# Minimax Nonparametric Classification—Part I: Rates of Convergence

Yuhong Yang

***Abstract*— This paper studies minimax aspects of nonparametric classification. We first study minimax estimation of the conditional probability of a class label, given the feature variable. This function, say $f$, is assumed to be in a general nonparametric class. We show the minimax rate of convergence under square $L_2$ loss is determined by the massiveness of the class as measured by metric entropy.**

**The second part of the paper studies minimax classification. The loss of interest is the difference between the probability of misclassification of a classifier and that of the Bayes decision. As is well known, an upper bound on risk for estimating $f$ gives an upper bound on the risk for classification, but the rate is known to be suboptimal for the class of monotone functions. This suggests that one does not have to estimate $f$ well in order to classify well. However, we show that the two problems are in fact of the same difficulty in terms of rates of convergence under a sufficient condition, which is satisfied by many function classes including Besov (Sobolev), Lipschitz, and bounded variation. This is somewhat surprising in view of a result of Devroye, Györfi, and Lugosi (1996).**

***Index Terms*—Conditional probability estimation, mean error probability regret, metric entropy, minimax rates of convergence, nonparametric classification, neural network classes, sparse approximation.**

## I. Introduction

In this paper, we study two related problems of minimaxity in nonparametric classification. For simplicity, consider the two-class case with class labels $Y \in \{0,1\}$. Direct extensions to cases with multiple classes are straightforward.

We observe $Z_i = (X_i, Y_i)$, $i = 1, \cdots, n$, which are independent copies of the random pair $Z = (X, Y)$. Let $f(x) = P\{Y = 1 | X = x\}$ be the conditional probability of the event $Y = 1$ given the feature variable $X = x \in \mathcal{X}$. Here $\mathcal{X}$ is the feature space, which could be of high dimension. Let $h(x)$ denote the marginal density of $X$ with respect to a $\sigma$-finite measure $\mu$. We are interested in both how well one can estimate $f$ and how well one can classify based on a feature value.

Many results have been obtained for nonparametric classification in the literature (see Devroye, Györfi, and Lugosi [16] for a review). A surprising result is that universally consistent estimators of $f$ and classifiers exist (see, e.g., [15], [17], [27], and [32]). However, the convergence could be arbitrarily slow ([16, Ch. 7]). If one knows *a priori* that the target function $f$ belongs to a nonparametric class of functions, uniform convergence rates are possible. A few methods have been shown to converge at certain rates when $f$ is in some nonparametric class (e.g., [2], [5], [6], [18], [19], and [24]). A general upper bound on mean error probability for classification is given in [16, Ch. 28] in terms of metric entropy. On the other hand, lower bound results seem to be rare. Optimal rates of convergence in probability for classification were identified in a related setting for some Lipschitz classes [29]. This paper aims to provide a general understanding of the minimax rate of convergence of the risks for general nonparametric function classes.

### A. Minimax Risk for Estimating $f$

We measure loss of an estimator of $f$ in terms of a square norm. Let $\|\cdot\|_{L_2(h)}$ denote the $L_2$ norm weighted by the density of the feature random variable $X$, i.e., for any $g$

$$\|g\|_{L_2(h)} = \left( \int (g(x))^2 h(x) \mu(dx) \right)^{1/2}.$$

Similarly, define the $L_q(h)$ norm $(q \geq 1)$. Call the corresponding distance $L_q(h)$ distance. Let $\hat{f}$ be an estimator of $f$ based on $Z^n = (X_i, Y_i)_{i=1}^n$. The risk then is

$$R(f; \hat{f}; n) = E \int (f(x) - \hat{f}(x))^2 h(x) \mu(dx)$$

where the expectation is taken with respect to the true $f$. Since $f$ is always between 0 and 1, the risk $R(f; \hat{f}; n)$ is always well-defined. Let $\mathcal{F}$ be a class of candidate conditional probability functions, i.e., every $g \in \mathcal{F}$ satisfies $0 \leq g(x) \leq 1$ for all $x \in \mathcal{X}$. Then the minimax risk under the square $L_2(h)$ loss for estimating a conditional probability in $\mathcal{F}$ is

$$R(\mathcal{F}; n) = \min_{\hat{f}} \max_{f \in \mathcal{F}} R(f; \hat{f}; n)$$

where $\hat{f}$ is over all valid estimators based on $Z^n$ (here "min" and "max" are understood to be "inf" and "sup," respectively, if the minimizer or maximizer does not exist). In this work, $\mathcal{F}$ is assumed to be a nonparametric class. The minimax risk describes how well one can estimate $f$ uniformly over the function class $\mathcal{F}$.

For density estimation and nonparametric regression under global losses, it is known that rates of convergence of minimax risks of function classes are determined by Kolmogorov's metric entropy of the classes ([7], [21], [36], [38], and others).

Manuscript received March 26, 1998; revised January 7, 1999.

The author is with the Department of Statistics, Iowa State University, Ames, IA 50011-1210 USA.

Communicated by P. Moulin, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(99)08515-6.

As will be shown, it is also the case for the estimation of the function of conditional probability. Our characterization of the minimax rate of convergence in terms of metric entropy enables one to derive minimax rates of convergence for many function classes. Examples are given in Section V.

### B. Minimax Risk for Classification

For classification (or pattern recognition), the goal is to have a classifier which predicts membership of $Y$ according to the feature variable $x$. Formally, a classifier $\delta$ based on the training data $Z^n$ is a mapping from $\mathcal{X} \times \{\mathcal{X} \times \{0,1\}\}^n$ to $\{0,1\}$. As is well known, given the true conditional probabilit function $f(x)$, an ideal optimal Bayes decision is to predict $Y$ as class 1 if $f(x) \geq 1/2$ and class 0 otherwise. This decision minimizes the probability of error $P\{Y \neq \nu(X)\}$ over all choices of classifier $\nu$. Let $e^*$ denote the corresponding error probability. For a given classifier $\delta = \delta(x; Z^n)$ based on $Z^n$, the mean error probability is

$$EP(Y \neq \delta(X; Z^n)|Z^n).$$

We will examine

$$r(f; \delta; n) = EP\{Y \neq \delta(X; Z^n)|Z^n)\} - e^* \tag{1}$$

which is the difference between the mean error probability of $\delta$ and the ideal one $e^*$. We call it mean error probability regret (MEPR). For a given class $\mathcal{F}$ of conditional probability functions, let the minimax MEPR be defined as

$$r(\mathcal{F}; n) = \min_{\delta} \max_{f \in \mathcal{F}} r(f; \delta; n)$$

where the minimization is over all classifiers based on $Z^n$. This risk describes how well one can classify $Y$ relative to the Bayes decision, uniformly over $\mathcal{F}$.

### C. Is Classification Much Easier Than the Estimation of Conditional Probability?

Based on an estimator $\hat{f}$ of $f$, one can define a plug-in classifier $\delta_{\hat{f}}$ pretending $\hat{f}$ is the true conditional probability, i.e.,

$$\delta_{\hat{f}}(x) = \begin{cases} 1, & \text{if } \hat{f}(x) \geq 1/2 \\ 0, & \text{if } \hat{f}(x) < 1/2. \end{cases}$$

Then it is well known that (see, e.g., [16, p. 93])

$$\begin{aligned} r(f; \delta_{\hat{f}}; n) &\leq 2E \int h(x)|f(x) - \hat{f}(x)|\mu(dx) \\ &\leq 2E\left( \int h(x)(f(x) - \hat{f}(x))^2 \mu(dx) \right)^{1/2} \\ &\leq 2\sqrt{R(f; \hat{f}; n)}. \end{aligned}$$

As a consequence,

$$r(\mathcal{F}; n) \leq 2\sqrt{R(\mathcal{F}; n)}.$$

Thus a minimax upper bound for estimating $f$ immediately gives an upper bound on the minimax classification risk MEPR over $\mathcal{F}$. But how good is this upper bound? Let $\mathcal{F}$ be the class of monotone functions. Then $\sqrt{R(\mathcal{F}; n)}$ is of order $n^{-1/3}$ (see Section V), but $r(\mathcal{F}; n) = O((\log n/n)^{1/2})$ [16, p. 485], which is much smaller than $\sqrt{R(\mathcal{F}; n)}$. (Here a referee pointed out that $r(\mathcal{F}; n)$ can be shown to be of order $n^{-1/2}$. The upper bound part can be derived by empirical risk minimization applying Alexander's inequality [16, p. 207] with the fact that the VC dimension of the class $\{I_{\{x \leq a\}}, a \in R\}$ is 1. The techniques used in [16, the proof of Theorem 14.5] may be used to show the rate $n^{-1/2}$ cannot be improved.)

This makes one wonder if this phenomenon is typical for classical function classes such as Sobolev, Lipschitz, Besov, and bounded variation. It turns out the answer is no. We give a sufficient condition for the minimax rates of convergence for the two problems to match. This condition is satisfied by many function classes. Then classification is no easier than estimating the conditional probability $f$ in a uniform sense. This result is rather surprising when compared with [16, Theorem 6.5], which says that for any fixed $f$ and a sequence of consistent estimators $\hat{f}_n$, one always has

$$\frac{r(f; \delta_{\hat{f}_n}; n)}{2\sqrt{R(f; \hat{f}_n; n)}} \to 0.$$

The explanation of this phenomenon is that the above ratio does not converge uniformly to 0 over the function classes.

### D. Model Selection for Minimax Adaptive Estimation of $f$

In the derivation of minimax upper bounds, $\epsilon$-nets are used in the construction of estimators achieving the optimal rates. These estimators are convenient for theoretic studies, but are not practical for applications. The sequel of this paper, "Minimax Nonparametric Classification—Part II: Model Selection for Adaptation" (see this issue) presents results on minimax adaptive estimation of the conditional probability $f$ by model selection over a countable collection of finite-dimensional approximating models. There it is shown that using a suitable model selection criterion, minimax rates of convergence are automatically achieved (or nearly achieved) simultaneously over different types of function classes and/or with different smoothness parameters.

This paper is organized as follows. In Section II, minimax rates for estimating $f$ are given; in Section III, minimax MEPR is given under a sufficient condition; in Section IV, we present some results on the relationship between approximation and classification; in Section V, examples are given as direct applications of the main results in Sections II and III; in Section VI, we give a simple result on minimax rates for some classes modified to allow some irregularity such as discontinuity. The proofs of the results are given in Section VII.

## II. Minimax Rates for Estimating the Conditional Probability

### A. Metric Entropy

A finite subset $N_\epsilon$ is called an $\epsilon$-packing set in $\mathcal{F}$ under a distance $d$ if $d(u, v) > \epsilon$ for any $u, v \in N_\epsilon$ with $u \neq v$. Let $M_2(\epsilon) = M_2(\epsilon; \mathcal{F})$ be the maximal logarithm of the cardinality of any $\epsilon$-packing set in $\mathcal{F}$ under $L_2(h)$ distance.

The asymptotic behavior of $M_2(\epsilon)$ when $\epsilon \to 0$ reflects how massive the class $\mathcal{F}$ is (under the chosen distance). We call $M_2(\epsilon)$ the packing $\epsilon$-entropy or simply the metric entropy of $\mathcal{F}$. Similarly, define $M_q(\epsilon)$ under the distance $L_q(h)$. A lot of results have been obtained on the orders of metric entropy for the classical function classes and many more under various norms (see, e.g., [9], [23], [25], [26], and [34]).

Assume $M_2(\epsilon) < \infty$ for every $\epsilon > 0$ (otherwise, the minimax risk typically does not converges to zero) and $M_2(\epsilon) \to \infty$ as $\epsilon \to 0$ (which excludes the trivial case when $\mathcal{F}$ is finite). For most function classes, the metric entropies are known only up to orders. For that reason, we assume that $M(\epsilon)$ is an available nonincreasing right-continuous function known to be of order $M_2(\epsilon)$. We call a class $\mathcal{F}$ rich in $L_2(h)$ distance if for some constant $0 < \tau < 1$,

$$\liminf_{\epsilon \to 0} M(\tau\epsilon)/M(\epsilon) > 1. \tag{2}$$

This condition is characteristic of usual nonparametric classes (see [36]), for which the metric entropy is often of order $\epsilon^{-\alpha} \log(1/\epsilon)^{\beta}$ for some $\alpha > 0$ and $\beta \in R$, see Section V.

### B. Minimax Risk Bounds

We need some conditions for our minimax risk bounds based on metric entropy.

