The Annals of Statistics
1999, Vol. 27, No. 5, 1564–1599

INFORMATION-THEORETIC DETERMINATION OF MINIMAX RATES OF CONVERGENCE[1]

BY YUHONG YANG AND ANDREW BARRON

Iowa State University and Yale University

We present some general results determining minimax bounds on statistical risk for density estimation based on certain information-theoretic considerations. These bounds depend only on metric entropy conditions and are used to identify the minimax rates of convergence.

1. Introduction. The metric entropy structure of a density class determines the minimax rate of convergence of density estimators. Here we prove such results using new direct metric entropy bounds on the mutual information that arises by application of Fano's information inequality in the development of lower bounds characterizing the optimal rate. No special construction is required for each density class.

We here study global measures of loss such as integrated squared error, squared Hellinger distance or Kullback–Leibler (K-L) divergence in nonparametric curve estimation problems.

The minimax rates of convergence are often determined in two steps. A good lower bound is obtained for the target family of densities, and a specific estimator is constructed so that the maximum risk is within a constant factor of the derived lower bound. For global minimax risk as we are considering here, inequalities for hypothesis tests are often used to derive minimax lower bounds, including versions of Fano's inequality. These methods are used in Ibragimov and Hasminskii (1977, 1978), Hasminskii (1978), Bretagnolle and Huber (1979), Efroimovich and Pinsker (1982), Stone (1982), Birgé (1983, 1986), Nemirovskii (1985), Devroye (1987), Le Cam (1986), Yatracos (1988), Hasminskii and Ibragimov (1990), Yu (1996) and others. Upper bounds on minimax risk under metric entropy conditions are in Birgé (1983, 1986), Yatracos (1985), Barron and Cover (1991), Van de Geer (1990), Wong and Shen (1995) and Birgé and Massart (1993, 1994). The focus of the present paper is on lower bounds determining the minimax rate, though some novel upper bound results are given as well. Parallel to the development of risk bounds for global measures of loss are results for point estimation of a density or functionals of the density; see, for example, Farrell (1972), Donoho and Liu (1991), Birgé and Massart (1995).

Received November 1995; revised July 1999.

[1]Supported in part by NSF Grants ECS-94-10760 and DMS-95-05168.

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

Key words and phrases. Minimax risk, density estimation, metric entropy, Kullback–Leibler distance.

In its original form, Fano's inequality relates average probability of error in a multiple hypothesis test to the Shannon mutual information for a joint distribution of parameter and random sample [Fano (1961); see also Cover and Thomas (1991), pages 39 and 205]. Beginning with Fano, this inequality has been used in information theory to determine the capacity of communication channels. Ibragimov and Hasminskii (1977, 1978) initiated the use of Fano's inequality in statistics for determination of minimax rates of estimation for certain classes of functions. To apply this technique, suitable control of the Shannon information is required. Following Birgé (1983), statisticians have stated and used versions of Fano's inequality where the mutual information is replaced by a bound involving the diameter of classes of densities using the Kullback–Liebler divergence. For success of this tactic to obtain lower bounds on minimax risk, one must make a restriction to small subsets of the function space and then establish the existence of a packing set of suitably large cardinality. As Birgé (1986) points out, in that manner, the use of Fano's inequality is similar to the use of Assouad's lemma. As it is not immediately apparent that such local packing sets exist, these techniques have been applied on a case by case basis to establish minimax rates for various function classes in the above cited literature.

In this paper we provide two means by which to reveal minimax rates from global metric entropies. The first is by use of Fano's inequality in its original form together with suitable information inequalities. This is the approach we take in Sections 2 through 6 to get minimax bounds for various measures of loss. The second approach we take in Section 7 shows that global metric entropy behavior implies the existence of a local packing set satisfying conditions for the theory of Birgé (1983) to be applicable.

Thus we demonstrate for nonparametric function classes (and certain measures of loss) that the minimax convergence rate is determined by the global metric entropy over the whole function class (or over large subsets of it). The advantage is that the metric entropies are available in approximation theory for many function classes [see, e.g., Lorentz, Golitzchek and Makovoz (1996)]. It is no longer necessary to uncover additional local packing properties on a case by case basis.

The following proposition is representative of the results obtained here. Let $\mathscr{F}$ be a class of functions and let $d(f, g)$ be a distance between functions [we require $d(f, g)$ nonnegative and equal to zero for $f = g$, but we do not require it to be a metric]. Let $N(\varepsilon; \mathscr{F})$ be the size of the largest packing set of functions separated by at least ε in $\mathscr{F}$, and let ε_n satisfy $\varepsilon_n^2 = M(\varepsilon_n)/n$, where $M(\varepsilon) = \log N(\varepsilon; \mathscr{F})$ is the Kolmogorov ε-entropy and n is the sample size. Assume the target class $\mathscr{F}$ is rich enough to satisfy $\liminf_{\varepsilon \to 0} M(\varepsilon/2)/M(\varepsilon) > 1$ [which is true, e.g., if $M(\varepsilon) = (1/\varepsilon)^r \kappa(\varepsilon)$ with $r > 0$ and $\kappa(\varepsilon/2)/\kappa(\varepsilon) \to 1$ as $\varepsilon \to 0$]. This condition is satisfied in typical nonparametric classes.

For convenience, we will use the symbols $\succeq$ and $\asymp$, where $a_n \succeq b_n$ means $b_n = O(a_n)$, and $a_n \asymp b_n$ means both $a_n \succeq b_n$ and $b_n \succeq a_n$. Probability densities are taken with respect to a finite measure μ in the following proposition.

PROPOSITION 1. *In the following cases, the minimax convergence rate is characterized by metric entropy in terms of the critical separation* ε_n *as follows*:

$$\min_f \max_{f \in \mathcal{F}} E_f d^2(f, \hat{f}) \asymp \varepsilon_n^2.$$

(i) $\mathcal{F}$ *is any class of density functions bounded above and below* $0 < \underline{C} \le f \le \overline{C}$ *for* $f \in \mathcal{F}$. *Here* $d^2(f, g)$ *is either integrated squared error* $\int (f(x) - g(x))^2 \, d\mu$, *squared Hellinger distance or Kullback–Leibler divergence.*

(ii) $\mathcal{F}$ *is a convex class of densities with* $f \le \overline{C}$ *for* $f \in \mathcal{F}$ *and there exists at least one density in* $\mathcal{F}$ *bounded away from zero and* d *is the* L_2 *distance.*

(iii) $\mathcal{F}$ *is any class of functions* f *with* $| f | \le \overline{C}$ *for* $f \in \mathcal{F}$ *for the regression model* $Y = f(X) + \varepsilon$, X *and* ε *are independent* $X \sim P_X$ *and* $\varepsilon \sim$ *Normal* $(0, \sigma^2)$, $\sigma > 0$ *and* d *is the* $L_2(P_X)$ *norm.*

From the above proposition, the minimax L^2 risk rate is determined by the metric entropy alone, whether the densities can be zero or not. For Hellinger and Kullback–Leibler (K-L) risk, we show that by modifying a nonparametric class of densities with uniformly bounded logarithms to allow the densities to approach zero or even vanish in some unknown subsets, the minimax rate may remain unchanged compared to that of the original class.

Now let us outline roughly the method of lower bounding the minimax risk using Fano's inequality. The first step is to restrict attention to a subset S_0 of the parameter space where minimax estimation is nearly as difficult as for the whole space and, moreover, where the loss function of interest is related locally to the K-L divergence that arises in Fano's inequality. (For example, the subset can in some cases be the set of densities with a bound on their logarithms.) As we shall reveal, the lower bound on the minimax rate is determined by the metric entropy of the subset.

The proof technique involving Fano's inequality first lower bounds the minimax risk by restricting to as large as possible a finite set of parameter values $\{\theta_1, \ldots, \theta_m\}$ in S_0 separated from each other by an amount ε_n in the distance of interest. The critical separation ε_n is the largest separation such that the hypotheses $\{\theta_1, \ldots, \theta_m\}$ are nearly indistinguishable on the average by tests as we shall see. Indeed, Fano's inequality reveals this indistinguishability in terms of the K-L divergence between densities $p_{\theta_j}(x_1, \ldots, x_n) = \prod_{i=1}^n p_{\theta_j}(x_i)$ and the centroid of such densities $q(x_1, \ldots, x_n) = (1/m) \sum_{j=1}^m p_{\theta_j}(x_1, \ldots, x_n)$ [which yields the Shannon mutual information $I(\Theta; X_1, \ldots, X_n)$ between θ and $X_1, \ldots, X_n$ with a uniform distribution on θ in $\{\theta_1, \ldots, \theta_m\}$]. Here the key question is to determine the separation such that the average of this K-L divergence is small compared to the distance $\log m$ that would correspond to maximally distinguishable densities (for which θ is determined by X^n). It is critical here that K-L divergence does not have a triangle inequality between the joint densities. Indeed, covering entropy properties of the K-L divergence (under the conditions of the proposition) show that the K-L divergence from every $p_{\theta_j}(x_1, \ldots, x_n)$ to the centroid is bounded by the right order $2n\varepsilon_n^2$, even though the distance between two such $p_{\theta_j}(x_1, \ldots, x_n)$ is as large as $n\beta$

where β is the K-L diameter of the whole set $\{p_{\theta_1}, \ldots, p_{\theta_m}\}$. The proper convergence rate is thus identified provided the cardinality of the subset m is chosen such that $n\varepsilon_n^2/\log m$ is bounded by a suitable constant less than 1. Thus ε_n is determined by solving for $n\varepsilon_n^2/M(\varepsilon_n)$ equal to such a constant, where $M(\varepsilon)$ is the metric entropy (the logarithm of the largest cardinality of an ε-packing set). In this way, the metric entropy provides a lower bound on the minimax convergence rate.

Applications of Fano's inequality to density estimation have used the K-L diameter $n\beta$ of the set $\{p_{\theta_1}^n, \ldots, p_{\theta_m}^n\}$ [see, e.g., Birgé (1983)] or similar rough bounds [such as $nI(\Theta; X_1)$ as in Hasminskii (1978)] in place of the average distance of $p_{\theta_1}^n, \ldots, p_{\theta_m}^n$ from their centroid. In that theory, to obtain a suitable bound, a statistician needs to find if possible a sufficiently large subset $\{\theta_1, \ldots, \theta_m\}$ for which the diameter of this subset (in the K-L sense) is of the same order as the separation between closest points in this subset (in the chosen distance). Apparently, such a bound is possible only for subsets of small diameter. Thus by that technique, knowledge is needed not only of the metric entropy but also of special localized subsets. Typical tools for smoothness classes involve perturbations of densities parametrized by vertices of a hypercube. While interesting, such involved calculations are not needed to obtain the correct order bounds. It suffices to know or bound the metric entropy of the chosen set S_0. Our results are especially useful for function classes whose global metric entropies have been determined but local packing sets have not been identified for applying Birgé's theory (e.g., general linear and sparse approximation sets of functions in Sections 4 and 5, and neural network classes in Section 6).

It is not our purpose to criticize the use of hypercube-type arguments in general. In fact, besides the success of such methods mentioned above, they are also useful in other applications such as determining the minimax rates of estimating functionals of densities [see, e.g., Bickel and Ritov (1988), Birgé and Massart (1995) and Pollard (1993)] and minimax rates in nonparametric classification [Yang (1999b)].

The density estimation problem we consider is closely related to a data compression problem in information theory (see Section 3). The relationship allows us to obtain both upper and lower bounds on the minimax risk from upper-bounding the minimax redundancy of data compression, which is related to the global metric entropy.

Le Cam (1973) pioneered the use of local entropy conditions in which convergence rates are characterized in terms of the covering or packing of balls of radius ε by balls of radius $\varepsilon/2$, with subsequent developments by Birgé and others, as mentioned above. Such local entropy conditions provide optimal convergence rates in finite-dimensional as well as infinite-dimensional settings. In Section 7, we show that knowledge of global metric entropy provides the existence of a set with suitable local entropy properties in infinite-dimensional settings. In such cases, there is no need to explicitly require or construct such a set.

The paper is divided into 7 sections. In Section 2, the main results are presented. Applications in data compression and regression are given in Section 3. In Sections 4 and 5, results connecting linear approximation and minimax rates, sparse approximation and minimax rates, respectively, are given. In Section 6, we illustrate the determination of minimax rates of convergence for several function classes. In Section 7, we discuss the relationship between the global entropy and local entropy. The proofs of some lemmas are given in the Appendix.

2. Main results. Suppose we have a collection of densities $\{p_\theta: \theta \in \Theta\}$ defined on a measurable space $\mathscr{X}$ with respect to a σ-finite measure μ. The parameter space Θ could be a finite-dimensional space or a nonparametric space (e.g., the class of all densities). Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from p_θ, $\theta \in \Theta$. We want to estimate the true density p_θ or θ based on the sample. The K-L loss, the squared Hellinger loss, the integrated squared error and some other losses will be considered in this paper. We determine minimax bounds for subclasses $\{p_\theta: \theta \in S\}$, $S \subseteq \Theta$. When the parameter is the density itself, we may use f and $\mathscr{F}$ in place of θ and S, respectively. Our technique is most appropriate for nonparametric classes (e.g., monotone, Lipschitz, or neural net; see Section 6).

Let $\overline{S}$ be an action space for the parameter estimates with $S \subseteq \overline{S} \subseteq \Theta$. An estimator $\hat{\theta}$ is then a measurable mapping from the sample space of $X_1, X_2, \ldots, X_n$ to $\overline{S}$. Let $\overline{S}_n$ be the collection of all such estimators. For nonparametric density estimation, $\overline{S} = \Theta$ is often chosen to be the set of all densities or some transform of the densities (e.g., square root of density). We consider general loss functions d, which are mappings from $\overline{S} \times \overline{S}$ to R^+ with $d(\theta, \theta) = 0$ and $d(\theta, \theta') > 0$ for $\theta \neq \theta'$. We call such a loss function a *distance* whether or not it satisfies properties of a metric.

The minimax risk of estimating $\theta \in S$ with action space $\overline{S}$ is defined as

$$r_n = \min_{\hat{\theta} \in \overline{S}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}).$$

Here "min" and "max" are understood to be "inf" and "sup," respectively, if the minimizer or maximizer does not exist.

DEFINITION 1. A finite set $N_\varepsilon \subset S$ is said to be an *ε-packing set* in S with separation $\varepsilon > 0$, if for any $\theta, \theta' \in N_\varepsilon$, $\theta \neq \theta'$, we have $d(\theta, \theta') > \varepsilon$. The logarithm of the maximum cardinality of ε-packing sets is called the *packing ε-entropy* or Kolmogorov capacity of S with distance function d and is denoted $M_d(\varepsilon)$.

DEFINITION 2. A set $G_\varepsilon \subset \overline{S}$ is said to be an *ε-net* for S if for any $\tilde{\theta} \in S$, there exists a $\theta_0 \in G_\varepsilon$ such that $d(\tilde{\theta}, \theta_0) \leq \varepsilon$. The logarithm of the minimum cardinality of ε-nets is called the *covering ε-entropy* of S and is denoted $V_d(\varepsilon)$.

From the definitions, it is straightforward to see that $M_d(\varepsilon)$ and $V_d(\varepsilon)$ are nonincreasing in ε and $M_d(\varepsilon)$ is right continuous. These definitions are slight generalizations of the metric entropy notions introduced by Kolmogorov and Tihomirov (1959). In accordance with common terminology, we informally call these ε-entropies "metric" entropies even when the distance is not a metric. One choice is the square root of the Kullback–Leibler (K-L) divergence which we denote by d_K, where $d_K^2(\theta, \theta') = D(p_\theta \| p_{\theta'}) = \int p_\theta \log(p_\theta / p_{\theta'})\, d\mu$. Clearly $d_K(\theta, \theta')$ is asymmetric in its two arguments, so it is not a metric. Other distances we consider include the Hellinger metric $d_H(\theta, \theta') = \left(\int \left(\sqrt{p_\theta} - \sqrt{p_{\theta'}}\right)^2 d\mu\right)^{1/2}$ and the $L_q(\mu)$ metric $d_q(\theta, \theta') = (\int |p_\theta - p_{\theta'}|^q\, d\mu)^{1/q}$ for $q \geq 1$. The Hellinger distance is upper bounded by the square root of the K-L divergence, that is, $d_H(\theta, \theta') \leq d_K(\theta, \theta')$. We assume the distance d satisfies the following condition.

CONDITION 0 (Local triangle inequality). There exist positive constants $A \leq 1$ and ε_0 such that for any $\theta,\ \theta' \in S,\ \tilde{\theta} \in \overline{S}$, if $\max\left(d(\theta, \tilde{\theta}), d(\theta', \tilde{\theta})\right) \leq \varepsilon_0$, then $d(\theta, \tilde{\theta}) + d(\theta', \tilde{\theta}) \geq A d(\theta, \theta')$.

REMARKS. (i) If d is a metric on Θ, then the condition is always satisfied with $A = 1$ for any $\overline{S} \subseteq \Theta$.

