Bernoulli **10**(1), 2004, 25–47

Aggregating regression procedures to improve performance

YUHONG YANG

Department of Statistics, Iowa State University, 102 Snedecor Hall, Ames IA 50011, USA. E-mail: yyang@iastate.edu

A fundamental question regarding combining procedures concerns the potential gain and how much one needs to pay for it in terms of statistical risk. Juditsky and Nemirovski considered the case where a large number of procedures are to be combined. We give upper and lower bounds for complementary cases. Under an l_1 constraint on the linear coefficients, it is shown that for pursuing the best linear combination of n^τ procedures, in terms of rate of convergence under the squared L_2 loss, one can pay a price of order $O(\log n/n^{1-\tau})$ when $0 < \tau < \frac{1}{2}$ and a price of order $O((\log n/n)^{1/2})$ when $\frac{1}{2} \leqslant \tau < \infty$. These rates cannot 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 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.

Keywords: aggregating procedures; adaptive estimation; linear combining; nonparametric regression

1. Introduction

New ideas on combining different procedures for estimation, coding, forecasting and learning have recently been considered in statistics and several related fields, leading to a number of very interesting results. The common theme behind these ideas 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, for example, 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 on the generation of the data. In the field of forecasting, combined forecasts have been shown to work better in various examples; see Clemen (1989) for a review of work in that direction. In information theory, the 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 *et al*. (1998) for reviews. In statistics, several methods have recently been proposed for linearly combining regression estimators. These include a model selection criterion based method by Buckland *et al*. (1997), cross-validation based 'stacking' by Wolpert (1992) and Breiman (1996) (an earlier

1350–7265

version is in Stone 1974), a bootstrap based method by LeBlanc and Tibshirani (1996), a stochastic approximation based method by Juditsky and Nemirovski (2000), and information-theoretic based methods to combine density and regression estimators by Yang (2000a; 2000b; 2001) – see also Catoni (1997, 1999) – using an idea of Barron (1987). 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 (2000b; 2001) 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 works in statistics are pulling 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 might be too conservative. However, since the best coefficients are unknown, one may need to pay a 'price' for it in terms of statistical risk.

Suppose that we have M_n candidate regression procedures and consider the squared L_2 risk as a performance measure in estimating the regression function. In Yang (2000b; 2001) it is shown that a suitable data-dependent convex combination of these procedures results in an estimator that (under certain conditions) has a risk within a multiple of the smallest risk among the candidate procedures plus a small penalty of order $(\log M_n)/n$. Thus, in terms of rate of convergence, with M_n candidate procedures to be combined, one only needs to pay the price basically of order $(\log M_n)/n$ for performing nearly as well as the best candidate procedure (which, of course, is unknown to the statistician). As long as M_n does not increase exponentially fast in n, the discrepancy $(\log M_n)/n$ 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 constraint that the l_1 norm of the linear coefficients is bounded above by 1, Juditsky and Nemirovski (2000) proposed algorithms and showed that with M_n estimators to be combined, the aggregated estimator has a risk within a multiple of $((\log M_n)/n)^{1/2}$ of the smallest risk over all the linear combinations of the estimators. Furthermore, they showed that, in general, this order $((\log M_n)/n)^{1/2}$ cannot be overcome uniformly by any combining methods. Thus, compared to combining for attaining the best performance, one has to pay a much higher price, $((\log M_n)/n)^{1/2}$, for searching for the best linear combination of the original procedures.

The work of Juditsky and Nemirovski (2000) is targeted at the case where M_n is large; for example, their results are applied to restore a certain neural network class with M_n of a polynomial order in n. They derived the above-mentioned lower bound when M_n and n have the relationship $C_1 \log M_n \leqslant n \leqslant C_2 M_n \log M_n$ (where the constants C_1 and C_2 depend on

the variance of the error and the upper bound, assumed known, on the supremum norm of the regression function f). The relationship implies that M_n is of order at least $n/\log n$. It is unclear what happens when M_n is of a smaller order. For such a case, the order $((\log M_n)/n)^{1/2}$ may no longer be a valid lower bound. In the extreme case with M_n fixed (M_n does not grow as $n \to \infty$), one would expect a penalty of order close to the parametric rate n^{-1} instead of order $n^{-1/2}$. In this paper, we show that when M_n is of order n^τ, one only needs to pay a price of order $\log n/n^{1-\tau}$ for $0 \leqslant \tau < \frac{1}{2}$, and of order $(\log n/n)^{1/2}$ for $\tau \geqslant \frac{1}{2}$. The rate cannot be improved uniformly beyond a logarithmic factor for the first case, and cannot be improved for the second one. Thus the rate $((\log M_n)/n)^{1/2}$ given by Juditsky and Nemirovski (2000) is still optimal as long as M_n is of order $(n)^{1/2}$ or higher.

Note that the order of the penalty increases dramatically as τ increases from 0, but after $\tau \geqslant \frac{1}{2}$ it stays at the rate $(\log n/n)^{1/2}$ as long as $\tau < \infty$. In fact, under the l_1 constraint on the linear coefficients, there cannot be too many (relative to M_n) large coefficients and combining sparsely selected procedures with suitably large coefficients achieves the optimal performance (see the proof of Theorem 1 and the Remark 7 in the next section for details). This phenomenon is closely related to the advantage of sparse approximations as observed in wavelet estimation (see Donoho and Johnstone 1998), neural networks and subset selection (Barron 1994; Yang and Barron 1998; Barron *et al.*, 1999).

In applications, one does not know if the best linear combination can substantially improve the estimation accuracy so that the high price of order, for example, $(\log n)/n^{1/2}$ is worthwhile. Accordingly, it is not clear which direction to take when combining the candidate procedures. Fortunately, we show, that, 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_n 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. Some proofs of the results are given in Section 6.

2. Risk bounds on linear aggregation

Consider the regression model

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

where $(X_i, Y_i)_{i=1}^n$ are independent and identically distributed copies from the joint distribution of (X, Y) with $Y = f(X) + \sigma \cdot \varepsilon$. The explanatory variable X (which could be high-dimensional) has an unknown distribution P_X unless otherwise stated. The variance parameter $\sigma^2 > 0$ 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. We further assume that X and ε are independent. 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 an estimator $\hat{f}_i(x) = \hat{f}_i(x; Z^i)$ for each sample size $i \geq 1$. Let $\|\cdot\|$ denote the L_2 norm with respect to the distribution of X, that is, $\|g\| = \sqrt{\int g^2(x) P_X(\mathrm{d}x)}$. Let $R(f; n; \delta) = \mathrm{E}\|f - \hat{f}_n\|^2$ denote the risk of the procedure δ at the sample size n under the squared L_2 loss.

Our strategy of linearly combining a list of procedures depends on the following method of combining for adaptation. It serves as a building block for the main 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.

2.1. A three-stage algorithm to combine procedures for adaptation

Let $\Delta = \{\delta_j, j \geq 1\}$ be a collection of regression procedures, and let $\hat{f}_{j,i}(x) = \hat{f}_{j,i}(x; Z^i)$ denote the estimator of f based on δ_j given the observations Z^i, for $i \geq 1$. The index set $\{j \geq 1\}$ is allowed to degenerate to a finite set. Let π_j be positive numbers summing to one, $\sum_{j=1}^{\infty} \pi_j = 1$. These will be used as prior weights on the procedures. The following algorithm, called 'adaptive regression by mixing' (ARM), for combining candidate procedures for adaptation, is essentially as given in Yang (2001).

