The Annals of Statistics
2000, Vol. 28, No. 1, 75–87

MIXING STRATEGIES FOR DENSITY ESTIMATION

BY YUHONG YANG

Iowa State University

General results on adaptive density estimation are obtained with respect to any countable collection of estimation strategies under Kullback-Leibler and squared L_2 losses. It is shown that without knowing which strategy works best for the underlying density, a single strategy can be constructed by mixing the proposed ones to be adaptive in terms of statistical risks. A consequence is that under some mild conditions, an asymptotically minimax-rate adaptive estimator exists for a given countable collection of density classes; that is, a single estimator can be constructed to be simultaneously minimax-rate optimal for all the function classes being considered. A demonstration is given for high-dimensional density estimation on $[0, 1]^d$ where the constructed estimator adapts to smoothness and interaction-order over some piecewise Besov classes and is consistent for all the densities with finite entropy.

1. Introduction. In recent years, there has been an increasing interest in adaptive function estimation. The main objective, if possible, is to construct a single estimator so that it is automatically asymptotically optimal in terms of a minimax risk for each function class in a given collection. Adaptive function estimators were constructed, for example, by Efroimovich and Pinsker (1984) and Efroimovich (1985) for ellipsoidal classes; by Härdle and Marron (1985) using adaptive kernel estimators for some Lipschitz classes; and by, for example, Donoho, Johnstone, Kerkyacharian and Picard (1996) and others using wavelet analysis for Besov classes. General schemes have also been proposed for the construction of adaptive estimators. Barron and Cover (1991) derived general adaptation risk bounds for density estimation based on minimum description length (MDL) criterion. These bounds were used to demonstrate adaptive properties of MDL criterion including adaptation over classical function classes using metric entropies. Lepskii (1991) gave some sufficient conditions to ensure existence of minimax-rate adaptive estimators and constructed adaptive estimators specifically for ellipsoidal classes under L_p loss for $2 < p \leq \infty$. In addition to the use of MDL criterion, other adaptation schemes by model selection have been developed, including very general penalized contrast criteria in Birgé and Massart (1996), and Barron, Birgé and Massart (1999) with a variety of interesting applications; penalized maximum likelihood criteria in Yang and Barron (1998) and complexity penalized criteria based on V-C theory; see, for example, Devroye, Györfi and Lugosi

Received June 1997; revised September 1999.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62B10, 62C20, 94A29.

Key words and phrases. Density estimation, rates of convergence, adaptation with respect to estimation strategies, minimax adaptation.

[(1996), Chapter 18] and Lugosi and Nobel (1999). Functional aggregation of estimators to adapt to within order $n^{-1/4}$ in L_2 risk is proposed in Juditsky and Nemirovski (1996).

Our interest in this work concerns adaptivity in a more general sense. The questions we plan to address is, given a countable collection of estimation strategies (regardless of how they have been obtained), is it possible to find an adaptive strategy so that it automatically performs as well as the best one in the list in an asymptotic sense? Such a strategy will be said to be adaptive with respect to the collection of the original ones. In the related context of estimating a functional, negative results have been obtained showing that optimal rate adaptation may not be possible [see Lepskii (1991) and Brown and Low (1996)]. Here we give positive results for global density estimation. Differently from the previous work on adaptation, no specific properties will be required here on the collection of strategies. Thus advantages of a list of possibly completely different strategies can be combined in terms of statistical risks, and, if desired, adaptive strategies constructed using various schemes available (e.g, automated kernel smoothing, wavelet procedures, smoothing splines, neural net estimation, etc.) can also be included in the list for even more adaptivity. The benefit of considering such a list of very different procedures could be substantial, especially for high-dimensional density estimation, where to overcome the curse of dimensionality, searching over different characterizations of functions is desired for better accuracy (see Section 4 for a demonstration).

Estimation strategies are often derived for specific function classes. For a collection of such strategies which are constructed to be minimax optimal for the corresponding target classes, adaptation with respect to the strategies as explained above implies minimax adaptation with respect to the target classes. In this sense, the notion of adaptation with respect to a collection of strategies is more general than minimax adaptation with respect to a collection of density classes. Results on minimax adaptation will be given as consequences of the main results on combining strategies.

In the revision of an earlier version of this paper, an editor and an associate editor brought to our attention an independent research of Catoni (1997) completed after our submission of this work. A result similar in spirit to our Theorem 1 under K-L loss was given.

Density estimation is closely related to universal coding as illustrated in Barron (1987), Clark and Barron (1990), Barron and Cover (1991), Yang (1996) (a formal statement is given as Lemma 2.6) and Haussler and Opper (1997). This relationship, discussed in Barron (1987) and Barron and Cover (1991), will be used for our construction of an adaptive strategy. For adaptation under the squared L_2 loss, some results used in our analysis come from Yang and Barron (1999), which derives minimax rate of convergence for a fixed general function class. Some recent results on universal coding are redundancy bounds for Bayes hierarchical coding in Feder and Merhav (1996) and redundancy bounds for individual sequences using a sequential procedure for binary tree sources in Willems, Shtarkov and Tjalkens (1995).

