How Powerful Can Any Regression Learning Procedure Be?

Yuhong Yang
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Abstract

Efforts have been directed at obtaining flexible learning procedures that optimally adapt to various possible characteristics of the data generating mechanism. A question that addresses the issue of how far one can go in this direction is: Given a regression procedure, however sophisticated it is, how many regression functions are estimated accurately? In this work, for a given sequence of prescribed estimation accuracy (in sample size), we give an upper bound (in terms of metric entropy) on the number of regression functions for which the accuracy is achieved. Interesting consequences on adaptive and sparse estimations are also given.

1 Introduction

Powerful regression procedures have been proposed and certainly more will come to deal with challenging issues such as the curse of dimensionality. A main spirit in modern regression learning is to construct statistical procedures that intelligently learn and adapt to the real characteristics of the data. The hope is that such a learning procedure is highly flexible to work well for various scenarios. Universally consistent estimators have been found (e.g., Stone (1977) and Devroye and Wagner (1980)). However, the convergence of such an estimator to the true regression function can be arbitrarily slow (see, e.g., Devroye (1982)). In contrast to consistency, minimax risks are often considered for better describing the performance of an estimator for target functions with certain characteristics (e.g., monotonicity or smoothness). Such characteristics usually determine how fast the minimax risk converges to zero. For example, the smoother the function to be estimated is, the faster the minimax risk converges (see, e.g., Ibragimov and Hasminskii (1977), Bretagnolle and Huber (1979), Stone (1982)). More generally, relationship between the minimax rate of convergence and the largeness of the function class as measured in terms of metric entropy is now well understood under familiar loss functions (e.g., Birgé (1986), Yatracos (1988) and Yang and Barron (1999)).

Since early original work of Efromovich and Pinsker (1984), adaptive function estimation has received considerable attention (see, e.g., Donoho and Johnstone (1998) and Barron, Birgé and Massart (1999) for some references). The main goal is to have a learning procedure that automatically adapt to various possible characteristics of interest (e.g., smoothness degree of the function or interaction order of the predictors) in terms of minimax rate of convergence.

Clearly adaptation is a desirable property for function estimation. If possible, one would wish to use a super adaptive procedure that can intelligently utilize the information in the data to produce most accurate estimators suitable for as many situations as possible. A natural question then is: how adaptive can any procedure be? This motivates the study of the problem that for a given regression leaning procedure, how many functions are estimated well. Kerkyacharian and Picard (e.g., 2002) answered this question for certain types of well-known nonparametric procedures (e.g., wavelet thresholding and smoothing with localized bandwidth selection). In this paper, we deal with an arbitrarily sophisticated procedure and show that the set of regression functions that can be well estimated is fundamentally limited in size.

The rest of the paper is organized as follows. In Section 2, we recall a result that for a given class of probability density functions, a good news in estimation directly forecasts a bad news. In Section 3, we present the main result that no estimator can converge fast for many underlying functions. In Section 4, we give a negative consequence of the main result on adaptive function estimation. In Section 5, we show that in some sense, every regression procedure is essentially no better than

a method based on a certain sparse approximation.

2 Lower bounding minimax risk through upper bounds

In this section, we provide a preliminary result, which will be used for deriving the main results. Let X_1, X_2, ..., X_n be i.i.d. observations with probability density function $p(x), x \in \mathcal{X}$ with respect to a σ-finite measure μ. Here the space $\mathcal{X}$ is general and could be any dimensional. One needs to estimate the unknown density p based on the data.

The Kullback-Leibler (K-L) divergence between two densities p and q is defined as $D(p \parallel q) = \int p \log(p/q)\, d\mu$. Let d be a distance (metric) between densities. Examples are Hellinger distance $d_H(p,q) = \left(\int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mu\right)^{1/2}$ and the L_2 distance $\| p - q \| = \left(\int (p-q)^2 d\mu\right)^{1/2}$. The loss $d^2(p, \widehat{p})$ will be considered for density estimation in this section.

Let $\mathcal{F}$ be a class of density functions. Then the minimax risk for estimating a density in $\mathcal{F}$ at sample size n under the d^2 loss is defined as

$$R(\mathcal{F}; d; n) = \inf_{\widehat{p}} \sup_{p \in \mathcal{F}} E d^2(p, \widehat{p}),$$

where the infimum is taken over all density estimators.

Metric entropy is a fundamental concept describing massiveness of a set (see, Kolmogorov and Tihomirov (1959) or Lorentz et al (1996, Chapter 15)). A set N is said to be an ϵ-packing set in $\mathcal{F}$ if $N \subset \mathcal{F}$ and any two distinct members in N are more than ϵ apart under the metric d. The packing ϵ-entropy of the set $\mathcal{F}$ under the metric d, denoted by $M(\epsilon; \mathcal{F})$, is then defined to be the logarithm of the size of the largest ϵ-packing set in $\mathcal{F}$.

The result below is essentially in Yang and Barron (1999) (not formally given there but contained in the proofs).

THEOREM 1: *For a sequence of estimators $\widehat{p}_k$ based on $X_1, ...X_k$, $1 \le k \le n$, if $\sup_{p \in \mathcal{F}} E D(p \parallel \widehat{p}_k) \le b_k^2$, then for the minimax risk, we have*

$$R(\mathcal{F}; d; n) \ge \frac{\underline{\sigma}_{n,d}^2}{8},$$

where $\underline{\sigma}_{n,d}^2$ is chosen such that

$$M(\underline{\sigma}_{n,d}; \mathcal{F}) = \left\lceil 2 \left(\sum_{i=0}^{n-1} b_i^2 + \log 2 \right) \right\rceil.$$

3 How many regression functions can be served well by any given regression procedure?

We now show that the collection of functions that have small risks by any given estimation procedure is fundamentally limited in size. We study the problem in a nonparametric regression setting for convenience, but similar results can be obtained for density estimation as well.