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 each procedure δ_j.

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

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

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

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

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

$$\tilde{f}_n(x) = \sum_{j=1}^{\infty} \overline{W}_j \hat{f}_{j,n_1}(x). \tag{1}$$

The combined estimator has the theoretical property given in Proposition 1 below, under the follwing conditons:

A1. The regression function and the estimators are uniformly bounded: there exists a constant $A > 0$ such that $\|f\|_\infty \leqslant A$ and $\|\hat{f}_{j,i}\|_\infty \leqslant A$ with probability one for all i, j.

A2. The variance parameter σ is bounded above and below by known positive constants $\overline{\sigma} < \infty$ and $\underline{\sigma} > 0$, respectively.

A3. The known error distribution h has a finite fourth moment and is such that, for each pair $0 < s_0 < 1$ and $T > 0$, there exists a constant $B_{s_0,T}$ (depending on s_0 and T) such that

$$\int h(x) \log \frac{h(x)}{s^{-1} h((x-t)s^{-1})} \mathrm{d}x \leqslant B_{s_0,T}((1-s)^2 + t^2),$$

for all $s_0 \leqslant s \leqslant s_0^{-1}$ 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 process, though knowledge of A may be needed to ensure that the $\|\hat{f}_{j,i}\|_\infty$ are uniformly bounded from above. Assumption A3 is satisfied by Gaussian, double exponential, t (with degrees of freedom greater than 2) and many other smooth distributions supported on the whole real line. For any distribution with a compact support, however, the assumption cannot be satisfied directly. For such a case, as far as the rate of convergence for regression estimation is concerned, one can artificially add a weak Gaussian noise to the response and the assumption may become satisfied.

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 estimators $\hat{\sigma}_j$ in step 3 are bounded above and below by the constants $\overline{\sigma}$ and $\underline{\sigma}$ (otherwise one needs to clip the estimator to be in that range).

Proposition 1. *Assume conditions* A1–A3 *hold. Then the above convexly combined estimator* $\tilde{f}_n$ *satisfies*

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

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

$$\mathrm{E}\|f - \tilde{f}_n\|^2 \leqslant C \left(\frac{\sigma^2 \log M_n}{n} + \inf_j \mathrm{E}\|f - \hat{f}_{j,n/2}\|^2 \right).$$

Remark 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, for example, 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.

Remark 2. As discussed in Yang (2001), the estimator $\tilde{f}_n$ depends on the order of observations. For an improvement, one can randomly permute the order of observations a number of times and average the corresponding estimators.

Remark 3. In the definition of the final estimator $\tilde{f}_n(x) = \sum_{j=1}^{\infty} \overline{W}_j \hat{f}_{j,n/2}(x)$, 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.

Remark 4. The constant C can be taken to be proportional to $(2(1 + \overline{\sigma}^2/\underline{\sigma}^2) + 13A^2)B_{s_0,T}$, where $s_0 = \underline{\sigma}/\overline{\sigma}$ and $T = A$.

Proof of Proposition 1. The result is proved in Yang (2001) for the case where there are finitely many, say J, candidate procedures with equal prior weight $\pi_j = 1/J$ for $1 \leqslant j \leqslant J$. The proof for the general case can be done similarly.

2.2. Linearly combining a finite number of procedures

Now let $\Delta = \{\delta_1, \delta_2, \ldots, \delta_{M_n}\}$ denote a finite collection of candidate procedures to be aggregated. The number of procedures, M_n, changes according to the sample size n. In particular, we will consider the case where M_n is of order n^τ for some $0 \leqslant \tau < \infty$. When the sample size increases, one is allowed to consider more candidate procedures (possibly more and more complicated).

As in Juditsky and Nemirovski (2000), the coefficients for linear combination are suitably constrained. Let $\mathbf{F}_n = \{\sum_{1 \leqslant j \leqslant M_n} \theta_j \hat{f}_{j,n}(x) : \sum_{1 \leqslant j \leqslant M_n} |\theta_j| \leqslant 1\}$ be the collection of linear combinations of the original estimators in Δ with coefficients whose absolute values sum to no more than 1. 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 (see Bates and Granger 1969; Breiman 1996). Let $\|\cdot\|_1^{M_n}$ denote the l_1 norm on R^{M_n}, that is, $\|\theta\|_1^{M_n} = \sum_{1 \leqslant j \leqslant M_n} |\theta_j|$. Define

$$R^*(f; n; \Delta) = \inf_{\|\theta\|_1^{M_n} \leqslant 1} \mathrm{E} \left\| f - \sum_{1 \leqslant j \leqslant M_n} \theta_j \hat{f}_{j,n} \right\|^2.$$

This is the smallest risk over all the estimators in the linear aggregation class $\mathbf{F}_n$. Obviously, $R^*(f; n; \Delta) \leqslant \inf_{1 \leqslant j \leqslant M_n} R(f; n; \delta_j)$.

Let us now describe our strategy of linear combining. There are two main steps. 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 a suitable discretization, some results on metric entropy are very helpful. In the second step, we combine these estimators for adaptation using the ARM algorithm described in the previous subsection. When M_n is large, however, an additional difficulty arises and sparse combining solves the problem.

We consider first the case where $M_n < (n)^{1/2}$. Let $G = \{\theta = (\theta_1, \ldots, \theta_{M_n}) : \sum_{i=1}^{M_n} |\theta_i| \leqslant 1\}$ be the M_n-dimensional unit ball under the $l_1^{M_n}$ distance. Let N_ϵ be an ϵ-net in G under the $l_1^{M_n}$ distance, that is, for each $\theta \in G$, there exists $\theta' \in N_\epsilon$ such that $\|\theta - \theta'\|_1^{M_n} = \sum_{i=1}^{M_n} |\theta_i - \theta_i'| \leqslant \epsilon$. We choose a best ϵ_n-net of size of 2^k points (which minimizes the maximum approximation error at the given size), where k is chosen so that

$$k = \left\lceil \frac{M_n(\log(n/M_n) + 2\log 2)}{2\log 2} \right\rceil.$$

For simplicity in notation, let $\hat{f}^1, \ldots, \hat{f}^{M_n}$ denote the original estimators at the sample size n. Let F_{ϵ_n} be the set of the linear combinations of the estimators $\hat{f}^1, \ldots, \hat{f}^{M_n}$ with coefficients in N_{ϵ_n}. Then we combine all the estimators in F_{ϵ_n} using the ARM algorithm with uniform prior weight $1/|N_{\epsilon_n}|$. Let $\hat{f}_n$ denote the combined estimator and δ^* denote this final procedure.

Now consider the other case: $M_n \geqslant (n)^{1/2}$. It turns out that the method above leads to a suboptimal rate of convergence. For this case, due to the l_1 constraint, the number of large coefficients is small relative to M_n when $M_n \gg (n)^{1/2}$. An appropriate search of the large coefficients can result in optimal rate of convergence, as will be seen.