Finally, it is worth mentioning that Bayesian model averaging methods have also been proposed to combine various models [see, e.g., Kass and Raftery (1995) and Berger and Pericchi (1996)]. Our method permits but does not require the estimators in the models to be obtained in a Bayesian framework, which sometimes has difficulties in the choice of insensitive priors on the parameters. In addition, our adaptation recipe works for combining estimation strategies even when some or all of them are not model-based procedures.

1.1. *Some notation.* Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations with density $f(x)$, $x \in \mathscr{X}$ with respect to a σ -finite measure μ. Here the space $\mathscr{X}$ is general and could be any dimensional. The goal is to estimate the unknown density f based on the data.

Let $\| g_1 - g_2 \|_2 = \left(\int |g_1 - g_2|^2 d\mu\right)^{1/2}$ be the L_2 distance between functions g_1 and g_2 with respect to μ. The Kullback–Leibler (K-L) divergence between two densities f and g is defined as $D(f \| g) = \int f \log(f/g)\, d\mu$. Both $D(f \| \hat{f})$ and $\| f - \hat{f} \|_2^2$ will be considered as loss functions.

In this paper, a density estimation strategy δ refers to an estimation procedure producing density estimators $f_{\delta,0}$, $\hat{f}_{\delta,1}(x; X_1), \ldots, \hat{f}_{\delta,n-1}(x; X_1, \ldots, X_{n-1}), \ldots$ based on observation(s) $X^0, X^1, \ldots, X^{n-1}, \ldots$, respectively [here $f_{\delta,0}$ is an initial guess based on no data (X^0) and $X^i = (X_1, \ldots, X_i)$ for $i \geq 1$]. Let

$$R_{\text{seq}}(f; n; \delta) = \frac{1}{n+1} \sum_{i=0}^{n} ED(f \| \hat{f}_{\delta,i})$$

denote the average cumulative risk for estimating f using strategy δ up to n observations. This notion of risk (sometimes called redundancy or regret) is considered by many others in the context of data compression, prediction, gambling and computational learning theory [see, e.g., Clarke and Barron (1990) and Barron and Xie (1996) for asymptotics on finite-dimensional models, Yang and Barron (1999), Section 3, and Haussler and Opper (1997) for rates of convergence over a given density class]. It is a reasonable and stable discrepancy measure to evaluate different strategies. The individual risk $ED(f \| \hat{f}_{\delta,n})$ at sample size n denoted by $R(f; n; \delta)$ will also be considered. Similarly define $r_{\text{seq}}(f; n; \delta)$ and $r(f; n; \delta)$ for the squared L_2 loss.

A minimax risk measures difficulty in estimation in a uniform sense. Let l be a chosen loss function; then for a density estimator $\hat{f}$, the risk is $El(f, \hat{f})$. Let $\mathscr{F}$ be a class of densities. Then the minimax risk for estimating a density in $\mathscr{F}$ at sample size n is defined as

$$R(\mathscr{F}; l; n) = \min_{\hat{f}} \max_{f \in \mathscr{F}} El(f, \hat{f}),$$

where the minimization is over all density estimators.

The symbol "$\asymp$" will be used to mean the same order, that is, $a_n \asymp b_n$ if a_n/b_n is bounded above and away from zero.

The paper is organized as follows. In Section 2, we present results on adaptation with respect to estimation strategies; in Section 3, minimax adaptation

results for function classes are given. A demonstration of the results is provided in Section 4. A generalization for predictation for dependent data is given in Section 5. The proofs of the results are given in Section 6.

2. Adaptation with respect to estimation strategies. Let $\{\delta_j, j \geq 1\}$ be any collection of density estimation strategies. Here the index set $\{j \geq 1\}$ is allowed to degenerate to a finite set. As mentioned earlier, there is no restriction at all on the choice of the strategies and they could be proposed for different purposes, classes, and/or under different assumptions. Some of them could be based only on heuristics but with practical significance. Strategy δ_j produces density estimators $f_{j,0}, \hat{f}_{j,1}(x; X_1), \ldots$ based on observation(s) $X^0, X^1, \ldots$ respectively.

2.1. *Adaptation under K-L risk.* The following is a simple yet powerful recipe to get an adaptive strategy by mixing $\{\delta_j, j \geq 1\}$ given in Yang (1996). Let $\underline{\pi} = \{\pi_j, j \geq 1\}$ be a set of positive numbers satisfying $\sum_{j\geq 1} \pi_j = 1$. They may be viewed as weights or prior probabilities of the strategies.