*Assumption 1:* The class $\mathcal{F}$ is convex and contains at least one member $f^*$ that is bounded away from 0 and 1, i.e., there exist constants $0 < \underline{c} \le \overline{c} < 1$ such that $\underline{c} \le f^* \le \overline{c}$.

A couple of quantities will appear in our minimax risk bounds. Choose $\epsilon_n$ such that

$$M_2(2\epsilon_n; \mathcal{F}) = n\epsilon_n^2. \tag{3}$$

Let $\underline{\epsilon}_n$ be chosen to satisfy

$$M_2(2\underline{\epsilon}_n; \mathcal{F}) = 2\left(1 + \frac{4}{\underline{c}(1-\overline{c})}\right) n\underline{\epsilon}_n^2 + 2\log 2. \tag{4}$$

(Since the packing entropy is right-continuous, under the assumption $M_2(\epsilon) \to \infty$, both $\epsilon_n$ and $\underline{\epsilon}_n$ are well-defined.) Similarly, define $\underline{\epsilon}_{n,q}$ with $M_2(\epsilon; \mathcal{F})$ replaced by $M_q(\epsilon; \mathcal{F})$.

The notation $a_n \preceq b_n$ will be used to mean

$$\limsup(a_n/b_n) < \infty.$$

When $a_n \preceq b_n$ and $b_n \preceq a_n$, i.e., $a_n$ and $b_n$ are of the same order, we use expression $a_n \asymp b_n$. If $a_n$ and $b_n$ are asymptotically equivalent, i.e., $\lim(a_n/b_n) = 1$, then we write $a_n \sim b_n$.

*Lemma 1:* Assume Assumption 1 is satisfied. For any $l > 0$, we have the following minimax lower bound on $L_q(h)$ $(q \ge 1)$ risk for estimating $f$:

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E\|f - \hat{f}\|_{L_q(h)}^l \ge \underline{\epsilon}_{n,q}^l / 2^{l+1}.$$

For upper bound, if $h$ is known, then we have

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E\|f - \hat{f}\|_{L_2(h)}^2 \le 102\epsilon_n^2.$$

*Remarks:*

1) For the upper bound part, Assumption 1 is not needed.
2) When $M_2(\epsilon; \mathcal{F})$ in (3) and (4) is replaced by an upper bound and a lower bound, respectively, the resulting bounds in Lemma 1 are valid.

In Lemma 1, for the upper bound, $h$ is assumed to be known. This rather restrictive condition is used for the purpose that an $\epsilon$-net can be constructed (theoretically) for $\mathcal{F}$ under $L_2(h)$ distance as needed in the proof. In practical problems, of course, $h$ is not known. To get the right rate of convergence, it suffices to be able to construct an $\epsilon$-net with log-cardinality of the same order. The following assumption will be used instead.

*Assumption 2:* For each (small) $\epsilon$, without knowing $h$ exactly, one can construct an $\epsilon$-net of $\mathcal{F}$ under $L_2(h)$ distance with log-cardinality of order $M(\epsilon)$.

When $h$ is unknown but known to be in a class $\mathcal{H}$, it is also appropriate to study minimax risk over $\mathcal{H}$ as well as over the class of conditional probability $\mathcal{F}$. Let

$$R(\mathcal{F}; \mathcal{H}; n) = \min_{\hat{f}} \max_{h \in \mathcal{H}} \max_{f \in \mathcal{F}} E_{h,f}\|f - \hat{f}\|_{L_2(h)}^2.$$

*Assumption 2′:* For each (small) $\epsilon$, one can construct an $\epsilon$-net for $\mathcal{F}$ under the $L_2(h)$ distance simultaneously over $h \in \mathcal{H}$ with log-cardinality of order $\overline{M}(\epsilon)$ for some right-continuous function $\overline{M}$. In addition, the packing entropy of $\mathcal{F}$ is of order $\overline{M}(\epsilon)$ under $L_2(h)$ distance for at least one $h \in \mathcal{H}$.

Satisfaction of Assumption 2 or 2′ requires some knowledge of $h$. See Section V for examples that satisfy these conditions.

Now we give results on minimax rates based on Lemma 1. Let $\epsilon_n$ satisfy

$$\overline{M}(\epsilon_n) = n\epsilon_n^2.$$

*Theorem 1:* Assume Assumptions 1 and 2 are satisfied and that $\mathcal{F}$ is rich in $L_2(h)$ distance.

1) For the square $L_2(h)$ risk, we have
$$R(\mathcal{F}; n) \asymp \epsilon_n^2.$$
2) If $M_q(\epsilon)$ and $M_2(\epsilon)$ are of the same order as $\epsilon \to 0$, then
$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E\|f - \hat{f}\|_q \asymp \epsilon_n.$$
3) Under Assumption 2′, we have
$$R(\mathcal{F}; \mathcal{H}; n) \asymp \epsilon_n^2. \tag{5}$$

*Remarks:*

1) Instead of Assumption 1, it suffices to assume that $\mathcal{F}$ contains a subset $\mathcal{F}_0$ that is uniformly bounded away from 0 and 1, and $\mathcal{F}_0$ has the same order metric entropy as $\mathcal{F}$.
2) The exact risk bounds may sometimes be of interest. Let $\tilde{M}(\epsilon)$ be the log-cardinality of the $\epsilon$-net that one can construct under Assumption 2. Let $\tilde{\epsilon}_n$ be determined by $\tilde{M}(2\tilde{\epsilon}_n) = n\tilde{\epsilon}_n^2$. Then we have
$$\underline{\epsilon}_n^2/8 \le R(\mathcal{F}; n) \le 102\tilde{\epsilon}_n^2.$$
Similar bounds hold for $R(\mathcal{F}; \mathcal{H}; n)$.

The condition that the packing entropies under $L_2$ and $L_q$ are of the same order are satisfied in case of many familiar nonparametric classes (see [26]).

The following corollary is obtained under some simple sufficient conditions for Assumption 2′.

*Corollary 1:* Suppose that Assumption 1 is satisfied. Assume there exists a density $h^*$ and a constant $C>0$ such that for every $h \in \mathcal{H}$, $h/h^* \le C$, and that $\mathcal{F}$ is rich in $L_2(h^*)$ distance with metric entropy of order $M^*(\epsilon)$. In addition, there exists at least one $h \in \mathcal{H}$ such that the $L_2(h)$ metric entropy of $\mathcal{F}$ is of order $M^*(\epsilon)$. Then

$$R(\mathcal{F};\mathcal{H};n) \asymp \epsilon_n^2$$

where $\epsilon_n$ is determined by $M^*(\epsilon_n) = n\epsilon_n^2$.

In particular, if there exists a $h^*$ such that

$$\sup_{h\in\mathcal{H}} ||\log(h/h^*)||_\infty < \infty$$

(i.e., $h/h^*$ is uniformly bounded above and away from zero), and $\mathcal{F}$ is rich in $L_2(h^*)$ distance, then

$$R(\mathcal{F};\mathcal{H};n) \asymp \epsilon_n^2$$

where $\epsilon_n$ is determined as above.

## III. Minimax Rates of Convergence of Error Probability

From the upper-bound rate $\epsilon_n^2$ on $R(\mathcal{F};n)$ in Theorem 1, we know the minimax MEPR $r(\mathcal{F};n)$ is upper-bounded by order $\epsilon_n$. A similar upper bound in terms of metric entropy was obtained in [16, Ch. 28] using an argument directly for MEPR. However, as mentioned in Section I, $r(\mathcal{F};n)$ may converge much faster than $\sqrt{R(\mathcal{F};n)}$. The interest is then on lower-bounding $r(\mathcal{F};n)$. One difficulty is that the loss

$$l(f;\delta) = P(Y \ne \delta(X;Z^n)|Z^n) - e^*$$

(which yields risk $r(f;\delta;n)$ when averaged with respect to $Z^n$) is not a metric in $\delta$ (it is not symmetric and there is no triangle-like inequality). It makes the notions of covering and packing (as used for estimating $f$) very tricky and hard to compute even if they make sense. The next result takes advantage of the observation that for many function classes, on a subclass (hypercubes) constructed from a suitable perturbation, the loss does behave like a metric and, therefore, the previous lower-bounding technique for $L_2(h)$ risk to estimate $f$ also works. The following setup of a hypercube class is in [8], [11], and others used for density estimation.

We assume $h$ is upper-bounded by a known constant $\overline{c}_1<\infty$ on $\mathcal{X}$.

*Assumption 3:* For any $\epsilon \in (0,1)$, there exist some function $g_\epsilon$ with support on $I \subset \mathcal{X}$, and $r = r_\epsilon$ disjoint translates $I+x_i$ such that the hypercube family

$$\mathcal{F}_{\text{cube}} = \left\{ f_\tau(x) = 1/2 + \sum_{i=1}^{r} \tau_i g_\epsilon(x-x_i): \right.$$
$$\left. \tau = (\tau_1,\cdots,\tau_r) \in \{-1,1\}^r \right\}$$

belongs to $\mathcal{F}$. Also $g_\epsilon$ satisfies $||g_\epsilon||_\infty \le \epsilon$ and $\mu(|g_\epsilon| \ge A_1\epsilon) \ge A_2 v_\epsilon$ for some constants $0<A_1,\ A_2<1$ with $\mu(I) = v_\epsilon$. Furthermore, there exists a nonincreasing right-continuous function $\underline{M}(\epsilon)$ with $\underline{M}(\epsilon) \to \infty$ as $\epsilon \to 0$ such that $v_\epsilon \le A_3/\underline{M}(\epsilon),\ 0<A_4 \le rv_\epsilon \le 1$ for some constants $A_3$ and $A_4$.

Assumption 3 is verified in [7], [11], and others for many function classes (see Section V).

The subclass of hypercubes is intended to capture the difficulty of classification for the whole class $\mathcal{F}$. The subset is very simple and easy to handle. Note that the perturbations are around $f^*(x) \equiv 1/2$. When $\underline{M}(\epsilon)$ is of the same order as the $L_2(h)$ metric entropy of $\mathcal{F}$, classification on $\mathcal{F}_{\text{cube}}$ alone has difficulty already matching that of estimating $f \in \mathcal{F}$, resulting in the determination of rate of $r(\mathcal{F};n)$ as in the following theorem.

*Theorem 2:* Assume Assumptions 2 is satisfied and that $\mathcal{F}$ is rich in $L_2(h)$ distance with metric entropy of order $M(\epsilon)$. Suppose Assumption 3 is satisfied with $\underline{M}(\epsilon)$ of the same order as $M(\epsilon)$ and that there exists a constant $\underline{c}_1>0$ such that $h \ge \underline{c}_1$ on the $r_\epsilon$ translates $I+x_i$ for each (small) $\epsilon$. Then the minimax mean error probability regret of $\mathcal{F}$ has rate

$$r(\mathcal{F};n) \asymp \epsilon_n$$

where $\epsilon_n$ is determined by $M(\epsilon_n) = n\epsilon_n^2$. If instead of Assumption 2, Assumption 2′ is satisfied for a class of densities $\mathcal{H}$ with $\overline{M}(\epsilon) \asymp \underline{M}(\epsilon)$ and containing at least one member $h$ with $h \le \overline{c}_1<\infty$ on $\mathcal{X}$ and $h \ge \underline{c}_1>0$ on $\cup_{1\le i\le r_\epsilon}\{I+x_i\}$ for each (small) $\epsilon$, then

$$r(\mathcal{F};\mathcal{H};n) =: \min_\delta \max_{h\in\mathcal{H}} \max_{f\in\mathcal{F}} r(f;\delta;n) \asymp \epsilon_n.$$

## IV. Approximation and Classification

Many different approximating systems can be used to estimate the function of conditional probability $f$ for classification. For instance, polynomial, trigonometric, spline, wavelet, or neural net approximations have different advantages and are useful in model building for the estimation of $f$. It is intuitively clear that given an approximation system, how well one can estimate $f$ based on a training data is related to how well the function can be approximated by the system. In this section, we establish formal conclusions on relationships between linear approximation, sparse approximation, and minimax rates of convergence for classification. Similar, but more general and detailed treatments in the context of density estimation and regression are in [36]. Let $h^*$ be a known density on $\mathcal{X}$. We assume that the true density satisfies

$$h/h^* \le A$$

for a known constant $A \ge 1$. Let $\mathcal{H}(A)$ be the collections of all such densities.

