INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Minimax Optimal Density Estimation

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Yuhong Yang

Dissertation Director: Professor Andrew R. Barron

May 1996

UMI Number: 9635410

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

ABSTRACT

Minimax Optimal Density Estimation

Yuhong Yang
Yale University
May 1996

Information-theoretic tools are used to derive minimax risk bounds for density estimation. A metric entropy condition alone determines the minimax rate of convergence in each class of density functions. To achieve the minimax rates simultaneously for multiple function classes, we consider lists of finite-dimensional approximating models and use model selection criteria related to AIC and MDL to select adaptively a good model based on data. The use of many candidate models, as in the case of subset selection, provides more flexibility for adaptation, yet significant selection bias can occur with criteria such as AIC. We incorporate a model complexity term in the model selection criteria to handle this selection bias. It is shown that the risk of the estimated density is bounded by an index of resolvability, which characterizes the best tradeoff among approximation error, estimation error, and model complexity. As an application, we show that the optimal rate of convergence is simultaneously achieved for density in the Sobolev spaces $W_2^s(U)$ without knowing the smooth parameter s and norm parameter U in advance. Applications in neural network models and sparse density estimation are also provided.

To Dr. Yan Xin

Acknowledgements

I am deeply indebted to my advisor, Professor Andrew Barron. I feel extremely lucky to have been a student of his, learning from him and working with him throughout these past five years. I have benefited so much from his deep insight into the fields of statistics and information theory. I appreciate the immeasurable amount of time he has spent talking with me. Professor Barron has been tremendously generous, caring, and patient. I have learned so much from him in every aspect of my life.

My sincere thanks also go to other members of the faculty: Professor David Pollard for his fabulous teaching and encouragement which gave me much inspiration for statistical research; Professor Joseph Chang for his enthusiasm and his wonderful courses; Professor John Hartigan for teaching me applied statistics and his valuable suggestions; and Professor Nicolas Hengartner for his time and effort in reading my dissertation.

I appreciate all the help I had from the department during a difficult time in my life. I will cherish forever the help, support and friendship from Professor Andrew Barron, Barbara Amato-Kuslan, Lucy Kennedy and many others.

Finally, I want to thank my family; my parents, sisters and my wife Zhaohui, for their boundless love and support. My parents have given me everything, asking nothing in return. This work would never have been completed without Zhaohui's love, emotional support, and encouragement.

Contents

Chapter 1

Introduction

We are interested in estimating a density function based on an independent and identically distributed sample. Density estimation provides a good way to understand the distribution that governs the data and is also used in some other statistical procedures such as nonparametric discriminant analysis and clustering analysis.

The simplest density estimation approaches use parametric models and then the estimation problems reduce to parameter estimations. When data become more irregular than the usual parametric models can handle, to provide enough flexibility to capture the complexities of the density curves, nonparametric methods have been proposed and widely used in statistical applications. Roughly speaking, there are two kinds of nonparametric procedures for density estimation in practice. Some are fully nonparametric in the sense that there is no operating finite-dimensional parametric models involved in the statistical procedures, thus no usual parameter estimation is required. For instance, in kernel density estimation, only choices of a kernel function and a bandwidth are involved. Some smoothing splines density estimation procedures also belong to this category. Some other nonparametric procedures do not abandon the parametric structure completely. Rather, they use parametric models to perform estimation, but give the freedom to use more and more complicated parametric models as data suggest the necessity of doing so. The main difference between these procedures and the traditional parametric approaches is that the underlying curve is not assumed to be in any of the finite-dimensional models, and the parametric models are used for approximation of the true function.

Of course, any statistical procedure should be evaluated for us to understand how well

and when the procedure works. Many evaluation criteria have been used for density estimation depending on the nature of the problems, particular interests, or convenience for handling. For instance, one could consider the statistical risk of an estimator at a specific sample point, or consider global performance of the estimator under some loss function. In this dissertation, we focus on density estimation under global risks using Hellinger loss, L_q loss, or Kullback-Leibler loss.

Suppose the true density function is in a certain function class (a target class). For a given estimator, the worst case risk (or supreme risk if the maximum is not achievable) is the maximum risk of the estimator over the target class. Minimax risk is then the smallest (or infimum) possible worst case risk over all density estimators based on the sample. This quantity characterizes how well we can estimate a density function using a sample in a uniform sense. A minimax procedure (which produces the minimax risk) gives the best protection for the worst case risk. Intuitively, a minimax procedure might be too conservative and indeed they are for some cases. But with certain choices of intrinsically invariant loss functions (e.g., K-L loss), the minimax risk may characterize the typical risk for functions in the target class (see, e.g., Barron and Hengartner (1995)).

Among well-studied nonparametric density classes are classical smooth function classes such as Sobolev spaces and Besov classes. For such nonparametric classes, a lot of asymptotic results have been obtained about the minimax risks and for various nonparametric procedures (see Donoho, Johnstone, et al (1995) for references). Roughly speaking, the minimax risk of a class of densities having more derivatives converges to zero faster than that of a class of densities assumed to have fewer derivatives as the sample size increases. The smoothness conditions are often used in the construction of minimax-rate estimators whose worst case risks converge within a constant factor of the minimax risk. For instance, for some smooth classes, a kernel density estimator with bandwidth suitably chosen according to the smoothness condition of the target class converges to the true density at the minimax rate (see e.g., Devroye (1987)). Similar results are also obtained for some sieve estimators with the parametric model sizes suitably chosen again according to smoothness conditions (e.g., see Stone (1990, 1994), Birgé and Massart (1993), Barron and Sheu (1991), and Wong and Shen (1995)).

The above mentioned procedures that are constructed depending on smoothness conditions of the target classes entail a great difficulty in application because it is impossible to

know how smooth the true density function is in advance in practical situations. The procedures with a predetermined bandwidth in kernel estimation or a predetermined parametric model size in sieve estimation is unlikely to always produce a very good estimator. This suggests that in practice, different bandwidths or model sizes should be considered and a suitable one needs to be chosen somehow based on data (instead of subjective assumptions).

The above consideration calls for adaptive estimation procedures in density estimation. In this work, we are interested in minimax-rate adaptiveness over multiple density classes. The true underlying density function is assumed to be in any of a countable collection of classes. For each given class, a good estimator (e.g., a minimax-rate optimal for this class) could be obtained. Without knowing which class contains the true density, can we have a single estimation procedure that works optimally in the minimax-rate sense simultaneously for all the classes being considered? Such an estimator is minimax-rate adaptive because it automatically adjusts according to the nature of the true density based only on data so that it converges at the right minimax rate for all the classes.

Many adaptive procedures have been proposed for different statistical problems. For nonparametric regression, methods have been introduced to adaptively select the bandwidth for kernel estimator (e.g., Härdle, Hall and Marron (1985)), or smoothness parameters for smoothing splines (e.g., Craven and Wahba (1979)), or model size for linear estimators using parametric models (Shibata (1981), Li (1987)), basis selection for wavelet estimators (Donoho, Johnstone, et al (1995)). For density estimation, Efroimovich (1985) considered linear procedures using projection estimators for the trigonometric coefficients and proposed a final estimator which was shown to be adaptive among some ellipsoidal classes with different smoothness conditions. Donoho, Johnstone, et al (1993) considered adaptive wavelet estimators for density estimation. Recently, Birgé, and Massart (1995), and Barron, Birgé, and Massart (1995) have obtained general model selection results using various contrast functions.

In this dissertation, I will address several issues on density estimation: minimax rates of convergence for a given density class; adaptation among a general collection of density classes; and model selection for adaptive density estimation. The following give some description of the problems we will deal with and summarize the main results we have obtained in these directions.

1. Minimax rates of convergence.

Due to Le Cam (1973), Birgé (1983, 1986) and other researchers' work, metric entropy of a target class is believed by many to determine the minimax rates of convergence. Indeed, Birgé (1986) gives good minimax upper bounds based only on local Hellinger metric entropy, but in deriving lower bounds, he takes a rough bound on Shannon's mutual information for the use of Fano's inequality and uses an additional condition other than entropy on the target class in his work. This extra condition is not always easy to check and is not necessary for checking once the metric entropy structure is known. We use some better bounds on Shannon's mutual information using some information theoretic tools to derive minimax lower bounds based only on global metric entropy. Some upper bounds on minimax risk are also provided based on Kullback-Leibler distance. As a result, it is shown that metric entropy is indeed essentially the only quantity needed to determine the minimax rate of convergence for a general density class. The key fact used in the derivation of the minimax results is the connection between density estimation and data compression in an information theory context.

2. Adaptation over a general collection of density classes.

   Under some mild conditions, we construct minimax-rate adaptive estimators using metric entropies of a general collection of nonparametric density classes. The construction of the adaptive estimators involve mixing densities shown to be minimax rate optimal for each class. The result is that one estimator is simultaneously minimax-rate optimal for all the classes.

3. Model selection for adaptive density estimation.

   A practical way to get adaptive estimators is through the use of model selection criteria to come up with models that produce good estimators for the given sample size. Consider approximating the true density function by some finite-dimensional parametric models. Given the approximating models, we use a model selection criterion related to AIC and MDL to compare the models and show that risk of the density estimator based on the selected model is upper bounded by an index of resolvability which characterizes the best trade-off between the approximation error (bias) and estimation error among all the models being considered. Thus upper bounds on the worst case risk of the estimator based on model selection for a density class could be obtained

by examining the index of resolvability for this class. With the approximating models suitably chosen for the target classes, this approach can produce adaptive estimators and the adaptation property can be shown by evaluating the index of resolvability. As an example, we show that the minimax rates of convergence are simultaneously achieved by density estimator based on model selection for Sobolev spaces without knowing the smoothness parameter and norm parameter in advance.

When exponentially many models are considered (as in subset selection problems), significant selection bias may occur with empirical based model selection criteria. To handle the selection bias, we incorporate a model complexity penalty term in model selection criteria and show that the risk of the density estimator based on these criteria is upper bounded by the best trade-off among approximation error, estimation error and model complexity. Applications in subset selection for density estimation and in some neural network models will be provided.

The dissertation is accordingly divided into 4 chapters.

Chapter 2

Minimax rates of convergence

2.1 Background

Let $X_1, X_2, ..., X_n$ be an independent, identically distributed sample from some distribution on a measurable space $\mathcal{X}$. We assume the probability distribution is dominated by a σ-finite measure μ on $\mathcal{X}$ with a density function which is assumed to be in a density class $\{p_\theta : \theta \in \Theta\}$ with respect to μ. The parameter space Θ could be a finite-dimensional space or a nonparametric space (e.g., the class of all densities, or square root densities). We want to estimate the true density p_θ or θ based on the sample.

Density estimation is important because it may extract useful information about the distribution that governs the data. For instance, with a good density estimator, we can have some visual understanding on whether the distribution is skewed, or has multiple modes or not. Some other statistical procedures also require estimation of certain densities. For example, density estimation is required in some nonparametric discriminant analysis methods and clustering analysis methods. For more details about the applications of density estimation, see Silverman (1986).

Nonparametric density estimation has been studied intensively in the past few decades. Rosenblatt (1956) proposed a moving window estimator and it was later generalized to kernel estimators (Parzen (1962), Cacoullos (1966)). Minimum distance estimators for density estimation were studied by Le Cam (1966), Pfanzagl (1968), Beran (1977), Pollard (1980), Millar (1981, 1983), and Yatracos (1985). Sieve density estimators using suitably chosen approximating models are studied by Cencov (1982), Portnoy (1988), Stone (1990, 1994), Barron and Sheu (1991), Birgé and Massart (1993, 1995), Shen and Wong (1994), Wong

and Shen (1995), and others. Adaptive density estimation has been studied by Efroimovich (1985), Donoho, Johnstone, et al (1993), Birgé and Massart (1995), Barron, Birgé and Massart (1995). More discussions and some results about adaptive density estimation will be given in Chapter 3.

In this chapter, we study some essential questions about density estimation. Several global loss functions will be considered for the evaluation of density estimators. They include Kullback-Leibler (K-L) (also called relative entropy) loss, the Hellinger loss and L_2. We are interested in minimax risks of density (or other parameters) estimators. For a given density class, we study how fast the minimax risk goes to zero and what essential property of the target class determines the minimax rate of convergence.

In statistical decision theory, minimax methods are of interest. Minimax risk characterizes how well we can estimate a parameter or the whole density in a target class in a uniform sense. Thus the minimax risk provides us with the insight into the limitation we can not overcome for the worst case. Furthermore, in many situations, the minimax rate not only captures the worst case but also the typical rate of convergence. Indeed, an intrinsic homogeneity of some of the loss functions we consider here leads to existence of (minimax rate optimal) estimators that have nearly equivalent risk throughout the parameter space.

We determine minimax risk bounds for subclasses $\{p_\theta : \theta \in S\}$, $S \subset \Theta$, which may be parametric or nonparametric (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 $S \subset \overline{S} \subset \Theta$. An estimator of θ is then a measurable mapping from the sample space of $X_1, X_2, ..., X_n$ 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 a general loss function d, which is a mapping from $\overline{S} \times \overline{S}$ to R^+ with $d(\theta, \theta') > 0$ for $\theta \neq \theta'$. We call d a distance whether or not it satisfies properties of a metric.

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

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

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

Related problems are point estimations such as estimating density or regression value

at some point $x_0 \in \mathcal{X}$ or some functional of the density function. We here study a global measure of loss. For characteristics of minimax rates of point estimation, see Donoho and Liu (1991), Bickel and Ritov (1988), and Birgé and Massart (1992).

The minimax rates of convergence are often determined by deriving a minimax upper bound using a specific estimator and obtaining a minimax lower bound in some way. If the maximum risk of the estimator is within a constant factor of the derived lower bound, then the minimax rate of convergence is obtained. 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. The first one is used in Hasminskii (1978), Ibragimov and Hasminskii (1980, 1982), Efroimovich and Pinsker (1982), Nemirovskii (1986), and Hasminskii and Ibragimov (1990). The second one is utilized in Bretagnolle and Huber (1979), Birgé (1986), and Devroye (1987). Birgé (1986) claims that Fano's inequality is more general and could replace Assouad's Lemma in almost all practical situations. Yu (1995) gives a lower bound similar to Assouad's in terms of Kullback-Leibler distance using Fano's inequality.

Both Assouad's lemma and Fano's inequality as it has previously been used involve first restriction to a local subset of the function space with 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. In such cases, it is not necessary to uncover local packing properties.

We prove the following result characterizing minimax convergence rate in terms of metric entropy. Let $d(f,g)$ be a distance and let $N(\epsilon;\mathcal{F})$ be the size of the largest packing set of density functions in the class separated by ϵ and let ϵ_n satisfy $\epsilon_n^2 = \frac{M(\epsilon_n;\mathcal{F})}{n}$, where $M(\epsilon;\mathcal{F}) = \log N(\epsilon;\mathcal{F})$ is the metric entropy and n is the sample size. Assume the target class is rich enough to satisfy $\underline{\lim}_{\epsilon\to 0}\frac{M(\frac{\epsilon}{2};\mathcal{F})}{M(\epsilon;\mathcal{F})} > 1$ (which is true if $M(\epsilon;\mathcal{F}) = \left(\frac{1}{\epsilon}\right)^r \kappa(\epsilon)$ with $r > 0$ and $\frac{\kappa(\frac{\epsilon}{2})}{\kappa(\epsilon)} \to 1$ as $\epsilon \to 0$). This condition is satisfied by the usual smooth nonparametric classes.

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

Proposition: In the following cases, the minimax convergence rate is characterized by metric entropy in terms of the critical radius ϵ_n as follows:

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

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

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

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

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 Kullback-Leibler 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 a finite set of parameter values $\{\theta_1, ..., \theta_m\}$ separated from each other by an amount ϵ_n in the distance of interest. The critical separation ϵ_n is the largest separation such that the hypothesis $\{\theta_1, ..., \theta_m\}$ are nearly indistinguishable on the average by tests. Fano's inequality reveal this indistinguishability in terms of the Kullback-Leibler divergence between densities $p_{\theta_j}(x_1, ..., x_n) = \Pi_{i=1}^n p_{\theta_j}(x_i)$ and the centroid of such densities $q(x_1, ..., x_n) = \frac{1}{m} \sum_{j=1}^m p_{\theta_j}(x_1, ..., x_n)$. Here the key question is to determine the separation such that the average of this K-L divergence is small compared to the distance $\log m$ that would correspond to maximally distinguishable densities (for which θ is determined by X^n). It is critical here that K-L divergence does not have a triangle inequality between the joint densities. We show that the K-L divergence from every $p_{\theta_j}(x_1, ..., x_n)$ to the centroid is 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 β 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. The metric entropy (logarithm of the largest cardinality of an ϵ_n-packing set) determines when this can be done.

Previous uses of Fano's inequality used the coarse bound $n\beta$ (or a similar rough bound) on the K-L diameter of the set $\{p_{\theta_1}^n, ..., p_{\theta_m}^n\}$. In that theory, to obtain a suitable bound, a statistician needs to find a suitable subset $\{\theta_1, ..., \theta_m\}$ with diameter β of the order of ϵ_n and m of the order of the metric entropy. 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 to determine the minimax rates of convergence. In fact, besides the success of such methods in deriving minimax rates as demonstrated in Birgé (1983, 1986), 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 (1992), and Pollard (1993)). Our point here is that for function classes with metric entropy of known order, there is no need to identify a special subset to get the right order lower bound.

The density estimation problem we consider is closely related to a data compression problem in information theory (see section 2.3). The relationship allows us to obtain both upper and lower bounds on the minimax risk from upper bounding a maximum redundancy which can be easily related to the global metric entropy. Combining the new lower bounds with upper bounds, the minimax rates of convergence are determined from the metric entropy properties alone.

In previous analyses, the techniques used for upper bounds seem to be unrelated to those for lower bounds. One of our findings is that given a certain metric entropy of a density class, an upper bound on the minimax K-L risk immediately results in a lower bound on the minimax risk.

This chapter is divided into 5 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, we demonstrate the determination of minimax rates of convergence for several classes of densities.

2.2 Main results

We first give definitions of "metric" entropies.

Definition 2.1: A finite set $N_\epsilon \subset S$ is said to be an ϵ-packing set ($\epsilon > 0$) in S if for any $\theta, \theta' \in N_\epsilon$, $\theta \neq \theta'$, we have $d(\theta, \theta') > \epsilon$, and for any $\tilde{\theta} \in S$, there exists a $\theta_0 \in N_\epsilon$ such that $d(\tilde{\theta}, \theta_0) \leq \epsilon$. We call ϵ the packing radius.

Definition 2.2: A set $G_\epsilon \subset \overline{S}$ is said to be an ϵ-net for S if for any $\tilde{\theta} \in S$, there exists a $\theta_0 \in G_\epsilon$ such that $d(\tilde{\theta}, \theta_0) \leq \epsilon$.

Definition 2.3: Let $M_d(\epsilon)$ be the logarithm of the maximum cardinality of any ϵ-packing set in S. We call $M_d(\epsilon)$ the packing ϵ-entropy of S.

Definition 2.4: Let $V_d(\epsilon)$ be the logarithm of the minimum cardinality of any ϵ-net for set S. We call $V_d(\epsilon)$ the covering ϵ-entropy of S.

From the definitions, it is clear that $M_d(\epsilon)$ and $V_d(\epsilon)$ are nonincreasing in ϵ. 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 any distance d.

The above definitions are slight generalizations of the metric entropy notions introduced by Kolmogorov and Tihomirov (1959). We do not require the distance d to be a metric. In fact, one choice of d will be the square root of the relative entropy or Kullback-Leibler (K-L) distance. Let $d_K^2(\theta, \theta') = D(p_\theta \parallel p_{\theta'}) = \int p_\theta \log(p_\theta / p_{\theta'})\, d\mu$. Clearly $d_K(\theta, \theta')$ is asymmetric in its two arguments, so it can not be a metric. The major 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 any θ, θ' and θ'' in S. Such an inequality would usually be used to rule out the possibility that one density estimator is too close in d to too many densities which are far away from each other in the same distance d. This might happen with K-L distance. It seems necessary to have some additional conditions enabling a triangle-like inequality to obtain a reasonable lower bound in terms of d distance. The following example demonstrates this point.

Example 2.1: Consider densities on [0,1] with respect to the Lebesgue measure. Let $S = \{\theta : 0 \leq \theta \leq \frac{1}{2}\}$ and $p_\theta(x) = 2I_{\{\theta \leq x \leq \theta + \frac{1}{2}\}}$. Then for any $\epsilon > 0$, the ϵ-packing set under d_K must be S itself. Let $\hat{p} = I_{\{0 \leq x \leq 1\}}$. Then $D(p_\theta \parallel \hat{p}) = \log 2$ 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.

