.R. Barron, L. Birgé and P. Massart, "Risk bounds for model selection via penalization," *Probability Theory and Related Fields,* **113**, 301-413, 1999.

[6] L. Breiman, J.H. Friedman, R. Olshen, and C.J. Stone, *Classification and Regression Trees,* Belmont, CA: Wadsworth, 1984.

[7] L.D. Brown and M.G. Low, "A constrained risk inequality with applications to nonparametric functional estimation", *Ann. Statistics.*, **24**, 2524-2535, 1996.

[8] A. Buja, T. Hastie and R. Tibshirani, "Linear smoothing and additive models," *Ann. Statistics.*, **17**, 453-555, 1989.

[9] O. Catoni, "The mixture approach to universal model selection," Technical Report LIENS-97-22, Ecole Normale Superieure, Paris, France, 1997.

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

[11] B. Delyon and A. Juditsky, "Wavelet estimators, global error measures revisited," Technical Report, IRISA, 1994.

[12] R.A. DeVore and G.G. Lorentz, *Constructive Approximation*, Springer-Verlag, New York, 1993.

[13] L.P. Devroye and G. Lugosi, "Nonparametric universal smoothing factors, kernel complexity, and Yatracos classes," *Ann. Statist.,* **25**, 2626-2637, 1997.

[14] L.P. Devroye and T.J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Ann. Statist.,* **8**, 231-239, 1980.

[15] D.L. Donoho, "CART and best-ortho-basis: a connection," *Ann. Statistics.*, **25**, 1870-1911, 1997.

[16] D.L. Donoho, I.M. Johnstone, "Minimax estimation via wavelet shrinkage," *Ann. Statistics.*, **26**, 879-921, 1998.

[17] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian and D. Picard, "Wavelet shrinkage: asymptopia?" (with discussion), *J. R. Statist. Soc. B*, **57**, 301-369, 1995.

[18] N. Duan and K.-C. Li, "Slicing regression, a link-free regression method," *Ann. Statist.,* **19**, 505-530, 1991.

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

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

[21] J. Fan and I. Gijbels, *Local Polynomial Modeling and its Applications*, Chapman and Hall, London, 1996.

[22] J. Friedman, "Multivariate adaptive regression splines" (with discussion), *Ann. Statist.,* **19**, 1-67, 1991.

[23] J. Friedman and W. Stuetzel, "Projection pursuit regression," *J. Amer. Statist. Assoc.*, **76**, 817-823, 1981.

[24] A. Goldenshluger and A. Nemirovski, "Adaptive de-noising of signals satisfying differential inequalities," *IEEE Transaction on Information Theory*, **43**, 872-889, 1997.

[25] G.K. Golubev and M. Nussbaum, "Adaptive spline estimates for nonparametric regression models," *Theory Probab. Appli.* **37**, 521-529.

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

[27] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric estimation," *Publication Interne, IRISA*, N. 993, 1996.

[28] O.V. Lepski, "Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates," *Theory probab. Appl.* **36**, 682-697, 1991.

[29] O.V. Lepski, E. Mammen, and V.G. Spokoiny, "Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection," *Ann. Statist.*, **25**, 929-947, 1997.

[30] G. Lugosi and A. Nobel, "Adaptive model selection using empirical complexities," unpublished manuscript, 1996.

[31] E. Mammen and S. van de Geer , "Locally adaptive regression splines," *Ann. Statistics*, **25**, 387-413, 1997.

[32] T. Nicoleris and Y.G. Yatracos, "Rate of convergence of estimates, Kolmogrov's entropy and the dimensionality reduction principle in regression," *Ann. Statistics*, **25**, 2493-2511, 1997.

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

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

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

[36] H. Triebel, *Theory of Function Spaces II*, Birkhauser, Basel and Boston, 1992.

[37] A.B. Tsybakov, "Pointwise and sup-norm adaptive signal estimation on the Sobolev classes," preprint, 1995.

[38] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.

[39] M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman and Hall, London, 1995.

[40] Y. Yang, *Minimax Optimal Density Estimation*, Ph.D. Dissertation, Department of Statistics, Yale University, May, 1996.

[41] Y. Yang, "On adaptive function estimation," Technical Report #30, Department of Statistics, Iowa State University, 1997.

[42] Y. Yang, "Model selection for nonparametric regression," *Statistica Sinica*, **9**, 475-499, 1999a.

[43] Y. Yang, "Mixing strategies for density estimation," accepted by *Ann. Statististics*, 1999b.

[44] Y. Yang, "Regression with multiple models: selecting or mixing?" Technical Report #8, Department of Statistics, Iowa State University, 1999c.

[45] Y. Yang and A.R. Barron, "An asymptotic property of model selection criteria," *IEEE Trans. on Information Theory*, **44**, 95-116, 1998.

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