Information-Theoretic Determination of Minimax Rates of Convergence *

Yuhong Yang and Andrew Barron
Iowa State University and Yale University
March 5, 1997

Abstract

In this paper, 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, two methods are often used to derive the minimax lower bounds: Fano’s inequality and Assouad’s lemma. These methods are used in Ibragimov and Hasminskii (1977, 1978, 1980, and 1982), Hasminskii (1978), Bretagnolle and Huber (1979), Efroimovich and Pinsker (1982), Stone (1982), Birgé (1983, 1986), Nemirovskii (1985), Devroye (1987), Yatracos (1988), Hasminskii and Ibragimov (1990) and others. Birgé (1986) claims that Fano’s

---

*This work was supported by NSF Grants ECS-9410760 and DMS-9505168.
Key words and phrases. minimax risk, density estimation, metric entropy, Kullback-Leibler distance.

Both Assouad’s lemma and Fano’s inequality as they have previously been used to obtain lower bounds involve restriction to a local subset of the function space and the assumption of special properties of packing sets in such a subset. The purpose of our work here is to demonstrate situations under which the convergence rate is determined by the global metric entropy over the whole function class (or over large subsets of it). The advantage of this approach is that the metric entropies are available in approximation theory for many function classes (see, e.g., Lorentz, Golitschek and Makovoz (1996)). In such cases, it is not necessary to uncover additional local packing properties.

The following proposition is representative of the results obtained here. Let \( \mathcal{F} \) be a class of functions, let \( d(f, g) \) be a distance between functions, let \( N(\epsilon; \mathcal{F}) \) be the size of the largest packing set of functions separated by at least \( \epsilon \) in \( \mathcal{F} \), and let \( \epsilon_n \) satisfy \( \epsilon_n^2 = M(\epsilon_n)/n \), where \( M(\epsilon) = \log N(\epsilon; \mathcal{F}) \) is the Kolmogorov \( \epsilon \)-entropy and \( n \) is the sample size. Assume the target class \( \mathcal{F} \) is rich enough to satisfy \( \lim_{n \to \infty} M(\epsilon/2)/M(\epsilon) > 1 \) (which is true, e.g., if \( M(\epsilon) = (1/\epsilon)^r \kappa(\epsilon) \) with \( r > 0 \) and \( \kappa(\epsilon/2)/\kappa(\epsilon) \to 1 \) as \( \epsilon \to 0 \)). This condition is satisfied in typical nonparametric classes.

For convenience, we will use the symbols \( \gtrsim \) and \( \asymp \), where \( a_n \gtrsim b_n \) means \( b_n = O(a_n) \), and \( a_n \asymp b_n \) means both \( a_n \gtrsim b_n \) and \( b_n \gtrsim a_n \).

**Proposition 1:** In the following cases, the minimax convergence rate is characterized by metric entropy in terms of the critical separation \( \epsilon_n \) as follows:

\[
\min_{f} \max_{\hat{f} \in \mathcal{F}} E_{f} d^2(f, \hat{f}) \asymp \epsilon_n^2.
\]

1. \( \mathcal{F} \) is any class of density functions bounded above and below \( 0 < \underline{\mathcal{C}} \leq f \leq \overline{\mathcal{C}} \) 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.
2. \( \mathcal{F} \) is a convex class of densities with \( f \leq \mathcal{C} \) 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.

3. \( \mathcal{F} \) is any class of functions \( f \) with \( |f| \leq \mathcal{C} \) for \( f \in \mathcal{F} \) for the regression model \( Y = f(X) + \epsilon \), \( X \) and \( \epsilon \) are independent \( X \sim P_X \) and \( \epsilon \sim \text{Normal}(0, \sigma^2) \), \( \sigma > 0 \) 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 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 subsets, the minimax rate may remain unchanged compared to that of the original class. An interesting application is bounding K-L risk of density estimation with support being an interval with unknown boundaries.

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 \( \{\theta_1, ..., \theta_m\} \) in \( S_0 \) separated from each other by an amount \( \epsilon_n \) in the distance of interest. The critical separation \( \epsilon_n \) is the largest separation such that the hypothesis \( \{\theta_1, ..., \theta_m\} \) 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 \( p_{\theta_i}(x_1, ..., x_n) = \prod_{i=1}^{n} p_{\theta_i}(x_i) \) and the centroid of such densities \( q(x_1, ..., x_n) = (1/m) \sum_{i=1}^{m} p_{\theta_i}(x_1, ..., x_n) \) (which is the Shannon mutual information \( I(\Theta; X_1, ..., X_n) \) between \( \theta \) and \( X_1, ..., X_n \) with a uniform distribution on \( \theta \) in \( \{\theta_1, ..., \theta_m\} \)). 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 \( \theta \) is determined by \( X^n \)). It is critical here that K-L divergence does not have a triangle inequality between the joint densities. Indeed, the K-L divergence from every \( p_{\theta_j}(x_1, ..., x_n) \) to the centroid is shown to be bounded by the right order \( 2n\epsilon_n^2 \) even though the distance between two such \( p_{\theta_j}(x_1, ..., x_n) \) is as large as \( n\beta \) where \( \beta \) is the K-L diameter of the whole set \( \{p_{\theta_1}, ..., p_{\theta_m}\} \). The proper convergence rate is thus identified provided the cardinality
of the subset $m$ is chosen such that $n\epsilon_n^2 / \log m$ is bounded by a suitable constant less than 1.
Thus $\epsilon_n$ is determined by solving for $n\epsilon_n^2 / M(\epsilon_n)$ equal to such a constant, where $M(\epsilon)$ is the metric entropy (the logarithm of the largest cardinality of an $\epsilon$-packing set). In this way, the metric entropy provides a lower bound on the minimax convergence rate.

Previous applications of Fano's inequality to density estimation used the K-L diameter $n\beta$
of the set $\{p_{\theta_1}^n, ..., p_{\theta_m}^n\}$ (see, e.g., Birgé (1983)) or similar rough bounds (such as $nI(\Theta; X_1)$
as in Hasinskii (1978)) on the average distance of $p_{\theta_1}^n, ..., p_{\theta_m}^n$ from their centroid. In that
theory, to obtain a suitable bound, a statistician needs to find if possible a sufficiently large
subset $\{\theta_1, ..., \theta_m\}$ 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
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 $S_0$.

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)).

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 $\epsilon$ by balls of radius $\epsilon/2$, 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 8, 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 8 sections. In Section 2, the main results are presented. Applications
in data compression and regression are given in Section 3 and 4, respectively. In Section
5 and 6, results connecting linear approximation and minimax rates, sparse approximation and
minimax rates are given respectively. In Section 7, we illustrate the determination of minimax
rates of convergence for several function classes for both density estimation and regression. In Section 8, we discuss the relationship between the global entropy and local entropy. Proofs of lemmas are given in an appendix.

2 Main results

Suppose we have a collection of densities \( \{ p_\theta : \theta \in \Theta \} \) defined on a measurable space \( \mathcal{X} \) with respect to a \( \sigma \)-finite measure \( \mu \). The parameter space \( \Theta \) could be a finite-dimensional space or a nonparametric space (e.g., the class of all densities). When the parameter is the density itself, we may use \( f \) and \( \mathcal{F} \) in place of \( \theta \) and \( \Theta \) respectively. Let \( X_1, X_2, \ldots, X_n \) be an i.i.d. sample from \( p_\theta, \theta \in \Theta \). We want to estimate the true density \( p_\theta \) or \( \theta \) 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 \( \{ p_\theta : \theta \in S \} \), \( S \subseteq \Theta \). Our technique is most appropriate for nonparametric classes (e.g., the class of densities with certain derivative satisfying a Lipschitz condition).

Let \( \overline{S} \) be an action space for the parameter estimates with \( S \subseteq \overline{S} \subseteq \Theta \). An estimator of \( \theta \) is then a measurable mapping from the sample space of \( X_1, X_2, \ldots, X_n \) to \( \overline{S} \). Let \( \mathcal{A}_n \) 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 \( d(\theta, \hat{\theta}) = 0 \) and \( d(\theta, \hat{\theta}') > 0 \) for \( \theta \neq \hat{\theta}' \). 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

\[
\min_{\theta \in \mathcal{A}_n} \max_{\hat{\theta} \in \overline{S}} E_\theta d^2(\theta, \hat{\theta}).
\]

Here “\( \min \)” and “\( \max \)” are understood to be “\( \inf \)” and “\( \sup \)” respectively if the minimizer or maximizer does not exist.

We first give definitions of \( \epsilon \)-entropies.

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

**Definition 2**: A set \( G_\epsilon \subseteq \overline{S} \) is said to be an \( \epsilon \)-net for \( S \) if for any \( \hat{\theta} \in S \), there exists a \( \theta_0 \in G_\epsilon \) such that \( d(\hat{\theta}, \theta_0) \leq \epsilon \). The logarithm of the minimum cardinality of \( \epsilon \)-nets is called the covering \( \epsilon \)-entropy of \( S \) and is denoted \( V_d(\epsilon) \).
From the definitions, it is clear that $M_d(\epsilon)$ and $V_d(\epsilon)$ are nonincreasing in $\epsilon$. Kolmogorov and Tihomirov (1959) showed that $M_d(\epsilon)$ and $V_d(\epsilon)$ are right continuous when $d$ is a metric. The same proof works to show $M_d(\epsilon)$ is also right continuous for the more general notion of distance $d$ used here.

The above definitions are slight generalizations of the metric entropy notions introduced by Kolmogorov and Tihomirov (1959). In accordance with common terminology, we informally call these $\epsilon$-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 $d_K(\theta, \theta') = D(\theta \| \theta') = \int p_\theta \log (p_\theta / p_{\theta'}) \, d\mu$. Clearly $d_K(\theta, \theta')$ is asymmetric in its two arguments, so it can not be a metric. A presumed shortcoming of this distance is that there is no triangle-like inequality in general, that is, there might not exist a constant $c > 0$ such that $d_K(\theta, \theta') + d_K(\theta, \theta'') \geq c d_K(\theta', \theta'')$ for all $\theta, \theta'$ and $\theta''$ in $S$. The absence of such an inequality can also be an advantage for K-L distance in other regards as we have mentioned. Nevertheless, it is appropriate that conditions enabling a triangle inequality hold at least locally. The following example demonstrates what can happen in an extreme case in the absence of such a condition.

**Example 1:** Consider densities on $[0,1]$ with respect to the Lebesgue measure. Let $S = \{ \theta : 0 \leq \theta \leq \frac{1}{2} \}$ and $p_\theta(x) = 2 I_{\{\theta \leq x \leq \theta + \frac{1}{2}\}}$. Then for any $\epsilon > 0$, the largest $\epsilon$-packing set under $d_K$ must be $S$ itself. Let $\hat{\theta} = I_{\{0 \leq x \leq 1\}}$. Then $D(\theta \| \hat{\theta}) = \log 2$ for all $0 \leq \theta \leq \frac{1}{2}$. Clearly there can not be any triangle-like inequality and the packing number alone can not determine the minimax rate of convergence.

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 $L_q$ metric $d_q(\theta, \theta') = \left( \int |p_\theta - p_{\theta'}|^q \, d\mu \right)^{1/q}$ for $q \geq 1$.

We assume the distance $d$ satisfies the following condition.

**Condition 0:** (local triangle inequality) There exist positive constants $A \leq 1$ and $c_0$ such that for any $\theta, \theta' \in S, \hat{\theta} \in \overline{S}$, if $\max \left( d(\theta, \hat{\theta}), d(\theta', \hat{\theta}) \right) \leq c_0$, then

$$d(\theta, \hat{\theta}) + d(\theta', \hat{\theta}) \geq A d(\theta, \theta').$$

**Remarks:**

1. This condition holds for $d_K$ 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 (see Section 2.3).

2. If $d$ is a metric on $\Theta$, then Condition 0 is always satisfied with $A = 1$ for any $\overline{S} \subseteq \Theta$. 

6
When Condition 0 is satisfied, the packing entropy and covering entropy have the following simple relationship for $\epsilon \leq \epsilon_0$:

$$M_d(2\epsilon/A) \leq V_d(\epsilon) \leq M_d(A\epsilon).$$

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 distance, Hellinger distance and $L_q$ distance. We assume $M_d(\epsilon) < \infty$ for all $\epsilon > 0$. The square root K-L, Hellinger and $L_q$ packing entropies are denoted $M_K(\epsilon)$, $M_H(\epsilon)$ and $M_q(\epsilon)$ respectively.

In subsection 1, we give minimax bounds under global entropy conditions. In subsections 2, 3, and 4, more results are given for $L_2$ risk and K-L risks.

### 2.1 Minimax risk under a global entropy condition

Suppose a good upper bound on the covering $\epsilon$-entropy under the square root K-L distance is available. That is, assume $V_K(\epsilon) \leq V(\epsilon)$. Ideally $V_K(\epsilon)$ and $V(\epsilon)$ are of the same order. Similarly, let $M(\epsilon) \leq M_d(\epsilon)$ be an available lower bound on $\epsilon$-packing entropy with distance $d$, and ideally, $M(\epsilon)$ is of the same order as $M_d(\epsilon)$. Suppose $V(\epsilon)$ and $M(\epsilon)$ are nonincreasing and right continuous functions. To avoid a trivial case (in which $S$ is a small finite set), we assume $M(\epsilon) > 2\log 2$ for $\epsilon$ small enough. Let $\epsilon_n = \inf\{\epsilon > 0 : V(\epsilon) \leq n\epsilon^2\}$ denote what we call the critical covering radius. Then because $V(\epsilon)$ is right continuous, the radius $\epsilon_n$ satisfies

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

The squared radius is the same as the covering entropy divided by the sample size. The trade-off here between $V(\epsilon)/n$ and $\epsilon^2$ is analogous to that between the squared bias and variance of an estimator. As will be shown later, $2\epsilon_n^2$ is an upper bound on the minimax K-L risk. Let $\xi_{n,d}$ be a separation $\epsilon$ such that

$$M(\xi_{n,d}) = 4n\epsilon_n^2 + 2 \log 2.$$

The existence of $\xi_{n,d}$ follows from the right continuity of $M(\epsilon)$ and the assumption that $M(\epsilon) > 2\log 2$ for small $\epsilon$. Roughly, $\xi_{n,d}$ is the separation at which the packing entropy under $d$ distance divided by the sample size $n$ is approximately four times the square of the covering radius. We call $\xi_{n,d}$ the packing separation commensurate with the critical covering radius $\epsilon_n$. This $\xi_{n,d}^2$ determines a lower bound on the minimax risk.
2.1.1 Minimax Lower bound

**Theorem 1:** Suppose Condition 0 is satisfied for the distance \( d \). Then when the sample size \( n \) is large enough such that \( L_{n,d} \leq 2\epsilon_0 \), the minimax risks for estimating \( \theta \in S \) satisfies

\[
\min_{\hat{\theta}} \max_{\theta \in S} P_{\theta} \{ d(\theta, \hat{\theta}) \geq (A/2)L_{n,d} \} \geq 1/2
\]

and consequently

\[
\min_{\hat{\theta}} \max_{\theta \in S} E_{\theta} d^2(\theta, \hat{\theta}) \geq (A^2/8)L_{n,d}^2,
\]

where the minimum is over all estimators mapping from \( X^n \) to \( \overline{S} \).

**Proof:** Let \( N_{s,d} \) be an \( \epsilon_{n,d} \)-packing set with the maximum cardinality in \( S \) under the given distance \( d \) and let \( G_{\epsilon_n} \) be an \( \epsilon_n \)-net for \( S \) under \( d_K \). For any estimator \( \hat{\theta} \) taking values in \( \overline{S} \), define \( \overline{\theta} = \arg \min_{\theta \in N_{s,d}} d(\theta', \hat{\theta}) \) (if there are more than one minimizer, choose any one), so that \( \overline{\theta} \) takes values in the finite packing set \( N_{s,d} \). Let \( \theta \) be any point in \( N_{s,d} \). If \( d(\theta, \hat{\theta}) < A_{s,d}/2 \), then \( \max \left( d(\theta, \hat{\theta}), d(\overline{\theta}, \hat{\theta}) \right) < A_{s,d}/2 \leq \epsilon_0 \) and hence by Condition 0, \( d(\theta, \hat{\theta}) + d(\overline{\theta}, \hat{\theta}) \geq A d(\theta, \overline{\theta}) \), which is at least \( A_{s,d} \) if \( \theta \neq \overline{\theta} \). Thus if \( \theta \neq \overline{\theta} \), we must have \( d(\theta, \hat{\theta}) \geq A_{s,d}/2 \), and

\[
\min_{\hat{\theta}} \max_{\theta \in S} P_{\theta} \{ d(\theta, \hat{\theta}) \geq (A/2)L_{n,d} \} \geq \min_{\hat{\theta}} \max_{\theta \in N_{s,d}} P_{\theta} \{ d(\theta, \hat{\theta}) \geq (A/2)L_{n,d} \}
\]

\[
= \min_{\hat{\theta}} \max_{\theta \in N_{s,d}} P_{\theta} \left( \theta \neq \overline{\theta} \right)
\]

\[
\geq \min_{\hat{\theta}} \sum_{\theta \in N_{s,d}} w(\theta) P_{\theta} \left( \theta \neq \overline{\theta} \right)
\]

\[
= \min_{\hat{\theta}} P_w \left( \theta \neq \overline{\theta} \right),
\]

where in the last line, \( \theta \) is randomly drawn according to a discrete prior probability \( w \) restricted to \( N_{s,d} \), and \( P_w \) denotes the Bayes average probability with respect to the prior \( w \). Moreover, since \( (d(\theta, \hat{\theta}))^\ell \) is not less than \(( (A/2)L_{n,d})^\ell 1_{\{\theta \neq \overline{\theta}\}} \), taking the expected value it follows that for all \( \ell > 0 \),

\[
\min_{\hat{\theta}} \max_{\theta \in S} E_{\theta} d^\ell(\theta, \hat{\theta}) \geq ((A/2)L_{n,d})^\ell \min_{\hat{\theta}} P_w \left( \theta \neq \overline{\theta} \right).
\]

By Fano’s inequality (see e.g., Fano (1961), Ash (1965, page 80) or Cover and Thomas (1991, pages 39 and 205)), with \( w \) being the discrete uniform prior on \( \theta \) in the packing set \( N_{s,d} \), we have

\[
P_w \left( \theta \neq \overline{\theta} \right) \geq 1 - \frac{I(\Theta; X^n) + \log 2}{\log |N_{s,d}|},
\]

where \( I(\Theta; X^n) \) is Shannon’s mutual information between the random parameter and the random sample, when \( \theta \) 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 distance between the product densities \( p(x^n|\theta) \) and any joint density \( q(x^n) \) on the sample space \( X^n \). Indeed,

\[
I(\Theta;X^n) = \sum_\theta w(\theta) \int p(x^n|\theta) \log(p(x^n|\theta)/p_w(x^n)) \mu(dx^n) \\
\leq \sum_\theta w(\theta) \int p(x^n|\theta) \log(p(x^n|\theta)/q(x^n)) \mu(dx^n) \\
\leq \max_{\theta \in N_{n,d}} D\left( P_{X^n|\theta} \| Q_{X^n} \right).
\]

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 \( N_{n,d} \). Now choose \( w_1 \) to be the uniform prior on \( G_{e_n} \) 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 \( G_{e_n} \) is an \( e_n \)-net in \( S \) under \( d_K \), for each \( \theta \in S \), there exists \( \tilde{\theta} \in G_{e_n} \) such that \( D(p_{\theta} \| p_{\tilde{\theta}}) = d_K^2(\theta, \tilde{\theta}) \leq e_n^2 \).

Also by definition, \( \log |G_{e_n}| \leq V_K(e_n) \). It follows that

\[
D\left( P_{X^n|\theta} \| Q_{X^n} \right) = E \log \left( \frac{p(X^n|\theta)}{\left(1/|G_{e_n}| \sum_{\theta' \in G_{e_n}} p(X^n|\theta') \right)} \right) \\
\leq E \log \left( \frac{p(X^n|\theta)}{\left|G_{e_n}\right| p(X^n|\theta)} \right) \\
= \log |G_{e_n}| + D\left( P_{X^n|\theta} \| P_{X^n|\tilde{\theta}} \right) \\
\leq V(e_n) + n e_n^2.
\]

Thus, by our choice of \( L_{n,d} \),

\[
\frac{I(\Theta;X^n) + \log 2}{\log |N_{n,d}|} \leq \frac{1}{2}.
\]

The conclusion follows. This completes the proof of Theorem 1.

Remarks:

1. Up to the point (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 \( I(\Theta;X^n) \leq n I(\Theta;X_1) \) or \( I(\Theta;X^n) \leq n \max_{\theta,\theta' \in \Theta} D(P_{X_1|\theta} \| P_{X_1|\theta'}) \) (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.

2. 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 \( C_n = \min_{Q_{X^n}} \max_{\theta \in S} D(P_{X^n|\theta} \| Q_{X^n}) = \max_{\theta} I_w(\theta;X^n) \), where the maximum is over priors supported on \( S \). The universal data compression interpretations of these quantities can be found in Davisson (1973) and
Davison and Leon-Garcia (1980) [see Clarke and Barron (1994), Yu (1996), Haussler and Opper (1995) and Haussler (1995) for some of the recent work in that area]. The bound 
\[ D(P_{X^n|\theta} \parallel P_{X^n}) \leq V(\epsilon_n) + nc_n^2 \] has roots in Barron (1987, pp. 89), where it is given in a more general form for arbitrary priors, that is 
\[ D(P_{X^n|\theta} \parallel P_{X^n}) \leq \log \frac{1}{w(N_{\theta, \epsilon})} + nc_n^2, \]
where \( N_{\theta, \epsilon} = \{ \theta : D(p_\theta \parallel p_{\theta'}) \leq \epsilon^2 \} \) and \( P_{X^n} \) has density \( p_w(x^n) = \int p(x^n|\theta) w(d\theta) \).
The redundancy bound \( V(\epsilon_n) + nc_n^2 \) can also be obtained from use of a two stage code of length 
\[ \log |G_{\epsilon_n}| + \min_{\theta' \in G_{\epsilon_n}} \log 1/p(x^n|\theta'), \]
see Barron and Cover (1991, Section V).