Consider the regression model

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

where $(X_i, Y_i)_{i=1}^n$ are i.i.d. copies from the joint distribution of (X, Y) with $Y = f(X) + \varepsilon$. The explanatory variable X (could be any dimensional) has a distribution P_X and the random error ε is assumed to have a normal distribution with mean zero and variance $\sigma^2 > 0$. One needs to estimate the regression function f based on the data $Z^n = (X_i, Y_i)_{i=1}^n$. Since our main interest in this work is on the negative side (limit of estimation), unless stated otherwise, σ^2 is assumed to be known and then taken to be 1 without loss of generality. When σ^2 is unknown, the problem of estimating f certainly can not be easier. For the same reason, we assume P_X is known in this paper unless stated otherwise.

Let δ be a regression estimation procedure producing estimator $\widehat{f}_i(x) = \widehat{f}_i(x; Z^i)$ at each sample size $i \ge 1$. 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. For a class of regression functions $\mathcal{F}$, let $R(\mathcal{F}; n) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E \| f - \widehat{f} \|^2$ denote its minimax risk.

Fix a regression procedure δ. Let b_n^2 be a non-increasing sequence with $b_n^2 \to 0$ as $n \to \infty$. Assume that the true regression function has the L_2 norm bounded by a known constant $A > 0$. Consider the class of regression function $\mathcal{F}(\{b_n^2\}; \delta)$:

$$\{f : \| f \| \le A \text{ and } R(f; n; \delta) \le b_n^2 \text{ for all } n \ge 1\}. \quad (1)$$

It is the collection of the regression functions for which the estimation procedure δ achieves the given accuracy b_n^2 at each sample size n. Ideally, one wants this class to be as large as possible.

As is mentioned in the introduction, it is well-known that one can not demand any universal convergence rate, i.e., for any given sequence $b_n^2 \downarrow 0$, for every

estimation procedure δ, there exists at least one regression function f such that $R(f;n;\delta) \geq b_n^2$ for each $n \geq 1$ (see, e.g., Devroye (1982)). Thus one knows that $\mathcal{F}(\{b_n^2\};\delta)$ can not be the class of all possible regression functions. But it is still unclear, however, how much smaller $\mathcal{F}(\{b_n^2\};\delta)$ is compared to the class of all regression functions. We will provide an upper order of the largeness of $\mathcal{F}(\{b_n^2\};\delta)$ in terms of metric entropy. The order can also be achieved by familiar regression procedures for smoothness function classes. Thus the largeness bound (order) that we will give can not be improved in general.

The problem of upper bounding the metric entropy of $\mathcal{F}(\{b_n^2\};\delta)$ is closely related to the problem of lower bounding the minimax risk of a general class of regression functions. Intuitively if $\mathcal{F}(\{b_n^2\};\delta)$ is too large, then the rate of convergence b_n^2 can not be achieved uniformly. However, it seems that no general lower bounds in the literature have direct implications on the size of $\mathcal{F}(\{b_n^2\};\delta)$. In particular, it is not feasible to apply hyper-cube methods (often used in deriving minimax lower bounds) because little can be said about the structure and local properties of the class $\mathcal{F}(\{b_n^2\};\delta)$ since no specific conditions are put on δ. Theorem 1 in the previous section is handy for upper bounding the metric entropy of $\mathcal{F}(\{b_n^2\};\delta)$.

Note that in the definition of $\mathcal{F}(\{b_n^2\};\delta)$, the risk bounds are required to hold for each sample size. One might wonder why not consider one sample size at a time. Actually, if one modifies the definition of $\mathcal{F}(\{b_n^2\};\delta)$ this way, i.e., define $\mathcal{F}_n(b_n^2;\delta) = \{f : \| f \| \leq A \text{ and } R(f;n;\delta) \leq b_n^2\}$, then no general nontrivial bound is possible as seen from the following simple example.

EXAMPLE 1: Consider the procedure δ that produces the trivial estimator $\widehat{f}_i(x) \equiv 0$ for all sample sizes. For any $\epsilon > 0$, the minimax risk of the class $\mathcal{F} = \{f : \| f \|_2 \leq \epsilon\}$ is obviously no bigger than ϵ by using δ, but the metric entropy of $\mathcal{F}$ is always ∞. This indicates that it is impossible to have a non-trivial upper bound on the metric entropy of the class of functions that δ serves well at a given sample size.

Let $b_0^2 = A^2 + 2\log 2$ and define $B_k = \sum_{i=0}^{k} b_i^2$ for $k \geq 1$ and $B_0 = b_0^2$.