### A. Linear Approximation and Rates of Convergence

Let $\Phi = \{\phi_1 = 1,\cdots,\phi_k,\cdots\}$ be a chosen fundamental sequence in

$$L^2(\mathcal{X};h^*) = \{g: ||g||_{L_2(h^*)} < \infty\}$$

(that is, linear combinations are dense in $L^2(\mathcal{X};h^*)$ with respect to $L_2(h^*)$ distance). Let $\Gamma = \{\gamma_0, \cdots, \gamma_k, \cdots\}$ for which $\gamma_k \downarrow 0$ as $k \to \infty$. Let $\eta_0(g) = ||g||_{L_2(h^*)}$, and for $k \geq 1$

$$\eta_k(g) = \min_{\{a_i\}} ||g - \sum_{i=1}^{k} a_i \phi_i||_{L_2(h^*)}$$

be the $k$th degree of approximation of $g \in L^2(\mathcal{X};h^*)$ by the system $\Phi$. Let $\mathcal{F}(\Gamma,\Phi)$ be all functions in $L^2(\mathcal{X};h^*)$ with the approximation errors bounded by $\Gamma$, i.e.,

$$\mathcal{F}(\Gamma,\Phi) = \{g \in L^2(\mathcal{X};h^*)\colon \eta_k(g) \leq \gamma_k, k = 0, 1, \cdots\}.$$

They are called full approximation sets of functions. Some classical function classes (e.g., Sobolev, general ellipsoidal) are essentially of this type.

Assume $\Gamma$ satisfies a condition that there exist $0 < c' < c < 1$ such that

$$c'\gamma_k \leq \gamma_{2k} \leq c\gamma_k \tag{6}$$

as is true for $\gamma_k \sim k^{-\alpha}(\log k)^{\beta}$, $\alpha > 0$, $\beta \in R$. Lorentz gives order of the $L_2(h^*)$ metric entropy and shows that $\mathcal{F}(\Gamma,\Phi)$ is rich [25, Theorem 4]. Then $\epsilon_n$ determined by $M(\epsilon_n;\mathcal{F}(\Gamma,\Phi)) = n\epsilon_n^2$ balances the approximation error bound $\gamma_k^2$ and the dimension over sample size $k/n$ [36].

Assume that the functions in $\mathcal{F}(\Gamma,\Phi)$ are uniformly bounded (see [36] on satisfaction of this condition). Let $\tilde{\mathcal{F}}(\Gamma,\Phi)$ be all the functions in $\mathcal{F}(\Gamma,\Phi)$ that are nonnegative and bounded above by 1. It can be shown that the $L_2(h^*)$ metric entropies of $\tilde{\mathcal{F}}(\Gamma,\Phi)$ and $\mathcal{F}(\Gamma,\Phi)$ are of the same order (see Lemma 4 in Section VII). By Corollary 1, we have the following result.

*Corollary 2:* Let $k_n$ be determined by $\gamma_k^2 \asymp k/n$. Then the minimax rates for classification satisfy

$$R(\mathcal{F}(\Gamma,\Phi);\mathcal{H}(A);n) \asymp k_n/n$$
$$r(\mathcal{F}(\Gamma,\Phi);\mathcal{H}(A);n) \preceq \sqrt{k_n/n}.$$

*Remark:* The condition that $\mathcal{F}(\Gamma,\Phi)$ is uniformly bounded is not needed for the upper bound rate on minimax MEPR.

An illustration of this result is as follows. For a system $\Phi$ (e.g., polynomial, trigonometric, or wavelet), consider the functions that can be approximated with polynomially decreasing approximation error $\gamma_k \sim k^{-\alpha}$, $\alpha > 0$. When $\alpha$ is not too small, $\mathcal{F}(\Gamma,\Phi)$ is often bounded. Then solving $\gamma_k^2 \asymp k/n$ gives order $n^{-2\alpha/(2\alpha+1)}$. Thus

$$R(\mathcal{F}(\Gamma,\Phi);\mathcal{H}(A);n) \asymp n^{-2\alpha/(1+2\alpha)}$$
$$r(\mathcal{F}(\Gamma,\Phi);\mathcal{H}(A);n) \preceq n^{-\alpha/(1+2\alpha)}.$$

Note that for the classification risk $r(\mathcal{F}(\Gamma,\Phi);\mathcal{H}(A);n)$, the rate $\sqrt{k_n/n}$ is not shown to be optimal in general. However, for some choices of smooth approximating systems such as polynomials or trigonometric functions, Assumption 3 is still satisfied, and the rate $n^{-\alpha/(1+2\alpha)}$ is optimal for classification for such cases.

From Corollary 2, the optimal convergence rate in the full approximation setting for estimating $f$ is of the same order as $k_n/n \asymp \min_k(\gamma_k^2 + k/n)$, which represents the familiar bias-squared plus variance tradeoff for mean-squared error. Of course, in applications, one does not know how well the underlying function can be approximated by the chosen system, which makes it impossible to know the optimal size $k_n$. This suggests the need of a good model selection criterion to choose a suitable size model to balance the two kinds of errors automatically based on data. Results on model selection for classification are in the sequel of this paper.

### B. Sparse Approximations and Minimax Rates

Full approximation utilizes all the basis terms up to certain orders. For slowly converging sequences $\gamma_k$, such as arise especially in high-dimensional function approximation, very large $k$ (e.g., exponential in dimension) is needed to get a good accuracy. It becomes of interest to examine approximation using a manageable-size subset of terms. This subset is sparse in comparison to the total that would be needed with full approximation.

Let $\Phi$ and $\Gamma$ be as in the previous section. Let $I_k > k$, $k \geq 1$ be a given sequence of integers satisfying $\liminf I_k/k = \infty$ and let $\mathcal{I}_k = \{1, 2, \cdots, I_k\}$ $(I_0 = 0)$. Denote by

$$\tilde{\eta}_k(g) = \min_{l_1 \in \mathcal{I}_1, \cdots, l_k \in \mathcal{I}_k} \min_{\{a_i\}} ||g - \sum_{i=1}^{k} a_i \phi_{l_i}||_{L_2(h^*)}$$

the $k$th degree of sparse approximation of $g \in L^2(\mathcal{X};h^*)$ by the system $\Phi$ (for a fixed choice of $\{I_1, I_2, \cdots\}$). Note here that roughly the $k$ terms used to approximate $g$ are allowed to be from $I_k$ basis functions. Let $\mathcal{S}(\Gamma,\Phi) = \mathcal{S}(\Gamma,\Phi,\{I_k\})$ be the set of all functions in $L^2(\mathcal{X};h^*)$ with sparse approximation errors bounded by $\Gamma$, i.e.,

$$\mathcal{S}(\Gamma,\Phi) = \{g \in L^2(\mathcal{X};h^*)\colon \tilde{\eta}_k(g) \leq \gamma_k, k = 0, 1, \cdots\}.$$

We call it a sparse approximation set of functions. Large $I_k$'s provide considerable more freedom of approximation.

The $\epsilon$-entropy of $\mathcal{S}(\Gamma,\Phi)$ satisfies [36]

$$M(\epsilon;\mathcal{F}(\Gamma,\Phi)) \leq M(\epsilon;\mathcal{S}(\Gamma,\Phi)) \preceq M(\epsilon;\mathcal{F}(\Gamma,\Phi))\log(\epsilon^{-1})$$

under the assumption $I_k \leq k^{\tau}$ for some possibly large constant $\tau > 1$. Note that the upper and lower bounds differ in a logarithmic factor. (We tend to believe that $M(\epsilon;\mathcal{S}(\Gamma,\Phi))$ is of higher order than $M(\epsilon;\mathcal{F}(\Gamma,\Phi))$ and the ratio might be right at order $\log(\epsilon^{-1})$ or a similar logarithmic factor (e.g., $\log^{1/2}(\epsilon^{-1})$)).

Suppose the functions in $\mathcal{F}(\Gamma,\Phi)$ are uniformly bounded. Let $\tilde{\mathcal{S}}(\Gamma,\Phi)$ be the set of all valid conditional probability functions in $\mathcal{S}(\Gamma,\Phi)$. Let $k_n$ satisfy

$$k_n/n \asymp \gamma_{k_n}^2. \tag{7}$$

Then based on Lemma 1 and Theorem 2, we have the following corollary. For simplicity, we assume that $k_n$ determined above satisfies $k_n \preceq n^{1-\tau}$ for some arbitrarily small positive $\tau$, as is true for $\gamma_k \sim k^{-\alpha}(\log k)^{\beta}$, $\alpha > 0$, $\beta \in R$.

*Corollary 3:* For the sparse approximation set $\tilde{\mathcal{S}}(\Gamma,\Phi)$, if $I_k \le k^\tau$ for some $\tau > 1$, then

$$k_n/n \preceq R(\tilde{\mathcal{S}}(\Gamma,\Phi);\mathcal{H}(A);n) \preceq k_{\lfloor n/\log n\rfloor}\log n/n.$$

For minimax MEPR, we have

$$r(\tilde{\mathcal{S}}(\Gamma,\Phi);\mathcal{H}(A);n) \preceq \sqrt{k_{\lfloor n/\log n\rfloor}\log n/n}.$$

As a special case, if $\gamma_k \sim k^{-\alpha}$, $\alpha > 0$, then

$$n^{-2\alpha/(1+2\alpha)} \preceq R(\tilde{\mathcal{S}}(\Gamma,\Phi);n) \preceq (n/\log n)^{-2\alpha/(1+2\alpha)}.$$

The upper and lower bound rates differ only in a logarithmic factor (we tend to believe that an extra logarithmic factor is necessary here for $R(\tilde{\mathcal{S}}(\Gamma,\Phi);n)$).

Sparse approximation provides much more flexibility yet does not give up much linearity. To achieve the same degree of approximation for all functions in a sparse approximation set, full approximation has to use many more terms. This has an implication for statistical estimation using the nested models corresponding to full approximation compared to subset models corresponding to sparse approximation. Since many more coefficients need to be estimated for the nested models, the variance of the final estimator is much bigger in general, resulting in a worse rate of convergence. For example, if $I_k = k^2$ and $\gamma_k \sim k^{-1}$, then under some conditions, it can be shown (see [37]) that the rate for estimating $f$ in the sparse approximation set based on the subset models is $O(n/\log n)^{-2/3}$, whereas the rate based on nested models is no faster than $n^{-1/2}$. Results on subset selection for classification are presented in the sequel of this paper [37].

We finally mention that from [25, Theorem 9], a full approximation set $\mathcal{F}(\Gamma,\Phi)$ cannot be better approximated beyond a constant factor with any other choices of basis $\Phi'$. The same conclusion carries over to a sparse approximation set $\mathcal{S}(\Gamma,\Phi)$.

## V. Examples

### A. Function Classes

We consider a few function classes including classical ones (ellipsoidal, Besov, etc.) and some other relatively new ones (e.g., neural network classes). For results on metric entropy orders of various function classes, see [26] and references cited there.

1) *Ellipsoidal classes in $L_2$*: Let $\{\phi_1,\phi_2,\cdots,\phi_k,\cdots\}$ be a complete orthonormal system in $L^2[0,1]$. For an increasing sequence of constants $b_k$ with $b_k \to \infty$, define an ellipsoidal class

$$\mathcal{E}(\{b_k\},C) = \left\{g = \sum_{i=1}^{\infty}\xi_i\phi_i\colon \sum_{i=1}^{\infty}\xi_i^2 b_i^2 \le C\right\}.$$

We here only consider the special case with $b_k = k^\alpha$ ($\alpha > 0$), for which the metric entropy is of order $\epsilon^{-1/\alpha}$ [30]. General treatment is similar to that in [36]. When $\alpha > 1/2$ and $\sup_{k\ge 1}||\phi_k||_\infty < \infty$, the functions in $\mathcal{E}(\{k^\alpha\},C)$ are uniformly bounded.

2) *Monotone, bounded variation, and Lipschitz classes*: The function class $BV(C)$ consists of all functions $g(x)$ on $[0,1]$ satisfying $||g||_\infty \le C$ and

$$V(g) := \sup\sum_{i=1}^{m}|g(x_{i+1}) - g(x_i)| \le C$$

where the supremum is taken over all finite sequences $x_1 < x_2 < \cdots < x_m$ in $[0,1]$. For $0 < \alpha \le 1$, let

$$\mathrm{Lip}_{\alpha,q}(C) = \{g\colon ||g(x+h)-g(x)||_q \le Ch^\alpha \text{ and } ||g||_q \le C\}$$

be a Lipschitz class (similar results hold when the Lipschitz condition applies to a derivative). When $\alpha > (1/q - 1/2)^+$, $q \ge 1$, the $L_2$ metric entropy is of order $\epsilon^{-1/\alpha}$ [9]. For $BV(C)$, with suitable modification of the value assigned at discontinuity points as in [14, Ch. 2], one has

$$\mathrm{Lip}_{1,\infty}(C) \subset BV(C) \subset \mathrm{Lip}_{1,1}(C).$$

So the $L_2$ metric entropy of $BV(C)$ is also of order $1/\epsilon$. Let $MI(C)$ be the set of all nondecreasing functions $g$ on $[0,1]$ such that $||g||_\infty \le C$. Using the fact that a function with bounded variation can be expressed as a difference between two monotone nondecreasing functions, it is easy to show that $MI(C)$ has $L_2$ metric entropy also of order $1/\epsilon$.

