Aggregating Regression Procedures for a Better Performance

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

December, 1999

Abstract

Methods have been proposed to linearly combine candidate regression procedures to improve estimation accuraccy. Applications of these methods in many examples are very succeesful, pointing to the great potential of combining procedures. A fundamental question regarding combining procedure is: What is the potential gain and how much one needs to pay for it?

A partial answer to this question is obtained by Juditsky and Nemirovski (1996) for the case when a large number of procedures are to be combined. We attempt to give a more general solution. Under a l_1 constrain on the linear coefficients, we show that for pursuing the best linear combination over n^τ procedures, in terms of rate of convergence under the squared L_2 loss, one can pay a price of order $O\left(\log n/n^{1-\tau}\right)$ when $0 < \tau < 1/2$ and a price of order $O\left(\left(\log n/n\right)^{1/2}\right)$ when $1/2 \leq \tau < \infty$. These rates can not be improved or essentially improved in a uniform sense. This result suggests that one should be cautious in pursuing the best linear combination, because one may end up with paying a high price for nothing when linear combination in fact does not help. We show that with care in aggregation, the final procedure can automatically avoid paying the high price for such a case and then behaves as well as the best candidate procedure in terms of rate of convergence.

Keywords and phrases: Aggregating procedures, adaptive estimation, linear combining, nonparametric regression.

1 Introduction

Recently, new ideas on combining different procedures for estimation, coding, forecasting, or learning have been considered in statistics and several related fields, resulting a number of very interesting results. The common theme behind these work is to automatically share the strength of the individual procedures in some sense. In the context of machine learning, it has been shown that with an appropriate weighting method, a combined procedure can behave close to the best procedure in terms of a certain cumulative loss, see, e.g., Vovk (1990), Littlestone and Warmuth (1994), Cesa-Bianchi *et al* (1997), and Cesa-Bianchi and Lugosi (1999). The focus has been on deriving mixed strategies with optimal performance without any probabilistic assumptions at all on the generation of the data. In the field of forecasting, combined forecasts have been shown to work better in various examples, see, e.g., Clemen (1989) for a review of work in that direction. In information theory, study of universal coding in the spirit of adaptation results in very interesting and powerful techniques also useful in other related fields such as machine learning and statistics. See Merhav and Feder (1998) and Barron, Rissanen and Yu (1998) for reviews of work in that field. In statistics, several methods have been recently proposed to linearly combine regression estimators. They include a model selection criterion based method by Buckland *et al* (1995), cross-validation based "stacking" by Wolpert (1992) and Breiman (1996a) (an earlier version is in Stone (1974)), a bootstrap based method by LeBlanc and Tibshirani (1996), a stochastic approximation based method by Juditsky and Nemirovski (1996), and information-theoretic based methods to combine density and regression estimators by Yang (1996, 1998, 1999bc) and Catoni (1997) for density estimation. Juditsky and Nemirovski proposed algorithms and derived interesting theoretical upper and lower bounds for linear aggregation in pursuing the best performance among the linearly combined estimators (with coefficients subject to an appropriate constraint). Yang (1998, 1999c) shows that with proper weighting, a combined procedure has a risk bounded above by a multiple of the smallest risk over the original procedures plus a small penalty.

The above mentioned theoretical work in statistics are in two related but different directions: one aiming at automatically achieving the best possible performance among the given collection of candidate procedures, and the other aiming at improving the performance of the original procedures. For the latter, the hope is that an aggregated procedure (through a convex or linear combination of the original procedures with data dependent coefficients) will significantly outperform each individual candidate procedure. Clearly the second direction is more aggressive. If one could identify the best linearly combined procedure, pursuing the best performance among the candidate procedures would be too conservative. On the other hand, common sense would suggest if one asks for more, one needs to pay

more. The present paper intends to contribute to the theoretical understanding on the gain and price for pursuing the best linear combination.

Suppose that we have M candidate regression procedures and consider the squared L_2 risk as a performance measure in estimating the regression function. In Yang (1998, 1999c) it is shown that a suitable data-dependent convex combination of these procedures results in an estimator that (under a minor condition) has a risk within a multiple of the smallest risk among the candidate procedures plus a small penalty of order $(\log M)/n$. Thus in terms of rate of convergence, with M candidate procedures to be combined, one only needs to pay the price basically of order $(\log M)/n$ for performing nearly as well as the best candidate procedure (which, of course, is unknown to the statistician). As long as M does not increase exponentially fast in n, the discrepancy $(\log M)/n$ is of order $\log n/n$, which does not affect the rate of convergence for typical nonparametric regression. As a consequence, when polynomially many nonparametric procedures are suitably combined, the estimator automatically converges at the best rate offered by the individual procedures. For the more aggressive goal of pursuing the best linear combination of the candidate procedures, under the constrain that the l_1 norm of the linear coefficients is bounded above by 1, Juditsky and Nemirovski (1996) proposed algorithms and showed that with M estimators to be combined, the aggregated estimator has a risk within a multiple of $\sqrt{(\log M)/n}$ of the smallest risk over all the linear combinations of the estimators. Furthermore, they show that, in general, this order $\sqrt{(\log M)/n}$ can not be overcome uniformly by any combining methods. Thus compared to combining for attaining the best performance, one has to pay a much higher price, $\sqrt{(\log M)/n}$, for searching for the best linear combination of the original procedures.

The work of Juditsky and Nemirovski (1996) is targeted at the case when M is large (e.g., their results are applied to restore Barron's class with M of a polynomial order in n). They derived the above mentioned lower bound when M and n have the relationship: $C_1 \log M \leq n \leq C_2 M \log M$ (where the constants C_1 and C_2 depend on the variance of the error and the assumed known upper bound on the supremum norm of the regression function f). The relationship implies that M is at least at order $n/\log(n)$. It is unclear then what happens when M is of a smaller order. For such a case, the order $\sqrt{(\log M)/n}$ may no longer be a valid lower bound. In the extreme case with M fixed (M does not grow as $n \to \infty$), one would expect a penalty of order close to the parametric rate $1/n$ instead of order $n^{-1/2}$. In this paper, we show that when M is of order n^τ, one only needs to pay the price of order $\log n/n^{1-\tau}$ for $0 \leq \tau < 1/2$. This rate can not be improved uniformly beyond a logarithmic factor.

Note that the order of the price increases dramatically as τ increases from 0, but after $\tau \geq 1/2$, it stays at the rate $\sqrt{(\log n)/n}$ as long as $\tau < \infty$. This phenomenon is closely related to the advantage

of sparse approximations as observed in wavelet estimation (see, e.g., Donoho and Johnstone (1998)), neural networks and subset selection (see, e.g., Barron (1994), Yang and Barron (1998), Yang (1999a), and Barron, Birgé and Massart (1999)). Under the l_1 constraint on the linear coefficients, when $\tau > 1/2$, there can not be too many (relative to M) large coefficients and combining sparsely selected procedures with suitably large coefficients achieves the optimal performance.

In applications, one does not know if the best linear combination can substantially improve the estimation accuracy so that the high price of order, e.g., $(\log n)/n^{1/2}$ is justified. Accordingly, it is not clear which direction to go when combining the candidate procedures. We show, fortunately, with some care in combining, an estimator can be aggressive and conservative automatically in the right way. For convenience in discussion, we will call the conservative goal *combining for adaptation*, and the aggressive goal *combining for improvement*.

The paper is organized as follows. In Section 2, we derive general risk bounds for combining M procedures. In Section 3, we study a combined procedure suitable for different purposes at the same time. In Section 4, we give an illustration using linear and sparse approximations. We briefly mention a generalization of the main results in Section 5. In Section 6, a basic combining algorithm and its property are presented, which provides a tool for the main results in this paper. The proofs of the results are in Section 7.

2 Risk bounds on linear aggregation

Consider the regression model

$$Y_i = f(X_i) + \sigma \cdot \varepsilon_i,\ i = 1, ...n,$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d. copies from the joint distribution of (X, Y) with $Y = f(X) + \sigma \cdot \varepsilon$. The explanatory variable X (could be high-dimensional) has an unknown distribution P_X. The variance parameter $\sigma > 0$ is unknown and the random variable ε is assumed to have a known density function $h(x)$ (with respect to Lebesgue or a general measure μ) with mean 0 and variance 1. The goal is to estimate the regression function f based on the data $Z^n = (X_i, Y_i)_{i=1}^n$.