Let

$$\begin{aligned} q_0(x) &= \sum_{j\geq 1} \pi_j f_{j,0}(x), \\ q_1(x; x_1) &= \frac{\sum_{j\geq 1} \pi_j f_{j,0}(x_1)\hat{f}_{j,1}(x; x_1)}{\sum_{j\geq 1} \pi_j f_{j,0}(x_1)}, \\ q_2(x; x_1, x_2) &= \frac{\sum_{j\geq 1} \pi_j f_{j,0}(x_1)\hat{f}_{j,1}(x_2; x_1)\hat{f}_{j,2}(x; x_1, x_2)}{\sum_{j\geq 1} \pi_j f_{j,0}(x_1)\hat{f}_{j,1}(x_2; x_1)} \\ &\cdots \end{aligned}$$

$$\begin{aligned} &q_{n-1}(x; x_1, x_2, \ldots, x_{n-1}) \\ &\quad = \frac{\begin{array}{r}\sum_{j\geq 1} \pi_j f_{j,0}(x_1)\hat{f}_{j,1}(x_2; x_1)\cdots\hat{f}_{j,n-2}(x_{n-1}; x_1, x_2, \ldots, x_{n-2}) \\ \times\hat{f}_{j,n-1}(x; x_1, x_2, \ldots, x_{n-1})\end{array}}{\sum_{j\geq 1} \pi_j f_{j,0}(x_1)\hat{f}_{j,1}(x_2; x_1)\cdots\hat{f}_{j,n-2}(x_{n-1}; x_1, x_2, \ldots, x_{n-2})} \\ &\quad \ldots. \end{aligned}$$

Define estimators for $i \geq 0$ based on $X_1, \ldots, X_i$ as follows:

$$\hat{f}_{\text{seq},i}(x) = q_i(x; X_1, \ldots, X_i). \tag{1}$$

They are valid probability density estimators at each sample size. Call this estimation strategy δ^*_{seq}. This strategy will be shown to be adaptive in terms of the average cumulative risk. For adaptation under the individual risk, let

$$\hat{f}_n(x) = \frac{1}{n+1}\sum_{i=0}^{n} q_i(x; X^i).$$

It is a valid density estimator of f based on X^n. The strategy producing $\hat{f}_n$, $n \geq 1$ will be called a combined strategy denoted by δ^*.

Consider

$$\inf_{j\geq 1}\left(\frac{1}{n+1}\log\frac{1}{\pi_j} + R_{\text{seq}}(f;n;\delta_j)\right).$$

It is the best trade-off between the average cumulative risk and the logarithm of the inverse weight (or prior probability) relative to the sample size over all the estimation strategies.

THEOREM 1. *For any given countable collection of estimation strategies* $\{\delta_j, j \geq 1\}$ *and* $\underline{\pi}$, *we can construct a single estimation strategy* δ^*_{seq} *as given in the above recipe such that for any underlying density* f,

$$R_{\text{seq}}(f;n;\delta^*_{\text{seq}}) \leq \inf_{j\geq 1}\left(\frac{1}{n+1}\log\frac{1}{\pi_j} + R_{\text{seq}}(f;n;\delta_j)\right). \tag{2}$$

The combined strategy δ^**has individual risk bounded by the same quantity*

$$R(f;n;\delta^*) \leq \inf_{j\geq 1}\left(\frac{1}{n+1}\log\frac{1}{\pi_j} + R_{\text{seq}}(f;n;\delta_j)\right). \tag{3}$$

REMARKS. (i) A similar adaptation bound is given in Yang (1997) for nonparametric regression under Gaussian errors with known variance using a connection between estimating the regression function and the joint density of the observation. Adaptation risk bounds for regression by model selection are in Barron, Birgé and Massart (1999) and Yang (1999).

(ii) An individual risk bound is given in Catoni (1997) for a similarly defined strategy. His formulation has a computational advantage and can avoid an extra logarithmic term in the risk bound for parametric estimation.

From (2), up to an additive penalty of order $1/n$, the adaptive strategy δ^*_{seq} performs as well as any strategy in the list in terms of the average cumulative risk. For a strategy δ with regularly decreasing risk converging essentially more slowly than the parametric case, $R_{\text{seq}}(f;n;\delta)$ and $R(f;n;\delta)$ are of the same order (see Section 3). When such strategies are combined, (3) ensures adaptation in terms of individual risk.

For applications, we may assign smaller weights (or prior probabilities) π_j for more complex estimation strategies. Then the risk bounds in the theorem are trade-offs between accuracy and complexity. For a complex strategy (with a small weight), its role in the risk bound becomes significant only when the sample size becomes large.

A strategy is said to be *consistent* for f under loss l, if $El(f, \hat{f}_{\delta,n}) \to 0$ as $n \to \infty$. A simple consequence of Theorem 1 is that if any of the strategy in the list is consistent for the unknown density, so is the combined adaptive strategy δ^*.

2.2. *Adaptation under L_2 loss.* For the adaptation results under the squared L_2 loss, unlike the bounds in Theorem 1, some mild technical conditions will be used (which arise from relating the K-L and L_2 distances). Throughout the paper, for squared L_2 adaptation, we assume that the dominating measure μ is finite and is normalized to be a probability measure, and the unknown density is uniformly upper bounded, that is, $\| f \|_\infty \le A < \infty$ for a known constant A.