(ii) For d_K, the condition is not automatically satisfied. It holds when S is a class of densities with bounded logarithms and $\overline{S} \subseteq \Theta$ is any action space, including the class of all densities. It is also satisfied by regression families as considered in Section 3. See Section 2.3 for more discussion.

When Condition 0 is satisfied, the packing entropy and covering entropy have the simple relationship for $\varepsilon \leq \varepsilon_0,\ M_d(2\varepsilon/A) \leq V_d(\varepsilon) \leq M_d(A\varepsilon)$.

We will obtain minimax results for such general d and then special results will be given with several choices of d: the square root K-L divergence, Hellinger distance and L_q distance. We assume $M_d(\varepsilon) < \infty$ for all $\varepsilon > 0$. The square root K-L, Hellinger and L_q packing entropies are denoted $M_K(\varepsilon)$, $M_H(\varepsilon)$ and $M_q(\varepsilon)$, respectively.

2.1. *Minimax risk under a global entropy condition.* Suppose a good upper bound on the covering ε-entropy under the square root K-L divergence is available. That is, assume $V_K(\varepsilon) \leq V(\varepsilon)$. Ideally, $V_K(\varepsilon)$ and $V(\varepsilon)$ are of the same order. Similarly, let $M(\varepsilon) \leq M_d(\varepsilon)$ be an available lower bound on ε-packing entropy with distance d, and ideally, $M(\varepsilon)$ is of the same order as $M_d(\varepsilon)$. Suppose $V(\varepsilon)$ and $M(\varepsilon)$ are nonincreasing and right-continuous functions. To avoid a trivial case (in which S is a small finite set), we assume $M(\varepsilon) > 2 \log 2$ for ε small enough. Let ε_n (called the critical covering radius) be determined by

$$\varepsilon_n^2 = V(\varepsilon_n)/n.$$

The trade-off here between $V(\varepsilon)/n$ and ε^2 is analogous to that between the squared bias and variance of an estimator. As will be shown later, $2\varepsilon_n^2$ is an

upper bound on the minimax K-L risk. Let $\underline{\varepsilon}_{n,d}$ be a separation ε such that

$$M(\underline{\varepsilon}_{n,d}) = 4n\varepsilon_n^2 + 2\log 2.$$

We call $\underline{\varepsilon}_{n,d}$ the packing separation commensurate with the critical covering radius ε_n. This $\underline{\varepsilon}_{n,d}^2$ determines a lower bound on the minimax risk.

In general, the upper and lower rates ε_n^2 and $\underline{\varepsilon}_{n,d}^2$ need not match. Conditions under which they are of the same order are explored at the end of this subsection and are used in Section 2.2 to identify the minimax rate for L_2 distance and in Sections 2.3 and 2.4 to relate the minimax K-L risk to the minimax Hellinger risk and the Hellinger metric entropy.

THEOREM 1 (Minimax lower bound). *Suppose Condition* 0 *is satisfied for the distance* d. *Then when the sample size* n *is large enough such that* $\underline{\varepsilon}_{n,d} \le 2\varepsilon_0$, *the minimax risks for estimating* $\theta \in S$ *satisfies*

$$\min_{\hat{\theta}} \max_{\theta \in S} P_\theta\{d(\theta, \hat{\theta}) \ge (A/2)\underline{\varepsilon}_{n,d}\} \ge 1/2$$

and consequently,

$$\min_{\hat{\theta}} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) \ge (A^2/8)\underline{\varepsilon}_{n,d}^2,$$

where the minimum is over all estimators mapping from $\mathcal{X}^n$ *to* $\overline{S}$.

PROOF. Let $N_{\underline{\varepsilon}_{n,d}}$ be an $\underline{\varepsilon}_{n,d}$-packing set with the maximum cardinality in S under the given distance d and let G_{ε_n} be an ε_n-net for S under d_K. For any estimator $\hat{\theta}$ taking values in $\overline{S}$, define $\tilde{\theta} = \arg\min_{\theta' \in N_{\underline{\varepsilon}_{n,d}}} d(\theta', \hat{\theta})$ (if there are more than one minimizer, choose any one), so that $\tilde{\theta}$ takes values in the finite packing set $N_{\underline{\varepsilon}_{n,d}}$. Let θ be any point in $N_{\underline{\varepsilon}_{n,d}}$. If $d(\theta, \hat{\theta}) < A\underline{\varepsilon}_{n,d}/2$, then $\max\big(d(\theta, \hat{\theta}), d(\tilde{\theta}, \hat{\theta})\big) < A\underline{\varepsilon}_{n,d}/2 \le \varepsilon_0$ and hence by Condition 0, $d(\theta, \hat{\theta}) + d(\tilde{\theta}, \hat{\theta}) \ge Ad(\theta, \tilde{\theta})$, which is at least $A\underline{\varepsilon}_{n,d}$ if $\theta \ne \tilde{\theta}$. Thus if $\theta \ne \tilde{\theta}$, we must have $d(\theta, \hat{\theta}) \ge A\underline{\varepsilon}_{n,d}/2$, and

$$\begin{aligned}
\min_{\hat{\theta}} \max_{\theta \in S} P_\theta\{d(\theta, \hat{\theta}) \ge (A/2)_{\underline{\varepsilon}_{n,d}}\} &\ge \min_{\hat{\theta}} \max_{\theta \in N_{\underline{\varepsilon}_{n,d}}} P_\theta\{d(\theta, \hat{\theta}) \ge (A/2)\underline{\varepsilon}_{n,d}\} \\
&= \min_{\hat{\theta}} \max_{\theta \in N_{\underline{\varepsilon}_{n,d}}} P_\theta\Big(\theta \ne \tilde{\theta}\Big) \\
&\ge \min_{\hat{\theta}} \sum_{\theta \in N_{\underline{\varepsilon}_{n,d}}} w(\theta) P_\theta\Big(\theta \ne \tilde{\theta}\Big) \\
&= \min_{\hat{\theta}} P_w\Big(\theta \ne \tilde{\theta}\Big),
\end{aligned}$$

where in the last line, θ is randomly drawn according to a discrete prior probability w restricted to $N_{\underline{\varepsilon}_{n,d}}$, and P_w denotes the Bayes average probability with respect to the prior w. Moreover, since $(d(\theta, \hat{\theta}))'$ is not less than

$((A/2)\underline{\varepsilon}_{n,d})^{\prime} 1_{\{\theta \neq \tilde{\theta}\}}$, taking the expected value it follows that for all $\prime > 0$,

$$\min_{\hat{\theta}} \max_{\theta \in S} E_{\theta} d^{\prime}(\theta, \hat{\theta}) \geq ((A/2)\underline{\varepsilon}_{n,d})^{\prime} \min_{\hat{\theta}} P_w(\theta \neq \tilde{\theta}).$$

By Fano's inequality [see, e.g., Fano (1961) or Cover and Thomas (1991), pages 39 and 205], with w being the discrete uniform prior on θ in the packing set $N_{\underline{\varepsilon}_{n,d}}$, we have

$$P_w(\theta \neq \tilde{\theta}) \geq 1 - \frac{I(\Theta; X^n) + \log 2}{\log |N_{\underline{\varepsilon}_{n,d}}|}, \tag{1}$$

where $I(\Theta; X^n)$ is Shannon's mutual information between the random parameter and the random sample, when θ is distributed according to w. This mutual information is equal to the average (with respect to the prior) of the K-L divergence between $p(x^n|\theta)$ and $p_w(x^n) = \sum_{\theta} w(\theta) p(x^n|\theta)$, where $p(x^n|\theta) = p_{\theta}(x^n) = \prod_{i=1}^{n} p_{\theta}(x_i)$ and $x^n = (x_1, \ldots, x_n)$. It is upper bounded by the maximum K-L divergence between the product densities $p(x^n|\theta)$ and any joint density $q(x^n)$ on the sample space $\mathscr{X}^n$. Indeed,

$$\begin{aligned} I(\Theta; X^n) &= \sum_{\theta} w(\theta) \int p(x^n|\theta) \log(p(x^n|\theta)/p_w(x^n)) \mu(dx^n) \\ &\leq \sum_{\theta} w(\theta) \int p(x^n|\theta) \log(p(x^n|\theta)/q(x^n)) \mu(dx^n) \\ &\leq \max_{\theta \in N_{\underline{\varepsilon}_{n,d}}} D(P_{X^n|\theta} \| Q_{X^n}). \end{aligned}$$

The first inequality above follows from the fact that the Bayes mixture density $p_w(x^n)$ minimizes the average K-L divergence over choices of densities $q(x^n)$ [any other choice yields a larger value by the amount $\int p_w(x^n) \log(p_w(x^n)/q(x^n)) \mu(dx^n) > 0$]. We have w uniform on $N_{\underline{\varepsilon}_{n,d}}$. Now choose w_1 to be the uniform prior on G_{ε_n} and let $q(x^n) = p_{w_1}(x^n) = \sum_{\theta} w_1(\theta) p(x^n|\theta)$ and Q_{X^n} be the corresponding Bayes mixture density and distribution, respectively. Because G_{ε_n} is an ε_n-net in S under d_K, for each $\theta \in S$, there exists $\bar{\theta} \in G_{\varepsilon_n}$ such that $D(p_{\theta} \| p_{\bar{\theta}}) = d_K^2(\theta, \bar{\theta}) \leq \varepsilon_n^2$. Also by definition, $\log |G_{\varepsilon_n}| \leq V_K(\varepsilon_n)$. It follows that

$$\begin{aligned} D(P_{X^n|\theta} \| Q_{X^n}) &= E \log \frac{p(X^n|\theta)}{(1/|G_{\varepsilon_n}|) \sum_{\theta' \in G_{\varepsilon_n}} p(X^n|\theta')} \\ &\leq E \log \frac{p(X^n|\theta)}{(1/|G_{\varepsilon_n}|) p(X^n|\bar{\theta})} \\ &= \log |G_{\varepsilon_n}| + D(P_{X^n|\theta} \| P_{X^n|\bar{\theta}}) \\ &\leq V(\varepsilon_n) + n\varepsilon_n^2. \end{aligned} \tag{2}$$

Thus, by our choice of $\underline{\varepsilon}_{n,d}$, $(I(\Theta; X^n) + \log 2)/\log |N_{\underline{\varepsilon}_{n,d}}| \leq 1/2$. The conclusion follows. This completes the proof of Theorem 1. $\square$

REMARKS. (i) Up to inequality (1), the development here is standard. Previous use of Fano's inequality for minimax lower bound takes one of the following weak bounds on mutual information: $I(\Theta; X^n) \le nI(\Theta; X_1)$ or $I(\Theta; X^n) \le n \max_{\theta, \theta' \in \Theta} D(P_{X_1|\theta} \| P_{X_1|\theta'})$ [see, e.g., Hasminskii (1978) and Birgé (1983), respectively]. An exception is work of Ibragimov and Hasminskii (1978) where a more direct evaluation of the mutual information for Gaussian stochastic process models is used.

(ii) Our use of the improved bound is borrowed from ideas in universal data compression for which $I(\Theta; X^n)$ represents the Bayes average redundancy and $\max_{\theta \in S} D(P_{X^n|\theta} \| P_{X^n})$ represents an upper bound on the minimax redundancy $C_n = \min_{Q_{X^n}} \max_{\theta \in S} D(P_{X^n|\theta} \| Q_{X^n}) = \max_w I_w(\theta; X^n)$, where the maximum is over priors supported on S. The universal data compression interpretations of these quantities can be found in Davisson (1973) and Davisson and Leon-Garcia (1980) [see Clarke and Barron (1994), Yu (1996), Haussler (1997) and Haussler and Opper (1997) for some of the recent work in that area]. The bound $D(P_{X^n|\theta} \| P_{X^n}) \le V(\varepsilon_n) + n\varepsilon_n^2$ has roots in Barron [(1987), page 89], where it is given in a more general form for arbitrary priors, that is $D(P_{X^n|\theta} \| P_{X^n}) \le \log 1/w(\mathcal{N}_{\theta, \varepsilon}) + n\varepsilon^2$, where $\mathcal{N}_{\theta, \varepsilon} = \{\theta' : D(p_\theta \| p_{\theta'}) \le \varepsilon^2\}$ and P_{X^n} has density $p_w(x^n) = \int p(x^n|\theta) w(d\theta)$. The redundancy bound $V(\varepsilon_n) + n\varepsilon_n^2$ can also be obtained from use of a two-stage code of length $\log |G_{\varepsilon_n}| + \min_{\theta' \in G_{\varepsilon_n}} \log 1/p(x^n|\theta')$ [see Barron and Cover (1991), Section V].

(iii) From inequality (1), the minimax risk is bounded below by a constant times $\underline{\varepsilon}_{n,d}^2 (1 - (C_n + \log 2)/K_n)$, where $C_n = \max_w I_w(\Theta; X^n)$ is the Shannon capacity of the channel $\{p(x^n|\theta), \theta \in S\}$ and $K_n = \log |N_{\underline{\varepsilon}_{n,d}}|$ is the Kolmogorov $\underline{\varepsilon}_{n,d}$-capacity of S. Thus $\underline{\varepsilon}_{n,d}^2$ lower bounds the minimax rate provided the Shannon capacity is less than the Kolmogorov capacity by a factor less than 1. This Shannon–Kolmogorov characterization is emphasized by Ibragimov and Hasminskii (1977, 1978).

(iv) For applications, the lower bounds may be applied to a subclass of densities $\{p_\theta : \theta \in S_0\}$ $(S_0 \subset S)$ which may be rich enough to characterize the difficulty of the estimation of the densities in the whole class yet is easy enough to check the conditions. For instance, if $\{p_\theta : \theta \in S_0\}$ is a subclass of densities that have support on a compact space and $\|\log p_\theta\|_\infty \le T$ for all $\theta \in S_0$, then the square root K-L divergence, Hellinger distance and L_2 distance are all equivalent in the sense that each of them is both upper bounded and lower bounded by multiples of each other.

We now turn our attention to obtaining a minimax upper bound. We use a uniform prior w_1 on the ε-net G_{ε_n} for S under d_K. For $n = 1, 2, \ldots$, let

$$p(x^n) = \sum_{\theta \in G_{\varepsilon_n}} w_1(\theta) p(x^n|\theta) = \frac{1}{|G_{\varepsilon_n}|} \sum_{\theta \in G_{\varepsilon_n}} p(x^n|\theta)$$

be the corresponding mixture density. Let

$$\bar{p}(x) = n^{-1} \sum_{i=0}^{n-1} \hat{p}_i(x) \tag{3}$$

be the density estimator constructed as a Cesaro average of the Bayes predictive density estimators $\hat{p}_i(x) = p(X_{i+1}|X^i)$ evaluated at $X_{i+1} = x$, which equal $p(X^i, x)/p(X^i)$ for $i > 0$ and $\hat{p}_i(x) = p(x) = (1/|G_{\varepsilon_n}|)\sum_{\theta\in G_{\varepsilon_n}} p(x|\theta)$ for $i = 0$. Note that $\hat{p}_i(x)$ is the average of $p(x|\theta)$ with respect to the posterior distribution $p(\theta|X^i)$ on $\theta \in G_{\varepsilon_n}$.

THEOREM 2 (Upper bound). *Let $V(\varepsilon)$ be an upper bound on the covering entropy of S under d_K and let ε_n satisfy $V(\varepsilon_n) = n\varepsilon_n^2$. Then*

$$\min_{\hat{p}} \max_{\theta\in S} E_\theta D(p_\theta \| \hat{p}) \le 2\varepsilon_n^2,$$

where the minimization is over all density estimators. Moreover, if Condition 0 is satisfied for a distance d and $A_0 d^2(\theta, \theta') \le d_K^2(\theta, \theta')$ for all $\theta, \theta' \in \overline{S}$, and if the set $\overline{S}_n$ of allowed estimators (mappings from X^n to $\overline{S}$) contains $\bar{p}$ constructed in (3) above, then when $\underline{\varepsilon}_{n,d} \le 2\varepsilon_0$,

$$(A_0 A/8)\underline{\varepsilon}_{n,d}^2 \le A_0 \min_{\hat{\theta}\in\overline{S}_n} \max_{\theta\in S} E_\theta d^2(\theta, \hat{\theta}) \le \min_{\hat{\theta}\in\overline{S}_n} \max_{\theta\in S} E_\theta d_K^2(\theta, \hat{\theta}) \le 2\varepsilon_n^2.$$

The condition that $\overline{S}_n$ contains $\bar{p}$ in Theorem 2 is satisfied if $\{p_\theta : \theta \in \overline{S}\}$ is convex (because $\bar{p}$ is a convex combination of densities p_θ). In particular, this holds if the action space $\overline{S}$ is the set of all densities on $\mathscr{X}$. In which case, when d is a metric the only remaining condition needed for the second set of inequalities is $A_0 d^2(\theta, \theta') \le d_K^2(\theta, \theta')$ for all θ, θ'. This is satisfied by Hellinger distance and L_1 distance (with $A_0 = 1$ and $A_0 = 1/2$, respectively). When d is the square root K-L divergence, Condition 0 restricts the family $p_\theta : \theta \in S$ (e.g., to consist of densities with uniformly bounded logarithms), though it remains acceptable to let the action space $\overline{S}$ consist of all densities.