3) *Besov classes*: Let

$$\Delta_h^r(g,x) = \sum_{k=0}^{r}\binom{r}{k}(-1)^{r-k}g(x+kh).$$

Then the $r$th modulus of smoothness of $g \in L_q[0,1]$ $(0 < q < \infty)$ or of $g \in C[0,1]$ if $q = \infty$ is defined by

$$\omega_r(g,t)_q = \sup_{0<h\le t}||\Delta_h^r(g,\cdot)||_q.$$

Let $\alpha > 0$, $r = [\alpha] + 1$, and

$$|g|_{B^\alpha_{\sigma,q}} = ||\omega_r(g,\cdot)||_{\alpha,\sigma} = \begin{cases} \left(\int_0^\infty (t^{-\alpha}\omega_r(g,t)_q)^\sigma \frac{dt}{t}\right)^{1/\sigma}, & \text{for } 0 < \sigma < \infty \\ \sup_{t>0} t^{-\alpha}\omega_r(g,t)_q, & \text{for } \sigma = \infty. \end{cases}$$

Then the Besov norm is defined as

$$||g||_{B^\alpha_{\sigma,q}} = ||g||_q + |g|_{B^\alpha_{\sigma,q}}$$

(see e.g., [14]). For definitions and characterizations of Besov classes in the $d$-dimensional case, see [35]. These classes and similarly defined $F$-classes (which can be handled the same way as Besov classes) include many well-known function spaces such as Hölder–Zygmund spaces, Sobolev spaces, fractional Sobolev spaces or Bessel potential spaces, and inhomogeneous Hardy spaces [35]. By [9], [12], and [34], for $1 \le \sigma \le \infty$, $1 \le q \le \infty$, and $\alpha/d > 1/q - 1/2$, the $L_p$ metric entropy of $B^\alpha_{\sigma,q}(C)$ is of order $\epsilon^{-d/\alpha}$ for $1 \le p \le 2$.

4) *Classes of functions with moduli of continuity of derivatives bounded by fixed functions*: Consider general bounds on the moduli of continuity of derivatives.

Let $\Lambda_{r,\omega}^{d,2} = \Lambda_{r,\omega}^{d,2}(C_0, C_1, \cdots, C_r)$ be the collection of all functions $g$ on $[0,1]^d$ which have all partial derivatives $||D^{\boldsymbol{k}} g||_2 \le C_k$, $|\boldsymbol{k}| = k = 0, 1, \cdots, r$, and the modulus of continuity in $L_2$ norm of each $r$th derivative is bounded by $\omega$. Here $\omega$ is any given modulus of continuity (for definition, see [14, p. 41]. Let $\delta = \delta(\epsilon)$ be defined by equation $\delta^r \omega(\delta) = \epsilon$. Then if $r \ge 1$, from [25], the $L_2$ metric entropy of $\Lambda_{r,\omega}^{d,2}$ is of order $(\delta(\epsilon))^{-d}$.

5) *Classes of functions with different moduli of smoothness with respect to different variables*: Let $k_1, \cdots, k_d$ be positive integers and $0 < \beta_i \le k_i$, $1 \le i \le d$. Let $\boldsymbol{k} = (k_1, \cdots, k_d)$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)$. Let $V(\boldsymbol{k}, \boldsymbol{\beta}, C)$ be the collection of all functions $g$ on $[0,1]^d$ with $||g||_\infty \le C$ and $\sup_{|h| \le t} ||\Delta_{i,h}^{k_i} g||_2 \le C t^{\beta_i}$, where $\Delta_{i,h}^{k_i}$ is the $k_i$th difference with step $h$ in variable $x_i$. As stated in [25, p. 921], from the metric entropy results on full approximation sets together with polynomial approximation results in [33, Sec. V-C], the $L_2$ metric entropy order of $V(\boldsymbol{k}, \boldsymbol{\beta}, C)$ is $(1/\epsilon)^{\sum_{i=1}^d \beta_i^{-1}}$.

6) *Classes $E_d^{\alpha,k}(C)$ and $G_d^\alpha(C)$*: Let $E_d^{\alpha,k}(C) (\alpha > 1/2$ and $k \ge 0)$ be the collection of periodic functions

$$g(x_1, \cdots, x_d) = \sum_{m_1, \cdots, m_d = -\infty}^{+\infty} \left( a_{m_1, \cdots, m_d} \cos\left( \sum_{i=1}^d 2\pi m_i x_i \right) + b_{m_1, \cdots, m_d} \sin\left( \sum_{i=1}^d 2\pi m_i x_i \right) \right)$$

on $[0,1]^d$ with

$$\sqrt{a_{m_1, \cdots, m_d}^2 + b_{m_1, \cdots, m_d}^2} \le C(\overline{m}_1 \cdots \overline{m}_d)^{-\alpha} (\log^k(\overline{m}_1 \cdots \overline{m}_d) + 1)$$

where $\overline{m} = m$ if $m \ne 0$ and $\overline{0} = 1$. Similarly, define $G_d^\alpha(C)$ $(\alpha > 0)$ with the constraint

$$\sum (\overline{m}_1 \cdots \overline{m}_d)^{2\alpha} (a_{m_1, \cdots, m_d}^2 + b_{m_1, \cdots, m_d}^2) \le C^2.$$

The functions in $E_d^{\alpha,k}(C)$ and $G_d^\alpha(C)$ are uniformly bounded if $\alpha > 1$ and $\alpha > 1/2$, respectively. From [31], the $L_2$ metric entropies of $E_d^{\alpha,k}(C)$ and $G_d^\alpha(C)$ are of order

$$(1/\epsilon)^{1/(\alpha - 1/2)} \log^{(2k + 2\alpha(d-1))/(2\alpha - 1)} (1/\epsilon)$$

and

$$(1/\epsilon)^{1/\alpha} \log^{d-1} (1/\epsilon)$$

respectively.

7) *Neural network classes*: Let $N(C)$ be the closure in $L_2[0,1]^d$ of the set of all functions $g: R^d \to R$ of the form

$$g(x) = c_0 + \sum_i c_i \sigma(v_i \cdot x + b_i)$$

with $|c_0| + \sum_i |c_i| \le C$, and $|v_i| = 1$, where $\sigma$ is a fixed sigmoidal function with $\sigma(t) \to 1$ as $t \to \infty$ and $\sigma(t) \to 0$ as $t \to -\infty$. It is further required that $\sigma$ is either the step function $\sigma^*(t) = 1$ for $t \ge 0$, and $\sigma^*(t) = 0$ for $t < 0$, or satisfies the Lipschitz requirements $|\sigma(t) - \sigma(t')| \le C_1 |t - t'|$ for some $C_1$, and $|\sigma(t) - \sigma^*(t)| \le C_2 |t|^{-\gamma}$ for some $C_2$, and $\gamma > 0$, for all $t \ne 0$. From [4] and [28], one knows that the $L_2$ metric entropy of $N(C)$ satisfies

$$(1/\epsilon)^{1/(1/2 + 1/d)} \preceq M(\epsilon) \preceq (1/\epsilon)^{1/(1/2 + 1/(2d))} \log(1/\epsilon).$$

Note that the exponents of $1/\epsilon$ in the upper and lower bounds are different. For both $d = 1$ (in which case, $N(C)$ is a subset of a bounded variation class) and $d = 2$ (see [28]), the exponent in the upper bound cannot be improved. It remains to be seen if this is the case for $d > 2$.

### B. Minimax Rates for the Examples

For classification, the conditional probability $f$ is between 0 and 1. So the classes of interest are the corresponding subsets of the above function classes. For convenience, we will not use another symbol for each of them. Similarly to Lemma 4 in Section VII, it can be shown that the restriction does not change metric entropy orders for the examples.

*1) Assumptions:* We assume that the true density $h(x)$ of the feature variable $X$ belongs to $\mathcal{H}(A)$, which consists of all densities on $[0,1]^d$ that are upper-bounded by $A > 1$. Assumption 1 is clearly satisfied for each of the function classes in the previous subsection. Note that except for the neural network classes (for which the metric entropy order is not exactly identified), each of those classes is rich in $L_2$ distance, and Assumption 2′ is satisfied for $\mathcal{H}(A)$. Thus Theorem 1 (or Corollary 1) is applicable.

*2) Rates for estimating $f$:* The following table summarizes the rates of convergence for estimation of $f$. The rates are for $R(\mathcal{F}; \mathcal{H}(A); n)$.

| Classes | Rate | Condition |
|---|---|---|
| $\mathcal{E}(\{k^\alpha\}, C)$ | $n^{-2\alpha/(2\alpha+1)}$ | $\alpha > 1/2$ |
| $BV(C)$ | $n^{-2/3}$ | |
| $MI(C)$ | $n^{-2/3}$ | |
| $\text{Lip}_{\alpha,q}(C)$ | $n^{-2\alpha/(2\alpha+1)}$ | $\alpha > (1/q - 1/2)^+$ |
| $B_{\sigma,q}^\alpha(C))$ | $n^{-2\alpha/(2\alpha+d)}$ | $\alpha/d > 1/q - 1/2$ |
| $\Lambda_{r,\omega}^{d,2}$ | $\epsilon_n^2$ | $(\delta(\epsilon_n))^{-d} = n\epsilon_n^2$ |
| $V(\boldsymbol{k}, \boldsymbol{\beta}, C)$ | $n^{-2\alpha/(2\alpha+1)}$ | $\alpha^{-1} = \sum_{i=1}^d \beta_i^{-1}$ |
| $E_d^{\alpha,k}(C)$ | $n^{-\alpha-1/2)/\alpha} \cdot (\log n)^{(k+\alpha(d-1))/\alpha}$ | $\alpha > 1$ |
| $G_d^\alpha(C)$ | $n^{-2\alpha/(2\alpha+1)} \cdot (\log n)^{2\alpha(d-1)/(2\alpha+1)}$ | $\alpha > 1/2$ |

Based on Lemma 1, for the neural network class $N(C)$, the minimax rate $R(N(C); \mathcal{H}(A); n)$ is between

$$n^{-(1+2/d)/(2+1/d)} (\log n)^{-(1+1/d)(1+2/d)/(2+1/d)}$$

and

$$(n/\log n)^{-(1+1/d)/(2+1/d)}.$$

The exponential terms in the upper and lower rates do not match exactly because the metric entropy order is not identified. Believing the upper bound on the metric entropy of $N(C)$ given earlier to be tight, one would expect the above upper rate to be the right rate of convergence (possibly ignoring a logarithmic factor). For large $d$ (that is where the neural network models are of main interest), they are very close to each other. Note that $R(N(C);\mathcal{H}(A);n)=o(n^{-1/2})$.

Using Theorem 1, conclusions can also be made for $L_p$ loss $(1\le p<2)$ for some of the classes. For example, for the Besov classes $B^{\alpha}_{\sigma,q}(C)$ with $\alpha/d>(1/q-1/2)^+$, we have

$$\min_{\hat{f}}\max_{f\in B^{\alpha}_{\sigma,q}(C)} E||f-\hat{f}||_{L_p(h)} \asymp n^{-\alpha/(2\alpha+1)}.$$

*3) Satisfaction of Assumption 3:* Based on previous work of others on density estimation, we know Assumption 3 is satisfied with $\underline{M}(\epsilon)$ of the same order as the metric entropy of the class (as required for Theorem 2) for the following classes: $\mathcal{E}(\{k^{\alpha}\},C)$ $(\alpha>1/2)$ with trigonometric basis $\phi_1=1,\phi_{2i}=\sin(2\pi ix)$, $\phi_{2i+1}=\cos(2\pi ix)$ for $i\ge 1$ [7]; $\mathrm{Lip}_{\alpha,\infty}(C)$ [7], also $\mathrm{Lip}_{\alpha,q}(C)(q\ge 1)$, and $BV(C)$ (because they contain $\mathrm{Lip}_{\alpha,\infty}(C)$ and $\mathrm{Lip}_{1,\infty}(C)$ with the same order metric entropy respectively); $B^{\alpha}_{\sigma,q}(C)$ [22]; $\Lambda^{d,2}_{r,\omega}$ with $\omega$ being a concave function [7]; $V(\boldsymbol{k},\boldsymbol{\beta},C)$ [7].