THEOREM 2: *Take $b_n^2 = Cn^{-\gamma}$ for some constant $C > 0$ and $0 < \gamma \leq 1$. When $\gamma < 1$, for every regression procedure δ, for $\epsilon \leq 3C^{1/2}$, we must have*

$$M\left(\epsilon; \mathcal{F}(\{b_n^2\};\delta)\right) \leq C^{'} \left(\frac{1}{\epsilon}\right)^{\frac{2(1-\gamma)}{\gamma}},$$

where $C^{'}$ is a constant depending only on γ, C and A. When $\gamma = 1$, for every regression procedure δ, for $\epsilon \leq 3C^{1/2}$, we have

$$M\left(\epsilon; \mathcal{F}(\{b_n^2\};\delta)\right) \leq C^{''} \log\left(\frac{1}{\epsilon}\right)$$

for some constant $C^{''}$ depending only on A and C. For a general sequence $\{b_n^2\}$, we have

$$M\left(3b_k; \mathcal{F}(\{b_n^2\};\delta)\right) \leq \lceil B_{k-1} \rceil \text{ for all } k \geq 1.$$

REMARKS:

1. The normality assumption on the errors is not essential. A similar result holds if the regression function is bounded in a known range $[-A, A]$ and the error distribution satisfies a mild condition as used for Theorem 1 in Yang (2001).

2. The dependence of $C^{'}$ or $C^{''}$ on γ, C *and* A is given in the proof of the theorem below.

3. For certain specific procedures (such as wavelet shrinkage), Kerkyacharian and Picard (2002) successfully characterized the set $\mathcal{F}(\{b_n^2\};\delta)$.

PROOF: For a function g, let $p_g(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-g(x))^2}$ denote the joint density of (X,Y) with respect to the product measure of P_X and Lebesgue measure when the regression function is g. It can be easily verified that the K-L divergence between two such densities satisfies $D(p_f \parallel p_g) = \frac{1}{2} \parallel f - g \parallel^2$.

Let $\widetilde{f}_k$, $k \geq 1$ be the estimators of f by the procedure δ at each sample size respectively. Let $\widetilde{f}_0(x) \equiv 0$ and then $E \parallel f - \widetilde{f}_0 \parallel^2 \leq A^2$ by assumption. Let

$$q_n(z^n) = p_{\widetilde{f}_0}(x_1,y_1) \cdot p_{\widetilde{f}_1}(x_2,y_2) \cdots p_{\widetilde{f}_{n-1}}(x_n,y_n).$$

It is a probability density function in z^n with respect to the n-fold product measure of the distribution of X and Lebesgue measure. Then as in the proof of Lemma 2,

$$D(p_f^n \parallel q_n) = \sum_{i=1}^{n} ED(p_f \parallel p_{\widetilde{f}_{i-1}}) = \frac{1}{2} \sum_{i=1}^{n} E \parallel f - \widetilde{f}_{i-1} \parallel^2.$$

It follows that for $f \in \mathcal{F}(\{b_n^2\};\delta)$, we have $D(p_f^n \parallel q_n) \leq \frac{1}{2}\left(A^2 + \sum_{i=1}^{n-1} b_i^2\right)$. Let $\mathcal{F}$ denote $\mathcal{F}(\{b_n^2\};\delta)$ for convenience. Choose $\underline{\epsilon}_n$ such that

$$M(\underline{\epsilon}_n; \mathcal{F}) = \left\lceil 2\left(\frac{1}{2}\left(A^2 + \sum_{i=1}^{n-1} b_i^2\right) + \log 2\right)\right\rceil = \lceil B_{n-1} \rceil.$$

Then by Lemma 1, we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} E \parallel f - \widehat{f} \parallel^2 \geq \frac{\underline{\epsilon}_n^2}{8}.$$

By definition of $\mathcal{F}(\{b_n^2\};\delta)$, we have $\sup_{f\in\mathcal{F}} E \parallel f - \widetilde{f}_n \parallel^2 \leq b_n^2$, and thus the minimax risk of $\mathcal{F}$ satisfies

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} E \parallel f - \widehat{f} \parallel^2 \leq b_n^2. \quad (2)$$

From above, if $M(3b_n;\mathcal{F}) > \lceil B_{n-1} \rceil$, then $3b_n \leq \epsilon_n$ and

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} E \parallel f - \widehat{f} \parallel^2 \geq \frac{(3b_n)^2}{8} > b_n^2,$$

which would contradict with (2). Thus we must have

$$M(3b_n;\mathcal{F}) \leq \lceil B_{n-1} \rceil.$$

For the case $b_n^2 = Cn^{-\gamma}$ for some $0 < \gamma < 1$, $B_n \leq A^2 + 2\log 2 + \frac{Cn^{1-\gamma}}{1-\gamma}$. For any $0 < \epsilon \leq 3C^{1/2}$, there exists $n_\epsilon \geq 1$ such that $3C^{1/2}(n_\epsilon+1)^{-\gamma/2} < \epsilon \leq 3C^{1/2} n_\epsilon^{-\gamma/2}$. Then $\left(\frac{3C^{1/2}}{\epsilon}\right)^{2/\gamma} - 1 < n_\epsilon \leq \left(\frac{3C^{1/2}}{\epsilon}\right)^{2/\gamma}$. It follows from these inequalities that when $0 < \epsilon \leq 3C^{1/2}$, $M(\epsilon;\mathcal{F})$ is upper bounded by

$$\begin{aligned} & M\left(3C^{1/2}(n_\epsilon+1)^{-\gamma/2};\mathcal{F}\right) \leq B_{n_\epsilon} + 1 \\ \leq\ & A^2 + 2\log 2 + 1 + \frac{Cn_\epsilon^{1-\gamma}}{1-\gamma} \\ \leq\ & A^2 + 2\log 2 + 1 + \frac{3^{2(1-\gamma)/\gamma} C^{1/\gamma}}{1-\gamma} \left(\frac{1}{\epsilon}\right)^{\frac{2(1-\gamma)}{\gamma}}. \end{aligned}$$