PROOF. By convexity and the chain rule [as in Barron (1987)],

$$\begin{aligned} E_\theta D(p_\theta \| \bar{p}) &\le E_\theta\big(n^{-1}\textstyle\sum_{i=0}^{n-1} D(P_{X_{i+1}|\theta} \| P_{X_{i+1}|X^i})\big) \\ &= n^{-1}\textstyle\sum_{i=0}^{n-1} E \log \dfrac{p(X_{i+1}|\theta)}{p(X_{i+1}|X^i)} \\ &= n^{-1} E \log \dfrac{p(X^n|\theta)}{p(X^n)} \\ &= n^{-1} D\big(P_{X^n|\theta} \| P_{X^n}\big) \\ &\le n^{-1}(V(\varepsilon_n) + n\varepsilon_n^2) = 2\varepsilon_n^2, \end{aligned} \tag{4}$$

where the last inequality is as derived as in (2). Combining this upper bound with the lower bound from Theorem 1, this completes the proof of Theorem 2. □

If $\underline{\varepsilon}_{n,d}^2$ and $2\varepsilon_n^2$ converge to 0 at the same rate, then the minimax rate of convergence is identified by Theorem 2. For $\underline{\varepsilon}_{n,d}^2$ and ε_n^2 to be of the same

order, it is sufficient that the following two conditions hold (for a proof of this simple fact, see Lemma 4 in the Appendix).

CONDITION 1 (Metric entropy equivalence). There exist positive constants a, b and c such that when ε is small enough, $M(\varepsilon) \le V(b\varepsilon) \le cM(a\varepsilon)$.

CONDITION 2 (Richness of the function class). For some $0 < \alpha < 1$,

$$\liminf_{\varepsilon \to 0} M(\alpha\varepsilon)/M(\varepsilon) > 1.$$

Condition 1 is the equivalence of the entropy structure under the square root K-L divergence and under d distance when ε is small (as is satisfied, for instance, when all the densities in the target class are uniformly bounded above and away from 0 and d is taken to be Hellinger distance or L_2 distance). Condition 2 requires the density class to be large enough, namely, $M(\varepsilon)$ approaches ∞ at least polynomially fast in $1/\varepsilon$ as $\varepsilon \to 0$, that is, there exists a constant $\delta > 0$ such that $M(\varepsilon) \succeq \varepsilon^{-\delta}$. This condition is typical of nonparametric function classes. It is satisfied in particular if $M(\varepsilon)$ can be expressed as $M(\varepsilon) = \varepsilon^{-r}\kappa(\varepsilon)$, where $r > 0$ and $\kappa(\alpha\varepsilon)/\kappa(\varepsilon) \to 1$ as $\varepsilon \to 0$. In most situations, the metric entropies are known only up to orders, which makes it generally not tractable to check $\liminf_{\varepsilon \to 0} M_d(\alpha\varepsilon)/M_d(\varepsilon) > 1$ for the exact packing entropy function. That is why Condition 2 is stated in terms of a presumed known bound $M(\varepsilon) \le M_d(\varepsilon)$. Both Conditions 1 and 2 are satisfied if, for instance, $M(\varepsilon) \asymp V(\varepsilon) \asymp \varepsilon^{-r}\kappa(\varepsilon)$ with r and $\kappa(\varepsilon)$ as mentioned above.

COROLLARY 1. *Assume Condition* 0 *is satisfied for a distance* d *satisfying* $A_0 d^2(\theta, \theta') \le d_K^2(\theta, \theta')$ *in* $\overline{S}$ *and assume* $\{p_\theta : \theta \in \overline{S}\}$ *is convex. Under Conditions* 1 *and* 2, *we have*

$$\min_{\hat{\theta} \in \overline{S}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) \asymp \varepsilon_n^2,$$

where ε_n *is determined by the equation* $M_d(\varepsilon_n) = n\varepsilon_n^2$. *In particular*, *if* μ *is finite and* $\sup_{\theta \in S} \|\log p_\theta\|_\infty < \infty$ *and Condition* 2 *is satisfied*, *then*

$$\min_{\hat{\theta} \in \overline{S}_n} \max_{\theta \in S} E_\theta d_K^2(\theta, \hat{\theta}) \asymp \min_{\hat{\theta} \in \overline{S}_n} \max_{\theta \in S} E_\theta d_H^2(\theta, \hat{\theta}) \asymp \min_{\hat{\theta} \in \overline{S}_n} \max_{\theta \in S} E_\theta \| p_\theta - p_{\hat{\theta}} \|_2^2 \asymp \varepsilon_n^2,$$

where ε_n *satisfies* $M_2(\varepsilon_n) = n\varepsilon_n^2$ *or* $M_H(\varepsilon_n) = n\varepsilon_n^2$.

Corollary 1 is applicable for many smooth nonparametric classes as we shall see. However, for not very rich classes of densities (e.g., finite-dimensional families or analytical densities), the lower bound and the upper bound derived in the above way do not converge at the same rate. For instance, for a finite-dimensional class, both $M_K(\varepsilon)$ and $M_H(\varepsilon)$ may be of order $\log(1/\varepsilon)^m$ for some constant $m \ge 1$, and then ε_n and $\underline{\varepsilon}_{n,H}$ are not of the same order with $\varepsilon_n \asymp \sqrt{(\log n)/n}$ and $\underline{\varepsilon}_{n,H} = o(1/\sqrt{n})$. For smooth finite-dimensional models, the minimax risk can be solved using some traditional statistical methods (such as Bayes procedures, Cramér–Rao inequality, Van Tree's inequality,

etc.), but these techniques require more than the entropy condition. If local entropy conditions are used instead of those on global entropy, results can be obtained suitable for both parametric and nonparametric families of densities (see Section 7).

2.2. *Minimax rates under L_2 loss.* In this subsection, we derive minimax bounds for L_2 risk without requiring K-L covering entropy.

Let $\mathscr{F}$ be a class of density functions f with respect to a probability measure μ on a measurable set $\mathscr{X}$ such as [0,1]. (Typically $\mathscr{X}$ will be taken to be a compact set, though it need not be; we assume only that the dominating measure μ is finite, then normalized to be a probability measure.) Let the packing entropy of $\mathscr{F}$ be $M_q(\varepsilon)$ under the $L_q(\mu)$ metric.

To derive minimax upper bounds, we derive a lemma that relates L_2 risk for densities that may be zero to the corresponding risk for densities bounded away from zero.

In addition to the observed i.i.d sample $X_1, X_2, \ldots, X_n$ from f, let $Y_1, Y_2, \ldots Y_n$ be a sample generated i.i.d from the uniform distribution on $\mathscr{X}$ with respect to μ (generated independently of $X_1, \ldots, X_n$). Let Z_i be X_i or Y_i with probability (1/2, 1/2) according to the outcome of Bernoulli(1/2) random variables V_i generated independently for $i = 1, \ldots, n$. Then Z_i has density $g(x) = (f(x) + 1)/2$. Clearly the new density g is bounded below (away from 0), whereas the family of the original densities need not be. Let $\tilde{\mathscr{F}} = \{g: g = (f + 1)/2, f \in \mathscr{F}\}$ be the new density class.

LEMMA 1. *The minimax L_2 risks of the two classes $\mathscr{F}$ and $\tilde{\mathscr{F}}$ have the following relationship*:

$$\min_{\hat{f}} \max_{f \in \mathscr{F}} E_{X^n} \|f - \hat{f}\|_2^2 \le 4 \min_{\hat{g}} \max_{g \in \tilde{\mathscr{F}}} E_{Z^n} \|g - \hat{g}\|_2^2,$$

where the minimization on the left-hand side is over all estimators based on $X_1, \ldots, X_n$ and the minimization on the right-hand side is over all estimators based on n independent observations from g. Generally, for $q \ge 1$ and $q \ne 2$, we have

$$\min_{\hat{f}} \max_{f \in \mathscr{F}} E_{X^n} \|f - \hat{f}\|_q^q \le 4^q \min_{\hat{g}} \max_{g \in \tilde{\mathscr{F}}} E_{Z^n} \|g - \hat{g}\|_q^q.$$

PROOF. We change the estimation of f to another estimation problem and show that the minimax risk of the original problem is upper bounded by the minimax risk of the new class. From any estimator in the new class (e.g., a minimax estimator), an estimator in the original problem is determined for which the risk is not greater than a multiple of the risk in the new class.

Let $\tilde{g}$ be any density estimator of g based on Z_i, $i = 1, \ldots n$. Fix $q \ge 1$. Let $\hat{g}$ be the density that minimizes $\|h - \tilde{g}\|_q$ over functions in the set $\mathscr{H} = \{h: h(x) \ge 1/2, \int h(x)\, d\mu = 1\}$. (If the minimizer does not exist, then one can choose $\hat{g}_\varepsilon$ within ε of the infimum and proceed similarly as follows and finally obtain the same general upper bound in Lemma 1 by letting $\varepsilon \to 0$.) Then

by the triangle inequality and because $g \in \mathscr{H}$, $\|g - \hat{g}\|_q^q \le 2^{q-1}\|g - \tilde{g}\|_q^q + 2^{q-1}\|\hat{g} - \tilde{g}\|_q^q \le 2^q\|g - \tilde{g}\|_q^q$. For $q = 2$, this can be improved to $\|g - \hat{g}\|_2 \le \|g - \tilde{g}\|_2$ by Hilbert space convex analysis [see, e.g., Lemma 9 in Yang and Barron (1997)]. We now focus on the proof of the assertion for the L_2 case. The proof for general L_q is similar. We construct a density estimator for f. Note that $f(x) = 2g(x) - 1$, let $\hat{f}_{\text{rand}}(x) = 2\hat{g}(x) - 1$. Then $\hat{f}_{\text{rand}}(x)$ is a nonnegative and normalized probability density estimator and depends on $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ and the outcomes of the coin flips $V_1, \ldots, V_n$. So it is a randomized estimator. The squared L_2 loss of $\hat{f}_{\text{rand}}$ is bounded as follows:

$$\int \left(f(x) - \hat{f}_{\text{rand}}(x)\right)^2 d\mu = \int (2g(x) - 2\hat{g}(x))^2 \, d\mu \le 4\|g - \tilde{g}\|_2^2.$$

To avoid randomization, we may replace $\hat{f}_{\text{rand}}(x)$ with its expected value over $Y_1, \ldots, Y_n$ and coin flips $V_1, \ldots, V_n$ to get $\hat{f}(x)$ with

$$\begin{aligned} E_{X^n}\|f - \hat{f}\|_2^2 &= E_{X^n}\|f - E_{Y^n, V^n}\hat{f}_{\text{rand}}\|_2^2 \\ &\le E_{X^n}E_{Y^n, V^n}\|f - \hat{f}_{\text{rand}}\|_2^2 \\ &= E_{Z^n}\|f - \hat{f}_{\text{rand}}\|_2^2 \\ &\le 4E_{Z^n}\|g - \tilde{g}\|_2^2, \end{aligned}$$

where the first inequality is by convexity and the second identity is because $\hat{f}_{\text{rand}}$ depends on X^n, Y^n, V^n only through Z^n. Thus $\max_{f \in \mathscr{F}} E_{X^n}\|f - \hat{f}\|_2^2 \le 4 \max_{g \in \tilde{\mathscr{F}}} E_{Z^n}\|g - \tilde{g}\|_2^2$. Taking the minimum over estimators $\tilde{g}$ completes the proof of Lemma 1. □

Thus the minimax risk of the original problem is upper bounded by the minimax risk on $\tilde{\mathscr{F}}$. Moreover, the ε-entropies are related. Indeed, since $\|(f_1 + 1)/2 - (f_2 + 1)/2\|_2 = (1/2)\|f_1 - f_2\|_2$, for the new class $\tilde{\mathscr{F}}$, the ε-packing entropy under L_2 is $\widetilde{M}_2(\varepsilon) = M_2(2\varepsilon)$.

Now we give upper and lower bounds on the minimax L_2 risk. Let us first get an upper bound. For the new class, the square root K-L divergence is upper bounded by multiples of L_2 distance. Indeed, for densities $g_1, g_2 \in \tilde{\mathscr{F}}$,

$$D(g_1\|g_2) \le \int \frac{(g_1 - g_2)^2}{g_2} d\mu \le 2\int (g_1 - g_2)^2 \, d\mu,$$

where the first inequality is the familiar relationship between K-L divergence and chi-square distance, and the second inequality follows because g_2 is lower bounded by 1/2. Let $\tilde{V}_K(\varepsilon)$ denote the d_K covering entropy of $\tilde{\mathscr{F}}$. Then $\tilde{V}_K(\varepsilon) \le \widetilde{M}_2(\varepsilon/\sqrt{2}) = M_2(\sqrt{2}\varepsilon)$. Let ε_n be chosen such that $M_2(\sqrt{2}\varepsilon_n) = n\varepsilon_n^2$. From Theorem 2, there exists a density estimator $\hat{g}_0$ such that $\max_{g \in \tilde{\mathscr{F}}} E_{Z^n} D(g\|\hat{g}_0) \le 2\varepsilon_n^2$. It follows that $\max_{g \in \tilde{\mathscr{F}}} E_{Z^n} d_H^2(g, \hat{g}_0) \le 2\varepsilon_n^2$ and $\max_{g \in \tilde{\mathscr{F}}} E_{Z^n}\|g - \hat{g}_0\|_1^2 \le 8\varepsilon_n^2$. Consequently, by Lemma 1, $\min_{\hat{f}} \max_{f \in \mathscr{F}} E_{X^n}\|f - \hat{f}\|_1 \le 8\sqrt{8}\varepsilon_n$. To get a good estimator in terms of L_2 risk, we assume $\sup_{f \in \mathscr{F}}\|f\|_\infty \le L < \infty$. Let $\hat{g}$ be

the density in $\tilde{\mathscr{F}}$ that is closest to $\hat{g}_0$ in Hellinger distance. Then by triangle inequality,

$$\max_{g\in\tilde{\mathscr{F}}} E_{Z^n} d_H^2(g,\hat{g}) \le 2\max_{g\in\tilde{\mathscr{F}}} E_{Z^n} d_H^2(g,\hat{g}_0) + 2\max_{g\in\tilde{\mathscr{F}}} E_{Z^n} d_H^2(\hat{g},\hat{g}_0)$$
$$\le 4\max_{g\in\tilde{\mathscr{F}}} E_{Z^n} d_H^2(g,\hat{g}_0) \le 8\varepsilon_n^2.$$

Now because both $\|g\|_\infty$ and $\|\hat{g}\|_\infty$ are bounded by $(L+1)/2$,

$$\int (g-\hat{g})^2\,d\mu = \int\left(\sqrt{g}-\sqrt{\hat{g}}\right)^2\left(\sqrt{g}+\sqrt{\hat{g}}\right)^2 d\mu \le 2(L+1)d_H^2(g,\hat{g}).$$

Thus $\max_{g\in\tilde{\mathscr{F}}} E_{Z^n}\|g-\hat{g}\|_2^2 \le 16(L+1)\varepsilon_n^2$. Using Lemma 1 again, we have an upper bound on the minimax squared L_2 risk. The action space $\overline{S}$ consists of all probability densities.

THEOREM 3. *Let $M_2(\varepsilon)$ be the L_2 packing entropy of a density class $\mathscr{F}$ with respect to a probability measure. Let ε_n satisfy $M_2(\sqrt{2}\varepsilon_n) = n\varepsilon_n^2$. Then*

$$\min_{\hat{f}}\max_{f\in\mathscr{F}} E_{X^n}\|f-\hat{f}\|_1 \le 8\sqrt{8}\varepsilon_n.$$

If in addition, $\sup_{f\in\mathscr{F}}\|f\|_\infty \le L < \infty$, *then*

$$\min_{\hat{f}}\max_{f\in\mathscr{F}} E_f\|f-\hat{f}\|_2^2 \le 256(L+1)\varepsilon_n^2.$$

The above result upper bounds the minimax L_1 risk and L_2 risk (under $\sup_{f\in\mathscr{F}}\|f\|_\infty < \infty$ for L_2) using only the L_2 metric entropy.

Using the relationship between L_q norms, namely, $\|f-\hat{f}\|_q \le \|f-\hat{f}\|_2$ for $1 \le q < 2$, under $\sup_{f\in\mathscr{F}}\|f\|_\infty < \infty$, we have

$$\min_{\hat{f}}\max_{f\in\mathscr{F}} E_{X^n}\|f-\hat{f}\|_q^2 \asymp \varepsilon_n^2 \quad \text{for } 1 \le q \le 2.$$

To get a minimax lower bound, we use the following assumption, which is satisfied by many classical classes such as Besov, Lipschitz, the class of monotone densities and more.

CONDITION 3. There exists at least one density $f^* \in \mathscr{F}$ with $\min_{x\in\mathscr{X}} f^*(x) = \underline{C} > 0$ and a positive constant $\alpha \in (0,1)$ such that $\mathscr{F}_0 = \{(1-\alpha)f^* + \alpha g : g \in \mathscr{F}\} \subset \mathscr{F}$.