3. From inequality (1), the minimax risk is bounded below by a constant times 
\[ \xi_{n,d}(1 - (C_n + \log 2)/K_n), \]
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 \( K_n = \log |N_{\xi, d}| \) is the Kolmogorov \( \xi_{n,d} \)-capacity of \( \Theta \). Thus 
\[ \xi_{n,d} \] 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).

When K-L distance is lower bounded by a multiple of a suitable distance \( d \), then a minimax lower bound on the K-L risk is obtained. That is, if there exists a constant \( A_0 \) such that 
\[ A_0 d^2(\theta, \theta') \leq d_K^2(\theta, \theta') \] for any \( \theta, \theta' \in \overline{S} \) with \( d_K(\theta, \theta') \leq \epsilon_0 \), and if \( d \) satisfies condition 0, then when \( n \) is suitably large (so that \( (A_0 A^2/8) \xi_{n,d} \leq \epsilon_0^2 \)),
\[ \min_{\hat{\theta} \in A_n} \max_{\theta \in S} E_\theta d_K^2(\theta, \hat{\theta}) \leq (A_0 A^2/8) \xi_{n,d}. \]
A natural choice for \( d \) is the Hellinger distance \( d_H \). (Since Hellinger distance does satisfy the triangle inequality between densities and \( d_H(\theta, \theta') \leq d_K(\theta, \theta') \) for all \( \theta, \theta' \); moreover, locally square root K-L distance behaves like Hellinger for bounded log-density ratios.) Let \( \xi_{n,K} \) and \( \xi_{n,H} \) be the packing separations commensurate with the critical covering radius \( \epsilon_n \) under \( d_K \) and \( d_H \), determined by \( M_K(\epsilon_{n,K}) = 4n c_n^2 + 2 \log 2 \) and \( M_H(\epsilon_{n,H}) = 4n c_n^2 + 2 \log 2 \), respectively, with \( V(\epsilon_n) = nc_n^2 \). We have the following corollary.

**Corollary 1:** If there exist constants \( A \) and \( \epsilon_0 > 0 \) such that \( d_K \) satisfies Condition 0 on \( S, \overline{S} \), then when \( \xi_{n,K} \leq 2\epsilon_0 \),
\[ \min_{\hat{\theta} \in A_n} \max_{\theta \in S} E_\theta d_K^2(\theta, \hat{\theta}) \geq (A^2/8) \xi_{n,K}^2. \]
For the square Hellinger risk, we have
\[ \min_{\hat{\theta} \in A_n} \max_{\theta \in S} E_\theta d_H^2(\theta, \hat{\theta}) \geq (1/8) \xi_{n,H}^2. \]

Note in the first conclusion, \( \xi_{n,K}^2 \) is determined by packing entropy \( M_K(\epsilon) \) only (with the choice \( V(\epsilon) = M_K(\epsilon) \)). However, for a general distance \( d \) (specifically the Hellinger distance
in Corollary 1), $\mathcal{L}_{n,d}^2$ is determined by two quantities: both $M_H(\epsilon)$ and $V_K(\epsilon)$ without any assumption on the relationship between the distances.

When $d_K$ is locally upper bounded by a multiple of $d$, the minimax risk bound for $d^2$ can be expressed exclusively in terms of packing entropy under $d$ distance, yielding the following corollary.

**Corollary 2:** Assume Condition 0 is satisfied for distance $d$ and there exists a constant $A$ such that $D(p_\theta||p_{\theta'}) \leq A d^2(\theta, \theta')$ for any $\theta, \theta' \in S$ with $d(\theta, \theta') \leq \epsilon_{n,d}/\sqrt{A}$, where $\epsilon_{n,d}$ is determined from $M_d\left(\sqrt{\epsilon_{n,d}}/\sqrt{A}\right) = n\epsilon_{n,d}^2$. Let $\epsilon_{n,d}$ be chosen such that $M_d(\epsilon_{n,d}) = 4n\epsilon_{n,d}^2 + 2 \log 2$. Then, if $\epsilon_{n,d} \leq 2\epsilon_0$, we have

$$\min_{\theta \in \mathcal{A}_n} \max_{\theta' \in S} E_{\theta'} d^2(\theta, \theta') \geq A^2 \epsilon_{n,d}^2/8.$$  

**Proof:** Under the assumption between distances $d$ and $d_K$, the largest $\sqrt{\epsilon_{n,d}}/\sqrt{A}$-packing set in $S$ under $d$ also serves as an $\epsilon_{n,d}$-covering set for $S$ under $d_K$. Thus when $\epsilon \leq \epsilon_{n,d}$, $V_K(\epsilon) \leq M_d\left(\sqrt{\epsilon}/\sqrt{A}\right)$. The result follows from Theorem 1.

For applications, the lower bounds above may be applied to a subclass of densities $\{p_\theta : \theta \in S_0\}$ ($S_0 \subset S$) which is 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 the densities $\{p_\theta : \theta \in S_0\}$ have support on a compact space and $\|\log p_\theta\|_\infty \leq T$ for all $\theta \in S_0$, then the square root K-L distance, 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.

### 2.1.2 Upper bound

To provide an upper bound on the minimax rate of convergence, we construct an estimator as follows. The technique applies in general to bound the cumulative Kullback-Leibler risk for sequences of Bayes estimates of densities. To identify the minimax rate, we use a uniform prior on an $\epsilon$-net. Consider the $\epsilon_n$-net $G_{\epsilon_n}$ for $S$ under $d_K$ and the uniform prior $w_1$ on $G_{\epsilon_n}$. For $n = 1, 2, ...$, let

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

be the corresponding mixture density. Let

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

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 $\frac{p(X_{i+1}|X^i)}{p(X^i)}$ for $i > 0$ and
\( \hat{p}_i(x) = p(x) = \left( \frac{1}{G_{G_n}} \right) \sum_{\theta \in G_n} p(x|\theta) \) for \( i = 0 \). Then by convexity and the chain rule (as in Barron (1987)),

\[
E_\theta D(p_\theta || \overline{p}) \leq E_\theta \left( n^{-1} \sum_{i=0}^{n-1} D(P_{X_{i+1}|\theta} || P_{X_{i+1}|X_i}) \right)
= n^{-1} \sum_{i=0}^{n-1} E \log \frac{P(X_{i+1}|\theta)}{p(X_{i+1}|X_i)}
= n^{-1} E \log \frac{p(X_{i+1}|\theta)}{p(X_{i+1})}
= n^{-1} D \left( P_{X^n|\theta} || P_{X^n} \right)
\leq n^{-1} \left( V(\epsilon_n) + n\epsilon_n^2 \right) = 2\epsilon_n^2,
\]

where the last inequality is as derived as in equation (2). Combining this upper bound with the lower bound from section 2.1.1 we have the following result.

**Theorem 2:** Let \( V(\epsilon) \) upper bound the covering entropy of \( S \) under \( d_K \) and let \( \epsilon_n \) satisfy \( V(\epsilon_n) = n\epsilon_n^2 \). Then

\[
\min_{\theta} \max_{\overline{p} \in \mathcal{P}} E_\theta D(p_\theta || \overline{p}) \leq 2\epsilon_n^2,
\]

where the minimization is over all density estimators. Moreover, if Condition 0 is satisfied for a distance \( d \) and \( A_0 d^2(\theta, \theta') \leq d^2_K(\theta, \theta') \) for all \( \theta, \theta' \in \overline{S} \), and if the set \( \mathcal{A}_n \) of allowed estimators (mappings from \( X^n \) to \( \overline{S} \)) contains \( \overline{p} \) constructed above, then when \( \epsilon_{n,d} \leq \epsilon_0 \),

\[
(A_0 A/8) \mathcal{E}_{\mathcal{A}_n,d}^2 \leq A_0 \min_{\theta} \max_{\overline{p}} E_\theta d^2(\theta, \hat{\theta}) \leq \min_{\theta} \max_{\overline{p}} E_\theta d^2_K(\theta, \hat{\theta}) \leq 2\epsilon_n^2.
\]

The condition that \( \mathcal{A}_n \) contains \( \overline{p} \) in Theorem 2 is satisfied if \( \{ p_\theta : \theta \in \overline{S} \} \) is convex. In particular, this holds if the action space \( \overline{S} \) is the set of all densities on \( \mathcal{X} \). In which case, when \( d \) is a metric the only remaining condition needed for the second set of inequalities is \( A_0 d^2(\theta, \theta') \leq d^2_K(\theta, \theta') \) for all \( \theta, \theta' \). 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 Kullback-Leibler distance, 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 (see the remark after Lemma 3 in subsection 2.3).

In obtaining the minimax risk bounds the quantity \( D \left( P_{X^n|\theta} || P_{X^n} \right) \) plays an important role. Indeed the derivation of both upper and lower bounds uses uniform upper bounds on this quantity. For the lower bound, it is used to bound \( I(\Theta; X^n) \), and for the upper bound, it bounds the risk of a specific estimator.

Averages with respect to uniform priors on nets of two different radius choices are used in these bounds. As we shall see, asymptotically these two radii typically have the same rate.

If \( \mathcal{E}_{n,d}^2 \) and \( 2\epsilon_n^2 \) converge to 0 at the same rate, then the minimax rate of convergence is identified by Theorem 2. For \( \mathcal{E}_{n,d}^2 \) and \( \epsilon_n^2 \) to be of the same order, it is sufficient that the following two conditions hold (for a proof, see Lemma 8 in the appendix).
**Condition 1:** (metric entropy equivalence) There exist positive constants $a$, $b$ and $c$ such that when $\epsilon$ is small enough,

$$M(\epsilon) \leq V(b\epsilon) \leq cM(ac).$$

**Condition 2:** (richness of the function class) For some $0 < \alpha < 1$,

$$\lim_{\epsilon \to 0} \frac{M(\alpha\epsilon)}{M(\epsilon)} > 1.$$

Condition 1 is the equivalence of the entropy structure under the square root K-L distance and under $d$ distance when $\epsilon$ is small. The equivalence of entropy structure 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, $M(\epsilon)$ approaches $\infty$ at least polynomially fast in $\frac{1}{\epsilon}$ as $\epsilon \to 0$, i.e., there exists a constant $\delta > 0$ such that $M(\epsilon) \geq \left(\frac{1}{\epsilon}\right)^{\delta}$. This condition is typical of nonparametric function classes. It is satisfied in particular, if $M(\epsilon)$ can be expressed as $M(\epsilon) = e^{-r}\kappa(\epsilon)$, where $r > 0$ and $\kappa(\alpha\epsilon)/\kappa(\epsilon) \to 1$ as $\epsilon \to 0$. In most situations, the metric entropies are known only up to orders, which makes it generally not tractable to check $\lim_{\epsilon \to 0} \frac{M(\alpha\epsilon)}{M(\epsilon)} > 1$ for the exact packing entropy function. That is why Condition 2 is stated in terms of a presumed known bound $M(\epsilon) \leq M_d(\epsilon)$.

Both conditions 1 and 2 are satisfied if, for instance, $M(\epsilon) \asymp V(\epsilon) \asymp e^{-r}\kappa(\epsilon)$ with $r$ and $\kappa(\epsilon)$ as mentioned above.

**Corollary 3:** Assume Condition 0 is satisfied for a distance $d$ satisfying $A_0d^2(\theta, \theta') \leq d^2_K(\theta, \theta')$ in $\mathcal{S}$ and assume $\{p_\theta : \theta \in \mathcal{S}\}$ is convex. Under Conditions 1 and 2, we have

$$\min_{\theta \in \mathcal{A}_n} \max_{\theta' \in \mathcal{S}} E_{\theta} d^2(\theta, \hat{\theta}) \asymp \epsilon_n^2,$$

where $\epsilon_n$ is determined by the equation $M_d(\epsilon_n) = n\epsilon_n^2$. In particular, if $\sup_{\theta \in \mathcal{S}} \|\log p_{\hat{\theta}}\| < \infty$ and Condition 2 is satisfied, then

$$\min_{\theta \in \mathcal{A}_n} \max_{\theta' \in \mathcal{S}} E_{\theta} d^2_K(\theta, \hat{\theta}) \asymp \min_{\theta \in \mathcal{A}_n} \max_{\theta' \in \mathcal{S}} E_{\theta} d^2_H(\theta, \hat{\theta}) \asymp \min_{\theta \in \mathcal{A}_n} \max_{\theta' \in \mathcal{S}} E_{\theta} \|p_{\theta} - p_{\hat{\theta}}\|_2^2 \asymp \epsilon_n^2,$$

where $\epsilon_n$ satisfies $M_2(\epsilon_n) = n\epsilon_n^2$ or $M_H(\epsilon_n) = n\epsilon_n^2$.

Corollary 3 is applicable for many smooth nonparametric classes as we shall see. However, for not very rich classes of densities (for example, 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 $M_K(\epsilon)$ and $M_H(\epsilon)$ may be of
order \( \log(1/\varepsilon)^m \) for some constant \( m \geq 1 \), and then \( \varepsilon_n \) and \( \mathcal{E}_{n,H} \) are not of the same order with \( \varepsilon_n \asymp \sqrt{\left(\log n\right)/n} \) and \( \mathcal{E}_{n,H} = o(1/\sqrt{n}) \). For smooth finite-dimensional models, the minimax risk can be solved using some traditional statistical methods (such as Bayes procedures, Cramer-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 8).

2.2 Minimax rates under \( L_2 \) loss

For general classes of densities, the assumption of upper boundedness of square root K-L distance by a multiple of \( d \) distance for the whole density class in Corollary 2 may not hold. Theorem 1 is applicable but the resulting minimax lower bounds involve metric entropies under both \( d_K \) and \( d \). In this subsection, we derive minimax bounds for \( L_2 \) risk without appealing to K-L covering entropy.

Let \( \mathcal{F} \) be a class of density functions \( f \) with respect to a finite measure \( \mu \) on a measurable set \( \mathcal{X} \) such as \([0,1]\). (Typically \( \mathcal{X} \) will be taken to be a compact set, though it need not be; we assume only that the dominating measure \( \mu \) is a finite.) We normalize \( \mu \) to be a probability measure. Let the packing entropy of \( \mathcal{F} \) be \( M_q(\varepsilon) \) under the \( L_q \) 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 \( X_1, X_2, ..., X_n \) from \( f \), let \( Y_1, Y_2, ..., Y_n \) be a sample generated i.i.d. from the uniform distribution on \( \mathcal{X} \) with respect to \( \mu \) (generated independently of \( X_1, ..., X_n \)). Let \( Z_i \) be \( X_i \) or \( Y_i \) with probability \((1/2, 1/2)\) according to the outcome of \( \text{Bernoulli}(1/2) \) random variables \( V_i \) generated independently for \( i = 1, ..., n \). 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 \( \tilde{\mathcal{F}} = \{ g : g = (f + 1)/2, f \in \mathcal{F} \} \) be the new density class.

**Lemma 1:** The minimax \( L_2 \) risks of the two classes \( \mathcal{F} \) and \( \tilde{\mathcal{F}} \) have the following relationship,

\[
\min_{f} \max_{\hat{f} \in \tilde{\mathcal{F}}} \left\| f - \hat{f} \right\|_2^2 \leq 4 \min_{g} \max_{\hat{g} \in \tilde{\mathcal{F}}} \left\| g - \hat{g} \right\|_2^2,
\]

where the minimization on the left hand side is over all estimators based on \( X_1, ..., X_n \) and the minimization on the right hand side is over all estimators based on \( n \) independent observations from \( g \). Generally, for \( q \geq 1 \), we have

\[
\min_{f} \max_{\hat{f} \in \tilde{\mathcal{F}}} \left\| f - \hat{f} \right\|_q^q \leq A^q \min_{g} \max_{\hat{g} \in \tilde{\mathcal{F}}} \left\| g - \hat{g} \right\|_q^q.
\]

14
Proof: We focus on the proof of the assertion for the \( L_2 \) case. The proof for general \( L_q \) is similar.

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 \( \bar{g} \) be any density estimator of \( g \) based on \( Z_i, i = 1,...n \). Let \( \hat{g} \) be the density that minimizes \( \| h - \bar{g} \|^2 \) over functions in the set \( \{ h : h(x) \geq 1/2, \int h(x) d\mu = 1 \} \). Then from Lemma 9 in Appendix, \( \| g - \hat{g} \|^2 \leq \| g - \bar{g} \|^2 \). (For general \( L_q \) norm, by the triangle inequality and because \( g \in \{ h : h(x) \geq 1/2, \int h(x) d\mu = 1 \} \), \( \| g - \hat{g} \|^q \leq 2^{q-1} \| g - \bar{g} \|^q + 2^{q-1} \| \hat{g} - g \|^q \leq 2^q \| g - \bar{g} \|^2 \).) Now 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 \( X_1, ..., X_n, Y_1, ..., Y_n \) and the outcomes of the coin flips \( V_1, ..., V_n \). So it is a randomized estimator. The squared \( L_2 \) loss of \( \hat{f}_{\text{rand}} \) is bounded as follows:

\[
\int \left( f(x) - \hat{f}_{\text{rand}}(x) \right)^2 d\mu = \int (2g(x) - 2\hat{g}(x))^2 d\mu
\]

\[
= 4 \int (g - \hat{g})^2 d\mu
\]

\[
\leq 4 \| g - \bar{g} \|^2.
\]

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

\[
EX^n \| f - \hat{f} \|^2 \leq EX^n \| f - EY^n,V^n\hat{f}_{\text{rand}} \|^2
\]

\[
\leq EX^n EY^n,V^n \| f - \hat{f}_{\text{rand}} \|^2
\]

\[
= EZ^n \| f - \hat{f}_{\text{rand}} \|^2
\]

\[
\leq 4EZ^n \| g - \bar{g} \|^2.
\]

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

\[
\max_{\hat{f} \in \mathcal{F}} EX^n \| f - \hat{f} \|^2 \leq 4 \max_{g \in \mathcal{F}} EZ^n \| g - \bar{g} \|^2.
\]

Taking the minimum over estimators \( \bar{g} \) completes the proof of Lemma 1.

Thus the minimax risk of the original problem is upper bounded by the minimax risk on \( \mathcal{F} \). Moreover, the \( c \)-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 \( \mathcal{F} \), the \( c \)-packing entropy under \( L_2 \) is \( \tilde{M}_2(c) = M_2(2c) \).
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 distance is upper bounded by a multiples of $L_2$ distance. Indeed, for densities $g_1, g_2 \in \tilde{\mathcal{F}}$,

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

where the first inequality is the familiar relationship between K-L distance and chi-square distance, and the second inequality follows because $g_1$ is lower bounded by 1/2. Let $\tilde{V}_K(\epsilon)$ denote the $d_K$ covering entropy of $\tilde{\mathcal{F}}$. Then $\tilde{V}_K(\epsilon) \leq \tilde{M}_2(\epsilon/\sqrt{2}) = M_2(\sqrt{2}\epsilon)$. Let $\epsilon_n$ be chosen such that $M_2(\sqrt{2}\epsilon_n) = n\epsilon_n^2$. From Theorem 2, there exists a density estimator $\hat{g}_0$ such that $\max_{g \in \tilde{\mathcal{F}}} E_{Z^n} D(g \parallel \hat{g}_0) \leq 2\epsilon_n^2$. It follows that $\max_{g \in \mathcal{F}} E_{Z^n} d_H^2(g, \hat{g}_0) \leq 2\epsilon_n^2$, and $\max_{g \in \mathcal{F}} E_{Z^n} \parallel \hat{g}_0 \parallel_2^2 \leq 8\epsilon_n^2$. Consequently, by Lemma 1,

$$\min_{f} \max_{\hat{f} \in \mathcal{F}} E_{X^n} \parallel f - \hat{f} \parallel_1 \leq 8\sqrt{8}\epsilon_n.$$

To get a good estimator in terms of $L_2$ risk, we assume $\sup_{f \in \mathcal{F}} \parallel f \parallel_\infty \leq L < \infty$. Let $\hat{g}$ be the density in $\tilde{\mathcal{F}}$ that is closest to $\hat{g}_0$ in Hellinger distance. Then by triangle inequality,

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

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

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

Thus $\max_{g \in \mathcal{F}} E_{Z^n} \parallel g - \hat{g} \parallel_2^2 \leq 16(L+1)\epsilon_n^2$. Using Lemma 1 again, we have an upper bound on the minimax squared $L_2$ risk.

**Theorem 3:** Let $M_2(\epsilon)$ be the $L_2$ packing entropy of a density class $\mathcal{F}$ with respect to a probability measure. Let $\epsilon_n$ satisfy $M_2(\sqrt{2}\epsilon_n) = n\epsilon_n^2$. Then

$$\min_{f} \max_{\hat{f} \in \mathcal{F}} E_{X^n} \parallel f - \hat{f} \parallel_1 \leq 8\sqrt{8}\epsilon_n.$$

If in addition, $\sup_{f \in \mathcal{F}} \parallel f \parallel_\infty \leq L < \infty$, then

$$\min_{f} \max_{\hat{f} \in \mathcal{F}} E_{f} \parallel f - \hat{f} \parallel_2^2 \leq 256(L+1)\epsilon_n^2.$$
The above result upper bounds the minimax $L_1$ risk and $L_2$ risk (under $\sup_{f \in \mathcal{F}} \| f \|_\infty < \infty$ for $L_2$) using only the $L_2$ metric entropy.