Another distance we consider is Hellinger distance $d_H(\theta, \theta') = \sqrt{\int \left(\sqrt{p_\theta} - \sqrt{p_{\theta'}}\right)^2 d\mu}$. Hellinger distance is a metric. We will also consider L_q distance $d_q(\theta, \theta') = \left(\int |p_\theta - p_{\theta'}|^q d\mu\right)^{\frac{1}{q}}$ for $q \geq 1$.

We assume the distance d satisfies the following condition, which is used in the derivation of minimax lower bounds.

Assumption 2.0: There exists a positive constant $A \leq 1$ such that for any $\theta, \theta' \in S$, $\widetilde{\theta} \in \overline{S}$,

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

Remarks:

1. Here we have assumed this triangle-like inequality to hold for all indicated parameter values. The assumption can be relaxed to require that it holds only for $\widetilde{\theta}$ close to θ and θ', specifically, that there exist positive constants $A \leq 1$ and $\epsilon_0 > 0$ such that for any $\theta, \theta' \in S$, $\widetilde{\theta} \in \overline{S}$, if $\max\left(d(\theta, \widetilde{\theta}), d(\theta', \widetilde{\theta})\right) \leq \epsilon_0$, then $d(\theta, \widetilde{\theta}) + d(\theta', \widetilde{\theta}) \geq A d(\theta, \theta')$. If one uses local entropy conditions, then a local version of this triangle-like inequality can be used to get similar results.

2. For general distance d, satisfaction of the above inequality may depend on the choice of $\overline{S}$. However, if d is a metric on Θ, then Assumption 2.0 is always satisfied with $A = 1$ for any $\overline{S} \subset \Theta$.

When Assumption 2.0 is satisfied, the packing entropy and covering entropy have the following relationship.

Lemma 2.0: Suppose Assumption 2.0 is satisfied for distance d. Then

$$M_d(\frac{2}{A}\epsilon) \leq V_d(\epsilon) \leq M_d(\epsilon).$$

The proof of the lemma is similar to that given for d being a metric by Kolmogorov and Tihomirov (1959).

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 $\epsilon > 0$ and $M_d(\epsilon) \to \infty$ as $\epsilon \to 0$ (the latter requirement is used to avoid the triviality of S being a finite set). 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 and 3, more results are given for L_2 risk and K-L risk respectively. In subsection 4, we present interesting results connecting minimax upper bounds with minimax lower bounds.

2.2.1 Minimax under global entropy condition

Suppose a good upper bound on the covering ϵ-entropy under the square root K-L distance is available. That is, assume $V_K(\epsilon) \leq V(\epsilon)$ with V being a nonincreasing and right continuous function. Ideally $V_K(\epsilon)$ and $V(\epsilon)$ are of the same order. 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 ϵ_n satisfies

$$\epsilon_n^2 = \frac{V(\epsilon_n)}{n}.$$

The squared radius is the same as the covering entropy divided by the sample size. The trade-off here between $\frac{V(\epsilon)}{n}$ and ϵ^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 $\underline{\epsilon}_{n,d}$ be a radius ϵ such that

$$M_d(\underline{\epsilon}_{n,d}) = 4n\epsilon_n^2 + 2\log 2.$$

The existence of $\underline{\epsilon}_{n,d}$ follows from the right continuity of $M_d(\epsilon)$ and the assumption $M_d(\epsilon) \to \infty$ as $\epsilon \to 0$. Roughly, $\underline{\epsilon}_{n,d}$ is the packing radius 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 $\underline{\epsilon}_{n,d}$ the packing radius commensurate with the critical covering radius ϵ_n. The speed at which $\underline{\epsilon}_{n,d}^2$ converges to 0 determines an lower bound on the minimax risk.

Minimax Lower bound

Theorem 2.1: Suppose Assumption 2.0 is satisfied for the distance d. Then the minimax risk for estimating $\theta \in S$ satisfies

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) \geq \frac{A \underline{\epsilon}_{n,d}^2}{8},$$

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

Proof: Let $N_{\underline{\epsilon}_{n,d}}$ be an $\underline{\epsilon}_{n,d}$-packing set with the maximum cardinality in S under the given distance d and let G_{ϵ_n} be an ϵ_n-net for S under d_K. For any estimator $\hat{\theta} \in \mathcal{A}_n$, define $\widetilde{\theta} = arg \min_{\theta' \in N_{\underline{\epsilon}_{n,d}}} d(\theta', \hat{\theta})$ (if there are more than one minimizer, choose any one), where the minimum is over θ' in the finite packing set $N_{\underline{\epsilon}_{n,d}}$. Then we have

$$d(\theta, \hat{\theta}) \geq d(\widetilde{\theta}, \hat{\theta}) \geq A d(\theta, \widetilde{\theta}) - d(\theta, \hat{\theta}).$$

Thus if $\theta \neq \widetilde{\theta}$, $2d(\theta, \hat{\theta}) \geq A d(\widetilde{\theta}, \theta) \geq A \underline{\epsilon}_{n,d}$, i.e., $d^2(\theta, \hat{\theta}) \geq \frac{A^2}{4} \underline{\epsilon}_{n,d}^2$. Then

$$\begin{aligned}
\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) &\geq \min_{\hat{\theta}} \max_{\theta \in N_{\underline{\epsilon}_{n,d}}} E_\theta d^2(\theta, \hat{\theta}) \\
&\geq \min_{\hat{\theta}} \max_{\theta \in N_{\underline{\epsilon}_{n,d}}} \frac{A^2 \underline{\epsilon}_{n,d}^2}{4} E_\theta 1_{\{\theta \neq \widetilde{\theta}\}} \\
&= \min_{\hat{\theta}} \max_{\theta \in N_{\underline{\epsilon}_{n,d}}} \frac{A^2 \underline{\epsilon}_{n,d}^2}{4} P_\theta \left(\theta \neq \widetilde{\theta}\right) \\
&\geq \frac{A^2 \underline{\epsilon}_{n,d}^2}{4} \min_{\hat{\theta}} \sum_{\theta \in N_{\underline{\epsilon}_{n,d}}} w(\theta) P_\theta \left(\theta \neq \widetilde{\theta}\right) \\
&= \frac{A^2 \underline{\epsilon}_{n,d}^2}{4} \min_{\hat{\theta}} P_w \left(\theta \neq \widetilde{\theta}\right),
\end{aligned}$$

where in the last line, θ is randomly drawn according to a prior probability w restricted to $N_{\underline{\epsilon}_{n,d}}$, and P_w denotes the Bayes average probability with respect to the prior w. By Fano's inequality (see e.g., Cover and Thomas (1991), Chapter 8), with w_0 being the uniform prior on $\Theta_0 = N_{\underline{\epsilon}_{n,d}}$, we have

$$P_{w_0} \left(\theta \neq \widetilde{\theta}\right) \geq 1 - \frac{I(\Theta_0; X^n) + \log 2}{\log |N_{\underline{\epsilon}_{n,d}}|}, \tag{2.1}$$

where $I(\Theta_0; X^n)$ is Shannon's mutual information between the random parameter and the observation X^n. This mutual information is upper bounded by the maximum K-L distance

between the product measure $p(x^n|\theta)$ and any density $q(x^n)$ on the sample space. Indeed,

$$\begin{aligned} I(\Theta_0; X^n) &= \textstyle\sum_\theta w_0(\theta) \int p(x^n|\theta) \log \frac{w_0(\theta)p(x^n|\theta)}{w_0(\theta)p_{w_0}(x^n)} \mu(dx^n) \\ &= \textstyle\sum_\theta w_0(\theta) \int p(x^n|\theta) \log \frac{p(x^n|\theta)}{p_{w_0}(x^n)} \mu(dx^n) \\ &\le \textstyle\sum_\theta w_0(\theta) \int p(x^n|\theta) \log \frac{p(x^n|\theta)}{q(x^n)} \mu(dx^n) \\ &\le \max_{\theta \in N_{\underline{\epsilon}_{n,d}}} D\left(P_{X^n|\theta} \parallel Q_{X^n}\right), \end{aligned}$$

where Q_{X^n} has density $q(x^n)$, and $p_{w_0}(x^n) = \sum_\theta w_0(\theta) p(x^n|\theta)$. The first inequality above follows the fact that the Bayes mixture density $p_{w_0}(x^n)$ minimizes the average relative entropy $\sum_\theta w_0(\theta) \int p(x^n|\theta) \log \frac{p(x^n|\theta)}{q(x^n)} \mu(dx^n)$ over all densities $q(x^n)$ (any other choice yields a larger value by the amount $\int p_{w_0}(x^n) \log \frac{p_{w_0}(x^n)}{q(x^n)} \mu(dx^n) > 0$). Choose w_1 be the uniform prior on G_{ϵ_n} and let $q(x^n) = p_{w_1}(x^n) = \sum_\theta w_1(\theta) p(x^n|\theta)$ and $Q_{X^n} = P_{w_1, X^n}$ be the corresponding Bayes mixture density and distribution respectively. Because G_{ϵ_n} is an ϵ_n-net in S under d_K, for each $\theta \in S$, there exists $\tilde{\theta} \in G_{\epsilon_n}$ such that $D(p_\theta \parallel p_{\tilde{\theta}}) = d_K^2(\theta, \tilde{\theta}) \le \epsilon_n^2$. Also by definition, $\log |G_{\epsilon_n}| \le V_K(\epsilon_n)$. It follows that

$$\begin{aligned} D\left(P_{X^n|\theta} \parallel P_{X^n}\right) &= E \log \frac{p(X^n|\theta)}{\frac{1}{|G_{\epsilon_n}|} \sum_{\theta' \in G_{\epsilon_n}} p(X^n|\theta')} \\ &\le E \log \frac{p(X^n|\theta)}{\frac{1}{|G_{\epsilon_n}|} p(X^n|\tilde{\theta})} \\ &\le \log |G_{\epsilon_n}| + D\left(P_{X^n|\theta} \parallel P_{X^n|\tilde{\theta}}\right) \\ &\le V(\epsilon_n) + n\epsilon_n^2. \end{aligned} \tag{2.2}$$

Thus, by our choice of $\underline{\epsilon}_{n,d}$,

$$\frac{I(\Theta_0; X^n) + \log 2}{\log |N_{\underline{\epsilon}_{n,d}}|} \le \frac{1}{2}.$$

Therefore

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) \ge \frac{A^2 \underline{\epsilon}_{n,d}^2}{8}.$$

Remark: Up to the point (2.1), the development here is standard. Previous use of Fano's inequality for minimax lower bound takes one of the following weak bounds on mutual information $I(\Theta; X^n) \le nI(\Theta; X_1)$ or $I(\Theta; X^n) \le n \max_{\theta, \theta' \in \Theta} D(P_{X_1|\theta} \parallel P_{X_1|\theta'})$. 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 $\max_{\theta \in S} D(P_{X^n|\theta} \parallel P_{X^n}) \le V(\epsilon_n) + n\epsilon_n^2$ represents an upper bound on the minimax redundancy $\min_{Q_{X^n}} \max_{\theta \in S} D(P_{X^n|\theta} \parallel Q_{X^n})$. The data compression interpretations of these quantities originate with Davisson (1973); see Clarke and Barron (1994), Haussler and Opper (1995) for some recent work in that area. The bound $D(P_{X^n|\theta} \parallel P_{X^n}) \le V(\epsilon_n) + n\epsilon_n^2$ has its roots in Barron (1987, pp.

89), where it is given in a more general form for arbitrary priors ($D(P_{X^n|\theta} \parallel P_{w,X^n}) \leq \log \frac{1}{w(\mathcal{N}_{\theta,\epsilon})} + n\epsilon^2$, where $\mathcal{N}_{\theta,\epsilon} = \{\theta' : D(p_\theta \parallel p_{\theta'}) \leq \epsilon^2\}$). The redundancy bound $V(\epsilon_n) + n\epsilon_n^2$ can also be obtained from use of a two stage code of length $\log |G_{\epsilon_n}| + \min_{\theta' \in G_{\epsilon_n}} \log \frac{1}{p(x^n|\theta')}$, see Barron and Cover (1991, Section V).

When K-L distance is lower bounded by a multiple of the chosen distance d on $\overline{S}$, 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 $\theta, \theta' \in \overline{S}$, then under the previous condition,

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d_K^2(\theta, \hat{\theta}) \geq \frac{A_0 A^2 \underline{\epsilon}_{n,d}^2}{8}.$$

A natural choice for d is the Hellinger distance (since Hellinger distance does satisfy the triangle inequality between densities and locally square root K-L distance behaves like Hellinger for bounded log-density ratios). Let $\underline{\epsilon}_{n,K}$ and $\underline{\epsilon}_{n,H}$ be the packing radius commensurate with the critical covering radius ϵ_n under d_K and d_H, determined by $M_K(\underline{\epsilon}_{n,K}) = 4n\epsilon_n^2 + 2\log 2$ and $M_H(\underline{\epsilon}_{n,H}) = 4n\epsilon_n^2 + 2\log 2$, respectively. We have the following two corollaries.

Corollary 2.1: Assume there exists a constant A such that for any θ, $\theta' \in S$ and $\tilde{\theta} \in \overline{S}$, $D(p_\theta \parallel p_{\tilde{\theta}}) + D(p_{\theta'} \parallel p_{\tilde{\theta}}) \geq A D(p_\theta \parallel p_{\theta'})$. Then

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d_K^2(\theta, \hat{\theta}) \geq \frac{A^2 \underline{\epsilon}_{n,K}^2}{8}.$$

Corollary 2.2: For the square Hellinger risk, we have

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d_H^2(\theta, \hat{\theta}) \geq \frac{\underline{\epsilon}_{n,H}^2}{8}.$$

Note in Corollary 2.1, $\underline{\epsilon}_{n,K}^2$ may be determined by packing entropy $M_K(\epsilon)$ only (with the choice $V(\epsilon) = M_K(\epsilon)$). However, for general distance d (specifically the Hellinger distance in Corollary 2), $\underline{\epsilon}_{n,d}^2$ is determined by two quantities: both $M_H(\epsilon)$ and $V_K(\epsilon)$ without any assumption on the relationship between the distances.

When the distance d_K is locally upper bounded by a multiple of distance d on S, the minimax lower bound under d^2 risk can be expressed in terms of packing entropy under d distance.

Corollary 2.3: Assume Assumption 2.0 is satisfied for distance d and there exists a constant $\overline{A}$ such that $D(p_\theta, p_{\theta'}) \leq \overline{A}d^2(\theta, \theta')$ for any $\theta, \theta' \in S$ with $d(\theta, \theta') \leq \frac{\tau_{n,d}}{\sqrt{\overline{A}}}$, where $\tau_{n,d}$ is determined from $M_d(\frac{\tau_{n,d}}{\sqrt{\overline{A}}}) = n\tau_{n,d}^2$. Then,

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) \geq \frac{A^2 \underline{\tau}_{n,d}^2}{8},$$

where $\underline{\tau}_{n,d}$ is chosen such that $M_d(\underline{\tau}_{n,d}) = 4n\tau_{n,d}^2 + 2\log 2$.

Proof: Under the assumption between distances d and d_K, a $\frac{\tau_{n,d}}{\sqrt{\overline{A}}}$-packing set in S under d also serves as a $\tau_{n,d}$-covering set for S under d_K. Thus when $\epsilon \leq \tau_{n,d}$, $V_K(\epsilon) \leq M_d\left(\frac{\epsilon}{\sqrt{\overline{A}}}\right)$. The result follows from Theorem 2.1.

The advantage of this bound is that it is determined by the radius $\underline{\tau}_{n,d}$ using exclusively the chosen distance d.

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 any other distance.

Upper bound

To provide an upper bound on the minimax rate of convergence, we construct an estimator as follows. Consider the ϵ_n-net G_{ϵ_n} for S under d_K and the uniform prior w_1 on G_{ϵ_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

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

be the density estimator constructed as a Cesaro average of the Bayes predictive density estimators $\hat{p}_i(x) = p\left(X_{i+1}|X^i\right)$ evaluated at $X_{i+1} = x$, which equal $\frac{p(X^i,x)}{p(X^i)}$ for $i > 0$ and $\hat{p}_i(x) = p(x) = \frac{1}{|G_{\epsilon_n}|}\sum_{\theta\in G_{\epsilon_n}} p(x|\theta)$ for $i = 0$. Then by convexity and the chain rule (as in Barron (1987)),

$$\begin{aligned} E_\theta D(p_\theta \parallel \overline{p}) &\leq \tfrac{1}{n} E_\theta \left(\textstyle\sum_{i=0}^{n-1} D(P_{X_{i+1}|\theta} \parallel P_{X_{i+1}|X^i})\right) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n-1} E\log\frac{p(X_{i+1}|\theta)}{p(X_{i+1}|x^i)} \\ &= \tfrac{1}{n} E\log\frac{p(X^n|\theta)}{p_{w_1}(X^n)} \\ &= \tfrac{1}{n} D\left(P_{X^n|\theta} \parallel P_{w_1,X^n}\right) \\ &\leq \tfrac{1}{n}\left(V(\epsilon_n) + n\epsilon_n^2\right) = 2\epsilon_n^2, \end{aligned}$$

where the last line is as derived as in equation (2.2). Thus

$$\min_{\hat{p}} \max_{\theta\in S} E_\theta D(p_\theta \parallel \hat{p}) \leq 2\epsilon_n^2,$$

where minimization is over all density estimators.

From the above lower and upper bounds, we have the following theorem on the minimax risk.

Theorem 2.2: Assume Assumption 2.0 is satisfied for distance d and $A_0 d^2(\theta,\theta') \leq d_K^2(\theta,\theta')$ for any $\theta,\theta' \in \overline{S}$. Assume also that $\mathcal{A}_n$ (the set of all allowed estimators) contains the estimator corresponding to $\overline{p}$ constructed above. Then

$$\frac{A_0 A\epsilon_{n,d}^2}{8} \leq A_0 \min_{\hat{\theta}\in\mathcal{A}_n} \max_{\theta\in S} E_\theta d^2(\theta,\hat{\theta}) \leq \min_{\hat{\theta}\in\mathcal{A}_n} \max_{\theta\in S} E_\theta d_K^2(\theta,\hat{\theta}) \leq 2\epsilon_n^2.$$

The condition that $\mathcal{A}_n$ contains $\overline{p}$ in Theorem 2.2 is satisfied if $\{p_\theta : \theta \in \overline{S}\}$ is convex. Specifically, if the action space $\overline{S}$ is the set of all densities on $\mathcal{X}$ and d is a pseudo-metric on densities and if we allow all estimators for competition, then the only remaining condition needed for the above inequalities is $A_0 d^2(\theta,\theta') \leq d_K^2(\theta,\theta')$ for all θ,θ'. This is satisfied by Hellinger distance and L_1 distance (with $A_0 = 1$ and $A_0 = \frac{1}{2}$ respectively).

Remark: In obtaining both the upper bound and lower bound on the minimax risk, $D\left(P_{X^n|\theta} \parallel P_{X^n}\right)$ plays an important role. For the lower bound, the quantity is used for bounding $I\left(\Theta_0; X^n\right)$, and for the upper bound, it bounds the risk of a specific estimator.

It is interesting that $D\left(P_{X^n|\theta} \parallel P_{X^n}\right)$ is upper and lower bounded in the very same way but with different radius choices for the two cases. As we shall see, asymptotically these two radii typically have the same rate.

If $\underline{\epsilon}_{n,d}^2$ and $2\epsilon_n^2$ converge to 0 at the same rate, then the minimax rate of convergence is identified. For $\underline{\epsilon}_{n,d}^2$ and ϵ_n^2 to be of the same order, it is sufficient that the following two conditions hold:

(1). There exist two positive constants a and b such that when ϵ is small enough,

$$M_d(b\epsilon) \leq V_K(\epsilon) \leq M_d(a\epsilon); \tag{2.3}$$

(2).

$$\underline{\lim}_{\epsilon \to 0} \frac{M_d(\frac{\epsilon}{2})}{M_d(\epsilon)} > 1. \tag{2.4}$$

The condition (2.3) is the equivalence of the entropy structure under the square root K-L distance and that under d distance when ϵ is small, which 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 either Hellinger distance or L_2 distance. It is also satisfied by the nonparametric regression example in Section 4. The second condition requires the density class to be large enough, namely, $M_d(\epsilon)$ approaches ∞ at least polynomially fast in $\frac{1}{\epsilon}$ as $\epsilon \to 0$, i.e., there exists a constant $\delta > 0$ such that $M_d(\epsilon) \succeq \left(\frac{1}{\epsilon}\right)^{\delta}$. The second condition is satisfied if $M_d(\epsilon)$ can be expressed as

$$M_d(\epsilon) = \left(\frac{1}{\epsilon}\right)^r \kappa(\epsilon),$$

where $r > 0$ and $\frac{\kappa(\frac{\epsilon}{2})}{\kappa(\epsilon)} \to 1$ as $\epsilon \to 0$.