*4) Rates for classification:* Applying Theorem 2, we have the following rates for $r(\mathcal{F};\mathcal{H}(A);n)$.

| Classes | Rate | Condition |
|---|---|---|
| $\mathcal{E}(\{k^{\alpha}\},C)$ | $n^{-\alpha/(2\alpha+1)}$ | $\alpha>1/2$ with trigonometric basis |
| $BV(C)$ | $n^{-1/3}$ | |
| $MI(C)$ | $O(n/\log n)^{-1/2}$ | |
| $\mathrm{Lip}_{\alpha,q}(C)$ | $n^{-\alpha/(2\alpha+1)}$ | $\alpha>(1/q-1/2)^+$ |
| $B^{\alpha}_{\sigma,q}(C))$ | $n^{-\alpha/(2\alpha+d)}$ | $\alpha/d>1/q-1/2$ |
| $\Lambda^{d,2}_{r,\omega}$ | $\epsilon_n$ | $(\delta(\epsilon_n))^{-d}=n\epsilon_n^2$ and $\omega$ concave |
| $V(\boldsymbol{k},\boldsymbol{\beta},C)$ | $n^{-\alpha/(2\alpha+1)}$ | $\alpha^{-1}=\sum_{i=1}^{d}\beta_i^{-1}$ |

The rate $O(n/\log n)^{-1/2}$ for $MI(C)$ is obtained in [16] as mentioned before. It is interesting to observe the difference between estimating the conditional probability $f$ and classification for bounded variation classes and monotone nondecreasing classes. In terms of metric entropy order, $BV(C)$ and $MI(C)$ are equally massive, resulting in the same rate of convergence for estimating $f$. However, metric entropy order is not sufficient to determine the rate of convergence for classification, for which case what matters most is the freedom of $f$ around the value $1/2$ (corresponding to hard classification problems). The difference between $BV(C)$ and $MI(C)$ then shows up drastically.

The class $N(C)$ with, e.g., step sigmoidal, contains a Sobolev class with smoothness parameter $\alpha=d/2+1+\gamma$ for every $\gamma>0$ [3]. Using the lower rate for Sobolev (special case of Besov), together with the upper bound for estimating $f$, we obtain

$$\begin{aligned} n^{-(1+2/d)/(4+4/d)-\gamma'} &\preceq r(N(C);\mathcal{H}(A);n)\\ &\preceq (n/\log n)^{-(1+1/d)/(4+2/d)}\end{aligned}$$

where $\gamma'$ in the lower bound rate can be arbitrarily close to zero. We conjecture that the upper rate (probably without the logarithmic factor) is in fact the optimal rate of convergence, at least for $d=1$ and $d=2$. When $d$ is large, the upper and lower rates of convergence are roughly $n^{-1/4}$.

## VI. Rates of Convergence for Function Classes Modified to Allow Some Irregularity

Included in the previous section are some smoothness classes. In some applications, the target function $f$ is not smooth, but not so only at a few change points of unknown locations. Here we show that such a small modification of a nonparametric class of $f$ does not change rates of convergence. The result is applicable to the following example.

*Example. Functions on $[0,1]$ with a piecewise property*: Let $\mathcal{S}$ be an original nonparametric class of nonnegative functions upper-bounded by 1. Let

$$\begin{aligned}\mathcal{F}=\Big\{f(x)=s(x)\cdot\sum_{i=0}^{k-1}b_{i+1}1_{\{a_i\le x<a_{i+1}\}}&: s\in\mathcal{S},\\ 0=a_0<a_1<a_2<\cdots<a_k&=1,\\ 0\le b_i\le 1, 1\le i\le k&\Big\}.\end{aligned}$$

Here $k$ is a positive integer. If the functions in $\mathcal{S}$ are continuous (or differentiable), then the functions in $\mathcal{F}$ are piecewise continuous (or differentiable).

Let $\mathcal{S}$ and $\mathcal{G}$ be two classes of nonnegative functions on $[0,1]^d$ upper-bounded by 1. Consider a new class of functions

$$\mathcal{F}=\{f(x)=s(x)g(x)\colon s\in\mathcal{S},g\in\mathcal{G}\}.$$

Then it can be easily shown that if $\mathcal{G}$ contains the constant function 1, the $L_2$ metric entropies of these classes have the following relationship:

$$M_2(\epsilon;\mathcal{S})\le M_2(\epsilon;\mathcal{F})\le M_2(\epsilon/\sqrt{2};\mathcal{S})+M_2(\epsilon/\sqrt{2};\mathcal{G}).$$

As a consequence, if $\mathcal{G}$ has the same or a smaller order metric entropy compared to a rich nonparametric class $\mathcal{S}$, then $M_2(\epsilon;\mathcal{F})\asymp M_2(\epsilon;\mathcal{S})$, which is the case for the above example.

*Corollary 4:* Suppose $\mathcal{G}$ contains the constant function 1, and $M_2(\epsilon;\mathcal{G})$ is of the same or a smaller order of $M_2(\epsilon;\mathcal{S})$. Then if $\mathcal{S}$ satisfies the conditions in Theorem 1, we have

$$R(\mathcal{F};\mathcal{H}(A);n)\asymp R(\mathcal{S};\mathcal{H}(A);n)\asymp\epsilon_n^2$$

where $\epsilon_n$ is determined by $M_2(\epsilon_n;\mathcal{S})=n\epsilon_n^2$. If $\mathcal{S}$ satisfies the corresponding conditions in Theorem 2

$$r(\mathcal{F};\mathcal{H}(A);n)\asymp\epsilon_n.$$

## VII. Proofs of the Results

The idea for the proof of Lemma 1 is similar to that used in Yang and Barron [36] on density estimation and nonparametric regression.

*Proof of Lemma 1:* For the lower bound result, we consider a subset $\mathcal{F}^0$ of $\mathcal{F}$. A minimax lower bound for $\mathcal{F}^0$ is clearly a lower bound for $\mathcal{F}$. Let $\mathcal{F}^0 = \{f/2 + f^*/2 : f \in \mathcal{F}\}$. Under the convexity assumption on $\mathcal{F}$, $\mathcal{F}^0$ is indeed a subset of $\mathcal{F}$. Note that by Assumption 1, the functions in $\mathcal{F}^0$ are uniformly bounded above away from 1 and below away from zero

$$0 < \underline{c}/2 \le f \le (\overline{c}+1)/2 < 1 \tag{8}$$

for every $f \in \mathcal{F}^0$. The boundedness property will be used in our analysis to relate Kullback–Leibler (K-L) divergence to $L_2$ distance. It is easy to verify that the $L_q(h)$ packing metric entropy of $\mathcal{F}^0$ is $M_q(2\epsilon; \mathcal{F})$ [36].

Let $N_{\underline{\epsilon}_{n,q}}$ be an $\underline{\epsilon}_{n,q}$-packing set with maximum cardinality in $\mathcal{F}^0$ under distance $d_q = L_q(h)$, and let $G_{\epsilon_n}$ be an $\epsilon_n$-net for $\mathcal{F}^0$ under distance $d_2$ (i.e., for any $f_1 \in \mathcal{F}^0$, there exists $f_2 \in G_{\epsilon_n}$ such that $d_2(f_1, f_2) \le \epsilon_n$). Since an $\epsilon$-packing set in $\mathcal{F}^0$ with maximum cardinality serves also as an $\epsilon$-net in $\mathcal{F}^0$, we can choose $G_{\epsilon_n}$ to have log-cardinality

$$M_2(\epsilon_n; \mathcal{F}^0) = M_2(2\epsilon_n; \mathcal{F}).$$

For any estimator $\hat{f}$ based on $Z^n$, define

$$\tilde{f} = \arg \min_{f \in N_{\underline{\epsilon}_{n,q}}} d_q(f, \hat{f})$$

(if there is more than one minimizer, choose one based on any tie-breaking rule) so that $\tilde{f}$ takes values in the packing set $N_{\underline{\epsilon}_{n,q}}$. Let $f$ be any point in $N_{\underline{\epsilon}_{n,q}}$. By the triangle inequality,

$$d_q(f, \hat{f}) + d_q(\tilde{f}, \hat{f}) \ge d_q(f, \tilde{f})$$

which is at least $\underline{\epsilon}_{n,q}$ if $f \ne \tilde{f}$ since $N_{\underline{\epsilon}_{n,q}}$ is a packing set. Thus if $f \ne \tilde{f}$, we must have $d_q(f, \hat{f}) \ge \underline{\epsilon}_{n,q}/2$, and

$$\begin{aligned}
&\min_{\hat{f}} \max_{f \in \mathcal{F}} P_f\{d_q(f, \hat{f}) \ge \underline{\epsilon}_{n,q}/2\} \\
&\quad \ge \min_{\hat{f}} \max_{f \in N_{\underline{\epsilon}_{n,q}}} P_f\{d_q(f, \hat{f}) \ge \underline{\epsilon}_{n,q}/2\} \\
&\quad \ge \min_{\hat{f}} \max_{f \in N_{\underline{\epsilon}_{n,q}}} P_f(f \ne \tilde{f}) \\
&\quad \ge \min_{\hat{f}} \sum_{f \in N_{\underline{\epsilon}_{n,q}}} P_f(f \ne \tilde{f}) / |N_{\underline{\epsilon}_{n,q}}| \\
&\quad = \min_{\hat{f}} P^w(\Theta \ne \tilde{f})
\end{aligned}$$

where $P_f$ denotes probability under $f$, and in the last line, $\Theta$ is randomly drawn according to a discrete uniform prior $w$ on $N_{\underline{\epsilon}_{n,q}}$, and $P^w$ denotes the Bayes average probability with respect to the prior $w$. Since $(d_q(f, \hat{f}))^\ell$ is not less than $(\underline{\epsilon}_{n,q}/2)^\ell 1_{\{f \ne \tilde{f}\}}$, taking the expected value it follows that for all $\ell > 0$

$$\begin{aligned}
&\min_{\hat{f}} \max_{f \in \mathcal{F}} E_f d_q^\ell(f, \hat{f}) \\
&\quad \ge (\underline{\epsilon}_{n,q}/2)^\ell \min_{\hat{f}} P^w(\Theta \ne \tilde{f}) \\
&\quad \ge (\underline{\epsilon}_{n,q}/2)^\ell \left(1 - \frac{I(\Theta; Z^n) + \log 2}{\log |N_{\underline{\epsilon}_{n,q}}|}\right)
\end{aligned} \tag{9}$$

by Fano's inequality (see, e.g., [13, pp. 39 and 205]), where $I(\Theta; Z^n)$ is Shannon's mutual information between the random parameter $\Theta$ and the sample $Z^n$.