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

$$\min \max_{f, \hat{f} \in \mathcal{F}} E_{X^n} \| f - \hat{f} \|_q^2 \leq \epsilon_n^2, \text{ for } 1 \leq q \leq 2.$$ 

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 $f^* \in \mathcal{F}$ with $\min_{x \in \mathcal{X}} f^*(x) = C > 0$ and a positive constant $\alpha \in (0, 1)$ such that $\mathcal{F}_0 = \{(1 - \alpha) f^* + \alpha g : g \in \mathcal{F}\} \subset \mathcal{F}$.

For a convex class of densities, Condition 3 is satisfied if there is at least one density bounded away from zero.

**Lemma 2:** Under Condition 3, the subclass $\mathcal{F}_0$ has $L_2$ packing entropy $M_2^0(\epsilon) = M_2(\epsilon/\alpha)$.

**Proof:** Because $\|(1 - \alpha) f^* + \alpha g_1\| - \|(1 - \alpha) f^* + \alpha g_2\| = \alpha \| g_1 - g_2 \|$, an $\epsilon$-packing set in $\mathcal{F}$ corresponds to an $\alpha \epsilon$-packing set in $\mathcal{F}_0$ and vice versa.

Under Condition 3, for two densities $f_1$ and $f_2$ in $\mathcal{F}_0$,

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

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

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

$$\min \max_{f, \hat{f} \in \mathcal{F}} E_{X^n} \| f - \hat{f} \|_2^2 \geq \epsilon_n^2/8.$$ 

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

1. if $M_2(\epsilon) \asymp M_1(\epsilon)$ holds,

   $$\min \max_{f, \hat{f} \in \mathcal{F}} E_{X^n} \| f - \hat{f} \|_1 \asymp \epsilon_n.$$ 

2. if $\sup_{f \in \mathcal{F}} \| f \|_\infty < \infty$,

   $$\min \max_{f, \hat{f} \in \mathcal{F}} E_{X^n} \| f - \hat{f} \|_2^2 \asymp \epsilon_n^2.$$ 

17
Using the relationship between $L_2$ and $L_q$ ($1 \leq q < 2$) distances and applying Theorem 1, we have the following corollary.

**Corollary 4:** Suppose $\mathcal{F}$ 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 $\xi_n$ satisfy $M_q(\xi_n) = n\epsilon_n^q$. Then

$$
\xi_n^2 \leq \min_{\hat{f}} \max_{f \in \mathcal{F}} \| E_{X^n} \| f - \hat{f} \|_q^2 \leq \epsilon_n^2.
$$

If the packing entropies under $L_2$ and $L_q$ are equivalent (which is the case for many familiar nonparametric classes, see Section 7 for examples), then the above upper and lower bounds converge at the same rate. Generally for a uniformly upper bounded density class $\mathcal{F}$ on a compact set, because $\int (f - g)^2 d\mu \leq (\| f + g \|_\infty)^2 \int |f - g| d\mu$, we know $M_1(\epsilon) \leq M_2(\epsilon) \leq M_1(\frac{\epsilon^2}{\sup_{p} \| f \|_\infty})$. Then the corresponding lower bound for $L_1$ risk may vary from $\epsilon_n$ to $\epsilon_n^2$ depending on how different the two entropies are (see also Birgé (1986)).

### 2.3 Minimax rates under K-L loss

For the square root K-L distance, Condition 0 is not necessarily satisfied for general classes of densities. We next discuss Condition 0 for $d_K$ and present some more results concerning the K-L risk.

**Lemma 3:** Assume $D(p_\theta \| p_{\theta'}) \leq \overline{d}^2(\theta, \theta')$ for $\theta, \theta' \in S$ with $d(\theta, \theta') \leq \epsilon^*$ and $D(p_\theta, p_{\theta'}) \geq A_0d^2(\theta, \theta')$ for all $\theta, \theta' \in \Theta$, where $d$ is a metric on $\Theta$. Then Condition 0 is satisfied for $d_K$ with $A = \sqrt{\overline{A}/A_0}$ and $\epsilon_0 = \sqrt{A_0} \epsilon^*/2$ for any choice of $\overline{\Theta} \subseteq \Theta$.

The conditions in the lemma are satisfied by regression families as considered in Section 4.

As we have discussed the conditions are also satisfied when the log-densities are uniformly bounded for $\theta \in S$, so the conclusion of the lemma then shows that we are permitted to take the action space $\overline{\Theta}$ to consist of all densities, so that Theorem 2 is applicable to the square-root K-L as well as Hellinger distance.

**Proof of Lemma 3:** Assume $\max\{d_K(\theta, \overline{\theta}), d_K(\theta', \overline{\theta})\} \leq \epsilon_0$. Then $\max\{d(\theta, \overline{\theta}), d(\theta', \overline{\theta})\} \leq \epsilon_0/\sqrt{A_0}$. Since $d$ is a metric, $d(\theta, \theta') \leq d(\theta, \overline{\theta}) + d(\theta', \overline{\theta}) \leq 2\epsilon_0/\sqrt{A_0} = \epsilon^*$. Thus $d_K(\theta, \theta') \leq \sqrt{A} \left( d(\theta, \overline{\theta}) + d(\theta', \overline{\theta}) \right) \leq \sqrt{A/A_0} \left( d_K(\theta, \overline{\theta}) + d_K(\theta', \overline{\theta}) \right)$.

**Lemma 4:** For the square root K-L distance $d_K$, either of the following two equivalent conditions is sufficient for the satisfaction of Condition 0 with $0 < A < 1$ and $\epsilon_0 = \epsilon_*/\sqrt{2} > 0$.

1. $D(p_\theta \| (p_\theta + p_{\theta'})/2) + D(p_{\theta'} \| (p_\theta + p_{\theta'})/2) \geq A^2 D(p_\theta \| p_{\theta'})$ for all $\theta, \theta' \in S$ with $\max\{D(p_\theta \| (p_\theta + p_{\theta'})/2), D(p_{\theta'} \| (p_\theta + p_{\theta'})/2)\} \leq \epsilon_*^2$.
2. \( D\left((p_\theta + p_{\theta'})/2 \parallel p_\theta\right) \leq (1/2)(A^2 - 1)D\left((p_\theta \parallel (p_\theta + p_{\theta'})/2)\right) \) for all \( \theta, \theta' \in S \) with
\[
\max\{D\left((p_\theta \parallel (p_\theta + p_{\theta'})/2)\right), D\left((p_{\theta'} \parallel (p_\theta + p_{\theta'})/2)\right)\} \leq \epsilon^2.
\]

If \( \{p_\theta : \theta \in S\} \) is a convex family (so that \( (p_\theta + p_{\theta'})/2 = p_{\tilde{\theta}} \) for some \( \tilde{\theta} \in S \)), then the conditions are necessary as well as sufficient.

When \( \{p_\theta : \theta \in \tilde{S}\} \) is convex, then from the above lemma, a sufficient condition for the satisfaction of Condition 0 is that there exist constants \( C > 1 \) and \( \epsilon_0 > 0 \) such that \( D(p_\theta \parallel p_{\theta'}) \leq CD(p_{\theta'} \parallel p_\theta) \) for any \( \theta, \theta' \in \tilde{S} \) with \( \max\{D(p_\theta \parallel p_{\theta'}), D(p_{\theta'} \parallel p_\theta)\} \leq \epsilon_0^2 \).

**Corollary 5:** Suppose the conditions in Lemma 3 or Lemma 4 are satisfied. Let \( \epsilon_n \) and \( \varepsilon_n \) be determined by \( M_K(\epsilon_n) = n\epsilon_n^2 \) and \( M_K(\varepsilon_n) = 4n\epsilon_n^2 + 2\log 2 \). Then when \( \varepsilon_n \leq 2\epsilon_0 \),
\[
(A^2/8)\epsilon_n^2 \leq \min_{\theta \in A_n} \max_{\tilde{\theta} \in \tilde{S}} E_{\theta} d_K^2(\theta, \tilde{\theta}) \leq 2\epsilon_n^2.
\]

If also \( \lim_{\alpha \to 0} \frac{M_K(\alpha)}{M_K(\varepsilon)} > 0 \) for some \( 0 < \alpha < 1 \) (that is, \( S \) is rich using the distance \( d_K \)), then \( \epsilon_n \) and \( \varepsilon_n \) are of the same order and
\[
\min_{\theta \in A_n} \max_{\tilde{\theta} \in \tilde{S}} E_{\theta} d_K^2(\theta, \tilde{\theta}) \approx \epsilon_n^2.
\]

As mentioned before, the conditions in both lemmas are satisfied if the log-densities in the class are uniformly bounded. When these conditions are not satisfied (for instance, if the densities in the class have different supports the conditions in Lemma 3 cannot be satisfied), the following result provides a minimax lower bound involving only the Hellinger metric entropy.

We consider estimating a density defined on \( \mathcal{X} \) with respect to a measure \( \mu \) with \( \mu(\mathcal{X}) = 1 \).

**Lemma 5:** Assume for a density \( f \), that \( ||f||_{\infty} \leq T \). Then if for a density \( g \), \( d_H(f, g) \leq \epsilon \) for some \( 0 \leq \epsilon \leq \sqrt{2} \), there exists a density \( \tilde{g} \) on \( \mathcal{X} \) depending only on \( g, T \) and \( \epsilon \) (but not on \( f \)) such that
\[
D(f \parallel \tilde{g}) \leq 2 \left( 2 + \log \left( 9T/4 \epsilon^2 \right) \right) \left( 9 + 8(8T - 1)^2 \right) \epsilon^2.
\]

Bounds analogous to Lemma 5 are in Barron, Birgé and Massart (1995, Proposition 1), Wong and Shen (1995, Theorem 5).

For classes whose metric entropy structure is known under the Hellinger distance but hard to know under K-L distance, the lemma is useful to give a bound on the covering entropy under K-L distance.

For a density class \( \mathcal{F} \) for which \( ||f||_{\infty} \leq T \) for each \( f \in \mathcal{F} \), let \( M_H(\epsilon) \) be the packing entropy under \( d_H \). By the lemma, an \( \epsilon \)-net under \( d_H \) can always result in an \( \eta \)-net under \( d_K \), where \( \eta^2 = 2 \left( 2 + \log \left( 9T/4 \epsilon^2 \right) \right) \left( 9 + 8(8T - 1)^2 \right) \epsilon^2 \), which is not greater than \( T_1^2 \epsilon^2 \log^2 \left( 2/\epsilon \right) \) for \( 0 < \epsilon \leq \sqrt{2} \) with \( T_1 \) being a constant depending only on \( T \). Thus for \( \epsilon \geq \sqrt{(\log 2)/n} \), we have
\[ V_K \left( (T_1/2) \epsilon \log (4n/\log 2) \right) \leq M_H(\epsilon) \text{ or equivalently, } V_K(\epsilon) \leq M_H \left( 2\epsilon / (T_1 \log(4n/\log 2)) \right). \]

Let \( \epsilon_n \) satisfy
\[
M_H \left( \frac{2\epsilon_n}{T_1 \log (4n/\log 2)} \right) = n \epsilon_n^2
\]
and let \( \epsilon_n \) be chosen such that \( M_H(\epsilon_n) = 4n \epsilon_n^2 + 2 \log 2 \). From Theorem 2, we have the following result.

**Theorem 5:** For a density class \( \mathcal{F} \) with \( \|f\|_\infty \leq T \) for \( f \in \mathcal{F} \), with \( \epsilon_n \) and \( \epsilon_n \) as defined above,
\[
\frac{\epsilon_n^2}{8} \leq \min_{\hat{f}} \max_{f \in \mathcal{F}} d_H^2 (f, \hat{f}) \leq \min_{\hat{f}} \max_{f \in \mathcal{F}} ED(\|f\|_\mathcal{F}, \hat{f}) \leq 2 \epsilon_n^2.
\]

**Remark:** Due to the presence of the \( \log n \) term in the determination of \( \epsilon_n \), the quantities \( \epsilon_n^2 \) and \( \epsilon_n^2 \) are typically of order \( \tau_n^2 / \log^\beta_1 n \) and \( \tau_n^2 \log^\beta_2 n \) for some \( \beta_1, \beta_2 > 0 \) respectively, for nonparametric smooth families, where \( \tau_n \) is chosen such that \( M_H(\tau_n) = n \tau_n^2 \). We suspect the extra \( \log n \) might be necessary for the upper bound without any regularity condition relating K-L distance to Hellinger distance. See Barron, Birgé and Massart (1995) for related conclusions.

### 2.4 Some more results for K-L 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. The result is applicable to the following examples.

1. **Densities with unknown interval of support.**

   Let \( \mathcal{F} = \{ f(x) 1_{a \leq x \leq b} / c : f \in \mathcal{F}_0, c = \int_0^1 f(t) dt, (a,b) \in [0,1] \text{ and } b-a \geq \Delta \} \), where \( \mathcal{F}_0 \) is a nonparametric class of positive functions on \([0,1]\) which are uniformly bounded above and away from 0, and \( \Delta \) is a fixed constant with \( 0 < \Delta < 1 \) (\( \Delta \) is used to force the densities in \( \mathcal{F} \) to be uniformly upper bounded). Note that the densities in this class are uniformly bounded above and away from 0 on their supports but the supports are unknown.

2. **Densities with polynomial tails at unknown boundaries.**

   Let \( \mathcal{F} = \{ f(x) g(x; a,b, \alpha, \beta) / c : f \in \mathcal{F}_0, c = \int_0^1 f(t) g(t; a,b, \alpha, \beta) dt, (a,b) \in [0,1] \text{ and } b-a \geq \Delta, (\alpha, \beta) \in [0,\gamma] \} \), where \( \mathcal{F}_0 \) and \( \Delta \) are as before, \( \gamma \) is any fixed positive number,
and
\[ g(x; a, b, \alpha, \beta) = \begin{cases} (x - a)^\alpha(b - x)^\beta & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}. \]

In this class, the densities are allowed to approach zero at polynomial rates at the end points of the supports.

3. Densities with support on unknown set of \( k \) intervals.

Let \( F = \{ f(x) \cdot \sum_{i=0}^{k-1} b_{i+1} 1_{a_i \leq x < a_{i+1}} / c : f \in F_0, c = \int_0^1 f(t) \cdot \sum_{i=0}^{k-1} b_{i+1} 1_{a_i \leq t < a_{i+1}} dt, 0 = a_0 < a_1 < a_2 < \cdots < a_k = 1, \sum_{i=0}^{k-1} b_{i+1} (a_{i+1} - a_i) \geq \gamma_1 \text{ and } 0 \leq b_i \leq \gamma_2, 1 \leq i \leq k \}. \)

Here \( k \) is a positive integer, \( \gamma_1 \) and \( \gamma_2 \) are positive constants (the two constants are used to force the densities in \( F \) to be uniformly bounded). If the densities in \( F_0 \) are continuous, then the densities in \( F \) have at most \( k - 1 \) discontinuous points. This example is a generalization of the first example.

Examples of the above type are handled by the following theory.

Let \( \mathcal{H} \) be a function class with \( 0 < \epsilon \leq h \leq \overline{C} < \infty \) for all \( h \in \mathcal{H} \). Let \( \mathcal{G} \) be a class of nonnegative functions satisfying \( \| g \|_{\infty} \leq \overline{C} \) and \( \int g d\mu \geq \epsilon \) for all \( g \in \mathcal{G} \). Though the \( g \)'s are nonnegative, they may equal zero on some subsets. We take the reference measure \( \mu \) to be a probability measure. Consider a class of densities with respect to \( \mu \)

\[ F = \left\{ f(x) = \frac{h(x)g(x)}{\int h(x)g(x) d\mu} : h \in \mathcal{H}, g \in \mathcal{G} \right\}. \]

Let \( \tilde{\mathcal{H}} = \{ \tilde{h} = h / \int h d\mu : h \in \mathcal{H} \} \) and \( \tilde{\mathcal{G}} = \{ \tilde{g} = g / \int g d\mu : g \in \mathcal{G} \} \) be density classes corresponding to \( \mathcal{H} \) and \( \mathcal{G} \) respectively. It will be assumed that the class of densities \( \tilde{\mathcal{G}} \) is smaller than the class \( \mathcal{H} \) in an \( \epsilon \)-entropy sense. Let \( M_2(\epsilon; \mathcal{H}), M_2(\epsilon; \tilde{\mathcal{G}}) \) be packing entropies of \( \mathcal{H} \) and \( \tilde{\mathcal{G}} \) respectively under \( L_2 \) distance and let \( V_K(\epsilon; F), V_K(\epsilon; \tilde{\mathcal{G}}) \), and \( V_K(\epsilon; \tilde{\mathcal{H}}) \) be covering entropies (logarithms of cardinalities of smallest \( \epsilon \)-nets in \( F, \tilde{\mathcal{G}} \) and \( \tilde{\mathcal{H}} \) respectively) under \( d_K \).

For the new density class, we have the following result.

**Theorem 6:** Suppose \( V_K(\epsilon; \tilde{\mathcal{G}}) \leq A_1 M_2(A_2 \epsilon; \mathcal{H}) \) for some positive constants \( A_1, A_2 \) and \( \liminf_{\epsilon \to 0} M_2(\alpha \epsilon; \mathcal{H}) / M_2(\epsilon; \mathcal{H}) > 1 \) for some \( 0 < \alpha < 1 \) (that is, \( \mathcal{H} \) is rich using the \( L_2 \) distance). If \( \mathcal{G} \) contains a constant function, then

\[ \min_{f} \max_{\hat{f} \in \mathcal{F}} E_f D(f \| \hat{f}) \leq \min_{f} \max_{\hat{f} \in \mathcal{F}} E_f d_H^2(f, \hat{f}) \leq \min_{f} \max_{\hat{f} \in \mathcal{F}} E_f \| f - \hat{f} \|_2^2 \leq \epsilon_n^2, \]

where \( \epsilon_n \) is determined by \( M_2(\epsilon_n; \mathcal{H}) = n \epsilon_n^2. \)

**Proof:** By Theorem 2, \( \min_{f} \max_{\hat{f} \in \mathcal{F}} E_f D(f \| \hat{f}) \leq \tilde{\epsilon}_n^2 \), where \( \tilde{\epsilon}_n \) satisfies \( V_K(\tilde{\epsilon}_n; F) = n \epsilon_n^2. \)

From Lemma 11 in the appendix, we have \( V_K(\epsilon; F) \leq \tilde{A}_1 M_2(\tilde{A}_2 \epsilon; \mathcal{H}) \) for some constants \( \tilde{A}_1, \tilde{A}_2 \).
and \( \tilde{A}_2 \). Then under the assumption of the richness of \( \mathcal{H} \), we have \( \tilde{e}_n = O(\epsilon_n) \) (see the proof of Lemma 8). Thus \( \epsilon_n^2 \) upper bounds the minimax risk rates under both \( d_{K}^2 \) and \( d_{\mathcal{H}}^2 \). Because \( \mathcal{G} \) contains a constant function, \( \mathcal{H} \) is a subset of \( \mathcal{F} \). Note that the log-densities in \( \mathcal{H} \) are uniformly bounded, and from Lemma 14, the \( L_2 \) metric entropies of \( \mathcal{H} \) and \( \mathcal{H} \) are of the same order. Thus taking \( S_0 = \mathcal{H} \), the lower bound rate under K-L or squared Hellinger or square \( L_2 \) distance is of order \( \epsilon_n^2 \) by Corollary 3. Because the densities in \( \mathcal{F} \) are uniformly upper bounded, the \( L_2 \) distance between two densities in \( \mathcal{F} \) 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 \( \mathcal{F} \) satisfies \( M_2(\epsilon; \mathcal{F}) \leq V_K(A\epsilon; \mathcal{F}) \leq M_2(A'\epsilon; \mathcal{H}) \) for some positive constants \( A \) and \( A' \). Consequently, the minimax \( L^2 \) risk is upper bounded by order \( \epsilon_n^2 \) by Theorem 3. This completes the proof of Theorem 6.

To apply this theorem, we still need to bound the covering entropy of \( \mathcal{F} \) under \( d_K \). In each of the following situations, we can obtain the desired inequality \( V_K(\epsilon; \mathcal{G}) \leq A_1 M_2(\epsilon; \mathcal{H}) \) for some constants \( A_1 \) and \( A_2 \).

1. Suppose the sup-norm metric entropy of \( \mathcal{G} \) satisfies \( M_\infty(\epsilon; \mathcal{G}) \leq B_1 M_2(B_2\epsilon; \mathcal{H}) \) for some constants \( B_1, B_2 > 0 \), and \( \sup_{g \in \mathcal{G}} \int (1/g) d\mu \leq \xi < \infty \). For \( g, g' \in \mathcal{G} \), let \( \bar{g} = g_1/\int g_1 d\mu \) and \( \bar{g}' = g_2/\int g_2 d\mu \) be corresponding densities. Then

\[
D(\bar{g}_1 \parallel \bar{g}_2) \leq \int \frac{\bar{g}_1 - \bar{g}_2}{g_2} d\mu \leq \frac{4\xi}{\epsilon^2} \int \frac{1}{g_2} d\mu \leq \frac{4\xi}{\epsilon^2} \|g_1 - g_2\|_2^2.
\]

It follows that \( V_K(\epsilon; \mathcal{G}) \leq M_\infty(\epsilon'; \mathcal{G}) \leq B_1 M_2(B_2\epsilon'; \mathcal{H}) \), where \( \epsilon' = \epsilon^{3/2} / \left(2c\xi^{1/2}\right) \). A sufficient condition for \( \sup_{g \in \mathcal{G}} \int (1/g) d\mu < \infty \) is that there is a function \( g_L \in \mathcal{G} \) such that \( g \geq g_L \) for all \( g \in \mathcal{G} \) and \( \int (1/g_L) d\mu < \infty \), such as \( g_L(x) = x^{\beta_1}(1 - x)^{\beta_2} \) for some \( 0 \leq \beta_1, \beta_2 < 1 \) with \( X = [0, 1] \) and \( \mu \) = Lebesgue measure.