Corollary 2.4: Assume Assumption 2.0 is satisfied for distance d and $\{p_\theta : \theta \in \overline{S}\}$ is convex. Under conditions (2.3) and (2.4), we have

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

where γ_n is determined by the equation $M_d(\gamma_n) = n\gamma_n^2$.

Corollary 2.4 is applicable for many smooth nonparametric classes. 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 $M_H(\epsilon)$ might be of order $\log\left(\frac{1}{\epsilon}\right)^m$ for some constant $m \geq 1$. Then ϵ_n and $\underline{\epsilon}_{n,H}$ are not of the same order with $\epsilon_n \asymp \frac{\sqrt{\log n}}{\sqrt{n}}$ and $\underline{\epsilon}_{n,H} = o(\frac{1}{\sqrt{n}})$. Thus both the upper bound and lower bound provided by the theorem are near but not the optimal rate. 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.

2.2.2 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 2.3 may not hold. Theorem 2.1 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 μ on a compact set $\mathcal{X}$ such as [0,1]. More generally, we may assume that μ is a finite dominating measure. We normalize μ to be a probability measure. Let the packing entropy of $\mathcal{F}$ be $M_q(\epsilon)$ under the L_q metric.

To derive minimax upper bounds, we need a lemma.

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), a randomized estimator in the original problem is determined for which the risk is not greater than a multiple of the risk in the new class.

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 generated i.i.d. sample from uniform distribution on $\mathcal{X}$ with respect to μ (independent of $X_1, ..., X_n$). Let Z_i be X_i or Y_i with probability $(\frac{1}{2}, \frac{1}{2})$ using $V_i \sim Bernoulli(\frac{1}{2})$ indepen-

dently for $i = 1, ..., n$. Then Z_i has density $g(x) = \frac{1}{2}(f + 1)$. Clearly the new density g is bounded below (away from 0), whereas the family of the original densities need not be. Let $\tilde{\mathcal{F}} = \{g : g = \frac{f+1}{2}, f \in \mathcal{F}\}$ be the new density class.

Lemma 2.1: The minimax L_2 risks of the two classes $\mathcal{F}$ and $\tilde{\mathcal{F}}$ have the following relationship.

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E_{X^n} \parallel f - \hat{f} \parallel_2^2 \;\leq\; 16 \min_{\hat{g}} \max_{g \in \tilde{\mathcal{F}}} E_{Z^n} \parallel g - \hat{g} \parallel_2^2,$$

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

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E_{X^n} \parallel f - \hat{f} \parallel_q^q \;\leq\; 4^q \min_{\hat{g}} \max_{g \in \tilde{\mathcal{F}}} E_{Z^n} \parallel g - \hat{g} \parallel_q^q .$$

Proof: We only prove the assertion for L_2. The proof for general L_q is similar. Let $\hat{g}$ be any density estimator of g based on Z_i, $i = 1, ...n$. Let $\hat{g}$ be the density that minimizes $\parallel h - \hat{\hat{g}} \parallel_2^2$ over $h \in \{k : k(x) \geq \frac{1}{2}, \int k(x)d\mu = 1\}$. Then by triangle inequality and because $g \in \{k : k(x) \geq \frac{1}{2}, \int k(x)d\mu = 1\}$, $\parallel g - \hat{g} \parallel_2^2 \leq 2 \parallel g - \hat{\hat{g}} \parallel_2^2 + 2 \parallel \hat{g} - \hat{\hat{g}} \parallel_2^2 \leq 4 \parallel g - \hat{\hat{g}} \parallel_2^2$. Now we construct a density estimator for f. Note that $f(x) = 2g(x) - 1$, let

$$\hat{f}_{rand}(x) = 2\hat{g}(x) - 1.$$

Then $\hat{f}_{rand}(x)$ is a nonnegative and normalized probability density estimate and depends on $X_1, ..., X_n, Y_1, ..., Y_n$ and outcomes of coin flips $V_1, ..., V_n$. So it is a randomized estimator. The squared L_2 loss of $\hat{f}_{rand}$ is bounded as follows:

$$\begin{aligned} \int \left(f(x) - \hat{f}_{rand}(x)\right)^2 d\mu &= \int \left(2g(x) - 2\hat{g}(x)\right)^2 d\mu \\ &= 4 \int (g - \hat{g})^2 \, d\mu \\ &\leq 16 \parallel g - \hat{\hat{g}} \parallel_2^2 . \end{aligned}$$

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

$$\begin{aligned} E_{X^n} \parallel f - \hat{f} \parallel_2^2 &= E_{X^n} \parallel f - E_{Y^n, V^n} \hat{f}_{rand} \parallel_2^2 \\ &\leq E_{X^n} E_{Y^n, V^n} \parallel f - \hat{f}_{rand} \parallel_2^2 \\ &= E_{Z^n} \parallel f - \hat{f}_{rand} \parallel_2^2 \\ &\leq 16 E_{Z^n} \parallel g - \hat{g} \parallel_2^2, \end{aligned}$$

where the first inequality is by convexity and the second identity is because $\hat{f}_{rand}$ depends on X^n, Y^n, V^n only through Z^n. Thus

$$\max_{f\in\mathcal{F}} E_{X^n} \parallel f - \hat{f} \parallel_2^2 \leq 16 \max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} \parallel g - \hat{g} \parallel_2^2 .$$

Taking the minimum over estimators $\hat{g}$, we proved the lemma.

Now, since $\parallel \frac{f_1+1}{2} - \frac{f_2+1}{2} \parallel_2 = \frac{1}{2} \parallel f_1 - f_2 \parallel_2$, for the new class $\widetilde{\mathcal{F}}$, the ϵ-packing entropy under L_2 is $\widetilde{M}_2(\epsilon) = M_2(2\epsilon)$.

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 $g_1, g_2 \in \widetilde{\mathcal{F}}$,

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

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 $\frac{1}{2}$. Let $\widetilde{V}_K(\epsilon)$ denote the d_K covering entropy of $\widetilde{\mathcal{F}}$. Then $\widetilde{V}_K(\epsilon) \leq \widetilde{M}_2(\frac{\epsilon}{\sqrt{2}}) = M_2(\sqrt{2}\epsilon)$. Let ϵ_n be chosen such that

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

From Theorem 2.2, there exists a density estimator $\hat{g}_0$ such that

$$\max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} D(g \parallel \hat{g}_0) \leq 2\epsilon_n^2.$$

It follows that

$$\max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} d_H^2(g, \hat{g}_0) \leq 2\epsilon_n^2,$$

and

$$\max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} \parallel g - \hat{g}_0 \parallel_1^2 \leq 8\epsilon_n^2.$$

By Lemma 2.1,

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

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 $\widetilde{\mathcal{F}}$ that is closest to $\hat{g}_0$ in Hellinger distance. Then by triangle inequality,

$$\max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} d_H^2(g, \hat{g}) \leq 2 \max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} d_H^2(g, \hat{g}_0) + 2 \max_{g\in\widetilde{\mathcal{F}}} E_{Z^n} d_H^2(\hat{g}, \hat{g}_0)$$

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

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

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

Thus $\max_{g \in \widetilde{\mathcal{F}}} E_{Z^n} \| g - \hat{g} \|_2^2 \leq 16(L+1)\epsilon_n^2$. Using Lemma 2.1 again, we have an upper bound on minimax squared L_2 risk.

Proposition 2.1: Let $M_2(\epsilon)$ be the L_2 metric entropy of a density class $\mathcal{F}$ with respect to a probability measure. Let ϵ_n satisfy $M_2(\sqrt{2}\epsilon_n) = n\epsilon_n^2$. Then

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E_{X^n} \| f - \hat{f} \|_1 \leq 8\sqrt{8}\epsilon_n.$$

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

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E_f \| f - \hat{f} \|_2^2 \leq 256(L+1)\epsilon_n^2.$$

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 risk) using only the L_2 metric entropy. For a related result under local entropy assumptions, see Birgè (1986, Theorem 3.1).

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

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

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

Assumption 2.1: There exists at least one density $f^* \in \mathcal{F}$ with $\min_{x \in \mathcal{X}} f^*(x) = \underline{C} > 0$ and a positive constant $\alpha \in (0,1)$ such that $\mathcal{F}_0 = \{\alpha f^* + (1-\alpha)g : g \in \mathcal{F}\} \subset \mathcal{F}$.

For a convex class of densities, Assumption 2.1 is satisfied if there is at least one density bounded away from zero.

Lemma 2.2: Under Assumption 2.1, the subclass $\mathcal{F}_0$ has L_2 metric entropy $M_2^0(\epsilon) = M_2(\frac{\epsilon}{1-\alpha})$.

Proof: Because $\sqrt{\int ((\alpha f^* + (1-\alpha)g_1) - (\alpha f^* + (1-\alpha)g_2))^2 \, d\mu} = (1-\alpha)\sqrt{\int (g_1 - g_2)^2 \, d\mu}$, an ϵ-packing set in $\mathcal{F}$ corresponds to an $(1-\alpha)\epsilon$-packing set in $\mathcal{F}_0$ and vise versa.

Under Assumption 2.1, for two densities f_1 and f_2 in $\mathcal{F}_0$,

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

Thus applying Theorem 2.1 on $\mathcal{F}_0$, we have the following conclusion.

Proposition 2.2: Let $M_2(\epsilon)$ be the L_2 metric entropy of a density class $\mathcal{F}$ with respect to a probability measure. Let $\bar{\epsilon}_n$ satisfy $M_2(\frac{\sqrt{\alpha \underline{C}}}{1-\alpha}\bar{\epsilon}_n) = n\bar{\epsilon}_n^2$ and $\underline{\epsilon}_n$ be chosen such that $M_2(\frac{1}{1-\alpha}\underline{\epsilon}_n) = 4n\bar{\epsilon}_n^2 + 2\log 2$. Then

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} E_{X^n} \parallel f - \hat{f} \parallel_2^2 \geq \frac{\underline{\epsilon}_n^2}{8}.$$

If Assumption 2.1 is not satisfied, similar lower bound can be obtained under the following condition.

Assumption 2.2: Suppose there exists a subclass $\mathcal{F}_0 \subset \mathcal{F}$ and $f_0 \in \mathcal{F}_0$ such that $\max_{f \in \mathcal{F}_0} \parallel \log \frac{f}{f_0} \parallel_\infty < \infty$ and the L_2 metric entropy $M_2^0(\epsilon)$ of $\mathcal{F}_0$ satisfies $M_2^0(\epsilon) \geq M_2(C\epsilon)$ for some constant C (independent of ϵ).

For some classes, a choice of $\mathcal{F}_0$ in Assumption 2.2 might be $\mathcal{F}^{v_1,v_2} = \{f \in \mathcal{F} : v_1 \leq f(x) \leq v_2\}$ for some constants $v_2 > v_1 > 0$.

Combining the lower bounds with upper bounds, the minimax L_2 rate is determined under some conditions.

Theorem 2.3: Suppose $\sup_{f\in\mathcal{F}} \| f \|_\infty < \infty$ and Assumption 2.1 (or Assumption 2.2) is satisfied. If $\underline{\lim}_{\epsilon\to 0}\frac{M_2(\frac{\epsilon}{2})}{M_2(\epsilon)} > 1$, then

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

where ϵ_n is determined by $M_2(\epsilon_n) = n\epsilon_n^2$.

Using the relationship between L_2 and L_q $(1 \le q < 2)$ distances and applying Theorem 2.1, we have the following corollary.

Corollary 2.5: Suppose the conditions in Theorem 2.3 are satisfied and $\underline{\lim}_{\epsilon\to 0}\frac{M_q(\frac{\epsilon}{2})}{M_q(\epsilon)} > 1$ for some $q \in [1, 2)$. Let $\underline{\epsilon}_{n,q}$ satisfy $M_q(\underline{\epsilon}_{n,q}) = n\epsilon_n^2$. Then

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

If the packing entropies under L_2 and L_q are equivalent, 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 $\int (f-g)^2 d\mu \le (\| f + g \|_\infty) \int |f - g| d\mu$, we know $M_1(\epsilon) \le M_2(\epsilon) \le M_1(\frac{\epsilon^2}{\sup_{f\in\mathcal{F}}\|f\|_\infty})$. Then the corresponding lower bound for L_1 risk may vary from ϵ_n to ϵ_n^2 depending on how different the two entropies are.

2.2.3 Minimax rates under K-L loss

For the square root K-L distance, Assumption 2.0 is not necessarily satisfied for general classes of densities. We next discuss Assumption 2.0 for d_K and present some more results concerning the K-L risk.

Lemma 2.3: Assume $D(p_\theta, p_{\theta'}) \le \overline{A} d^2(\theta, \theta')$ for all $\theta, \theta' \in S$ and $D(p_\theta, p_{\theta'}) \ge A_0 d^2(\theta, \theta')$ for all $\theta, \theta' \in \Theta$, where d is a metric on Θ. Then Assumption 2.0 is satisfied for d_K with $A = \sqrt{\frac{\overline{A}}{A_0}}$ for any choice of $\overline{S} \subset \Theta$.

Remark: It suffices to assume $D(p_\theta, p_{\theta'}) \le \overline{A} d^2(\theta, \theta')$ for $\theta, \theta' \in S$ when $d^2(\theta, \theta')$ is small. Further more, if the condition is satisfied only locally, then local entropy can be used to derive minimax lower bounds.

The conditions in the lemma are satisfied by the normal location family and the regression family considered in Section 4.

Proof of Lemma 2.3: From the assumptions,

$$\begin{aligned} d_K(\theta,\theta') &\le \sqrt{\overline{A}}\, d(\theta,\theta') \\ &\le \sqrt{\overline{A}}\left(d(\theta,\widetilde{\theta}) + d(\theta,\widetilde{\theta})\right) \\ &\le \sqrt{\frac{\overline{A}}{A_0}}\left(d_K(\theta,\widetilde{\theta}) + d_K(\theta,\widetilde{\theta})\right). \end{aligned}$$

Lemma 2.4: For the square root K-L distance, each of the following two equivalent conditions is sufficient for the satisfaction of Assumption 2.0 with $0 < A < 1$.

1. $D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \frac{p_\theta + p_{\theta'}}{2}\right) \ge AD(p_\theta \parallel p_{\theta'})$ for all $\theta, \theta' \in S$.
2. $D\left(\frac{p_\theta + p_{\theta'}}{2} \parallel p_\theta\right) \le \frac{1}{2}(\frac{1}{A} - 1)D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right)$ for all $\theta, \theta' \in S$.

Remark: Because $D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \frac{p_\theta + p_{\theta'}}{2}\right) \le 2\log 2$, the above condition 1 necessarily enforce the family to be totally bounded in K-L distance. If $\{p_\theta : \theta \in S\}$ is a convex family (so that $\frac{p_\theta + p_{\theta'}}{2} = p_{\widetilde{\theta}}$ for some $\widetilde{\theta} \in S$), then the conditions are necessary as well as sufficient. It is enough to assume condition 1 is satisfied when $D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \frac{p_\theta + p_{\theta'}}{2}\right)$ is small or satisfied only locally (for θ' close to a fixed point θ in terms of K-L distance) if local entropy condition is used.

Proof of Lemma 2.4: The sufficiency of condition 1 comes from the fact that $\frac{p_\theta + p_{\theta'}}{2}$ minimizes $\frac{1}{2}D(p_\theta \parallel p) + \frac{1}{2}D(p_{\theta'} \parallel p)$ over all densities p. For the second condition, because

$$\begin{aligned} & D\left(p_\theta \parallel \tfrac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \tfrac{p_\theta + p_{\theta'}}{2}\right) - D(p_\theta \parallel p_{\theta'}) \\ = & \int p_\theta \log \frac{p_\theta}{\frac{p_\theta + p_{\theta'}}{2}} + \int p_{\theta'} \log \frac{p_{\theta'}}{\frac{p_\theta + p_{\theta'}}{2}} - \int p_\theta \log \frac{p_\theta}{p_{\theta'}} \\ = & \int p_\theta \log \frac{\frac{p_{\theta'}}{2}}{\frac{p_\theta + p_{\theta'}}{2}} + \int p_{\theta'} \log \frac{\frac{p_{\theta'}}{2}}{\frac{p_\theta + p_{\theta'}}{2}} \\ = & -2\int \tfrac{p_\theta + p_{\theta'}}{2} \log \frac{\frac{p_\theta + p_{\theta'}}{2}}{p_{\theta'}} = -2D\left(\tfrac{p_\theta + p_{\theta'}}{2} \parallel p_{\theta'}\right), \end{aligned}$$

we have

$$D(p_\theta \parallel p_{\theta'}) = D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + 2D\left(\frac{p_\theta + p_{\theta'}}{2} \parallel p_{\theta'}\right).$$

(This equality is a special case of a parallelogram identity, see Csiszár and Körner (1981, pp. 59)). Thus,

$$\begin{aligned} D(p_\theta \parallel p_{\theta'}) & \\ &\leq D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + (\frac{1}{A} - 1)D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) \\ &\leq \frac{1}{A}\left(D\left(p_\theta \parallel \frac{p_\theta + p_{\theta'}}{2}\right) + D\left(p_{\theta'} \parallel \frac{p_\theta + p_{\theta'}}{2}\right)\right). \end{aligned}$$

When $\{p_\theta : \theta \in \overline{S}\}$ is convex, then from the above lemma, a sufficient condition for the satisfaction of Assumption 2.0 is that there exists a constant $0 < A < 1$ such that $D(p_\theta \parallel p_{\theta'}) \leq \frac{1}{2}(\frac{1}{A} - 1)D(p_{\theta'} \parallel p_\theta)$ for any $\theta, \theta' \in \overline{S}$.

Corollary 2.6: Suppose the conditions in Lemma 2.3 or Lemma 2.4 are satisfied. Then

$$\frac{A^2\underline{\epsilon}_n^2}{8} \leq \min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d_K^2(\theta, \hat{\theta}) \leq 2\epsilon_n^2,$$

where ϵ_n and $\underline{\epsilon}_n$ are determined by $M_K(\epsilon_n) = n\epsilon_n^2$ and $M_K(\underline{\epsilon}_n) = 4n\epsilon_n^2 + 2\log 2$.

As discussed before, if $\underline{\lim}_{\epsilon \to 0} \frac{M_K(\frac{\epsilon}{2})}{M_K(\epsilon)} > 0$, then ϵ_n and $\underline{\epsilon}_n$ are of the same order, which determines the minimax rate of convergence.

As mention before, the conditions in both lemmas are satisfied if the log-density in the class are uniformly bounded. When these conditions are not satisfied (for instance, if the densities in the class have different supports, then the conditions in Lemma 2.3 can not be satisfied), the following result provides minimax lower bound involving only the Hellinger metric entropy.

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

Lemma 2.5: 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 ϵ (but not on f) such that

$$D(f \parallel \tilde{g}) \leq 2\left(2 + \log\left(\frac{9T}{4\epsilon^2}\right)\right)\left(9 + 8(8T - 1)^2\right)\epsilon^2.$$

Bounds analogous to Lemma 2.5 are in Barron, Birgé and Massart (1995), Wong and Shen (1995).

Proof of Lemma 2.5: The proof is by a truncation of g from above and below. Let $G = \{x : g(x) \leq 4T\}$. Let $\overline{g} = gI_G + 4TI_{G^c}$. 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 $f(x) \leq T \leq \frac{1}{4}g(x)$ for $x \in G^c$, it follows that $\int_{G^c} \left(\sqrt{g} - \sqrt{\frac{g}{4}}\right)^2 \leq \int_{G^c} \left(\sqrt{f} - \sqrt{g}\right)^2 d\mu \leq \epsilon^2$. Thus $\int_{G^c} g d\mu \leq 4\epsilon^2$, which implies $1 - 4\epsilon^2 \leq \int \overline{g} d\mu \leq 1$ and $\int \left(\sqrt{g} - \sqrt{\overline{g}}\right)^2 d\mu \leq \int_{G^c} g d\mu \leq 4\epsilon^2$. Let $\widetilde{g} = \frac{\overline{g}+4\epsilon^2}{\int \overline{g} d\mu + 4\epsilon^2}$. Clearly $\widetilde{g}$ is a probability density function with respect to μ. For $0 \leq z \leq 4T$, by simple calculation using $1 - 4\epsilon^2 \leq \int \overline{g} d\mu \leq 1$, we have $|\sqrt{z} - \sqrt{\frac{z+\epsilon^2}{\int \overline{g} d\mu + \epsilon^2}}| \leq 2(8T-1)\epsilon$. Thus $\int \left(\sqrt{\overline{g}} - \sqrt{\widetilde{g}}\right)^2 d\mu \leq 4(8T-1)^2\epsilon^2$. Therefore, by triangle inequality,

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

That is $d_H^2(f,\widetilde{g}) \leq 2\left(9 + 8(8T-1)^2\right)\epsilon^2$. Because $\frac{f}{\widetilde{g}} \leq \frac{T}{\frac{4\epsilon^2}{\int \overline{g} d\mu + 4\epsilon^2}} \leq \frac{9T}{4\epsilon^2}$, by Lemma 4.5 in Section 6 of Chapter 4,

$$D(f \parallel \widetilde{g}) \leq 2\left(2 + \log\left(\frac{9T}{4\epsilon^2}\right)\right)\left(9 + 8(8T-1)^2\right)\epsilon^2,$$

which completes the proof.