This mutual information is equal to the average (with respect to the prior) of the K-L divergence between $p_f(z^n)$ and

$$p^w(z^n) = \sum_{f \in N_{\underline{\epsilon}_{n,q}}} p_f(z^n) / |N_{\underline{\epsilon}_{n,q}}|.$$

Here the density of $Z^n$ is

$$p_f(z^n) = \left(\prod_{i=1}^n h(x_i)\right) \prod_{i=1}^n (f(x_i)^{y_i}(1 - f(x_i))^{1-y_i})$$

with respect to the product measure $(\mu \otimes \nu)^n$, where $\nu$ is the counting measure (for $Y$). Since the Bayes mixture density $p^w(z^n)$ minimizes the average K-L divergence over all choices of density $q(z^n)$, the mutual information is upper-bounded by the maximum K-L divergence between $p_f(z^n)$ and any joint density $q(z^n)$, i.e.,

$$I(\Theta; Z^n) \le \max_{f \in N_{\underline{\epsilon}_{n,q}}} \int p_f(z^n) \log \frac{p_f(z^n)}{q(z^n)}$$

where the integral is with respect to the product measure $(\mu \otimes \nu)^n$. Now choose $w_1$ to be the uniform prior on $G_{\epsilon_n}$ and let

$$q(z^n) = p^{w_1}(z^n) = \sum_{f \in G_{\epsilon_n}} p_f(z^n) / |G_{\epsilon_n}|.$$

For any $f \in \mathcal{F}^0$, there is $f' \in G_{\epsilon_n}$ such that $||f - f'||_{L_2(h)} \le \epsilon_n$, and then

$$\begin{aligned}
\int p_f(z^n) \log \frac{p_f(z^n)}{q(z^n)} &\le \int p_f(z^n) \log \frac{p_f(z^n)}{(1/|G_{\epsilon_n}|) p_{f'}(z^n)} \\
&= \log |G_{\epsilon_n}| + \int p_f(z^n) \log \frac{p_f(z^n)}{p_{f'}(z^n)}.
\end{aligned}$$

We now bound the K-L divergence

$$\begin{aligned}
&\int p_f(z^n) \log \frac{p_f(z^n)}{p_{f'}(z^n)} \\
&\quad = n \int h(x) \left( f(x) \log \frac{f(x)}{f'(x)} + (1 - f(x)) \right. \\
&\qquad \left. \cdot \log \frac{1 - f(x)}{1 - f'(x)} \right) \mu(dx) \\
&\quad \le n \int \left( \frac{(f(x) - f'(x))^2}{f'(x)} + \frac{(f(x) - f'(x))^2}{1 - f'(x)} \right) h(x) \mu(dx) \\
&\quad \le \frac{4n}{\underline{c}(1 - \overline{c})} ||f - f'||_{L_2(h)}^2
\end{aligned} \tag{10}$$

where the first inequality follows from the familiar bound on K-L divergence by chi-square distance, i.e.,

$$\int p \log(p/q) \le \int (p - q)^2 / q$$

for densities $p$ and $q$, and the second inequality follows from (8). Together with our choice of $\epsilon_n$ in (3), we have

$$\begin{aligned}
\int p_f(z^n) \log \frac{p_f(z^n)}{q(z^n)} &\le M_2(2\epsilon_n; \mathcal{F}) + \frac{4n}{\underline{c}(1 - \overline{c})} \epsilon_n^2 \\
&= \left(1 + \frac{4}{\underline{c}(1 - \overline{c})}\right) n \epsilon_n^2.
\end{aligned} \tag{11}$$

Thus we have shown that $I(\Theta; Z^n) \le (1 + (4/\underline{c}(1 - \overline{c}))) n \epsilon_n^2$.

By our choice of $\underline{\epsilon}_{n,q}$ in (4)

$$(I(\Theta; Z^n) + \log 2)/\log |N_{\underline{\epsilon}_{n,q}}| \le \frac{1}{2}.$$

The lower bound follows.

Now let us derive the upper bound. We will manipulate the data in some way for the purpose of relating K-L risk of density estimation to the risk of interest.

In addition to the observed independent and identically distributed (i.i.d.) sample $(X_1, Y_1), \cdots, (X_n, Y_n)$, we generate some random variables. At the observed $X_i = x_i$ value, let $W_i$ be an independently generated Bernoulli random variable with success probability $f_*(x) \equiv 1/2$. Let $\tilde{Y}_i$ be $Y_i$ or $W_i$ with probability $(1/2, 1/2)$ according to the outcome of Bernoulli$(1/2)$ random variables $V_i$ generated independently. Then the conditional probability of $\tilde{Y}_i$ taking value 1 is $g(x) = (f(x) + f_*(x))/2$. The new conditional probability $g$ is bounded between $1/4$ and $3/4$ as in (8), whereas the family of the original densities need not be. Let

$$\mathcal{F}_0 = \{g\colon g = (f + f_*)/2, f \in \mathcal{F}\}$$

be the new class of conditional probability for $\tilde{Z}^n = (X_i, \tilde{Y}_i)_{i=1}^n$. The next lemma relates the risk of estimating $f \in \mathcal{F}$ to that of estimating $g \in \mathcal{F}_0$. See [36] for an analogous result in density estimation.

*Lemma 2:* The minimax $L_2(h)$ risks of the two classes $\mathcal{F}$ and $\mathcal{F}^0$ have the following relationship:

$$R(\mathcal{F}; n) \le 4R(\mathcal{F}_0; n).$$

*Proof of Lemma 2:* Let $\tilde{g}(x; \overline{Z}^n)$ be any estimator of $g \in \mathcal{F}_0$ based on an i.i.d. sample $\overline{Z}^n$ with density $h(x)g(x)^y(1-g(x))^{1-y}$. Let $\hat{g}(x) = \hat{g}(x; \overline{Z}^n)$ be the function that minimizes $||s - \tilde{g}||^2_{L_2(h)}$ over functions in the set

$$S = \{s\colon 1/4 \le s \le 3/4\}.$$

Since $\hat{g}$ is the projection of $\tilde{g}$ into $S$ which contains $g$, it is not hard to show (see [36, Lemma 9]) that

$$||g - \hat{g}||^2_{L_2(h)} \le ||g - \tilde{g}||^2_{L_2(h)}.$$

Now we construct an estimator for $f$. Since $\overline{Z}^n$ and the constructed sample $\tilde{Z}^n$ have the same distribution, replacing $\overline{Z}^n$ by $\tilde{Z}^n$ in $\hat{g}$, we get an estimator of $g(x) = (f + f_*)/2$ based on $\tilde{Z}^n$. From $f(x) = 2g(x) - f_*$, let

$$\hat{f}_{\text{rand}}(x) = 2\hat{g}(x; \tilde{Z}^n) - f_*.$$

Then $\hat{f}_{\text{rand}}(x)$ is a valid estimator of $f$ and depends on $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ and the outcomes of the coin flips $V_1, \cdots, V_n$ as well as $W_1, \cdots, W_n$. So it is a randomized estimator. The squared $L_2(h)$ loss of $\hat{f}_{\text{rand}}$ is bounded as follows:

$$\begin{aligned}\int h(x)(f(x) - \hat{f}_{\text{rand}}(x))^2\, d\mu &= \int (2g(x) - 2\hat{g}(x))^2 h(x)\, d\mu \\ &= 4\int (g - \hat{g})^2 h(x)\, d\mu \\ &\le 4||g - \tilde{g}||^2_{L_2(h)}.\end{aligned}$$

To avoid randomization, we may replace $\hat{f}_{\text{rand}}(x)$ with its expected value over $W_1, \cdots, W_n$ and $V_1, \cdots, V_n$ to get $\hat{f}(x)$ with

$$\begin{aligned}E_{Z^n}||f - \hat{f}||^2_{L_2(h)} &= E_{Z^n}||f - E_{W^n, V^n}\hat{f}_{\text{rand}}||^2_{L_2(h)} \\ &\le E_{Z^n}E_{W^n, V^n}||f - \hat{f}_{\text{rand}}||^2_{L_2(h)} \\ &= E_{\tilde{Z}^n}||f - \hat{f}_{\text{rand}}||^2_{L_2(h)} \\ &= E_{\overline{Z}^n}||f - \hat{f}_{\text{rand}}||^2_{L_2(h)} \\ &\le 4E_{\overline{Z}^n}||g - \tilde{g}||^2_{L_2(h)}\end{aligned}$$

where the first inequality is by convexity and the second and third identities are because $\hat{f}_{\text{rand}}$ depends on $X^n, Y^n, W^n, V^n$ only through $\tilde{Z}^n$, which has the same distribution as $\overline{Z}^n$. Thus

$$\max_{f \in \mathcal{F}} E_{Z^n}||f - \hat{f}||^2_{L_2(h)} \le 4 \max_{g \in \mathcal{F}_0} E_{\overline{Z}^n}||g - \tilde{g}||^2_{L_2(h)}.$$

Taking the minimum over all estimators $\tilde{g}$ completes the proof of Lemma 2.

Thus the minimax risk of the original problem is upper-bounded by a multiple of the minimax risk on $\mathcal{F}_0$. Moreover, the $\epsilon$-entropies are related. Indeed, since

$$||(f_1 + f_*)/2 - (f_2 + f_*)/2||_{L_2(h)} = (1/2)||f_1 - f_2||_{L_2(h)}$$

for the new class $\mathcal{F}_0$, the $\epsilon$-packing entropy under $L_2(h)$ distance is $M_2(\epsilon; \mathcal{F}_0) = M_2(2\epsilon; \mathcal{F})$.

Now we derive an upper bound on the minimax risk for the new class $\mathcal{F}_0$.

Consider an $\epsilon_n$-net $\tilde{G}_{\epsilon_n}$ with log-cardinality $M_2(2\epsilon_n; \mathcal{F})$ for $\mathcal{F}_0$ (just as $G_{\epsilon_n}$ for $\mathcal{F}^0$ used in the proof of the minimax lower bound in Lemma 1) and the uniform prior $w_2$ on $\tilde{G}_{\epsilon_n}$. Since we assume that $h$ is known, the $\epsilon$-net can be constructed, theoretically speaking. Let the Bayes predictive density estimators of $p_f(z) = h(x)f(x)^y(1 - f(x))^{1-y}$ be $\hat{p}_i(z) = p(Z_{i+1}|Z^i)$ evaluated at $Z_{i+1} = z$, which equal the expression at the bottom of this page for $i > 0$ and

$$\hat{p}_i(z) = p^{w_2}(z) = (1/|\tilde{G}_{\epsilon_n}|) \sum_{f \in \tilde{G}_{\epsilon_n}} p_f(z)$$

$$p^{w_2}(Z^i, z)/p^{w_2}(Z^i) = \frac{\displaystyle\sum_{f \in \tilde{G}_{\epsilon_n}} \left( \prod_{j=1}^{i} h(X_i)f(X_i)^{Y_i}(1 - f(X_i))^{1-Y_i} \right) h(x)f(x)^y(1 - f(x))^{1-y}}{\displaystyle\sum_{f \in \tilde{G}_{\epsilon_n}} \prod_{j=1}^{i} h(X_i)f(X_i)^{Y_i}(1 - f(X_i))^{1-Y_i}}$$

for $i = 0$. Let

$$\beta_f(Z^i) = \frac{\prod_{j=1}^{i} h(X_i) f(X_i)^{Y_i} (1 - f(X_i))^{1-Y_i}}{\sum_{g \in \tilde{G}_{\epsilon_n}} \prod_{j=1}^{i} h(X_i) g(X_i)^{Y_i} (1 - g(X_i))^{1-Y_i}} = \frac{\prod_{j=1}^{i} f(X_i)^{Y_i} (1 - f(X_i))^{1-Y_i}}{\sum_{g \in \tilde{G}_{\epsilon_n}} \Pi_{j=1}^{i} g(X_i)^{Y_i} (1 - g(X_i))^{1-Y_i}}.$$

Then

$$\begin{aligned} p^{w_2}(Z^i, z)/p^{w_2}(Z^i) &= h(x) \cdot \sum_{f \in \tilde{G}_{\epsilon_n}} \beta_f(Z^i) f(x)^y (1 - f(x))^{1-y} \\ &= h(x) \cdot \left( \sum_{f \in \tilde{G}_{\epsilon_n}} \beta_f(Z^i) f(x) \right)^y \\ &\quad \cdot \left( 1 - \sum_{f \in \tilde{G}_{\epsilon_n}} \beta_f(Z^i) f(x) \right)^{1-y}. \end{aligned}$$

Let

$$\hat{f}_i(x) = \sum_{f \in \tilde{G}_{\epsilon_n}} \beta_f(Z^i) f(x).$$

It is an estimator of the conditional probability of $Y$ taking 1 given $X = x$ based on $Z^i$. Note that $\hat{f}_i(x)$ is between $1/4$ and $3/4$ and does not depend on $h$ except through the $\epsilon$-net $\tilde{G}_{\epsilon_n}$. As in Barron [1] and [36], by the chain rule of relative entropy, for any $f \in \mathcal{F}_0$

$$\begin{aligned} \sum_{i=0}^{n-1} E \log \frac{p_f(Z_{i+1})}{\hat{p}_i(Z_{i+1})} &= E \log \frac{p_f(Z^n)}{p^{w_2}(Z^n)} \\ &\le M_2(2\epsilon_n; \mathcal{F}) + 16 n \epsilon_n^2 \\ &= 17 n \epsilon_n^2, \end{aligned}$$

where the inequality is as in (11). Since the squared Hellinger distance satisfies

$$d_H^2(p_1, p_2) = \int (p_1^{1/2} - p_2^{1/2})^2 \le \int p_1 \log(p_1/p_2)$$

for two densities $p_1$ and $p_2$, and observing that

$$E \log(p_f(Z_{i+1})/\hat{p}_i(Z_{i+1})) = E \int p_f(z) \log(p_f(z)/\hat{p}_i(z)) (\mu \otimes \nu)(dz)$$

from above, we have

$$\max_{f \in \mathcal{F}_0} \sum_{i=0}^{n-1} E d_H^2(p_f, p_{\hat{f}_i}) \le 17 n \epsilon_n^2.$$