Let δ be a regression estimation procedure producing estimator $\widehat{f}_i(x) = \widehat{f}_i(x; Z^i)$ for each $i \geq 1$. Let $\|\cdot\|$ denote the L_2 norm with respect to the distribution of X, i.e., $\|g\| = \sqrt{\int g^2(x) P_X(dx)}$. Let $R(f; n; \delta) = E\|f - \widehat{f}_n\|^2$ denote the risk of the procedure δ at the sample size n under the squared L_2 loss.

Let $\Delta = \{\delta_1, \delta_2, ..., \delta_M\}$ denote a collection of candidate procedures to be aggregated. Let $\widehat{f}_{j,i}(x) = \widehat{f}_{j,i}(x; Z^i)$ denote the estimator of f based on procedure δ_j given the observations Z^i for $i \geq 1$. Assume

$M = M_n$ changes according to the sample size n. In particular, we will consider the case when $M = Cn^\tau$ for some $0 \leq \tau < \infty$. When the sample size increases, one is allowed to consider more candidate procedures (possibly more and more complicated).

As in Juditsky and Nemirovski (1996), the coefficients for linear combination are suitably constrained. Let $\mathbf{F}_n = \{\sum_{1\leq j\leq M} \theta_j \widehat{f}_{j,n}(x) : \sum_{1\leq j\leq M} |\theta_j| \leq 1\}$ be the collection of linear combinations of the original estimators in Δ with coefficients summing up no more than 1 in absolute values. The hope behind the consideration of the linear aggregation is that a certain combination of the original estimators might have a much better performance than the individual ones. Advantages of such combining have been empirically demonstrated in several related fields (e.g., Bates and Granger (1969), Breiman (1996)). Let $\|\cdot\|_1^M$ denote the l_1 norm on R^M, i.e., $\|\theta\|_1^M = \sum_{1\leq j\leq M} |\theta_j|$. Define

$$R^*(f; n; \Delta) = \inf_{\|\theta\|_1^M \leq 1} E\|f - \sum_{1\leq j\leq M} \theta_j \widehat{f}_{j,n}\|^2.$$

It is the smallest risk over all the estimators in the linear aggregation class $\mathbf{F}_n$. Obviously, $R^*(f; n; \Delta) \leq \inf_{1\leq j\leq M_n} R(f; n; \delta_j)$. In this paper, unless stated otherwise, by linear combination, we mean linear combination with the coefficients satisfying the above l_1 constraint.

We need the following assumptions for our results.

A1. The regression function $f(x)$ is uniformly bounded, i.e., $\|f\|_\infty \leq A < \infty$. The variance parameter σ is bounded above and below by known positive constants $\overline{\sigma} < \infty$ and $\underline{\sigma} > 0$.

A2. The error distribution h has a finite fourth moment and satisfies that for each pair $0 < s_0 < 1$ and $T > 0$, there exists a constant B (depending on s_0 and T) such that

$$\int h(x) \log \frac{h(x)}{\frac{1}{s} h(\frac{x-t}{s})} dx \leq B\left((1-s)^2 + t^2\right)$$

for all $s_0 \leq s \leq s_0^{-1}$ and $-T < t < T$.

The constants A and B in the above assumptions are involved in the derivation of the risk bounds, but they need not to be known in our aggregation procedure. The Assumption A2 is mild and is satisfied by Gaussian, double-exponential, and many other smooth distributions.

An algorithm, named ARM in Yang (1999c), to combine procedures for adaptation is given in Section 6. This algorithm serves as a building block for the results in this paper. Through a suitable discretization of the linear coefficients together with a sparse approximation, the problem of combining for improvement becomes the problem of combining for adaptation over a (much) larger class of procedures. We have the following performance upper bound.

THEOREM 1: *Assume that Conditions A1 and A2 are satisfied. For any given collection of estimation procedures* $\Delta = \{\delta_j, 1 \le j \le M_n\}$, *we can construct a combined procedure* δ^* *such that*

$$R(f;n;\delta^*) \le C \begin{cases} R^*\left(f;\frac{n}{2};\Delta\right) + \frac{M_n \log(1+n/M_n)}{n} & \text{when } M_n < \sqrt{n} \\ R^*\left(f;\frac{n}{4};\Delta\right) + \frac{\log M_n}{\sqrt{n \log n}} & \text{when } M_n \ge \sqrt{n} \end{cases},$$

where C *is a constant depending on* A, $\underline{\sigma}$, $\overline{\sigma}$, *and* h. *In particular, if* $M_n \le C_0 n^\tau$ *for some* $\tau > 0$ *and* $C_0 > 0$, *then*

$$R(f;n;\delta^*) \le C' \begin{cases} R^*\left(f;\frac{n}{4};\Delta\right) + \left(\frac{\tau \log n}{n}\right)^{1/2} & \text{when } 1/2 \le \tau < \infty \\ R^*\left(f;\frac{n}{2};\Delta\right) + \frac{\log n}{n^{1-\tau}} & \text{when } 0 \le \tau < 1/2, \end{cases} \tag{1}$$

where the constant C' *depends on* A, $\underline{\sigma}$, $\overline{\sigma}$, C_0, *and* h.

REMARK: The condition on σ in Assumption A1 is mainly technical (it is not really needed to perform the procedure). The lower bound condition on σ is not essential even from a technical point of view, since one can always add a little bit noise to the observations to satisfy the condition usually without affecting the rate of convergence.

The constructed procedure δ^* is given in the proof of Theorem 1 in Section 7. Note that for both parametric and nonparametric regression, for a good procedure, $R(f;n;\delta)$ and $R(f;n/2;\delta)$ are usually of the same order. Thus it is typically the case that $R^*(f;n;\Delta)$ and $R^*\left(f;\frac{n}{2};\Delta\right)$ converge at the same rate. From the result, when $\tau \ge 1/2$, the penalty term for pursuing the best linear combination of n^τ procedures is of order $((\log n)/n)^{1/2}$ (independent of τ). This rate is obtained by Juditsky and Nemirovski (1996) with a weaker assumption on the errors (finite variance), but requiring the knowledge of A. When $\tau < 1/2$, our result above shows that the penalty is smaller in order, resulting in a possibly much faster rate of convergence. For an extreme example, when M_n is fixed, the price we pay is only of order $\log n/n$.

How good are the upper bounds derived here? Juditsky and Nemirovski (1996) show that when M and n satisfy $C_1 \log M \le n \le C_2 M \log M$ for some constants C_1 and C_2 (i.e., M is no smaller than order $n/\log n$ but not too large), the order $((\log n)/n)^{1/2}$ can not be improved in a minimax sense. We show in general, the rates given in Theorem 1 can not be improved up to possibly a logarithmic factor for some cases. For simplicity, assume that the errors are normally distributed with variance 1.

THEOREM 2: *Consider* $M_n = \lfloor C_0 n^\tau \rfloor$ *for some* $\tau > 0$. *There exist* M_n *procedures* $\Delta_{M_n} = \{\delta_j, 1 \le j \le M_n\}$ *such that for any aggregated procedure* $\delta^{(n)}$ *based on* Δ_{M_n}, *one can find a regression function* f *with* $\| f \|_\infty \le \sqrt{2}$ *satisfying*

$$R\left(f;n;\delta^{(n)}\right) - R^*(f;n;\Delta_{M_n}) \ge C \begin{cases} \left(\frac{\log n}{n}\right)^{1/2} & \text{when } 1/2 < \tau < \infty \\ \frac{1}{n^{1-\tau}} & \text{when } 0 \le \tau \le 1/2, \end{cases}$$

where the constant C does not depend on n.

Thus no aggregation method can achieve the smallest risk over all the linear combinations within an order smaller than the ones given above in accordance with τ uniformly over all bounded regression functions. Note that the lower rate matches the upper rate when $\tau > 1/2$ and the upper and lower rates differ only in logarithmic factors when $0 \leq \tau \leq 1/2$.