For each fixed subset $I \subset \{1, \ldots, M_n\}$ of size $m^* = \lceil (n/\log n)^{1/2} \rceil$, consider a best ϵ-net in $B_I = \{\theta_I : \sum_{i \in I} |\theta_i| \leqslant 1\}$ under the $l_1^{m^*}$ distance with size of 2^k points, where $k = \lceil (m^*(\log(n/m^*) + 2\log 2))/2\log 2 \rceil$. Then (with uniform prior weight) combine the corresponding linear combinations of the procedures in I. Then combine these (combined) procedures over all possible choices of I – there are $\binom{M_n}{m^*}$ many such I altogether – with uniform prior weight. Let δ^* denote this final procedure.

2.3. An upper bound for linear combining

Theorem 1. *Assume that conditions* A1–A3 *are satisfied. For any given collection of estimation procedures* $\Delta = \{\delta_j, 1 \leqslant j \leqslant M_n\}$, *the combined procedure* δ^* *constructed in the previous subsection satisfies*

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

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

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

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

Remark 5. The technical condition A2 is used for deriving a risk bound for the built-in variance estimation in step 3 of the three-stage combining algorithm in Section 2.1. It can be dropped if one has a good alternative estimator available (e.g., by the nearest-neighbour method), and then the risk of the estimator appears in the performance bound; see Yang (2000b, Section 6) for details. Under mild continuity assumptions, it typically does not affect the rate of convergence of the combined procedure.

Remark 6. The discretization-based combined estimator is computationally very costly. Thus the combining method is difficult to implement for applications.

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 $R^*(f;\, n/2;\, \Delta)$ converge at the same rate. From the result, when $\tau \geqslant \frac{1}{2}$, 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 (2000) with a weaker assumption on the errors (finite variance), while requiring the knowledge of A. When $\tau < \frac{1}{2}$, 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 $\log n/n$.

Proof of Theorem 1. Consider first the case where $M_n < (n)^{1/2}$. Note that an ϵ-net in G yields a suitable net in the set $\mathbf{F}_n$ of the linear combinations of the original estimators: for any estimator $\hat{f} = \sum_{i=1}^{M_n} \theta_i \hat{f}^i$ with $\theta \in G$, there exists $\theta' \in N_\epsilon$ such that

$$\left\| \hat{f} - \sum_{i=1}^{M_n} \theta_i' \hat{f}^i \right\| = \left\| \sum_{i=1}^{M_n} (\theta_i - \theta_i') \hat{f}^i \right\| \leqslant A \|\theta - \theta'\|_1^{M_n} \leqslant A\epsilon. \tag{3}$$

By Proposition 1, for any f with $\|f\|_\infty \leqslant A$, we have

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

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 the triangle inequality, for any f, we have $\inf_{\hat{f} \in F_\epsilon} R(f;\, \hat{f};\, n/2) \leqslant 2 \inf_{\hat{f} \in \mathbf{F}_n} R(f;\, \hat{f};\, n/2) + 2A^2\epsilon^2$. It follows that

$$\mathrm{E}\|f-\hat{f}_n\|^2 \leqslant \frac{C\log(|N_\epsilon|)}{n} + 2C \inf_{\hat{f}\in\mathbf{F}_n} R(f;\hat{f};n/2) + 2A^2C\epsilon^2. \tag{4}$$

To obtain the best upper bound (in order), we need to minimize $\log(|N_\epsilon|)/n + 2A^2\epsilon^2$ when discretizing G. Note that the logarithm of the smallest size of N_ϵ is the covering entropy of the set G under the $l_1^{M_n}$ distance; see Kolmogorov and Tikhomirov (1959) for properties of metric entropy. For this case, the metric entropy is easy to compute. The following result is given in terms of the entropy number, that is, the worst-case approximation error with the best net of size 2^k points. Let ϵ_k denote the entropy number of G. From Edmunds and Triebel (1989, Proposition 3.1.3), when $k \geqslant M_n$, we know that $\epsilon_k \leqslant c2^{-k/M_n}$ for some constant c independent of k and M_n (note that the results of Edmunds and Triebel are much more general than what is needed here, and they can be useful for considering linear combining under other l_p $(p \neq 1)$ constraints). Take

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

(note that $k \geqslant M_n$). Then

$$\frac{\log(|N_\epsilon|)}{n} + 2A^2\epsilon^2 \leqslant \frac{M_n(\log(n/M_n)+2\log 2)}{2n} + \frac{\log 2}{n} + \frac{(Ac)^2 M_n}{2n} \leqslant \frac{c' M_n \log(1+n/M_n)}{n},$$

where c' depends only on A and c. The upper bound in Theorem 1 for $M_n < (n)^{1/2}$ then follows.

Now consider the other case: $M_n \geqslant (n)^{1/2}$. Note that for $\|\theta\|_1^{M_n} \leqslant 1$, $\|\sum_{i=1}^{M_n}\theta_i\hat{f}^i\| \leqslant A$. Then by a sampling argument (see Lemma 1 in Barron 1993), for each m, there exist a subset $I \subset \{1,\ldots,M_n\}$ of size m and $\theta_I' = (\theta_i', i \in I)$ such that $\|\sum_{i=1}^{M_n}\theta_i\hat{f}^i - \sum_{i\in I}\theta_i'\hat{f}^i\| \leqslant A/(m)^{1/2}$. With the choice of $m^* = \lceil (n/\log n)^{1/2}\rceil$, we have $\|\sum_{i=1}^{M_n}\theta_i\hat{f}^i - \sum_{i\in I}\theta_i'\hat{f}^i\| \leqslant A(\log n/n)^{1/4}$. Consider an ϵ-net in $B_I = \{\theta_I : \sum_{i\in I}|\theta_i| \leqslant 1\}$ under the $l_1^{m^*}$ distance. Again by Edmunds and Triebel (1989), taking $k = \lceil (m^*(\log(n/m^*) + 2\log 2))/2\log 2\rceil$, the best ϵ-net has approximation accuracy $\epsilon \leqslant c/2(m^*/n)^{1/2}$. Then as in (3), we know that there exists θ_I'' in this ϵ-net such that $\|\sum_{i\in I}\theta_i'\hat{f}^i - \sum_{i\in I}\theta_i''\hat{f}^i\| \leqslant Ac/2(m^*/n)^{1/2}$. Thus for each $\hat{f} \in \mathbf{F}_n$, there exist $I^* \subset \{1,\ldots,M_n\}$ of size m^* and θ_I'' such that

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

where c'' depends only on A and c. Notice that, in general, I^* depends on f and therefore it should be sought adaptively. Applying Proposition 1 twice, we have that

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

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

Remark 7. In the above derivation, when $M_n > (n)^{1/2}$, combining a small number (relative to M_n) of procedures together with subset search yields a price of order $(\log n/n)^{1/2}$ for M_n of a polynomial order in n, which is the optimal rate based on Theorem 2 (to be given in the next subsection) in that case. Similar ideas on sparse subset selection are given in Barron (1994), Yang and Barron (1998) and Barron *et al*. (1999).

2.4. A lower bound for linear combining

How good are the upper bounds derived in the previous subsection? Juditsky and Nemirovski (2000) show that when M_n and n satisfy

$$C_1 \log M_n \leqslant n \leqslant C_2 M_n \log M_n \tag{5}$$

for some constants C_1 and C_2 (i.e., M_n is no smaller than order $n/\log n$ but not too large), the order $(\log n/n)^{1/2}$ cannot be improved in a minimax sense. We show in general that the rates given in Theorem 1 cannot be improved up to possibly a logarithmic factor for some cases. In particular, the lower rate $(\log n/n)^{1/2}$ derived by Juditsky and Nemirovski (2000) for the case (5) continues to hold as long as M_n is of order at least $(n)^{1/2}$.