Thus

$$M(\epsilon;\mathcal{F}) \leq C' \left(\frac{1}{\epsilon}\right)^{\frac{2(1-\gamma)}{\gamma}},$$

where C' is a constant depending only on γ, C and A. For the case $b_n^2 = Cn^{-1}$,

$$\begin{aligned} B_n &= A^2 + 2\log 2 + C\sum_{i=1}^{n-1} i^{-1} \\ &\leq A^2 + 2\log 2 + C\left(1 + \log(n-1)\right). \end{aligned}$$

Similarly as the case when $0 < \gamma < 1$, we have for $0 < \epsilon \leq 3C^{1/2}$, $M(\epsilon;\mathcal{F})$ is no larger than

$$\begin{aligned} & A^2 + 2\log 2 + 1 + C\left(1 + \log\left(\frac{3C^{1/2}}{\epsilon}\right)^2\right) \\ \leq\ & A^2 + 2\log 2 + 1 \\ & + C\left(1 + 2\log\left(3C^{1/2}\right)\right) + 2C\log\left(\frac{1}{\epsilon}\right). \quad (3) \end{aligned}$$

Thus $M(\epsilon;\mathcal{F}) \leq C'' \log\left(\frac{1}{\epsilon}\right)$ for some C'' depending only on C and A. This completes the proof of Theorem 2.

It is well-known that for smoothness function classes (e.g., Sobolev or Besov), the minimax rate of convergence is usually determined by a certain smoothness parameter α (e.g., the number of derivatives that the unknown regression function is assumed to have) with the rate $n^{-2\alpha/(2\alpha+d)}$, where d is the dimension of the function being estimated. The metric entropy order of such a class is typically $(1/\epsilon)^{d/\alpha}$ as $\epsilon \to 0$. Note that for $\gamma = 2\alpha/(2\alpha+d)$, $2(1-\gamma)/\gamma = d/\alpha$ and accordingly the entropy upper bound given in Theorem 2 for $\mathcal{F}(\{b_n^2\};\delta)$ is of order $(1/\epsilon)^{d/\alpha}$. Notice that this order matches the metric entropy of smoothness classes with convergence rates $n^{-2\alpha/(2\alpha+d)}$. As is well-known, the convergence rate $1/n$ corresponds to parametric classes and they usually have metric entropies of order $\log(1/\epsilon)$, which is the order given in Theorem 2 for $\gamma = 1$. Thus in terms of order, the upper bounds in Theorem 2 can not be generally improved.

From Theorem 2, no matter how sophisticatedly a regression procedure is constructed, it can converge fast for only a limited set of regression functions.

Barron and Hengartner (1998) study super-efficiency in density estimation for both parametric and nonparametric classes. For a nonparametric class of densities, they show that for any given estimation procedure, the set of densities for which the procedure converges faster than the minimax rate of the class is asymptotically negligible compared to the whole class in terms of metric entropy order. The metric entropy bound in Theorem 2 can also be used to readily derive such a result for the regression problem.

4 Implications on adaptive estimation

The non-asymptotic nature of the upper bound in Theorem 2 is helpful to draw some conclusions on limitations of adaptive estimation.

Traditionally, adaptive estimation addresses the objective of achieving the minimax risk (often rate of convergence) over smoothness function classes with the smoothness parameter unknown. As mentioned earlier, different values of the smoothness parameter are usually associated with different rates of convergence. The function classes have different sizes (in terms of metric entropy) and are usually nested. Yang (2000ab, 2001) shows that adaptive rate of convergence can be obtained for a general countable collection of function classes. The function classes are allowed to be completely different, which may be desirable when very distinct scenarios are explored in situations such as high-dimensional estimation.

Let $\mathcal{F}_1, \mathcal{F}_2 ...$ be a countable collection of regression classes. Assume that the regression functions in the classes are uniformly bounded between $-A$ and A for some $A > 0$, and the variance parameter σ^2 is upper bounded by a known constant $\overline{\sigma}^2$. Let $\{\pi_j, j \geq 1\}$ be

positive numbers satisfying $\sum_{j=1}^{\infty} \pi_j = 1$. Yang (2001) shows that one can construct an adaptive estimator $\widehat{f}_n^*$ such that for each $j \geq 1$,

$$\sup_{f \in \mathcal{F}_j} E \parallel f - \widehat{f}_n^* \parallel^2 \leq C \left(\frac{1}{n} \log \frac{1}{\pi_j} + \frac{1}{n} + R(\mathcal{F}_j; \lfloor n/2 \rfloor) \right), \tag{4}$$

where the constant C depends only on A and $\overline{\sigma}$.

For a typical parametric or nonparametric function class, the minimax risk sequence is rate-regular in the sense that $R(\mathcal{F}_j; \lfloor n/2 \rfloor)$ and $R(\mathcal{F}_j; n)$ converge at the same order. If the regression function classes $\mathcal{F}_j$ are all rate-regular, then from (4), since for each fixed j, $\frac{1}{n} \log \frac{1}{\pi_j} + \frac{1}{n}$ does not change the rate of $R(\mathcal{F}_j; n)$, we conclude that the minimax rate of convergence is automatically achieved for every class $\mathcal{F}_j$. This is called minimax-rate adaptation.