For a convex class of densities, Condition 3 is satisfied if there is at least one density bounded away from zero. Under Condition 3, the subclass $\mathscr{F}_0$ has L_2 packing entropy $M_2^0(\varepsilon) = M_2(\varepsilon/\alpha)$ and for two densities f_1 and f_2 in $\mathscr{F}_0$,

$$D(f_1\|f_2) \le \int \frac{(f_1-f_2)^2}{f_2}\,d\mu \le \frac{1}{(1-\alpha)\underline{C}}\int (f_1-f_2)^2\,d\mu.$$

Thus applying Theorem 1 on $\mathscr{F}_0$, and then applying Theorem 2, we have the following conclusion.

THEOREM 4. *Suppose Condition* 3 *is satisfied. Let* $M_2(\varepsilon)$ *be the* L_2 *packing entropy of a density class* $\mathscr{F}$ *with respect to a probability measure, let* $\bar{\varepsilon}_n$ *satisfy* $M_2(\sqrt{(1-\alpha)}\underline{C}\bar{\varepsilon}_n/\alpha) = n\bar{\varepsilon}_n^2$ *and* $\underline{\varepsilon}_n$ *be chosen such that* $M_2(\underline{\varepsilon}_n/\alpha) = 4n\bar{\varepsilon}_n^2 + 2\log 2$. *Then*

$$\min_{\hat{f}} \max_{f\in\mathscr{F}} E_{X^n}\|f - \hat{f}\|_2^2 \geq \underline{\varepsilon}_n^2/8.$$

Moreover, if the class $\mathscr{F}$ *is rich using the* L_2 *distance* (*Condition* 2), *then with* ε_n *determined by* $M_2(\varepsilon_n) = n\varepsilon_n^2$:

(i) *If* $M_2(\varepsilon) \asymp M_1(\varepsilon)$ *holds, then* $\min_{\hat{f}} \max_{f\in\mathscr{F}} E_{X^n}\|f - \hat{f}\|_1 \asymp \varepsilon_n$.

(ii) *If* $\sup_{f\in\mathscr{F}}\|f\|_\infty < \infty$, *then* $\min_{\hat{f}} \max_{f\in\mathscr{F}} E_{X^n}\|f - \hat{f}\|_2^2 \asymp \varepsilon_n^2$.

Using the relationship between L_2 and $L_q(1 \leq q < 2)$ distances and applying Theorem 1, we have the following corollary.

COROLLARY 2. *Suppose* $\mathscr{F}$ *is rich using both the* L_2 *distance and the* L_q *distance for some* $q \in [1, 2)$. *Assume Condition* 3 *is satisfied. Let* $\underline{\varepsilon}_{n,q}$ *satisfy* $M_q(\underline{\varepsilon}_{n,q}) = n\varepsilon_n^2$. *Then*

$$\underline{\varepsilon}_{n,q}^2 \preceq \min_{\hat{f}} \max_{f\in\mathscr{F}} E_{X^n}\|f - \hat{f}\|_q^2 \preceq \varepsilon_n^2.$$

If the packing entropies under L_2 and L_q are equivalent (which is the case for many familiar nonparametric classes; see Section 6 for examples), then the above upper and lower bounds converge at the same rate. Generally for a uniformly upper bounded density class $\mathscr{F}$ on a compact set, because $\int(f-g)^2\,d\mu \leq (\|f+g\|_\infty)\int|f-g|\,d\mu$, we know $M_1(\varepsilon) \leq M_2(\varepsilon) \leq M_1(\varepsilon^2/\sup_{\mathscr{F}}\|f\|_\infty)$. Then the corresponding lower bound for L_1 risk may vary from ε_n to ε_n^2 depending on how different the two entropies are [see also Birgé (1986)].

2.3. *Minimax rates under K-L and Hellinger loss.* For the square root K-L divergence, Condition 0 is not necessarily satisfied for general classes of densities. When it is not satisfied (for instance, if the densities in the class have different supports), the following lemma is helpful to lower bound as well as upper bound the K-L risk involving only the Hellinger metric entropy by relating it to the covering entropy under d_K.

We consider estimating a density defined on $\mathscr{X}$ with respect to a measure μ with $\mu(\mathscr{X}) = 1$ for the rest of this section and Section 2.3.

LEMMA 2. *For each probability density* g, *positive* T *and* $0 < \varepsilon \leq \sqrt{2}$, *there exists a probability density* $\tilde{g}$ *such that for every density* f *with* $\|f\|_\infty \leq T$ *and* $d_H(f, g) \leq \varepsilon$, *the K-L divergence* $D(f\|\tilde{g})$ *satisfies*

$$D(f\|\tilde{g}) \leq 2\Big(2 + \log\Big(9T/4\varepsilon^2\Big)\Big)\Big(9 + 8(8T-1)^2\Big)\varepsilon^2.$$

Bounds analogous to Lemma 2 are in Barron, Birgé and Massart [(1999), Proposition 1], Wong and Shen [(1995), Theorem 5]. A proof of this lemma is given in the Appendix.

We now show how Lemma 2 can be used to give rates of convergence under Hellinger and K-L loss without forcing the density to be bounded away from zero. Let $\mathscr{F}$ be a density class with $\|f\|_\infty \le T$ for each $f \in \mathscr{F}$. Assume, for simplicity, that the metric entropy $M_H(\varepsilon)$ under d_H is of order $\varepsilon^{-1/\alpha}$ for some $\alpha > 0$. Then by replacing each g in a Hellinger covering with the associated $\tilde{g}$, Lemma 2 implies that we obtain a d_K covering with $V_K(\varepsilon) \le \left((\log(1/\varepsilon))^{1/2}/\varepsilon\right)^{1/\alpha}$ and we obtain the associated upper bound on the K-L risk from Theorem 2. For the lower bound, use the fact that d_K is no smaller than the Hellinger distance for which we have the packing entropy. (Note that the d_K packing entropy, which may be infinity, may not be used here since Condition 0 is not satisfied.) Consequently, from Theorem 1 and Theorem 2, we have

$$(n \log n)^{-2\alpha/(2\alpha+1)} \preceq \min_{\hat{f}} \max_{f\in\mathscr{F}} E d_H^2(f, \hat{f}) \le \min_{\hat{f}} \max_{f\in\mathscr{F}} E D(f\|\hat{f}) \preceq n^{-2\alpha/(2\alpha+1)} (\log n)^{1/(2\alpha+1)}.$$

Here, the minimization is over *all* estimators, $\hat{f}$; that is, $\overline{S}$ is the class of all densities.

Thus even if the densities in $\mathscr{F}$ may be 0 or near 0, the minimax K-L risk is within a logarithmic factor of the squared Hellinger risk. See Barron, Birgé and Massart (1999) for related conclusions.

2.4. *Some more results for risks when densities may be 0.* In this section, we show that by modifying a nonparametric class of densities with uniformly bounded logarithms to allow the densities to approach zero at some points or even vanish in some subsets, the minimax rates of convergence under K-L and Hellinger (and L_2) may remain unchanged compared to that of the original class. Densities in the new class are obtained from the original nonparametric class by multiplying by members in a smaller class (often a parametric class) of functions that may be near zero. The result is applicable to the following example.

EXAMPLE. (*Densities with support on unknown set of k intervals.*) Let $\mathscr{F} = \{h(x) \cdot \sum_{i=0}^{k-1} b_{i+1} 1_{\{a_i \le x < a_{i+1}\}}/c : h \in \mathscr{H},\ 0 = a_0 < a_1 < a_2 < \cdots < a_k = 1,\ \sum_{i=0}^{k-1} b_{i+1}(a_{i+1} - a_i) \ge \gamma_1 \text{ and } 0 \le b_i \le \gamma_2,\ 1 \le i \le k\}$ (c is the normalizing constant). Here $\mathscr{H}$ is a class of functions with uniformly bounded logarithms, k is a positive integer, γ_1 and γ_2 are positive constants. The constants γ_1 and γ_2 force the densities in $\mathscr{F}$ to be uniformly upper bounded. These densities may be 0 on some intervals and arbitrarily close to 0 on others.

Note that if the densities in $\mathscr{H}$ are continuous, then the densities in $\mathscr{F}$ have at most $k - 1$ discontinuous points. For instance, if $\mathscr{H}$ is a Lipschitz class, then the functions in $\mathscr{F}$ are piecewise Lipschitz.

Somewhat more generally, let $\mathscr{G}$ be a class of nonnegative functions satisfying $\|g\|_\infty \le \overline{C}$ and $\int g\,d\mu \ge \underline{c}$ for all $g \in \mathscr{G}$. Though the g's are nonnegative, they may equal zero on some subsets. Suppose $0 < \underline{c} \le h \le \overline{C} < \infty$ for all $h \in \mathscr{H}$. Consider a class of densities with respect to μ:

$$\mathscr{F} = \left\{ f(x) = \frac{h(x)g(x)}{\int h(x)g(x)d\mu} \cdot\ h \in \mathscr{H},\ g \in \mathscr{G} \right\}.$$

Let $\tilde{\mathscr{G}} = \{\tilde{g} = g/\int g\,d\mu : g \in \mathscr{G}\}$ be the density class corresponding to $\mathscr{G}$ and similarly define $\tilde{\mathscr{H}}$. Let $M_2(\varepsilon;\mathscr{H})$ be the packing entropy of $\mathscr{H}$ under L_2 distance and let $V_K(\varepsilon;\mathscr{F})$ and $V_K(\varepsilon;\tilde{\mathscr{G}})$ be covering entropies as defined in Section 2 of $\mathscr{F}$ and $\tilde{\mathscr{G}}$, respectively, under d_K. The action space $\overline{S}$ consists of all probability densities.

THEOREM 5. *Suppose* $V_K(\varepsilon;\tilde{\mathscr{G}}) \le A_1 M_2(A_2\varepsilon;\mathscr{H})$ *for some positive constants* A_1, A_2 *and suppose the class* $\mathscr{H}$ *is rich in the sense that* $\liminf_{\varepsilon\to 0} M_2(\alpha\varepsilon;\mathscr{H})/M_2(\varepsilon;\mathscr{H}) > 1$ *for some* $0 < \alpha < 1$. *If* $\mathscr{G}$ *contains a constant function, then*

$$\min_{\hat f}\max_{f\in\mathscr{F}} E_f D(f\|\hat f) \asymp \min_{\hat f}\max_{f\in\mathscr{F}} E_f d_H^2(f,\hat f) \asymp \min_{\hat f}\max_{f\in\mathscr{F}} E_f\|f - \hat f\|_2^2 \asymp \varepsilon_n^2,$$

where ε_n *is determined by* $M_2(\varepsilon_n;\mathscr{H}) = n\varepsilon_n^2$.

The result basically says that if $\tilde{\mathscr{G}}$ is smaller than the class $\mathscr{H}$ in an ε-entropy sense, then for the new class, being 0 or close to 0 due to $\mathscr{G}$ does not hurt the K-L risk rate. To apply the result, we still need to bound the covering entropy of $\tilde{\mathscr{G}}$ under d_K. One approach is as follows. Suppose the Hellinger metric entropy of $\tilde{\mathscr{G}}$ satisfies $M_H(\varepsilon/\log(1/\varepsilon);\tilde{\mathscr{G}}) \le M_2(A\varepsilon;\mathscr{H})$ for some constant $A > 0$. From Lemma 2, $V_K(\varepsilon;\tilde{\mathscr{G}}) \le M_H(A'\varepsilon/\log(1/\varepsilon);\tilde{\mathscr{G}})$ for some constant $A' > 0$ when ε is sufficiently small (note that the elements $\tilde{g}$ of the cover from Lemma 2 are not necessarily in $\mathscr{F}$ though they are in the set $\overline{S}$ of all probability densities). Then we have $V_K(\varepsilon;\tilde{\mathscr{G}}) \le M_2(AA'\varepsilon;\mathscr{H})$. This covers the example above (for which case, $\mathscr{G}$ is a parametric family) and it even allows $\mathscr{G}$ to be almost as large as $\mathscr{H}$. An example is $\mathscr{H} = \{h : \log h \in B^{\alpha}_{\sigma,q}(C)\}$ and $\mathscr{G} = \{g : g \in B^{\alpha'}_{\sigma',q'}(C),\ g \ge 0 \text{ and } \int g\,d\mu = 1\}$ with α' bigger than but arbitrarily close to α (for a definition of Besov classes, see Section 6). Note that here the density can be 0 on infinitely many subintervals. Other examples are given in Yang and Barron (1997).

PROOF. By Theorem 2, we have $\min_{\hat f}\max_{f\in\mathscr{F}} E_f D(f\|\hat f) \le \tilde{\varepsilon}_n^2$, where $\tilde{\varepsilon}_n$ satisfies $V_K(\tilde{\varepsilon}_n;\mathscr{F}) = n\tilde{\varepsilon}_n^2$. Under the entropy condition that $\tilde{\mathscr{G}}$ is smaller than $\mathscr{H}$, it can be shown that $V_K(\varepsilon;\mathscr{F}) \le \tilde{A}_1 M_2(\tilde{A}_2\varepsilon;\mathscr{H})$ for some constants $\tilde{A}_1$ and $\tilde{A}_2$ [see Yang and Barron (1997) for details]. Then under the assumption of the richness of $\mathscr{H}$, we have $\tilde{\varepsilon}_n = O(\varepsilon_n)$. Thus ε_n^2 upper bounds the minimax risk rates under both d_K^2 and d_H^2. Because $\mathscr{G}$ contains a constant function, $\tilde{\mathscr{H}}$ is a subset of $\mathscr{F}$. Since the log-densities in $\tilde{\mathscr{H}}$ are uniformly bounded, the

L_2 metric entropies of $\mathscr{H}$ and $\tilde{\mathscr{H}}$ are of the same order. Thus taking $S_0 = \tilde{\mathscr{H}}$, the lower bound rate under K-L or squared Hellinger or square L_2 distance is of order ε_n^2 by Corollary 1. Because the densities in $\mathscr{F}$ are uniformly upper bounded, the L_2 distance between two densities in $\mathscr{F}$ is upper bounded by a multiple of the Hellinger distance and d_K. Thus under the assumptions in the theorem, the L_2 metric entropy of $\mathscr{F}$ satisfies $M_2(\varepsilon; \mathscr{F}) \le V_K(A\varepsilon; \mathscr{F}) \le M_2(A'\varepsilon; \mathscr{H})$ for some positive constants A and A'. Consequently, the minimax L^2 risk is upper bounded by order ε_n^2 by Theorem 3. This completes the proof of Theorem 5. □

3. Applications in data compression and regression. Two cases when the K-L divergence plays a direct role include data compression, where total K-L divergence provides the redundancy of universal codes, and regression with Gaussian errors where the individual K-L divergence between two Gaussian models is the customary squared L_2 distance. A small amount of additional work is required in the regression case to convert a predictive density estimator into a regression estimator via a minimum Hellinger distance argument.

3.1. *Data compression.* Let $X_1, \ldots, X_n$ be an i.i.d. sample of discrete random variables from $p_\theta(x)$, $\theta \in S$. Let $q_n(x_1, \ldots x_n)$ be a density (probability mass) function. The redundancy of the Shannon code using density q_n is the difference of its expected codelength and the expected codelength of the Shannon code using the true density $p_\theta(x_1, \ldots, x_n)$, that is, $D(p_\theta^n \| q_n)$. Formally, we examine the minimax properties of the game with loss $D(p_\theta^n \| q_n)$ for continuous random variables also. In that case, $D(p_\theta^n \| q_n)$ corresponds to the redundancy in the limit of fine quantization of the random variable [see, e.g., Clarke and Barron (1990), pages 459 and 460].

The asymptotics of redundancy lower bounds have been considered by Rissanen (1984), Clarke and Barron (1990, 1994), Rissanen, Speed and Yu (1992) and others. These results were derived for smooth parametric families or a specific smooth nonparametric class. We here give general redundancy lower bounds for nonparametric classes. The key property revealed by the chain rule is the relationship between the minimax value of the game with loss $D(p_\theta^n \| q_n)$ and the minimax cumulative K-L risk,

$$\min_{q_n} \max_{\theta \in S} D(p_\theta^n \| q_n) = \min_{\{\hat{p}_i\}_{i=0}^{n-1}} \max_{\theta \in S} \sum_{i=0}^{n-1} E_\theta D(p_\theta \| \hat{p}_i),$$

where the minimization on the left is over all joint densities $q_n(x_1, \ldots, x_n)$ and the minimization on the right is over all sequences of estimators $\hat{p}_i$ based on samples of size $i = 0, 1, \ldots, n-1$ (for $i = 0$, it is any fixed density). Indeed, one has $D(p_\theta^n \| q_n) = \sum_{i=0}^{n-1} E_\theta D(p_\theta \| \hat{p}_i)$ when $q_n(x_1, \ldots, x_n) = \prod_{i=0}^{n-1} \hat{p}_i(x_{i+1})$ [cf. Barron and Hengartner (1998)]. Let $R_n = \min_{q_n \in \Omega_n} \max_{\theta \in S} D(p_\theta^n \| q_n)$ be the minimax total K-L divergence and let $r_n = \min_{\hat{p}_n \in \mathscr{P}_n} \max_{\theta \in S} E_\theta D(p_\theta \| \hat{p})$

be the individual minimax K-L risk. Then

$$nr_{n-1} \le R_n \le \sum_{i=0}^{n-1} r_i. \tag{5}$$

Here the first inequality $r_{n-1} \le n^{-1}R_n$ follows from noting that as in (4), for any joint density q_n on the sample space of $X_1, \ldots, X_n$, there exists an estimator $\hat{p}$ such that $E_\theta D(p_\theta \| \hat{p}) \le n^{-1} D(p_\theta^n \| q_n)$ for all $\theta \in S$. The second inequality $R_n \le \sum_{i=0}^{n-1} r_i$ is from the fact that the maximum (over θ in S) of the sum of risks is not greater than the sum of the maxima.