For simplicity, consider the case where $X_1, X_2, \ldots$ are independent and uniformly distributed on [0, 1]. Let $\{\varphi_i(x), i \geqslant 1\}$ be the orthonormal trigonometric basis functions on [0, 1]. Take δ_j, $j \geqslant 1$ to be the procedure that always estimates f by $\varphi_j(x)$.

Theorem 2. *Assume that the errors are normally distributed with variance 1. Consider* $M_n = \lfloor C_0 n^\tau \rfloor$ *for some* $\tau > 0$ *and* $C_0 > 0$. *For the* M_n *procedures* $\Delta_{M_n} = \{\delta_j, 1 \leqslant j \leqslant M_n\}$, *for any aggregated procedure* $\delta^{(n)}$ *based on* Δ_{M_n}, *one can find a regression function* f *with* $\|f\|_\infty \leqslant \sqrt{2}$ *satisfying*

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

where the constant C *does not depend on* n *or* f.

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 $\tau > \frac{1}{2}$ and the upper and lower rates differ only in logarithmic factors when $0 \leqslant \tau \leqslant \frac{1}{2}$.

It is worth noting 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_n reaches $(n)^{1/2}$, it increases much more slowly in a logarithmic fashion. Accordingly, it stays at rate $(\log n/n)^{1/2}$ as long as M_n increases polynomially in n.

In a different direction, Yang (2000b; 2001) shows that one only needs to pay a penalty of order $(\log M_n)/n$ to pursue the less ambitious goal of achieving the best performance among the original M_n procedures (see also Proposition 1 above). Observing the dramatic difference between the two penalties, one naturally faces the question whether one should combine for adaptation or for improvement. If one of the original procedures happens to behave best (or close to 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 $0 \leqslant \tau < \frac{1}{2}$) or $(\log n/n)^{1/2}$ (for $\tau \geqslant \frac{1}{2}$), if one aggregates to improve performance, one could be unfortunately paying too high a price for nothing but adversely affecting the convergence rate in estimating f. In terms of rate of convergence, combining for improvement is worth the effort for certain only if $R^*(f; n/2; \Delta)$ plus the penalty in (2) is of a smaller order than $(\log M_n)/n + \inf_j R(f; n/2; \delta_j)$. In applications, since the risks are of course unknown, one does not know in advance whether to combine for adaptation or for improvement. An indiscriminate choice can lead to a much worse rate of convergence. In the next section we show that one can actually handle the two goals optimally at the same time.

Finally, we briefly mention an interesting observation in the proof of Theorem 2 (to be given in Section 6). It is well known that metric entropy plays a determining role in rate of convergence for function estimation. Both local entropy (Le Cam 1975; Birgé 1986) and global entropy (Yang and Barron 1999, and references therein) have been used for obtaining upper rates and lower rates of convergence. Here, in the proof of Theorem 2, we see the advantage of each over the other in different scenarios. See the proof of the theorem and Remark 12 in Section 6 for more details.

3. Multi-direction aggregation

We now show that, 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 in considering sophisticated linear combinations.

Let us consider a slightly different setting than that of the previous section. Suppose that we have a countable collection of candidate procedures $\Delta = \{\delta_1, \delta_2, \ldots\}$. We may combine some or all of the procedures. We consider three different approaches to combining 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 given in Section 2. Since Δ is not (necessarily) a finite collection, one cannot use the uniform prior weight for combining. The prior weight π_j is taken to be $c\mathrm{e}^{-\log^* j}$, where $\log^*$ is defined by $\log^* x = \log(x+1) + 2\log\log(x+1)$ for $x > 0$, and the constant c is chosen to normalize the weights to sum to 1. Based on Proposition 1, we have that, for any f with $\|f\|_\infty \leqslant A$,

$$R(f;\, n;\, \delta_A^*) \leqslant C_1 \inf_j \left(\frac{\log(j+1)}{n} + R(f;\, n/2;\, \delta_j) \right) =: C_1 R_1^*(f;\, n;\, \Delta), \tag{6}$$

where the constant C_1 depends on A, $\underline{\sigma}$, $\overline{\sigma}$, and h. In the rest of the paper, unless otherwise stated, a constant C (with or without subscript) may depend on A, $\underline{\sigma}$, $\overline{\sigma}$, and h. For convenience, we may use the same symbol C for different such constants in different places. From (6), if one procedure, say δ_{j^*} behaves best, then the penalty is of order n^{-1}. If the best estimator changes according to n, then $\inf_j((\log(j+1))/n + R(f;\, n/2;\, \delta_j))$ is a trade-off between complexity and estimation accuracy.

The second approach targets the best performance among all the l_1-constrained linear combinations of the original procedures up to different orders. For each integer $L \geqslant 1$, let δ^L denote the combined (for improvement) procedure based on the first L procedures $\delta_1, \ldots, \delta_L$ as used for Theorem 1. Then combine (for adaptation) the procedures $\{\delta^1, \delta^2, \ldots\}$ with prior weight $c\mathrm{e}^{-\log^* L}$ for $L \geqslant 1$. Let δ_B^* denote this combined procedure. Let Δ_L denote the set of the first L procedures in Δ. Let

$$\psi_n(L) = \begin{cases} \dfrac{L\log(1+n/L)}{n} & 1 \leqslant L < \sqrt{n}, \\ \dfrac{\log L}{\sqrt{n\log n}} & L \geqslant \sqrt{n}. \end{cases}$$

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

$$R(f;\, n;\, \delta_B^*) \leqslant C_2 \inf_L \left(R^*\left(f; \frac{n}{2}; \Delta_L\right) + \psi_n(L) \right) =: C_2 R_2^*(f;\, n;\, \Delta). \tag{7}$$

The third approach recognizes that in many cases, when combining a lot of procedures, the best linear combination may concentrate on only a few of them. For such a case, working with these important procedures only leads to a much smaller penalty when combining for improvement. This calls for additional care in aggregation and it can be done as follows. For each integer $L > 1$, $1 \leqslant k < L$, and a subset S of $\{1, 2, \ldots, L\}$ of size k, let $\delta(S)$ be the combined (for improvement) procedure based on $\{\delta_j : j \in S\}$ as for Theorem 1. Then let $\delta_{L,k}$ be the combined (for adaptation) procedure based on all such $\delta(S)$ with uniform prior 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 $\delta_{L,1}, \ldots, \delta_{L,L-1}$ using the uniform prior weight $1/(L-1)$. Let δ_C^* denote the combined (for adaptation) procedure based on $\delta^{(L)}$, $L \geqslant 2$ with prior weight $c' \exp(-\log^* L)$, where the constant c' is chosen such that $\sum_{L=2}^{\infty} c'\mathrm{e}^{-\log^* L} = 1$. 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 $\|f\|_\infty \leqslant A$,

$$R(f;\, n;\, \delta_C^*) \leqslant C_3 \inf_{L \geqslant 2} \left(\inf_{1 \leqslant k \leqslant L-1} \left(\inf_{|S|=k, S \subset \{1,2,\ldots,L\}} R^*\left(f; \frac{n}{16}; \Delta_S\right) + \psi_n(k) + \frac{\log\binom{L}{k}}{n} \right) \right) \tag{8}$$

$$=: C_3 R_3^*(f;\, n;\, \Delta).$$

Now we combine these three procedures δ_A^*, δ_B^*, and δ_C^* with equal prior weight $\frac{1}{3}$ for

adaptation. 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–A3 *are satisfied. Then, for each f with $\|f\|_\infty \leqslant A$, we have*

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

where $R_1^*(f; n; \Delta)$, $R_2^*(f; n; \Delta)$ and $R_3^*(f; n; \Delta)$ are given in (6)*,* (7) *and* (8)*, respectively.*

This 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 a price of order at most $\psi_n(L_n)$ for the better performance. When L_n is small, the gain is substantial. When a certain linear combination of a small number of procedures performs 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), whichever is better.