This notion of adaptation addresses the asymptotic performance for each class as $n \to \infty$. However, for each fixed n, since for a countable collection of classes, π_j necessarily goes to zero and accordingly $\frac{1}{n} \log \frac{1}{\pi_j} \to \infty$ as $j \to \infty$. Thus the penalty term $\frac{1}{n} \log \frac{1}{\pi_j}$ in (4) is large except for a few classes with high prior weights. A question then is: Can one construct a better adaptive method such that for a constant C,

$$\sup_{f \in \mathcal{F}_j} E \parallel f - \widehat{f}^* \parallel^2 \leq C R(\mathcal{F}_j; \lfloor n/2 \rfloor)$$

for all $j \geq 1$ and $n \geq 1$? If so, the adaptive estimator achieves the minimax risk up to a constant factor uniformly over both the classes and the sample sizes. Based on the known fact that no uniform rate of convergence is possible, it seems intuitively clear that the objective is simply too ambitious to achieve in general. We give a formal result below using the metric entropy bound in the previous section.

COROLLARY 1: *There exist a collection of uniformly bounded classes of regression functions* $\{\mathcal{F}_j, j \geq 1\}$ *and a constant* $B > 0$ *such that for any regression procedure* δ^**, we have that for each* $\lambda > 1$*, there can be at most* $e^{B\lambda}$ *many classes for which*

$$\frac{\sup_{f \in \mathcal{F}_j} R(f; n; \delta^*)}{R(\mathcal{F}_j; n)} \leq \lambda \text{ for all } n \geq 1.$$

REMARK: With a similar argument, it can be shown that if one is willing to loose a logarithmic factor $\log n$ in risk for each $n \geq 1$, then in general one can achieve that for not more than n^{κ} many classes for some constant $\kappa > 0$.

PROOF: Consider that $X = (X_1, ...)$ takes values in $\mathcal{X} = [0,1]^{\infty}$ with independent and uniformly distributed components. Let $\mathcal{G} = \{g(x) : g(x) = \theta x, 1 \leq \theta \leq 2\}$ be a parametric class of functions on [0,1]. Let $\mathcal{F}_j$ be the collection of functions that actually depend only on x_j and the univariate function belongs to $\mathcal{G}$. For the parametric family $\mathcal{F}_j$, it is not hard to show that $\frac{C_1}{n} \leq R(\mathcal{F}_j; n) \leq \frac{C_2}{n}$ for some positive constants C_1 and C_2. Note that the functions in different classes are well separated: for $f_1 \in \mathcal{F}_{j_1}$ and $f_2 \in \mathcal{F}_{j_2}$ with $j_1 \neq j_2$, one always has

$$\parallel f_1 - f_2 \parallel \geq 1/6.$$

As a consequence, the ϵ-packing sets in different classes $\mathcal{F}_j$ are at least 1/6 away from each other, and accordingly, the packing entropy of $\cup_{j \geq 1} \mathcal{F}_j$ is infinity when ϵ is smaller than 1/6. Fix a constant $\lambda > 1$. For any given regression procedure δ^*, consider the classes that each has risk

$$\sup_{f \in \mathcal{F}_j} R(f; n; \delta^*) \leq \frac{\lambda C_2}{n} \text{ for all } n \geq 1.$$

Let Γ be the collection of all such classes. Then we have

$$\sup_{f \in \cup_{j \in \Gamma} \mathcal{F}_j} R(f; n; \delta^*) \leq \frac{\lambda C_2}{n} \text{ for all } n \geq 1.$$

Thus $\cup_{j \in \Gamma} \mathcal{F}_j \subset \mathcal{F}(\{\frac{\lambda C_2}{n}\}; \delta^*)$. It follows from Theorem 2 and (3) that $\cup_{j \in \Gamma} \mathcal{F}_j$ has packing entropy bounded above by $2\lambda C_2 \log(1/\epsilon)$ asymptotically as $\epsilon \to 0$. It can be easily verified that $\mathcal{F}_j$ has metric entropy uniformly lower bounded by $C_3 \log(1/\epsilon)$ for some constant $C_3 > 0$. Since the classes $\mathcal{F}_j$ are well separated, when $\epsilon < 1/6$, the maximum packing number of $\cup_{j \in \Gamma} \mathcal{F}_j$ is the sum of the maximum packing numbers of the classes. It follows that the packing entropy of $\cup_{j \in \Gamma} \mathcal{F}_j$ is the packing entropy of $\mathcal{G}$ plus the logarithm of the size of Γ. As a consequence, the logarithm of the size of Γ is upper bounded by $B\lambda$ for a constant $B > 0$. Thus $|\Gamma| \leq e^{B\lambda}$. This completes the proof of Corollary 1.

From Corollary 1, adaptation up to a uniform constant factor can not be achieved in general except for finitely many function classes. Note that the result does not exclude the possibility of adaptation within a constant factor for particular collections of function classes. See Barron, Birgé and Massart (1999) for examples of such adaptation results for some smoothness classes.

5 Sparse approximation and estimation

In recent years, sparse estimation has attracted an increasing attention in statistical leaning (for some theoretical results, see, e.g., Donoho (1993), Barron (1993), Donoho and Johnstone (1998), Yang and Barron (1998, 1999), Johnstone (1999) and Barron, Birgé

and Massart (1999)). This is particularly important for learning a high-dimensional function, especially when the sample size is small relative to the dimension. In such a case, one seeks a sparse representation of the target function, which makes the estimation both feasible and reliable. It has been shown that sparse approximation together with suitable statistical methods lead to better estimation compared to traditional linear approximation in situations such as orthogonal wavelet expansion and neural network modeling. In this section, we show that in some sense (to be made clear), each regression procedure is essentially no better than a method based on sparse approximation.