2. Suppose the Hellinger metric entropy of \( \mathcal{G} \) satisfies \( M_{H}(\epsilon/\log(1/\epsilon); \mathcal{G}) \leq M_2(A\epsilon; \mathcal{H}) \) for some constant \( A > 0 \). From Lemma 5, \( V_K(\epsilon; \mathcal{G}) \leq M_{H}(A'\epsilon/\log(1/\epsilon); \mathcal{G}) \) for some constant \( A' > 0 \) when \( \epsilon \) is sufficiently small. Then we have \( V_K(\epsilon; \mathcal{G}) \leq M_2(AA'\epsilon; \mathcal{H}) \).

3. Suppose \( \mathcal{G} \) is small compared to \( \mathcal{H} \) in the sense that the \( L_2 \) metric entropy of \( \mathcal{G} \) satisfies \( M_2(\epsilon^2; \mathcal{G}) \leq M_2(\epsilon; \mathcal{H}) \). Then by Lemma 12 in the appendix, \( V_K(\epsilon; \mathcal{G}) \leq M_2(A_3\epsilon^2; \mathcal{G}) \) for some positive constant \( A_3 \). Using Lemma 13 in the appendix together with \( M_2(\epsilon^2; \mathcal{G}) \leq M_2(\epsilon; \mathcal{H}) \), we have \( M_2(A_3\epsilon^2; \mathcal{G}) \leq M_2(A_4\epsilon; \mathcal{H}) \) for some constant \( A_4 > 0 \). Thus \( V_K(\epsilon; \mathcal{G}) \leq M_2(A_4\epsilon; \mathcal{H}) \).
The first case allows $G$ to be as large as $\mathcal{H}$, but puts an integrability condition to regulate the behavior when density values approach zero. The second case allows $G$ to be almost as large as $\mathcal{H}$ without any additional condition. An example for the first case is: $\mathcal{H} = \{ h : \log h \in B_{\alpha}^\mu (C) \}$ and $G = \{ g : g \in B_{\sigma, q}^\nu (C), g(x) \geq \beta_0 x^{\beta_1} (1-x)^{\beta_2} \}$ for some $0 \leq \beta_1, \beta_2 < 1$ and $\beta_0 \leq 1$, where $B_{\sigma, q}^\nu (C)$ ($\alpha > 1/q$) is a Besov class defined in Section 7. An example for the second case is: $\mathcal{H} = \{ h : \log h \in B_{\alpha}^\mu (C) \}$ and $G = \{ g : g \in B_{\sigma, q}^\nu (C), g \geq 0 \text{ and } \int g \mu = 1 \}$ ($\alpha' > 1/q$) with $\alpha' > \alpha$ and $\alpha'$ arbitrarily close to $\alpha$. Case 3 includes the situation when $\mathcal{H}$ is a rich nonparametric class and $G$ is a parametric class with $M_2(\varepsilon; G) \leq \log^\beta (1/\varepsilon)$ for some $\beta > 0$. The condition for case 3 is satisfied by the three examples in the beginning of the section.

3 Application in data compression

The obtained theorems can be used to get bounds on the minimax redundancy for data compression. Let $X_1, \ldots, X_n$ be an i.i.d. sample of discrete random variable from $p_\theta (x_1, \ldots, x_n)$, $\theta \in \mathcal{S}$. 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 code length and the expected code length 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, pp. 459-460)).

The minimax redundancy lower bounds have been previously considered by Rissanen (1984), Clarke and Barron (1990), Rissanen, Speed and Yu (1992), Yu (1996) 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 following lemma relates the game with loss $D(p_\theta^n || q_n)$ to the cumulative K-L risk.

**Lemma 6:**

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

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).

Moreover, for any density $q_n$ on the sample space of $X_1, \ldots, X_{n-1}$, there exists an estimator $\hat{p}$ such that

$$E_\theta D(p_\theta || \hat{p}) \leq n^{-1} D(p_\theta^n || q_n) \text{ for all } \theta \in \mathcal{S}.$$  \hspace{1cm} (3)

23
Proof: Given the estimator sequence, define \( q(x_{k+1}|x^k) = \hat{p}_k(x_{k+1}) \) for \( k \geq 0 \). Let \( q_n(x_1, ..., x_n) = \Pi_{k=0}^{n-1} q(x_{k+1}|x^k) \). Then \( q_n \) is a joint probability density function on \( \mathcal{X}^n \). Following an argument similar to the one for proving the upper bound part of Theorem 2,

\[
D(p^n_\theta \parallel q_n) = E_\theta \log \frac{p_\theta(X_1) \cdots p_\theta(X_n)}{q_n(X_1, ..., X_n)} = \sum_{k=0}^{n-1} E_\theta \log \frac{p_\theta(X_k)}{q(X_k|X^{k-1})} d\mu \leq \sum_{k=0}^{n-1} E_\theta D(p_\theta \parallel \hat{p}_k).
\]

In the same way, for any density \( q_n \), we can rewrite it as \( q_n(x_1, ..., x_n) = q_1(x_1)q_2(x_2|x_1) \cdots q_n(x_n|x^{n-1}) \), where \( q_k(x|x^{k-1}) \) is the conditional density of \( X_k \) given \( X^{k-1} = x^{k-1} \) according to the joint density \( q_n \). Then defining \( \hat{p}_{k-1}(x) = q_k(x|x^{k-1}) \), the identity (4) again holds. Finally, Let \( \hat{p} = n^{-1} \sum_{k=1}^{n} \hat{p}_{k-1}(x) \). Then \( \hat{p} \) is a density estimator of \( p_\theta \). Then by convexity as in the proof of the upper bound in Theorem 2,

\[
E_\theta D(p_\theta \parallel \hat{p}) \leq n^{-1} \sum_{k=1}^{n} E_\theta D(p_\theta \parallel \hat{p}_{k-1}) = n^{-1} D(p^n_\theta \parallel q_n).
\]

This completes the proof of Lemma 6.

Let \( \Omega_n \) be the collection of all density functions on the sample space \( \mathcal{X}^n \) of \( (X_1, ..., X_n) \). From Lemma 6 and the fact that the maximum (over \( \theta \) in \( S \)) of the sum of risks is not greater than the sum of the maxima, we have the following result connecting the minimax redundancy with the minimax risk.

Corollary 6: Let \( R_n = \min_{q_\theta \in \Omega_n} \max_{\theta \in S} D(p^n_\theta \parallel q_n) \) and \( r_n = \min_{\hat{p}_n \in R_n} \max_{\theta \in S} E_\theta D(p_\theta \parallel \hat{p}) \).

Then

\[
r_{n-1} \leq R_n \leq \sum_{i=0}^{n-1} r_i.
\]

Moreover, \( r_n \asymp R_n/n \) when \( r_n \asymp n^{-1} \sum_{i=0}^{n-1} r_i \).

For smooth nonparametric density classes, \( nr_n \) often gives the right order of the minimax redundancy. However, for parametric classes or other less “rich” families, this lower bound may be suboptimal. For instance, for smooth parametric families, it is known (see, e.g., Clarke and Barron (1994)) that the minimax redundancy is of order \( \frac{m}{2} \log n \), where \( m \) is the number of parameters in the family. But \( nr_n \) is bounded by a constant. When \( r_n \) converges to zero at a polynomial rate \( n^{-\rho} \) with \( 0 < \rho < 1 \), we have that \( r_n \asymp \frac{1}{n} \sum_{i=1}^{n} r_i \) and hence the minimax redundancy satisfies \( R_n \asymp nr_n \). In particular, when \( r_n \asymp \epsilon_n^2 \) is determined by solving \( n \epsilon_n^2 = M(\epsilon) \) for some \( M(\epsilon) \asymp \epsilon^{-\alpha} \) with \( \alpha > 0 \), then \( n^{-1} \sum_{i=0}^{n-1} \epsilon_i^2 \asymp \epsilon_n^2 \) and \( R_n \asymp nr_n \) holds with \( r_n \asymp n^{-2\alpha/(2\alpha+1)} \).
Now let $d(\theta, \theta')$ be a metric on $S$ and assume that $\{p_\theta : \theta \in S\}$ contains all probability densities on $X$. Let $M_d(\epsilon)$ be the packing entropy of $S$ under $d$ and let $V(\epsilon)$ be an upper bound on the covering entropy $V_K(\epsilon)$ of $S$ under $d_K$. Choose $\epsilon_n$ such that $\epsilon_n^2 = V(\epsilon_n)/n$ and choose $\zeta_n$ such that $M_d(\zeta_n) = 4n\epsilon_n^2 + 2\log 2$.

**Theorem 7:** Assume that $D(p_\theta \| p_{\theta'}) \geq A_0 d^2(\theta, \theta')$ for all $\theta, \theta' \in S$. Then we have

$$(A_0/8)n\zeta_n^2 \leq \min_{\zeta_n} \max_{\theta \in S} D(p_\theta^n \| q_n) \leq 2n\epsilon_n^2,$$

where the minimization is over all densities on $X^n$.

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

**Proof of Theorem 7:** The lower bound follows from Theorem 1 and Lemma 6 (Equation (3)). For the upper bound, consider the choice $q(x^n) = (1/|G_{\epsilon_n}|) \sum_{p \in G_{\epsilon_n}} p(x^n)$, the mixture with respect to the uniform prior on the $\epsilon_n$-net $G_{\epsilon_n}$ of $S$. Then $D(p^n \| q_n) \leq V(\epsilon_n) + n\epsilon_n^2 \leq 2n\epsilon_n^2$ as in equation (2). So the proof of Theorem 7 is complete.

When interest is focused on the cumulative K-L risk (or on the individual risk $r_n$ in the case that $r_n \approx n^{-1} \sum_{i=0}^{n-1} r_i$), direct proof of suitable bounds are possible without the use of Fano’s inequality. See Haussler and Opper (1995) for new results in that direction. The following is another proof using an idea of Rissanen (1984) (see also Barron and Hengartner (1995)).

**Theorem 8:** Assume $D(p_\theta \| p_{\theta'}) \leq \overline{A}d_H^2(\theta, \theta')$ for any $\theta, \theta' \in S$ with $d_H^2(\theta, \theta')$ small. Let $\epsilon_n$ be chosen such that $V(\epsilon_n) = n\epsilon_n^2$. Then when $n$ is large enough,

$$\min_{\zeta_n} \max_{\theta \in S} D(p_\theta^n \| q_n) \geq (1/2)M_H(4\epsilon_n/\overline{A})^{1/2} - \log 2.$$

**Proof:** As above, the mixture $q_n(x^n) = 1/|G_{\epsilon_n}| \sum_{p \in G_{\epsilon_n}} p(x^n)$ achieves $\max_{\theta \in S} D(p_\theta^n \| q_n) \leq 2n\epsilon_n^2$. Let $\hat{p}$ be the density estimator constructed in section 2.1.2, then $E_\theta D(p_\theta \| \hat{p}) \leq 2\epsilon_n^2$ in accordance with (3). Let $B_{n, \theta} = \{X^n : D(p_\theta \| \hat{p}) \leq 4\epsilon_n^2\}$. Then by Markov inequality,$$ P_\theta(B_{n, \theta}) \geq 1/2 \text{ uniformly for } \theta \in S. \text{ Let } q_n(x^n) \text{ be any other density on } (x_1, ..., x_n) \text{ and } \overline{Q} \text{ be the corresponding probability measure. Since discretization reduces the Kullback divergence (Kullback and Leibler (1951)), we have}\n
$$D(p_\theta^n \| q_n) = D(P_{X^n|\theta} \| \overline{Q} X^n)$$

$$\geq P_{X^n|\theta}(B_{n, \theta}) \log \frac{P_{X^n|\theta}(B_{n, \theta})}{\overline{Q} X^n(B_{n, \theta})} + \left(1 - P_{X^n|\theta}(B_{n, \theta})\right) \log \frac{1 - P_{X^n|\theta}(B_{n, \theta})}{1 - \overline{Q} X^n(B_{n, \theta})}$$

$$\geq P_{X^n|\theta}(B_{n, \theta}) \log \left(1/\overline{Q} X^n(B_{n, \theta})\right) - \log 2$$

$$\geq (1/2) \log \left(1/\overline{Q} X^n(B_{n, \theta})\right) - \log 2.$$
Now let $N_{\xi_n}$ be an $\xi_n$-packing set under $d_H$ (one may also consider a more general distance satisfying Condition 0). Since $D(p_\theta \mid p_{\theta'}) \leq \overline{A}d_H^2(\theta, \theta')$ for any $\theta, \theta' \in S$ with $d_H^2(\theta, \theta')$ small, when $n$ is large enough, with the choice of $\xi_n = 4\epsilon_n/A^{1/2}$, the set $B_{n, \theta}$ is contained in $\{X^n : d_H(p_{\theta}, \overline{\pi}) \leq \xi_n/2 \}$. Since $d_H(\theta, \theta') \geq \xi_n$ for distinct $\theta, \theta'$ in the packing set $N_{\xi_n}$, we know $B_{n, \theta} \cap B_{n, \theta'}$ is empty. Thus $\sum_{\theta \in N_{\xi_n}} \overline{Q}X^n(B_{n, \theta}) \leq 1$, and hence there exists at least one $\theta_0 \in N_{\xi_n}$ with $\overline{Q}X^n(B_{n, \theta}) \leq 1/|N_{\xi_n}|$. Now we lower bound the maximal total risk as follows

$$\max_{\theta \in S} D(p_{\theta}^n || \overline{q}_n) \geq \max_{\theta \in S} \left( (1/2) \log \left( 1/\overline{Q}X^n(B_{n, \theta}) \right) - \log 2 \right)$$

$$\geq (1/2) \log \left( 1/\overline{Q}X^n(B_{n, \theta_0}) \right) - \log 2$$

$$\geq (1/2) \log \left( |N_{\xi_n}| \right) - \log 2.$$  

This bound holds for all $\overline{q}_n$. The assertion of Theorem 8 follows.

### 4 Application in nonparametric regression

Consider the regression model

$$y_i = u(x_i) + \varepsilon_i, \; i = 1, \ldots, n.$$  

Suppose the errors $\varepsilon_i, 1 \leq i \leq n$, are i.i.d. with the $\text{Normal}(0, \sigma^2)$ distribution. The explanatory variables $x_i, 1 \leq i \leq n$, are i.i.d. with a fixed density function $h(x)$. The regression function $u$ is assumed to be in a function class $\mathcal{U}$. For this case, the square root K-L distance between the joint densities of $(X, Y)$ in the family is a metric. Let $\|u - v\|_{L_2(h)} = (\int (u - v)^2 h \, d\mu)^{1/2}$ be the $L_2$ distance with respect to the measure induced by $X$. Let $M_2(\epsilon)$ be the maximum of the logarithm of the cardinality of any $\epsilon$-packing set under $L_2(h)$ norm. Similarly let $M_q(\epsilon)$ be the $\epsilon$-packing entropy of $\mathcal{U}$ under $L_q(h)$ norm. Assume $M_2(\epsilon) < \infty$ for every $\epsilon > 0$ and $M_2(\epsilon) \to \infty$ as $\epsilon \to 0$. Choose $\epsilon_n$ such that

$$M_2(\sqrt{2\epsilon_n}) = n\epsilon_n^2.$$  

Let $\xi_n = \xi_{n, 2}$ satisfy

$$M_2(\xi_n) = 4n\epsilon_n^2 + 2\log 2.$$  

Similarly let $\xi_{n, q}$ be determined by equation $M_q(\xi_{n, q}) = 4n\epsilon_n^2 + 2\log 2$.

**Theorem 9:** The minimax squared $L_2(h)$ risk for the regression function estimation is lower bounded by a rate determined by the $L_2$ packing entropy of $\mathcal{U}$ as follows:

$$\min_{\hat{u}} \max_{u \in \mathcal{U}} \mathbb{E} \| u - \hat{u} \|^2_{L_2(h)} \geq \sigma^2 \xi_{n, 2}/8.$$  

26
Generally, for the minimax $L_q(h)$ risk, we have

$$
\min_{\hat{u}} \max_{u \in \mathcal{U}} E \| u - \hat{u} \|_{L_q(h)} \geq \sigma_{\mathcal{L}_n, q}/4.
$$

**Proof:** Let $p_u(x, y) = (2\pi \sigma^2)^{-1/2}e^{-(y-u(x))^2/2\sigma^2}h(x)$ denote the joint density of $(X, Y)$ with regression function $u$. Then $D(p_u || p_v) = (1/2\sigma^2)E \{(Y - u(X) - (Y - v(X))^2\right\})$ reduces as is well know to $(1/2\sigma^2)\int (u(x) - v(x))^2h(x)dx$, so that $d_K$ is equivalent to the $L_2(h)$ distance.

Let $S = \{ u : u \in \mathcal{U} \}$ and $\overline{S} = \{ u : \| u \|_{L_2(h)}^2 < \infty \}$. Let $\mathcal{A}_n$ be the collection of the regression estimators which maps from the sample space to $\overline{S}$. From Theorem 1 with $d = d_K$, we have

$$
\min_{\hat{u} \in \mathcal{A}_n} \max_{u \in \mathcal{U}} E \| u - \hat{u} \|_{L_2(h)}^2 \geq \sigma_{\mathcal{L}_n, q}^2/8.
$$

With $d$ equal to the $L_q(h)$ distance, Theorem 1 yields the bound for the $L_q(h)$ risk, where now $\mathcal{L}_{n, q}$ depends on the metric entropy under both $L_2$ and $L_q$ as indicated. This ends the proof of Theorem 9.

We next determine upper bounds for regression by specializing the bound from Theorem 2. However, it is not immediately clear that the risk of estimating $u(x)$ with loss $D(p_u || p_{\hat{u}_n}) = (1/2\sigma^2)\| u - \hat{u}_n \|_{L_2(h)}^2$ can be bounded by the risk of estimating the density $p_u$ with loss $D(p_u || \hat{p}_n)$, since the estimator $\hat{p}_n$ in Theorem 2 is not of the form $p_{\hat{u}_n}$. We show that a suitable bound on the risk holds when the regression functions are uniformly bounded, using a minimum Hellinger distance argument, in the case that $\sigma^2$ is known.

We assume $\| u \|_{\infty} \leq L$ uniformly for $u \in \mathcal{U}$. Theorem 2 provides a density estimator $\hat{p}_n$ such that $\max_{u \in \mathcal{U}} E D(p_u || \hat{p}_n) \leq 2\epsilon_n^2$. It follows that $\max_{u \in \mathcal{U}} E d_H^2(p_u, \hat{p}_n) \leq 2\epsilon_n^2$. (A similar conclusion is available in B"{u}rg's (1986), Theorem 3.1.) Here we take advantage of the fact that when the density $h(x)$ is fixed, $p_{\hat{u}_n}(x, y)$ 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 $\epsilon$-nets). For given $x$ and $(X_i, Y_i)_{i=1}^n$, let $\overline{u}_n(x)$ 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 $|z| \leq L$. Then $\overline{u}_n(x)$ is an estimator of $u(x)$ based on $(X_i, Y_i)_{i=1}^n$. By the triangle inequality, given $x$ and $(X_i, Y_i)_{i=1}^n$,

$$
d_H\left( \phi_{\overline{u}_n(x)}, \phi_{\overline{u}_n(z)} \right) \leq d_H\left( \phi_{\overline{u}_n(x)}, \hat{g}_n(\cdot | x) \right) + d_H\left( \phi_{\overline{u}_n(x)}, \hat{g}_n(\cdot | x) \right).
$$

It follows that

$$
\max_{u \in \mathcal{U}} E d_H^2(p_u, p_{\overline{u}_n}) \leq 4 \max_{u \in \mathcal{U}} E d_H^2(p_u, \hat{p}_n) \leq 8\epsilon_n^2.
$$
Now $Ed^2_h(p_u, p_n) = 2E \int h(x) \left( 1 - e^{-\langle u(x) - \tilde{u}_n(x) \rangle^2 / 8\sigma^2} \right) d\mu$. The concave function $(1 - e^{-v})$ is above the chord $(v/B)(1 - e^{-B})$ for $0 \leq v \leq B$. Thus using $v = (u(x) - \tilde{u}_n(x))^2 / 8\sigma^2$ and $B = L^2 / 2\sigma^2$ we obtain

$$
\max_{u \in \mathcal{U}} E \int (u(x) - \tilde{u}_n(x))^2 h(x) d\mu \leq 2L^2 (1 - e^{-L^2 / 2\sigma^2})^{-1} \max_{u \in \mathcal{U}} Ed^2_h(p_u, p_n) \leq 16L^2 (1 - e^{-L^2 / 2\sigma^2})^{-1} \epsilon_n^2.
$$

Together with Theorem 9, we have the following result.

**Theorem 10:** Assume $\sup_{u \in \mathcal{U}} || u ||_\infty \leq L$ for some $L \geq \sigma$. Then

$$
\min \max_{\bar{u}} \max_{u \in \mathcal{U}} || u - \bar{u} ||_{L_2(h)}^2 \leq 16L^2 (1 - e^{-1/2})^{-1} \epsilon_n^2,
$$

for $\epsilon_n$ satisfying $M_2(\sqrt{2} \epsilon_n) = n \epsilon_n^2$. If further $\lim_{\epsilon \to 0} M_2(\epsilon/2) / M_2(\epsilon) > 1$ (i.e., $\mathcal{U}$ is rich in $L_2$), then

$$
\min \max_{\bar{u}} \max_{u \in \mathcal{U}} || u - \bar{u} ||_{L_2(h)}^2 \leq \epsilon_n^2,
$$

and if further $M_2(\epsilon) \asymp M_q(\epsilon)$ for $1 \leq q < 2$, then for the $L_q(h)$ risk, we have

$$
\min \max_{\bar{u}} \max_{u \in \mathcal{U}} || u - \bar{u} ||_{L_q(h)} \leq \epsilon_n.
$$

**Remark:** Let $\overline{M}_2(\epsilon)$ be the $\epsilon$-packing entropy of $\mathcal{U}$ under $L_2$ norm $|| u ||_{L_2} = (\int u^2 d\mu)^{1/2}$. If $|| \log h ||_{\infty} \leq C < \infty$, then $\overline{M}_2(b \epsilon) \leq M_2(\epsilon) \leq \overline{M}_2(a \epsilon)$ holds for some constants $a, b$ depending on $C$. Note that the constructed estimator $\tilde{u}_n$ (which achieves the minimax optimal rate) depends on the density $h(x)$ of the independent variable only through finding an $\epsilon$-net under $L_2(h)$ norm.