For each f, let $g = (f+1)/2$ be a mixture of f and the uniform density 1. We have the following conclusion.

THEOREM 2. *For any given countable collection of strategies $\{\delta_j, j \ge 1\}$, we can construct a strategy Δ^* such that*

$$r(f; n; \Delta^*) \le C \inf_{j \ge 1} \left(\frac{1}{n+1} \log \frac{1}{\pi_j} + \frac{1}{n+1} \sum_{i=0}^{n} r(g; i; \delta_j) \right), \tag{4}$$

where the constant C depends only on A.

Note that the risk of the combined strategy at an unknown density f is bounded in terms of the risks of the original strategies at $g = (f+1)/2$ instead of f itself. For usual nonparametric procedures, the risks at f and g are most likely to be bounded at the same rate. Formally, this does not cause trouble for applications where minimax risks are considered for nonparametric classes including f and g at the same time as is the case for the classical convex classes. This technical difficulty is avoided if one is wiling to assume that the unknown density is bounded away from zero, for which case the K-L divergence and the squared L_2 distance are equivalent and thus $r(g; i; \delta_j)$ can be replaced by $r(f; i; \delta_j)$ directly in the theorem.

In light of Theorems 1 and 2, adaptive estimators can be obtained using the adaptation recipe for a countable collection of function classes as will be given in the next section. The results are also useful for combining estimation procedures with hyperparameters (e.g., bandwidth for a kernel estimator). Various conclusions can be made for a combined strategy with a suitable discretization of the hyperparameters.

3. Adaptation with respect to function classes. Let $\{\mathcal{F}_j, j \ge 1\}$ be a collection of density classes. Assume the true function is in one of the classes, that is, $f \in \cup_{j \ge 1} \mathcal{F}_j$. The question we want to address is without knowing which class contains f , can we have one estimator (not depending on j) such that it converges asymptotically at the minimax rate of the class containing f? If such an estimator exists, we call it a minimax-rate adaptive estimator with respect to the classes $\{\mathcal{F}_j, j \ge 1\}$. This concept of adaptation can be obviously extended to any given collection of classes not necessarily countable.

We need a regularity condition for our results. The familiar rates of convergence for function estimation are $n^{-\alpha} (\log n)^{\beta}$ for some $0 \le \alpha < 1$ and $\beta \in R$. When $0 < \alpha < 1$ [then $R(\mathcal{F}; l; n)$ converges essentially more slowly than the

parametric case], we have that $(1/n)\sum_{i=0}^{n} R(\mathscr{F}; l; i)$ is of the same order as $R(\mathscr{F}; l; n)$. For such a case, we say the class has *a regular nonparametric risk rate*.

THEOREM 3. *Let $\{\mathscr{F}_j,\ j \geq 1\}$ be any collection of density classes each with a regular nonparametric risk rate under the loss being considered.*

1. *There exists a combined minimax-rate adaptive strategy under the K-L loss.*
2. *Assume that $\{\mathscr{F}_j,\ j \geq 1\}$ is uniformly bounded with*

$$\sup_{j\geq 1} \sup_{f\in\mathscr{F}_j} \| f \|_\infty \leq A < \infty.$$

If, in addition, each $\mathscr{F}_j$ is convex including the uniform density 1 or the classes are uniformly bounded away from zero, then there exists a minimax-rate adaptive estimator over the classes under the squared L_2 loss.

4. An illustration. Consider estimating a density function on $[0, 1]^d$ with respect to Lebesgue measure μ. We focus on the K-L loss.

Densities with finite entropy. Let $\mathscr{H}$ consist of all densities that have finite entropy, that is, $\mathscr{H} = \{f : \int f \log f d\mu < \infty\}$. Note that densities in $\mathscr{H}$ are not necessarily bounded above or away from zero. The class is very large and no uniform rate of convergence is possible under K-L or L_2 loss.

Piecewise Besov classes with different interaction order and smoothness. To allow discontinuities, consider the following modification of a function class. Let $\mathscr{G}$ be a class of functions on $[0, 1]^d$ that are uniformly upper bounded and lower bounded away from zero. For an integer k, and positive constants γ_1 and γ_2, let $\mathscr{G}_{k,\gamma_1,\gamma_2} = \{h(x)\cdot \sum_{i=1}^{k} b_i 1_{B_i}/c : h \in \mathscr{G},\ B_i$'s are hypercubes partitioning $[0, 1]^d$ satisfying $\sum_{i=1}^{k} b_i \mu(B_i) \geq \gamma_1$ and $0 \leq b_i \leq \gamma_2, 1 \leq i \leq k\}$. Here c is the normalizing constant to make $\mathscr{G}_{k,\gamma_1,\gamma_2}$ a class of probability density functions. Note that with the constraints on b_i's and the volumes of B_i's, the densities in $\mathscr{G}_{k,\gamma_1,\gamma_2}$ are uniformly upper bounded. However, they are allowed to be 0 on some hypercubes and arbitrarily close to 0 on others. If the functions in $\mathscr{G}$ are smooth, then the densities in $\mathscr{G}_{k,\gamma_1,\gamma_2}$ are piecewise smooth. The modification provides somewhat more flexibility than the original class.