In particular, we see that $r_n \asymp R_n/n$ when $r_n \asymp n^{-1} \sum_{i=0}^{n-1} r_i$. Effectively, this means that the minimax individual K-L risk and $1/n$ times minimax total K-L divergence match asymptotically when r_n converges at a rate sufficiently slower than $1/n$ (e.g., n^{-p} with $0 < \rho < 1$).

Now let $d(\theta, \theta')$ be a metric on $\overline{S}$ and assume that $\{p_\theta : \theta \in \overline{S}\}$ contains all probability densities on $\mathscr{X}$. Let $M_d(\varepsilon)$ be the packing entropy of S under d and let $V(\varepsilon)$ be an upper bound on the covering entropy $V_K(\varepsilon)$ of S under d_K. Choose ε_n such that $\varepsilon_n^2 = V(\varepsilon_n)/n$ and choose $\varepsilon_{n,d}$ such that $M_d(\varepsilon_{n,d}) = 4n\varepsilon_n^2 + 2\log 2$. Based on Theorem 1, (2) and inequality (5), we have the following result.

COROLLARY 3. *Assume that* $D(p_\theta \| p_{\theta'}) \ge A_0 d^2(\theta, \theta')$ *for all* $\theta, \theta' \in \overline{S}$. *Then we have*

$$(A_0/8) n\varepsilon_{n,d}^2 \le \min_{q_n} \max_{\theta \in S} D(p_\theta^n \| q_n) \le 2n\varepsilon_n^2,$$

where the minimization is over all densities on $\mathscr{X}^n$.

Two choices that satisfy the requirements are the Hellinger distance and the L_1 distance.

When interest is focused on the cumulative K-L risk (or on the individual risk r_n in the case that $r_n \asymp n^{-1} \sum_{i=0}^{n-1} r_i$), direct proof of suitable bounds are possible without the use of Fano's inequality. See Haussler and Opper (1997) for new results in that direction. Another proof using an idea of Rissanen (1984) is in Yang and Barron (1997).

3.2. *Application in nonparametric regression.* Consider the regression model

$$y_i = u(x_i) + \epsilon_i, \qquad i = 1, \ldots n.$$

Suppose the errors ϵ_i, $1 \le i \le n$, are i.i.d. with the $Normal(0, \sigma^2)$ distribution. The explanatory variables x_i, $1 \le i \le n$, are i.i.d. with a fixed distribution P. The regression function u is assumed to be in a function class $\mathscr{U}$. For this case, the square root K-L divergence between the joint densities of (X, Y) in the family is a metric. Let $\|u - v\|_2 = (\int (u(x) - v(x))^2 \, dP)^{1/2}$ be the $L_2(P)$ distance with respect to the measure induced by X. Let $M_2(\varepsilon)$ and similarly for $q \ge 1$ let $M_q(\varepsilon)$ be the ε-packing entropy under $L_2(P)$ and $L_q(P)$,

respectively. Assume $M_2(\varepsilon)$ and $M_q(\varepsilon)$ are both rich (in accordance with Condition 2). Choose ε_n such that $M_2(\varepsilon_n) = n\varepsilon_n^2$. Similarly let $\underline{\varepsilon}_{n,q}$ be determined by $M_q(\underline{\varepsilon}_{n,q}) = n\varepsilon_n^2$.

THEOREM 6. *Assume* $\sup_{u\in\mathscr{U}} \|u\|_\infty \le L$. *Then*

$$\min_{\hat{u}} \max_{u\in\mathscr{U}} E\|u-\hat{u}\|_2^2 \asymp \varepsilon_n^2.$$

For the minimax $L_q(P)$ *risk, we have*

$$\min_{\hat{u}} \max_{u\in\mathscr{U}} E\|u-\hat{u}\|_q \succeq \underline{\varepsilon}_{n,q}.$$

If further, $M_2(\varepsilon) \asymp M_q(\varepsilon)$ *for* $1 \le q < 2$, *then for the* $L_q(P)$ *risk, we have*

$$\min_{\hat{u}} \max_{u\in\mathscr{U}} E\|u-\hat{u}\|_q \asymp \varepsilon_n.$$

PROOF. With no loss of generality, suppose that P_X has a density $h(x)$ with respect to a measure λ. Let $p_u(x,y) = (2\pi\sigma^2)^{-1/2}\exp(-(y-u(x))^2/2\sigma^2)h(x)$ denote the joint density of (X,Y) with regression function u. Then $D(p_u\|p_v) = (1/2\sigma^2)E((Y-u(X))^2 - (Y-v(X))^2)$ reduces as is well known to $(1/2\sigma^2)\int(u(x)-v(x))^2h(x)\,d\lambda$, so that d_K is equivalent to the $L_2(P)$ distance. The lower rates then follow from Theorem 1 together with the richness assumption on the entropies.

We next determine upper bounds for regression by specializing the bound from Theorem 2. We assume $\|u\|_\infty \le L$ uniformly for $u \in \mathscr{U}$. Theorem 2 provides a density estimator $\hat{p}_n$ such that $\max_{u\in\mathscr{U}} ED(p_u\|\hat{p}_n) \le 2\varepsilon_n^2$. It follows that $\max_{u\in\mathscr{U}} Ed_H^2(p_u,\hat{p}_n) \le 2\varepsilon_n^2$. [A similar conclusion is available in Birgé (1986), Theorem 3.1.] Here we take advantage of the fact that when the density $h(x)$ is fixed, $\hat{p}_n(x,y)$ takes the form of $h(x)\hat{g}(y|x)$, where $\hat{g}(y|x)$ is an estimate of the conditional density of y given x (it happens to be a mixture of Gaussians using a posterior based on a uniform prior on ε-nets). For given x and $(X_i,Y_i)_{i=1}^n$, let $\tilde{u}_n(x)$ be the minimizer of the Hellinger distance $d_H(\hat{g}_n(\cdot|x),\phi_z)$ between $\hat{g}_n(y|x)$ and the normal $\phi_z(y)$ density with mean z and the given variance over choices of z with $|z| \le L$. Then $\tilde{u}_n(x)$ is an estimator of $u(x)$ based on $(X_i,Y_i)_{i=1}^n$. By the triangle inequality, given x and $(X_i,Y_i)_{i=1}^n$,

$$\begin{aligned} d_H\big(\phi_{u(x)},\phi_{\tilde{u}_n(x)}\big) &\le d_H\big(\phi_{u(x)},\hat{g}_n(\cdot|x)\big) + d_H\big(\phi_{\tilde{u}_n(x)},\hat{g}_n(\cdot|x)\big) \\ &\le 2d_H\big(\phi_{u(x)},\hat{g}_n(\cdot|x)\big). \end{aligned}$$

It follows that $\max_{u\in\mathscr{U}} Ed_H^2(p_u,p_{\tilde{u}_n}) \le 4\max_{u\in\mathscr{U}} Ed_H^2(p_u,\hat{p}_n) \le 8\varepsilon_n^2$. Now $Ed_H^2(p_u,p_{\tilde{u}_n}) = 2E\int h(x)\big(1-\exp(-(u(x)-\tilde{u}_n(x))^2/8\sigma^2)\big)\,d\lambda$. The concave function $(1-e^{-v})$ is above the chord $(v/B)(1-e^{-B})$ for $0 \le v \le B$. Thus using $v = (u(x)-\tilde{u}_n(x))^2/8\sigma^2$ and $B = L^2/2\sigma^2$, we obtain

$$\begin{aligned} &\max_{u\in\mathscr{U}} E\int (u(x)-\tilde{u}_n(x))^2h(x)\,d\lambda \\ &\qquad \le 2L^2(1-\exp(-L^2/2\sigma^2))^{-1}\max_{u\in\mathscr{U}} Ed_H^2(p_u,p_{\tilde{u}}) \asymp \varepsilon_n^2. \end{aligned}$$

This completes the proof of Theorem 6. □

4. Linear approximation and minimax rates. In this section, we apply the main results and known metric entropy to give a general conclusion on the relationship between linear approximation and minimax rates of convergence. A more general and detailed treatment is in Yang and Barron (1997).

Let $\Phi = \{\phi_1 = 1, \phi_2 \ldots, \phi_k, \ldots\}$ be a fundamental sequence in $L^2[0,1]^d$ (that is, linear combinations are dense in $L^2[0,1]^d$). Let $\Gamma = \{\gamma_0, \ldots, \gamma_k, \ldots\}$ be a decreasing sequence of positive numbers for which there exist $0 < c' < c < 1$ such that

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

as is true for $\gamma_k \sim k^{-\alpha}$ and also for $\gamma_k \sim k^{-\alpha}(\log k)^{\beta}$, $\alpha > 0, \beta \in R$. Let $\eta_0(f) = \|f\|_2$ and $\eta_k(f) = \min_{\{a_i\}} \|f - \sum_{i=1}^{k} a_i\phi_i\|_2$ for $k \ge 1$ be the kth degree of approximation of $f \in L^2[0,1]^d$ by the system Φ. Let $\mathcal{F}(\Gamma, \Phi)$ be all functions in $L_2[0,1]^d$ with the approximation errors bounded by Γ; that is,

$$\mathcal{F}(\Gamma, \Phi) = \{f \in L_2[0,1]^d : \eta_k(f) \le \gamma_k,\ k = 0, 1, \ldots\}.$$

They are called the full approximation sets. Lorentz (1966) gives metric entropy bounds on these classes (actually in more generality than stated here) and the bounds are used to derive metric entropy orders for a variety of function classes including Sobolev classes.

Suppose the functions in $\mathcal{F}(\Gamma, \Phi)$ are uniformly bounded, that is, $\sup_{g \in \mathcal{F}(\Gamma,\Phi)} \|g\|_\infty \le \rho$ for some positive constant ρ. Let $\tilde{\mathcal{F}}(\Gamma, \Phi)$ be all the probability density functions in $\mathcal{F}(\Gamma, \Phi)$. When γ_0 is large enough, the L_2 metric entropies of $\tilde{\mathcal{F}}(\Gamma, \Phi)$ and $\mathcal{F}(\Gamma, \Phi)$ are of the same order. Let k_n be chosen such that $\gamma_k^2 \asymp k/n$.

THEOREM 7. *The minimax rate of convergence for a full approximation set of functions is determined simply as follows*:

$$\min_{\hat{f}} \max_{f \in \tilde{\mathcal{F}}(\Gamma,\Phi)} E\|f - \hat{f}\|_2^2 \asymp k_n/n.$$

A similar result holds for regression. Note that there is no special requirement on the bases Φ (they may even not be continuous). For such a case, it seems hard to find a local packing set for the purpose of directly applying the lower bounding results of Birgé (1983).

The conclusion of the theorem follows from Theorem 4. Basically, from Lorentz, the metric entropy of $\mathcal{F}(\Gamma, \Phi)$ is order $k_\varepsilon = \inf\{k : \gamma_k \le \varepsilon\}$ and $\mathcal{F}(\Gamma, \Phi)$ is rich in L_2 distance. As a consequence, ε_n determined by $M(\varepsilon_n) \asymp n\varepsilon_n^2$ also balances γ_k^2 and k/n.

As an illustration, for a system Φ, consider the functions that can be approximated by the linear system with polynomially decreasing approximation error $\gamma_k \sim k^{-\alpha}$, $\alpha > 0$. Then $\min_{\hat{f}} \max_{f \in \tilde{\mathcal{F}}(\Gamma,\Phi)} E\|f - \hat{f}\|_2^2 \asymp n^{-2\alpha/(1+2\alpha)}$. Similarly we have rate $n^{-2\alpha/(1+2\alpha)}(\log n)^{-2\beta/(1+2\alpha)}$ if $\gamma_k \sim k^{-\alpha}(\log k)^{\beta}$.

As we have seen, the optimal convergence rate in this full approximation setting is of the same order as $\min_k(\gamma_k^2 + k/n)$, which we recognize as the familiar bias-squared plus variance trade-off for mean squared error. Indeed, for regression as in Section 3, with $y_i = u(x_i) + \epsilon_i$, $u \in \mathscr{F}(\Gamma, \Phi)$, approximation systems yield natural and well-known estimates that achieve this rate. This trade-off is familiar in the literature [see, e.g., Cox (1988) for least squares regression estimates, Cenov (1982) or Barron and Sheu (1991), for maximum likelihood log-density estimates, and Birgé and Massart (1996) for projective density estimators and other contrasts].

The best rate k_n is of course, unknown in applications, suggesting 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. For recent results on model selection, see for instance, Barron, Birgé, and Massart (1999) and Yang and Barron (1998). There knowledge of the optimal convergence rates for various situations is still of interest, because it permits one to gauge the extent to which an automatic procedure adapts to multiple function classes.

To sum up this section, if a function class $\mathscr{F}$ is contained in $\mathscr{F}(\Gamma, \Phi)$ and contains $\mathscr{F}(\Gamma', \Phi')$ for some pair of fundamental sequences Φ and Φ' for which the γ_k, γ'_k sequences yield $\varepsilon_n \asymp \varepsilon'_n$, then ε_n provides the minimax rate for $\mathscr{F}$ and moreover (under the conditions discussed above) minimax optimal estimates are available from suitable linear estimates. However, some interesting function classes do not permit linear estimators to be minimax rate optimal [see Nemirovskii (1985), Nemirovskii, Polyak and Tsybakov (1985), Donoho, Johnstone, Kerkyacharian and Picard (1996)]. Lack of a full approximation set characterization does not preclude determination of the metric entropy by other approximation-theoretic means in specific cases as will be seen in the next two sections.

5. Sparse approximations and minimax rates. In the previous section, full approximation sets of functions are defined through linear approximation with respect to a given system Φ. There, to get a given accuracy of approximation δ, one uses the first k_δ basis functions with $k_\delta = \min\{i: \gamma_i \le \delta\}$. This choice works for all $g \in \mathscr{F}(\Gamma, \Phi)$ and these basis functions are needed to get the accuracy δ for some $g \in \mathscr{F}(\Gamma, \Phi)$. For slowly converging sequences γ_k, very large k_δ is needed to get accuracy δ. This phenomenon occurs especially in high-dimensional function approximation. For instance, if one uses full approximation with any chosen basis for a d-dimensional Sobolev class with all $\alpha\,(\alpha \ge 1)$ partial derivatives well behaved, the approximation error with k terms can not converge faster than $k^{-\alpha/d}$. It then becomes of interest to examine approximation using manageable size subsets of terms sparse in comparison to the total that would be needed with full approximation. We next give minimax results for some sparse function classes.

Let Φ and Γ be as in the previous section. Let $I_k > k$, $k \ge 1$ be a given nondecreasing sequence of integers satisfying $\liminf I_k/k = \infty\,(I_0 = 0)$ and let $\mathscr{I} = \{I_1, I_2, \ldots\}$. Let $\tilde{\eta}_k(g) = \min_{l_1 \le I_1, \ldots, l_k \le I_k} \min_{\{a_i\}} \|g - \sum_{i=1}^k a_i \phi_{l_i}\|_2$ be called the kth degree of sparse approximation of $g \in L^2[0,1]^d$ by the system

Φ. Here for $k = 0$, there is no approximation and $\tilde{\eta}_0(g) = \|g\|_2$. The kth term used to approximate g is selected from I_k basis functions. Let $\mathcal{S}(\Gamma, \Phi) = \mathcal{S}(\Gamma, \Phi, \mathcal{I})$ be all functions in $L_2[0, 1]^d$ with the sparse approximation errors bounded by Γ, that is,

$$\mathcal{S}(\Gamma, \Phi) = \{g \in L_2[0, 1]^d : \tilde{\eta}_k(g) \le \gamma_k, \qquad k = 0, 1, \ldots\}.$$

We call it a sparse approximation set of functions (for a fixed choice of $\mathcal{I}$). Larger I_k's provide considerable more freedom of approximation.

In terms of metric entropy, a sparse approximation set is not much larger than the corresponding full approximation set $\mathcal{F}(\Gamma, \Phi)$ it contains. Indeed, as will be shown, its metric entropy is larger by at most a logarithmic factor under the condition $I_k \le k^\tau$ for some possibly large $\tau > 1$. If full approximation is used instead to approximate a sparse approximation set, one is actually approximating a class with a much larger metric entropy than $\mathcal{S}(\Gamma, \Phi)$ if I_k is much bigger than k [see Yang and Barron (1997)].