So under the condition that $\mathcal{U}$ is rich with $L_2$ distance, one can show for any $C > 0$,

$$
\min \max_{\bar{u}} \max_{u: || \log h ||_{\infty} \leq C} \max_{u \in \mathcal{U}} || u - \bar{u} ||_{L_2}^2 \leq C \epsilon_n^2.
$$

### 5 Linear approximation and minimax rates

Let $\Phi = \{ \phi_1, ..., \phi_k, ... \}$ be a fundamental sequence in $L^2[0, 1]^d$ (that is, linear combinations are dense in $L^2[0, 1]^d$). Let $\Gamma = \{ \gamma_0, ..., \gamma_k, ... \}$ for which $\gamma_k \downarrow 0$ as $k \to \infty$. Let $\eta_0(f) = \| f \|_2$ and $\eta_k(f) = \min \{ a_i \} \| f - \sum_{i=1}^k a_i \phi_i \|_2$ for $k \geq 1$ be the $k$-th degree of approximation of $f \in L^2[0, 1]^d$ by the system $\Phi$. Let $\mathcal{F}(\Gamma, \Phi)$ be all functions in $L^2[0, 1]^d$ with the approximation errors bounded by $\Gamma$, i.e.,

$$
\mathcal{F}(\Gamma, \Phi) = \{ f \in L^2[0, 1]^d : \eta_k(f) \leq \gamma_k, k = 0, 1, ... \}.
$$
They are called the full approximation sets. Lorentz (1966) gives metric entropy bounds on
these classes (he actually treats general Banach spaces, not only $L_2$). These bounds have
been used to derive metric entropy orders for a variety of function classes (see Lorentz (1966))
including Sobolev classes.

Define $N_0 = 0$, $N_i = \min\{k : \gamma_k \leq e^{-i}\}$, $i = 1, 2, \ldots$. For a given $\epsilon > 0$, let $j(\epsilon)$ be defined
by $e^{-(j-1)} < \epsilon \leq e^{-j-2}$. Then from Lorentz (1966), the $L_2$ metric entropy of $\mathcal{F}(\Gamma, \Phi)$ satisfies
$M_2(\epsilon) \geq \sum_{i=1}^{j(\epsilon)-3} N_i$, and if further $\gamma_k$ satisfies $\gamma_{2k} \leq c \gamma_k$, $k = 0, 1, \ldots$ for some $0 < c < 1$, then
there exists $\lambda > 1$ such that $M_2(\epsilon) \leq \lambda \sum_{i=1}^{j(\epsilon)-3} N_i$.

We next illustrate the use of these results. For a system $\Phi$, consider the functions that
can be approximated by the linear system with polynomially decreasing approximation error
$\gamma_k \sim k^{-\alpha}, \alpha > 0$. Then using the above upper and lower bounds, by simple calculations,
it is seen that the $L_2$ metric entropy of $\mathcal{F}(\Gamma, \Phi)$ is of order $M_2(\epsilon) \sim (1/\epsilon)^{1/\alpha}$. Similarly if
$\gamma_k \sim k^{-\alpha} (\log k)^{\beta}$ then $M_2(\epsilon) \sim (1/\epsilon)^{1/\alpha} (\log (1/\epsilon))^{-\beta/\alpha}$ for $\beta \in R$. More restrictive function
classes correspond to faster decays of the approximation error (for instance, if $\gamma_k \sim e^{-\lambda k}$ for
some $\lambda > 0$, then the corresponding metric entropy is only of order $M_2(\epsilon) \sim (\log (1/\epsilon))^2$, in
which case $\mathcal{F}(\Gamma, \Phi)$ would not satisfy Condition 2).

We focus attention on approximation error sequences $\Gamma$ that satisfy conditions introduced
by Lorentz (1966). First, suppose there exist $0 < c' < c < 1$ such that

$$c' \gamma_k \leq \gamma_{2k} \leq c \gamma_k,$$

(5)
as is true for $\gamma_k \sim k^{-\alpha}$ and also for $\gamma_k \sim k^{-\alpha} (\log k)^{\beta}$, $\alpha > 0$, $\beta \in R$. Then Lorentz (1966,
Theorem 4) shows that $\mathcal{F}(\Gamma, \Phi)$ is rich using the $L_2$ distance, that is, Condition 2 is satisfied.
Secondly, Lorentz assumed that $\gamma_t(1+\delta)k \geq e^{-1} \gamma_k$ for some $\delta > 0$.

Let $\epsilon_n$ be chosen such that $n \epsilon_n^2 \geq M_2(\epsilon_n)$ or equivalently $n \epsilon_n^2 \geq \sum_{i=1}^{j(\epsilon_n)-3} N_i$. Then under
Lorentz’s conditions the rate $\epsilon_n$ is also determined by $\epsilon_n^2 \leq k_n/n$ where $k_n$ is the solution to
$\gamma_k^2 \geq k/n$. In this setting, the metric entropy $M_2(\epsilon_n)$ is of the same order as the dimension $k_n$
at which the approximation error is of order $\epsilon_n$.

For density estimation, suppose for convenience that $\phi_1 = 1$ and assume that the functions
in $\mathcal{F}(\Gamma, \Phi)$ are uniformly bounded, i.e., $\sup_{g \in \mathcal{F}(\Gamma, \Phi)} \| g \|_{\infty} \leq \rho$ for some positive constant $\rho$.
Let $\tilde{\mathcal{F}}(\Gamma, \Phi)$ be all the probability density functions in $\mathcal{F}(\Gamma, \Phi)$. When $\gamma_0$ is large enough,
the $L_2$ metric entropies of $\tilde{\mathcal{F}}(\Gamma, \Phi)$ and $\mathcal{F}(\Gamma, \Phi)$ are of the same order. In fact, let $\Gamma' =
\{\gamma'_0, \gamma_1, \ldots, \gamma_k, \ldots\}$. It can be verified that $\{f = (g + \rho + 1)/(\int g \, d\mu + \rho + 1) : g \in \mathcal{F}(\Gamma', \Phi)\} \subset
\tilde{\mathcal{F}}(\Gamma, \Phi) \subset \mathcal{F}(\Gamma, \Phi)$ provided $\gamma_0 \geq \gamma'_0 + \rho + 1$. It is not hard to see that the class $\{f =
(g + \rho + 1)/(\int g \, d\mu + \rho + 1) : g \in \mathcal{F}(\Gamma', \Phi)\}$ has $\epsilon$-entropy lower bounded by order $M_2(\zeta \epsilon)$ for
some $\zeta > 0$, where $M_2(\epsilon)$ is the $\epsilon$-entropy of $\mathcal{F}(\Gamma, \Phi)$. 29
Let \( c_n \) be determined as indicated above. Assuming Lorentz’s first condition (equation 5 above) and assuming the uniform boundedness of functions in \( \mathcal{F}(\Gamma, \Phi) \), we have from Theorem 4 that the minimax rate for density estimation in \( L_2 \) is given by

\[
\min_{f} \max_{f \in \mathcal{F}(\Gamma, \Phi)} E \| f - \hat{f} \|_2^2 \leq c_n^2.
\]

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

\[
\min_{f} \max_{f \in \mathcal{F}(\Gamma, \Phi)} E \| f - \hat{f} \|_2^2 \leq n^{-2\alpha/(1+2\alpha)}.
\]

Similarly we have rate \( n^{-2\alpha/(1+2\alpha)} (\log n)^{-2\beta/(1+2\alpha)} \) if \( \gamma_k \sim k^{-\alpha} (\log k)^\beta \).

Our requirement that the functions in \( \mathcal{F}(\Gamma, \Phi) \) are uniformly bounded is satisfied if \( \Phi = \{\phi_1, \phi_2, ..., \phi_k, ...\} \) is a complete orthonormal system in \( L^2[0,1]^d \) and the approximation error bounds \( \gamma_k \) satisfy \( \sum_{i=0}^{\infty} \gamma_i \| \phi_{i+1} \|_\infty \leq \rho < \infty \). (For \( g(x) = \sum_{i=0}^{\infty} \xi_i \phi_i \in \mathcal{F}(\Gamma, \Phi) \), the orthonormality yields \( \sum_{i=0}^{\infty} \xi_i^2 \leq \gamma_k^2 \), which implies \( |\xi_{k+1}| \leq \gamma_k \) and together with the latter condition, it implies \( \| g \|_\infty \leq \rho \). The condition \( \sum_{i=0}^{\infty} \gamma_i \| \phi_{i+1} \|_\infty < \infty \) is equivalent to \( \sum_{i=0}^{\infty} \gamma_i < \infty \) for the trigonometric system, and is equivalent to \( \sum_{i=0}^{\infty} \gamma_i \sqrt{i} < \infty \) for the system of one-dimensional Legendre polynomials.

Now consider estimating a regression function in \( \mathcal{F}(\Gamma, \Phi) \). Let \( \xi_n \) be determined from \( 4n\xi_n^2 + 2 \log 2 = \sum_{i=1}^{\tilde{N}} N_i^{-1/2} \). Then from Theorem 9 we have

\[
\min_{u} \max_{u \in \mathcal{F}(\Gamma, \Phi)} E \| u - \hat{u} \|_2^2 \geq \frac{\xi_n^2}{4}.
\]

This lower bound holds without requiring the functions in \( \mathcal{F}(\Gamma, \Phi) \) to have a uniform bound.

When the class is uniformly bounded and the requirement \( c' \gamma_k \leq \gamma_{2k} \leq c \gamma_k \) is satisfied, we also have the upper bound from Theorem 10 and the minimax rate is identified as follows

\[
\min_{\hat{u}} \max_{u \in \mathcal{F}(\Gamma, \Phi)} E \| u - \hat{u} \|_2^2 \leq c_n^2.
\]

As before, we get rate \( n^{-2\alpha/(1+2\alpha)} \) when \( \gamma_k \sim k^{-\alpha} \).

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, in the regression example, with \( y_i = u(x_i) + \varepsilon_i, u \in \mathcal{F}(\Gamma, \Phi) \), approximation systems yields natural and well-known estimates that achieve this rate.

In particular, let \( \phi_1(x), \ldots, \phi_k(x), \ldots \) be chosen (by the Gram-Schmidt process preserving \( \mathcal{F} \)) to be orthonormal with respect to a given density \( h(x) \). Then the model \( u_k(x) = \sum_{j=1}^{k} a_j \phi_j(x) + \varepsilon_j \) with \( a_j = \int u(x) \phi_j(x) h(x) d\mu \) is estimated by \( \hat{u}_k(x) = \sum_{j=1}^{k} \hat{a}_j \phi_j(x) \) with \( \hat{a}_j = n^{-1} \sum_{i=1}^{n} y_i \phi_j(x_i) \),
which has approximation error (squared bias) \( \|u - u_k\|^2 \leq \gamma_k^2 \) and estimation error (variance) \( E[\|\tilde{u}_k - u_k\|^2] \) bounded by \( (\sigma^2 + \|u\|_\infty^2) k/n \), in general, or by \( (\sigma^2 + \|u\|_\infty^2 \epsilon^2) k/n \) if the basis functions are uniformly bounded by \( \epsilon \). This trade-off is familiar in the literature [see for instance, Cox (1988) for least squares regression estimates, Cencov (1982) or Barron and Sheu (1991), for maximum likelihood log-density estimates, and Birgé and Massart (1994) for projective density estimators and other contrasts].

The best rate occurs when \( k_n \) is chosen such that \( \gamma_k^2 \approx k_n/n \). Of course, in applications, one does not know how well the underlying function can be approximated by a chosen system, which makes it impossible to know the optimal size \( k_n \). This suggests the need of a good model selection criterion to choose a suitable size model to balance the two kinds of errors automatically based on data. For recent results on model selection, see for instance, Barron, Birgé, and Massart (1995) and Yang and Barron (1996). 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 \( \mathcal{F} \) is contained in \( \mathcal{F}(\Gamma', \Phi') \) for some pair of fundamental sequences \( \Phi \) and \( \Phi' \) for which the \( \gamma_k, \gamma_k' \) sequences yield \( \epsilon_n \approx \epsilon_n' \), then \( \epsilon_n \) provides the minimax rate for \( \mathcal{F} \) 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, et al (1993)). 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.

6 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 \( \Phi \). There to get a given accuracy of approximation \( \delta \), one uses the first \( k_\delta \) basis functions with \( k_\delta = \min \{i : \gamma_i \leq \delta\} \). This choice works for all \( g \in \mathcal{F}(\Gamma, \Phi) \) and these basis functions are needed to get the accuracy \( \delta \) for some \( g \in \mathcal{F}(\Gamma, \Phi) \). For slowly converging sequences \( \gamma_k \), very large \( k_\delta \) is needed to get accuracy \( \delta \). 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 \( \alpha (\alpha \geq 1) \) 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. It is quite possible that for a subclass $G$ of functions, there is a choice (depending on $g \in G$) of sparse subsets that produce nearly as good approximation error as the full approximation. Such a function class is called sparse in terms of approximation by the system $\Phi$. We next give minimax results for some sparse function classes.

Let $\Phi = \{\phi_1, \phi_2, \ldots\}$ and $\Gamma = \{\gamma_0, \ldots, \gamma_k, \ldots\}$ be as in the previous section with $\Gamma$ satisfying the condition in equation (5). Let $I_k > k$, $k \geq 1$ be a given nondecreasing sequence of integers satisfying $\liminf I_k/k = \infty$ ($I_0 = 0$) and let $I = \{I_1, I_2, \ldots\}$. Let $\tilde{\eta}_k(g) = \min_{i \leq I_1, \ldots, i \leq I_k} \|g - \sum_{i=1}^{k} a_i \phi_i\|_2$ be called the $k$-th degree of sparse approximation of $g \in L^2[0,1]^d$ by the system $\Phi$. Here for $k = 0$, there is no approximation and $\tilde{\eta}_0(g) = \|g\|_2$. The $k$-th term used to approximate $g$ is selected from $I_k$ basis functions. Let $S(\Gamma, \Phi) = S(\Gamma, \Phi, I)$ be all functions in $L_2[0,1]^d$ with the sparse approximation errors bounded by $\Gamma$, i.e.,

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

We call it a sparse approximation set of functions (for a fixed choice of $I$). Note the smallest full approximation set that contains $S(\Gamma, \Phi)$ is $F(\Gamma', \Phi)$, where $\Gamma' = \{\gamma_0, \ldots, \gamma_0, \gamma_1, \gamma_1, \ldots, \gamma_1, \gamma_2, \gamma_2, \ldots, \gamma_2, \gamma_3, \ldots\}$ with $\gamma_k$ repeated $I_{k+1} - I_k$ times for $k \geq 0$. Then we have the following containment relationship

$$F(\Gamma, \Phi) \subseteq S(\Gamma, \Phi) \subseteq F(\Gamma', \Phi).$$

Here the smaller class $F(\Gamma, \Phi)$ corresponds to the sparse approximation set with the choice of $I_k = k$, $k \geq 1$. 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 it contains. Indeed, as will be shown, its metric entropy is larger by at most a logarithmic factor under the condition $I_k \leq k^\tau$ for some possibly large $\tau > 1$. If full approximation is used instead to approximate a sparse approximation set, one is actually approximating the class $F(\Gamma', \Phi)$, which has a much larger metric entropy than $S(\Gamma, \Phi)$ if $I_k$ is much bigger than $k$. For example, with $\gamma_k = k^{-\alpha}$ ($\alpha > 0$) and $I_k = k^\tau$ ($\tau > 1$), the metric entropy of $F(\Gamma', \Phi)$ is lower bounded by order $e^{-\tau/\alpha}$ while as will be seen the metric entropy of $S(\Gamma, \Phi)$ is between order $e^{-1/\alpha}$ and order $e^{-1/\alpha} \log (e^{-1})$.

For a class of examples, let $\Psi$ be a class of functions uniformly bounded by $c$ with a $L_2$ cover of cardinality of order $(1/c)^d$ (metric entropy $d \log (1/c)$) for some positive constant $d$. Let $C_\Psi$ be the closure of its convex hull. For $k \geq 1$, let $\varphi_{I_{k-1}+1}, \ldots, \varphi_{I_k}$ in $\Phi$ be the members of a $(1/(2k^{1/2}))$-cover of $\Psi$. Here $I_k - I_{k-1}$ is of order $k^{d/2}$. Then $C_\Psi \subseteq S(\Gamma, \Phi)$ with $\gamma_k = 4c/k^{1/2}$. 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 $C_{\Psi} \subset \mathcal{S}(\Gamma, \Phi)$ can be verified using greedy approximation (see, Jones (1992) or Barron (1993, Section VIII)), which we omit here. A specific example of $\Psi$ is $\{\sigma(a \cdot x + b) : a \in [-1, 1]^d, b \in \mathbb{R}\}$, where $\sigma$ is a fixed sigmoidal function satisfying a Lipschitz condition such as $\sigma(z) = (e^z - 1)/(e^z + 1)$, or a sinusoidal function $\sigma(z) = \sin(z)$.

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 on the metric entropy of $\mathcal{F}(\Gamma, \Phi)$ is a lower bound for that of $\mathcal{S}(\Gamma, \Phi)$. We next derive an upper bound. Let $k_\epsilon = \min\{k : \gamma_k \leq \epsilon/2\}$. Let $l_i \leq I_i, 1 \leq i \leq k_\epsilon$ be fixed for a moment. Consider the subset $\mathcal{G}_{\gamma_i, \ldots, \gamma_k} \subset \mathcal{S}(\Gamma, \Phi)$ in the span of $\phi_{l_1}, \ldots, \phi_{l_k}$ that has approximation errors bounded by $\gamma_1, \ldots, \gamma_k$ using the basis $\phi_{l_1}, \ldots, \phi_{l_k}$ (i.e., $g \in \mathcal{G}_{\gamma_i, \ldots, \gamma_k}$ if and only if $g = \sum_{i=1}^k a_i^* \phi_i$ for some coefficients $a_i^*$ and $\min_{a_1, \ldots, a_m} \|g - \sum_{i=1}^m a_i \phi_i\|_2 \leq \gamma_m$ for $1 \leq m \leq k_\epsilon$). From the previous section, we know the $\epsilon$-entropy of $\mathcal{G}_{\gamma_i, \ldots, \gamma_k}$ is upper bounded by order $\sum_{i=1}^{j(\epsilon)} N_i \approx k_\epsilon$. Based on the construction, it is not hard to see that an $\epsilon/2$-net in $\cup_{l_i \leq I_i, 1 \leq i \leq k_\epsilon} \mathcal{G}_{\gamma_i, \ldots, \gamma_k}$ is an $\epsilon$-net for $\mathcal{S}(\Gamma, \Phi)$. There are fewer than $\binom{k_\epsilon}{l_i}$ choices of the basis $\phi_i$, $1 \leq i \leq k_\epsilon$, thus the $\epsilon$-entropy of $\mathcal{S}(\Gamma, \Phi)$ is upper bounded by order $k_\epsilon + \log \binom{k_\epsilon}{l_i} = O(k_\epsilon \log (\epsilon^{-1}))$ under the assumption $k_\epsilon \leq k^\tau$ for some $\tau > 1$. As seen in Section 5, if $\gamma_k \sim k^{-\alpha} (\log k)^{-\beta}$, we have that $k_\epsilon$ is of order $\epsilon^{-1/\alpha} (\log (\epsilon^{-1}))^{-\beta/\alpha}$. Then the metric entropy of $\mathcal{S}(\Gamma, \Phi)$ is bounded by order $\epsilon^{-1/\alpha} (\log (\epsilon^{-1}))^{1-\beta/\alpha}$.

Thus we pay at most a price of a $\log (\epsilon^{-1})$ 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 bounded. Let $\mathcal{S}(\Gamma, \Phi)$ be all the probability density functions in $\mathcal{S}(\Gamma, \Phi)$. When $\gamma_0$ is large enough, the $L_2$ metric entropies of $\mathcal{F}(\Gamma, \Phi)$ and $\mathcal{S}(\Gamma, \Phi)$ are of the same order. Let $\epsilon_n$ satisfy $n \epsilon_n^2 = k_{\epsilon_n} \log (\epsilon_n^{-1})$ and $\epsilon_n$ satisfy $k_{\epsilon_n} = n \epsilon_n^2$. Then applying Theorem 1 and 3, we have

$$\epsilon_n^2 \leq \min_{\tilde{f}} \max_{f \in \mathcal{S}(\Gamma, \Phi)} E \| f - \tilde{f} \|_2^2 \leq \epsilon_n^2,$$

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

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

Note that upper and lower bound rates differ only in a logarithmic factor. Similarly, for estimating a regression function in $\mathcal{S}(\Gamma, \Phi)$, under the same conditions, by Theorem 10, we have

$$\epsilon_n^2 \leq \min_{\tilde{w}} \max_{w \in \mathcal{S}(\Gamma, \Phi)} E \| u - \tilde{w} \|_2^2 \leq \epsilon_n^2,$$

33
From the proofs of the minimax upper bounds, the estimators there are constructed based on Bayes averaging over the $\epsilon_n$-net of the sparse approximation set $S(\Gamma, \Phi)$. In this context, it is also natural to consider estimators based on subset selection. Upper bound results in this direction can be found in Yang and Barron (1996) and Barron, Birgé, and Massart (1996).

A general theory of sparse approximation should avoid requiring an assumption of orthogonality of the basis functions $\phi_1, \phi_2,$ etc.. In contrast with the story for full approximation sets where $F(\Gamma, \Phi)$ 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 (1994) of sparse orthonormal approximation for minimax statistical risks.