It is interesting to notice how the price (in rate) for combining for improvement changes according to M_n. In the beginning, it basically increases linearly in M_n, but after M reaches $\sqrt{n}$, it increases much more slowly in a logarithmic fashion. Accordingly, it stays at rate $\left(\frac{\log n}{n}\right)^{1/2}$ as long as M_n increases polynomially in n.

In a different direction, Yang (1998, 1999c) shows that one only needs to pay the price of order $(\log M)/n$ to pursuit the less ambitious goal of achieving the best performance among the original M procedures. Observing the dramatic difference between the two penalties, one naturally faces the question: Should we combine for adaptation or for improvement? If one of the original procedures happen to behave the best (or close to the best) among all the linear combinations, or at least one of the original procedures converges at a rate faster than $(\log n)/n^{1-\tau}$ (for $0 \leq \tau < 1/2$) or $\sqrt{\log n/n}$ (for $\tau \geq 1/2$), if one aggregates for better performance, one could be unfortunately paying too high a price for nothing but hurting the convergence rate in estimating f. In terms of rate of convergence, combining for improvement is worth the effort only if $R^*(f; n/2; \Delta)$ plus the penalty in (1) is of a smaller order than $(\log M)/n + \inf_j R(f; n/2; \delta_j)$. Since the risks are of course unknown, in applications, one does not know in advance whether to combine for adaptation or combine for improvement. A wrong choice can lead to a much worse rate of convergence. In the next section, we show one can actually handle the two goals optimally at the same time.

3 Multi-purpose aggregation

Here we show when combining the procedures properly, one can have the potential of obtaining a large gain in estimation accuracy yet without losing much when there happens to be no advantage considering sophisticated linear combinations.

Let us consider a slightly different setting compared to the previous section. Suppose that we have a countable collection of candidate procedures $\Delta = \{\delta_1, \delta_2, ...\}$. Under this setting, one does not need to decide before hand how many procedures should be included at a given sample size. Consider three different approaches to combine the procedures in Δ.

The first approach is to combine the procedures for adaptation. Here one intends to capture the

best performance in terms of rate of convergence among the candidate procedures. Let δ_A^* denote this combined procedure based on Δ using the Three-Stage ARM Algorithm as given in the Section 6. Since Δ is not (necessarily) a finite collection, one can not use the uniform weight. The prior weight π_j is taken to be $ce^{-\log^* j}$, where $\log^*$ is defined by $\log^* x = \log(x+1) + 2\log\log(x+1)$ and the constant c is chosen to normalize the weights to add up to 1. Based on Proposition 1 in Section 6, we have that for any f with $\| f \|_\infty < \infty$,

$$R(f;n;\delta_A^*) \le C_1 \inf_j \left(\frac{\log(j+1)}{n} + R(f;n/2;\delta_j) \right) =: C_1 R_1^*(f;n;\Delta), \tag{2}$$

where the constant C_1 depends on $\| f \|_\infty$, $\underline{\sigma}$, $\overline{\sigma}$, and h. In the rest of the paper, unless stated otherwise, a constant C (with or without subscript) may depend on $\| f \|_\infty$, $\underline{\sigma}$, $\overline{\sigma}$, and h. For convenience, we may use the same symbol C for different such constants in different places. From above, if one procedure, say δ_{j^*} behaves the best, then the penalty is of order $\frac{1}{n}$. If the best estimator changes according to n, then $\inf_j \left(\frac{\log(j+1)}{n} + R(f;n/2;\delta_j) \right)$ is a trade-off between complexity and estimation accuracy.

The second approach targets at the best performance among all the linear combinations of the original procedures up to different orders. For each integer $L \ge 1$, let δ^L denote the combined (for improvement) procedure based on the first L procedures $\delta_1, , ..., \delta_L$ as used for Theorem 1. Then combine the procedures $\{\delta^1, \delta^2, ...\}$ with weight $ce^{-\log^* j}$ for $j \ge 1$ as defined earlier. Let δ_B^* denote this combined procedure. Let Δ_L denote the set of the first L procedures in Δ. Let

$$\psi_n(L) = \begin{cases} \frac{L\log(1+n/L)}{n} & 1 \le L < \sqrt{n} \\ \frac{\log L}{\sqrt{n\log n}} & L \ge \sqrt{n}. \end{cases}$$

By Theorem 1 and Proposition 1, we have that for any f with $\| f \|_\infty < \infty$,

$$R(f;n;\delta_B^*) \le C_2 \inf_L \left(R^*\left(f;\frac{n}{2};\Delta_L\right) + \psi_n(L) \right) =: C_2 R_2^*(f;n;\Delta). \tag{3}$$

The third approach recognizes that in many cases, when combining a lot of procedures, the best linear combination may concentrate on only a few procedures. For such a case, working with these important procedures only leads to a much smaller price when combining for improvement. This calls for additional care in aggregation and it can be done as follows. For each integer $L > 1$, $1 \le k < L$, and a subset S of $\{1, 2, ..., L\}$ of size k, let $\delta(S)$ be the combined (for improvement) procedure based on $\{\delta_j : j \in S\}$ as for (1). Then let $\delta_{L,k}$ be the combined (for adaptation) procedure based on all such $\delta(S)$ with uniform weight $1/\binom{L}{k}$ (there are $\binom{L}{k}$ many such procedures). Then let $\delta^{(L)}$ be the combined (for adaptation) procedure based on $\delta_{L,1}, ..., \delta_{L,L-1}$ using the uniform weight $1/(L-1)$. Let δ_C^* denote the combined (for adaptation) procedure based on $\delta^{(L)}$, $L \ge 2$ with weight $c' \log^* j$, where the constant c'

is chosen such that $\sum_{j=2}^{\infty} c^{'} e^{-\log^* j} = 1$. Let Δ_S denote the collection of procedures $\{\delta_j : j \in S\}$. Based on Proposition 1 and Theorem 1, we have that for any f with $\| f \|_\infty < \infty$,

$$\begin{aligned} & R(f; n; \delta_C^*) \\ \leq \; & C_3 \inf_{L \geq 2} \left(\inf_{1 \leq k \leq L-1} \left(\inf_{|S|=k, S \subset \{1,2,\ldots,L\}} R^* \left(f; \frac{n}{16}; \Delta_S \right) + \psi_n(k) + \frac{\log(\frac{L}{k})}{n} \right) \right) \qquad (4) \\ =: \; & C_3 R_3^* \left(f; n; \Delta \right). \end{aligned}$$

Now we combine these three procedures δ_A^*, δ_B^*, and δ_C^* with equal weight 1/3. And let δ_F denote the final combined procedure. Note that it is still a linear combination of the original procedures. We have the following result.

COROLLARY 1: *Assume Conditions A1 and A2 are satisfied. Then for each f with $\|f\|_\infty < \infty$, we have*

$$R\left(f; n; \delta_F\right) \leq C \min\left(R_1^*\left(f; n/2; \Delta\right), R_2^*\left(f; n/2; \Delta\right), R_3^*\left(f; n/2; \Delta\right)\right),$$

where $R_1^*\left(f; n; \Delta\right), R_2^*\left(f; n; \Delta\right), R_3^*\left(f; n; \Delta\right)$ are given in (2), (3), and (4).

The above result characterizes good performance of the final estimator simultaneously in three directions in terms of rate of convergence. First of all, the final estimator converges as fast as any original procedure. Secondly, when linear combinations of the first L_n procedures (for some $L_n > 1$) can improve estimation accuracy dramatically, one pays the price at most of order $\psi_n(L_n)$ for the better performance. When L_n is small, the gain is substantial. When certain linear combinations of a small number of procedures perform well, the final estimator can also take advantage of that. In summary, the final estimator can behave both aggressively (combining for improvement) and conservatively (combining for adaptation) which ever is better.

4 An illustration via linear approximation

We illustrate the result of multi-purpose aggregation studied in the previous section through an example with linear and sparse approximations. We assume that $x \in [0,1]^d$ $(1 \leq d \leq \infty)$.