For a class of examples, let Ψ be a class of functions uniformly bounded by v with a L_2 ε-cover of cardinality of order $(1/\varepsilon)^{d'}$ [metric entropy $d' \log(1/\varepsilon)$] for some positive constant d'. Let $\mathcal{C}_\Psi$ be the closure of its convex hull. For $k \ge 1$, let $\varphi_{I_{k-1}+1}, \ldots, \varphi_{I_k}$ in Φ be the members of a $v/(2k^{1/2})$—cover of Ψ. Here $I_k - I_{k-1}$ is of order $k^{d'/2}$. Then $\mathcal{C}_\Psi \subset \mathcal{S}(\Gamma, \Phi)$ with $\gamma_k = 4v/k^{1/2}$. That is, the closure of the convex hull of the class can be uniformly sparsely approximated by a system consisting of suitably chosen members in the class at rate $k^{-1/2}$ with k sparse terms out of about $k^{d'/2}$ many candidates. This containment result, $\mathcal{C}_\Psi \subset \mathcal{S}(\Gamma, \Phi)$, can be verified using greedy approximation [see Jones (1992) or Barron (1993), Section 8], which we omit here. A specific example of Ψ is $\{\sigma(ax + b) : a \in [-1, 1]^d,\ b \in R\}$, where σ is a fixed sigmoidal function satisfying a Lipschitz condition such as $\sigma(z) = (e^z - 1)/(e^z + 1)$, or a sinusoidal function $\sigma(z) = \sin(z)$. For metric entropy bounds on $\mathcal{C}_\Psi$, see Dudley (1987), with refinements in Ball and Pajor (1990).

Now let us prove metric entropy bounds on $\mathcal{S}(\Gamma, \Phi)$. Because $\mathcal{F}(\Gamma, \Phi) \subset \mathcal{S}(\Gamma, \Phi)$, the previous lower bound for $\mathcal{F}(\Gamma, \Phi)$ is a lower bound for $\mathcal{S}(\Gamma, \Phi)$. We next derive an upper bound. Let $l_i \le I_i$, $1 \le i \le k_\varepsilon$ be fixed for a moment. Consider the subset $\mathcal{G}_{l_1, \ldots, l_{k_\varepsilon}}$ of $\mathcal{S}(\Gamma, \Phi)$ in the span of $\phi_{l_1}, \ldots, \phi_{l_{k_\varepsilon}}$ that has approximation errors bounded by $\gamma_1, \ldots, \gamma_{k_\varepsilon}$ using the basis $\phi_{l_1}, \ldots, \phi_{l_{k_\varepsilon}}$ (i.e., $g \in \mathcal{G}_{l_1, \ldots, l_{k_\varepsilon}}$ if and only if $g = \sum_{i=1}^{k_\varepsilon} a_i^* \phi_{l_i}$ for some coefficients a_i^* and $\min_{a_1, \ldots, a_m} \|g - \sum_{i=1}^m a_i \phi_{l_i}\|_2 \le \gamma_m$ for $1 \le m \le k_\varepsilon$). From the previous section, we know the ε-entropy of $\mathcal{G}_{l_1, \ldots, l_{k_\varepsilon}}$ is upper bounded by order $k_\varepsilon = \inf\{k : \gamma_k \le \varepsilon/2\}$. Based on the construction, it is not hard to see that an $\varepsilon/2$-net in $\cup_{l_i \le I_i,\, 1 \le i \le k_\varepsilon} \mathcal{G}_{l_1, \ldots, l_{k_\varepsilon}}$ is an ε-net for $\mathcal{S}(\Gamma, \Phi)$. There are fewer than $\binom{I_{k_\varepsilon}}{k_\varepsilon}$ many choices of the basis ϕ_{l_i}, $1 \le i \le k_\varepsilon$, thus the ε-entropy of $\mathcal{S}(\Gamma, \Phi)$ is upper bounded by order $k_\varepsilon + \log \binom{I_{k_\varepsilon}}{k_\varepsilon} = O(k_\varepsilon) \log(\varepsilon^{-1})$ under the assumption $I_k \le k^\tau$ for some $\tau > 1$. As seen in Section 4, if $\gamma_k \sim k^{-\alpha} (\log k)^{-\beta}$, we have that k_ε is of order $\varepsilon^{-1/\alpha} (\log(\varepsilon^{-1}))^{-\beta/\alpha}$. Then the metric entropy of $\mathcal{S}(\Gamma, \Phi)$ is bounded by order $\varepsilon^{-1/\alpha} (\log(\varepsilon^{-1}))^{1-\beta/\alpha}$.

Thus we pay at most a price of a $\log(\varepsilon^{-1})$ factor to cover the larger class $\mathcal{S}(\Gamma, \Phi)$ with a greater freedom of approximation.

For density estimation, suppose the functions in $\mathcal{S}(\Gamma, \Phi)$ are uniformly upper bounded. Let $\tilde{\mathcal{S}}(\Gamma, \Phi)$ be all the probability density functions in $\mathcal{S}(\Gamma, \Phi)$. When γ_0 is large enough, the L_2 metric entropies of $\tilde{\mathcal{S}}(\Gamma, \Phi)$ and $\mathcal{S}(\Gamma, \Phi)$ are of the same order. Let ε_n satisfy $n\varepsilon_n^2 = k_{\varepsilon_n} \log(\varepsilon_n^{-1})$ and $\underline{\varepsilon}_n$ satisfy $k_{\underline{\varepsilon}_n} = n\underline{\varepsilon}_n^2$. Then applying Theorem 3 together with the lower bound on $\mathcal{F}(\Gamma, \Phi)$, we have derived the following result.

THEOREM 8. *The minimax rate of convergence for a sparse approximation set satisfies*

$$\underline{\varepsilon}_n^2 \preceq \min_{\hat{f}} \max_{f \in \tilde{\mathcal{S}}(\Gamma, \Phi)} E\|f - \hat{f}\|_2^2 \preceq \varepsilon_n^2.$$

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

$$n^{-2\alpha/(1+2\alpha)} \preceq \min_{\hat{f}} \max_{f \in \tilde{\mathcal{S}}(\Gamma, \Phi)} E\|f - \hat{f}\|_2^2 \preceq (n/\log n)^{-2\alpha/(1+2\alpha)}.$$

Note that upper and lower bound rates differ only in a logarithmic factor.

From the proofs of the minimax upper bounds, the estimators there are constructed based on Bayes averaging over the ε_n-net of $\mathcal{S}(\Gamma, \Phi)$. In this context, it is also natural to consider estimators based on subset selection. Upper bound results in this direction can be found in Barron, Birgé, and Massart (1999), Yang (1999a) and Yang and Barron (1998).

A general theory of sparse approximation should avoid requiring an assumption of orthogonality of the basis functions ϕ_i, $i \geq 1$. In contrast with the story for full approximation sets where $\mathcal{F}(\Gamma, \Phi)$ is unchanged by the Gram–Schmidt process, sparse approximation is not preserved by orthogonalization. Nonetheless, consideration of those functions that are approximated well by sparse combinations of orthonormal basis has the advantage that conditions can be more easily expressed directly in terms of the coefficients. Here we discuss consequences of Donoho's treatment (1993) of sparse orthonormal approximation for minimax statistical risks.

Let $\{\phi_1, \phi_2, \ldots\}$ be a given orthonormal basis in $L_2[0, 1]$. For $0 < q < 2$, let

$$\mathcal{S}_q(C_1, C_2, \beta) = \left[\sum_{i=1}^{\infty} \xi_i \phi_i : \sum_{i=1}^{\infty} |\xi_i|^q \leq C_1 \text{ and } \sum_{i=1}^{\infty} |\xi_i|^2 \leq C_2 l^{-\beta} \text{ for all } l \geq 1\right].$$

Here C_1, C_2 and β are positive constants though β may be quite small, for example, s/d with s smaller than d. The condition $\sum_{i=l}^{\infty} |\xi_i|^2 \leq C_2 l^{-\beta}$ is used to make the target class small enough to have convergent estimators in L_2 norm. Roughly speaking, the sparsity of the class comes from the condition that $\sum_{i=1}^{\infty} |\xi_i|^q \leq C_1$ which implies that the ith largest coefficient satisfies $|\xi_{(i)}|^q \leq C_1/i$ and that selection of the k largest coefficients are sufficient to achieve a small remaining sum of squares. The condition $\sum_{i=l}^{\infty} |\xi_i|^2 \leq C_2 l^{-\beta}$ is used to ensure that it suffices to select the k largest coefficients from the first $I_k = k^{\tau}$ terms with sufficiently large τ.

The class $\mathscr{S}_q(C_1, C_2, \beta)$ is a special case of function classes with unconditional basis. Let $\mathscr{G}$ be a uniformly bounded function class and $\Phi = \{\phi_1 = 1, \phi_2, \ldots\}$ be an orthonormal basis in L_2. The basis Φ is said to be an unconditional for $\mathscr{G}$ if for any $g = \sum_{i=1}^{\infty} \xi_i \phi_i \in \mathscr{G}$, then $\sum_{i=1}^{\infty} \tilde{\xi}_i \phi_i \in \mathscr{G}$ for all $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ with $|\tilde{\xi}_i| \le |\xi_i|$. Donoho (1993, 1996) gives results on metric entropy of these classes and proves that an unconditional basis for a function class gives essentially best sparse representation of the functions and shows that simple thresholding estimators are nearly optimal. We here apply the main theorems on minimax rates in this paper to these sparse function classes using his metric entropy results.

Let $q^* = q^*(\mathscr{G}) = \inf\{q: \sup_{i \ge 1} i^{1/q} |\xi_{(i)}| < \infty\}$, where $\xi_{(1)}, \xi_{(2)}, \ldots$ are the coefficients ordered in decreasing magnitude $|\xi_{(1)}| \ge |\xi_{(2)}| \ge \ldots$. This q^* is called the sparsity index in Donoho (1993, 1996). Let $\alpha^* = \alpha^*(\mathscr{G}) = \sup\{\alpha: M_2(\varepsilon) = O(\varepsilon^{-1/\alpha})\}$ be the optimal exponent of the L_2 metric entropy $M_2(\varepsilon)$ of $\mathscr{G}$. From Donoho (1996), if $\mathscr{G}$ further satisfies the condition $\sum_{i=l}^{\infty} |\xi_i|^2 \le C l^{-\beta}$ for all $l \ge 1$, then $\alpha^* = 1/q^* - 1/2$, that is, for any $0 < \alpha_1 < \alpha^* < \alpha_2$, $\varepsilon^{-1/\alpha_2} \preceq M_2(\varepsilon) \preceq \varepsilon^{-1/\alpha_1}$. Let $\mathscr{F}_{C'} = \{f: f/C' \in \mathscr{G} \text{ and } f \text{ is a density}\}$. Then when C' is large enough, $\mathscr{F}_{C'}$ has the same metric entropy order as that of $\mathscr{G}$. Thus under the same conditions, we have

$$n^{-2\alpha_2/(2\alpha_1+1)} \preceq \min_{\hat{f}} \max_{f \in \mathscr{F}_{C'}} E\|f - \hat{f}\|_2^2 \preceq n^{-2\alpha_1/(2\alpha_1+1)}$$

for any $0 < \alpha_1 < \alpha^* < \alpha_2$. The first order in the exponent of the minimax risk is $-2\alpha^*/(2\alpha^* + 1)$.

For the special case of $\mathscr{S}_q(C_1, C_2, \beta)$, better entropy bounds are available based on Edmunds and Triebel [(1987), page 141] [see Yang and Barron (1997)], resulting in a sharper upper rate by $n^{-2\alpha/(2\alpha+1)}$ when $\beta/2 \ge 1/q - 1/2$ and by $(n/\log n)^{-2\alpha/(2\alpha+1)}$ when $0 < \beta/2 < 1/q - 1/2$, where $\alpha = 1/q - 1/2$.

6. Examples. In this section, we demonstrate the applications of the theorems developed in the previous sections. As will be seen from the following examples, once we know the order of metric entropy of a target class, the minimax rate can be determined right away for many nonparametric classes without additional work. For results on metric entropy orders of various function classes, see Lorentz, Golitschek and Makovoz (1996) and references cited there.

6.1. *Ellipsoidal classes in L_2.* Let $\{\phi_1 = 1, \phi_2, \ldots, \phi_k, \ldots\}$ be a complete orthonormal system in $L^2[0, 1]$. For an increasing sequence of constants b_k with $b_1 \ge 1$ and $b_k \to \infty$, define an ellipsoidal class $\mathscr{E}(\{b_k\}, C) = \{g = \sum_{i=1}^{\infty} \xi_i \phi_i: \sum_{i=1}^{\infty} \xi_i^2 b_i^2 \le C\}$. Define $m(t) = \sup\{i: b_i \le t\}$ and let $l(t) = \int_0^t m(t)/t\, dt$. Then from Mitjagin (1961), one knows that the L_2 covering metric entropy of $\mathscr{E}(\{b_k\}, C)$ satisfies $l(1/2\varepsilon) \le V_2(\varepsilon) \le l(8/\varepsilon)$. For the special case with $b_k = k^{\alpha}$ $(\alpha > 0)$, the metric entropy order $\varepsilon^{-1/\alpha}$ determined above was previously obtained by Kolmogorov and Tihomirov (1959) for the trigonometric basis. [When $\alpha > 1$ or b_k increases suitably fast, the

entropy rate can also be derived using the results of Lorentz (1966) on full approximation sets.]

6.2. *Classes of functions with bounded mixed differences.* As in Temlyakov (1989), define the function classes $H_q^{\mathbf{r}}$ on $\pi_d = [-\pi, \pi]^d$ having bounded mixed differences as follows. Let $\mathbf{k} = (k_1, \ldots, k_d)$ be a vector of integers, $q \geq 1$, and $\mathbf{r} = (r_1, \ldots, r_d)$ with $r_1 = \cdots = r_v < r_{v+1} \leq \cdots \leq r_d$. Let $L_q(\pi_d)$ denote the periodic functions on π_d with finite norm $\|g\|_{L_q(\pi_d)} := (2\pi)^{-d}\big(\int_{\pi_d} |g(\mathbf{x})|^q \, d\mathbf{x}\big)^{1/q}$. Denote by $H_{q,l}^{\mathbf{r}}$ the class of functions $g(\mathbf{x}) \in L_q(\pi_d)$, $\int_{\pi_d} g(\mathbf{x})\, dx_i = 0$ for $1 \leq i \leq d$, and $\|\Delta_{\mathbf{t}}^l g(\mathbf{x})\|_{L_q(\pi_d)} \leq \prod_{j=1}^d |t_j|^{r_j}$ $(l > \max_j r_j)$, where $\mathbf{t} = (t_1, \ldots, t_d)$ and $\Delta_{\mathbf{t}}^l$ is the mixed lth difference with step t_j in the variable x_j, that is, $\Delta_{\mathbf{t}}^l g(\mathbf{x}) = \Delta_{t_d}^l \cdots \Delta_{t_1}^l g(x_1, \ldots, x_d)$. From Temlyakov (1989), for $\mathbf{r} = (r_1, \ldots, r_d)$ with $r_1 > 1$, $1 \leq p < \infty$ and $1 \leq q \leq \infty$,

$$M_p(\varepsilon; H_q^{\mathbf{r}}) \asymp (1/\varepsilon)^{1/r_1} (\log 1/\varepsilon)^{(1+1/2r_1)(v-1)}.$$

Functions in this class are uniformly bounded.

6.3. *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 $\omega_r(g, t)_q = \sup_{0<h\leq t} \|\Delta_h^r(g, \cdot)\|_q$. Let $\alpha > 0$, $r = [\alpha] + 1$, and

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

Then the Besov norm is defined as $\|g\|_{B_{\sigma,q}^\alpha} = \|g\|_q + |g|_{B_{\sigma q}^\alpha}$ [see DeVore and Lorentz (1993)]. Closely related are Triebel or F classes, which can be handled similarly. For definitions and characterizations of Besov (and F) classes in the d-dimensional case, see Triebel (1975). They include many well-known function spaces such as Hölder–Zygmund spaces, Sobolev spaces, fractional Sobolev spaces or Bessel potential spaces and inhomogeneous Hardy spaces. For $0 < \sigma, q \leq \infty$ $(q < \infty$ for $F)$ and $\alpha > 0$, let $B_{\sigma,q}^\alpha(C)$ be the collections of all functions $g \in L_q[0,1]^d$ such that $\|g\|_{B_{\sigma,q}^\alpha} \leq C$. Building on the conclusions previously obtained for Sobolev classes by Birman and Solomajak (1974), Triebel (1975), with refinements in Carl (1981), showed that for $1 \leq \sigma \leq \infty$, $1 \leq p, q \leq \infty$ and $\alpha/d > 1/q - 1/p$, $M_p(\varepsilon; B_{\sigma,q}^\alpha(C)) \asymp \varepsilon^{-d/\alpha}$.