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

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

Here $C_1, C_2$ and $\beta$ are positive constants though $\beta$ may be quite small, e.g., $s/d$ with $s$ smaller than $d$. The condition $\sum_{i=1}^\infty |\xi_i|^2 \leq C_2 l^{-\beta}$ 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 $i$th 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 $\sum_{i=1}^\infty |\xi_i|^2 \leq C_2 l^{-\beta}$ is used to ensure that it suffices to select the $k$ largest coefficients from the first $I_k = k^\tau$ terms with sufficiently large $\tau$.

The class $S_q(C_1, C_2, \beta)$ is a special case of function classes with unconditional basis. Let $\mathcal{G}$ be a uniformly bounded function class and $\Phi = \{\phi_1 = 1, \phi_2, \ldots\}$ be an orthonormal basis in $L_2$. The basis $\Phi$ is said to be an unconditional for $\mathcal{G}$ if for any $g = \sum_{i=1}^\infty \xi_i \phi_i \in \mathcal{G}$, then $\sum_{i=1}^\infty \tilde{\xi}_i \phi_i \in \mathcal{G}$ for all $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ with $|\tilde{\xi}_i| \leq |\xi_i|$. 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 $q^* = q^*(\mathcal{G}) = \inf \{ q : \sup_{i \geq 1} i^{1/q} |\xi_i| < \infty \}$, where $\xi_1, \xi_2, \ldots$ are the coefficients ordered in decreasing magnitude $|\xi_1| \geq |\xi_2| \geq \ldots$. This $q^*$ is called the sparsity index in Donoho (1993, 1996). Let $\alpha^* = \alpha^*(\mathcal{G}) = \sup \{ \alpha : M_2(\epsilon) = O(\epsilon^{-1/\alpha}) \}$ be the optimal exponent of the $L_2$ metric entropy $M_2(\epsilon)$ of $\mathcal{G}$. From Donoho (1996), if $\mathcal{G}$ further satisfies the condition $\sum_{i=1}^\infty |\xi_i|^2 \leq C l^{-\beta}$ for all $l \geq 1$, then $\alpha^* = 1/q^* - 1/2$, i.e., for any $0 < \alpha_1 < \alpha^* < \alpha_2$, $\epsilon^{-1/\alpha_1} \leq M_2(\epsilon) \leq \epsilon^{-1/\alpha_2}$.
Applying Theorems 9 and 10, we have for regression,

\[ n^{-2\alpha_2/(2\alpha_1+1)} \leq \min_{\delta} \max_{u \in \mathcal{G}} E \| u - \hat{u} \|^2 \leq n^{-2\alpha_1/(2\alpha_1+1)} \].

Because \( \alpha_1 \) and \( \alpha_2 \) can be arbitrarily close to \( \alpha^* \), we know the exponential component of the minimax risk is \( n^{-2\alpha^*/(2\alpha^*+1)} \). For density estimation, let \( \mathcal{F}_{C'} = \{ f : f/C' \in \mathcal{G} \text{ and } f \text{ is a density} \} \). Then when \( C' \) is large enough, \( \mathcal{F}_{C'} \) has the same metric entropy order as that of \( \mathcal{G} \). Thus under the same conditions, we have for density estimation,

\[ n^{-2\alpha_2/(2\alpha_1+1)} \leq \min_{f} \max_{f \in \mathcal{F}_{C'}} E \| f - \hat{f} \|^2 \leq n^{-2\alpha_1/(2\alpha_1+1)} \]

for any \( 0 < \alpha_1 < \alpha^* < \alpha_2 \). Again the exponential component of the minimax risk is \( n^{-2\alpha^*/(2\alpha^*+1)} \).

For the special case of \( S_q(C_1, C_2, \beta) \), better entropy bounds are available. Indeed, based on Edmunds and Triebel (1987, p. 141), it can be shown that the \( L_2 \) metric entropy of \( S_q(C_1, C_2, \beta) \) is upper bounded by order \( e^{-1/(q-1/2)} \) when \( \beta/2 \geq 1/q - 1/2 \) and by order \( e^{-1/(q-1/2)} \log(e^{-1}) \) when \( 0 < \beta/2 < 1/q - 1/2 \). As a consequence, if \( S_q(C_1, C_2, \beta) \) is uniformly bounded, the minimax rate for the squared \( L_2 \) risk for estimating a regression function in \( S_q(C_1, C_2, \beta) \) or a density in \( S_q(C_1, C_2, \beta) \) (assuming \( C_1 \) and \( C_2 \) are suitably large for density estimation) is upper bounded by \( n^{-2\alpha/(2\alpha+1)} \) when \( \beta/2 \geq 1/q - 1/2 \) and by \( (n/\log n)^{-2\alpha/(2\alpha+1)} \) when \( 0 < \beta/2 < 1/q - 1/2 \), where \( \alpha = 1/q - 1/2 \).

7 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 smooth 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.

We will consider several function classes in the examples. Let \( \mu \) denote the Lebesgue measure.

1. Ellipsoidal classes in \( L_2 \).

Let \( \{ \phi_1, \phi_2, ..., \phi_k, ... \} \) be a complete orthonormal system in \( L_2[0,1] \). For an increasing sequence of constants \( b_k \) with \( b_1 \geq 1 \) and \( b_k \to \infty \), define an ellipsoidal class \( \mathcal{E}(\{b_k\}, C) = \{ g = \sum_{i=1}^\infty \xi_i \phi_i : \sum_{i=1}^\infty \xi_i^2 b_i^2 \leq C \} \). Define \( m(t) = \sup \{ i : b_i \leq t \} \) 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 $\mathcal{E}\{\{b_k\}, C\}$ satisfies

$$l(1/2\epsilon) \leq V_2(\epsilon) \leq l(8/\epsilon).$$

For the special case with $b_k = k^\alpha$ ($\alpha > 0$), the metric entropy order $\epsilon^{-1/\alpha}$ 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.)

2. Classes of functions with bounded mixed differences.

As in Temlyakov (1989), define the function classes $H^r_q$ on $\pi_d = [-\pi, \pi]^d$ having bounded mixed differences as follows. Let $k = (k_1, ..., k_d)$ be a vector of integers $q \geq 1$, and $r = (r_1, ..., r_d)$ with $r_1 = \cdots = r_v < r_{v+1} \leq \cdots \leq r_d$. Let $L_q(\pi_d)$ denote the periodic functions on $\pi_d$ with finite norm $\|g\|_{L_q(\pi_d)} = (2\pi)^{-d} \left( \int_{\pi_d} |g(x)|^q dx \right)^{1/q}$. Denote by $H^r_q$ the class of functions $g(x) \in L_q(\pi_d)$, $\int_{\pi_d} g(x) dx = 0$ for $1 \leq i \leq d$, and $\Delta^l\xi g(\chi) \|_{L_q(\pi_d)} \leq \prod_{j=1}^d |t_j|^{l_j}$, where $t = (t_1, ..., t_d)$ and $\Delta^l\xi$ is the mixed $l$th difference with step $t_j$ in the variable $x_j$, i.e., $\Delta^l\xi g(\chi) = \Delta^l_t g(x_1, ..., x_d)$.

From Temlyakov (1990), for $r = (r_1, ..., r_d)$ with $r_1 > 1$, $1 \leq p < \infty$, and $1 \leq q \leq \infty$,

$$M_p(\epsilon; H^r_q) \approx (1/\epsilon)^{1/r_1} (\log 1/\epsilon)^{(1+1/2r_1)(v-1)}.$$ 

Functions in this class are uniformly bounded.

3. Besov and Triebel classes.

Let $\Delta^r_h g, x) = \sum_{k=0}^r (-1)^r k g(x + kh)$. 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 $\omega_r(g, t_q) = \sup_{0 < h \leq t} \|\Delta^r_h g, \cdot\|_q$. Let $\alpha > 0$, $r = [\alpha] + 1$, and

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

Then the Besov norm is defined as $\|g\|_{B^\alpha,q} = |g|_{B^\alpha,q} + |g|_{B^\alpha,q}^\sigma$ (see DeVore and Lorentz (1993)). For Triebel or $F$ classes, let

$$|g|_{F^\alpha,q} = \left\{ \begin{array}{ll} (\int_0^\infty (t^{-\alpha} \sup_{0 < h \leq t} \|\Delta^r_h g, \cdot\|_q)^\sigma dt)^{1/\sigma} & \text{for } 0 < \sigma < \infty \\ \sup_{t > 0} t^{-\alpha} \sup_{0 < h \leq t} \|\Delta^r_h g, \cdot\|_q & \text{for } \sigma = \infty. \end{array} \right.$$ 

Then the $F$ norm is defined as $\|g\|_{F^\alpha,q} = |g|_{F^\alpha,q} + |g|_{F^\alpha,q}^\sigma$. For definitions and characterizations of Besov and $F$ classes in the $d$-dimensional case see Triebel (1992). 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 $B_{\sigma,q}^\alpha(C)$ and $F_{\sigma,q}^\alpha(C)$ be the collections of all functions $g \in L_q[0,1]^{d}$ such that $\| g \|_{B_{\sigma,q}^\alpha} \leq C$ and $\| g \|_{F_{\sigma,q}^\alpha} \leq C$ respectively. Building on the conclusions previously obtained for Sobolev classes by Birman and Solomjak (1967, 1974), Triebel (1975), with refinements in Carl (1981), showed that for $1 \leq \sigma \leq \infty$, $1 \leq p, q \leq \infty$, and $\alpha/d > 1/q - 1/p$,

$$M_p(\epsilon; B_{\sigma,q}^\alpha(C)) \sim \epsilon^{-d/\alpha}.$$  

Using the inclusion relationship between Besov and $F$ classes: $B_{\min(q,\sigma),q}^\alpha(C_1) \subset F_{\sigma,q}^\alpha(C) \subset B_{\max(q,\sigma),q}^\alpha(C_2)$ for some constants $C_1$ and $C_2$, under the above condition on the parameters except now requiring $q < \infty$, we have

$$M_p(\epsilon; F_{\sigma,q}^\alpha(C)) \sim \epsilon^{-d/\alpha}.$$  

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 $V(g) := \sup_{i=0}^{\infty} \frac{1}{m} \sum_i |g(x_{i+1}) - g(x_i)| \leq C$, where the supremum is taken for all finite sequence $x_1 < x_2 < \cdots < x_m$ in $[0, 1]$. For $0 < \alpha \leq 1$, let $\text{Lip}_{\alpha,q}(C) = \{ g : \| g(x+h) - g(x) \|_{q} \leq C h^\alpha \text{ and } \| g \|_{q} \leq C \}$ be a Lipschitz class. When $\alpha > 1/q - 1/p$, $1 \leq p, q \leq \infty$, the metric entropy satisfies $M_p(\epsilon; \text{Lip}_{\alpha,q}(C)) \sim \epsilon^{-1/\alpha}$, see Birman and Solomjak (1967, 1974); related conclusions are in Kolmogorov and Tihomirov (1959) and Clements (1963). This conclusion for Lipschitz classes can be understood now as special cases of the conclusions for Sobolev and Besov classes, though the Lipschitz classes play a special role in determining the metric entropy for the larger classes, especially in obtaining lower bounds, see Lorentz, Golitschek and Makovoz (1996), Chapter 15. 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

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

and since the $L_p$ ($1 \leq p < \infty$) metric entropies of $\text{Lip}_{1,1}(C)$ and $\text{Lip}_{1,\infty}(C)$ are both of order $1/\epsilon$, the $L_p$ ($1 \leq p < \infty$) metric entropy of $BV(C)$ is also of order $1/\epsilon$.

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 modulus of continuity of derivatives. Let $\Lambda_{r,\omega}^{d,2} = \Lambda_{r,\omega}^{d,2}(C_0, C_1, ..., C_r)$ be the collection of all functions
$g$ on $[0, 1]^d$ which have all partial derivatives $\| D^k g \|_2 \leq C_k$, $|k| = k = 0, 1, \ldots, r$, and the modulus of continuity in the $L_q$ norm of each $r$th derivative is bounded by a function $\omega$. Here $\omega$ is any given modulus of continuity (for definition, see DeVore and Lorentz (1993, p. 41)). Let $\delta = \delta(\epsilon)$ be defined by equation $\delta^r \omega(\delta) = \epsilon$. Then if $r \geq 1$, again from Lorentz (1966), the $L_2$ metric entropy of $\Lambda_{\omega, d}^r$ is of order $O^{\alpha}(r \delta^{-d})$. When $r = 0$ and $\omega$ is a concave modulus of continuity, the sup-norm metric entropy order is identified by Timan (1964) for similar classes.

6. Classes of functions with different moduli of smoothness with respect to different variables.

Let $k_1, \ldots, k_d$ be positive integers and $0 < \beta_i \leq k_i, 1 \leq i \leq d$. Let $k = (k_1, \ldots, k_d)$ and $\beta = (\beta_1, \ldots, \beta_d)$. Let $V(k, \beta, C)$ be the collection of all functions $g$ on $[0, 1]^d$ with $\| g \|_{L^\beta} \leq C$ and $\sup_{|k| \leq \beta} \| \Delta_{i, h}^k g \|_2 \leq C \kappa_i h$, where $\Delta_{i, h}^k$ is the $k_i$th difference with step $h$ in variable $x_i$. As stated in Lorentz (1966, p. 921), from the metric entropy results on full approximation sets together with polynomial approximation results of Timan (1963, Section 5.3), the $L_2$ metric entropy order of $V(k, \beta, C)$ is $(1/\epsilon)^{\sum_{i=1}^d \beta_i^{-1}}$. Similar results in terms of sup-norm metric were given in Brudnyi and Kotliar (1963).

7. Classes $E_{d, k}^{\alpha, \beta}(C)$ and $G_{d, k}^{\alpha, \beta}(C)$.

Let $E_{d, k}^{\alpha, \beta}(C)$ ($\alpha > 1/2$ and $k \geq 0$) be the collection of periodic functions

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

on $[0, 2\pi]$ with $\sqrt{a_{m_1, \ldots, m_d}^2 + b_{m_1, \ldots, m_d}^2} \leq C (m_1 \cdots m_d)^{-\alpha} \log^k (m_1 \cdots m_d) + 1$, where $\overline{m} = m$ if $m \neq 0$ and $\overline{0} = 1$. Similarly define $G_{d, k}^{\alpha, \beta}(C)$ ($\alpha > 0$) with the constraint $\sum (m_1 \cdots m_d)^{2\alpha} \left( a_{m_1, \ldots, m_d}^2 + b_{m_1, \ldots, m_d}^2 \right) \leq C^2$. From Smoljak (1960), the $L_2$ metric entropies of $E_{d, k}^{\alpha, \beta}(C)$ and $G_{d, k}^{\alpha, \beta}(C)$ are of order $(1/\epsilon)^{1/(\alpha-1/2) \log (2k+2\alpha(d-1))/(2\alpha-1)}(1/\epsilon)$ and $(1/\epsilon)^{1/\alpha \log^d(1/\epsilon)}$ respectively. Note that for these two classes, the dependence of entropy orders on the input dimension $d$ is only through logarithmic factors.

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 $g(x) = c_0 + \sum_{i=1}^d c_i (v_i \cdot x + b_i)$, with $|c_0| + \sum_{i} |c_i| \leq C$, and $|v_i| = 1$, where $\sigma$ is a fixed sigmoidal function with $\sigma(t) \to 1$ as $t \to \infty$ and $\sigma(t) \to 0$ as $t \to -\infty$. We further require that $\sigma$ is either the step function $\sigma^s(t) = 1$ for $t \geq 0$, and $\sigma^s(t) = 0$ for $t < 0$, or satisfies the Lipschitz requirement that $|\sigma(t) - \sigma(t')| \leq C_1 |t - t'|$ for some $C_1$ and $|\sigma(t) - \sigma^s(t)| \leq C_2 |t|^{-\gamma}$ for some $C_2$ and $\gamma > 0$ for all $t \neq 0$. 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, p. 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, pp. 108-109) 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 $(1/\epsilon)^{1/(1/2+1/d)} \leq M_2(\epsilon) \leq (1/\epsilon)^{1/(1/2+1/(2d))} \log (1/\epsilon)$. For $d = 2$, a better lower bound matches the upper bound in the exponent $(1/\epsilon)^{4/3} \log^{-2/3} (1/\epsilon) \leq M_2(\epsilon) \leq (1/\epsilon)^{4/3} \log (1/\epsilon)$. For $d = 1$, $N(C)$ is equivalent to a bounded variation class.

Density estimation.

1. Assume the functions in $E(\{b_k\}, C)$ 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 $\tilde{E}(\{b_k\}, C)$ be the collection of probability density functions in $E(\{b_k\}, C)$. It is easily seen that a linear transform of $E(\{b_k\}, C)$, e.g., $\{(g + C_0 + 1)/(f g d\mu + C_0 + 1) : g \in E(\{b_k\}, C)\}$ is contained in $\tilde{E}(\{b_k\}, C')$ for some $C'$ (as in Section 5). Thus when $C$ is suitably large, $\tilde{E}(\{b_k\}, C)$ and $E(\{b_k\}, C)$ have the same order of $L_2$ metric entropies. By Theorem 4, let $\epsilon_n$ be determined by $l(1/\epsilon_n) = n\epsilon_n^2$, we have

$$\min \max_{f \in \tilde{E}(\{b_k\}, C)} E \| f - \hat{f} \|_2^2 \approx \epsilon_n^2.$$ 

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 $E(\{b_k\}, C)$ are uniformly bounded. Then

$$\min \max_{f \in \tilde{E}(\{k^\alpha\}, C)} E \| f - \hat{f} \|_2^2 \approx n^{-2\alpha/(2\alpha+1)}.$$ 

See Efroimovich and Pinsker (1982) for more detailed asymptotics with the trigonometric basis and $b_k = k^\alpha$, and see Birgé (1983) for similar conclusions using a hypercube construction for the trigonometric basis and Barron, Birgé and Massart (1996, Section 3) for general ellipsoids.

2. Let $\tilde{H}_r^\delta(C) = \{ f = e^\delta \int e^\delta d\mu : g \in H_r^\delta(C) \}$ be a uniformly bounded density class. From Lemma 14, the metric entropy of $\tilde{H}_r^\delta(C)$ is of the same order as for $H_r^\delta(C)$. So for $r_1 > 1$ and $1 \leq q \leq \infty$, $1 \leq p \leq 2$, by Corollary 3, we have

$$\min \max_{f \in \tilde{H}_r^\delta} E \| f - \hat{f} \|_p \approx n^{-r_1/(2r_1+1)} (\log n)^{(r-1)/2}.$$ 

39
For this density class, the minimax risk under K-L distance or squared Hellinger distance is also of the same order as that under the squared $L_2$ distance, $n^{-2r_1/(2r_1+1)} (\log n)^{(r_1-1)}$.

3. Let $\tilde{B}_{\sigma,q}^\alpha (C)$ be the set of probability density functions in the Besov class $B_{\sigma,q}^\alpha (C)$. Similarly define $\tilde{F}_{\sigma,q}^\alpha (C)$. When $\alpha/d > 1/q$, the functions in $B_{\sigma,q}^\alpha (C)$ and $F_{\sigma,q}^\alpha (C)$ are uniformly bounded. Similarly to $\tilde{E}(\{b_k\}, C)$, when $C$ is suitably large, the metric entropy of $\tilde{B}_{\sigma,q}^\alpha (C)$ or $\tilde{F}_{\sigma,q}^\alpha (C)$ is of the same order as for $B_{\sigma,q}^\alpha (C)$ or $F_{\sigma,q}^\alpha (C)$. Thus from Theorem 4, we have, when $C$ is large enough, for $1 \leq \sigma \leq \infty$, $1 \leq q \leq \infty$, $1 \leq p \leq 2$, and $\alpha/d > 1/q$,

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

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

The above conclusions also hold for $\tilde{F}_{\sigma,q}^\alpha (C)$ ($q < \infty$). When $1/q - 1/2 < \alpha/d \leq 1/q$, some functions in $B_{\sigma,q}^\alpha (C)$ or $F_{\sigma,q}^\alpha (C)$ are not bounded, nevertheless this class has the same order metric entropy as the subclass $B_{\sigma,\infty}^\alpha (C)$ or $F_{\sigma,\infty}^\alpha (C)$ with $q^* > d/\alpha$ which is uniformly bounded. Consequently for $C$ sufficiently large the metric entropy of $\tilde{B}_{\sigma,q}^\alpha (C)$ or $\tilde{F}_{\sigma,q}^\alpha (C)$ is still the same order as for $B_{\sigma,q}^\alpha (C)$ or $F_{\sigma,q}^\alpha (C)$. From Theorem 4, for $L_1$ risk we have, when $\alpha/d > 1/q - 1/2$, for $\mathcal{F} = \tilde{B}_{\sigma,q}^\alpha (C)$ or $\tilde{F}_{\sigma,q}^\alpha (C)$,

$$\min_{f} \max_{\hat{f} \in \mathcal{F}} E \| f - \hat{f} \|_1 \asymp n^{-\alpha/(2\alpha+d)}.$$ 

From the monotonicity property of the $L_p$ norm in $p$, we have for $p > 1$, and $\alpha/d > 1/q - 1/2$,

$$\min_{f} \max_{\hat{f} \in \mathcal{F}} E \| f - \hat{f} \|_p \asymp n^{-\alpha/(2\alpha+d)}.$$ 

In particular when $q \geq 2$ the last two conclusions hold for all $\alpha > 0$. Donoho, Johnstone, Kerkyacharian and Picard (1993) obtained suitable minimax bounds for the case of $p \geq q$, $\alpha > 1/q$. Here we permit smaller $\alpha$. The rates for Lipschitz classes are 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 $B_{\sigma,q}^\alpha (C)$, then by Corollary 3, for $\alpha/d > 1/q$, the minimax risk under K-L distance is of the same order $n^{-2\alpha/(2\alpha+d)}$, which was previously shown by Koo and Kim (1996).