Let $\{\Phi_j : j = 1, 2, ...\}$ be a collection of linear approximation systems. For each j, $\Phi_j = \{\varphi_{j,1}(x), \varphi_{j,2}(x), ...\}$ is a chosen collection of linearly independent functions in $L^2[0,1]^d$. Traditionally orthonormal bases (or at least with some frame properties) have been emphasized. Recently non-orthogonal and/or over complete bases have been advocated and studied. Relaxation of orthogonality enables one to consider e.g., trigonometric expansions with fractional frequencies and neural network models. Considering different bases provides much more flexibility that gives a great potential to improve estimation accuracy, especially in high-dimensional settings. See Barron and Cover (1991), Mallat and Zhang (1993), Barron

(1994), Donoho and Johnstone (1994), Juditsky and Nemirovski (1996), Yang and Barron (1998), Yang (1999a), and Barron, Birgé and Massart (1999) for some work in those directions.

For a fixed j, the (squared L_2) approximation error of f using the first N terms is

$$\eta_{j,N}(f) = \inf_{\{a_l\}} \| f - \sum_{l=1}^{N} a_l \varphi_{j,l} \|^2 .$$

We call this individual approximation. The approximation error of f using linear combinations of the individual approximations of f up to N terms based on the first L systems is

$$\eta_N^L(f) = \inf_{\{a_{j,l}\}} \| f - \sum_{j=1}^{L} \sum_{l=1}^{N} a_{j,l} \varphi_{j,l} \|^2 .$$

We call this linearly combined approximation. Obviously $\eta_N^L(f) \le \eta_{j,N}(f)$ for $1 \le j \le L$. When $\eta_N^L(f) \ll \eta_{j,N}(f)$ for $1 \le j \le L$ with the right size, the advantage of considering linear combinations over different systems can be substantial. The approximation error of f based on sparse approximation using k out of the first L systems is

$$\eta_N^{L,k}(f) = \inf_{S \subset \{1,\ldots,M\}, |S|=k} \inf_{\{a_{j,l}\}} \| f - \sum_{j \in S} \sum_{l=1}^{N} a_{j,l} \varphi_{j,l} \|^2 .$$

We call this sparsely combined approximation. The sparse approximation can improve estimation accuracy compared to the linearly combined approximation if only a few approximation systems are actually needed in the linearly combined approximation, i.e., one can find $k \ll L$ such that $\eta_N^{L,k}(f)$ is close to $\eta_N^L(f)$.

For a given j and N, traditional linear model estimators (e.g., based on the least squares principle or projection estimators with orthogonal basis functions) can be used to estimate the best parameters in the linear approximation, resulting in the familiar bias-squared (approximation error) plus variance (estimation error) trade-off for the mean squared error. As is well-known, the variance is typically of order N/n under minor conditions.

Combining the approximation error and estimation error, one can bound $R_1^*(f;n;\Delta)$, $R_2^*(f;n;\Delta)$, and $R_3^*(f;n;\Delta)$ as defined in (2), (3), and (4) as follows

$$R_1^*(f;n;\Delta) = O\left(\inf_j \left(\eta_{j,N}(f) + \frac{N}{n} + \frac{\log j}{n} \right) \right), \tag{5}$$

$$R_2^*(f;n;\Delta) = O\left(\inf_{L,N} \left(\eta_N^L(f) + \frac{LN}{n} + \psi_n(L) \right) \right), \tag{6}$$

$$R_3^*(f;n;\Delta) = O\left(\inf_{L,N} \left(\inf_{1 \le k \le L-1} \left(\eta_N^{L,k}(f) + \psi_n(k) + \frac{k \log L}{n} + \frac{kN}{n} \right) \right) \right). \tag{7}$$

Based on Corollary 1 and the above bounds, one can derive rate of convergence for the final aggregated procedure δ_F under various assumptions on the approximation errors $\eta_{j,N}(f)$, $\eta_N^L(f)$, and $\eta_N^{L,k}(f)$. The conclusion is basically that, in terms of rate of convergence, the final estimator behaves as well as the best estimator based on an individual approximation system, or as the linearly combined estimator, or as the sparsely combined estimator, which ever is the best.

When the basis functions are orthonormal, conditions on the L_2 approximation errors typically correspond to conditions on the coefficients, resulting in simple characterization of the functions. Here we give an example. Suppose $d = \infty$ and assume $X = (X_1, X_2, ...)$ has independent, uniformly distributed components (or after suitable transformations). We assume the true regression function is additive, i.e.,

$$f(x) = c_0 + f_1(x_1) + f_2(x_2) + \cdots. \tag{8}$$

To estimate the additive component $f_j(x_j)$, a linear approximation system $\Phi_j = \{\varphi_{j,1}(x_j), \varphi_{j,2}(x_j), ...\}$ is used. Assume the basis functions are orthonormal with mean zero. For a given j, let $\widehat{f}_{j,N}(x_k)$ be the projection estimator of $f_j(x_j)$ based on the first N basis functions in Φ_j. That is, $\widehat{f}_{j,N}(x_k) = \sum_{i=1}^N \widehat{\theta}_{j,i}\varphi_{j,i}(x_k)$, where $\widehat{\theta}_{j,i} = \frac{1}{n}\sum_{l=1}^n Y_l\varphi_{j,l}(X_{k,l})$. For simplicity, assume $\| f \|_\infty \leq A$ for some known constant $A > 0$ and the estimators are accordingly clipped into the range. Let $\delta_{j,N}$, $j \geq 1$, $N \geq 1$ denote these regression procedures. Let δ_A^*, δ_B^*, and δ_C^* be the differently combined procedures as constructed in the previous section and let δ_F denote the final procedure combining them together.

Assume $f_j(x_j) = \sum_{l=1}^\infty \theta_{j,l}\varphi_{j,l}(x_j)$ for $j \geq 1$ and assume the coefficients satisfy the following condition B0:

$$\sum_{j=1}^\infty j^{2\beta}\left(\sum_{i=1}^\infty i^{2s}\theta_{j,i}^2\right) < \infty. \tag{9}$$

for some $s > 0$ and $\beta > 0$. When the true regression function is actually univariate in one variable, say x_{j_0}, then $\theta_{j,i} = 0$ for all j and i except $j = j_0$. If one knew this is the case, one can ignore the other variables. Let B1 denote this condition. Another condition, denoted B2, is that $\theta_{j,i} = 0$ for all $j \geq 1$ and $i > i_0$ for some unknown integer i_0, and in addition,

$$\sum_{j=1}^\infty \sum_{i=1}^{i_0} |\theta_{j,i}| < \infty.$$

COROLLARY 2: *Assume the errors are normally distributed with σ^2 bounded above and below by known constants. If f satisfies Condition B0, we have*

$$R(f; n; \delta_F) = O\left(n^{-\frac{2s}{1+s(2+1/\beta)}}\right). \tag{10}$$

If f satisfies conditions B0 and B1, we have

$$R(f; n; \delta_F) = O\left(n^{-\frac{2s}{1+2s}}\right). \tag{11}$$

If f satisfies conditions B0 and B2, we have

$$R\left(f;n;\delta_F\right)=O\left(\left(\log n/n\right)^{1/2}\right). \tag{12}$$

Note that the procedure δ_F does not require knowledge of the constants s and β (or i_0). Thus the rate $n^{-\frac{2s}{1+s(2+1/\beta)}}$ is adaptively achieved. When s or β is very small, the rate of convergence is very slow. Under the additional assumption of B1 or B2, a much better rate of convergence is automatically achieved by the aggregated procedure.

REMARKS:

1. In the construction of the aggregated procedure δ_C^*, sparseness is in terms of the number of procedures being combined. One can also consider sparseness in terms of the number of terms in the linear approximation within each approximation system. Then the same convergence rate $(\log n/n)^{1/2}$ can be obtained under Condition B2 without assuming that for each j, there are only finitely many non-zero coefficients. See Yang and Barron (1998) for such a treatment in density estimation based on models selection.