6.4. *Bounded variation and Lipschitz classes.* The function class $BV(C)$ consists of all functions $g(x)$ on [0,1] satisfying $\|g\|_\infty \leq C$ and $V(g) := \sup \sum_{i=1}^m |g(x_{i+1}) - g(x_i)| \leq C$, where the supremum is taken for all finite sequences $x_1 < x_2 < \cdots < x_m$ in [0,1]. For $0 < \alpha \leq 1$, let $\mathrm{Lip}_{\alpha,q}(C) = \{g : \|g(x + h) - g(x)\|_q \leq Ch^\alpha$ and $\|g\|_q \leq C\}$ be a Lipschitz class. When $\alpha > 1/q - 1/p$, $1 \leq p, q \leq \infty$, the metric entropy satisfies $M_p(\varepsilon; \mathrm{Lip}_{\alpha,q}(C)) \asymp \varepsilon^{-1/\alpha}$

[see Birman and Solomajak (1974)]. For the class of functions in $BV(C)$, with suitable modification of the value assigned at discontinuity points as in DeVore and Lorentz [(1993), Chapter 2] one has $\mathrm{Lip}_{1,\infty}(C) \subset BV(C) \subset \mathrm{Lip}_{1,1}(C)$ and since the $L_p\,(1 \le p < \infty)$ metric entropies of $\mathrm{Lip}_{1,1}(C)$ and $\mathrm{Lip}_{1,\infty}(C)$ are both of order $1/\varepsilon$, the $L_p\,(1 \le p < \infty)$ metric entropy of $BV(C)$ is also of order $1/\varepsilon$.

6.5. *Classes of functions with moduli of continuity of derivatives bounded by fixed functions.* Instead of Lipschitz requirements, one may consider more general bounds on the moduli of continuity of derivatives. Let $\Lambda_{r,\omega}^{d,2} = \Lambda_{r,\omega}^{d,2}(C_0, C_1, \ldots, C_r)$ be the collection of all functions g on $[0,1]^d$ which have all partial derivatives $\|D^{\mathbf{k}} g\|_2 \le C_k$, $|k| = k = 0, 1, \ldots, r$, and the modulus of continuity in the L_2 norm of each rth derivative is bounded by a function ω. Here ω is any given modulus of continuity [for a definition, see DeVore and Lorentz (1993), page 41]. Let $\delta = \delta(\varepsilon)$ be defined by the equation $\delta^r \omega(\delta) = \varepsilon$. Then if $r \ge 1$, again from Lorentz (1966), the L_2 metric entropy of $\Lambda_{r,\omega}^{d,2}$ is of order $(\delta(\varepsilon))^{-d}$.

6.6. *Classes of functions with different moduli of smoothness with respect to different variables.* Let $k_1, \ldots, k_d$ be positive integers and $0 < \beta_i \le k_i$, $1 \le i \le d$. Let $\mathbf{k} = (k_1, \ldots, k_d)$ and $\beta = (\beta_1, \ldots, \beta_d)$. Let $V(\mathbf{k}, \beta, C)$ be the collection of all functions g on $[0,1]^d$ with $\|g\|_\infty \le C$ and $\sup_{|h| \le t} \|\Delta_{i,h}^{k_i} g\|_2 \le Ct^{\beta_i}$, where $\Delta_{i,h}^{k_i}$ is the k_ith difference with step h in variable x_i. From Lorentz [(1996), page 921], the L_2 metric entropy order of $V(\mathbf{k}, \beta, C)$ is $(1/\varepsilon)^{\sum_{i=1}^d \beta_i^{-1}}$.

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

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

on $[0, 2\pi]$ with $\sqrt{a_{m_1,\ldots,m_d}^2 + b_{m_1,\ldots,m_d}^2} \le C(\overline{m}_1 \cdots \overline{m}_d)^{-\alpha} \left(\log^k(\overline{m}_1 \cdots \overline{m}_d) + 1\right)$, where $\overline{m} = \max(m, 1)$. The L_2 metric entropy is of order $(1/\varepsilon)^{1/(\alpha - 1/2)} \log^{(2k + 2\alpha(d-1))/(2\alpha-1)}(1/\varepsilon)$ by Smoljak (1960). Note that for these classes, the dependence of entropy orders on the input dimension d is only through logarithmic factors.

6.8. *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 $g(x) = c_0 + \sum_i c_i \sigma(v_i x + b_i)$, with $|c_0| + \sum_i |c_i| \le C$, and $|v_i| = 1$, where σ is a fixed sigmoidal function with $\sigma(t) \to 1$ as $t \to \infty$ and $\sigma(t) \to 0$ as $t \to -\infty$. We further require that σ is either the step function $\sigma^*(t) = 1$ for $t \ge 0$, and $\sigma^*(t) = 0$ for $t < 0$, or satisfies the Lipschitz requirement that $|\sigma(t) - \sigma(t')| \le C_1 |t - t'|$ for some C_1 and $|\sigma(t) - \sigma^*(t)| \le C_2 |t|^{-\gamma}$ for some C_2 and $\gamma > 0$ for all $t \ne 0$. Approximations to functions in $N(C)$ using k sigmoids achieves L_2 error bounded by $C/k^{1/2}$ [as

shown in Barron (1993)], and using this approximation bound, Barron [(1994), page 125] gives certain metric entropy bounds. The approximation error can not be made uniformly smaller than $(1/k)^{1/2+1/d+\delta}$ for $\delta > 0$ as shown in Barron [(1991), Theorem 3]. Makovoz [(1996), pages 108 and 109] improves the approximation upper bound to a constant times $(1/k)^{1/2+1/(2d)}$ and uses these bounds to show that the L_2 metric entropy of the class $N(C)$ with either the step sigmoid or a Lipschitz sigmoid satisfies $(1/\varepsilon)^{1/(1/2+1/d)} \preceq M_2(\varepsilon) \preceq (1/\varepsilon)^{1/(1/2+1/(2d))} \log(1/\varepsilon)$. For $d = 2$, a better lower bound matches the upper bound in the exponent $(1/\varepsilon)^{4/3} \log^{-2/3}(1/\varepsilon) \preceq M_2(\varepsilon) \preceq (1/\varepsilon)^{4/3} \log(1/\varepsilon)$. For $d = 1$, $N(C)$ is equivalent to a bounded variation class.

Convergence rates. We give convergence rates for density estimation. Unless stated otherwise, attention is restricted to densities in the corresponding class. Norm parameters of the function classes are assumed to be large enough so that the restriction does not change the metric entropy order. For regression, the rates of convergence are the same for each of these function classes assuming the design density (with respect to Lebesgue measure) is bounded above and away from zero.

1. Assume the functions in $\mathscr{C}(\{b_k\}, C)$ are uniformly bounded by C_0 and assume that $l(t)$ satisfies $\liminf_{t\to 0} l(\beta t)/l(t) > 1$ for some fixed constant $\beta > 1$. Let ε_n be determined by $l(1/\varepsilon_n) = n\varepsilon_n^2$; we have

$$\min_{\hat{f}} \max_{f\in\mathscr{C}(\{b_k\}, C)} E\|f - \hat{f}\|_2^2 \asymp \varepsilon_n^2.$$

   Specially, if $b_k = k^\alpha$, $k \geq 1$ and the basis functions satisfy $\sup_{k\geq 1} \|\phi_k\|_\infty < \infty$, then when $\alpha > 1/2$, functions in $\mathscr{C}(\{b_k\}, C)$ are uniformly bounded. Then the rate of convergence is $n^{-2\alpha/(2\alpha+1)}$. See Efroimovich and Pinsker (1982) for more detailed asymptotics with the trigonometric basis and $b_k = k^\alpha$, see Birgé (1983) for similar conclusions using a hypercube construction for the trigonometric basis and Barron, Birgé and Massart [(1999), Section 3] for general ellipsoids.
2. Let $\tilde{H}_q^{\mathbf{r}}(C) = \{f = e^g / \int e^g \, d\mu : g \in H_q^{\mathbf{r}}(C)\}$ be a uniformly bounded density class. The metric entropy of $\tilde{H}_q^{\mathbf{r}}(C)$ is of the same order as for $H_q^{\mathbf{r}}(C)$. So for $r_1 > 1$ and $1 \leq q \leq \infty$, $1 \leq p \leq 2$, by Corollary 1, we have

$$\min_{\hat{f}} \max_{f\in\tilde{H}_q^{\mathbf{r}}} E\|f - \hat{f}\|_p \asymp n^{-r_1/(2r_1+1)} (\log n)^{(v-1)/2}.$$

   For this density class, the minimax risk under K-L divergence or squared Hellinger distance is also of the same order as that under the squared L_2 distance, $n^{-2r_1/(2r_1+1)} (\log n)^{(v-1)}$.
3. When $\alpha/d > 1/q$, the functions in $B_{\sigma,q}^\alpha(C)$ are uniformly bounded. For $1 \leq \sigma \leq \infty$, $1 \leq q \leq \infty$, $1 \leq p \leq 2$ and $\alpha/d > 1/q$,

$$\min_{\hat{f}} \max_{f\in B_{\sigma,q}^\alpha(C)} E\|f - \hat{f}\|_p^2 \asymp n^{-2\alpha/(2\alpha+d)},$$

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

When $1/q - 1/2 < \alpha/d \le 1/q$, some functions in $B^{\alpha}_{\sigma,q}(C)$ are not bounded; nevertheless, this class has the same order metric entropy as the subclass $B^{\alpha}_{\sigma,\infty}(C)$ with $q^* > d/\alpha$ which is uniformly bounded. Consequently for C sufficiently large the metric entropy of class of densities in $B^{\alpha}_{\sigma,q}(C)$ is still the same order as for $B^{\alpha}_{\sigma,q}(C)$. From Theorem 4, for L_1 risk we have, when $\alpha/d > 1/q - 1/2$, the rate is also $n^{-\alpha/(2\alpha+d)}$. From the monotonicity property of the L_p norm in p, we have for $p > 1$ and $\alpha/d > 1/q - 1/2$, $\min_{\hat{f}} \max_{f \in \mathcal{F}} E\|f - \hat{f}\|_p \succeq n^{-\alpha/(2\alpha+d)}$. In particular, when $q \ge 2$, the last two conclusions hold for all $\alpha > 0$. Donoho, Johnstone, Kerkyacharian and Picard (1993) obtained suitable minimax bounds for the case of $p \ge q$, $\alpha > 1/q$. Here we permit smaller α. The rates for Lipschitz classes were previously obtained by Birgé (1983) and Devroye (1987) (the latter for special cases with $p = 1$). If the log-density is assumed to be in $B^{\alpha}_{\sigma,q}(C)$, then by Corollary 1, for $\alpha/d > 1/q$, the minimax risk under K-L divergence is of the same order $n^{-2\alpha/(2\alpha+d)}$, which was previously shown by Koo and Kim (1996).

4. For $1 \le p \le 2$, the minimax rate for square L_p risk or L_p risk is of order $n^{-2/3}$ or $n^{-1/3}$. Since a class of bounded monotone functions has the same order metric entropy as a bounded variation class, we conclude immediately that it has the same minimax rate of convergence. The rate is obtained in Birgé (1983) by setting up a delicate local packing set for the lower bound.
5. Let ε_n be chosen such that $(\delta(\varepsilon_n))^{-d} = n\varepsilon_n^2$. Then for $r \ge 1$, the minimax square L_2 risk is at rate ε_n^2.
6. Let $\alpha^{-1} = \sum_{i=1}^{d} \beta_i^{-1}$; then the minimax square L_2 risk is at rate $n^{-2\alpha/(2\alpha+1)}$.
7. The class $E_d^{\alpha,k}(C)$ is uniformly bounded when $\alpha > 1$. Then

$$\min_{\hat{f}} \max_{f \in E_d^{\alpha,k}(C)} E\|f - \hat{f}\|_2^2 \asymp n^{-(\alpha-1/2)/\alpha} \log^{(k+\alpha(d-1))/\alpha} n.$$

8. We have upper and lower rates

$$n^{-(1+2/d)/(2+1/d)} (\log n)^{-(1+1/d)(1+2/d)(2+1/d)} \preceq \min_{\hat{f}} \max_{f \in N(C)} E\|f - \hat{f}\|_2^2 \preceq (n/\log n)^{-(1+1/d)/(2+1/d)},$$

and for $d = 2$, using the better lower bound on the metric entropy, we have

$$n^{-3/5} (\log n)^{-19/10} \preceq \min_{\hat{f}} \max_{f \in N(C)} E\|f - \hat{f}\|_2^2 \preceq (n/\log n)^{-3/5}.$$

For this case, it seems unclear how one could set up a local packing set for applying Birgé's lower bounding result. Though the upper and lower rates do not agree (except $d = 2$, ignoring a logarithmic factor), when d is moderately large, the rates are roughly $n^{-1/2}$, which is independent of d.

7. Relationship between global and local metric entropies. As stated in the introduction, the previous use of Fano's inequality to derive the minimax rates of convergence involved local metric entropy calculation. To apply this technique, constructions of special local packing sets capturing the essential difficulty of estimating a density in the target class are seemingly required and they are usually done with a hypercube argument. We have shown in Section 2 that the global metric entropy alone determines the minimax lower rates of convergence for typical nonparametric density classes. Thus there is no need to put in efforts in search of special local packing sets. After the distribution of an early version of this paper which contained the main results in Section 2, we realized a connection between the global metric entropy and local metric entropy. In fact, the global metric entropy ensures the existence of at least one local packing set which has the property required for the use of Birgé's argument. This fact also allows one to bypass the special constructions. Here we show this connection between the global metric entropy and local metric entropy and comment on the uses of these metric entropies.

For simplicity, we consider the case when d is a metric. Suppose the global packing entropy of S under distance d is $M(\varepsilon)$.

DEFINITION (Local metric entropy). The local ε-entropy at $\theta \in S$ is the logarithm of the largest $(\varepsilon/2)$-packing set in $B(\theta, \varepsilon) = \{\theta' \in S: d(\theta', \theta) \leq \varepsilon\}$. The local ε-entropy at θ is denoted by $M(\varepsilon \mid \theta)$. The local ε-entropy of S is defined as $M^{\text{loc}}(\varepsilon) = \max_{\theta \in S} M(\varepsilon \mid \theta)$.

LEMMA 3. *The global and local metric entropies have the following relationship*:

$$M(\varepsilon/2) - M(\varepsilon) \leq M^{\text{loc}}(\varepsilon) \leq M(\varepsilon/2).$$

PROOF. Let N_ε and $N_{\varepsilon/2}$ be the largest ε-packing set and the largest $\varepsilon/2$-packing set respectively in S. Let us partition $N_{\varepsilon/2}$ into $|N_\varepsilon|$ parts according to the minimum distance rule (the Voronoi partition). For $\theta \in N_\varepsilon$, let $R_\theta = \{\theta': \theta' \in S, \theta = \arg\min_{\tilde{\theta} \in N_\varepsilon} d(\theta', \tilde{\theta})\}$ be the points in $N_{\varepsilon/2}$ that are closest to θ (if a point in $N_{\varepsilon/2}$ has the same distance to two different points in N_ε, any rule can be used to ensure $R_\theta \cap R_{\tilde{\theta}} = \phi$). Note that $N_{\varepsilon/2} = \cup_{\theta \in N_\varepsilon} R_\theta$. Since N_ε is the largest packing set, for any $\theta' \in N_{\varepsilon/2}$, there exists $\theta \in N_\varepsilon$ such that $d(\theta', \theta) \leq \varepsilon$. It follows that $d(\theta', \theta) \leq \varepsilon$ for $\theta' \in R_\theta$. From above, we have

$$\frac{|N_{\varepsilon/2}|}{|N_\varepsilon|} = \frac{1}{|N_\varepsilon|} \sum_{\theta \in N_\varepsilon} |R_\theta|.$$

Roughly speaking, the ratio of the numbers of points in the two packing sets characterizes the average local packing capability.

From the above identity, there exists at least one $\theta^* \in N_\varepsilon$ with $|R_{\theta^*}| \geq |N_{\varepsilon/2}|/|N_\varepsilon|$. Thus we have $M(\varepsilon \mid \theta^*) \geq M(\varepsilon/2) - M(\varepsilon)$. On the other hand,

by concavity of the log function,

$$M\Big(\frac{\varepsilon}{2}\Big) - M(\varepsilon) = \log\left(\frac{1}{|N_\varepsilon|}\sum_{\theta\in N_\varepsilon}|R_\theta|\right) \geq \frac{1}{|N_\varepsilon|}\sum_{\theta\in N_\varepsilon}\log(|R_\theta|).$$

Thus an average of the local ε-entropies is upper bounded by the difference of two global metric entropies $M(\varepsilon/2) - M(\varepsilon)$. An obvious upper bound on the local ε-entropies is $M(\varepsilon/2)$. This ends the proof of Lemma 3. $\square$

The sets R_θ in the proof are local packing sets which have the property that the diameter of the set is of the same order as the smallest distance between any two points. Indeed, the points in R_θ are $\varepsilon/2$ apart from each other and are within ε from θ. This property together with the assumption that locally d_K is upper bounded by a multiple of d enables the use of Birgé's result [(1983), Proposition 2.8] to get the lower bound on the minimax risk.