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 ϵ-net under d_H can always result in an η-net under d_K, where $\eta = \sqrt{2\left(2 + \log\left(\frac{9T}{4\epsilon^2}\right)\right)(9 + 8(8T-1)^2)\epsilon} \leq T_1 \epsilon \log\left(\frac{2}{\epsilon}\right)$ for $0 < \epsilon \leq \sqrt{2}$ with T_1 being a constant depending only on T. Thus for $\epsilon \geq \sqrt{\frac{\log 2}{n}}$, $V_K\left(\frac{T_1}{2}\epsilon \log\left(\frac{4n}{\log 2}\right)\right) \leq M_H(\epsilon)$ or equivalently,

$$V_K(\epsilon) \leq M_H\left(\frac{2\epsilon}{T_1 \log\left(\frac{4n}{\log 2}\right)}\right). \tag{2.5}$$

Let ϵ_n satisfy $M_H\left(\frac{2\epsilon_n}{T_1\log\left(\frac{4n}{\log 2}\right)}\right) = n\epsilon_n^2$ (then $\epsilon_n \geq \frac{\sqrt{\log 2}}{\sqrt{n}}$ under the assumption $M_d(\epsilon) \to \infty$ as $\epsilon \to 0$, hence (2.5) is satisfied with $\epsilon = \epsilon_n$) and let $\underline{\epsilon}_n$ be chosen such that $M_H(\underline{\epsilon}_n) = 4n\epsilon_n^2 + 2\log 2$. From Theorem 2.2, we have the following result.

Theorem 2.4: Assume the packing entropy $M_H(\epsilon) < \infty$ and $M_H(\epsilon) \to \infty$ as $\epsilon \to 0$ for the density class $\mathcal{F}$ with $||f||_\infty \leq T$ for each $f \in \mathcal{F}$. Then with ϵ_n and $\underline{\epsilon}_n$ as defined above,

$$\frac{\underline{\epsilon}_n^2}{8} \leq \min_{\hat{f}} \max_{f\in\mathcal{F}} E d_H^2(f, \hat{f}) \leq \min_{\hat{f}} \max_{f\in\mathcal{F}} E D(f \parallel \hat{f}) \leq 2\epsilon_n^2.$$

Remark: Due to the presence of $\log n$ term in the determination of ϵ_n, $\underline{\epsilon}_n^2$ and ϵ_n^2 are typically of order $\frac{\tau_n^2}{\log n}$ and $\tau_n^2 \log n$ respectively for nonparametric smooth families, where τ_n is chosen such that $M_H(\tau_n) = n\tau_n^2$. See Barron, Birgé and Massart (1995) for related conclusions. We suspect the extra $\log n$ might be necessary for the upper bound without any regularity condition relating K-L distance to Hellinger distance.

2.2.4 Lower bounding minimax risk through upper bounds

From the proof of Theorem 2.1, we see that an upper bound on $\max_{\theta\in S} D(p_{X^n|\theta} \parallel q_{X^n})$ with any choice of q_{X^n} together with the packing entropy determines an minimax lower bound as stated in the following corollary.

Corollary 2.7: Assume Assumption 2.0 is satisfied for distance d and there exists a density function q_n such that

$$\max_{\theta\in S} D(p_{X^n|\theta} \parallel q_{X^n}) \leq n\delta_n^2.$$

Let $\underline{\eta}_n$ be chosen such that

$$M_d(\underline{\eta}_{n,d}) = 2(n\delta_n^2 + \log 2).$$

Then the minimax risk satisfies

$$\min_{\hat{\theta}\in\mathcal{A}_n} \max_{\theta\in S} E_\theta d^2(\theta, \hat{\theta}) \geq \frac{A^2\underline{\eta}_{n,d}^2}{8}.$$

If a good upper bound on $\max_{\theta\in S} D(p_{X^n|\theta} \parallel q_{X^n})$ is available with a suitable choice of q (not necessarily constructed as in the proof of Theorem 2.1), then a lower bound on the minimax risk is given by the corollary.

The term $\max_{\theta\in S} D(p_{X^n|\theta} \parallel q_{X^n})$ is the maximum redundancy of Shannon codes based on densities q for an i.i.d. sequence of data from density $p_\theta, \theta \in S$. This redundancy is closely related to the K-L risk. In fact, based on a good sequence of estimators, a good data compression strategy could be constructed and vise versa. More precisely, we have the following lemma.

Lemma 2.6: Suppose there exists a sequence of estimators $\widetilde{\theta}_k$ based on $X_1, ...X_k$ for $k \geq 1$ such that

$$\max_{\theta\in S} ED(p_\theta \parallel p_{\widetilde{\theta}_k}) \leq b_k^2,\ k = 1, ...n-1.$$

Let $b_0 = \max_{\theta\in S} D(p_\theta \parallel p^0)$ for any given p^0. Then there exists a density q_n such that

$$\max_{\theta\in S} D(p_{X^n|\theta} \parallel q_{X^n}) \leq \sum_{i=0}^{n-1} b_k^2.$$

Conversely, for any density q_n on the sample space of $X_1, ..., X_n$, there exists an estimator $\hat{p}$ such that

$$E_\theta D(p_\theta \parallel \hat{p}) \leq \frac{1}{n} D(p_\theta^n \parallel q_n) \text{ for all } \theta \in S.$$

Proof: Given the estimator sequence $\widetilde{\theta}_k$, define $q(x_{k+1}|x^k) = p_{\widetilde{\theta}_k}(x_{k+1})$ for $k \geq 1$ and $q(x_1|x^0) = p^0$. Let $q_n(x_1, ..., x_n) = \Pi_{k=0}^{n-1} q(x_{k+1}|x^k)$. Then q_n is a probability density function. Following an argument similar to the one for proving the upper bound part of Theorem 2.2, we have $D(p_{X^n|\theta} \parallel q_{X^n}) \leq \sum_{i=0}^{n-1} b_k^2$.

For the second assertion, for any $q_n \in Q_n$, we can rewrite it as $q_n(x_1, ...x_n) = h_1(x_1)h_2(x_2|x_1) \cdot \cdot \cdot h_n(x_n|x^{n-1})$, where $h_i(x) = h_i(x|x^{i-1})$ is the conditional density of X_i given $X^{i-1} = x^{i-1}$ according to the joint density q_n. Then

$$\begin{aligned} D(p_\theta^n \parallel q_n) &= \int p_\theta(x_1) \cdots p_\theta(x_n) \log \frac{p_\theta(x_1) \cdots p_\theta(x_n)}{h_1(x_1)h_2(x_2|x_1) \cdots h_n(x_n|x^{n-1})} d\mu \\ &= \sum_{i=1}^{n} \int p_\theta(x_1) \cdots p_\theta(x_n) \log \frac{p_\theta(x_i)}{h_i(x_i|x^{i-1})} d\mu \\ &= \sum_{i=1}^{n} E_\theta D(p_\theta \parallel h_i). \end{aligned}$$

Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} h_i(x)$. Then $\hat{p}$ is a density estimator of p_θ. From the proof of the upper bound in Theorem 2.2,

$$E_\theta D(p_\theta \parallel \hat{p}) \leq \frac{1}{n} D(p_\theta^n \parallel q_n).$$

This completes the proof of the lemma.

From Lemma 2.6 and Corollary 2.7, we have the following corollary.

Corollary 2.8: For a sequence of estimators $\hat{\theta}_k$ based on $X_1, ...X_k$, $1 \leq k \leq n$, let $\max_{\theta \in S} ED(p_\theta \parallel p_{\hat{\theta}_k}) = b_k^2$, then for the minimax risk, we have

$$\min_{\hat{\theta} \in \mathcal{A}_n} \max_{\theta \in S} E_\theta d^2(\theta, \hat{\theta}) \geq \frac{A^2 \underline{\sigma}_{n,d}^2}{8},$$

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

$$M_d(\underline{\sigma}_{n,d}) = 2\left(\sum_{i=0}^{n-1} b_i^2 + \log 2\right).$$

Remark: For smooth nonparametric classes, $\sum_{i=0}^{n-1} b_i^2$ is often of the same order of nb_{n-1}^2 for a sequence of estimators converging at the optimal rate. Then $\underline{\sigma}_{n,d}$ gives the right rate.

It may seem mysterious that an upper bound also forecasts a lower bound. Our explanation is as follows. The smaller the upper bound is, the closer the densities in the class to a fixed $q(x_1, ..., x_n)$ in terms of the Kullback divergence, which suggests the densities are harder to distinguish as revealed by Fano's inequality.

2.3 Application in data compression

The obtained theorems can be used to get bounds on the minimax redundancy for data compression. Let $X_1, ..., X_n$ be an i.i.d. sample of discrete random variable from p_θ, $\theta \in S$. Let $q_n(x_1, ...x_n)$ be a density (probability mass) function. The redundancy of the Shannon code using density q_n is the difference of its expected codelength and the expected codelength of the Shannon code using the true density p_θ, that is, $D(p_\theta^n \parallel q_n)$. Formally, we examine the minimax properties of the game with loss $D(p_\theta^n \parallel q_n)$ for continuous random variables also.

In that case, $D(p_\theta^n \parallel q_n)$ 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 (1986), 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.

Let Q_n be the collection of all density functions on the sample space $\mathcal{X}^n$ of $(X_1, ..., X_n)$. From Lemma 2.6, we have the following result connecting the minimax redundancy with the minimax risk.

Corollary 2.9:

$$n\min_{\hat{p}\in\mathcal{P}_n}\max_{\theta\in S}E_\theta D(p_\theta \parallel \hat{p}) \le \min_{q_n\in Q_n}\max_{\theta\in S}D(p_\theta^n \parallel q_n) \le \sum_{i=0}^{n-1}\min_{\hat{p}_i\in\mathcal{P}_i}\max_{\theta\in S}E_\theta D(p_\theta \parallel \hat{p}_i),$$

where for $i = 0$, $\hat{p}_i$ is any fixed density.

Remark: For smooth nonparametric density classes, $n\cdot\min_{\hat{p}\in\mathcal{P}_n}\max_{\theta\in S}E_\theta D(p_\theta \parallel \hat{p})$ 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 $\frac{m}{n}\log n$, where m is the number of parameters in the family. But $n\cdot\min_{\hat{p}\in\mathcal{P}_n}\max_{\theta\in S}E_\theta D(p_\theta \parallel \hat{p})$ is bounded by a constant.

Now we take $\Theta = \mathcal{P}$ to be the set of all probability densities on $\mathcal{X}$ and $S \subset \mathcal{P}$ to be a subclass. The action space is assumed to be the set of all densities $\overline{S} = \mathcal{P}$ (so as to include estimates such as $\hat{p}$ constructed in the proof of Theorem 2.2). Let $d(p, p')$ be a metric on $\mathcal{P}$. 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 ϵ_n such that $\epsilon_n^2 = \frac{V(\epsilon_n)}{n}$ and choose $\underline{\epsilon}_{n,d}$ be a radius ϵ such that $M_d(\underline{\epsilon}_{n,d}) = 4n\epsilon_n^2 + 2\log 2$.

Theorem 2.5: Assume that $D(p \parallel p') \geq A_0 d^2(p, p')$ for all $p, p' \in \mathcal{P}$. Then we have

$$\frac{nA_0 A\underline{\epsilon}_{n,d}^2}{8} \leq \min_{q_n} \max_{p \in S} D(p^n \parallel q_n) \leq 2n\epsilon_n^2,$$

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

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

Proof of Theorem 2.5: The lower bound follows from Theorem 2.1 and Lemma 2.6. For the upper bound, consider the code based on $q(x^n) = \frac{1}{|G_{\epsilon_n}|} \sum_{p \in G_{\epsilon_n}} p(x^n)$, the mixture with respect to the uniform prior on an ϵ_n-net G_{ϵ_n} of S. Then the redundancy is $D(p^n \parallel q_n) \leq V(\epsilon_n) + n\epsilon_n^2 \leq 2n\epsilon_n^2$ as in equation (2.2).

The redundancy for data compression is connected with the cumulative risk of density estimation under K-L loss (see, e.g., Clark and Barron (1990)). This risk is natural for consideration when we estimate the density sequentially based on observations obtained so far, predict the next observation, and then adjust the estimator once a new observation is obtained.

Let δ be an estimation procedure. That is, for each sample size n, it produces an estimator $\hat{f}_n$ based on X_1, ..., X_n. Let $\hat{f}_0 = f_0$ be an initial guess density without any observations. Then the cumulative risk under the K-L distance up to $n-1$ observations $R_{cum}(f, \delta, n)$ is defined as

$$R_{cum}(f, \delta, n) = \sum_{i=0}^{n-1} E_f D(f \parallel \hat{f}_i).$$

This is the cumulative redundancy of predictive codes (using Shannon code based on $\hat{f}_i$ to encode the next observation X_{i+1}) in a information theory context. This cumulative redundancy is exactly the redundancy of data compression for X_1, ..., X_n. As seen in Lemma 2.6, any estimation procedure δ can result in a density on sample space of X_1, ..., X_n, which can be used to construct a data compression scheme. Let $q_\delta^{(n)}(x_1, ..., x_n) = f_0(x_1) \cdot \hat{f}_1(x_2|x_1) \cdots \hat{f}_{n-1}(x_n|x_1, ..., x_{n-1})$. Then As shown before,

$$D(f^n \parallel q_\delta^{(n)}) = R_{cum}(f, \delta, n).$$

Thus

$$\min_\delta \max_{f\in\mathcal{F}} R_{cum}(f,\delta,n) = \min_{q^{(n)}} \max_{f\in\mathcal{F}} D(f^n \parallel q^{(n)}),$$

where $q^{(n)}$ is over all densities on sample space of X_1, ..., X_n.

Suppose the minimizer of $\max_{f\in\mathcal{F}} D(f^n \parallel q^{(n)})$ exists, say $q_*^{(n)}$. In some sense, $q_*^{(n)}$ is a center of the product density class $\{f^n,\ f\in\mathcal{F}\}$ and $\max_{f\in\mathcal{F}} D(f^n \parallel q_*^{(n)})$ is the "radius" of the class $\{f^n,\ f\in\mathcal{F}\}$. Theorem 2.5 provides useful bounds on this radius.

2.4 Application in nonparametric regression

Consider the regression model

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

Suppose the errors ε_i, $1 \le i \le n$ are i.i.d. with $N(0,1)$ distribution. The explanatory variables x_i, $1 \le i \le n$ are i.i.d. with 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 $\parallel u-v \parallel_{L_2(h)} = \sqrt{\int (u-v)^2 h d\mu}$ 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 ϵ-packing set under $L_2(h)$ norm. Assume $M_2(\epsilon) < \infty$ for every $\epsilon > 0$ and $M_2(\epsilon) \to \infty$ as $\epsilon \to 0$. Choose ϵ_n such that

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

Let $\underline{\epsilon}_n$ satisfy

$$M_2(\sqrt{2}\underline{\epsilon}_n) = 4n\epsilon_n^2 + 2\log 2.$$

Theorem 2.6: The minimax $L_2(h)$ risk for the regression function estimation is lower bounded by a rate determined by the L_2 packing entropy of $\mathcal{U}$ as follows:

$$\min_{\hat{u}} \max_{u\in\mathcal{U}} \parallel u - \hat{u} \parallel^2_{L_2(h)} \ \ge \frac{\epsilon_n^2}{4}.$$

Proof: Denote the joint density $\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-u(x))^2}h(x)$ of (X,Y) with regression function u by p_u. Then

$$\begin{aligned} D(p_u \parallel p_v) &= E_u \frac{1}{2}\left((Y-u(X)^2-(Y-v(X)^2\right) \\ &= E_u \frac{1}{2}(u(X)-v(X))^2 \\ &= \frac{1}{2}\int (u(x)-v(x))^2 h(x)dx. \end{aligned}$$

Let $S = \{u : u \in \mathcal{U}\}$ and $\overline{S} = \{u : \| u \|^2_{L_2(h)} < \infty\}$. Let $\mathcal{A}_n$ be the collection of the regression estimators which maps from the sample space to $\overline{S}$. Let $d^2(u,v) = D(p_u \parallel p_v)$. From Theorem 2.1, we have

$$\min_{\hat{u}\in\mathcal{A}_n}\max_{u\in\mathcal{U}} ED(p_u \parallel p_{\hat{u}}) \geq \frac{\epsilon_n^2}{8}.$$

The conclusion follows.

To get good upper bounds, a little more work is needed. The upper bound in Theorem 2.2 is not directly applicable, because it is less clear whether the minimax K-L risk of estimating the joint density of (X,Y) is lower bounded by a multiple of the minimax risk of estimating the regression function. Under some conditions, we show that it is indeed the case. To that end, Hellinger risk of estimating the density of (X,Y) is used as an intermediate quantity.

We assume $\sup_{u\in\mathcal{U}} \| u \|_\infty \leq L < \infty$, which will be needed in our analysis. From Theorem 2.2, there exists a density estimator $\hat{p}_n$ of the joint density such that $\max_{u\in\mathcal{U}} ED(p_u \parallel \hat{p}_n) \leq 2\epsilon_n^2$. It follows that $\max_{u\in\mathcal{U}} Ed_H^2(p_u,\hat{p}_n) \leq 2\epsilon_n^2$. More precisely, let $u_1(x)$, $u_2(x)$, ..., $u_N(x)$ be a covering set in $\mathcal{U}$ under $L_2(h)$ norm with covering radius ϵ_n, then the estimator constructed in the proof of Theorem 2.2 has the form $\hat{p}_n = \frac{1}{n}\sum_{i=0}^{n-1}\hat{p}_i$, where

$$\begin{aligned} \hat{p}_i\left(x,y \mid (X_l,Y_l)_{l=1}^i\right) &= \frac{h(x)\sum_{j=1}^N \left(\frac{1}{\sqrt{2\pi}}\right)^{i+1} e^{-\frac{1}{2}\left(\sum_{l=1}^i (y_l-u_j(x_l))^2+(y-u_j(x))^2\right)}}{\sum_{j=1}^N \left(\frac{1}{\sqrt{2\pi}}\right)^{i} e^{-\frac{1}{2}\sum_{l=1}^i (y_l-u_j(x_l))^2}} \\ &= h(x)\hat{g}_i\left(y \mid x,(X_l,Y_l)_{l=1}^i\right) \end{aligned}$$

for $i \geq 1$ and $\hat{p}_0(x,y) = h(x)\cdot\frac{1}{N}\sum_{j=1}^N \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-u_j(x))^2}$. Thus the marginal density of X in the joint density $\hat{p}_n$ is $h(x)$ and the conditional density of Y given $X=x$ is $\hat{g}_n(y \mid x,(X_l,Y_l)_{l=1}^n) = \frac{1}{n}\sum_{i=0}^{n-1}\hat{g}_i(y \mid x,(X_l,Y_l)_{l=1}^i)$. For given x and $(X_l,Y_l)_{l=1}^n$, let $\tilde{u}_n$ be

of X in the joint density $\hat{p}_n$ is $h(x)$ and the conditional density of Y given $X=x$ is $\hat{g}_n(y \mid x,(X_l,Y_l)_{l=1}^n) = \frac{1}{n}\sum_{i=0}^{n-1}\hat{g}_i(y \mid x,(X_l,Y_l)_{l=1}^i)$. For given x and $(X_l,Y_l)_{l=1}^n$, let $\tilde{u}_n$ be

the minimizer of $d_H\left(\hat{g}_n(y), \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-z)^2}\right)$ over z with $|z| \leq L$. Then $\widetilde{u}_n(x)$ is an estimator of $u(x)$ based on $(X_l, Y_l)_{l=1}^n$. By triangle inequality, given x and $(X_l, Y_l)_{l=1}^n$,

$$\begin{aligned} & d_H\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-u(x))^2}, \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-\widetilde{u}_n(x))^2}\right) \\ \leq\ & d_H\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-u(x))^2}, \hat{g}_n(y)\right) + d_H\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-\widetilde{u}_n(x))^2}, \hat{g}_n(y)\right) \\ \leq\ & 2d_H\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-u(x))^2}, \hat{g}_n(y)\right). \end{aligned}$$

Because for two joint densities $h(x)g_1(y \mid x)$ and $h(x)g_2(y \mid x)$ with the same marginal density $h(x)$ of X, the Hellinger distance between the joint densities equals

$$\int h(x) d_H^2\left(g_1(y \mid x), g_2(y \mid x)\right) d\mu,$$

so from the above inequality, given $(X_l, Y_l)_{l=1}^n$,

$$d_H^2(p_u, p_{\widetilde{u}_n}) \leq 4 d_H^2(p_u, \hat{p}_n).$$

It follows that

$$\max_{u \in \mathcal{U}} E d_H^2(p_u, p_{\widetilde{u}_n}) \leq 4 \max_{u \in \mathcal{U}} E d_H^2(p_u, \hat{p}_n) \leq 8\epsilon_n^2.$$

Now

$$\begin{aligned} E d_H^2(p_u, p_{\widetilde{u}_n}) &= 2E\left(1 - \int \sqrt{p_{\widetilde{u}_n} p_u} dy \times d\mu\right) \\ &= 2E \int h(x)\left(1 - e^{-\frac{1}{8}(u(x)-\widetilde{u}_n(x))^2}\right) d\mu. \end{aligned}$$

Because $\frac{1}{8}(u(x) - \widetilde{u}_n(x))^2 \leq \frac{1}{2}L^2$, $\left(1 - e^{-\frac{1}{8}(u(x)-\widetilde{u}_n(x))^2}\right) \geq \frac{1}{8}e^{-\frac{1}{2}L^2} \cdot (u(x) - \widetilde{u}_n(x))^2$. It follows that

$$\max_{u \in \mathcal{U}} E \int (u(x) - \widetilde{u}_n(x))^2 d\mu \leq 16 e^{\frac{1}{2}L^2} \max_{u \in \mathcal{U}} E d_H^2(p_u, \hat{p}_n) \leq 128 e^{\frac{1}{2}L^2} \epsilon_n^2.$$

Thus we have the following result.