2. Under the assumptions that $X=(X_1,...)$ has independent and uniformly distributed components and that the basis functions have mean zero, $E\varphi_{j,l}(X_j)\varphi_{j',l'}(X_{j'})=0$ for all $j\neq j'$. These strong conditions make the approximation error readily bounded under Condition B0. Without these conditions, the convergence rates in Corollary 2 can be shown to still hold under the direct conditions $\inf_{\{\theta_{j,l}\}}\parallel f-\sum_{j=1}^{J}\sum_{i=1}^{\infty}\theta_{j,l}\varphi_{j,l}\parallel^2=O\left(J^{-2\beta}\right)$ and $\sum_{i=1}^{\infty}i^{2s}\theta_{j,i}^2<\infty$ for each j. Also the additive condition (8) can be expressed in terms of pre-specified linear combinations of the original explanatory variables rather than the original explanatory variables themselves.

3. If f happens to be "parametric" in the sense that it can be expressed as a linear combination of finitely many basis functions (possibly across different systems), then the convergence rate of the final procedure is $O\left(\left(\log n\right)/n\right)$, possibly losing a logarithmic factor.

4. When $\beta>1/2$, under Condition B0 and that $\theta_{j,i}=0$ for all j and i except $j=j_0$, the condition $\sum_{j=1}^{\infty}\sum_{i=1}^{i_0}|\theta_{j,i}|<\infty$ is automatically satisfied. For this case, it can be shown that $R\left(f;n;\delta_F\right)$ in fact converges at a better rate $n^{-2\beta/(2\beta+1)}$ than $\left(\log n/n\right)^{1/2}$.

5 Generalization

The main results in this paper can be generalized with little difficulty in two directions based on an analysis similar to that in Yang (1999c). Firstly, the error distribution h need not to be known completely. It suffices to assume that h is in a countable collection of candidate error distributions. This gives more

flexibility to hand errors with different degree of heavy tail. Secondly, one does not need to require that the random errors have a constant variance function. Assume instead that for each δ_j, in addition to having an estimator $\hat{f}_{j,n}$ of the regression function, we also have an estimator $\hat{\sigma}_{j,n}$ of the variance function. The procedures can share variance estimators if so desired. The procedures can be combined for estimating f using both the regression estimators and the variance estimators (see Yang (1999c)). A recent work on variance estimation is in Ruppert *et al* (1997), where a local polynomial method is proposed with a theoretical justification.

6 A Three-Stage algorithm to combine procedures for adaptation

Let $\Delta = \{\delta_j, j \geq 1\}$ be a collection of regression procedures. The index set $\{j \geq 1\}$ is allowed to degenerate to a finite set. Let π_j be positive numbers summing up to one, i.e., $\sum_{j=1}^{\infty} \pi_j = 1$. They will be used as prior weights on the procedures. The following is an algorithm to combine candidate procedures for adaptation as essentially given in Yang (1999c).

A Three-Stage ARM Algorithm

Step 1. Split the data into three parts $Z^{(1)} = (X_i, Y_i)_{i=1}^{n_1}$, $Z^{(2)} = (X_i, Y_i)_{i=n_1+1}^{n_1+n_2}$, and $Z^{(3)} = (X_i, Y_i)_{i=n_1+n_2+1}^{n}$. Let $n_3 = n - n_1 - n_2$.

Step 2. Obtain estimates $\hat{f}_{j,n_1}(x;Z^{(1)})$ of f based on $Z^{(1)}$ for $j \geq 1$.

Step 3. Estimate the variance σ^2 for each procedure by

$$\hat{\sigma}_j^2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} \left(Y_i - \hat{f}_{j,n_1}(X_i) \right)^2.$$

Step 4. For each j, evaluate predictions. For $n_1 + n_2 + 1 \leq k \leq n$, predict Y_k by $\hat{f}_{j,n_1}(X_k)$. For $n_1 + n_2 + 1 \leq k \leq n$, compute

$$E_{j,k} = \frac{\Pi_{i=n_1+n_2+1}^{k} h\left(\frac{(Y_i - \hat{f}_{j,n_1}(X_i))}{\hat{\sigma}_j} \right)}{\hat{\sigma}_j^{k-n_1-n_2}}.$$

Step 5. Let

$$W_{j,k} = \frac{\pi_j E_{j,k}}{\sum_{l \geq 1} \pi_l E_{l,k}}$$

and compute the final weight

$$\overline{W}_j = \frac{1}{n_3} \sum_{k=n_1+n_2+1}^{n} W_{j,k}$$

The final estimator is

$$\widetilde{f}_n(x) = \sum_{j=1}^{\infty} \overline{W}_j \hat{f}_{j,n/2}(x) \tag{13}$$

The combined estimator has the following theoretical property. For simplicity in notation, assume that n is a multiple of 4, and then take $n_1 = n/2$ and $n_2 = n_3 = n/4$. We assume that the estimator $\hat{\sigma}_j$ are bounded above and below by positive constants $\overline{\sigma}$ and $\underline{\sigma}$ (otherwise one needs to clip the estimator to be in that range).

PROPOSITION 1: *Assume Conditions A1 and A2 hold. Then the above convexly combined estimator* $\widetilde{f}_n$ *satisfies*

$$E\|f - \widetilde{f}_n\|^2 \le C \inf_j \left(\frac{1}{n}\left(1 + \log\frac{1}{\pi_j}\right) + E\|f - \hat{f}_{j,n/2}\|^2 \right),$$

where the constant C *depends only on* A, $\overline{\sigma}$, $\underline{\sigma}$, *and* h. *In particular, if there are* M *procedures to be combined with uniform weight, then*

$$E\|f - \widetilde{f}_n\|^2 \le C \left(\frac{\log M}{n} + \inf_j E\|f - \hat{f}_{j,n/2}\|^2 \right).$$

REMARKS:

1. In the ARM algorithm, the second stage is used to estimate σ^2. Here the estimators are derived in terms of predictions based on the individual regression procedures. The use of these variance estimators does not get in the way of estimating the regression function f in terms of rate of convergence. One can also use common model-independent estimators of σ^2 (see, e.g., Rice (1984)). Then one does not need this stage, and accordingly, the risk of the variance estimators will appear in the risk bound on estimating f.

2. As discussed in Yang (1999c), the estimator $\widetilde{f}_n$ depends on the order of observations. For improvement, one can randomly permute the order of observations a number of times and average the corresponding estimators.

3. In the definition of the final estimator $\widetilde{f}_n = \sum_{j=1}^{\infty} \overline{W}_j \hat{f}_{j,n/2}(x)$, we use $\hat{f}_{j,n/2}(x)$ instead of $\hat{f}_{j,n}(x)$ to have a cleaner risk bound. But $\hat{f}_{j,n}(x)$ should be a slightly better choice in terms of accuracy.

PROOF OF PROPOSITION 1: The result is proved in Yang (1999c) for the case when there are finitely many, say J, candidate procedures with equal weight $\pi_j = 1/J$ for $1 \le j \le J$. The proof for the general case can be done similarly.

7 Proof of the results

PROOF OF THEOREM 1: There are mainly two steps in our derivation of an aggregated procedure yielding the given risk bound. First, we discretize (with suitable accuracy) the coefficients for linear combinations and then treat the set of all the corresponding linearly discretely combined estimators as a new collection of candidate estimators. For suitable discretization, some results on metric entropy are very helpful. In the second step, we combine these estimators for adaptation using the algorithm ARM proposed in Yang (1999c) and described in Section 6. When M_n is large, however, an additional difficulty arises and an idea of sparse combining takes care of the problem.