For $1 \leq \sigma, q \leq \infty$ and $\alpha > 0$, let $B_{q,\sigma}^{\alpha,r}$ be the Besov space consisting of all functions $g \in L_q[0, 1]^r$ such that the Besov norm satisfies $\| g \|_{B_{q,\sigma}^{\alpha,r}} < \infty$ [see, e.g., Triebel (1975) and DeVore and Lorentz (1993)]. Let $B_{q,\sigma}^{\alpha,r}(C)$ denote the subset of positive functions g in the Besov space with $\| g \|_{B_{q,\sigma}^{\alpha,r}} + \| \log g \|_\infty \leq C$ (without the extra boundness assumption here, we could not identify the minimax rate of convergence under the K-L loss). Define

$$\begin{aligned}
S_{q,\sigma}^{\alpha,1}(C) &= \{\textstyle\sum_{i=1}^{d} g_i(x_i) :\ g_i \in B_{q,\sigma}^{\alpha,1}(C),\ 1 \leq i \leq d\},\\
S_{q,\sigma}^{\alpha,2}(C) &= \{\textstyle\sum_{1\leq i<j\leq d} g_{i,j}(x_i, x_j) :\ g_{i,j} \in B_{q,\sigma}^{\alpha,2}(C),\ 1 \leq i < j \leq d\}\\
&\cdots\\
S_{q,\sigma}^{\alpha,d}(C) &= B_{q,\sigma}^{\alpha,d}(C).
\end{aligned}$$

The simplest function class $S^{\alpha,1}_{q,\sigma}(C)$ contains additive functions (no interaction) and the complexity of the classes increases when r increases. To allow discontinuity, let $S^{\alpha,r}_{q,\sigma}(C)_{k,\gamma_1,\gamma_2}$ be the modified class (piecewise Besov in some sense) from $S^{\alpha,r}_{q,\sigma}(C)$ as defined earlier. It is easy to show that the metric entropy of $S^{\alpha,r}_{q,\sigma}(C)$ under the L_2 distance and its covering entropy under the K-L divergence are of the same orders as $B^{\alpha,r}_{q,\sigma}(C)$. By Theorem 5 of Yang and Barron (1999), the minimax rate of convergence under the K-L (or squared L_2) loss for estimating a density in $S^{\alpha,r}_{q,\sigma}(C)_{k,\gamma_1,\gamma_2}$ is $n^{-2\alpha/(2\alpha+r)}$ for $1 \le r \le d$.

4.1. *Desired properties on estimation.* Suppose we have the following wish list for the adaptive estimator $\hat{f}_n$, $n \ge 1$ to be constructed:

1. $\hat{f}_n$ is consistent for all $f \in \mathscr{H}$.
2. $\hat{f}_n$ converges automatically at the optimal rate $n^{-2\alpha/(2\alpha+r)}$ if $f \in S^{\alpha,r}_{q,\sigma}(C)_{k,\gamma_1,\gamma_2}$ without knowing any of the hyperparameters.
3. $\hat{f}_n$ behaves well if a projection pursuit density estimator happens to converge reasonably fast.

The rationale behind the wish list is as follows. Besov classes with different choices of the hyperparameters provide considerable flexibility in modeling a density [see, e.g., Donoho, Johnstone, Kerkyacharian and Picard (1996)]. The piecewise modification allows discontinuity of the density. When α is small relative to d, the rate of convergence is rather slow (well known as the curse of dimensionality). The consideration of different interaction order can lead to a substantial improvement if f happens to be in $S^{\alpha,r}_{q,\sigma}(C)_{k,\gamma_1,\gamma_2}$ with r much smaller than d. Projection pursuit [see, e.g., Huber (1985)] is another approach to high-dimensional density estimation by dimension reduction. Despite the lack of theory on convergence rate property, such procedures have practical merits. Hence the third wish above. (Of course, one could go on with more target classes or add different strategies such as neural nets in the wish list, sacrificing simplicity and computation ease.) Finally, since the true density f may well not be in any of these classes, we want at least consistency for every $f \in \mathscr{H}$.

4.2. *Method of adaptation.* To use the adaptation recipe, it suffices to construct a consistent estimator for $\mathscr{H}$ and optimal-rate estimators for the classes $S^{\alpha,r}_{q,\sigma}(C)_{k,\gamma_1,\gamma_2}$ and then combine them and the projection pursuit estimator appropriately.