Theorem 2.7: Assume $\sup_{u \in \mathcal{U}} \| u \|_\infty \leq L$. Then

$$\min_{\hat{u}} \max_{u \in \mathcal{U}} \| u - \hat{u} \|_{L_2(h)}^2 \leq 128 e^{\frac{1}{2}L^2} \epsilon_n^2.$$

If further $\underline{\lim}_{\epsilon \to 0} \frac{M_2(\frac{\epsilon}{2})}{M_2(\epsilon)} > 1$, then

$$\min_{\hat{u}} \max_{u \in \mathcal{U}} \| u - \hat{u} \|_{L_2(h)}^2 \asymp \epsilon_n^2.$$

Remark: Using a similar argument, it can be shown that the above conclusion is still true if the error distribution is assumed to be double exponential in stead of normal.

Previously, minimax rates of convergence for nonparametric regression are identified for specific smooth classes such as Lipschitz classes, Sobolev classes and Besov classes by Stone (1982), Nemirovskii (1986), Nemirovskii, Polyak, and Tsybakov (1985), Donoho, Johnstone, et al (1993), Pinsker (1980), Ibragimov and Hasminskii (1982) and others. Here we have shown that under normal (or double exponential) assumption on the error distribution, the minimax L_2 rates are determined by metric entropy.

2.5 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 (or bounds on) metric entropy of a target class, the minimax rates (or bounds) can be determined right away for many smooth nonparametric classes without additional work.

We will consider several function classes on $[0,1]^d$ for some $d \geq 1$ in the examples.

1. Take $d = 1$. Let $\omega^f(h) = \max_{|t|\leq h, 0\leq x\leq 1} |\Delta_t f(x)|$, where $\Delta_t f(x) = f(x+2t) - 2f(x+t) + f(x)$ is the second difference of f at x with increment t (maximum is taken over those t for which is defined). Let $\phi(h)$ be a concave increasing function and let $\Lambda_\phi(C) = \{f : f \text{ is continuous, } |f| \leq C, \text{ and } \omega^f(h) \leq \phi(h)\}$. Then from a result of Clements (1963), the sup-norm metric entropy of this class $M_\infty(\epsilon)$ satisfies $M_\infty(\epsilon) \succeq \frac{1}{\phi^{-1}(\epsilon)}$. An example of such a ϕ is $\phi(h) = h^\alpha$ for $0 < \alpha \leq 1$.

2. Let $\Lambda^d_{r,\alpha}(C_0, C_1, ..., C_r, C)$ be the class of functions f which have all partial derivatives $|D^{(k)}(f)| \leq C_k$ for $k = 0, 1, .., r$, and $|D^{(r)}(f)(x) - D^{(r)}(f)(x+h)| \leq Ch^\alpha$ $(0 < \alpha \leq 1)$. From results of Kolmogorov and Tihomirov (1959) and Clements (1963), for $1 \leq q \leq \infty$, the L_q packing entropy of $\Lambda^d_{r,\alpha}$ is of order $M_q(\epsilon) \asymp \left(\frac{1}{\epsilon}\right)^{\frac{d}{r+\alpha}}$.

3. Let $\Delta^{d,2}_{r,\alpha}(C_0, C_1, ..., C_r, C)$ be the class analogous to $\Lambda^d_{r,\alpha}(C_0, C_1, ..., C_r, C)$ defined in terms of L_2 norm. That is, $\Delta^{d,2}_{r,\alpha}(C_0, C_1, ..., C_r, C)$ consists of functions f which have all partial derivatives $\| D^{(k)}(f) \|_2 \leq C_k$ for $k = 0, 1, .., r$, and $\| D^{(r)}(f)(x) - D^{(r)}(f)(x+$

$h) \|_2 \leq Ch^\alpha$ $(0 < \alpha \leq 1)$. Then from Lorentz (1966), the L_2 metric entropy of this class is also of order $\left(\frac{1}{\epsilon}\right)^{\frac{d}{r+\alpha}}$.

4. Let $P_\alpha(C)$ be the class of real functions $f(x)$ on $[0, 2\pi]$, periodic with period 2π, with mean value zero and having a derivative of order α in L^2 (in the sense of Weyl) uniformly bounded in mean by C. It was shown by Kolmogorov and Tihomirov (1959) that the L_2 metric entropy of this class is of order $\left(\frac{1}{\epsilon}\right)^{\frac{1}{\alpha}}$.

Density estimation.

Assume $\underline{\lim}_{\epsilon \to 0} \frac{M_\infty(\frac{\epsilon}{2})}{M_\infty(\epsilon)} > 1$ for $\Lambda_\phi(C)$. Consider the density class of f with $\log f \in \Lambda_\phi(C)$. For this class, the log-densities are uniformly bounded and the sup-norm metric entropy is of the same order as $M_\infty(\epsilon)$. Note also that K-L distance is equivalent to L_2 distance and hence upper bounded by a multiple of the squared sup-norm distance. Applying Theorem 2.1, we have $\min_{\hat{f}} \max_{\log f \in \Lambda_\phi(C)} E \parallel f - \hat{f} \parallel_\infty \succeq \epsilon_n^2$, where ϵ_n satisfies $n\epsilon_n^2 = M_\infty(\epsilon_n)$. But $M_\infty(\epsilon) \succeq \frac{1}{\phi^{-1}(\epsilon)}$, we obtain

$$\min_{\hat{f}} \max_{\log f \in \Lambda_\phi(C)} E \parallel f - \hat{f} \parallel_\infty \succeq \underline{\epsilon}_n^2,$$

where $\underline{\epsilon}_n$ satisfies $n\underline{\epsilon}_n^2 = \frac{1}{\phi^{-1}(\underline{\epsilon}_n)}$.

Assume $\log f \in \Lambda_{r,\alpha}^d(C_0, C_1, ..., C_r, C)$. Then for this density class, the L_q $(q \geq 1)$ metric entropy is of order $\left(\frac{1}{\epsilon}\right)^{\frac{d}{r+\alpha}}$. From Corollary 2.4, we have

$$\min_{\hat{f}} \max_{\log f \in \Lambda_{r,\alpha}^d} E \parallel f - \hat{f} \parallel_2^2 \asymp n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

Because the log-densities are uniformly bounded, the minimax K-L risk or squared Hellinger risk converges at the same rate. Also because the metric entropies under L_q, $q \geq 1$ are of the same order, using Theorem 2.1, we have $\min_{\hat{f}} \max_{\log f \in \Lambda_{r,\alpha}^d} E \parallel f - \hat{f} \parallel_q^2 \succeq n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}$. Together with the upper bound rate on L^2 risk, we have for $1 \leq q \leq 2$,

$$\min_{\hat{f}} \max_{\log f \in \Lambda_{r,\alpha}^d} E \parallel f - \hat{f} \parallel_q^2 \asymp n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

Now we consider classes of densities which may not be bounded above or below from zero. Let $\widetilde{\Lambda}_{r,\alpha}^d(C_0, C_1, ..., C_r, C)$ and $\widetilde{\Delta}_{r,\alpha}^{d,2}(C_0, C_1, ..., C_r, C)$ be the density functions in $\Lambda_{r,\alpha}^d(C_0, C_1, ..., C_r, C)$ and $\Delta_{r,\alpha}^{d,2}(C_0, C_1, ..., C_r, C)$ respectively. When the constants $C, C_0, ..., C_r$ are large enough (which is assumed here), the orders of L_q metric entropies of $\widetilde{\Lambda}_{r,\alpha}^d(C_0, C_1, ..., C_r, C)$

are still $\left(\frac{1}{\epsilon}\right)^{\frac{d}{r+\alpha}}$. For these classes, Assumption 2.1 is satisfied. Therefore, by Theorem 2.3 and Corollary 2.5, for the class $\widetilde{\Lambda}^d_{r,\alpha}$, we have

$$\min_{\hat{f}} \max_{f \in \widetilde{\Lambda}^d_{r,\alpha}} E \parallel f - \hat{f} \parallel_q^2 \asymp n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}, \text{ for } 1 \leq q \leq 2.$$

The densities in $\widetilde{\Delta}^{d,2}_{r,\alpha}$ are not necessarily bounded. But from Proposition 2.1, for the squared L_1 risk,

$$\min_{\hat{f}} \max_{f \in \widetilde{\Delta}^{d,2}_{r,\alpha}} E \parallel f - \hat{f} \parallel_1^2 \preceq n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

Because $\widetilde{\Lambda}^d_{r,\alpha}(C_0, C_1, ..., C_r, C) \subset \widetilde{\Delta}^{d,2}_{r,\alpha}(C_0, C_1, ..., C_r, C)$, the lower bound obtained above on the squared L_1 risk is also a lower bound for the larger class. Combining the upper and lower bounds, we have

$$\min_{\hat{f}} \max_{f \in \widetilde{\Delta}^{d,2}_{r,\alpha}} E \parallel f - \hat{f} \parallel_1^2 \asymp n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

Birgé (1986) and Devroye (1987) obtained similar results for $\widetilde{\Lambda}^d_{r,\alpha}$ with additional constructions of special subsets.

Regression function estimation.

Consider the regression problem in Section 4. Let h be the density of the explanatory variable X. If $|\log h|$ is bounded, then for $\Lambda^d_{r,\alpha}(C_0, C_1, ..., C_r, C)$ and $P_\alpha(C)$, the metric entropies under $L_2(h)$ norm are of the same orders as given before. By Theorems 2.6 and 2.7, we have the following conclusion.

Corollary 2.10: Assume $|\log h|$ is bounded. Then the minimax L^2 rate or bound of estimating a function in $\Lambda^d_{r,\alpha}(C_0, C_1, ..., C_r, C)$ and $P_\alpha(C)$ are given below.

$$\min_{\hat{u}} \max_{u \in \Lambda^d_{r,\alpha}} E \parallel u - \hat{u} \parallel_2^2 \asymp n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

$$\min_{\hat{u}} \max_{u \in P_\alpha(C)} E \parallel u - \hat{u} \parallel_2^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}.$$

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 $\tilde{\Lambda}^d_{r,\alpha}$ is lower bounded in rate as follows (as a consequence of Theorem 2.5):

$$\min_{q_n \in Q_n} \max_{f \in \tilde{\Lambda}^d_{r,\alpha}} \frac{1}{n} D(f^n \parallel q_n) \succeq n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

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

If we assume $\log f \in \tilde{\Lambda}^d_{r,\alpha}$, then the minimax redundancy rate is identified from Theorem 2.5:

$$\min_{q_n \in Q_n} \max_{\log f \in \tilde{\Lambda}^d_{r,\alpha}} \frac{1}{n} D(f^n \parallel q_n) \asymp n^{-\frac{2(r+\alpha)}{2(r+\alpha)+d}}.$$

Chapter 3

Adaptation of Density Estimation

In this chapter, we estimate a density function assumed to be in a countable collection of nonparametric classes. We study the possibility of adaptation and provide adaptive estimators over the classes.

The minimax results in Chapter 2 deal with one general class of densities. In some statistical applications, target classes are chosen to be smooth nonparametric classes such as Sobolev spaces, Lipschitz spaces, etc. The smoothness condition of a function indicates how much the function value may change according to change of the independent variables. Smoothness is often measured by some kind of norm defined in terms of the derivatives of the function. As seen in the examples in Chapter 2, smoothness conditions of a target class affect how fast the minimax risk converges to zero and roughly speaking, a smoother class has a smaller order metric entropy, thus has a better minimax rate of convergence. For various smooth nonparametric classes, a lot of estimation procedures such as kernel methods with predetermined band widths (e.g., see Devroye (1987)) and some sieve methods (e.g., Stone (1990, 1994), Barron and Sheu (1991), Birgé and Massart (1993), Wong and Shen (1995)) have been proposed utilizing the smoothness information. But in practice, we only observe a sample and the smoothness condition of the density is not known to us. A statistical procedure specifically designed for one smoothness condition generally does not work optimally for other classes with different smoothness conditions. This consideration suggests the necessity of adaptation capability of estimators. Of course, one may first obtain an estimator under a smoothness condition somehow chosen based on some rough idea, and then readjust the assumption based on the estimator. Or some times it might

be possible to compare estimators under different smoothness assumptions and choose the best one according to some visual or ad hoc justifications. But these kinds of procedures might depend too much on the user's personal opinion. We here instead are interested in good data-driven strategies.

Though smooth nonparametric classes are the ones most often used in practice, conceptually we do not restrict ourselves to those classes. More generally, we assume that the true density can be from any of a countable collection of density classes. An interesting question is: Can we have one estimator that suits simultaneously for all classes in some sense? We wish to have one estimation procedure which can automatically adjust to the curvature (smoothness) of the true density based only on data. Such an estimation procedure is called an adaptive procedure.

Adaptation is desirable for an estimator, because such an estimator is more flexible and can work well without strong assumptions on the true density. The idea of adaptation dates back over ten years ago. For instance, a variable bandwidth was considered to make an adaptive kernel estimator (e.g., Härdle, Hall and Marron (1985)), and smoothness parameters were adaptively adjusted based on data in smoothing spline estimation (e.g., Craven and Wahba (1979)). Efroimovich (1985) made a great contribution in this direction for density estimation. He considered estimating a density having a Fourier representation satisfying a certain smoothness assumption with smoothness parameter not known. He considered projection estimators of the Fourier coefficients and proposed an adaptive strategy to achieve the minimax rates of convergence without knowing the smoothness parameter in advance. In later years, Donoho, Johnstone and some other researchers (see, e.g., [32] and [33]) advocated the use of wavelet shrinkage estimators in both nonparametric regression and density estimation. They showed that the wavelet shrinkage estimators converge near optimally simultaneously over the Besov spaces without knowing the hyper-parameters in advance. Birgé and Massart (1995), Barron, Birgé and Massart (1995) have had great success in providing model selection theory using general contrast functions.

In this chapter, we are interested in adaptation in terms of the minimax rates of convergence. From now on, we will concentrate on the estimation of density itself in stead of other parametrization (such as root density or log-density). We will only consider K-L loss in this chapter.

Let $\mathcal{F}_j$, $j \geq 1$ be a collection of target classes. Assume $f \in \cup_{j \geq 1} \mathcal{F}_j$. If the conditions in

Corollary 2.4 are satisfied for each class, then for a given class, the estimator constructed in the proof of Theorem 2.2 converges at the optimal rate for that class. Now, without knowing which class contains f, can we have one estimator (not depending on j) such that it converges at optimal rate of the class containing f? If such an estimator exists, we call it an adaptive estimator over the classes $\mathcal{F}_j$, $j \geq 1$.

3.1 Adaptation under entropy conditions

Let $M_j(\epsilon)$, $j \geq 1$ be packing entropies under d_H for the density classes $\mathcal{F}_j$, $j \geq 1$. For simplicity, we assume that for each class, there exist constants C_j and γ_j such that $D(f \parallel g) \leq C_j d_H^2(f,g)$ for $f, g \in \mathcal{F}_j$ with $d_H^2(f,g) \leq \gamma_j$. This condition necessarily requires K-L distance and squared Hellinger distance behave similarly when the densities are close to each other in Hellinger distance. Suppose also that the classes are rich enough such that $\liminf_{\epsilon \to 0} \frac{M_j(\frac{\epsilon}{2})}{M_j(\epsilon)} > 1$ for $j \geq 1$. Applying Corollary 2.3 and Theorem 2.2, for a fixed j, the minimax rate of convergence under K-L distance is $\epsilon_{n,j}^2$, where $\epsilon_{n,j}$ is determined by $\epsilon_{n,j}^2 = \frac{M_j\left(\frac{\epsilon_{n,j}}{\sqrt{C_j}}\right)}{n}$. Using the relationship between density estimation and data compression as discussed in Chapter 2, we obtain the following result.

Theorem 3.1: Under the above conditions, there is a minimax-rate adaptive estimator, that is, an estimator that is simultaneously minimax-rate optimal for $\{\mathcal{F}_j, j \geq 1\}$. Specifically, the estimator $\hat{f}_n$ given in (3.1) based on $X_1, ..., X_n$ has risk bound

$$\max_{f \in \mathcal{F}_j} E_f D(f \parallel \hat{f}_n) \leq const_j \cdot \epsilon_{n,j}^2$$

for all $j \geq 1$.

Proof: We construct an adaptive estimator using some Bayesian mixing idea. As before, for each class j, consider an $\epsilon_{n,j}$-net $G_{\epsilon_{n,j}}$ in $\mathcal{F}_j$ under d_K and the uniform prior on $G_{\epsilon_{n,j}}$ and let $q_j(x^n) = \frac{1}{|G_{\epsilon_{n,j}}|} \sum_{f \in G_{\epsilon_{n,j}}} f^n(x^n)$ be the mixture density over the covering set. Let $\pi(j)$ be positive prior probabilities on the classes satisfying $\sum_{j=1}^{\infty} \pi(j) = 1$. Then let us mix these mixtures over the classes according to the prior $\pi(j)$ on the classes. Let

$$q^{(n)}(x^n) = \sum_{j \geq 1} \pi(j) q_j(x^n).$$

Then this density is close to all densities in the classes in K-L distance sense. In fact, for any $f \in \mathcal{F}_{j^*}$,

$$\begin{aligned} D\left(f^n \parallel q^{(n)}\right) &= \int f^n(x^n)\log\frac{f^n(x^n)}{\sum_{j\geq 1}\pi(j)q_j(x^n)}d\mu \\ &\leq \int f^n(x^n)\log\frac{f^n(x^n)}{\pi(j^*)q_{j^*}(x^n)}d\mu \\ &\leq \log\frac{1}{\pi(j^*)} + \int f^n(x^n)\log\frac{f^n(x^n)}{q_{j^*}(x^n)}d\mu \\ &= \log\frac{1}{\pi(j^*)} + D\left(f^n \parallel q_{j^*}^{(n)}\right). \end{aligned}$$

From previous analysis, we know that $D\left(f^n \parallel q_{j^*}^{(n)}\right) \leq 2n\epsilon_{n,j^*}^2$. Thus

$$D\left(f^n \parallel q^{(n)}\right) \leq \log\frac{1}{\pi(j^*)} + n\epsilon_{n,j^*}^2.$$

As before, let

$$\hat{f}(x) = \frac{1}{n}\sum_{i=0}^{n-1} p\left(x_{i+1} = x|X^i\right) \tag{3.1}$$

be the estimator constructed as a Cesaro average of the Bayes predictive density estimators, where

$$p\left(x|X^i\right) = \frac{\sum_j \pi(j)\left(\frac{1}{|G_{\epsilon_{n,j}}|}\sum_{f\in G_{\epsilon_{n,j}}} f(x)f^i(X^i)\right)}{\sum_j \pi(j)\left(\frac{1}{|G_{\epsilon_{n,j}}|}\sum_{f\in G_{\epsilon_{n,j}}} f^i(X^i)\right)}$$

is Bayes predictive density based on X^i using two layers of priors. Then we obtain

$$\begin{aligned} ED(p_0 \parallel \hat{f}_n) &\leq \tfrac{1}{n}D\left(f^n \parallel q^{(n)}\right) \\ &\leq \tfrac{1}{n}\log\tfrac{1}{\pi(j^*)} + 2\epsilon_{n,j^*}^2. \end{aligned}$$

Thus for every $j^* \geq 1$,

$$\max_{f\in\mathcal{F}_{j^*}} E_f D(f \parallel \hat{f}_n) \leq \frac{1}{n}\log\frac{1}{\pi(j^*)} + 2\epsilon_{n,j^*}^2.$$

Note that $\frac{1}{n}\log\frac{1}{\pi(j^*)}$ does not affect the rate of convergence in $\frac{1}{n}\log\frac{1}{\pi(j^*)} + 2\epsilon_{n,j^*}^2$ and the estimator does not require the knowledge of which class contains the true density. We conclude that $\hat{f}$ is an adaptive estimator in terms of the minimax rates of convergence.