4. Aggregating estimators based on linear approximation

In this section, we illustrate the spirit of multi-direction aggregation through an example with linear and sparse approximations. We assume that $x \in [0, 1]^d$ $(1 \leqslant d \leqslant \infty)$.

Let $\{\Phi_j : j = 1, 2, \ldots\}$ be a countable collection of linear approximation systems. For each j, $\Phi_j = \{\varphi_{j,1}(x), \varphi_{j,2}(x), \ldots\}$ is a chosen collection of linearly independent functions in $L^2[0, 1]^d$. Bases that are orthonormal (or at least have some frame properties) have traditionally been emphasized, but non-orthogonal and/or over-complete bases have recently been advocated and studied. Relaxation of orthogonality enables one to consider, for example, trigonometric expansions with fractional frequencies and neural network models. Considering different bases at the same time provides much more flexibility and gives 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), Johnstone (1999), Juditsky and Nemirovski (2000), Yang and Barron (1998), and Barron *et al*. (1999) for work in these directions.

For a fixed j, the (squared L_2) approximation error of f using the first N terms (together with a constant term if needed) is

$$\eta_{(j,N)}(f) = \inf_{\{a_l\}} \left\| f - a_0 - \sum_{l=1}^{N} a_l \varphi_{j,l} \right\|^2.$$

This uses an individual approximation system. The approximation error of f using (unrestricted) linear combinations of the first N terms of each of the first L systems is

$$\eta_{L,N}(f) = \inf_{\{a_0,a_{j,l}\}} \left\| f - a_0 - \sum_{j=1}^{L}\sum_{l=1}^{N} a_{j,l}\varphi_{j,l} \right\|^2.$$

This is a mixed linear approximation using the first L systems. The approximation error of f based on the sparse (unrestricted) linear approximation using k out of the first L systems is

$$\eta^k_{L,N}(f) = \inf_{S\subset\{1,\ldots,L\},|S|=k} \inf_{\{a_0,a_{j,l}\}} \left\| f - a_0 - \sum_{j\in S}\sum_{l=1}^{N} a_{j,l}\varphi_{j,l} \right\|^2.$$

In addition, consider the approximation error of f using the l_1 constrained linear combinations:

$$\eta^L_N(f) = \inf_{\{a_{j,l}\}} \left\| f - a_0 - \sum_{j=1}^{L}\sum_{l=1}^{N} a_{j,l}\varphi_{j,l} \right\|^2,$$

where the coefficients $a_{j,l}$ and a_0 are such that the l_1 norm is upper bounded by 1. For different functions, one of the approximations above can be advantageous over the others. With various assumptions on the approximation errors and with appropriate handling of the estimation errors, results for the adaptive estimation of f can be derived.

For simplicity in illustration, we now focus on approximations based on orthonormal basis functions, where the relationship between the approximation error and the linear coefficients is often clear.

Suppose $d = \infty$ and assume that $X_i = (X_{i1}, X_{i2}, \ldots)$ has independent, uniformly distributed components X_{ij}, $j \geqslant 1$ (or after suitable transformation). We assume that the true regression function is additive, that is, for $x = (x_1, x_2, \ldots)$,

$$f(x) = c_0 + f_1(x_1) + f_2(x_2) + \ldots, \tag{9}$$

where f_i has mean zero (with respect to the Lebesgue measure on [0, 1]) for $i \geqslant 1$ and we assume that $\|f\|_\infty \leqslant A$ for some known constant $A > 0$.

To estimate the additive component $f_j(x_j)$, a linear approximation system $\Phi_j = \{\varphi_{j,1}(x_j), \varphi_{j,2}(x_j), \ldots\}$ is used. Assume that the basis functions are orthonormal with mean zero, and in addition, $\sup_{j,l}\|\varphi_{j,l}\|_\infty \leqslant A'$ for some constant $1 < A' < \infty$.

We consider several estimators. Let $\hat{c}_0 = n^{-1}\sum_{i=1}^n Y_i$. For a given j, let $\tilde{f}^{(j,N)}(x_j)$ be the projection estimator of $f_j(x_j)$ based on the first N basis functions in Φ_j. That is, $\tilde{f}^{(j,N)}(x_j) = \sum_{l=1}^N \hat{\theta}_{j,l}\varphi_{j,l}(x_j)$, where $\hat{\theta}_{j,l} = n^{-1}\sum_{i=1}^n Y_i\varphi_{j,l}(X_{ij})$. Then let $\hat{f}^{(j,N)}(x) = \hat{c}_0 + \tilde{f}^{(j,N)}(x_j)$, clipped to $[-A, A]$ if necessry. Let $\tilde{f}^{L,N}(x) = \hat{c}_0 + \sum_{j=1}^L \tilde{f}^{(j,N)}(x_j)$, also clipped to $[-A, A]$ if required. For a given L, N, and a subset $S \subset \{1, \ldots, L\}$, define $\hat{f}^{L,N,S}(x) = \hat{c}_0 + \sum_{j\in S}\tilde{f}^{(j,N)}(x_j)$, again clipped to $[-A, A]$.

Now we consider combining several estimators based on the different approximations. First we combine the estimators $\hat{f}^{(j,N)}$ over j and N for adaptation with prior weight proportional to $\mathrm{e}^{-\log j^* - \log N^*}$. Let δ_1 denote the combined procedure. Its risk is bounded from above by a multiple of

$$\inf_{j,N}\left(\eta_{(j,N)}(f)+\frac{N}{n}+\frac{\log j}{n}+\frac{\log N}{n}\right); \tag{10}$$

see the proof of Corollary 2 in Section 6.