Let k be an index. For each k, let $\Phi_k = \{\varphi_{k,1}, ..., \varphi_{k,L_k}\}$ be a collection of L_k linearly independent functions. Given k, $1 \le m \le L_K$ and $I = I_{k,m} = \{i_1, ..., i_m\}$ as a subset of $\{1, 2, ..., L_k\}$ with m terms in Φ_k, consider approximation of a function f by linear combinations

$$\sum_{l=1}^{m} \theta_l \varphi_{k,i_l}, \ (\theta_1, ..., \theta_m) \in R^m.$$

When m is small compared to L_k, the terms used in the linear combination is a sparse subset of Φ_k. The sparsity in fact is the key for improved accuracy in estimation compared to traditional linear approximation using all L_k terms when f has certain sparsity characteristics. We call such approximation that allows the use of sparse linear combinations *sparse approximation*. In general, the choices of approximation systems $\{\Phi_k\}$ are also allowed to depend on the sample size.

For estimating a regression function f based on the data $(X_i, Y_i)_{i=1}^n$, one can use sparse subset models corresponding to sparse approximation. For each choice of (k, m, I), one fits the model

$$Y_i = \sum_{l=1}^{m} \theta_l \varphi_{k,i_l}(X_i) + \varepsilon_i, \quad 1 \le i \le n$$

based on the observations. The use of sparse subsets can be advantageous in terms of estimation accuracy when a small number of terms can provide a good approximation of f, since using a sparse subset avoids large variability that arises when all the L_k coefficients are estimated. Since one does not know which subset provides a good approximation, one may select a model according to a certain appropriate criterion. For convergence rate results on model selection for nonparametric regression, see, e.g., Yang (1999), Lugosi and Nobel (1999), Barron, Birgé and Massart (1999), and Wegkamp (2003). Alternatively to selecting a single model, one can also average the sparse subset models. Proper averagings have been demonstrated to lead to reduced model uncertainty (e.g., Hoeting, et al (1999)), increased stability of the estimator (Breiman (1996)) and improved estimation accuracy in risk (Yang (2001)) from different angles.

In this section, we assume that P_X is dominated by a known probability measure μ with a density function $p_X(x)$, which is uniformly bounded above and below (away from zero). We assume σ^2 is upper bounded by a known constant $\overline{\sigma}^2 < \infty$.

THEOREM 3: *For any given regression procedure δ, there exists a procedure $\widetilde{\delta}$ based on sparse approximation (with a proper model averaging) such that for every regression function f with $\| f \|_\infty < \infty$, if $R(f; \delta; n) \le Cn^{-\gamma}$ under $\sigma^2 = \overline{\sigma}^2$ for all n for some constant $C > 0$ and $0 < \gamma < 1$, then $R(f; \widetilde{\delta}; n) \le \widetilde{C} n^{-\gamma}$ holds under $\sigma^2 \le \overline{\sigma}^2$ for all n for some constant $\widetilde{C} > 0$; if $R(f; \delta; n) \le Cn^{-1}$ under $\sigma^2 = \overline{\sigma}^2$ for all n for some constant $C > 0$, then $R(f; \widetilde{\delta}; n) \le \widetilde{C} n^{-1} \log n$ holds under $\sigma^2 \le \overline{\sigma}^2$ for some constant $\widetilde{C} > 0$.*

REMARKS:

1. The sparse approximation systems constructed in the theorem depend on the procedure δ.

2. The assumption on P_X is needed so that the procedure $\widetilde{\delta}$ can be constructed (in theory) based on δ.

3. If there exists an estimator of σ^2 converging at rate $1/n$ under the square error, then the condition that σ^2 is upper bounded by a known constant $\overline{\sigma}^2 < \infty$ is not needed.

The theorem says that as far as polynomial rates of convergence are concerned, under the squared L_2 loss, theoretically speaking, estimation based on a certain sparse approximation together with model averaging can do as well as any given regression procedure (but loosing a logarithmic factor for the parametric rate of convergence).

PROOF OF THEOREM 3: We first assume that P_X is known. Without loss of generality, assume $\overline{\sigma}^2 = 1$. For a given regression procedure δ, for each C and $0 < \gamma \le 1$, from Section 3, the set $\mathcal{F}(\{b_n^2\}; \delta)$ of regression functions as defined in (1) with $b_n^2 = Cn^{-\gamma}$ has metric entropy bounded above by order $\left(\frac{1}{\epsilon}\right)^{\frac{2(1-\gamma)}{\gamma}}$ when $0 < \gamma < 1$ and by order $\log(1/\epsilon)$ when $\gamma = 1$. Note that when P_X is known and $\sigma^2 = 1$, the set $\mathcal{F}(\{b_n^2\}; \delta)$ can be identified (theoretically speaking). We now denote $\mathcal{F}(\{b_n^2\}; \delta)$ by $\mathcal{F}(A; \{b_n^2\}; \delta)$ in this section since different values will be considered for A. For $0 < \gamma < 1$, with $\epsilon_n = C^{1/2} n^{-\gamma/2}$, from the derivation in the proof of Theorem 2, we have

$$M\left(\epsilon_n; \mathcal{F}(A; \{b_n^2\}; \delta)\right) \le C' \, n^{(1-\gamma)},$$