Barron (1988) [see also Barron, Györfi and van de Meulen (1992)] constructed a histogram estimator consistent for $\mathscr{H}$ under the K-L loss, that is, there is a strategy $\delta_{\mathscr{H}}$ such that $R(f;n;\delta_{\mathscr{H}}) \to 0$ for each $f \in \mathscr{H}$.

For adaptation among the piecewise Besov classes, we may first obtain adaptivity over the smoothness parameters $1 \le \sigma \le \infty$, $1 \le q \le \infty$, $\alpha > d/q$ for fixed r, C, k, γ_1 and γ_2. To that end, a suitable discretization of the smoothness parameters leaves us a countable collection of density classes to work with, for each of which a minimax-rate adaptive estimator can be constructed, for example, utilizing a covering set under K-L divergence as in Yang

and Barron (1999). Then the adaptation recipe together with an appropriate assignment of weights on the discretized values can result in adaptation for these classes following some "continuity" argument. Further adaptation is obtained by combining these estimators over integer values of r (between 1 and d), C, k, γ_1 and γ_2. This strategy, say δ_B, is minimax-rate adaptive over all the piecewise Besov classes.

Finally, we combine the above strategies δ_H, δ_B, and a chosen projection pursuit strategy (e.g., with equal weights). Then the overally combined strategy makes the three wishes come true.

A similar result holds for the squared L_2 risk applying Theorem 2, assuming the unknown density is bounded above by a known constant. Adaptation results over Besov classes by wavelets are in Donoho, Johnstone, Kerkyacharian and Picard (1996), Birgé and Massart (1996) and Juditsky (1997).

Adaptation over density classes with different characteristics is discussed in Barron and Cover (1991) using MDL criterion and later in Barron, Birgé and Massart (1999) and Yang and Barron (1998) by other model selection criteria.

5. A generalization for prediction for dependent data. A similar adaptation result holds for prediction with dependent data.

Let $X_1, X_2, \ldots, X_n$ be a stochastic process. One is interested in "estimating" or predicting the conditional density $f_i(x_{i+1}|x^i)$ of X_{i+1} given X^i. For an predictor $\hat{f}_i$, we measure the loss using K-L divergence $D(f_i \parallel \hat{f}_i) = \int f_i(x_{i+1}|x^i) \log \left(f_i(x_{i+1}|x^i)/\hat{f}_i(x_{i+1}) \right) \mu(dx_{i+1})$. The average cumulative prediction risk is

$$\frac{1}{n+1} \sum_{i=0}^{n} ED(f_i \parallel \hat{f}_i).$$

Let δ denote a prediction strategy which produces predictors $f_{\delta,0}$, $\hat{f}_{\delta,1}(x; X_1)$, $\ldots$, based on observation(s) $X^0, X^1, \ldots$, respectively. Let $R_{\text{pred}}(\{f_i,\ i \geq 1\}; n; \delta) = (1/(n+1)) \sum_{i=0}^{n} ED(f_i \parallel \hat{f}_{\delta,i})$ denote the average cumulative risk when the true conditional densities are f_i, $i \geq 1$.

Let $\{\delta_j,\ j \geq 1\}$ be a collection of density prediction strategies. We have the following conclusion on adaptive prediction.

PROPOSITION 1. *For any given countable collection of prediction strategies* $\{\delta_j,\ j \geq 1\}$ *and* $\underline{\pi}$*, we can construct a single adaptive prediction strategy* δ^* *such that*

$$R_{\text{pred}}(\{f_i,\ i \geq 1\}; n; \delta^*) \leq \inf_{j \geq 1} \left(\frac{1}{n+1} \log \frac{1}{\pi_j} + R_{\text{pred}}(\{f_i,\ i \geq 1\}; n; \delta_j) \right).$$

The proof of the proposition is similar to that of Theorem 1 and is omitted here.

6. Proof of the results.

PROOF OF THEOREM 1. Let $f^n(x^n)$ denote the product density $f(x_1)f(x_2)\cdots f(x_n)$. Let

$$q_j^{(n)} = f_{j,0}(x_1)\hat{f}_{j,1}(x_2;x_1)\cdots\hat{f}_{j,n-1}(x_n;x_1,\ldots x_{n-1}).$$

It is a joint density function on the product space of $X_1,\ldots,,X_n$. Then let $q^{(n)} = \sum_{j\geq 1}\pi_j q_j^{(n)}$ be a mixture from $q_j^{(n)}$'s. The cumulative risk of the constructed estimators satisfy