Note that

$$d_H^2(p_f, p_g) = \int h(x) d_Y^2(f(x), g(x)) \, \mu(dx)$$

where $d_Y^2(f(x), g(x))$ is the Hellinger distance between $f(x)^y (1 - f(x))^{1-y}$ and $g(x)^y (1 - g(x))^{1-y}$ with respect to $y$ at a given $x$, i.e.,

$$d_Y^2(f(x), g(x)) = (\sqrt{f(x)} - \sqrt{g(x)})^2 + (\sqrt{1 - f(x)} - \sqrt{1 - g(x)})^2.$$

For $f$ and $g$ between $1/4$ and $3/4$, this distance is bounded below by

$$\begin{aligned} &(\sqrt{f(x)} - \sqrt{g(x)})^2 + (\sqrt{1 - f(x)} - \sqrt{1 - g(x)})^2 \\ &\quad \ge \tfrac{1}{3} (\sqrt{f(x)} + \sqrt{g(x)})^2 (\sqrt{f(x)} - \sqrt{g(x)})^2 \\ &\qquad + \frac{1}{3} (\sqrt{1 - f(x)} + \sqrt{1 - g(x)})^2 \\ &\qquad \cdot (\sqrt{1 - f(x)} - \sqrt{1 - g(x)})^2 \\ &\quad \ge \frac{1}{3} (f(x) - g(x))^2 \\ &\qquad + \frac{1}{3} ((1 - f(x)) - (1 - g(x)))^2 \\ &\quad = \frac{2}{3} (f(x) - g(x))^2. \end{aligned}$$

As a consequence, we have

$$\begin{aligned} &\int h(x) d_Y^2(f(x), \hat{f}_i(x)) \mu(dx) \\ &\qquad \ge (2/3) \int h(x) (f(x) - \hat{f}_i(x))^2 \mu(dx) \\ &\qquad = (2/3) \|f - \hat{f}_i\|_{L_2(h)}^2, \end{aligned}$$

and

$$\max_{f \in \mathcal{F}_0} \sum_{i=0}^{n-1} E \|f - \hat{f}_i\|_{L_2(h)}^2 \le (51/2) n \epsilon_n^2.$$

Let

$$\hat{\hat{f}}(x) = \frac{1}{n} \sum_{i=0}^{n-1} \hat{f}_i(x)$$

be the final estimator of $f$. Then by convexity, we have

$$\begin{aligned} \max_{f \in \mathcal{F}_0} E \|f - \hat{\hat{f}}\|_{L_2(h)}^2 &\le \max_{f \in \mathcal{F}_0} \frac{1}{n} \sum_{i=0}^{n-1} E \|f - \hat{f}_i\|_{L_2(h)}^2 \\ &\le (51/2) \epsilon_n^2. \end{aligned}$$

The conclusion on upper bound in Lemma 1 then follows based on Lemma 2. This completes the proof of Lemma 1.

*Proof of Theorem 1:* In the derivation of the upper bound in Lemma 1, we needed to know the density $h(x)$ of $X$ only in the construction of the $\epsilon$-net $\tilde{G}_{\epsilon_n}$. When $h$ is not known, we cannot construct $\epsilon$-nets for $\mathcal{F}$ under the $L_2(h)$ distance. Assume that Assumption 2 is satisfied with log-cardinality of the $\epsilon$-net bounded by $\overline{M}_2(\epsilon)$. Following the same argument for the upper bound in the proof of Lemma 1, we have

$$R(\mathcal{F}; n) \le 102 \tilde{\epsilon}_n^2$$

where $\tilde{\epsilon}_n$ is determined by

$$\overline{M}_2(2\tilde{\epsilon}_n) = n \tilde{\epsilon}_n^2.$$

Under the richness assumption of $\mathcal{F}$, it can be easily shown (see [36]) that $\tilde{\epsilon}_n$ and $\underline{\epsilon}_n$ given in (4) are of the same order as $\epsilon'_n$ determined by $M(\epsilon'_n;\mathcal{F}) = n(\epsilon'_n)^2$ if $\overline{M}_2(\epsilon)$ and $M(\epsilon)$ are of the same order.

For the second claim, under the assumption that $M_q(\epsilon)$ and $M_2(\epsilon)$ are of the same order and the richness assumption, $\underline{\epsilon}_{n,q}$ as determined in (4) is of the same order as $\epsilon'_n$. Thus by Lemma 1

$$\min_{\hat{f}} \max_{f\in\mathcal{F}} E_f||f-\hat{f}||_{L_q(h)} \succeq \epsilon'_n.$$

The upper bound rate follows from the fact that for $1 \le q \le 2$

$$E_f||f-\hat{f}||_{L_q(h)} \le E_f||f-\hat{f}||_{L_2(h)} \le \sqrt{E_f||f-\hat{f}||^2_{L_2(h)}}.$$

The third claim follows similarly as above together with that the upper bound $102\tilde{\epsilon}_n^2$ on $R(\mathcal{F};n)$ is uniform over $h \in \mathcal{H}$ and that the estimator does not depend on $h$. This completes the proof of Theorem 1.

*Proof of Corollary 1:* Since $h/h^* \le C$, we have

$$\begin{aligned}||g||_{L_2(h)} &= \left(\int g^2 h\mu(dx)\right)^{1/2} \le \sqrt{C}\left(\int g^2 h^*\mu(dx)\right)^{1/2}\\ &= \sqrt{C}||g||_{L_2(h^*)}.\end{aligned}$$

Thus an $\epsilon$-net for $\mathcal{F}$ under the $L_2(h^*)$ distance is a $\sqrt{C}\epsilon$-net for $\mathcal{F}$ under the $L_2(h)$ distance. Together with the other conditions, Assumption $2'$ is satisfied. The conclusion then follows from Theorem 1.

Before we prove Theorem 2, we give a lemma for lower-bounding MEPR.

*Lemma 3:* Assume Assumption 3 is satisfied and that $h$ is upper-bounded by $\overline{c}$ on $\mathcal{X}$ and lower-bounded by $\underline{c}>0$ on $\cup_{1\le i\le r}\{I+x_i\}$. Let $\underline{\epsilon}_n$ be determined by

$$\frac{512\overline{c}n\underline{\epsilon}_n^2 + 8\log 2}{(A_4/A_3)\underline{M}(\underline{\epsilon}_n)} = \frac{1}{2}. \tag{12}$$

Then when $n$ is large enough

$$r(\mathcal{F}_{\text{cube}};n) \ge \frac{A_1A_2A_4\underline{c}\underline{\epsilon}_n}{8}.$$

*Proof of Lemma 3:* The idea of the proof is that under Assumption 3, the loss of interest behaves like $L_q$ distance on the subset $\mathcal{F}_{\text{cube}}$. Then we can again use Fano's inequality to derive a lower bound.

For two functions $f$ and $s$ bounded between 0 and 1, let

$$\begin{aligned}d(f,s) = &\int_{\{s<1/2\le f\}} h(x)(2f(x)-1)\mu(dx)\\ &+\int_{\{f<1/2\le s\}} h(x)(1-2f(x))\mu(dx).\end{aligned}$$

Note that a decision of the class membership at each $x$ corresponds to a function of $x$ of two values. For such a $g$, $d(f,g)$ is the loss $l(f;g)$ incurred by using $g$ to classify $Y$ when the true conditional probability is $f$. To see that, let $g^*(x)=1$ if $f(x)\ge 1/2$ and $g^*(x)=0$ if $f(x)<1/2$ be a Bayes decision, then

$$\begin{aligned}l(f;g) &= P(Y\ne g(X)) - P(Y\ne g^*(X))\\ &= \int h(x)(P(Y\ne g(x)|X=x)\\ &\quad - P(Y\ne g^*(x)|X=x))\mu(dx).\end{aligned}$$

Note

$$\begin{aligned}&P(Y\ne g(x)|X=x) - P(Y\ne g^*(x)|X=x)\\ &= \begin{cases} P(Y=1|X=x) - P(Y=0|X=x),\\ \quad \text{when } g(x)=0 \text{ and } g^*(x)=1\\ P(Y=0|X=x) - P(Y=1|X=x),\\ \quad \text{when } g(x)=1 \text{ and } g^*(x)=0\\ 0, \quad \text{otherwise}\end{cases}\\ &= \begin{cases} 2f(x)-1, & \text{when } g(x)<1/2\le f(x)\\ 1-2f(x), & \text{when } f(x)<1/2\le g(x)\\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

Thus $l(f;g) = d(f;g)$ as claimed.

For any

$$\begin{aligned}f_\tau &= 1/2 + \sum_{i=1}^r \tau_i g_\epsilon(x-x_i) =: f_1\\ f_{\tau'} &= 1/2 + \sum_{i=1}^r \tau'_i g_\epsilon(x-x_i) =: f_2\end{aligned}$$

let

$$\begin{aligned}B_i &= \{f_i \ge 1/2 + A_1\epsilon\}\\ D_i &= \{f_i \le 1/2 - A_1\epsilon\}\\ E_i &= \{s<1/2\le f_i\}\\ G_i &= \{f_i<1/2\le s\}\end{aligned}$$

for $i=1,2$. Then for any $s$, we have

$$\begin{aligned}d(f_i,s) &\ge \int_{E_i\cap B_i} h(x)(2f_i(x)-1)\mu(dx)\\ &\quad + \int_{G_i\cap D_i} h(x)(1-2f_i(x))\mu(dx)\\ &\ge \underline{c}\int_{E_i\cap B_i}(2f_i(x)-1)\mu(dx)\\ &\quad + \underline{c}\int_{G_i\cap D_i}(1-2f_i(x))\mu(dx)\\ &\ge 2A_1\underline{c}\epsilon\cdot\mu(E_i\cap B_i) + 2A_1\underline{c}\epsilon\cdot\mu(G_i\cap D_i).\end{aligned}$$

As a consequence, by regrouping the events involved above, we have

$$\begin{aligned}&d(f_1,s)+d(f_2,s)\\ &\quad\ge 2A_1\underline{c}\epsilon\cdot\mu((E_1\cap B_1)\cup(G_2\cap D_2))\\ &\qquad + 2A_1\underline{c}\epsilon\cdot\mu((G_1\cap D_1)\cup(E_2\cap B_2)).\end{aligned}$$

Consider such $j$ that $\tau_j\ne\tau'_j$ for a moment. We have

$$\{I+x_j\}\cap B_i = \{I+x_j\}\cap D_{i'}, \qquad \text{for } i\ne i'.$$

It is also true that for $i\ne i'$, in $\{I+x_j\}\cap B_i$, either $s<1/2\le f_i$ or $f_{i'}<1/2\le s$. Thus

$$\{I+x_j\}\cap((E_i\cap B_i)\cup(G_{i'}\cap D_{i'})) = \{I+x_j\}\cap B_i.$$

Furthermore,

$$\mu(\{I+x_j\}\cap B_1)+\mu(\{I+x_j\}\cap B_2) = \mu(|g_\epsilon|\ge A_1\epsilon))\ge A_2 v.$$

Now let $N$ be the number of different signs $f_1$ and $f_2$ have in the hypercube representation, i.e., $N=\#\{j\colon \tau_j\neq\tau_j'\}$. Then from Assumption 3 and above

$$d(f_1,s)+d(f_2,s)\ge 2A_1\underline{c}\epsilon N\cdot\mu\{|g_\epsilon|\ge A_1\epsilon\}\ge 2A_1A_2\underline{c}Nv\epsilon. \tag{13}$$

In particular, letting $s=f_2$, we have

$$d(f_1,f_2)\ge 2A_1A_2\underline{c}Nv\epsilon. \tag{14}$$

Now let us upper-bound the K-L divergence between $p_{f_1}$ and $p_{f_2}$ for $f_1,\ f_2\in\mathcal{F}_{\rm cube}$ as in (10)

$$\begin{aligned}\int p_{f_1}\log\frac{p_{f_1}}{p_{f_2}} &\le \overline{c}\int\frac{(f_1(x)-f_2(x))^2}{f_2(x)\cdot(1-f_2(x))}\mu(dx)\\ &\le \frac{\overline{c}}{(1/2-\epsilon)^2}\cdot 4N\int_I g_\epsilon^2\mu(dx)\\ &\le 64\overline{c}Nv\epsilon^2,\qquad \text{for } \epsilon\le 1/4\end{aligned} \tag{15}$$

where for the second inequality, we use the assumption that $|g_\epsilon|\le\epsilon$ and the fact that on $I+x_j$ with $\tau_j\neq\tau_j'$, $|f_1(x)-f_2(x)|=2|g_\epsilon(x-x_j)|$. Now consider a fraction of functions in $\mathcal{F}_{\rm cube}$, on which Fano's inequality will be used. From [20, p. 256], there exists a subset $\Gamma$ of $\{-1,1\}^r$ with at least $\lceil e^{r/8}\rceil$ elements so that any two distinct members have at least $\lceil r/4\rceil$ different coordinates. Let $\mathcal{F}_{\rm sub}=\{f_\tau\colon \tau\in\Gamma\}$. Then from (14), for any $f_1,\ f_1\in\mathcal{F}_{\rm sub}$, we have

$$d(f_1,f_2)\ge A_1A_2\underline{c}rv\epsilon/2. \tag{16}$$

Let $\Theta$ take values in $\Gamma$ with a uniform distribution. The Shannon mutual information between the random parameter $\Theta$ and the observations $Z^n$ satisfies

$$I(\Theta;Z^n)\le nI(\Theta;Z^1)\le n\max_{\tau\in\Gamma}D(p_{f_\tau}||p_{f_0})\le 64\overline{c}nrv\epsilon^2$$

for any $f_0\in\mathcal{F}_{\rm sub}$.