where C' can be taken as $\frac{A^2 + 2\log 2 + C}{1-\gamma} + 1$. Let

$s = (A, C, \gamma)$. It follows that, with P_X known, we can find (again theoretically speaking) a covering set $N(A, C, \gamma) = \{f_{s,1}, ..., f_{s,J_s}\}$, with $f_{s,i}$ bounded between $-A$ and A and $J_s \leq \exp(C^{'} n^{(1-\gamma)})$, satisfying that for any $f \in \mathcal{F}(A; \{b_n^2\}; \delta)$, there exists $1 \leq i \leq J_s$ such that $\| f - f_{s,i} \|_2 \leq \epsilon_n$. Similarly, when $\gamma = 1$, we can find a covering set $N(A, C, \gamma) = \{f_{s,1}, ..., f_{s,J_s}\}$, with $f_{s,i}$ bounded between $-A$ and A and $J_s \leq n^{C''}$ for some constant $C'' > 0$, such that for any $f \in \mathcal{F}(A; \{b_n^2\}; \delta)$, there exists $1 \leq i \leq J_s$ with $\| f - f_{s,i} \|_2 \leq C^{1/2} n^{-1/2}$. Let $\Phi_s = N(A, C, \gamma)$.

Let $S = \{s = (A, C, \gamma) : A \in (0, \infty),\ C \in (0, \infty),\ \gamma \in (0, 1]\}$. Let Q denote the set of positive dyadic rationals. Let S_D consist of all the points (A, C, γ) in S with $A, C, \gamma \in Q$.

Now consider $T = \{t = (s, j),\ 1 \leq j \leq J_s,\ s \in S_D\}$. Clearly T is countable. Let δ_t, $t \in T$ be the procedure that gives estimator $\widehat{f}_{\delta_t, i} \equiv f_{s,j}$ for all $1 \leq i \leq n$. Then for $f \in \mathcal{F}(A; \{Cn^{-\gamma}\}; \delta)$ with $s \in S_D$, there exists $1 \leq j \leq J_s$ such that

$$R(f; n; \delta_t) \leq Cn^{-\gamma}. \tag{5}$$

The procedures $\{\delta_t\}$ will be combined appropriately to have a small risk.

We assign prior weights $\{\pi_t, t \in T\}$ based on a description of the index t of the classes according to information theory (see, e.g., Rissanen (1983)). For every dyadic rational number q, it can be written as $q = i(q) + \sum_{j=1}^{l(q)} a_j(q) 2^{-j}$ for some $l \geq 1$, a_j's being either 0 or 1, and i is the integer part of q. To describe such a q, we just need to describe the integers i, l, and the a_j's. To describe integer $i \geq 0$, we may use $\log^*(i) =: \log_2(i+1) + 2\log_2(\log_2(i+2))$ bits. Then describe l using $\log^*(l)$ bits, and finally describe a_j's using l bits. By this way, we describe the hyperparameter components A, C, γ and use $\log_2 J_s$ bits to describe j for $t \in T$. The total description length for t then is

$$\log^*(i(A)) + \log^*(l(A)) + l(A) + \log^*(i(C)) + l(C) + \log^*(l(C)) + \log^*(i(\gamma)) + \log^*(l(\gamma)) + l(\gamma) + \log_2 J_s.$$

The prior weight of the procedure δ_t in the countable collection is then assigned to be π_t with $-\log_2 \pi_t$ equal the above expression. The coding interpretation guarantees $\{\pi_t : t \in T\}$ is a sub-probability (see, e.g., Cover and Thomas (1991, p. 52)), i.e., $\sum_{t \in T} \pi_t \leq 1$. One can either normalize π_t to be a probability or put the remaining probability on any chosen procedure δ_t, which does not have any effect on rates of convergence.

Now we combine these procedures based on the three-stage ARM method in Yang (2001) using the prior weights described above to get $\widehat{f}_n^*$. Note that the combined estimator $\widehat{f}_n^*$ is a convex combination of $f_{s,j}$'s with $1 \leq j \leq J_s$, $s \in S_D$. It is an estimator based on sparse approximation systems $\{\Phi_s : s \in S_D\}$. The risk of the combined procedure $\tilde{\delta}$ satisfies that for any f with $\| f \|_\infty \leq A$ (A needs not to be known) and if $\sigma^2 \leq \overline{\sigma}^2$, then

$$E\|f - \widehat{f}_n^*\|^2 \leq C_{A,\overline{\sigma}} \inf_{t \in T} \left(\frac{1}{n} \log \frac{1}{\pi_t} + \frac{1}{n} + R(f; \lfloor n/2 \rfloor; \delta_t) \right). \tag{6}$$

Fix $s_0 = (A_0, C_0, \gamma_0) \in S$ with $0 < \gamma_0 < 1$. For each $m \geq -\log_2 \gamma_0$, there exists $s^{(m)} = (A^{(m)}, C^{(m)}, \gamma^{(m)}) \in S_D$ such that $A_0 \leq A^{(m)} \leq A_0 + 2^{-m}$, $C_0 \leq C^{(m)} \leq C_0 + 2^{-m}$, and $\gamma_0 - 2^{-m} < \gamma^{(m)} \leq \gamma_0$. For $t = (s^{(m)}, j)$ with $1 \leq j \leq J_{s^{(m)}}$, noting that $i(\gamma) = 0$ for $0 < \gamma < 1$, the prior weight satisfies that $\log \frac{1}{\pi_t}$ is upper bounded by

$$\begin{aligned} & \log^*(A_0 + 1) + \log^*(C_0 + 1) + 1 + \\ & 3\log^* m + 3m + C^{'} n^{(1-\gamma^{(m)})} \\ \leq\ & Cm + C^{'} n^{(1-\gamma^{(m)})}, \end{aligned} \tag{7}$$