From the above analysis, due to not knowing which class contains the true density, we pay a price of an extra $\frac{1}{n}\log\frac{1}{\pi(j^*)}$ in the risk bound. Because $\log\frac{1}{\pi(j^*)} \to \infty$ as $\pi(j^*) \to 0$, the obtained upper bound becomes useful for a class with small prior probability, only when the sample size is large compared to $\log\frac{1}{\pi(j^*)}$.

3.2 Adaptation based on existing good estimators

In the above subsection, an adaptive estimator is constructed by considering suitable covering sets in the classes and mixing the corresponding densities on the product space. The construction is theoretically easy to handle, but hard to implement for practical applications. For specific classes, many nonparametric procedures have been proposed and have been shown to be minimax optimal (in terms of rate of convergence). Can we obtain adaptive estimators based on these specially constructed estimators for specific function classes?

We are going to answer this question in the positive direction. Again, we take advantage of the connection between estimating a density and data compression to construct a good density on the product space of $(X_1, ..., X_n)$ which is close to all densities in the considered classes in K-L sense and then use the relationship between the two problems in the reverse direction to get back one density estimator that is optimal for all the density classes under some conditions.

Suppose for each class $f \in \mathcal{F}_j$, we have an estimation strategy δ_j producing density estimators $\hat{f}_{j,1}(x|X_1)$, $\hat{f}_{j,2}(x|X_1, X_2)$, ..., $\hat{f}_{j,n-1}(x|X_1, ..., X_{n-1}), ...$ based on observation(s) $\{X_1\}$, $\{X_1, X_2\}$, ..., $\{X_1, X_2, ..., X_{n-1}\}$, and so on. Here we allow estimation strategies to be different for different classes. For instance, we may use kernel estimators for some classes and wavelet estimators for some others. A single estimator will be constructed by mixing these estimators somehow and shown to be good for all the classes in the asymptotic sense.

The following is the recipe to get a good estimator.

1. Construct a good density on the product space $(x_1, ..., x_n)$ for each class.

   Let $f_{j,0}(x)$ be a fixed density function on the sample space. We may take $f_{j,0}(x)$ to be the "centroid" density f^* (or a more easily obtained good one) of the density class $\mathcal{F}_j$ that minimizes $\max_{f \in \mathcal{F}_j} D(f \parallel f^*)$ over all densities (that is, $f_{j,0}(x)$ is close to the minimax density estimator based on no data). However, the choice of $f_{j,0}(x)$ will have no effect on the asymptotic results. Let

   $$q_j^{(n)} = f_{j,0}(x_1) \cdot \hat{f}_{j,1}(x_2|x_1) \cdot \cdots \cdot \hat{f}_{j,n-1}(x_n|x_1, ..., x_{n-1}).$$

   Then $q_j^{(n)}$ is a density function on the product space of $X_1, ..., , X_n$.

2. Average over j to get a mixture density.

Let π_j, $j \geq 1$ be prior probabilities of the density classes satisfying $\pi_j > 0$ for all $j \geq 1$. Let

$$q^{(n)} = \sum \pi_j q_j^{(n)}.$$

It is a mixture density of constructed densities on the product space.

3. Get conditional densities based on $q^{(n)}$.

Let us rewrite the density $q^{(n)}$ as product of conditional densities:

$$q^{(n)} = q_0(x_1) \cdot q_1(x_2|x_1) \cdots q_{n-1}(x_n|x_1, ..., x_{n-1}).$$

4. Final estimator.

Let $\hat{f}_i(x) = q_i(x|X_1, ...X_i)$, $i = 0, ..., n-1$ be final predictive density estimators based on current observations $X_1, ..., X_i$. Call this estimation strategy δ^*. Let

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

We use $\hat{\hat{f}}_n$ as our final estimator of the unknown density based on $X_1, ..., X_{n-1}$.

Let $R(f, \delta_j, n) = D(f \parallel \hat{f}_{j,n})$ and $WR(\delta_j, n) = \max_{f \in \mathcal{F}_j} D(f \parallel \hat{f}_{j,n})$ be the risk at f and worst case risk respectively of estimator $\hat{f}_{j,n}$ (produced using strategy δ_j) based on sample $X_1, ..., X_n$. We next bound the risk $D(f \parallel \hat{\hat{f}}_n)$ in terms of $R(f, \delta_j, i)$, $1 \leq i \leq n-1$.

As before, for any f,

$$\begin{aligned}
E_f D(f \parallel \hat{\hat{f}}_n) &\leq \frac{1}{n} \sum_{i=0}^{n-1} E_f D(f \parallel \hat{f}_i) \\
&\leq \frac{1}{n} D(f^n \parallel q^{(n)}) \\
&\leq \frac{1}{n} \int f^n(x^n) \log \frac{f^n(x^n)}{\pi(j^*) q_{j^*}^{(n)}(x^n)} d\mu \\
&= \frac{1}{n} \log \frac{1}{\pi(j^*)} + \frac{1}{n} \int f^n(x^n) \log \frac{f^n(x^n)}{q_{j^*}^{(n)}(x^n)} d\mu \\
&= \frac{1}{n} \log \frac{1}{\pi(j^*)} + \frac{1}{n} D\left(f^n \parallel q_{j^*}^{(n)}\right).
\end{aligned}$$

The term $D\left(f^n \parallel q_{j^*}^{(n)}\right)$ can be bounded in terms of risks of original estimators. Indeed,

$$D\left(f^n \parallel q_{j^*}^{(n)}\right) = \int f^n(x^n) \log \frac{f^n(x^n)}{q_{j^*}^{(n)}(x^n)} d\mu$$

$$
\begin{aligned}
&= \int f^n(x^n) \sum_{i=0}^{n-1} \log \frac{f(x_{i+1})}{\hat{f}_{j^*,i}(x_{i+1}|x_1,...,x_i)} d\mu \\
&= \sum_{i=0}^{n-1} \int f^i(x^i) \log \frac{f(x_{i+1})}{\hat{f}_{j^*,i}(x_{i+1}|x_1,...,x_i)} d\mu \\
&= \sum_{i=0}^{n-1} E_f D(f \parallel \hat{f}_{j^*,i}) \\
&= \sum_{i=0}^{n-1} R(f, \delta_{j^*}, i).
\end{aligned}
$$

Thus we have obtained the following inequality:

$$
E_f D(f \parallel \hat{\hat{f}}_n) \leq \frac{1}{n} \log \frac{1}{\pi(j^*)} + \frac{1}{n} \sum_{i=0}^{n-1} R(f, \delta_{j^*}, i).
$$

Let $R_{cum}(f, \delta, n) = \sum_{i=0}^{n-1} R(f, \delta, i)$. It is the cumulative K-L risk of the estimation strategy δ up to $n-1$ observations. With the given prior π on the strategies δ_j, let

$$
R_n(\pi, \{\delta_j, j \geq 1\}) = \inf_{j \geq 1} \left(\log \frac{1}{\pi(j)} + R_{cum}(f, \delta_j, n) \right).
$$

Then $R_n(\pi, \{\delta_j, j \geq 1\})$ is the best trade-off between the cumulative risk and the logarithm of the inverse prior probability over the estimation strategies.

Theorem 3.2: For any given countable collection of estimation strategies $\{\delta_j, j \geq 1\}$, we can construct one estimation strategy δ^* such that for any underlying density f, the cumulative risk of δ^* up to $n-1$ observations and the usual risk of $\hat{\hat{f}}_n$ based on $n-1$ observations are upper bounded by $R_n(\pi, \{\delta_j, j \geq 1\})$ and $\frac{1}{n} R_n(\pi, \{\delta_j, j \geq 1\})$ respectively. That is,

$$
E_f D(f \parallel \hat{\hat{f}}_n) \leq \frac{1}{n} R_n(\pi, \{\delta_j, j \geq 1\}),
$$

and

$$
R_{cum}(f, \delta^*, n) \leq R_n(\pi, \{\delta_j, j \geq 1\}).
$$

From the above theorem, by mixing the existing strategies designed for different classes using a prior, we have one single estimation strategy that shares the advantages of all the proposed strategies automatically in the asymptotic sense. More precisely, the cumulative risk of the mixed strategy is bounded by the best trade-off between minus logarithm of the

prior probability (weight) of a strategy and the cumulative risk of that strategy. It follows that the cumulative risk of the mixed strategy is bounded by minus logarithm of the weight put on the best strategy for the underlying density f (the strategy minimizes $R_{cum}(f, \delta_j, n)$ over δ_j) and the cumulative risk of the best strategy for f. Thus without knowing which strategy works best for the unknown density, the price we pay in terms of the cumulative risk is at most a constant (minus logarithm of the weight). In another word, one strategy can do the job of a countable collection of strategies designed for different target densities in terms of the rates of convergence of the cumulative risks.

From the theorem, if δ_j is minimax-rate optimal for class $\mathcal{F}_j$ in terms of the cumulative risk, then the mixed strategy δ^* is minimax-rate adaptive over the classes $\mathcal{F}_j$, $j \geq 1$.

It is more complicated to obtain minimax-rate adaptive estimators over a general collection of density classes for the usual risk instead of cumulative risk. Our approach of analysis on risk of $\hat{\hat{f}}_n$ is through the results on the cumulative risk. As will be seen next, this approach does not always guarantee the minimax-rate adaptivity for the usual risk as opposed to that for the cumulative risk, yet the result is satisfactory for many nonparametric classes of densities.

Suppose δ_j, $j \geq 1$ are minimax-rate procedures for the corresponding classes under the usual risks, that is, there exist constants β_j $(j \geq 1)$ such that

$$\max_{f \in \mathcal{F}_j} R(f, \delta_j, n) \leq \beta_j \min_{\hat{f}(X_1,\ldots,X_n)} \max_{f \in \mathcal{F}_j} D(f \parallel \hat{f})$$

for all n. Denote the minimax risk $\min_{\hat{f}(X_1,\ldots,X_n)} \max_{f \in \mathcal{F}_j} D(f \parallel \hat{f})$ by $R_{mm}(j, n)$ for the classes. Then, from Theorem 3.2, for each $j \geq 1$,

$$\begin{aligned} \max_{f \in \mathcal{F}_j} D(f \parallel \hat{\hat{f}}_n) &\leq \frac{1}{n} \log \frac{1}{\pi(j)} + \frac{1}{n} \max_{f \in \mathcal{F}_j} \sum_{i=0}^{n-1} R(f, \delta_j, i) \\ &\leq \frac{1}{n} \log \frac{1}{\pi(j)} + \beta_j \frac{1}{n} \sum_{i=0}^{n-1} R_{mm}(j, i). \end{aligned}$$

Thus, the estimator $\hat{\hat{f}}_n$ is adaptive if

$$\frac{\frac{1}{n} \sum_{i=0}^{n-1} R_{mm}(j, i)}{R_{mm}(j, n)}$$

is upper bounded by a constant c_j for every class j.

Typically for smooth nonparametric classes, the minimax risks converge around some polynomial rates in the sample size. For example, if $R_{mm}(j, n) \sim n^{-r_j}$ for some $0 < r_j < 1$, then $\sum_{i=0}^{n-1} R_{mm}(j, i) \sim n^{-r_j+1}$ and $\frac{\frac{1}{n} \sum_{i=0}^{n-1} R_{mm}(j,i)}{R_{mm}(j,n)}$ is indeed bounded.

We next give some sufficient conditions for the average $\frac{1}{n}\sum_{i=0}^{n-1} R_{mm}(j,i)$ to converge at the same order of $R_{mm}(j,n)$.

Let a_n be a decreasing sequence. Suppose $a_n = n^{-(1-\alpha)}\kappa(n)$, where $0 < \alpha < 1$ and $\kappa(n)$ satisfies one of the following conditions.

Case 1. $\kappa(n)$ is increasing to ∞. An example is $\kappa(n) = (\log n)^\eta$ for some $\eta > 0$. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} a_n &= \frac{1}{n}\sum_{i=1}^{n} i^{-(1-\alpha)}\kappa(i) \\ &= \sum_{i=1}^{n}\left(\frac{i}{n}\right)^{-(1-\alpha)}\kappa(i)\cdot\frac{1}{n}\cdot n^{-(1-\alpha)} \\ &\leq n^{-(1-\alpha)}\kappa(n)\cdot\left(\sum_{i=1}^{n}\left(\frac{i}{n}\right)^{-(1-\alpha)}\cdot\frac{1}{n}\right).\end{aligned}$$

But $\sum_{i=1}^{n}\left(\frac{i}{n}\right)^{-(1-\alpha)}\cdot\frac{1}{n} \to \int_0^1 x^{-(1-\alpha)}dx = \frac{1}{\alpha}$. So $\frac{1}{n}\sum_{i=1}^{n} a_i \preceq a_n$. Together with the monotonicity of a_n, we know for this case,

$$\frac{1}{n}\sum_{i=1}^{n} a_i \asymp a_n.$$

Case 2. $\kappa(n)$ stay bounded above and away from 0. For this case, using a similar argument, it is not hard to see that $\frac{1}{n}\sum_{i=1}^{n} a_i \asymp a_n$.

Case 3. $\kappa(n) \downarrow 0$ and assume that there exists $\tau > 0$ such that $\liminf \frac{\kappa(n)}{\kappa(n^{\frac{1}{1+\tau}})} > 0$. An example for this case is $\kappa(n) = \left(\frac{1}{\log n}\right)^\eta$ for some $\eta > 0$. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} a_i &= \frac{1}{n}\sum_{i=1}^{n} i^{-(1-\alpha)}\kappa(i) \\ &= \frac{1}{n}\sum_{i=1}^{n^{\frac{1}{1+\tau}}} i^{-(1-\alpha)}\kappa(i) + \frac{1}{n}\sum_{i=n^{\frac{1}{1+\tau}}}^{n} i^{-(1-\alpha)}\kappa(i) \\ &\leq \kappa(1)\frac{1}{n}\sum_{i=1}^{n^{\frac{1}{1+\tau}}} i^{-(1-\alpha)} + \kappa(n^{\frac{1}{1+\tau}})\frac{1}{n}\sum_{i=1}^{n} i^{-(1-\alpha)}.\end{aligned}$$

In the last expression, the first term is of order $n^{-(1-\frac{\alpha}{1+\tau})} = o\left(n^{-(1-\alpha)}\right)$ and the second term is of order $n^{-(1-\alpha)}\kappa(n)$ under the assumption. Thus $\frac{1}{n}\sum_{i=1}^{n} a_i \asymp a_n$.

From the above simple calculations, if for each class $\mathcal{F}_j$, the minimax risk behaves as one of the above situations, then our mixed strategy δ^* does provide an adaptive estimator over the considered classes.

Chapter 4

Model Selection for Density Estimation

4.1 Introduction

In Chapter 3, we constructed adaptive estimators based either on metric entropies of the density classes or on existing good estimators for each of the target classes. The adaptive estimator based on metric entropies require construction of suitable packing sets in the density classes, which is very hard to do in practice. The adaptive estimator based on good estimators for each of the classes is also hard to compute, because integration is involved in one step. Practically feasible adaptive density estimators are desired.

Two types of nonparametric density estimation procedures are often used in practice. One type is fully nonparametric, where no parametric models are assumed to do statistical inference. Another type is a compromise between full nonparametric and parametric procedures, where parametric models are still used to perform statistical estimation, but the parametric models are allowed to become more and more complicated as the sample size increases. In this chapter, we consider the use of the second approach to obtain adaptive estimators based on model selection.

To estimate the unknown density function $f(x)$, a sequence of finite-dimensional density families $f_k(x, \theta^{(k)})$, $\theta^{(k)} \in \Theta_k$ are suggested to approximate the true density $f(x)$. For example, one might approximate the logarithm of the density function by a basis function expansion using polynomials, trigonometric, or spline series (for a detailed review on this

topic, see Barron and Sheu (1991)). For a given model k, we consider the maximum likelihood estimator $\hat{\theta}^{(k)}$ of $\theta^{(k)}$. Barron and Sheu (1991) (later referred as B & S) show that $n^{-\frac{2s}{2s+1}}$ is the optimal rate of convergence of $f_k(x, \hat{\theta}^{(k)})$ to $f(x)$ in the sense of relative entropy (Kullback-Leibler distance) $\int f(x) \log\left(\frac{f(x)}{f_k(x,\hat{\theta}^{(k)})}\right)$ for densities whose logarithms have s square-integrable derivatives and that this rate is achieved by suitably choosing the model size according to the smoothness parameter s. Stone (1990) obtains similar results for one dimensional log-spline models and later (1994) develops convergence rates for multi-dimensional function estimation (including density estimation) using tensor products of splines with a given order of interaction. The convergence rates are also obtained with the knowledge of the smoothness property of the target function. These results are theoretically very useful but are not applicable when the smoothness condition of the logarithm of the true density is not known in advance. In practice, with the smoothness parameters unknown, the size of the model to be used should be chosen automatically from data. The completely data-driven estimation requires a model selection criterion to compare the different models and select a suitable size one.

AIC (Akaike (1973)) is a widely used model selection criterion in many statistical applications. This criterion is suggested by Akaike from considering the asymptotic behavior of the relative entropy between the true density and the estimated one from a model. From his analysis, a bias correction term should be added to -loglikelihood as a penalty term to provide an asymptotically unbiased estimate of a certain essential part of the relative entropy loss. The familiar *AIC* takes the form

$$AIC(k) = -\log \text{likelihood} + m_k,$$

where m_k is the number of parameters in model k, and the likelihood is maximized over each family.

In addition to *AIC*, some other criteria have received a lot of attention. Schwartz (1978) proposed *BIC* based on some Bayesian analysis; Rissanen (1984) suggested the minimum description length (*MDL*) criterion from an information-theoretic point of view. Usually the *MDL* criterion takes the form

$$MDL(k) = -\log \text{likelihood} + \frac{m_k}{2} \log n.$$

The term $\frac{m_k}{2} \log n$ is the description length of the parameters with precision of order $\frac{1}{\sqrt{n}}$ for each parameter, and the likelihood is maximized over the parameters represented with this

precision (addition terms that appear in refinements of AIC and MDL are in Bernardo (1979), Clark and Barron (1990, 1994)).

The asymptotic properties of these criteria have been studied. It is shown that if the true density $f(x)$ is in one of the finite-dimensional models, then BIC chooses the correct model with probability tending to 1 (see, e.g., Haughton (1989) and Speed and Yu (1993)). For AIC, however, under the same setting, the probability of selecting a wrong model does not vanish as the sample size approaches ∞.

In a related nonparametric regression setting, an asymptotic optimality property is shown for AIC with fixed design (Shibata (1981) and Li (1987)). Li shows that if the true regression function is not in any of the finite-dimensional models, then the average squared error of the selected model is asymptotically the same as that could be achieved with the knowledge of the size of the best model to be used in advance. For the above MDL criterion, however, the average squared error of the selected model converges at a slower rate due to the presence of the $\log n$ factor in the penalty term. In a density estimation setting, Barron and Cover (1991) show that the Hellinger distance between the true density and the estimated one from MDL converges at a rate within a logarithmic factor of the optimal rate.

The MDL principle requires that the criterion retains the Kraft's inequality requirement of a uniquely decodable code. This requirement puts a restriction on the choices of candidate parameter values. For some cases, with suitable restrictions on the parameters, the MDL principle can yield a minimax optimal criterion of the form $-\log$ likelihood $+$ constant $\cdot\, m_k$, whose penalty term is of the same order as that in AIC (see Barron, Yang and Yu (1994)).

In this work, we consider comparing models using criteria related to AIC and MDL in the density estimation setting. We demonstrate that the criteria have an asymptotic optimality property for certain nonparametric classes of densities, i.e., the optimal rate of convergence for density functions in various nonparametric classes is simultaneously achieved with the automatically selected model without knowing the smooth parameters in advance.