$$\begin{aligned}
&\sum_{i=0}^{n-1} E_f D(f \parallel \hat{f}_{\text{seq},i}) \\
&\quad= \sum_{i=0}^{n-1} E_f \int f(x)\log\frac{f(x)}{\hat{f}_{\text{seq},i}(x)}\mu(dx) \\
&\quad= \sum_{i=0}^{n-1} E_f \int f(x_{i+1})\log\frac{f(x_{i+1})}{\hat{f}_{\text{seq},i}(x_{i+1})}\mu(dx_{i+1}) \\
&\quad= \sum_{i=0}^{n-1} \int f(x_1)f(x_2)\cdots f(x_i)f(x_{i+1})\log\frac{f(x_{i+1})}{q_i(x_{i+1};x_1,\ldots,x_i)} \\
&\qquad\qquad \times\mu(dx_1)\mu(dx_2)\cdots\mu(dx_{i+1}) \\
&\quad= \sum_{i=0}^{n-1} \int f^n(x^n)\log\frac{f(x_{i+1})}{q_i(x_{i+1};x_1,\ldots,x_i)}\mu(dx_1)\mu(dx_2)\cdots\mu(dx_n) \\
&\quad= \int f^n(x^n)\left(\sum_{i=0}^{n-1}\log\frac{f(x_{i+1})}{q_i(x_{i+1};x_1,\ldots,x_i)}\right)\mu(dx_1)\mu(dx_2)\cdots\mu(dx_n) \\
&\quad= \int f^n(x^n)\left(\log\frac{f^n(x^n)}{q^{(n)}}\right)\mu(dx_1)\mu(dx_2)\cdots\mu(dx_n) \\
&\quad= D(f^n \parallel q^{(n)}).
\end{aligned}$$

Thus we have $nR_{\text{seq}}(f;n-1;\delta^*_{\text{seq}}) = D(f^n \parallel q^{(n)})$. To prove Theorem 1, our task is to bound $D(f^n \parallel q^{(n)})$. For any f and any $j \geq 1$, since $\log(x)$ is an increasing function, we have

$$\begin{aligned}
D(f^n \parallel q^{(n)}) &\leq \int f^n(x^n)\log\frac{f^n(x^n)}{\pi_j q_j^{(n)}(x^n)}d\mu(x^n) \\
&= \log\frac{1}{\pi_j} + \int f^n(x^n)\log\frac{f^n(x^n)}{q_j^{(n)}(x^n)}d\mu(x^n).
\end{aligned}$$

The term $\int f^n(x^n)\log\left(f^n(x^n)/q_j^{(n)}(x^n)\right)d\mu(x^n)$ can be bounded in terms of risks of the estimators produced by strategy δ_j. Indeed, as earlier,

$$\int f^n(x^n)\log\frac{f^n(x^n)}{q_j^{(n)}(x^n)}d\mu(x^n)=\sum_{i=0}^{n-1}E_f D(f\parallel\hat{f}_{j,i})=nR_{\text{seq}}(f;n-1;\delta_j).$$

Thus we have $R_{\text{seq}}(f;n-1;\delta^*_{\text{seq}})\le(1/n)\log\left(1/\pi_j\right)+R_{\text{seq}}(f;n-1;\delta_j)$. Since the inequality holds for all j, minimizing over j, we have the first inequality in Theorem 1. Let $\hat{f}_{n-1}=(1/n)\sum_{i=0}^{n-1}\hat{f}_{\text{seq},i}$, by convexity, as in Barron (1987), we have

$$E_f D(f\parallel\hat{f}_{n-1})\le\frac{1}{n}\sum_{i=0}^{n-1}E_f D(f\parallel\hat{f}_{\text{seq},i})\le\inf_{j\ge1}\left(\frac{1}{n}\log\frac{1}{\pi_j}+R_{\text{seq}}(f;n-1;\delta_j)\right).$$

This completes the proof of Theorem 1. □

PROOF OF THEOREM 2. We use an idea in Yang and Barron [(1999), Section 2] to change the problem to another one for which application of Theorem 1 gives bounds under square L_2 loss. In addition to the observed i.i.d. sample $X_1, X_2, \ldots, X_n$ from f, let $W_1, W_2, \ldots, W_n$ be an independent sample generated i.i.d. from the uniform distribution on $\mathscr{X}$ with respect to μ. Let $\widetilde{X}_i$ be X_i or W_i with probability $(1/2, 1/2)$ according to independent coin flips. Then $\widetilde{X}_i$ has density $g(x)=(f(x)+1)/2$. Clearly the new density g is bounded below (away from 0), whereas the family of the original densities need not be. Since the unknown density g is known to be bounded between $1/2$ and $1/2+A/2$, we can project (if necessary) the estimators produced by the original strategies into this range without increasing the squared L_2 risk (see, e.g., Yang and Baron [(1999), Section 2)]. Then we apply the adaptation recipe using the generated sample $\widetilde{X}_i, 1\le i\le n$ to get an adaptive strategy Δ^* with risk bound

$$R(g;n-1;\Delta^*)\le\inf_{j\ge1}\left(\frac{1}{n}\log\frac{1}{\pi_j}+\frac{1}{n}\sum_{i=0}^{n-1}R(g;i;\delta_j)\right).$$