We consider first the case when $M_n < \sqrt{n}$. Let $G = \{\theta = (\theta_1, ..., \theta_M) : \sum_{i=1}^M |\theta_i| \le 1\}$. Let N_ϵ be an ϵ-net in G under the l_1^M distance, i.e., for each $\theta \in G$, there exists $\theta' \in N_\epsilon$ such that $\|\theta - \theta'\|_1^M = \sqrt{\sum_{i=1}^M |\theta_i - \theta_i'|} \le \epsilon$. An ϵ-net in G yields a suitable net in the set $\mathbf{F}_n$ of the linear combinations of the original estimators. For simplicity in notation, let $\widehat{f}_1, ..., \widehat{f}_M$ denote the original estimators at the sample size n. Let F_ϵ be the set of the linear combinations of the estimators $\widehat{f}_1, ..., \widehat{f}_M$ with coefficients in N_ϵ. Then for any estimator $\widehat{f} = \sum_{i=1}^M \theta_i \widehat{f}_i$ with $\theta \in G$, there exists $\theta' \in N_\epsilon$ such that

$$\|\widehat{f} - \sum_{i=1}^M \theta_i' \widehat{f}_i\| = \|\sum_{i=1}^M \left(\theta_i - \theta_i'\right) \widehat{f}_i\| \le A\|\theta - \theta'\|_1^M \le A\epsilon. \tag{14}$$

Now we combine all the estimators in F_ϵ using the ARM algorithm given in Section 6 with uniform weight $1/|N_\epsilon|$. Let $\widehat{f}_n$ denote the combined estimator. By Proposition 1, for any f with $\|f\|_\infty < \infty$, we have

$$E\|f - \widehat{f}_n\|^2 \le \frac{C \log(|N_\epsilon|)}{n} + C \inf_{\widehat{f} \in F_\epsilon} R(f; \widehat{f}; n/2),$$

where C depends only on A, $\overline{\sigma}$, $\underline{\sigma}$, and h. Since F_ϵ is an $(A\epsilon)$-net in $\mathbf{F}_n$, by triangle inequality, for any f, we have $\inf_{\widehat{f} \in F_\epsilon} R(f; \widehat{f}; n/2) \le 2 \inf_{\widehat{f} \in \mathbf{F}_n} R(f; \widehat{f}; n/2) + 2A^2\epsilon^2$. It follows that

$$E\|f - \widehat{f}_n\|_2^2 \le \frac{C \log(|N_\epsilon|)}{n} + 2C \inf_{\widehat{f} \in \mathbf{F}_n} R(f; \widehat{f}; n/2) + 2A^2 C\epsilon^2. \tag{15}$$

To get the best upper bound (in order), we need to minimize $\frac{\log(|N_\epsilon|)}{n} + 2A^2\epsilon^2$ when discretizing G. Note that the logarithm of the smallest size of N_ϵ is the covering entropy of the set G under the l_1^M distance (see, e.g., Kolmogorov and Tihomirov (1959) for properties of metric entropies). For this case, metric entropy orders are known. The following result is given in terms of the entropy number, i.e., the worst case approximation error with the best net of size of 2^k points. Let ϵ_k denote the entropy number of G. From Edmunds and Triebel (1989, Proposition 3.1.3), when $k \ge M$, $\epsilon_k \le c2^{-k/M}$ for some constant c independent of k and M. Take

$$k = \frac{M\left(\log(n/M) + 2\log 2\right)}{2\log 2}$$

(note that $k \geq M$). (Strictly speaking, we need to round up or down to make k an integer.) Then

$$\frac{\log(|N_\epsilon|)}{n} + 2A^2\epsilon^2 \leq \frac{M\left(\log(n/M) + 2\log 2\right)}{(2\log 2)\, n} + \frac{(Ac)^2\, M}{2n} \leq \frac{c^{'} M \log(1 + n/M)}{n},$$

where $c^{'}$ depends only on A and c. The upper bound in Theorem 1 for $M < \sqrt{n}$ then follows.

Now consider the other case: $M \geq \sqrt{n}$. The argument above leads to a rate $(M/n)\log(1 + n/M)$ for $M \leq n$, which as will be seen is only sub-optimal. For this case, due to the l_1 constraint, the number of large coefficients is small relative to M when $M \gg \sqrt{n}$. An appropriate search of the large coefficients can result in optimal rate of convergence, as we derive below.

Note that for $\|\theta\|_1^M \leq 1$, $\|\sum_{i=1}^M \theta_i \widehat{f}_i\| \leq A$. Then by a sampling argument (see e.g., Lemma 1 in Barron (1993)), for each m, there exist a subset $I \subset \{1, ..., M\}$ of size m and $\theta_I^{'} = (\theta_i^{'}, i \in I)$ such that $\|\sum_{i=1}^M \theta_i \widehat{f}_i - \sum_{i \in I} \theta_i^{'} \widehat{f}_i\| \leq A/\sqrt{m}$. Taking $m^* = \sqrt{n/\log n}$, we have $\|\sum_{i=1}^M \theta_i \widehat{f}_i - \sum_{i \in I} \theta_i^{'} \widehat{f}_i\| \leq A\left(\log n/n\right)^{1/4}$. Consider an ϵ-net in $B_I = \{\theta_I : \sum_{i \in I} |\theta_i| \leq 1\}$ under the $l_1^{m^*}$ distance. Again by Edmunds and Triebel (1989), taking $k = \frac{m^*(\log(n/m^*) + 2\log 2)}{2\log 2}$, the best ϵ-net has approximation accuracy $\epsilon \leq c/2\sqrt{m^*/n}$. Then as in (14), we know that there exists $\theta_I^{''}$ in this ϵ-net such that $\|\sum_{i \in I} \theta_i^{'} \widehat{f}_i - \sum_{i \in I} \theta_i^{''} \widehat{f}_i\| \leq Ac/2\sqrt{m^*/n}$. Thus for each $\widehat{f} \in \mathbf{F}_n$, there exist $I^* \subset \{1, ..., M\}$ of size m^* and $\theta_I^{''}$ such that

$$\|\sum_{i=1}^M \theta_i \widehat{f}_i - \sum_{i \in I^*} \theta_i^{''} \widehat{f}_i\| \leq \frac{A\left(\log n\right)^{1/4}}{n^{1/4}} + \frac{Ac}{2n^{1/4}\left(\log n\right)^{1/4}} \leq \frac{c^{''}\left(\log n\right)^{1/4}}{n^{1/4}},$$

where $c^{''}$ depends only on A and c. Notice that, in general, I^* depends on f and therefore it should be chosen adaptively. The above analysis suggests the following method of sparse combining.

For each fixed subset $I \subset \{1, ..., M\}$ of size m^*, discretize the linear coefficients as described above. Then (with uniform weight) combine the corresponding linear combinations of the procedures in Δ. Then combine these (combined) procedures over all possible choices of I (there are $\binom{M}{m^*}$ many such I altogether) with uniform weight. Let δ^* denote this final procedure and let $\Delta_I = \{\delta_i, i \in I\}$. Applying Proposition 1 twice, we have that

$$\begin{aligned} R(f; n; \delta^*) &\leq C\left(R^*\left(f; \frac{n}{4}; \Delta\right) + \frac{\left(\log n\right)^{1/2}}{n^{1/2}} + \frac{m^* \log(n/m^*)}{n} + \frac{\log\binom{M}{m^*}}{n}\right) \\ &\leq C^{'}\left(R^*\left(f; \frac{n}{4}; \Delta\right) + \frac{\log M}{\sqrt{n \log n}}\right), \end{aligned}$$

where the constants C and $C^{'}$ depend on A, $\overline{\sigma}$, $\underline{\sigma}$, and h. This completes the proof of Theorem 1.

REMARK: In the above derivation, when $M > \sqrt{n}$, combining a small number (relative to M) of procedures together with subset search yields the price of order $\sqrt{\log n/n}$ for M of a polynomial order in n, which is the optimal rate based on Theorem 2 when M is of a higher order than $\sqrt{n}$. Similar ideas

on sparse subset selection are in e.g., Barron (1994), Yang and Barron (1998) and Barron, Birgé and Massart (1999).

We need a lemma on minimax lower bound for the proof of Theorem 2. Let d be a distance (metric) on a space S. For $D \subset S$, we say G is an ϵ-packing set in D ($\epsilon > 0$) if $G \subset D$ and any two distinct members in G are more than ϵ apart in the distance d. Now let F be a class of regression functions. The distance d here is the L_2 distance.

DEFINITION 1: (*Global metric entropy*) The packing ϵ-entropy of F is the logarithm of the largest ϵ-packing set in F. The packing ϵ-entropy of F is denoted $M(\epsilon)$.