4. Let $\overline{BV} (C)$ be the class of all density functions in $BV (C)$. When $C$ is suitably large, the $L_p \ (1 \leq p < \infty)$ metric entropy of $\overline{BV} (C)$ is of order $1/\epsilon$. By Theorem 4, for $1 \leq p \leq 2$,

$$\min_{f} \max_{\hat{f} \in \overline{BV} (C)} E \| f - \hat{f} \|_p^2 \asymp n^{-2/3},$$

40
5. Let $\Lambda_{r,\omega}^{d,2}$ be the density functions in moduli of derivative class $\Lambda_{r,\omega}^{d,2}$. For $r \geq 1$, when $C_0$ is suitably large, $\Lambda_{r,\omega}^{d,2}$ and $\Lambda_{r,\omega}^{d,2}$ have the same order $L_2$ metric entropies. Let $\epsilon_n$ be chosen such that $(\delta(\epsilon_n))^{-d} = n\epsilon_n^2$. Then
\[
\min_{f} \max_{\tilde{f} \in B(V(C))} E \| f - \tilde{f} \|_2 \leq n^{-1/3}.
\]
6. Let $V(k, \beta, C)$ be densities in $V(k, \beta, C)$. When $C$ is not too small, the $L_2$ metric entropy of $V(k, \beta, C)$ is also of order $(1/\epsilon)^{\sum_{i=1}^{d} \beta_i^{-1}}$. Let $\alpha = \sum_{i=1}^{d} \beta_i^{-1}$, then
\[
\min_{f} \max_{\tilde{f} \in V(k, \beta, C)} E \| f - \tilde{f} \|_2 \leq n^{-2\alpha/(2\alpha+d)}.
\]
7. Let $E_d^{\alpha,k}(C)$ and $G_d^{\alpha}(C)$ be densities functions in $E_d^{\alpha,k}(C)$ and $G_d^{\alpha}(C)$ respectively. They are uniformly bounded when $\alpha > 1$ and $\alpha > 1/2$ respectively. Then when $C$ is not too small, we have
\[
\min_{f} \max_{\tilde{f} \in E_d^{\alpha,k}(C)} E \| f - \tilde{f} \|_2 \leq n^{-(\alpha-1/2)/\alpha \log(k+\alpha(d-1))/\alpha} n,
\]
\[
\min_{f} \max_{\tilde{f} \in G_d^{\alpha}(C)} E \| f - \tilde{f} \|_2 \leq n^{-2\alpha/(2\alpha+1) \log(2\alpha(d-1)/(2\alpha+1))} n.
\]
Note that for both classes, the dimension $d$ does not appear in the exponents of sample size $n$.
8. Let $N(C)$ be the densities in $N(C)$. When $C$ is not too small, we have
\[
n^{-(1+2/d)/(2+1/d)} (\log n)^{-(1+1/d)(1+2/d)/(2+1/d)} \leq \min_{f} \max_{\tilde{f} \in N(C)} E \| f - \tilde{f} \|_2 \leq (n/\log n)^{-(1+1/d)/(2+1/d)},
\]
and for $d = 2$, using the better lower bound on the metric entropy, we have
\[
n^{-3/5} (\log n)^{-19/10} \leq \min_{f} \max_{\tilde{f} \in N(C)} E \| f - \tilde{f} \|_2 \leq (n/\log n)^{-3/5}.
\]
Previously Modha and Masry (1994) (see also Yang and Barron (1996)) obtain upper rate $n^{-1/2}$ (ignoring a logarithmic factor) for a similar class by feedforward neural network models using minimum description length criterion and approximation results of Barron (1993, 1994). Here the rates are slightly better. Though the upper and lower rates do not agree (except $d = 2$), when $d$ is moderately large, the rates are roughly $n^{-1/2}$, which is independent of $d$. 

41
Regression function estimation.

Consider the regression problem in Section 4. Let $h$ be the density of the explanatory variable $X$. Assume $\| \log h \|_\infty$ is finite, then for the considered function classes, the metric entropies under $L_2(h)$ norm are of the same orders as given before. By Theorems 9 and 10, we have the following conclusions.

1. Assume $l(t)$ satisfies $\liminf_{t \to 0} l(\beta t)/l(t) > 1$ for some constant $\beta > 1$. Let $\epsilon_n$ be determined by $l(1/\epsilon_n) = n \epsilon_n^2$, we have the lower bound

$$\min_u \max_{v \in \mathcal{E}(c_{b,k},C)} E \| u - \hat{u} \|_2^2 \geq \epsilon_n^2.$$

If the functions in $\mathcal{E}(\{b_k\},C)$ are uniformly bounded, then

$$\min_u \max_{v \in \mathcal{E}(c_{b,k},C)} E \| u - \hat{u} \|_2^2 \times \epsilon_n^2.$$

Specially, if $b_k = k^\alpha$, $k \geq 1$ for some $\alpha > 0$, we have

$$\min_u \max_{v \in \mathcal{E}(c_{k^\alpha},C)} E \| u - \hat{u} \|_2^2 \geq n^{-2\alpha/(2\alpha+1)}.$$

If $\alpha > 1/2$, then the functions in $\mathcal{E}(\{k^\alpha\},C)$ are uniformly bounded and

$$\min_u \max_{v \in \mathcal{E}(c_{k^\alpha},C)} E \| u - \hat{u} \|_2^2 \times n^{-2\alpha/(2\alpha+1)}.$$

For results giving the right constant in addition to the rate in this special ellipsoidal case with trigonometric basis, see Pinsker (1980).

2. For $H_q^r(C)$, with $r_1 > 1$ and $1 \leq q \leq \infty$, $1 \leq p \leq 2$, we have

$$\min_u \max_{v \in H_q^r(C)} E \| u - \hat{u} \|_p^2 \times n^{-2r_1/(2r_1+1)} (\log n)^{(q-1)}.$$

3. For $B_{\infty,q}^\alpha(C)$ or $F_{\infty,q}^\alpha(C)$, with $1 \leq \sigma \leq \infty$, $1 \leq q \leq \infty$ or $1 \leq q < \infty$, $1 \leq p \leq 2$, and $\alpha/d > 1/q$, the rates for $U = B_{\infty,q}^\alpha(C)$ or $F_{\infty,q}^\alpha(C)$ are

$$\min_u \max_{v \in U} E \| u - \hat{u} \|_p^2 \times n^{-2\alpha/(2\alpha+d)}.$$

When $\alpha/d > 1/q - 1/2$ and $p \geq 1$, we have,

$$\min_u \max_{v \in U} E \| u - \hat{u} \|_p^2 \times n^{-2\alpha/(2\alpha+d)}.$$

Donoho, Johnstone (1994) obtained the minimax rates for $p = 2$, $\alpha > 1/q$ and $d = 1$. As pointed out in their paper, whenever $q < 2$, traditional linear methods can not achieve the
minimax optimal rate. Previously many results have been obtained for Sobolev classes including the following. For fixed design regression estimation, when \( \alpha \) is a positive integer, Nemirovskii (1985) obtained both upper and lower bounds on the minimax risks not only suitable for \( 1 \leq p \leq 2 \), but also for \( p > 2 \); See Ibragimov and Hasminskii (1977a) for results on Gaussian white noise models for \( p \geq 2 \). Stone (1982) gives optimal rates for estimating a regression function or its derivatives for integer \( \alpha \) and \( 0 < p \leq \infty \). For some refinement results on the minimax risk for some Sobolev classes with \( p = 2 \) (not only the rate but also the right constant), see Nussbaum (1985).

4. For \( BV(C) \), we have for \( 1 \leq p \leq 2 \),

\[
\min_{\hat{u}} \max_{u \in BV(C)} E \| u - \hat{u} \|_p^2 \asymp n^{-2/3}.
\]

The result in also given in Donoho, Johnstone (1994).

5. Let \( \epsilon_n \) be chosen such that \( (\delta(\epsilon_n))^{-d} = n\epsilon_n^2 \). For \( r \geq 1 \), we have

\[
\min_{\hat{u}} \max_{u \in \mathcal{N}_{\epsilon_n^r}} E \| u - \hat{u} \|_2^2 \asymp \epsilon_n^2.
\]

6. Let \( \alpha^{-1} = \sum_{i=1}^{d} \beta_i^{-1} \), then

\[
\min_{\hat{u}} \max_{u \in V(k, \beta, C)} E \| u - \hat{u} \|_2^2 \asymp n^{-2\alpha/(2\alpha + d)}.
\]

7. When \( \alpha > 1/2 \) for \( E_{d}^{\alpha,k}(C) \) and \( \alpha > 0 \) for \( G_{d}^{\alpha}(C) \), we have lower bounds

\[
\min_{\hat{u}} \max_{u \in E_{d}^{\alpha,k}(C)} E \| u - \hat{u} \|_2^2 \asymp n^{-(\alpha-1/2)/\alpha \log((k^{1+d})/\alpha) n},
\]

\[
\min_{\hat{u}} \max_{u \in G_{d}^{\alpha}(C)} E \| u - \hat{u} \|_2^2 \asymp n^{-2\alpha/(2\alpha + 1) \log(2\alpha - 1)/(2\alpha + 1) n}.
\]

When \( \alpha > 1 \) and \( \alpha > 1/2 \) respectively for the two classes, we have

\[
\min_{\hat{u}} \max_{u \in E_{d}^{\alpha,k}(C)} E \| u - \hat{u} \|_2^2 \asymp n^{-(\alpha-1/2)/\alpha \log(k^{1+d})/\alpha) n},
\]

\[
\min_{\hat{u}} \max_{u \in G_{d}^{\alpha}(C)} E \| u - \hat{u} \|_2^2 \asymp n^{-2\alpha/(2\alpha + 1) \log(2\alpha - 1)/(2\alpha + 1) n}.
\]

8. For general \( d \geq 1 \), we have

\[
n^{-(1+2/d)/(2+1/d) (\log n)^{-(1+1/d)(1+2/d)/(2+1/d)} \leq \min_{\hat{u}} \max_{u \in V(C)} E \| u - \hat{u} \|_2^2 \asymp (n/\log n)^{-(1+1/d)/(2+1/d)},
\]

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

\[
n^{-3/5} (\log n)^{-19/10} \leq \min_{\hat{u}} \max_{u \in V(C)} E \| u - \hat{u} \|_2^2 \asymp (n/\log n)^{-3/5}.
\]

Similar upper bound results are initially obtained by Barron (1991, 1994) using neural network models.
Data compression.

Because K-L distance is lower bounded by half the squared $L_1$ distance, according to the relationship between density estimation and data compression, we know that the minimax redundancy for compressing an i.i.d. data string governed by a density in $\tilde{B}_{\alpha,q}^\sigma(C)$ or $\tilde{F}_{\alpha,q}^\sigma(C)$ with $\alpha/d > 1/q - 1/2$ is lower bounded in rate as follows (as a consequence of Theorem 7):

$$
\min_{q_n \in Q_n} \max_{f \in \tilde{B}_{\alpha,q}^\sigma(C)} \frac{1}{n} D(f^n \parallel q_n) \geq n^{-2\alpha}. \tag{7.1}
$$

This rate is also obtained by Yu (1996) for special Lipschitz classes through a hypercube argument to bound the mutual information between the parameter and the observations.

If we assume $\log f \in B_{\alpha,q}^\sigma(C)$ with $\alpha/d > 1/q$, then the minimax redundancy rate is identified from Theorem 7:

$$
\min_{q_n \in Q_n} \max_{f \in B_{\alpha,q}^\sigma(C)} \frac{1}{n} D(f^n \parallel q_n) \asymp n^{-\frac{\log q}{(2\alpha+d)}}. \tag{7.2}
$$

Similar results could be stated for the other function classes considered in this section.

8 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 involve local metric entropy calculations. 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 efforts in search for 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 $M(e)$.

**Definition:** (Local metric entropy) The local $\epsilon$-entropy at $\theta \in S$ is the logarithm of the largest $(\epsilon/2)$-packing set in $B(\theta, \epsilon) = \{\theta' \in S : d(\theta', \theta) \leq \epsilon\}$. The local $\epsilon$-entropy at $\theta$ is denoted by
$M(\epsilon \mid \theta)$. The local $\epsilon$-entropy of $S$ is defined as

$$M^{\text{loc}}(\epsilon) = \max_{\theta \in S} M(\epsilon \mid \theta).$$

**Lemma 7:** The global and local metric entropies have the following relationship:

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

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

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

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 $\theta^* \in N_\epsilon$ with $|R_{\theta^*}| \geq |N_{\epsilon/2}|/|N_\epsilon|$. Thus we have

$$M(\epsilon \mid \theta^*) \geq M(\epsilon/2) - M(\epsilon).$$

On the other hand, by concavity of the log function,

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

Thus an average of the local $\epsilon$-entropies is upper bounded by the difference of two global metric entropies $M(\epsilon/2) - M(\epsilon)$. An obvious upper bound on the local $\epsilon$-entropies is $M(\epsilon/2)$. This ends the proof of Lemma 7.

The sets $R_\theta$ 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 $R_\theta$ are $\epsilon/2$ apart from each other and are within $\epsilon$ from $\theta$. This property together with the assumption that locally $d_K$ is upper bounded by a multiple of $d$ enables the use of Birge’s result (Proposition 2.8, 1983) 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 $\epsilon_0 > 0$ such that

$$D(p_\theta \parallel p_{\theta'}) \leq \overline{A}d^2(\theta, \theta'), \text{ for } \theta, \theta' \in S \text{ with } d(\theta, \hat{\theta}) \leq \epsilon_0.$$  

From Lemma 7, there exist $\theta^* \in S$ and a subset $S_0 = R_{\theta^*}$ with a local packing set $N$ of log-cardinality at least $M(\epsilon/2) - M(\epsilon)$. Using Fano’s inequality together with the diameter bound on mutual information as in Birgé, yields with $\theta$ a uniformly distributed random variable on $N$, when $\epsilon \leq \epsilon_0$,

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

and hence, choosing $\epsilon_n$ to satisfy

$$M(\epsilon_n/2) - M(\epsilon_n) = 2 \left( n\overline{A}\epsilon_n^2 + \log 2 \right),$$

as in the proof of Theorem 1, but with $S$ replaced by $S_0 = R_{\theta^*}$, one gets

$$\min_{\hat{\theta}} \max_{\theta \in S_0} E_\theta d^2(\theta, \hat{\theta}) \geq \frac{\epsilon_n^2}{32},$$

which is similar to our conclusions from Section 2 (cf. Corollary 2). The difference in the bound just obtained compared with the previous work of Birgé and others is the use of Lemma 7 to avoid requiring explicit construction of the local packing set.

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

$$\min_{\hat{\theta}} \max_{\theta \in S_0} E_\theta d^2(\theta, \hat{\theta}) \geq \overline{\epsilon}_n^2,$$

where $\epsilon = \overline{\epsilon}_n$ is determined by

$$M^{loc}(\overline{\epsilon}_n) = n\overline{\epsilon}_n^2,$$

provided $D(p_\theta \parallel p_{\theta'}) \leq \overline{A}d^2(\theta, \theta')$ holds for $\theta, \theta'$ 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 $\epsilon$-entropy of a usual parametric class is often of order $m \log (1/\epsilon)$, where $m$ is the dimension of the model. Then the metric entropy difference $M(\epsilon/2) - M(\epsilon)$ or $M^{loc}(\epsilon)$ is of order of a constant, yielding the anticipated rate $\epsilon_n \asymp 1/\sqrt{n}$ and $\overline{\epsilon}_n \asymp 1/\sqrt{n}$. The achievability of the rate in (7) is due to Birgé (Theorem 3.1, 1986) under the following condition.

**Condition 5:** There is a nonincreasing function $U : (0, \infty) \to [1, \infty)$. For any $\epsilon > 0$ with $n\epsilon^2 \geq U(\epsilon)$, there exists an $\epsilon$-net $S_\epsilon$ for $S$ under $d$ such that for all $\lambda \geq 2$ and $\theta \in S$, we have
card(\(S \cap B_d(\theta, \lambda \epsilon)\)) \leq \lambda^{U(\epsilon)}. 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 \(\epsilon\), an \(\epsilon\)-packing set \(N_\epsilon\) for \(B_d(\theta_0, \lambda \epsilon)\) satisfying \(\text{card}(N_\epsilon \cap B_d(\theta_0, \lambda \epsilon)) \geq (\lambda/A_1)^{U(\epsilon)}\) and \(d^2_K(\theta, \bar{\theta}) \leq A_1^2 d^2(\theta, \bar{\theta})\) for all \(\bar{\theta}\) in the packing set \(N_\epsilon\).

**Proposition 2**: (Birgé) Suppose Condition 5 is satisfied. If the distance \(d\) is a metric bounded above by a multiple of the Hellinger distance, then

\[
\min_{\theta \in A_n} \max_{\bar{\theta} \in S} E_0 d^2(\theta, \bar{\theta}) < \epsilon_n^2,
\]

where \(\epsilon_n\) satisfies \(n \epsilon_n^2 = U(\epsilon_n)\).

The upper bound is from Birgé (Theorem 3.1, 1986) using the first part of Condition 5. From the second part of condition 5 with \(\lambda = 2\), we have a \(M^{loc}(2\epsilon) \asymp U(\epsilon)\) and a suitable relationship between \(d_K\) and \(d\) in a maximal order local \(\epsilon\)-net, so the lower bound follows from (8) by the Fano inequality bound as discussed above in accordance with Birgé (Proposition 2.8, 1983).

When \(\lim_{\epsilon \to 0} M(\epsilon/2)/M(\epsilon) > 1\) (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 considering global or local metric entropy. The condition \(\lim_{\epsilon \to 0} M(\epsilon/2)/M(\epsilon) > 1\) is characteristic of large function classes as we have seen. In that case, the three entropy quantities in Lemma 7 are asymptotically equivalent as \(\epsilon \to 0\),

\[
M(\epsilon/2) - M(\epsilon) \asymp M^{loc}(\epsilon) \asymp M(\epsilon/2).
\]

In contrast, when \(\lim_{\epsilon \to 0} M(\epsilon/2)/M(\epsilon) = 1\) as is typical of finite-dimensional parametric cases, \(M(\epsilon/2) - M(\epsilon)\) is of smaller order than \(M(\epsilon/2)\). In this case, we may use (8) 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 \(\lim_{\epsilon \to 0} M(\epsilon/2)/M(\epsilon) = 1\), under a stronger homogeneous entropy assumption (Condition 5), 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 (1995) and Yang and Barron (1996)).

47
9 Appendix: Proofs of some lemmas used for main results

Proof of Lemma 4: We are to show that Condition 0 is satisfied under the given conditions. Assume \( \max(D(p_\theta || p), D(p_{\theta'}) || p) \leq e^2 \). Since \( (p_\theta + p_{\theta'}) / 2 \) minimizes \( \frac{1}{2} D(p_\theta || p) + \frac{1}{2} D(p_{\theta'} || p) \) over all densities \( p \), then \( \max(D(p_\theta || (p_\theta + p_{\theta'}) / 2), D(p_{\theta'} || (p_\theta + p_{\theta'}) / 2) \leq 2e^2 \). So under the first condition, if \( \epsilon \leq \epsilon_0 = \epsilon_* / \sqrt{2} \), then \( D(p_\theta || (p_\theta + p_{\theta'}) / 2) + D(p_{\theta'} || (p_\theta + p_{\theta'}) / 2) \geq A^2 D(p_\theta || p_{\theta'}) \). As a consequence, \( d_K(\theta, \theta') \leq d_K((\theta, \epsilon_0 + \delta) + d_K(\theta, \theta) \) for any \( \theta \). For the second condition, because

\[
D \left( p_\theta || p_{\theta'} \right) + D \left( p_{\theta'} || p_\theta \right) = \int p_\theta \log \frac{p_\theta}{p_{\theta'}} + \int p_{\theta'} \log \frac{p_{\theta'}}{p_\theta} = -2 \int (p_\theta + p_{\theta'}) / 2 \log \frac{(p_\theta + p_{\theta'}) / 2}{p_\theta} = -2D \left( \frac{p_\theta + p_{\theta'}}{2} || p_\theta \right),
\]

we have

\[
D(p_\theta || p_{\theta'}) = D \left( p_\theta || \frac{p_\theta + p_{\theta'}}{2} \right) + D \left( p_{\theta'} || \frac{p_\theta + p_{\theta'}}{2} \right) + 2D \left( \frac{p_\theta + p_{\theta'}}{2} || p_\theta \right).
\]

(This equality is a special case of a parallelogram identity, see Csiszár and Körner (1981, pp. 59)). Thus, when \( \max(D(p_\theta || (p_\theta + p_{\theta'}) / 2), D(p_{\theta'} || (p_\theta + p_{\theta'}) / 2) \leq 2e_0^2 \),

\[
D(p_\theta || p_{\theta'}) \leq D \left( p_\theta || \frac{p_\theta + p_{\theta'}}{2} \right) + D \left( p_{\theta'} || \frac{p_\theta + p_{\theta'}}{2} \right) + \left( \frac{1}{A^2} - 1 \right) D \left( p_\theta || \frac{p_\theta + p_{\theta'}}{2} \right)
\]

\[
\leq \frac{1}{A^2} \left( D \left( p_\theta || \frac{p_\theta + p_{\theta'}}{2} \right) + D \left( p_{\theta'} || \frac{p_\theta + p_{\theta'}}{2} \right) \right).
\]

Thus the first condition is satisfied.