Next, we combine $\hat{f}^{L,N}$ over L and N for adaptation with prior weight proportional to $\mathrm{e}^{-\log L^*-\log N^*}$. Let δ_2 denote the combined procedure. Then the risk of δ_2 is bounded from above by a multiple of

$$\inf_{L,N}\left(\eta_{L,N}(f)+\frac{LN}{n}+\frac{\log L}{n}+\frac{\log N}{n}\right). \tag{11}$$

We also combine the sparse approximation based estimators $\hat{f}^{L,N,S}$ over (L, N, S) with prior weights proportional to $\mathrm{e}^{-\log L^*-\log N^*-\log(L-1)-\log\binom{L}{k}}$ $(L \geqslant 2)$. The combined procedure, δ_3, has a risk bounded from above by a multiple of

$$\inf_{L,N}\left(\inf_{1\leqslant k\leqslant L-1}\left(\eta^k_{L,N}(f)+\frac{kN}{n}+\frac{\log L}{n}+\frac{\log N}{n}+\frac{k\log L}{n}\right)\right). \tag{12}$$

For the use of the l_1-constrained combined approximation, let the basis functions themselves (together with 1) be the initial estimators and consider linearly combining them (as in Theorem 1): $a_0+\sum_{j=1}^{L}\sum_{l=1}^{N}a_{j,l}\varphi_{j,l}$, with the coefficients bounded from above by 1 in l_1 norm. Then we combine the estimators over L and N for adaptation (with prior weight proportional to $\mathrm{e}^{-\log L^*-\log N^*}$) and let δ_4 denote this combined procedure. The resulting risk bound for δ_4 is a multiple of

$$\inf_{L,N}\left(\eta^L_N(f)+\psi_n(LN+1)+\frac{\log L}{n}+\frac{\log N}{n}\right). \tag{13}$$

Let δ_F denote the final procedure combining δ_1, δ_2, δ_3, δ_4 together (for adaptation with equal prior weight). Based on the aforementioned risk bounds, one can derive rate of convergence for the final aggregated procedure δ_F under various assumptions on the approximation errors.

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

$$\sum_{j=1}^{\infty} j^{2\beta}\left(\sum_{l=1}^{\infty} l^{2s}\theta^2_{j,l}\right)<\infty, \tag{14}$$

for some $s>0$ and $\beta>0$. When the true regression function is actually univariate in one variable, say x_{j_0}, then $\theta_{j,l}=0$ for all j and l except $j=j_0$. Let B1 denote this condition. Under condition B0, we have $\eta_{L,N}(f)=O(N^{-2s}+L^{-2\beta})$; see the proof of Corollary 2 in Section 6. Let B2 denote the condition under which the earlier sparse approximation satisfies the requirement that there exist constants $0<\nu<1$ and $c>0$ such that, for any L, there exists a subset S of size $k \leqslant cL^{\nu}$ with $\eta^k_{L,N}(f)=O(N^{-2s}+L^{-2\beta})$ (i.e., a small fraction of terms can yield the same approximation error rate). Another condition, denoted B3, is that

$$|c_0| + \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} |\theta_{j,i}| \leqslant 1. \tag{15}$$

Corollary 2. *Assume that the errors are normally distributed with σ^2 bounded above and below by known constants. Assume also that $\|f\|_\infty \leqslant A$ for a known constant $A > 0$ and $\sup_{j,l} \|\varphi_{j,l}\|_\infty \leqslant A'$ for some constant $1 < A' < \infty$. If f satisfies condition* B0*, we have*

$$R(f;\, n;\, \delta_F) = O\Big(n^{-2s/(1+s(2+1/\beta))}\Big). \tag{16}$$

If f satisfies conditions B0 *and* B1*, we have*

$$R(f;\, n;\, \delta_F) = O\Big(n^{-2s/(1+2s)}\Big). \tag{17}$$

If f satisfies conditions B0 *and* B2*, we have*

$$R(f;\, n;\, \delta_F) = O\Big(n^{-2s/(1+s(2+\nu/\beta))}\Big). \tag{18}$$

If f satisfies conditions B0 *and* B3*, we have*

$$R(f;\, n;\, \delta_F) = O\Big(\min\Big((\log n/n)^{1/2},\, n^{-2s/(1+s(2+1/\beta))}\Big)\Big). \tag{19}$$

Note that the procedure δ_F does not require knowledge of the constants s and β. Thus the rate $n^{-2s/(1+s(2+1/\beta))}$ is adaptively achieved. When s or β is very small, the rate of convergence is very slow. If f is in fact univariate in one variable, a (possibly much) better rate of convergence $n^{-2s/(1+2s)}$ is automatically achieved by the aggregated procedure (the same rate also holds (under B0) if f depends on x only in a finite number of variables). The sparse approximation helps when ν is small in condition B2. Under B0 and B3, a good rate $O((\log n/n)^{1/2})$ is guaranteed regardless of how unfavourable s and β are.

Remark 8. In the construction of sparse linear combining, 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. See Yang and Barron (1998) for such a treatment in density estimation based on model selection.

Remark 9. For an integer j_0 and positive constants s and C, let $\mathcal{F}(j_0, s, C)$ denote the set of functions f with $\theta_{j,l} = 0$ for $j \neq j_0$ and $\sum_{l=1}^{\infty} l^{2s} \theta_{j_0,l}^2 \leqslant C$. Then for each $f \in \mathcal{F}(j_0, s, C)$, instead of the rate in (17), we in fact have that $R(f;\, n;\, \delta_F) = o(n^{-2s/(1+2s)})$. But the rate $o(n^{-2s/(1+2s)})$ cannot occur uniformly over $\mathcal{F}(j_0, s, C)$.

Remark 10. 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 $O(\log n/n)$, possibly losing a logarithmic factor.

Remark 11. With a proper modification of the combining method, condition (B3) can be relaxed to $|c_0| + \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} |\theta_{j,i}| < \infty$ without affecting the rate of convergence.

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 (2001). Firstly, the error distribution h need not be known completely. It suffices to assume that h is in a countable collection of candidate error distributions. This gives more flexibility for handling errors with different degrees of heavy-tailedness. Secondly, one need not 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 $\hat{\sigma}^2_{j,n}$ 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 2001). A recent work on variance estimation is Ruppert *et al*. (1997), where a local polynomial method is proposed with a theoretical justification.

6. Proofs of the results

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

Both global (for references, see Yang and Barron 1999) and local entropies (Le Cam 1975; Birgé 1986) have been used for deriving minimax upper and/or lower bounds. Here we focus on the lower bounds. Assume that $M^{\mathrm{loc}}(\epsilon)$ is lower-bounded by $\underline{M}^{\mathrm{loc}}(\epsilon)$, a continuous function. Let

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

Assume $M(\epsilon)$ is bounded from above by $\overline{M}(\epsilon)$ and from below by $\underline{M}(\epsilon)$, with $\overline{M}(\epsilon)$ and $\underline{M}(\epsilon)$ both being continuous. Let $\bar{\epsilon}_n$ be determined by

$$\overline{M}(\sqrt{2}\bar{\epsilon}_n) = n\bar{\epsilon}_n^2 \tag{20}$$

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

$$\underline{M}(\underline{\epsilon}_n) = 4n\bar{\epsilon}_n^2 + 2\log 2. \tag{21}$$

The following lemma is useful for deriving minimax lower bounds using either global or local metric entropy.

Lemma 1. *Assume the random errors in the regression model are normally distributed with variance* 1. *The minimax risk for estimating* f *in* F *is lower-bounded as follows:*

$$\min_{\hat{f}} \max_{f \in \mathsf{F}} \mathrm{E}\|f - \hat{f}\|^2 \geqslant \frac{\epsilon_n^2}{32},$$

$$\min_{\hat{f}} \max_{f \in \mathsf{F}} \mathrm{E}\|f - \hat{f}\|^2 \geqslant \frac{\epsilon_n^2}{8},$$

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). Earlier general results in terms of local entropy are given in Birgé (1986).