As opposed to AIC, we allow the bias correction penalty term to be a multiple of the number of parameters in the model, and the coefficient will depend on a dimensionality constant of the finite-dimensional model related to the metric entropy. This dependency is needed when the dimensionality constants for all the models are not uniformly bounded. In this paper, the coefficients are specified so that the asymptotic results hold. With this

consideration, the criteria take the form:

$$-\sum_{i=1}^{n} \log f_k(X_i, \hat{\theta}^{(k)}) + \lambda_k m_k, \tag{4.1}$$

where $\hat{\theta}^{(k)}$ is the maximum likelihood estimator in model k and λ_k is a positive constant. Let $\hat{k}$ be the selected model which minimizes the above criterion value.

In contrast to the minimum description length criterion, we do not discretize the parameter spaces and the criteria used here do not necessarily have a total description length interpretation. In addition, the results here can be applied to more classes of densities than that considered by Barron, Yang and Yu (1994). We should also note that our criteria are not necessarily Bayesian.

We evaluate the criteria by comparing the Hellinger distance $d_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) = \int(\sqrt{f} - \sqrt{f_{\hat{k},\hat{\theta}^{(\hat{k})}}})^2 d\mu$ with an index of resolvability. The concept of resolvability was introduced by Barron and Cover (1991). It naturally captures the capability of estimating a function by a sequence of models. The index of resolvability can be defined as

$$R_n(f) = \inf_k \{ \inf_{\theta^{(k)} \in \Theta_k} D(f \| f_{k,\theta^{(k)}}) + \frac{\lambda_k m_k}{n} \}.$$

The first term $\inf_{\theta^{(k)} \in \Theta_k} D(f \| f_{k,\theta^{(k)}})$ reflects the approximation capability of the model k to the true density function in the sense of relative entropy distance, and the second term $\frac{\lambda_k m_k}{n}$ reflects the variation of the estimator in the model due to the estimation of the best parameters in the model. The index of resolvability quantifies the best trade-off between the approximation error and the estimation error. It is shown in this work that with the use of the criterion, when the λ_k's are chosen large enough, the statistical risk $E d_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}})$ is bounded by a multiple of $R_n(f)$. Clearly, the resolvability gets the best rate with smallest allowable λ_k's, say λ_k^*, $k \in \Gamma$. The cases studied in this paper follow one of the two forms: 1. λ_k^*'s are constants independent of k (then it is like AIC but with a constant possibly different from 1); 2. $\lambda_k^* \leq$constant$\cdot \log m_k$ (then it is like BIC but with a constant possibly different from $\frac{1}{2}$).

To apply the above results, we can evaluate $R_n(f)$ for f in various nonparametric classes of functions, then an upper bound of the convergence rate can be easily obtained. Examples will be given to show these bounds correspond to optimal or near optimal rates of convergence for density functions in various nonparametric classes.

In statistical applications, models are needed to perform sensible analysis and draw useful conclusions. Usually, the models are suggested based on previous experience or

intuition on what class of functions the unknown function might be in. Due to the lack of such knowledge on the true function, it is often more flexible to consider more than one class of functions to do statistical analysis. For example, if we use series expansion method to estimate the logarithm of density, we might consider polynomial models, trigonometric models, and spline models at the same time and wish to choose whatever is the best in term of the statistical risk. For spline models, we might consider different orders and different numbers of knots. Even if we consider only one type of series expansion, it might be advantageous to consider sparse subset models if the true function is sparse in the sense that only a small fraction of the basis functions are useful for good approximation. Such an advantage will be demonstrated in Section 4. In high dimensional function estimation, the complete models which use all the basis functions up to certain orders often fail due to the "curse of dimensionality". On the other hand, the sparse subset models such as additive models and low order interaction models might yield reasonable estimates. Considering many different classes of functions or sparse models may result in exponentially many or even more models. When exponentially many models are considered, significant selection bias might occur with the bias-correction based criteria like AIC and the criteria we just proposed. The reason is that the criterion value can not estimate the targeted quantity (e.g., the relative entropy loss of the density estimator in each model) uniformly well for exponentially many models. For such cases, the previously obtained results for the selection among polynomially many models can not be applied any more. For example, for the nonparametric regression function estimation with fixed design, a condition for Li's results is no longer satisfied. To handle the selection bias in that regression setting, a model complexity based on an information-theoretical consideration is incorporated to AIC and a new criterion named ABC is suggested (Yang (1993)). There it is shown that ABC provides the best trade-off among the approximation error, the estimation error and the model complexity.

For the density estimation problem, we also take the model complexity into consideration to handle the possible selection bias when exponentially many or more models are presented for more flexibility. For each model, a complexity C_k is assigned with $L_k = (\log_2 e)C_k$ satisfying the Kraft's inequality: $\sum_k 2^{-L_k} \leq 1$; that is

$$\sum_k e^{-C_k} \leq 1.$$

The complexity $L_k = C_k \log_2 e$ can be interpreted as the a codelength of a uniquely decodable code to describe the models. Another interpretation is that e^{-C_k} is a prior probability of model k. Then the criteria we propose are

$$-\sum_{i=1}^{n} \log f_k(X_i, \hat{\theta}^{(k)}) + \lambda_k m_k + \nu C_k, \tag{4.2}$$

where ν is a nonnegative constant.

For the above more general criteria, we redefine the index of resolvability by adding the complexity term as follows:

$$R_n(f) = \inf_k \{ \inf_{\theta^{(k)} \in \Theta_k} D(f \| f_{k,\theta^{(k)}}) + \frac{\lambda_k m_k}{n} + \frac{\nu C_k}{n} \}. \tag{4.3}$$

It provides the best trade-off among the approximation error, estimation error, and the model complexity relative to sample size. We show

$$E d_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) = O(R_n(f)).$$

As an example, we will consider estimating a density function on [0,1]. We assume that the logarithm of the density is in the union of the classes of Sobolev space $W_2^s(U), s \in N$, $U > 0$. We approximate the logarithm of the density by spline functions. If we knew U and s , then by using suitably pre-determined order splines, the optimal rate of convergence is achieved. However, this rate of convergence of $n^{-\frac{2s}{2s+1}}$ is saturated for smoother densities. Without knowing U and s, we might consider all the spline models with different smoothness orders and let the criterion choose a suitable one automatically from data. Indeed, from our theorem, the optimal rate of convergence is obtained simultaneously for density functions with logarithms in the classes $W_2^s(U)$, $s \in N$, $U > 0$. In another word, the density estimator based on the model selection adapts to every class $W_2^s(U)$, $s \in N$, $U > 0$.

The above examples suggest that good model selection criteria can provide us with minimax optimal function estimation strategies simultaneously for many different classes. As some other applications of our results, neural network models and sparse density function estimation will be considered.

This chapter is organized as follows: in Section 2, we present a key lemma for the main result; in Section 3, we state and prove the main theorem; in Section 4, we provide some applications of the main results; in Section 5, we give the proofs of the key lemma and some other lemmas; and finally in Section 6, we prove several useful inequalities.

4.2 A key lemma

Let f be the true density function, and $f(x,\theta), \theta \in \Theta$ be a parametric family of densities. For $r > 0$, let $B_\Theta(f,r)$ be a Hellinger "ball" in Θ around f (f may not be in the parametric family) with radius r defined by $B_\Theta(f,r) = \{\theta : \theta \in \Theta, d_H^2(f, f_\theta) \leq r^2\}$.

Let P^* denote the outer measure of probability measure P on some measurable space (Ω, G) where $X_1, ..., X_n$ are defined. Outer measure is used later for possibly non-measurable sets of interests.

Our asymptotic results rely on an exponential inequality to control the probability of selecting a bad model. The inequality requires a dimensionality assumption on the parametric family. This type of assumptions were previously used by Le Cam (1973), Birgé (1983) and others.

In our analysis, we will consider sup-norm distance between the logarithms of densities. In this chapter, unless stated otherwise, by a δ-net, we mean a δ-net in the sense of sup-norm requirement for the logarithms of the densities. That is, for a class of densities B, we say a finite collection of densities F_δ is a δ-net if for any density $f \in B$, there exists $\tilde{f} \in F_\delta$ such that $\| \log \tilde{f} - \log f \|_\infty \leq \delta$. For convenience, the index set of F_δ might also be called a δ-net.

Assumption 4.0: For a fixed density f, there exist constants $A > 0$, $m \geq 1$ and $\rho > 0$ with $\rho \leq A$ (A, m, ρ are allowed to depend on f) such that for any $r > 0$ and $\delta \leq \rho r$, there exists a δ-net F_δ for $B_\Theta(f,r)$ satisfying the following requirement:

$$\text{card}(F_\delta) \leq \left(\frac{Ar}{\delta}\right)^m .$$

Remark: This dimensionality assumption necessarily requires that the densities in the parametric family share the same support. If the support of the true density is not known to us, we might consider families of densities with different supports and let the model selection criterion decide which one has a suitable support for the best estimation of the unknown density.

Lemma 4.0: Assume Assumption 4.0 is satisfied with $\rho \geq \frac{0.13\gamma}{\sqrt{1-4\gamma}}$ for some $0 < \gamma < \frac{1}{4}$. If

$\frac{\xi}{m} \geq \frac{4}{1-4\gamma} \log\left(\frac{15.4A\sqrt{1-4\gamma}}{\gamma}\right)$, then

$$P^*\{\text{for some } \theta \in \Theta, \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i)} \geq -\gamma d_H^2(f, f_\theta) + \frac{\xi}{n}\} \leq 15.1 \exp(-\frac{(1-4\gamma)\xi}{8}).$$

Remark: From the proof of Lemma 4.0, it is seen that the requirement in Assumption 4.0 needs only to be checked for $r \geq \sqrt{\frac{4\log 2}{1-4\gamma}\frac{m}{n}}$.

The proof of Lemma 4.0 is given in section 5.

4.3 Main results

We consider a list of parametric families of densities $f_k(x, \theta^{(k)}), \theta^{(k)} \in \Theta_k, k \in \Gamma$, where Γ is the collection of the indices of the models. The model list is assumed to be fixed and independent of sample size unless otherwise stated (e.g., in Subsection 4.4.2). Lemma 4.0 will be used to derive the main theorem with the choice of $\gamma = 0.039$ to have small penalty constants λ_k's (see Theorem 4.1). The corresponding value of ρ is 0.0056. We use this value in the following assumption.

Assumption 4.1: For a fixed density f, for each $k \in \Gamma$, Assumption 4.0 is satisfied with some constants A_k, m_k and $\rho \geq 0.0056$.

Assumption 4.1 may look hard to be checked because of the presence of the unknown function f as the center of the balls $B_{\Theta_k}(f, r)$, but actually this condition can be replaced by a condition involving only the operating families in Γ.

For $\theta_0^{(k)} \in \Theta_k$, consider Hellinger balls centered at density $f_{k,\theta_0^{(k)}}$ (instead of the true density) in family k defined by

$$B_k(\theta_0^{(k)}, r) = \{\theta^{(k)} : \theta^{(k)} \in \Theta_k, d_H^2(f_{k,\theta_0^{(k)}}, f_{k,\theta^{(k)}}) \leq r^2\}.$$

Assumption 4.1': For each $k \in \Gamma$, and $\theta_0^{(k)} \in \Theta_k$, Assumption 4.0 is satisfied for density $f_{k,\theta_0^{(k)}}$ with $B_k(\theta_0^{(k)}, r)$ in place of $B_{\Theta_k}(f, r)$ and with $\rho \geq 0.0056$ and constants $\tilde{A}_k, m_k$ not depending on $\theta_0^{(k)}$.

Lemma 4.1: If Assumption 4.1′ is satisfied with $\tilde{A}_k, m_k$ and ρ, then Assumption 4.1 is satisfied for any density f with $A_k = 3\tilde{A}_k$, m_k and ρ.

Proof : Fix any $\epsilon > 0$. Let $\theta_*^{(k)} \in \Theta_k$ satisfy $d_H(f, f_{k,\theta_*^{(k)}}) \leq \inf_{\theta^{(k)} \in \Theta_k} d_H(f, f_{k,\theta^{(k)}}) + \epsilon r$. Then because

$$\begin{aligned} d_H(f, f_{k,\theta^{(k)}}) &\geq \tfrac{1}{2}(d_H(f, f_{k,\theta_*^{(k)}}) + d_H(f, f_{k,\theta^{(k)}})) - \tfrac{\epsilon r}{2} \\ &\geq \tfrac{1}{2} d_H(f_{k,\theta^{(k)}}, f_{k,\theta_*^{(k)}}) - \tfrac{\epsilon r}{2}, \end{aligned}$$

we have

$$\begin{aligned} B_{\Theta_k}(f, r) &= \{\theta^{(k)} : \theta^{(k)} \in \Theta_k, d_H(f, f_{k,\theta^{(k)}}) \leq r\} \\ &\subset \{\theta^{(k)} : \theta^{(k)} \in \Theta_k, d_H(f_{k,\theta_*^{(k)}}, f_{k,\theta^{(k)}}) \leq (2+\epsilon) r\} \\ &= B_k(\theta_*^{(k)}, (2+\epsilon) r). \end{aligned}$$

Thus if Assumption 4.1′ is satisfied with $\tilde{A}_k, m_k$ and ρ, then Assumption 4.1 is satisfied with $A_k = (2+\epsilon)\tilde{A}_k$, m_k and ρ for any $\epsilon > 0$. For the statement of the Lemma, ϵ is taken to be 1. We note that if $\inf_{\theta^{(k)} \in \Theta_k} d_H^2(f, f_{k,\theta^{(k)}})$ is achievable for all $k \in \Gamma$, then we may set $\epsilon = 0$ and Assumption 4.1 is satisfied with $A_k = 2\tilde{A}_k$.

Let

$$V(k, \theta^{(k)}) = -\frac{1}{n} \sum_{i=1}^{n} \log f_k(X_i, \theta^{(k)}) + \frac{\lambda_k m_k}{n} + \frac{\nu C_k}{n},$$

where λ_k, ν are nonnegative numbers. Then the model selection criterion we consider is to choose $\hat{k}$ to minimize

$$\text{crit}(k) = V(k, \hat{\theta}^{(k)}), \tag{4.4}$$

where $\hat{\theta}^{(k)}$ is the maximum likelihood estimator in model k. The final density estimator $\hat{f}$ is $\hat{f} = f_{\hat{k},\hat{\theta}^{(\hat{k})}}$, i.e., the maximum likelihood density estimator in the selected model.

Let

$$R_n(k, \theta^{(k)}) = D(f \| f_{k,\theta^{(k)}}) + \frac{\lambda_k m_k}{n} + \frac{\nu C_k}{n}.$$

Then the index of resolvability is

$$R_n(f) = \inf_{k \in \Gamma, \theta^{(k)} \in \Theta_k} R_n(k, \theta^{(k)}).$$

As mentioned before, in a rough sense, $R_n(f)$ characterizes the best trade-off among three sources of discrepancy: approximation error, estimation error, and model complexity.

The asymptotic results we present requires suitable choice of the penalty constants λ_k (according to the cardinality constants A_k) and ν. Let

$$\Lambda(A) = 4.75 \log A + 27.93, \tag{4.5}$$

$$\nu^* = 9.49.$$

Theorem 4.1: Assume Assumption 4.1 is satisfied. Take $\lambda_k \geq \lambda_k^* = \Lambda(A_k)$ and $\nu \geq \nu^*$ in the model selection criterion given in (4.2). Then for the density estimator $f_{\hat{k},\hat{\theta}(\hat{k})}$, we have

$$E d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) \leq 2657 R_n(f),$$

where the resolvability $R_n(f)$ is defined as follows

$$R_n(f) = \inf_{k \in \Gamma} \{ \inf_{\theta^{(k)} \in \Theta_k} D(f \| f_{k,\theta^{(k)}}) + \frac{\lambda_k m_k}{n} + \frac{\nu C_k}{n} \}. \tag{4.6}$$

In general, if Assumption 4.1 holds with $\rho \geq \frac{0.13\gamma}{\sqrt{1-4\gamma}}$ for some $0 < \gamma < \frac{1}{4}$ as in Lemma 4.0, then for $\lambda_k \geq \frac{4}{1-4\gamma} \log \left(\frac{15.4 A_k \sqrt{1-4\gamma}}{\gamma} \right)$ and $\nu \geq \frac{8}{1-4\gamma}$,

$$E d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) \leq \frac{1}{\gamma} \left(\frac{85.6}{1-4\gamma} + 2.2 \right) R_n(f).$$

The choice $\gamma = 0.039$ minimizes $\frac{4}{1-4\gamma} \log \left(\frac{15.4 A_k \sqrt{1-4\gamma}}{\gamma} \right)$ at $A_k = 1$.

Corollary 4.1: Under the above conditions,

$$E \| f - f_{\hat{k},\hat{\theta}(\hat{k})} \|_{L_1} \leq 104 \sqrt{R_n(f)}.$$

Corollary 4.1 follows from Theorem 4.1 using the familiar relationship between the Hellinger distance and L_1 distance, namely, $\| f - g \|_{L_1} \leq 2 d_H(f, g)$ for densities f and g.

Corollary 4.2: Under the same conditions above, we have convergence in probability of $d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})})$ at rate $R_n(f)$, that is,

$$d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) = O_p(R_n(f)).$$

Remarks:

1. The resolvability bound in the theorem is valid for any sample size. So the model list Γ is allowed to change according to sample size.

2. In $R_n(f)$, the estimation error term $\frac{\lambda_k m_k}{n}$ is allowed to depend on the dimensionality constant A_k, which may not be uniformly bounded for all $k \in \Gamma$. For an unknown density function in a class, if the sequence of models k_n minimizing $\inf_{\theta^{(k)} \in \Theta_k} D(f \parallel f_{k,\theta^{(k)}}) + \frac{m_k}{n}$ have A_{k_n} bounded, then $R_n(f)$ is asymptotically comparable to

$$\inf_{k,\theta^{(k)}} \{D(f \parallel f_{k,\theta^{(k)}}) + \frac{m_k}{n} + \frac{C_k}{n}\}.$$

If furthermore, $C_{k_n} = O(m_{k_n})$, then

$$Ed_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) = O(\inf_{k,\theta^{(k)}} \{D(f \parallel f_{k,\theta}) + \frac{m_k}{n}\}),$$

which often gives the minimax optimal rate of convergence for density functions in many smooth nonparametric density classes. These conditions that A_{k_n} is bounded and $C_{k_n} = O(m_{k_n})$ will be verified in a spline estimation setting in Section 4.

3. For the case when m_k's are integers for $k \in \Gamma$, one way to assign the complexities for the models is by considering only the number of models for each dimension. Let $N(m) = \text{card}\{k \in \Gamma : m_k = m\}$ be the number of models with dimension m. If $N(m) < \infty$, then we may assign complexity $C_k = \log N(m) + 2\log(m+1)$ for the models with dimension m, which corresponds to the strategy of describing m first and then specifying the model among all the models with the same dimension m. Then we have

$$\begin{aligned} Ed_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) \leq \; & 2657 \inf_{k \in \Gamma} \{\inf_{\theta^{(k)} \in \Theta_k} D(f \| f_{k,\theta^{(k)}}) \\ & + (\frac{\lambda_k m_k}{n} + \frac{\nu_k(\log N(m_k) + 2\log(m_k+1))}{n})\}. \end{aligned}$$

If N_m grows slowerly than exponential in m, then $\frac{\log N(m_k)+2\log(m_k+1)}{m_k}$ goes to 0, i.e., the complexity is essentially negligible compared to the model dimension. Then the complexity part of the penalty term can be ignored in the model selection criteria. However, if there are exponentially many or more models in Γ, then the complexity term $\frac{C_k}{n}$ is not negligible compared to $\frac{m_k}{n}$ (for related discussions, see Yang (1993) and Birgé and Massart (1995)).

Proof of Theorem 4.1: Clearly the working criterion is theoretically equivalent to select $\hat{k}$ and $\hat{\theta}^{(\hat{k})}$ to minimize

$$\tilde{V}(k,\theta^{(k)}) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i)}{f_k(X_i,\theta^{(k)})} + \frac{\lambda_k m_k}{n} + \frac{\nu C_k}{n} + \frac{t}{n}$$

by adding $\frac{1}{n}\sum_{i=1}^{n}\log f(X_i) + \frac{t}{n}$ (which does not depend on k) to each criterion value. In our analysis, we will concentrate on the above theoretically equivalent criterion. We relate it to the resolvability. Indeed, for each fixed family k, we show that $\tilde{V}(k,\theta^{(k)})) \geq \gamma d_H^2(f, f_{k,\theta^{(k)}})$ for all θ except in a set of small probability. Then the probability bound is summed over k to obtain a corresponding bound uniformly over all the models.