DEFINITION 2: (*Local metric entropy*) The local ϵ-entropy at $f \in \mathsf{F}$ is the logarithm of the largest $(\epsilon/2)$-packing set in $B(f,\epsilon) = \{f^{'} \in \mathsf{F} : \| f^{'} - f \| \leq \epsilon\}$. The local ϵ-entropy at f is denoted by $M(\epsilon \mid f)$. The local ϵ-entropy of F is defined as $M^{\rm loc}(\epsilon) = \max_{f \in \mathsf{F}} M(\epsilon \mid f)$.

Both global and local entropies will be involved in our derivations of the lower bounds. Assume that $M^{\rm loc}(\epsilon)$ is lower bounded by $\underline{M}^{\rm loc}(\epsilon)$. Let

$$\underline{M}^{\rm loc}(\epsilon_n) = n\epsilon_n^2 + 2\log 2.$$

Assume $M(\epsilon)$ is upper bounded by $\overline{M}(\epsilon)$ and lower bounded by $\underline{M}(\epsilon)$. Let $\overline{\epsilon}_n$ be determined by

$$\overline{M}(\sqrt{2}\overline{\epsilon}_n) = n\overline{\epsilon}_n^2 \tag{16}$$

and $\underline{\epsilon}_n$ be determined by

$$\underline{M}(\underline{\epsilon}_n) = 4n\overline{\epsilon}_n^2 + 2\log 2. \tag{17}$$

Assume the random errors in the regression model are normally distributed with variance 1. The following lemma is useful for deriving minimax lower bounds using either global or local metric entropy.

LEMMA 1: *The minimax risk for estimating f in F is lower bounded as follows:*

$$\min_{\widehat{f}} \max_{f \in \mathsf{F}} E \parallel f - \widehat{f} \parallel^2 \geq \frac{\epsilon_n^2}{32},$$

$$\min_{\widehat{f}} \max_{f \in \mathsf{F}} E \parallel f - \widehat{f} \parallel^2 \geq \frac{\underline{\epsilon}_n^2}{8},$$

where the minimization (or infimum) is over all regression estimators based on $Z^n = (X_i, Y_i)_{i=1}^n$.

The first bound in the lemma is from Yang and Barron (1999, Section 7) and the second one is from Yang and Barron (1997, Section 4).

PROOF OF THEOREM 2: Let $\varphi_1(x)$, $\varphi_2(x)$, ... be a uniformly bounded orthonormal basis (with respect to the distribution of X). An example is the trigonometric basis on [0,1]. Take δ_i, $i \geq 1$ to be the procedure that always estimate f by $\varphi_i(x)$. For each $M = C_0 n^\tau$, consider the class of regression

functions $\mathsf{F} = \{f_\theta(x) = \theta_1\varphi_1(x) + ... + \theta_M\varphi_M(x) :\| \theta \|_1^M \leq 1\}$. It is obvious that $R^*(f; n; \Delta_{M_n}) = 0$ for $f \in \mathsf{F}$. Thus to prove Theorem 2, it suffices to show that $\min_{\widehat{f}} \max_{f\in\mathsf{F}} E \| f - \widehat{f} \|^2 \geq C\gamma(n)$ for some constant $C > 0$ not depending on n, where $\gamma(n) = (\log n/n)^{1/2}$ for $1/2 < \tau < \infty$ and $\gamma(n) = n^{-(1-\tau)}$ for $0 \leq \tau \leq 1/2$. Note that under the orthonormality assumption on the basis functions, the L_2 distance on F is the same as the l_2 distance on the coefficients $\Theta = \{\theta :\| \theta \|_1^M \leq 1\}$. Thus the entropy of F under the L_2 distance is the same as the that of Θ under the l_2^M distance. To apply Lemma 1, we lower bound the local entropy of F or Θ. Note that by Cauchy-Schwartz inequality, the l_1^M and l_2^M norms have the relationship: $\| \theta \|_1^M \leq \sqrt{M} \| \theta \|_2^M$. Thus for $\epsilon \leq M^{-1/2}$, taking $f \equiv 0$, we have

$$B(f, \epsilon) = \{f_\theta \in \mathsf{F} :\| f_\theta \| \leq \epsilon\} = \{f_\theta :\| \theta \|_1^M \leq 1, \| \theta \|_2^M \leq \epsilon\} = \{f_\theta :\| \theta \|_2^M \leq \epsilon\}.$$

Consequently, for $\epsilon \leq M^{-1/2}$, the $(\epsilon/2)$-packing of $B(f, \epsilon)$ under the L_2 distance is equivalent to the $(\epsilon/2)$-packing of $B_\epsilon = \{\theta :\| \theta \|_2^M \leq \epsilon\}$ under the l_2^M distance. Since a maximum $(\epsilon/2)$-packing set is an $(\epsilon/2)$-covering set, the union of the balls with radius $\epsilon/2$ and centered at points in a maximum packing set in B_ϵ should cover B_ϵ. It follows that the size of the maximum packing set is at least the ratio of volumes of the balls B_ϵ and $B_{\epsilon/2}$, which is 2^M. Thus we have shown that the local entropy $M^{\text{loc}}(\epsilon)$ of F under the L_2 distance is at least $\underline{M}^{\text{loc}}(\epsilon) = M \log 2$ for $\epsilon \leq M^{-1/2}$. For $M = C_0 n^\tau$ for some $0 \leq \tau \leq 1/2$, solving $\underline{M}^{\text{loc}}(\epsilon_n) = n\epsilon_n^2 + 2\log 2$ gives ϵ_n of order $n^{-(1-\tau)/2}$. Note that for such τ, by possibly reducing $\underline{M}^{\text{loc}}(\epsilon)$ by a constant factor, ϵ_n obtained this way can be made smaller than $M^{-1/2}$ (as required in the earlier derivation). By Lemma 1, we have proved the minimax lower rates for F when $0 \leq \tau \leq 1/2$. That is,

$$\min_{\widehat{f}} \max_{f\in\mathsf{F}} E \| f - \widehat{f} \|^2 \geq \underline{C}_1 n^{-(1-\tau)}$$

for some constant $\underline{C}_1$ independent of n. For $\tau > 1/2$, we use the global entropy to derive the minimax lower bound. It is known form Schütt (1984) that the entropy number satisfies

$$c_1\sqrt{\frac{\log(1 + M/k)}{k}} \leq \epsilon_k \leq c_2\sqrt{\frac{\log(1 + M/k)}{k}}$$

for some constants c_1 and c_2 independent of M and k when $\log M \leq k \leq M$. We can choose $\underline{\epsilon}_n$ and $\overline{\epsilon}_n$ both of order $(\log n/n)^{1/4}$ to satisfy (16) and (17). This gives the minimax lower rate for F when $\tau > 1/2$, i.e.,

$$\min_{\widehat{f}} \max_{f\in\mathsf{F}} E \| f - \widehat{f} \|^2 \geq \underline{C}_2 (\log n/n)^{1/2}$$

for some constant $\underline{C}_2$ independent of n. Finally, with the trigonometric basis, the functions in F satisfies $\| f \|_\infty \leq \sqrt{2}$. The conclusion of Theorem 2 follows. This completes the proof of Theorem 2.

REMARKS:

1. It is interesting to note that both the global and the local entropies are useful here for different cases. For $\tau > 1/2$, the application of global entropy gives the right rate of convergence. However, if one intends to use the minimax lower bound in terms of the local entropy, the above derivation of a local entropy bound does not work because for the critical ϵ of order $(\log n/n)^{1/4}$, it is of a higher order than $M^{-1/2}$ and accordingly $B(f,\epsilon) \neq \{f_\theta : \| \theta \|_2^M \leq \epsilon\}$. On the other hand, for $0 \leq \tau \leq 1/2$, the application of the local entropy method gives a rate that agrees with the upper bound up to a logarithmic factor. If one uses the global entropy, the lower bound by Lemma 1 differs substantially in rate from the upper bound. For general relationship between global and local entropies, see Yang and Barron (1999, Section 7).

2. In the derivation of the lower bounds in Theorem 2, we choose very special (nonrandom) original estimators. This is of course not a typical situation when one would consider combining estimation procedures. In applications, the candidate estimators (or many of them) are most likely somewhat highly correlated (they are estimating the same target), but probably not too highly correlated (otherwise one can gain little even by ideal combining). For such cases, the actual price paid by a good aggregation method is smaller than that given in Theorem 2, but probably not too much smaller.