Note that from the construction recipe, the adaptive estimators are convex combinations of the original estimators (but with random coefficients); thus they also stay between $1/2$ and $(A+1)/2$. With this boundedness property, the ratio of K-L divergence and squared L_2 distance is bounded above and below by constants depending only on A [see, e.g., Yang and Barron, Section 2)]. As a consequence, we have

$$r(g;n-1;\Delta^*)\le C_A\inf_{j\ge1}\left(\frac{1}{n}\log\frac{1}{\pi_j}+\frac{1}{n}\sum_{i=0}^{n-1}r(g;i;\delta_j)\right).$$

Finally, observing that for any estimator $\hat{g}$, the estimator $\hat{f}=2\hat{g}-1$ of f has risk bounded by $E\|f-\hat{f}\|^2\le4E\|g-\hat{g}\|^2$, from above, we have a combined strategy with the claimed risk bound in Theorem 2. Note that the combined strategy is randomized because of the dependence on the generated random

variables. Due to convexity of the loss, a nonrandomized strategy can be obtained with no bigger risk by averaging out the randomness in the generated sample W_i, $1 \le i \le n$ and the coin flips. This completes the proof of Theorem 2. □

PROOF OF THEOREM 3. For each class $\mathscr{F}_j$, let δ_j be an asymptotic minimax strategy. The existence of a minimax-rate adaptive estimator under K-L loss follows directly from Theorem 1 together with the assumption that the minimax risk is at a regular nonparametric rate.

For the proof of the conclusions under squared L_2 loss, observe that under the assumption on the density classes, for each f in a class, $g = (f + 1)/2$. Applying the risk bound given in Theorem 2, we have

$$\sup_{f \in \mathscr{F}_j} r(f; n-1; \Delta^*) \le C \inf_{j \ge 1} \left(\frac{1}{n} \log \frac{1}{\pi_j} + \frac{1}{n} \sum_{i=0}^{n-1} \sup_{f \in \mathscr{F}_j} r(f; n; \delta_j) \right).$$

Taking the strategies to be minimax-rate optimal for the classes, respectively, we have the minimax-rate adaptation under the squared L_2 loss. This completes the proof of Theorem 3. □

Acknowledgments. The author is grateful to Andrew Barron, from whom he learned the ideas on adaptive function estimation. The author thanks Mark Low for a helpful discussion. He also thanks the referees for their comments on earlier versions of this paper.

REFERENCES

BARRON, A. R. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (T. M. Cover and B. Gopinath, eds.) 85–91. Springer, New York.

BARRON, A. R. (1988). The convergence in information of probability density estimators. Presented at the IEEE International Symposium on Information Theory, Kobe, Japan.

BARRON, A. R. and COVER, T. M. (1991), Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054.

BARRON, A. R., BIRGÉ, L. and MASSART, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301–413.

BARRON, A. R., GYÖRFI, L. and VAN DE MEULEN, E. C. (1992). Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Trans. Inform. Theory* **38** 1437–1454.

BARRON, A. R. and XIE, Q. (1996). Asymptotic minimax regret for data compression, gambling, and prediction. Preprint.

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122.

BIRGÉ, L. and MASSART, P. (1996). From model selection to adaptive estimation. In *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.) 55–91. Springer, New York.

BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statis.* **24** 2524–2535.

CATONI, O. (1997). The mixture approach to universal model selection. Technical Report, LIENS-97-22, Ecole Normale Superieure, Paris, France.

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

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. Springer, New York.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539.

EFROIMOVICH, S. YU. (1995). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 557–568.

EFROIMOVICH, S. YU. and PINSKER, M. S. (1984). A self-educating nonparametric filtration algorithm. *Automat. Remote Control* **45** 58–65.

FEDER, M. and MERHAV, M. (1996). Hierarchical universal coding. *IEEE Trans. Inform. Theory* **42** 1354–1364.

HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

HAUSSLER, D. and OPPER, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist.* **25** 2451–2492.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

JUDITSKY, A. (1997). Wavelet estimators: adapting to unknown smoothness. *Math. Methods Statist.* **6** 1–25.

JUDITSKY, A. and NEMIROVSKI, A. (1996). Functional aggregation for nonparametric estimation. *Publication Interne, IRISA* 993.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

LEPSKII, O. V. (1991). Asymptotically minimax adaptive estimation I: upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

LUGOSI, G. and NOBEL, A. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27** 1830–1864.

TRIEBEL, H. (1975). Interpolation properties of ϵ -entropy and diameters. Geometric characteristics of embedding for function spaces of Sobolev–Besov type. *Mat. Sbornik* **98** 27–41.

WILLEMS, F. M. J., SHTARKOV, Y. M. and TJALKENS, T. J. (1998). The context-tree weighting method: basic properties. *IEEE Trans. Inform. Theory* **41** 653–664.

YANG, Y. (1996). Minimax optimal density estimation. Ph.D. dissertation, Dept. Statistics, Yale Univ.

YANG, Y. (1997). On adaptive function estimation. Technical Report 30, Dept. Statistics, Iowa State Univ.

YANG, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9** 475–499.

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

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

DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
SNEDECOR HALL
AMES, IOWA 50011-1210
E-mail: yyang@iastate.edu