Proof of Theorem 2. For each $M_n = \lceil C_0 n^\tau \rceil$, consider the class of regression functions $\mathsf{F} = \{f_\theta(x) = \theta_1 \varphi_1(x) + \ldots + \theta_{M_n} \varphi_{M_n}(x) : \|\theta\|_1^{M_n} \leqslant 1\}$. 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 $\min_{\hat{f}} \max_{f \in \mathsf{F}} \mathrm{E}\|f - \hat{f}\|^2 \geqslant C\gamma(n)$ for some constant $C > 0$ not depending on n, where $\gamma(n) = (\log n/n)^{1/2}$ for $\frac{1}{2} < \tau < \infty$ and $\gamma(n) = n^{-(1-\tau)}$ for $0 \leqslant \tau \leqslant \frac{1}{2}$. Since the basis functions are orthonormal, the L_2 distance on F is the same as the l_2 distance on the coefficients $\Theta = \{\theta : \|\theta\|_1^{M_n} \leqslant 1\}$. Thus the entropy of F under the L_2 distance is the same as the that of Θ under the $l_2^{M_n}$ distance. To apply Lemma 1, we bound the local entropy of F or Θ from below. Note that by the Cauchy–Schwarz inequality, the $l_1^{M_n}$ and $l_2^{M_n}$ norms have the relationship $\|\theta\|_1^{M_n} \leqslant (M_n)^{1/2} \|\theta\|_2^{M_n}$. Thus for $\epsilon \leqslant M_n^{-1/2}$, taking $f \equiv 0$, we have

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

Consequently, for $\epsilon \leqslant M_n^{-1/2}$, the $(\epsilon/2)$-packing of $B(f, \epsilon)$ under the L_2 distance is equivalent to the $(\epsilon/2)$-packing of $B_\epsilon = \{\theta\colon \|\theta\|_2^{M_n} \leqslant \epsilon\}$ under the $l_2^{M_n}$ distance. Since a maximum $(\epsilon/2)$-packing set is an $(\epsilon/2)$-covering set, the union of the balls with radius $\epsilon/2$ and centred 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_n}. Thus we have shown that the local entropy $M^{\mathrm{loc}}(\epsilon)$ of F under the L_2 distance is at least $\underline{M}^{\mathrm{loc}}(\epsilon) = M_n \log 2$ for $\epsilon \leqslant M_n^{-1/2}$. For $M_n = \lceil C_0 n^\tau \rceil$ for some $0 \leqslant \tau \leqslant \frac{1}{2}$, solving $\underline{M}^{\mathrm{loc}}(\epsilon_n) = n\epsilon_n^2 + 2\log 2$ gives ϵ_n of order $n^{-(1-\tau)/2}$. Note that for such τ, by possibly reducing $\underline{M}^{\mathrm{loc}}(\epsilon)$ by a constant factor, ϵ_n obtained this way can be made smaller than $M_n^{-1/2}$ (as required in the earlier derivation). Thus, by Lemma 1, we have identified a minimax lower rate for F when $0 \leqslant \tau \leqslant \frac{1}{2}$. That is,

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

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

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

for constants c_1 and c_2 independent of M_n and k when $\log M_n \leqslant k \leqslant M_n$. We can choose $\underline{\epsilon}_n$ and $\bar{\epsilon}_n$ both of order $(\log n/n)^{1/4}$ to satisfy (20) and (21). This gives the minimax lower rate for F when $\tau > \frac{1}{2}$, that is,

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

for some constant $\underline{C}_2$ independent of n. Finally, with the trigonometric basis, the functions in F satisfy $\|f\|_\infty \leqslant (2)^{1/2}$. The theorem follows. □

Remark 12. Both the global and the local entropies are useful here for different cases. For $\tau > \frac{1}{2}$, 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 $M_n^{-1/2}$ and accordingly $B(f,\epsilon) \neq \{f_\theta : \|\theta\|_2^{M_n} \leqslant \epsilon\}$. On the other hand, for $0 \leqslant \tau \leqslant \frac{1}{2}$, 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).

Remark 13. In the derivation of the lower bounds in Theorem 2, we choose very special (non-random) original estimators. This is, of course, not a typical situation where 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. We first examine the approximation errors under the different conditions. Assume that condition B0 is satisfied. For a given j, the approximation error of $f_j(x_j)$ using the first N terms satisfies

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

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

$$\eta_{L,N}(f) = \left\| f - c_0 - \sum_{j=1}^{L}\sum_{l=1}^{N} \theta_{j,l}\varphi_{j,l} \right\|^2 = \sum_{j=L+1}^{\infty}\sum_{l=1}^{\infty} \theta_{j,l}^2 + \sum_{j=1}^{L}\sum_{l=N+1}^{\infty} \theta_{j,l}^2$$

$$\leqslant \frac{1}{(L+1)^{2\beta}} \sum_{j=L+1}^{\infty} j^{2\beta} \sum_{l=1}^{\infty} l^{2s}\theta_{j,l}^2 + \frac{1}{(N+1)^{2s}} \sum_{j=1}^{L} j^{2\beta} \sum_{l=N+1}^{\infty} l^{2s}\theta_{j,l}^2.$$

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

Under conditions B0 and B3, from the above upper bound on $\eta_{L,N}(f)$, we know that the approximation error using the l_1-constrained linearly combined approximation is bounded from above by the same order,

$$\eta_N^L(f) = O(N^{-2s} + L^{-2\beta}). \tag{22}$$

We next examine the risks of the individual estimators. By orthonormality of the basis functions, for any f with $\|f\|_\infty \leqslant A$,

$$\mathrm{E}\|f - \hat{f}^{(j,N)}\|^2 \leqslant \mathrm{E}\|f - \hat{c}_0 - \tilde{f}^{(j,N)}\|^2 = \eta_{(j,N)}(f) + \mathrm{E}(\hat{c}_0 - c_0)^2 + \sum_{l=1}^{N} \mathrm{E}(\hat{\theta}_{j,l} - \theta_{j,l})^2,$$

$$\mathrm{E}\|f - \hat{f}^{L,N}\|^2 \leqslant \mathrm{E}\left\| f - \hat{c}_0 - \sum_{j=1}^{L} \tilde{f}^{(j,N)} \right\|^2 = \eta_{L,N}(f) + \mathrm{E}(\hat{c}_0 - c_0)^2 + \sum_{j=1}^{L}\sum_{l=1}^{N} \mathrm{E}(\hat{\theta}_{j,l} - \theta_{j,l})^2,$$

$$\mathrm{E}\|f - \hat{f}^{L,N,S}\|^2 \leqslant \mathrm{E}\left\| f - \hat{c}_0 - \sum_{j\in S} \tilde{f}^{(j,N)} \right\|^2$$

$$= \sum_{j\notin S}\sum_{l=1}^{\infty} \theta_{j,l}^2 + \sum_{j\in S}\sum_{l=N+1}^{\infty} \theta_{j,l}^2 + \mathrm{E}(\hat{c}_0 - c_0)^2 + \sum_{j\in S}\sum_{l=1}^{N} \mathrm{E}\left(\hat{\theta}_{j,l} - \theta_{j,l}\right)^2.$$