Proof of Corollary 2: Assume that Condition B0 is satisfied. For a given j, the approximation error of $f_j(x_j)$ using the best first N terms satisfies

$$\eta_{j,N}(f_j) = \| f_j - \sum_{l=1}^{N} \theta_{j,l}\varphi_{j,l} \|^2 = \sum_{i=N+1}^{\infty} \theta_{j,i}^2 \leq \sum_{i=N+1}^{\infty} \frac{i^{2s}\theta_{j,i}^2}{(N+1)^{2s}} \leq \frac{1}{(N+1)^{2s}} \sum_{i=1}^{\infty} i^{2s}\theta_{j,i}^2.$$

Thus under Condition B0 on f, we have $\eta_{j,N}(f_j) = O\left((N+1)^{-2s}\right)$ as $N \to \infty$. The approximation error of $f(x)$ using the basis functions $\varphi_{j,l}(x_j)$ with $1 \leq j \leq L$ and $1 \leq l \leq N$ satisfies

$$\begin{aligned} \eta_N^L(f) &= \| f - \sum_{j=1}^{L}\sum_{i=1}^{N} \theta_{j,i}\varphi_{j,i} \|^2 = \sum_{j=L+1}^{\infty}\sum_{i=1}^{\infty} \theta_{j,i}^2 + \sum_{j=1}^{L}\sum_{i=N+1}^{\infty} \theta_{j,i}^2 \\ &\leq \frac{1}{(L+1)^{2\beta}} \sum_{j=L+1}^{\infty} j^{2\beta} \sum_{i=1}^{\infty} i^{2s}\theta_{j,i}^2 + \frac{1}{(N+1)^{2s}} \sum_{j=1}^{L} j^{2\beta} \sum_{i=N+1}^{\infty} i^{2s}\theta_{j,i}^2. \end{aligned}$$

Thus the approximation error is $\eta_N^L(f) = O\left((N+1)^{-2s} + (L+1)^{-2\beta}\right)$.

Under Conditions B0 and B2, sparse approximation has the potential to perform much better. From Condition B2, $\sum_{i=i_0+1}^{\infty} |\theta_{j,i}|^2 = 0$. Let $\rho_j = \sum_{i=1}^{i_0} |\theta_{j,i}|$. Then Condition B2 implies that there exists a constant a such that $\| \sum_{j=1}^{L} \sum_{i=1}^{i_0} \theta_{j,i}\varphi_{j,l} \| \leq \sum_{j=1}^{L} \rho_j \leq a$ for all $L \geq 1$. From Lemma 1 in Barron (1993), there is a subset, $S \subset \{1, 2, ..., L\}$ of size k such that $\| \sum_{j=1}^{L} \sum_{i=1}^{i_0} \theta_{j,i}\varphi_{j,l} - \sum_{j \in S} \sum_{i=1}^{i_0} \theta_{j,i}\varphi_{j,l} \|^2 \leq Ck^{-1}$ for some constant $C > 0$. Thus taking $N = i_0$, the overall approximation error is upper bounded in order by

$$\eta_N^{L,k}(f) = O\left((L+1)^{-2\beta} + k^{-1}\right).$$

From (5), (6), and (7) and above, we have that under Conditions B0 and B1, with the choice of N of order $n^{1/(2s+1)}$,

$$R_1^*(f;n;\Delta) = O\left(\inf_j\left((N+1)^{-2s} + \frac{N}{n}\right)\right) = O\left(n^{-\frac{2s}{1+2s}}\right);$$

and under Condition B0, with the choice of N of order $n^{\frac{1}{1+s(2+1/\beta)}}$ and L of order $n^{\frac{s/\beta}{1+s(2+1/\beta)}}$,

$$R_2^*(f;n;\Delta) = O\left(\inf_{L,N}\left((N+1)^{-2s} + (L+1)^{-2\beta} + \frac{LN}{n} + \psi_n(L)\right)\right) = O\left(n^{-\frac{2s}{1+s(2+1/\beta)}}\right);$$

and under Conditions B0 and B2, with the choice of k of order $\sqrt{n/\log n}$, L of order $n^{1/(4\beta)}$, and $N = i_0$,

$$\begin{aligned} & R_3^*(f;n;\Delta) \\ = \; & O\left(\inf_{L,N}\left(\inf_{1\le k\le L-1}\left((L+1)^{-2\beta} + k^{-1} + \psi_n(k) + \frac{k\log L}{n} + \frac{kN}{n}\right)\right)\right) = O(\log n/n)^{1/2}. \end{aligned}$$

The conclusions of Corollary 2 follow. This completes the proof of Corollary 2.

References

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

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

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

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

[5] Barron, A.R., Rissanen, J., and Yu, B. (1998) The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, **44**, 2743-2760.

[6] Bates, J.M., and Granger, C.W.J. (1969) The combination of forecasts. *Operational Research Quarterly*, **20**, 451-468.

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

[8] Breiman, L. (1996) Stacked regressions. *Machine Learning*, **24**, 49-64.

[9] Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1995) Model selection: An integral part of inference. *Biometrics*, **53**, 603-618.

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

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

[12] Cesa-Bianchi, N. and Lugosi, G. (1999) On prediction of individual sequences. Accepted by *Ann. Statistics.*

[13] Clemen, R.T. (1989) Combining forecasts: a review and annotated bibliography. *Intl. J. Forecast.*, **5**, 559-583.

[14] Donoho, D.L. and Johnstone, I.M. (1994) Ideal denoising in an orthonormal basis chosen from alibrary of bases. *C. R. Acad. Sci. Paris,* **319**, 1317-1322.

[15] Donoho, D.L. and Johnstone, I.M. (1998) Minimax estimation via wavelet shrinkage. *Ann. Statistics.*, **26**, 879-921.

[16] Edmunds, D.E. and Triebel, H. (1989) Entropy numbers and approximation numbers in function spaces. *Proc. London Math. Soc.*, **58**, 137-152.

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

[18] Kolmogorov, A.N. and Tihomirov, V.M. (1959) ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspehi Mat. Nauk* **14**, 3-86.

[19] Merhav, N. and Feder, M. (1998) Universal prediction. *IEEE Trans. on Information Theory*, **44** 2124-2147.

[20] LeBlanc, M. and Tibshirani, R (1996) Combining estimates in regression and classification. *J. Amer. Statist. Asso.*, **91**, 1641-1650.

[21] Littlestone, N. and Warmuth, M.K. (1994) The weighted majority algorithm. *Information and Computation* **108**, 212-261.

[22] Mallat, S.G. and Zhang, Z. (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, **41** 3397-3415.

[23] Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

[24] Ruppert, D., Wand, M.P., Holst, U., and Hössjer, O. (1997) Local polynomial variance-function estimation. *J. Amer. Statist. Assoc.*, **39**, 262-273.

[25] Schütt, C. (1984) Entropy numbers of diagonal operators between symmetric Banach spaces. J. Approx. Theory, **40**, 121-128.

[26] Stone, M. (1974) Cross-validatory Choice and Assessment of Statistical Predictions (with Discussion). *J. Roy. Statist. Soc., Ser.* B, **36**, 111-147.

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

[28] Wolpert, D. (1992) Stacked generalization. *Neural Networks*, **5**, 241-259.

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

[30] Yang, Y. (1998) Combining Different Procedures for Adaptive Regression. Accepted by *Journal of Multivariate Analysis*.

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

[32] Yang, Y. (1999b) Mixing strategies for density estimation. To appear in the *Ann. Statististics*.

[33] Yang, Y. (1999c) Adaptive regression by mixing. Technical Report No. 12, Department of Statistics, Iowa State University.

[34] Yang, Y. and Barron, A.R. (1997) Information-theoretic determination of minimax rates of convergence. Tech. Report #28, Department of Statistics, Iowa State University, IA.

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

[36] Yang, Y. and Barron, A.R. (1999) Information-theoretic determination of minimax rates of convergence. To appear in the *Ann. Statististics*.