To identify the minimax rates of convergence, we assume there exist constants $\overline{A} > 0$ and $\varepsilon_0 > 0$ such that $D(p_\theta\|p_{\theta'}) \leq \overline{A}d^2(\theta, \theta')$, for $\theta, \theta' \in S$ with $d(\theta, \theta') \leq \varepsilon_0$. From Lemma 3, there exist $\theta^* \in S$ and a subset $S_0 = R_{\theta^*}$ with a local packing set N of log-cardinality at least $M(\varepsilon/2) - M(\varepsilon)$. Using Fano's inequality together with the diameter bound on mutual information as in Birgé, yields with θ a uniformly distributed random variable on N, when $\varepsilon \leq \varepsilon_0$,

$$I(\Theta; X^n) \leq n\max_{\theta,\,\theta'\in N} D(p_\theta\|p_{\theta'}) \leq n\overline{A}\max_{\theta,\,\theta'\in S_0} d^2(\theta, \theta') \leq n\overline{A}\varepsilon^2,$$

and hence, choosing ε_n to satisfy $M(\varepsilon_n/2) - M(\varepsilon_n) = 2(n\overline{A}\varepsilon_n^2 + \log 2)$, as in the proof of Theorem 1, but with S replaced by $S_0 = R_{\theta^*}$, one gets $\min_{\hat\theta}\max_{\theta\in S_0} E_\theta d^2(\theta, \hat\theta) \geq \varepsilon_n^2/32$, which is similar to our conclusions from Section 2 (cf. Corollary 1). The difference in the bound just obtained compared with the previous work of Birgé and others is the use of Lemma 3 to avoid requiring explicit construction of the local packing set.

If one does have knowledge of the entropy of an ε-ball with the largest order $\varepsilon/2$-packing set, then the same argument with S_0 equal to this ε-ball yields the conclusion

$$\min_{\hat\theta}\max_{\theta\in S_0} E_\theta d^2(\theta, \hat\theta) \succeq \tilde\varepsilon_n^2, \tag{7}$$

where $\varepsilon = \tilde\varepsilon_n$ is determined by $M^{\mathrm{loc}}(\tilde\varepsilon_n) = n\tilde\varepsilon_n^2$, provided $D(p_\theta\|p_{\theta'}) \leq \overline{A}d^2(\theta, \theta')$ holds for θ, θ' in the chosen packing set.

The above lower bound is often at the optimal rate even when the target class is parametric. For instance, the ε-entropy of a usual parametric class is often of order $m\log(1/\varepsilon)$, where m is the dimension of the model. Then the metric entropy difference $M(\varepsilon/2) - M(\varepsilon)$ or $M^{\mathrm{loc}}(\varepsilon)$ is of order of a constant, yielding the anticipated rate $\varepsilon_n \asymp 1/\sqrt{n}$ and $\tilde\varepsilon_n \asymp 1/\sqrt{n}$. The achievability of the rate in (7) is due to Birgé [(1986), Theorem 3.1] under the following condition.

CONDITION 4. There is a nonincreasing function $U:(0,\infty)\to[1,\infty)$. For any $\varepsilon>0$ with $n\varepsilon^2\geq U(\varepsilon)$, there exists an ε-net S_ε for S under d such that for all $\lambda\geq 2$ and $\theta\in S$, we have $\operatorname{card}(S_\varepsilon\cap B_d(\theta,\lambda\varepsilon))\leq\lambda^{U(\varepsilon)}$. Moreover, there are positive constants $A_1>1$, $A_2>0$, a $\theta_0\in S$ and for $\lambda\geq 2$ there exists for all sufficiently small ε, an ε-packing set N_ε for $B_d(\theta_0,\lambda\varepsilon)$ satisfying $\operatorname{card}(N_\varepsilon\cap B_d(\theta_0,\lambda_\varepsilon))\geq(\lambda/A_1)^{U(\varepsilon)}$ and $d_K^2(\theta,\theta')\leq A_2^2d^2(\theta,\theta')$ for all θ,θ' in N_ε.

PROPOSITION 2 (Birgé). *Suppose Condition* 4 *is satisfied. If the distance* d *is a metric bounded above by a multiple of the Hellinger distance, then*

$$\min_{\hat\theta\in\overline{S}_n}\max_{\theta\in S}E_\theta d^2(\theta,\hat\theta)\asymp\varepsilon_n^2,$$

where ε_n *satisfies* $n\varepsilon_n^2=U(\varepsilon_n)$.

The upper bound is from Birgé [(1986), Theorem 3.1] using the first part of Condition 4. From the second part of Condition 4 with $\lambda=2$, we have $M^{\mathrm{loc}}(2\varepsilon)\asymp U(\varepsilon)$ and a suitable relationship between d_K and d in a maximal order local ε-net, so the lower bound follows from (7) by the Fano inequality bound as discussed above in accordance with Birgé [(1983), Proposition 2.8].

When $\liminf_{\varepsilon\to0}M(\varepsilon/2)/M(\varepsilon)>1$ (as in Condition 2), the upper bounds given in Section 2 are optimal in terms of rates when d and d_K are locally equivalent. For such cases, there is no difference in considering global or local metric entropy. The condition $\liminf_{\varepsilon\to0}M(\varepsilon/2)/M(\varepsilon)>1$ is characteristic of large function classes, as we have seen. In that case, the three entropy quantities in Lemma 3 are asymptotically equivalent as $\varepsilon\to0$,

$$M(\varepsilon/2)-M(\varepsilon)\asymp M^{\mathrm{loc}}(\varepsilon)\asymp M(\varepsilon/2).$$

In contrast, when $\lim_{\varepsilon\to0}M(\varepsilon/2)/M(\varepsilon)=1$ as is typical of finite-dimensional parametric cases, $M(\varepsilon/2)-M(\varepsilon)$ is of smaller order than $M(\varepsilon/2)$. In this case, one may use (7) to determine a satisfactory lower bound on convergence rate.

From the above results, it seems that for lower bounds, it is enough to know the global metric entropy, but for upper bounds, when $\liminf_{\varepsilon\to0}M(\varepsilon/2)/M(\varepsilon)=1$, under a stronger homogeneous entropy assumption (Condition 4), the local entropy condition gives the right order upper bounds, while using the global entropy results in suboptimal upper bounds (within a logarithmic factor). However, for inhomogeneous finite-dimensional spaces, no general results are available to identify the minimax rates of convergence.

Besides being used for the determination of minimax rates of convergence, local entropy conditions are used to provide adaptive estimators using different kinds of approximating models [see Birgé and Massart (1993, 1995), Barron, Birgé and Massart (1999) and Yang and Barron (1998)].

APPENDIX

Proofs of some lemmas.

PROOF OF LEMMA 2. The proof is by a truncation of g from above and below. Let $G = \{x: g(x) \le 4T\}$. Let $\overline{g} = gI_G + 4TI_{G^c}$. Then because $d_H(f, g) \le \varepsilon$, we have $\int_{G^c}(\sqrt{f} - \sqrt{g})^2\, d\mu \le \varepsilon^2$. Since $f(x) \le T \le g(x)/4$ for $x \in G^c$, it follows that $\int_{G^c}(\sqrt{g} - \sqrt{g/4})^2 \le \int_{G^c}(\sqrt{f} - \sqrt{g})^2\, d\mu \le \varepsilon^2$. Thus $\int_{G^c} g\, d\mu \le 4\varepsilon^2$, which implies $1 - 4\varepsilon^2 \le \int \overline{g}\, d\mu \le 1$ and $\int(\sqrt{g} - \sqrt{\overline{g}})^2\, d\mu \le \int_{G^c} g\, d\mu \le 4\varepsilon^2$. Let $\tilde{g} = (\overline{g} + 4\varepsilon^2)/(\int \overline{g}\, d\mu + 4\varepsilon^2)$. Clearly $\tilde{g}$ is a probability density function with respect to μ. For $0 \le z \le 4T$, by simple calculation using $1 - 4\varepsilon^2 \le \int \overline{g}\, d\mu \le 1$, we have $|\sqrt{z} - \sqrt{(z + \varepsilon^2)/(\int \overline{g}\, d\mu + \varepsilon^2)}| \le 2(8T - 1)\varepsilon$. Thus $\int(\sqrt{\overline{g}} - \sqrt{\tilde{g}})^2\, d\mu \le 4(8T - 1)^2\varepsilon^2$. Therefore, by triangle inequality,

$$\begin{aligned}\int\left(\sqrt{f} - \sqrt{\tilde{g}}\right)^2 d\mu &\le 2\int\left(\sqrt{f} - \sqrt{g}\right)^2 d\mu + 4\int\left(\sqrt{g} - \sqrt{\overline{g}}\right)^2 d\mu \\ &\quad + 4\int\left(\sqrt{\overline{g}} - \sqrt{\tilde{g}}\right)^2 d\mu \\ &\le 2\varepsilon^2 + 16\varepsilon^2 + 16(8T - 1)^2\varepsilon^2.\end{aligned}$$

That is, $d_H^2(f, \tilde{g}) \le 2(9 + 8(8T - 1)^2)\varepsilon^2$. Because $f/\tilde{g}$ is upper bounded by $T/(4\varepsilon^2/(\int \overline{g}\, d\mu + 4\varepsilon^2)) \le 9T/(4\varepsilon^2)$, the K-L divergence is upper bounded by a multiple of the square Hellinger distance [see, e.g., Birgé and Massart (1994), Lemma 4 or Yang and Barron (1998), Lemma 4] as given in the lemma. This completes the proof of Lemma 2. □

LEMMA 4. *Assume Conditions* 1 *and* 2 *are satisfied. Let* ε_n *satisfies* $\varepsilon_n^2 = V(\varepsilon_n)/n$ *and* $\underline{\varepsilon}_{n,d}$ *be chosen such that* $M(\underline{\varepsilon}_{n,d}) = 4n\varepsilon_n^2 + 2\log 2$. *Then* $\underline{\varepsilon}_{n,d} \asymp \varepsilon_n$.

PROOF. Let $\sigma = \liminf_{\varepsilon \to 0} M(\alpha\varepsilon)/M(\varepsilon) > 1$. Under the assumption $M(\varepsilon) > 2\log 2$ when ε is small, we have $4n\varepsilon_n^2 + 2\log 2 \le 6n\varepsilon_n^2$ when n is large enough. Then under Condition 1, $M((a/b)\varepsilon_n) \ge (1/c)V(\varepsilon_n) \ge (6c)^{-1}M(\underline{\varepsilon}_{n,d})$. Take k large enough such that $\sigma^k \le 6c$. Then $M((a/b)\alpha^k\varepsilon_n) \ge \sigma^k M((a/b)\varepsilon_n) \ge M(\underline{\varepsilon}_{n,d})$. Thus $(a/b)\alpha^k\varepsilon_n \le \underline{\varepsilon}_{n,d}$, that is, $\varepsilon_n = O(\underline{\varepsilon}_{n,d})$. Similarly, $M(\varepsilon_n/b) \le V(\varepsilon_n) = n\varepsilon_n^2 \le (1/4)M(\underline{\varepsilon}_{n,d}) \le M(\underline{\varepsilon}_{n,d})$. So $\varepsilon_n/b \ge \underline{\varepsilon}_{n,d}$, that is, $\underline{\varepsilon}_{n,d} = O(\varepsilon_n)$. This completes the proof. □

Acknowledgments. The referees and an Associate Editor are thanked for encouraging improved focus and clarity of presentation in the paper.

REFERENCES

BALL, K. and PAJOR, A. (1990). The entropy of convex bodies with "few" extreme points. In *Geometry of Banach Spaces* 26–32. Cambridge Univ. Press.

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

BARRON, A. R. (1991). Neural net approximation. In *Proceedings of the Yale Workshop on Adaptive Learning Systems* (K. Narendra, ed.) Yale University.

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

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

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

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

BARRON, A. R. and HENGARTNER, N. (1998). Information theory and superefficiency. *Ann. Statist.* **26** 1800–1825.

BARRON, A. R. and SHEU, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BIRGÉ, L. (1983). Approximation dans les espaces metriques et theorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

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

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

BIRGÉ, L. and MASSART, P. (1994). Minimum contrast estimators on sieves. Technical report, Univ. Paris-Sud.

BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29.

BIRGÉ, L. and MASSART, P. (1996). From model selection to adaptive estimation. In *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Gam* (D. Pollard, E. Torgersen and G. Yang, eds.) 55–87. Springer, New York.

BIRMAN, M. S. and SOLOMJAK, M. (1974). Quantitative analysis in Sobolev embedding theorems and application to spectral theory. *Tenth Math. School Kiev* 5–189.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densites: risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137.

CARL, B. (1981). Entropy numbers of embedding maps between Besov spaces with an application to eigenvalue problems. *Proc. Roy. Soc. Edinburgh* **90A** 63–70.

CENCOV, N. N. (1972). *Statistical Decision Rules and Optimal Inference*. Nauka, Moscow; English translation in *Amer. Math. Soc. Transl.* **53** (1982).

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

CLARKE, B. and BARRON, A. R. (1994). Jeffrey's prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference* **41** 37–60.

COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

COX, D. D (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

DAVISSON, L. (1973). Universal noiseless coding. *IEEE Trans. Inform. Theory* **19** 783–795.

DAVISSON, L. and LEON-GARCIA, A. (1980). A source matching approach to finding minimax codes. *IEEE Trans. Inform. Theory* **26** 166–174.

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

DEVROYE, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston.

DONOHO, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.* **1** 100–115.

DONOHO, D. L. (1996). Unconditional bases and bit-level compression. Technical report 498, Dept. Statistics, Stanford Univ.

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

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence II. *Ann. Statist*. **19** 633–667.

DUDLEY, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probab*. **15** 1306–1326.

EDMUNDS, D. E. and TRIEBEL, H. (1987). Entropy numbers and approximation numbers in function spaces. *Proc. London Math. Soc*. **58** 137–152.

EFROIMOVICH, S. YU. and PINSKER, M. S. (1982). Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii* **18** 19–38.

FANO, R. M. (1961). *Transmission of Information: A Statistical Theory of Communication*. MIT Press.

FARRELL, R. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist*. **43** 170–180.

HASMINSKII, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Probab. Appl*. **23** 794–796.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist*. **18** 999–1010.

HAUSSLER, D. (1997). A general minimax result for relative entropy. *IEEE Trans. Inform. Theory* **40** 1276–1280.

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

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

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1978). On the capacity in communication by smooth signals. *Soviet Math. Dokl*. **19** 1043–1047.

JONES, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist*. **20** 608–613.

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14** 3–86.

KOO, J. Y. and KIM, W. C. (1996). Wavelet density estimation by approximation of log-densities. *Statist. Probab. Lett* **26** 271–278.

LE CAM, L. M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist*. **1**, 38–53.

LE CAM, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

LORENTZ, G. G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc*. **72** 903–937.

LORENTZ, G. G., GOLITSCHEK, M. V. and MAKOVOZ, Y. (1996). *Constructive Approximation: Advanced Problems*. Springer, New York.

MAKOVOZ, Y. (1996). Random approximants and neural networks. *J. Approx. Theory* **85** 98–109.

MITJAGIN, B. S. (1961). The approximation dimension and bases in nuclear spaces. *Uspekhi Mat. Nauk* **16** 63–132.

NEMIROVSKII, A. (1985). Nonparametric estimation of smooth regression functions. *J. Comput. System. Sci*. **23** 1–11.

NEMIROVSKII, A., POLYAK, B. T. and TSYBAKOV, A. B. (1985). Rates of convergence of nonparametric estimates of maximum-likelihood type. *Probl. Peredachi Inf*. **21** 17–33.

POLLARD, D. (1993). Hypercubes and minimax rates of convergence. Preprint.

RISSANEN, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **30** 629–636.

RISSANEN, J., SPEED, T. and YU, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Inform. Theory* **38** 315–323.

SMOLJAK, S. A. (1960). The ϵ-entropy of some classes $E_s^{\alpha,k}(B)$ and $W_s^{\alpha}(B)$ in the L_2 metric. *Dokl. Akad. Nauk SSSR* **131** 30–33.

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

TEMLYAKOV, V. N. (1989). Estimation of the asymptotic characteristics of classes of functions with bounded mixed derivative or difference. *Trudy Mat. Inst. Steklov* **189** 162–197.

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

VAN DE GEER, S. (1990). Hellinger consistency of certain nonparametric maximum likelihood estimates. *Ann. Statist*. **21** 14–44.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist*. **23** 339–362.

YANG, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9** 475–500.

YANG, Y. (1999). Minimax nonparametric classification I: rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284.

YANG, Y. and BARRON, A. R. (1997). Information-theoretic determination of minimax rates of convergence. Technical Report 28, Dept. Statistics, Iowa State Univ.

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

YATRACOS, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist*. **13** 768–774.

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

YU, B. (1996). Assouad, Fano, and Le Cam. In *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.) 423–435. Springer, New York.

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

YALE UNIVERSITY
DEPARTMENT OF STATISTICS
P.O. BOX 208290
YALE STATION
NEW HAVEN, CONNECTICUT 06520-8290
E-MAIL: barron@stat.yale.edu