Proof of Lemma 5: The proof is by a truncation of \( g \) from above and below. Let \( G = \{ x : g(x) \leq 4T \} \). Then because \( d_H(f, g) \leq \epsilon \), we have \( \int_{G^c} \left( \sqrt{f} - \sqrt{g} \right)^2 d\mu \leq \epsilon^2 \). Since \( f(x) \leq T \leq g(x) / 4 \) for \( x \in G^c \), it follows that \( \int_{G^c} \left( \sqrt{g} - \sqrt{f} \right)^2 d\mu \leq \int_{G^c} \left( \sqrt{f} - \sqrt{g} \right)^2 d\mu \leq \epsilon^2 \). Thus \( \int_{G^c} g d\mu \leq 4e^2 \), which implies \( 1 - 4e^2 \leq \int g d\mu \leq 1 \) and \( \int \left( \sqrt{g} - \sqrt{f} \right)^2 d\mu \leq \int_{G^c} g d\mu \leq 4e^2 \). Let \( \bar{g} = (g + 4e^2) / (\int g d\mu + 4e^2) \). Clearly \( \bar{g} \) is a probability density function with respect to \( \mu \). For \( 0 \leq z \leq 4T \), by simple calculation using \( 1 - 4e^2 \leq \int g d\mu \leq 1 \), we have \( |\sqrt{z} - \sqrt{\bar{g}}| \leq 2(8T - 1) \epsilon \). Thus \( \int (\sqrt{g} - \sqrt{\bar{g}})^2 d\mu \leq 4(8T - 1)^2 \epsilon^2 \). Therefore, by triangle inequality,

\[
\int \left( \sqrt{f} - \sqrt{\bar{g}} \right)^2 d\mu \leq 2 \int \left( \sqrt{f} - \sqrt{\bar{g}} \right)^2 d\mu + 4 \int \left( \sqrt{\bar{g}} - \sqrt{\bar{f}} \right)^2 d\mu + 4 \int \left( \sqrt{\bar{f}} - \sqrt{\bar{g}} \right)^2 d\mu
\]

\[
\leq 2 e^2 + 16e^2 + 16(8T - 1)^2 \epsilon^2.
\]
That is $d_H^2(f, \tilde{g}) \leq 2 \left(9 + 8(8T - 1)^2\right) \epsilon^2$. Because $f/\tilde{g} \leq T/(4\epsilon^2 / (\int \tilde{g}d\mu + 4\epsilon^2)) \leq 9T/(4\epsilon^2)$, by Lemma 10 in later this section,
\[
D(f \parallel \tilde{g}) \leq 2 \left(2 + \log \left(9T / (4\epsilon^2)\right)\right) \left(9 + 8(8T - 1)^2\right) \epsilon^2,
\]
which completes the proof.

**Lemma 8:** Assume Conditions 1 and 2 are satisfied. Let $\epsilon_n$ satisfies $\epsilon_n^2 = V(\epsilon_n)/n$ and $\xi_{n,d}$ be chosen such that $M(\xi_{n,d}) = 4n\epsilon_n^2 + 2\log 2$. Then
\[
\xi_{n,d} \sim \epsilon_n.
\]

**Proof:** Let $\sigma = \lim_{n \to 0} M(\sigma) / M(\epsilon) > 1$. Under the assumption $M(\epsilon) > 2\log 2$ when $\epsilon$ is small, we have $4n\epsilon_n^2 + 2\log 2 \leq 6n\epsilon_n^2$ when $n$ is large enough. Then under Condition 1, $M((a/b)\epsilon_n) \geq (1/e) V(\epsilon_n) \geq (6e)^{-1} M(\xi_{n,d})$. Take $k$ large enough such that $\sigma^k \geq 6e$. Then $M((a/b)\epsilon_n)^{\alpha} \geq \sigma^k M((a/b)\epsilon_n) \geq M(\xi_{n,d})$. Thus $(a/b)\alpha^{k}\epsilon_n \leq \xi_{n,d}$, i.e., $\epsilon_n = O(\xi_{n,d})$. Similarly, $M(\epsilon_n/b) \leq V(\epsilon_n) = n\epsilon_n^2 \leq (1/4)M(\xi_{n,d}) \leq M(\xi_{n,d})$. So $\epsilon_n/b \geq \xi_{n,d}$, i.e., $\xi_{n,d} = O(\epsilon_n)$. This completes the proof.

The following lemma is standard from convex analysis.

**Lemma 9:** Let $C$ be a convex class of densities. For a fixed density $f_0$, suppose there exists a density $f^* \in C$ minimizing $\| f - f_0 \|_2$ over $f \in C$. Then $\| f - f^* \|_2 \leq \| f - f_0 \|_2$ for all $f \in C$.

**Proof:** Fix a density $f_1 \in C$. Consider $f_\beta = \beta f_1 + (1 - \beta)f^*$. Then expanding the squares,
\[
\| f_0 - f_\beta \|^2 - \| f_0 - f^* \|^2 = \beta^2 \left( \| f_0 - f_1 \|^2 - \| f_0 - f^* \|^2 \right) + 2\beta(1 - \beta) \left( \langle f_0 - f_1, f_0 - f^* \rangle - \| f_0 - f^* \|^2 \right).
\]
If the coefficient $\langle f_0 - f_1, f_0 - f^* \rangle - \| f_0 - f^* \|^2$ is negative, then the second term in the right hand side of the above equality dominates the sum as $\beta \to 0^+$, which makes $\| f_0 - f_\beta \|^2 - \| f_0 - f^* \|^2$ negative eventually. Thus we have to have $\langle f_0 - f_1, f_0 - f^* \rangle \geq \| f_0 - f^* \|^2$.

Now by expanding squares and after some simplification,
\[
\| f_0 - f_\beta \|^2 - \| f_\beta - f^* \|^2 = \| f_0 - f^* \|^2 + 2\beta \langle f_0 - f_1, f_0 - f^* \rangle - \| f_0 - f^* \|^2 \geq 0.
\]
Because $f_1$ is arbitrary in $C$, the conclusion follows.
Lemma 10: Let $p$ and $q$ be two probability density functions with respect to some $\sigma$-finite measure $\mu$. If $p(x)/q(x) \leq V$ for all $x$, then

$$\phi_1(V) \int p \left( \log \frac{p}{q} \right)^2 d\mu \leq D(p \parallel q) \leq \phi_2(V) d_H^2(p, q).$$

where $\phi_1(V) = (\log V + 1/V - 1)/\log^2 V \geq 1/(2 + \log V)$ and $\phi_2(V) = (V \log V + 1 - V)/(\sqrt{V} - 1)^2 \leq (2 + \log V)$.


The following 4 lemmas are used in Section 2.4. We follow the notations used there.

Lemma 11: Assume $V_K(c; \tilde{G}) \leq A_1 M_2(A_2; \mathcal{H})$ for some positive constants $A_1$ and $A_2$. Then there exist positive constants $A'_1$ and $A'_2$ such that

$$V_K(c; \mathcal{F}) \leq A'_1 M_2(A'_2 c; \mathcal{H}).$$

Proof: We first relate distance $d_K$ on $\mathcal{F}$ to distances on $\tilde{G}$ and $\tilde{H}$.

Let $f_1 = \tilde{h}_1 \tilde{g}_1 / \int \tilde{h}_1 \tilde{g}_1 d\mu$ and $f_2 = \tilde{h}_2 \tilde{g}_2 / \int \tilde{h}_2 \tilde{g}_2 d\mu$ with $\tilde{h}_1, \tilde{h}_2 \in \tilde{H}$ and $\tilde{g}_1, \tilde{g}_2 \in \tilde{G}$. Let $a_1 = \int \tilde{h}_1 \tilde{g}_1 d\mu$ and $a_2 = \int \tilde{h}_2 \tilde{g}_2 d\mu$. Then by adding and subtraction,

$$D(f_1 \parallel f_2) = \int f_1 \left( \log \frac{f_1}{f_2} \right) d\mu$$

$$= \int f_1 \left( \frac{\tilde{h}_1}{\tilde{h}_2} + \frac{\tilde{h}_2}{\tilde{h}_1} - 1 \right) d\mu + \int f_1 \left( \frac{\tilde{g}_1}{\tilde{g}_2} + \frac{\tilde{g}_2}{\tilde{g}_1} - 1 \right) d\mu$$

$$+ \left( \log \frac{a_2}{a_1} - \int \frac{\tilde{g}_1 \tilde{h}_2 d\mu}{a_1} - \int \frac{\tilde{g}_2 \tilde{h}_1 d\mu}{a_1} + 2 \right).$$

Now we bound each of the three terms in the above sum. Using the boundness assumptions, one can easily show $\tilde{g}_1/a_1 \leq C/\sqrt{\mathcal{F}}$ and $\tilde{h}_1/a_1 \leq C/\sqrt{\mathcal{F}}$. It follows that

$$\int f_1 \left( \log \frac{\tilde{h}_1}{\tilde{h}_2} + \frac{\tilde{h}_2}{\tilde{h}_1} - 1 \right) d\mu \leq \frac{C^2}{\mathcal{F}} \int \tilde{h}_1 \left( \log \frac{\tilde{h}_1}{\tilde{h}_2} + \frac{\tilde{h}_2}{\tilde{h}_1} - 1 \right) d\mu \leq \frac{C^2}{\mathcal{F}^2} D(\tilde{h}_1 \parallel \tilde{h}_2),$$

and similarly, $\int f_1 \left( \log \frac{\tilde{g}_1}{\tilde{g}_2} + \frac{\tilde{g}_2}{\tilde{g}_1} - 1 \right) d\mu \leq \frac{C^2}{\mathcal{F}^2} D(\tilde{g}_1 \parallel \tilde{g}_2)$. For the last term in the summation expression of $D(f_1 \parallel f_2)$,

$$\log \frac{a_2}{a_1} - \int \frac{\tilde{g}_1 \tilde{h}_2 d\mu}{a_1} - \int \frac{\tilde{g}_2 \tilde{h}_1 d\mu}{a_1} + 2.$$
\[
\begin{align*}
&= \left( \log \frac{a_2}{a_1} - \frac{a_2}{a_1} + 1 \right) + \frac{1}{a_1} \int \left( |\tilde{g}_1 - \tilde{g}_2| + |\tilde{h}_1 - \tilde{h}_2| \right) d\mu \\
&\leq \frac{1}{a_1} \int |\tilde{g}_1 - \tilde{g}_2| \ ||\tilde{h}_1 - \tilde{h}_2| \ d\mu \\
&\leq \frac{1}{a_1} \left( \int (\tilde{g}_1 - \tilde{g}_2)^2 d\mu \right)^{1/2} \left( \int (\tilde{h}_1 - \tilde{h}_2)^2 d\mu \right)^{1/2} \\
&\leq \frac{C^2}{\epsilon^2} \left( \int (\tilde{g}_1 - \tilde{g}_2)^2 d\mu \right)^{1/2} \left( \int (\tilde{h}_1 - \tilde{h}_2)^2 d\mu \right)^{1/2} \\
&\leq \frac{4C^3}{\epsilon^3} \left( D(\tilde{g}_1 \ || \ \tilde{g}_2) D(\tilde{h}_1 \ || \ \tilde{h}_2) \right)^{1/2},
\end{align*}
\]

where the first inequality follows from that \( \log x - x + 1 \leq 0 \) for \( x > 0 \) and the last inequality follows from

\[
\int (\tilde{g}_1 - \tilde{g}_2)^2 d\mu = \int (\sqrt{\tilde{g}_1} - \sqrt{\tilde{g}_2})^2 (\sqrt{\tilde{g}_1} + \sqrt{\tilde{g}_2}) d\mu \\
\leq \left( \sqrt{\int |\tilde{g}_1|^2 d\mu} + \sqrt{\int |\tilde{g}_2|^2 d\mu} \right)^2 \int (\sqrt{\tilde{g}_1} - \sqrt{\tilde{g}_2})^2 d\mu \\
\leq \frac{4C}{\epsilon} \int (\sqrt{\tilde{g}_1} - \sqrt{\tilde{g}_2})^2 d\mu \\
\leq \frac{4C}{\epsilon} D(\tilde{g}_1 \ || \ \tilde{g}_2)
\]

and a similar inequality \( \int (\tilde{h}_1 - \tilde{h}_2)^2 d\mu \leq 4C/\epsilon D(\tilde{h}_1 \ || \ \tilde{h}_2) \). From all above, we have

\[
D(f_1 \ || \ f_2) \leq \frac{2C^2(\epsilon + 2C)}{\epsilon^3} \cdot \max \left( D(\tilde{g}_1 \ || \ \tilde{g}_2), D(\tilde{h}_1 \ || \ \tilde{h}_2) \right).
\]

Thus if \( \tilde{H}_\epsilon \) and \( \tilde{G}_\epsilon \) are \( \epsilon \)-nets in \( \tilde{H} \) and \( \tilde{G} \) under \( d_K \) distance respectively, then \( \tilde{F} = \{ \tilde{h}g / \int \tilde{h}g d\mu : h \in \tilde{H}_\epsilon, g \in \tilde{G}_\epsilon \} \) provides a \( (\sigma\epsilon) \)-net for \( F \) under \( d_K \) distance, where \( \sigma = \left( \frac{2C^2(\epsilon + 2C)}{\epsilon^3} \right)^{1/2} \).

It follows that

\[
V_K(\epsilon ; \tilde{F}) \leq V_K(\epsilon / \sigma; \tilde{H}) + V_K(\epsilon / \sigma; \tilde{G}).
\]

Because \( \tilde{H} \) is uniformly upper bounded and lower bounded away from 0, the distance \( d_K \) and \( L_2 \) distance are equivalent, thus \( V_K(\epsilon ; \tilde{H}) \leq M_2(A\epsilon; \tilde{H}) \) for some constant \( A > 0 \). Together with Lemma 14, we have \( V_K(\epsilon ; \tilde{H}) \leq M_2(A'\epsilon; \tilde{H}) \) for some constant \( A' \). The conclusion of Lemma 11 follows.

**Lemma 12:** For a nonnegative function class \( G \subset L^2(\mu) \), the covering entropy of \( \tilde{G} = \{ g / \int g d\mu : g \in \tilde{G} \} \) under \( d_K \) can be upper bounded in terms of the packing entropy of \( \tilde{G} \) under \( L_2 \) distance as follows for \( \epsilon \leq 1 \):

\[
V_K(\sqrt{\epsilon}; \tilde{G}) \leq M_2(A\epsilon; \tilde{G}),
\]

where \( A \) is a positive constant.
**Proof:** Suppose for two densities $g_1$ and $g_2$, $\int (g_1 - g_2)^2 \, d\mu \leq \epsilon^2$. We modify density $g_2$ to $\bar{g}_2$ such that $\int (g_1 - \bar{g}_2)^2 \, d\mu$ is well controlled. Let $B = \{ x : g_2(x) \leq \delta \}$ and $\bar{g}_2(x) = (\delta 1_B + g_2(x) 1_{B^c}) / (\mu(B) \delta + \int_{B^c} g_2(x) \, d\mu)$. Then because $\bar{g}_2(x) \geq \delta / (1 + \delta)$, $D(g_1 \parallel \bar{g}_2) \leq \int (g_1 - \bar{g}_2)^2 \, d\mu \leq (1 + \delta) / \delta \int (g_1 - \bar{g}_2)^2 \, d\mu$. Using triangle inequality, $\int (g_1 - \bar{g}_2)^2 \, d\mu \leq 2 \int (g_1 - g_2)^2 \, d\mu + 2 \int (g_2 - \bar{g}_2)^2 \, d\mu$. Now we bound $\int (g_2 - \bar{g}_2)^2 \, d\mu$ as follows

$$
\int (g_2 - \bar{g}_2)^2 \, d\mu = \int_B (g_2 - \bar{g}_2)^2 \, d\mu + \int_{B^c} (g_2 - \bar{g}_2)^2 \, d\mu
\leq \int_{\{g_2(x) \leq \delta\}} \delta^2 \, d\mu + \int_{\{g_2(x) \geq \delta\}} g_2^2(x)(1 - \frac{1}{\mu(B) \delta + \int_{B^c} g_2(x) \, d\mu})^2 \, d\mu
\leq \delta^2 + \frac{\int_{\{g_2(x) \leq \delta\}} (\delta - g_2(x)) \, d\mu)^2}{(\mu(B) \delta + \int_{B^c} g_2(x) \, d\mu)^2} \cdot \int g_2^2(x) \, d\mu
\leq \delta^2 (1 + \int g_2^2(x) \, d\mu).
$$

From all above, we have $D(g_1 \parallel \bar{g}_2) \leq 2 (1 + \delta) / \delta \left( \int (g_1 - g_2)^2 \, d\mu + (1 + \sup_{g \in G} \int g^2(x) \, d\mu) \delta^2 \right)$. Taking $\delta = \epsilon \leq 1$, we have $D(g_1 \parallel \bar{g}_2) \leq 4(2 + \sup_{g \in \bar{G}} \int g^2(x) \, d\mu) \cdot \epsilon$. Thus by modifying an $\epsilon$-net of $\bar{G}$ under $L_2$ distance as described above, we have a $2 \sqrt{(2 + \sup_{g \in \bar{G}} \int g^2(x) \, d\mu)} \epsilon$-net of $\bar{G}$ under $d_K$. This completes the proof.

**Lemma 13:** Let $G \subseteq L^2(\mu)$ be a class of nonnegative functions with $\int g^2 \, d\mu \leq C^2$ and $\int gd\mu \geq \epsilon > 0$ for all $g \in G$. Then the packing entropies of $G$ and $\bar{G}$ under the $L_2$ distance have the following relationship:

$$M_2(\epsilon; G) \leq M_2(\epsilon^* \epsilon; \bar{G}),$$

where $\epsilon^* = \epsilon^2 / 2 \epsilon$.

**Proof:** Consider $g_1, g_2 \in G$. Let $a_1 = \int g_1 \, d\mu$, $a_2 = \int g_2 \, d\mu$ and let $\bar{g}_1 = g_1 / a_1$, $\bar{g}_2 = g_2 / a_2$ be corresponding densities. Then

$$
\int (\bar{g}_1 - \bar{g}_2)^2 \, d\mu = \frac{1}{a_1^2 a_2^2} \int (a_2 (g_1 - g_2) + (a_2 - a_1) g_2)^2 \, d\mu
\leq \frac{2}{a_1^2 a_2^2} \left( a_2^2 \int (g_1 - g_2)^2 \, d\mu + (a_2 - a_1)^2 \int g_2^2 \, d\mu \right).
$$

Using that $(a_2 - a_1)^2 = (\int (g_1 - g_2) \, d\mu)^2 \leq \int (g_1 - g_2)^2 \, d\mu$, $\int g_2^2 \, d\mu \leq C^2$ and $\int gd\mu \geq \epsilon$, we have

$$
\int (\bar{g}_1 - \bar{g}_2)^2 \, d\mu \leq \frac{4C^2}{\epsilon^4} \int (g_1 - g_2)^2 \, d\mu.
$$
Thus, a subset of densities in $\tilde{G}$ separated by $(\epsilon^*)^{-1} \epsilon$ under $L_2$ distance corresponds to a set in $G$ separated by $\epsilon$ under $L_2$ distance. The conclusion follows.

**Lemma 14:** Let $G$ be a uniformly bounded class of measurable functions with respect to a probability measure $\mu$. Let $\tilde{G} = \{e^g / \int e^g \mu : g \in G\}$ be a corresponding density class. Let the $L_q$ ($1 \leq q \leq \infty$) packing $\epsilon$-entropies of $G$ and $\tilde{G}$ be $M_q(\epsilon; G)$ and $M_q(\epsilon; \tilde{G})$ respectively. Then there exist positive constants $c_1$, $c_2$ and $c_3$ such that for small $\epsilon$,

$$M_q(\epsilon; \tilde{G}) \leq M_q(\epsilon, 1; G) \leq M_q(\epsilon, 1; G) + \epsilon^3 \log \left( \frac{1}{\epsilon} \right).$$

**Proof:** Let $f_1 = e^{g_1} / \int e^{g_1} \mu$ and $f_2 = e^{g_2} / \int e^{g_2} \mu$ for $g_1, g_2 \in G$. Then it can be easily shown that $\| f_1 - f_2 \|_q \leq (1/c_1) \| g_1 - g_2 \|_q$, where constant $c_1$ depends only on $\sup_{g \in G} \| g \|_\infty$. As a consequence, $M_q(\epsilon, \tilde{G}) \leq M_q(\epsilon, G)$. Let $\{f_i, 1 \leq i \leq m\}$ be an $\epsilon$-packing set under $L_q$ distance in $\tilde{G}$. For any $f = e^g / \int e^g \mu$ with $g \in G$, there exists $1 \leq i \leq m$ such that $\| f - f_i \|_q \leq \epsilon$. Because of the uniform boundedness assumption, $\| \log f - \log f_i \|_q \leq c^3 \epsilon$ for some $c^3$ depending only on $\sup_{g \in G} \| g \|_\infty$. Also because for $g \in G$, $\| \log e^g \mu \|_q \leq \sup_{g \in G} \| g \|_\infty$, there exists $-J \leq j^* \leq J$ such that $\| \log e^g \mu - j^* \epsilon \|_q \leq c^3 \epsilon$ with $J = [\sup_{g \in G} \| g \|_\infty / \epsilon]$. Then $\| g - \log f_i - j^* \epsilon \|_q = \| \log f - \log f_i + \log e^g \mu - j^* \epsilon \|_q \leq \| \log f - \log f_i \|_q + \| \log e^g \mu - j^* \epsilon \|_q \leq (c^3 + 1) \epsilon$. As a consequence, $\{\log f_i + j \epsilon : 1 \leq i \leq m, -J \leq j \leq J\}$ is an $(c^3 + 1) \epsilon$ net for $G$ under $L_q$ distance. Using the relationship between covering and packing entropies (see Section 2), we have $M_q(2(c^3 + 1) \epsilon; G) \leq M_q(\epsilon; \tilde{G}) + \log (2J + 1)$. The conclusion follows.

**References**


54


Yuhong Yang  
Department of Statistics  
Iowa State University  
Ames, IA 50011

Andrew Barron  
Department of Statistics  
Yale University  
P.O. Box 208290  
New Haven, CT 06520