It is straightforward to bound the variances of the estimators of the coefficients. Clearly $\mathrm{E}(\hat{c}_0 - c_0)^2 = n^{-1}\mathrm{var}(Y_1) \leqslant n^{-1}(A^2 + \sigma^2)$, and, by expanding squares, we have that $\mathrm{E}(\hat{\theta}_{j,l} - \theta_{j,l})^2$ is bounded from above by

$$\frac{1}{n}\mathrm{E}\left(\sigma\varepsilon_1\varphi_{j,l}(X_{1j}) + c_0\varphi_{j,l}(X_{1j}) + \theta_{j,l}((\varphi_{j,l}(X_{1j}))^2 - 1) + \sum_{(j',l')\neq(j,l)} \theta_{j',l'}\varphi_{j,l}(X_{1j})\varphi_{j',l'}(X_1 j') \right)^2$$

$$= \frac{1}{n}\left(\sigma^2 + c_0^2 + \theta_{j,l}^2\mathrm{E}((\varphi_{j,l}(X_{1j}))^2 - 1)^2 + \sum_{(j',l')\neq(j,l)} \theta_{j',l'}^2 + 2c_0\theta_{j,l}\mathrm{E}(\varphi_{j,l}(X_{1j})((\varphi_{j,l}(X_{1j}))^2 - 1)) \right)$$

$$\leqslant \frac{1}{n}(\sigma^2 + A^2 + A^2((A')^2 + 1)^2 + 2A^2A'((A')^2 + 1)),$$

where the inequality follows from the boundness assumptions on the basis functions and f.

Now from (10), (11), (12), (13) and the above upper bounds on the approximation and estimation errors, we have that, under conditions B0 and B1, with the choice of N of order $n^{1/(2s+1)}$, the quantity in (10) is bounded in order from above by

$$\inf_N \left(N^{-2s} + \frac{N}{n} \right) = O\left(n^{-2s/(1+2s)} \right);$$

under condition B0, with the choice of N of order $n^{1/(1+s(2+1/\beta))}$ and L of order $n^{(s/\beta)/(1+s(2+1/\beta))}$, the quantity in (11) is bounded in order from above by

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

under conditions B0 and B2, with the choice of N of order $n^{1/(1+s(2+\nu/\beta))}$, L of order $n^{(s/\beta)/(1+s(2+\nu/\beta))}$ and k of order L^{ν}, the risk of δ_3 is bounded in order from above by

$$\inf_{L,N} \left(N^{-2s} + L^{-2\beta} + \frac{L^{\nu}N}{n} + \frac{L^{\nu}\log L}{n} \right) = O\left(n^{-2s/(1+s(2+\nu/\beta))} \right);$$

under conditions B0 and B3, with the choice of, for example, $L = O(n^{1/(4\beta)})$ and $N = O(n^{1/(4s)})$, together with (22), the quantity in (13) is bounded in order from above by

$$\inf_{L,N} (\eta_N^L(f) + \psi_n(LN+1)) = O\left(n^{-1/2} + \psi_n\left(n^{1/(4\beta)+1/(4s)} \right) \right) = O(\log n/n)^{1/2},$$

where, for the last equality, we use the fact that $\psi_n(n^{\tau})$ is bounded in order from above by $(\log n/n)^{1/2}$ for any $0 < \tau < \infty$. The corollary follows.

Acknowledgements

This research was partially supported by US National Security Agency Grant MDA9049910060 and US National Science Foundation CAREER Grant DMS0094323. The author sincerely thanks an associate editor and two anonymous reviewers whose very constructive comments led to the correction of some errors and improvements to the presentation of the paper.

References

Barron, A.R. (1987) Are Bayes rules consistent in information? In T.M. Cover and B. Gopinath (eds) *Open Problems in Communication and Computation*, pp. 85–91. Berlin: Springer-Verlag.

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

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

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

Barron, A.R., Rissanen, J. and Yu, B. (1998) The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, **44**, 2743–2760.

Barron, A.R., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113**, 301–413.

Bates, J.M. and Granger, C.W.J. (1969) The combination of forecasts. *Oper. Res. Quart.*, **20**, 451–468.
Birgé, L. (1986) On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields*, **71**, 271–291.
Breiman, L. (1996) Stacked regressions. *Machine Learning*, **24**, 49–64.
Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997) Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.
Catoni, O. (1997) The mixture approach to universal model selection. Technical Report LIENS-97-22, École Normale Superieure, Paris, France.
Catoni, O. (1999) 'Universal' aggregation rules with exact bias bounds. Preprint.
Cesa-Bianchi, N. and Lugosi, G. (1999) On prediction of individual sequences. *Ann. Statist.*, **27**, 1865–1895.
Cesa-Bianchi, N., Freund, Y., Haussler, D.P., Schapire, R. and Warmuth, M.K. (1997) How to use expert advice. *J. ACM*, **44**, 427–485.
Clemen, R.T. (1989) Combining forecasts: a review and annotated bibliography. *Internat. J. Forecasting*, **5**, 559–583.
Donoho, D.L. and Johnstone, I.M. (1994) Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sér. I Math.*, **319**, 1317–1322.
Donoho, D.L. and Johnstone, I.M. (1998) Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879–921.
Edmunds, D.E. and Triebel, H. (1989) Entropy numbers and approximation numbers in function spaces. *Proc. London Math. Soc. (3)*, **58**, 137–152.
Johnstone, I.M. (1999) Function estimation in Gaussian noise: sequence models. Manuscript.
Juditsky, A. and Nemirovski, A. (2000) Functional aggregation for nonparametric estimation. *Ann. Statist.*, **28**, 681–712.
Kolmogorov, A.N. and Tikhomirov, V.M. (1959) ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk*, **14**, 3–86.
Le Cam, L.M. (1975) On local and global properties in the theory of asymptotic normality of experiments. In M.L. Puri (ed.), *Proceedings of the Summer Research Institute on Statistical Inference for Stochastic Processes, Vol. 1: Stochastic Processes and Related Topics*, 1, pp. 13–54. New York: Academic Press.
LeBlanc, M. and Tibshirani, R (1996) Combining estimates in regression and classification. *J. Amer. Statist. Assoc.*, **91**, 1641–1650.
Littlestone, N. and Warmuth, M.K. (1994) The weighted majority algorithm. *Inform. and Comput.*, **108**, 212–261.
Mallat, S.G. and Zhang, Z. (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, **41**, 3397–3415.
Merhav, N. and Feder, M. (1998) Universal prediction. *IEEE Trans. Inform. Theory*, **44**, 2124–2147.
Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.
Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1997) Local polynomial variance-function estimation. *J. Amer. Statist. Assoc.*, **39**, 262–273.
Schütt, C. (1984) Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory*, **40**, 121–128.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**, 111–147.
Vovk, V.G. (1990) Aggregating strategies. In M. Fulk and J. Case (eds), *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pp. 372–383. San Mateo, CA: Morgan Kanfmann.

Wolpert, D. (1992) Stacked generalization. *Neural Networks*, **5**, 241–259.
Yang, Y. (2000a) Mixing strategies for density estimation. *Ann. Statistics*, **28**, 75–87.
Yang, Y. (2000b) Combining different procedures for adaptive regression. *J. Multivariate Anal.*, **74**, 135–161.
Yang, Y. (2001) Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, **96**, 574–588.
Yang, Y. and Barron, A.R. (1997) Information-theoretic determination of minimax rates of convergence. Technical Report no. 28, Department of Statistics, Iowa State University.
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.

Received May 2001 and revised August 2003