where C is a constant depending on A_0 and C_0. Note that $f \in \mathcal{F}(A_0; \{C_0 n^{-\gamma_0}\}; \delta)$ implies that $f \in \mathcal{F}(A^{(m)}; \{C^{(m)} n^{-\gamma^{(m)}}\}; \delta)$. Then from (5), (6) and (7), we have that for $f \in \mathcal{F}(A_0; \{C_0 n^{-\gamma_0}\}; \delta)$,

$$R(f; n; \widetilde{\delta}) \leq \widetilde{C} \left(\frac{m}{n} + n^{-\gamma^{(m)}} \right),$$

where $\widetilde{C}$ is a constant depending on A_0, C_0 and $\overline{\sigma}$. Take m of order $\log n$, observing that $n^{-\gamma^{(m)}}$ is then of order $n^{-\gamma_0}$, we have that for $f \in \mathcal{F}(A_0; \{C_0 n^{-\gamma_0}\}; \delta)$, $R(f; n; \tilde{\delta}) \leq \widetilde{\widetilde{C}} n^{-\gamma}$ for some constant $\widetilde{\widetilde{C}}$ not depending on n.

For $s_0 = (A_0, C_0, \gamma_0) \in S$ with $\gamma_0 = 1$, there exists $s_1 = (A_1, C_1, 1) \in S_D$ such that $A_0 \leq A_1$ and $C_0 \leq C_1$. For $t = (s_1, j)$ with $1 \leq j \leq J_{s_1}$, applying the metric entropy bound in Theorem 2 for the case $\gamma = 1$, we have that the prior weight satisfies

$$\log \frac{1}{\pi_t} \leq C''' \log n,$$

where C''' is a constant depending only on s_1. It follows, similarly as the case with $0 < \gamma < 1$, that for $f \in \mathcal{F}(A_0; \{C_0 n^{-1}\}; \delta)$, when $\sigma^2 \leq \overline{\sigma}^2$, we have $R(f; n; \tilde{\delta}) \leq \widetilde{C} \log n / n$ for some constant $\widetilde{C} > 0$ not depending on n.

Now we assume that P_X is unknown but known to have a probability density $f_X(x)$ with respect to a probability measure μ with f_X bounded above and away from zero. Then for any function g,

$$\underline{C} \int g(x)^2 \mu(dx) \leq \int g(x)^2 P_X(dx) \leq \overline{C} \int g(x)^2 \mu(dx)$$

for some constants $0 < \underline{C} < \overline{C} < \infty$. It follows that $R(f; n; \delta)$ under design distribution P_X is bounded above and below by multiples of $R(f; n; \delta)$ under design distribution μ. One can then modify the derivation above slightly to show the conclusion still holds with the relaxed condition on P_X. This completes the proof of Theorem 3.

Acknowledgments

Supported by NSF CAREER grant 0094323, the first version of the work appeared in the pre-print series of 2000 at Department of Statistics, Iowa State University.

References

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

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

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

A.R. Barron and N. Hengartner (1998). Information theory and superefficiency. *Ann. Statist.* **26**, 1800-1825.

L. Birgé (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields* **71**, 271-291.

J. Bretagnolle and C. Huber (1979). Estimation des densites: risque minimax. *Z. Wahrscheinlichkeitstheor. Verw. Geb.,* **47**, 119-137.

L. Devroye (1982). Necessary and Sufficient Conditions for the Pointwise Convergence of Nearest Neighbor Regression Function Estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 467–481.

L. Devroye and T.J. Wagner (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Stat.,* **8**, 231-239.

D.L. Donoho (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, **1**, 100-115.

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

S. Efromovich and M.S. Pinsker (1984). A self-educating nonparametric filtration algorithm. *Automation and Remote Control,* **45**, 58-65.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky (1999). Bayesian model averaging: A tutotial. *Statistical Science*, **14**, 382-417.

I.A. Ibragimov and R.Z. Hasminskii (1977). On the estimation of an infinite-dimensional parameter in Gaussian white noise. *Soviet Math. Dokl.,* **18**, 1307-1309.

I. Johnstone (1999). *Function Estimation in Gaussian Noise: Sequence Models.* Book Manuscript.

G. Kerkyacharian and D. Picard (2002). Minimax or maxisets? *Bernoulli*, **8**, 219-253.

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

G.G. Lorentz, M.v. Golitschek, and Y. Makovoz (1996). *Constructive Approximation: Advanced Problems*, Springer, New York.

G. Lugosi and A. Nobel (1999). Adaptive model selection using empirical complexities. *Ann. Statistics,* **27**, 1830-1864.

C.J. Stone (1977). Consistent nonparametric regression. *Ann. Statist.*, **8**, 1348-1360.

C.J. Stone (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.

M. Wegkamp (2003). Model selection in nonparametric regression. *Ann. Statist.*, **31**, 252-273.

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

Y. Yang (2000a). Mixing strategies for density estimation. *Ann. Statistics*, **28**, 75-87.

Y. Yang (2000b). Combining Different Procedures for Adaptive Regression. *J. Mult. Anal.*, **74**, 135-161.

Y. Yang (2001). Adaptive regression by mixing. *Journal of American Statistical Association*, **96**, 574-588.

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

Y. Yang and A.R. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statistics*, **27**, 1564-1599.

Y.G. Yatracos (1988). A lower bound on the error in nonparametric regression type problems. *Ann. Statist.*, **16**, 1180-1187.