Now for any given classifier $\delta$, define $\hat{s}_\delta(x)=0$ if $\delta(x;Z^n)=0$ and $\hat{s}_\delta(x)=1$ if $\delta(x;Z^n)=1$. Let $\hat{\tau}$ be the minimizer of $d(f_\tau,\hat{s}_\delta)$ over $\tau\in\Gamma$ (using any tie-breaking rule if necessary). Then if $\tau\neq\hat{\tau}$, from (13)

$$d(f_\tau,\hat{s}_\delta)+d(f_{\hat{\tau}},\hat{s}_\delta)\ge\frac{A_1A_2\underline{c}rv\epsilon}{2}.$$

Thus

$$d(f_\tau,\hat{s}_\delta)\ge\frac{A_1A_2\underline{c}rv\epsilon}{4}.$$

Proceeding as in (9), we have

$$\begin{aligned}\min_\delta\max_{\tau\in\Gamma}Ed(f_\tau,\hat{s}_\delta) &\ge \frac{A_1A_2\underline{c}rv\epsilon}{4}\left(1-\frac{I(\Theta,Z^n)+\log 2}{\log(|\Gamma|)}\right)\\ &\ge \frac{A_1A_2\underline{c}rv\epsilon}{4}\left(1-\frac{64n\overline{c}rv\epsilon^2+\log 2}{r/8}\right)\\ &\ge \frac{A_1A_2A_4\underline{c}\epsilon}{4}\left(1-\frac{512n\overline{c}\epsilon^2+8\log 2}{(A_4/A_3)\underline{M}(\epsilon)}\right)\end{aligned}$$

where for the last two inequalities, we assume $\epsilon\le 1/4$ as used in (15), and use the assumptions on the relationship between the constants involved in Assumption 3. Choosing $\underline{\epsilon}_n$ as in (12), when $n$ is large enough, $\underline{\epsilon}_n\le 1/4$, we have

$$r(\mathcal{F}_{\rm cube};n)\ge\min_\delta\max_{\tau\in\Gamma}Ed(f_\tau,\hat{s}_\delta)\ge\frac{A_1A_2A_4\underline{c}\underline{\epsilon}_n}{8}.$$

*Proof of Theorem 2:* Under Assumption 2, from Lemma 1,

$$r(\mathcal{F};n)\le\sqrt{R(\mathcal{F};n)}\preceq\epsilon_n.$$

By Lemma 3,

$$r(\mathcal{F};n)\ge r(\mathcal{F}_{\rm cube};n)\succeq\underline{\epsilon}_n.$$

Under the richness assumption and that $\underline{M}(\epsilon)$ is of the same order as $M(\epsilon)$, $\epsilon_n$ and $\underline{\epsilon}_n$ are of the same order. Thus $r(\mathcal{F};n)\asymp\epsilon_n$. The proof of the remaining statement is similar to that for Theorem 1.

*Proof of Corollary 3:* The lower bound rate on $R(\tilde{\mathcal{S}}(\Gamma,\Phi);\mathcal{H}(A);n)$ follows from Corollary 2 together with that $\mathcal{F}(\Gamma,\Phi)$ is a subset of $\mathcal{S}(\Gamma,\Phi)$. Let

$$k(\epsilon)=\min\{k\colon\gamma_k\le\epsilon/2\}.$$

Then from [36], the metric entropy of $\mathcal{S}(\Gamma,\Phi)$ is upper-bounded by order $k(\epsilon)\log(\epsilon^{-1})$. Thus based on Lemma 1 and Assumption 2′, $\epsilon_n^2$ determined by $n\epsilon_n^2=k(\epsilon_n)\log(\epsilon_n^{-1})$ gives an upper bound rate for the sparse approximation set. Under the assumption $k_n\preceq n^{1-\tau}$ for some $\tau>0$, it can be shown that $\log(\epsilon_n^{-1})$ is of order $\log n$, and $\epsilon_n$ is of the same order as $k_{\lfloor n/\log n\rfloor}\log n/n$ with $k_n$ defined in (7). This completes the proof of Corollary 3.

*Lemma 4:* Let $\mathcal{F}(\Gamma,\Phi)$ and $\tilde{\mathcal{F}}(\Gamma,\Phi)$ be as in Section IV. If $\sup_{g\in\mathcal{F}(\Gamma,\Phi)}||g||_\infty\le C<\infty$, then $\mathcal{F}(\Gamma,\Phi)$ and $\tilde{\mathcal{F}}(\Gamma,\Phi)$ have the same order metric entropies.

*Proof of Lemma 4:* Note that for $0<\tau\le 1$

$$\mathcal{F}_\tau(\Gamma,\Phi)=\{\tau g\colon g\in\mathcal{F}(\Gamma,\Phi)\}$$

is a subset of $\mathcal{F}(\Gamma,\Phi)$ but with the same order metric entropy. Since $g_0\equiv\min(1,\gamma_0)\in\mathcal{F}(\Gamma,\Phi)$ and $\mathcal{F}(\Gamma,\Phi)$ is convex, we have $\mathcal{G}=\{g_0/2+g/2\colon g\in\mathcal{F}_\tau(\Gamma,\Phi)\}\subset\mathcal{F}(\Gamma,\Phi)$. Again $\mathcal{G}$ and $\mathcal{F}_\tau(\Gamma,\Phi)$ have the same order metric entropies. When $\tau$ is small enough, under the uniform boundedness assumption, the functions in $\mathcal{G}$ are between 0 and 1, i.e., $\mathcal{G}\subset\tilde{\mathcal{F}}(\Gamma,\Phi)$. As a consequence, we know $\mathcal{F}(\Gamma,\Phi)$ and $\tilde{\mathcal{F}}(\Gamma,\Phi)$ have the same order metric entropies. This completes the proof of Lemma 4.

## ACKNOWLEDGMENT

The author thanks three referees and the associate editor for their very helpful comments and suggestions.

## REFERENCES

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

[2] ——, "Complexity regularization," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991, pp. 561–576.

[3] ———, "Neural net approximation," in *Proc. Yale Workshop Adaptive Learning Syst.*, K. Narendra, Ed. New Haven, CT: Yale Univ. Press, May 1992.
[4] ———, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–945, 1993.
[5] ———, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994.
[6] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.
[7] L. Birgé, "Approximation dans les espaces metriques et theorie de l'estimation," *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, vol. 65, pp. 181–237, 1983.
[8] ———, "On estimating a density using Hellinger distance and some other strange facts," *Probab. Theory Related Fields*, vol. 71, pp. 271–291, 1986.
[9] M. S. Birman and M. Solomajak, "Piecewise polynomial approximation of functions of the class $W_p^\alpha$," *Matem. Sbornik, N.S.*, vol. 2, pp. 295–317, 1967.
[10] ———, "Quantitative analysis in Sobolev imbedding theorems and application to spectral theory," in *10th Math. School Kiev*, 1974, pp. 5–189; English transl.: *Amer. Math. Soc. Transl.*, vol. 114, 1980, pp. 1–132.
[11] J. Bretagnolle and C. Huber, "Estimation des densites: Risque minimax," *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, vol. 47, pp. 119–137, 1979.
[12] B. Carl, "Entropy numbers of embedding maps between Besov spaces with an application to eigenvalue problems," *Proc. Roy. Soc. Edinburgh*, vol. 90A, pp. 63–70 1981.
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[14] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. New York: Springer-Verlag, 1993.
[15] L. Devroye and L. Györfi, "Distribution-free exponential bound on the $L_1$ error of partitioning estimates of a regression functions," in *Proc. 4th Pannonian Symp. Mathematical Statistics*, F. Konecny, J. Mogyoródi, and W. Wertz, Eds. Budapest, Hungary: Akadémiai Kiadó, 1983, pp. 67–76.
[16] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[17] L. Gorden and R. Olshen, "Asymptotically efficient solutions to the classification problem," *Ann. Statist.*, vol. 6, pp. 515–533, 1978.
[18] W. Greblicki, "Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 364–366, 1981.
[19] L. Györfi, "The rates of convergence of $k_n$-NN regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 362–364, 1981.
[20] C. Huber, "Lower bounds for function estimation," in *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. Yang, Eds. New York: Springer-Verlag, 1996, pp. 245–258.
[21] I. A. Ibragimov and R. Z. Hasminskii, "On the estimation of an infinite-dimensional parameter in Gaussian white noise," *Sov. Math.–Dokl.*, vol. 18, pp. 1307–1309, 1977.
[22] G. Kerkyacharian and D. Picard, "Density estimation in Besov spaces," *Statist. Probab. Lett.*, vol. 13, pp. 15–24, 1992.
[23] A. N. Kolmogorov and V. M. Tihomirov, "$\epsilon$-entropy and $\epsilon$-capacity of sets in function spaces," *Usp. Mat. Nauk*, vol. 14, pp. 3–86, 1959; English transl.: *Amer. Math. Soc. Transl.*, vol. 17, pp. 277–364, 1961.
[24] A. Krzyżak, "The rate of convergence of kernel regression and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 668–679, 1986.
[25] G. G. Lorentz, "Metric entropy and approximation," *Bull. Amer. Math. Soc.*, vol. 72, pp. 903–937, 1966.
[26] G. G. Lorentz, M. V. Golitschek, and Y. Makovoz, *Constructive Approximation: Advanced Problems*. New York: Springer-Verlag, 1996.
[27] G. Lugosi and A. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *Ann. Statist.*, vol. 24, pp. 687–706, 1996.
[28] Y. Makovoz, "Random approximants and neural networks," *J. Approx. Theory*, vol. 85, pp. 98–109, 1996.
[29] J. S. Marron, "Optimal rates of convergence to Bayes risk in nonparametric discrimination," *Ann. Statist.*, vol. 11, pp. 1142–1155, 1983.
[30] B. S. Mitjagin, "The approximation dimension and bases in nuclear spaces," *Usp. Mat. Nauk*, vol. 16, pp. 63–132, 1961.
[31] S. A. Smolyak, "The $\epsilon$-entropy of some classes $E_s^{\alpha,k}(B)$ and $W_s^\alpha(B)$ in the $L_2$ metric," *Dokl. Akad. Nauk SSSR*, vol. 131, pp. 30–33, 1960. English transl.: in *Sov. Math.–Dolk.*, vol. 1, pp. 192–195, 1960.
[32] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 8, pp. 1348–1360, 1977.
[33] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*. New York: MacMillan, 1963.
[34] H. Triebel, "Interpolation properties of $\epsilon$-entropy and diameters. Geometric characteristics of imbedding for function spaces of Sobolev–Besov type," *Mat. Sbornik*, vol. 98, pp. 27–41, 1975; English Transl.: in *Math. USSR Sb.*, vol. 27, pp. 23–37, 1977.
[35] ———, *Theory of Function Spaces II*. Basel and Boston: Birkhauser, 1992.
[36] Y. Yang and A. R. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statistics*, 1999, accepted for publication.
[37] Y. Yang, "Minimax nonparametric classification—Part II: Model selection for adaptation," this issue, pp. 0000–0000.
[38] Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy," *Ann. Statist.*, vol. 13, 768–774, 1985.