For a fixed k, let

$$L_n(k,\theta^{(k)}) = \sum_{i=1}^{n}\log\frac{f(X_i)}{f_k(X_i,\theta^{(k)})}\ .$$

Then

$$\begin{aligned} P^*\{ \quad & \text{for some } \theta^{(k)}\in\Theta_k, \quad \tilde{V}(k,\theta^{(k)})) \leq \gamma d_H^2(f, f_{k,\theta^{(k)}})\} \\ &= P^*\{\text{for some } \theta^{(k)}\in\Theta_k, \quad \tfrac{1}{n}L_n(k,\theta^{(k)}) \leq -\tfrac{\lambda_k m_k}{n} \\ &\qquad -\tfrac{\nu C_k}{n} - \tfrac{t}{n} + \gamma d_H^2(f, f_{k,\theta^{(k)}})\} \\ &= P^*\{\text{for some } \theta^{(k)}\in\Theta_k, \quad -\tfrac{1}{n}L_n(k,\theta^{(k)}) \geq \tfrac{\lambda_k m_k}{n} \\ &\qquad +\tfrac{\nu C_k}{n} + \tfrac{t}{n} - \gamma d_H^2(f, f_{k,\theta^{(k)}})\}. \end{aligned}$$

If $\lambda_k \geq \frac{4}{1-4\gamma}\log\left(\frac{15.4A_k\sqrt{1-4\gamma}}{\gamma}\right)$, then by Lemma 4.0 with $\xi = \lambda_k m_k + \nu C_k + t$,

$$\begin{aligned} &P^*\{\text{for some } \theta\in\Theta_k, \quad \tilde{V}(k,\theta^{(k)}) \leq \gamma d_H^2(f, f_{k,\theta^{(k)}})\} \\ &\leq 15.1\exp\left(-\tfrac{1-4\gamma}{8}(\lambda_k m_k + \nu C_k + t)\right). \end{aligned}$$

Now sum over $k\in\Gamma$,

$$\begin{aligned} q_n(t) \quad =: \quad & P^*\{\text{for some } k\in\Gamma, \quad \theta^{(k)}\in\Theta_k, \tilde{V}(k,\theta^{(k)}) \leq \gamma d_H^2(f, f_{k,\theta^{(k)}})\} \\ \leq \quad & 15.1\textstyle\sum_{k\in\Gamma}\exp\left(-\tfrac{1-4\gamma}{8}(\lambda_k m_k + \nu C_k + t)\right) \\ \leq \quad & 10.7\textstyle\sum_{k\in\Gamma}\exp\left(-\tfrac{(1-4\gamma)t}{8} - C_k\right) \\ \leq \quad & 10.7\exp\left(-\tfrac{(1-4\gamma)t}{8}\right). \end{aligned}$$

For the second inequality above, we use $\frac{(1-4\gamma)\lambda_k m_k}{4} \geq \log 2$ and $\nu \geq \frac{8}{1-4\gamma}$. For the last inequality, we use $\sum_{k\in\Gamma} e^{-C_k} \leq 1$. For expectation bounds, it will be helpful to bound the integral of the tail probability $q_n(t)$. From above,

$$\int_0^\infty q_n(t)dt \leq \tfrac{85.6}{(1-4\gamma)}\ .$$

The above bound suggests it is unlikely that the criterion values are much smaller than $d_H^2(f, f_{k,\theta^{(k)}})$ for any $k, \theta^{(k)}$ and the integral of tail probability is bounded. To obtain the

conclusion of the theorem, we next show that the criterion values for a sequence of nearly best choices of $k, \theta^{(k)}$ are not much greater than $R_n(f)$.

Assume $R_n(f) < \infty$ (otherwise the conclusion of the theorem is trivially true). Let $(k_n, \theta_n^{(k_n)})$ be a choice such that $R_n(k_n, \theta_n^{(k_n)}) \leq (1+\epsilon)R_n(f)$ for some positive constant ϵ. (If there is a minimizer of $R_n(k, \theta^{(k)})$, then we may set k_n, $\theta_n^{(k_n)}$ to achieve the minimum.) For simplicity, denote $L_n(k_n, \theta^{(k_n)})$ by L_n. Then for $t \geq t_0 = \frac{4\log 2}{2\log 2 - 1 + 4\gamma}$, we have

$$\begin{aligned} p_n(t) &=: P\{\tilde{V}(k_n, \theta^{(k_n)}) \geq tR_n(k_n, \theta^{(k_n)})\} \\ &\leq P\{L_n \geq n[tD(f \parallel f_{k_n,\theta^{(k_n)}}) + \tfrac{(t-1)\lambda_{k_n} m_{k_n}}{n} + \tfrac{(t-1)\nu C_{k_n}}{n} - \tfrac{t}{n}]\} \\ &\leq P\{L_n \geq \tfrac{nt}{2}(D(f \parallel f_{k_n,\theta^{(k_n)}}) + \tfrac{\lambda_{k_n} m_{k_n}}{n} + \tfrac{\nu C_{k_n}}{n})\} \\ &= P\{L_n \geq \tfrac{nt}{2} R_n(k_n, \theta^{(k_n)})\}. \end{aligned}$$

For the last inequality above, we use the fact $\lambda_k m_k \geq \lambda_k^* m_k \geq \frac{4\log 2}{1-4\gamma}$ for all $k \in \Gamma$. Note also $\frac{n}{2} R_n(k_n, \theta^{(k_n)}) \geq \frac{\lambda_{k_n} m_{k_n}}{2} \geq \frac{2\log 2}{1-4\gamma}$. Let $\tilde{L}_n = L_n I_{\{L_n \geq \frac{t_0 n}{2} R_n(k_n, \theta^{(k_n)})\}}$ and $S_t = \{L_n \geq \frac{nt}{2} R_n(k_n, \theta^{(k_n)})\}$. Then for $t \geq t_0$, $S_t = \{\tilde{L}_n \geq \frac{nt}{2} R_n(k_n, \theta^{(k_n)})\}$. Now,

$$\begin{aligned} \textstyle\int_{t_0}^{\infty} p_n(t)dt &\leq \textstyle\int_{t_0}^{\infty} E I_{S_t} dt \\ &= E(\textstyle\int_{t_0}^{\infty} I_{S_t} dt) \\ &= E\frac{\tilde{L}_n}{\frac{n}{2} R_n(k_n, \theta^{(k_n)})} - t_0 \\ &= \frac{1}{\frac{n}{2} R_n(k_n, \theta^{(k_n)})} \textstyle\int_{\{L_n \geq t_0 n R_n(k_n, \theta^{(k_n)})\}} L_n \cdot \Pi_{i=1}^n f(x_i) d\mu - t_0. \end{aligned}$$

Here

$$t_0 n R_n(k_n, \theta^{(k_n)}) \geq \frac{4\log 2}{2\log 2 - 1 + 4\gamma} \cdot \frac{4\log 2}{1-4\gamma} \geq \frac{(4\log 2)^2}{(\log 2)^2} = 16.$$

To bound the last integral involving the tail of an expected log-likelihood ratio, we apply Lemma 4.4 in Section 6 with $\alpha^* = \alpha(e^{16}) = 1.07$ and obtain

$$\begin{aligned} \textstyle\int_{t_0}^{\infty} p_n(t)dt &\leq \frac{\alpha^* D(f^n \| f^n_{k_n,\theta^{(k_n)}})}{\frac{n}{2} R_n(k_n, \theta^{(k_n)})} - t_0 \\ &= \frac{2\alpha^* D(f \| f_{k_n,\theta^{(k_n)}})}{D(f \| f_{k_n,\theta^{(k_n)}}) + \frac{\lambda_{k_n} m_{k_n}}{n} + \frac{\nu C_{k_n}}{n}} - t_0 \\ &\leq 2\alpha^* - t_0. \end{aligned}$$

Now, from the analysis above,

$$\gamma d_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) \leq \tilde{V}(\hat{k}, \hat{\theta}^{(\hat{k})}) \leq \tilde{V}(k_n, \theta_n^{(k_n)}) \leq tR_n(k_n, \theta^{(k_n)}) \leq (1+\epsilon)tR_n(f)$$

with exception probability no bigger than $q_n(t) + p_n(t)$. That is,

$$P\{\frac{d_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}})}{\gamma^{-1}(1+\epsilon)R_n(f)} \geq t\} \leq q_n(t) + p_n(t).$$

Let $Z = \frac{d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})})}{\gamma^{-1}(1+\epsilon)R_n(f)}$, then

$$\begin{aligned} EZ &= \int_0^\infty P\{Z \geq t\}dt \\ &\leq \int_0^\infty q_n(t)dt + \int_0^{t_0} p_n(t)dt + \int_{t_0}^\infty p_n(t)dt \\ &\leq \frac{85.6}{1-4\gamma} + t_0 + 2\alpha^* - t_0. \\ &= \frac{85.6}{1-4\gamma} + 2\alpha^*. \end{aligned}$$

Because $\epsilon > 0$ is arbitrary, by letting $\epsilon \to 0$, we conclude that

$$Ed_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) \leq \frac{1}{\gamma}\left(\frac{85.6}{1-4\gamma} + 2.2\right) R_n(f).$$

This completes the proof of Theorem 4.1.

Remark: In the proof of the theorem, for the (nearly) best models k_n, we just use the fact that $D(f \parallel f_{k_n,\theta(k_n)})$ is finite. For many cases, $\| \log \frac{f}{f_{k_n,\theta(k_n)}} \|_\infty$ is bounded. Then we can use Hoeffding's inequality to obtain exponential bound on the tail probability for these models. Then we can show that $d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})})$ is bounded by $R_n(f)$ in all moments, i.e.,

$$Ed_H^{2j}(f, f_{\hat{k},\hat{\theta}(\hat{k})}) = O(R_n^j(f))$$

for all $j > 0$.

The criteria in (4.2) can yield a criterion very similar to the familiar MDL criterion when applied to a sequence of candidate densities. Suppose we have a countable collection of densities $q \in \Gamma_n$. The description lengths of the indices are $L(q)$ satisfying the Kraft's inequality: $\sum_{q\in\Gamma_n} e^{-L(q)} \leq 1$. Treat each density in Γ_n as a model, then Assumption 4.0 is satisfied with $A_k = 1, m_k = 1$, and $\rho = 1$. Thus $\lambda_k^* m_k = \frac{1}{1-4\gamma}\log 2$ is a constant independent of k. Therefore when taking $\nu = \nu^*$, the criterion in (4.2) is equivalent to minimizing

$$-\sum_{i=1}^n \log f_k(X_i, \hat{\theta}^{(k)}) + \nu^* L(q)$$

over $q \in \Gamma_n$. This criterion is different from the MDL criterion only in that $\nu^* \neq 1$. The corresponding resolvability given in our expression (4.6) is essentially the same as the resolvability $\inf_{q\in\Gamma_n}\{D(f \parallel q) + \frac{L(q)}{n}\}$ considered by Barron and Cover (1991).

4.4 Applications

4.4.1 Sequences of exponential families

As an application of the theorem we develop in Section 3, we consider estimating an unknown density by sequences of exponential models. The log-density is modeled by sequences of finite dimensional linear spaces of functions.

Localized basis

Let $S_j, j \in J$ (J is an index set) be a linear function space on $[0,1]^d$. Assume for each $j \in J$, there is a basis $\varphi_{j,1}(x), \varphi_{j,2}(x), ..., \varphi_{j,m_j}(x)$ for S_j satisfying the following two conditions with constants T_1 and T_2 not depending on j :

$$\| \sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) \|_\infty \leq T_1 \max_i |\theta_i|, \tag{4.7}$$

$$\| \sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) \|_2 \geq \frac{T_2}{\sqrt{m_j}} \| \theta \|_2 . \tag{4.8}$$

Here $\| \|_\infty$and $\| \|_2$ denote the sup-norm and L_2-norm respectively. The first condition is satisfied with localized basis. The second one is part of the requirement that $\varphi_{j,1}(x), \varphi_{j,2}(x), ..., \varphi_{j,m_j}(x)$ forms a frame (see, e.g., Chui (1991), Chapter 3) (the other half of the frame property can be used to bound the approximation error). It is assumed that $1 \in S_j$.

For each S_j, consider the following family of densities with respect to Lebesgue measure μ:

$$f_j(x,\theta) = \exp(\sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) - \psi_j(\theta)),$$

where $\psi_j(\theta) = \log \int \exp(\sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x)) d\mu$ is the normalizing constant. If there is no restriction on the parameters $(\theta_1, ..., \theta_{m_j})$, the above parametrization is not identifiable. Since the interest is on the risk of density estimation instead of parameter estimation, identifiability is not an issue here. The model selection criterion will be used to choose an appropriate model.

To apply the results in Section 3, the models need to satisfy the cardinality assumption. For that purpose, we can not directly use the nature parameter space R^{m_j}. Instead, we consider a sequence of compact parameter spaces

$$\Theta_{j,L} = \{\theta \in R^{m_j} : \| \log f_j(\cdot, \theta) \|_\infty \leq L\},$$

where L takes positive integer values. We treat each choice of $\Theta_{j,L}$ as a model. The following lemma gives upper bounds on the cardinality constants $A_{(j,L)}$.

Lemma 4.2: There exists a constant $A(L,T_1,T_2) = 19.28\frac{T_1}{T_2}(L+1)e^{\frac{L}{2}} + 0.06$ such that Assumption 4.1′ is satisfied with $A_{(j,L)} = A(L,T_1,T_2)$ and $m_{(j,L)} = m_j$.

Note in Lemma 4.2, $A(L,T_1,T_2)$ does not depend on the number of parameters m_j in the models. So $A_{(j,L)}$ remain bounded for any fixed L. The proof of Lemma 4.2 is provided in Section 5.

In practice, we might consider many different "types" of localized basis which satisfy (4.7) and (4.8) for each type of basis. For example, different order splines are useful when the smoothness condition of the true function is unknown. For such cases, the constants $T_{q,1}$ and $T_{q,2}$ may not be bounded for all considered type q's, which leads to the unboundedness of $\lambda^*_{q,(j,L)}$. It is hoped that through the use of the model selection criterion, good values of q, j, and L will be chosen with corresponding penalty constants λ^*'s being bounded so that the optimal rate of convergence could be achieved.

Assume for each q in an index set Q, we have a collection of models J_q satisfying the conditions (4.7) and (4.8) with $T_{q,1}$ and $T_{q,2}$. Let $k = (j,q,L)$ be the index of the models $f_j(x,\theta)$, $\theta \in \Theta_{j,L}$, $j \in J_q$, $q \in Q$ and let Γ be the collection of the indices k. Let C_k, $k \in \Gamma$ be a complexity assigned for the models in Γ satisfying $\sum_{k\in\Gamma} e^{-C_k} \le 1$.

Let $\lambda^*(q,L) = \Lambda(2A(L,T_{q,1},T_{q,2}))$. Let $\hat{k}$ be the model minimizing

$$-\sum_{i=1}^{n} \log f_k(X_i,\hat{\theta}^{(k)}) + \lambda^*(q,L)m_j + 9.49C_k. \tag{4.9}$$

Then from Theorem 4.1, we have the following conclusion.

Corollary 4.3: For localized basis models for the log-density satisfying conditions (4.7) and (4.8), for any underline density f,

$$Ed_H^2(f, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) \le 2657R_n(f),$$

where

$$R_n(f) = \inf_{L\ge 1}\inf_{q\in Q}\inf_{j\in J_q}\{\inf_{\theta\in\Theta_{j,q,L}} D(f \parallel f_{j,\theta}) + \frac{\lambda^*(q,L)m_j}{n} + \frac{9.49C_k}{n}\}.$$

Corollary 4.3 can yield minimax optimal rates of convergence simultaneously for many nonparametric classes of densities when the sup-norms of the log-densities in each class are uniformly bounded (with the bound possibly unknown) and the log-densities in each class can be "well" approximated by the models $f_{(j,q),\theta}$, $j \in J_q$ for some fixed q. For such a class of densities, when L is sufficiently large, a sequence of densities in $\Theta_{j_n,q^*,L}$ for some j_n and a fixed q^* achieves the resolvability. With these L and q^*, the penalty constants λ_k^* are bounded for the particular sequence of densities. Suitable assignment of the complexities might give us $C_k = O(m_k)$, then $R_n(f) = O\left(\inf_{j\in J_{q^*}}\{\inf_{\theta\in\Theta_{j,q^*,L}} D(f \parallel f_{j,\theta}) + \frac{m_j}{n}\}\right)$, which usually gives the minimax optimal rate of convergence for the density in the class.

Example 4.1: Univariate Log-spline models.

Let $S_{m,q}$ $(m \geq q)$ be the linear function space of splines of order q (piecewise polynomial of order less than q) with $m-q+2$ equally spaced knots. Let $\varphi_{m,q,1}(x), \varphi_{m,q,2}(x), ..., \varphi_{m,q,m}(x)$ be the B-spline basis. Let

$$f_{m,q}(x,\theta) = \exp(\sum_{i=1}^{m} \theta_i \varphi_{m,q,i}(x) - \psi_{m,q}(\theta)),$$

where $\psi_{m,q}(\theta) = \log \int \exp(\sum_{i=1}^{m} \theta_i \varphi_{m,q,i}(x)) d\mu$. To make the family identifiable, we assume $\sum_{i=1}^{m} \theta_i = 0$. The model selection criterion will be used to choose appropriate number of knots and spline order q.

Consider

$$\Theta_{m,q,L} = \{\theta \in R^m : \parallel \log f_{m,q}(\cdot,\theta) \parallel_\infty \leq L\},$$

where $L \geq 1, q \geq 1, m \geq q$ are integers. Each parameter space $\Theta_{m,q,L}$ corresponds to a model.

The B-spline basis is known to satisfy the two conditions (4.7) and (4.8). In fact, the sup-norm of spline expressed by B-splines is bounded by the sup-norm of the coefficients (see, de Boor (1978, pp. 155)), that is,

$$\parallel \sum_{i=1}^{m} \theta_i \varphi_{m,q,i}(x) \parallel_\infty \leq \max_{1\leq i\leq m} |\theta_i|.$$

The second requirement follows from the frame property of the B-splines. From (12) of Stone (1986),

$$\int (\sum_{i=1}^{m} (\beta_i - \beta_i^*)\varphi_{m,q,i}(x))^2 dx \geq \frac{\gamma_q}{m} \sum_{i=1}^{m} (\beta_i - \beta_i^*)^2 ,$$

for some constant γ_q depending only on q. Thus, the two requirements are satisfied with $T_{q,1} = 1$ and $T_{q,2} = \gamma_q$. Therefore, Corollary 4.3 is applicable to the log-spline models. Let us index our models by $k = (m, q, L)$. We specify the model complexity in a natural way to describe the index as follows:

1. describe L using $\log_2^* L$ bits
2. describe q using $\log_2^* q$ bits
3. describe m using $\log_2^* m$ bits,

where the function $\log^*$ is defined by $\log^* i = \log(i+1) + 2\log\log(i+1)$ for $i > 0$. Then the total number of bits needed to describe k is $\log_2^* L + \log_2^* q + \log_2^* m$. Thus a natural choice of C_k is $C_k = \log^* L + \log^* q + \log^* m$.

Assume the logarithm of the target density belongs to $W_2^{s^*}(U^*)$ for some $s^* \geq 1$ and $U^* > 0$, where $W_2^s(U)$ is the Sobolev space of functions g on [0,1] for which $g^{(s-1)}$ is absolute continuous and $\int (g^{(s)}(x))^2 dx \leq U$. The parameters s^* and U^* are not known.

Corollary 4.4: Let $\hat{f} = f_{\hat{k},\hat{\theta}(\hat{k})}$ be the density estimator with $\hat{k}$ selected by the criterion in (4.9) with $\lambda^*(q, L) = 42.0 + 4.75 \log\left(\frac{1}{\gamma_q}(L+1)e^{\frac{L}{2}}\right)$. Then for any f with $\log f \in W_2^{s^*}(U^*)$,

$$Ed_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) \leq M \cdot n^{-\frac{2s^*}{2s^*+1}} ,$$

where the constant M depend only on s^*, $\| \log f \|_\infty$ and $\| (\log f)^{(s^*)} \|_2$.

This corollary guarantees the optimal rate of convergence for densities with logarithms in Sobolev balls without knowing U and s in advance. It shows that with a good model selection criterion, we could perform asymptotically as well as we knew the smoothness parameters. This theorem demonstrates an example of success of a completely data-driven strategy for nonparametric density estimation.

Proof of Corollary 4.4: We examine the resolvability bounds for the classes of density functions considered. To do so, we need to upper-bound the approximation error for a good sequence of models. By Theorem 5.2 and Theorem 2 .1 of de Boor and Fix (1973), for $\log f \in W_2^{s^*}(U^*)$ and for each $m \geq s^*$, there exists $g(x, \beta) = \sum_{i=1}^m \beta_i \varphi_{m,s^*,i}(x)$ such that

$$\| \log f - g \|_2 \leq \frac{K}{(m - s^* + 2)^{s^*}} \| \log^{(s^*)} f \|_2,$$

$$\| \log f - g \|_\infty \leq \frac{K'}{(m - s^* + 2)^{s^* - 0.5}} \| \log^{(s^*)} f \|_2,$$

where K and K' are absolute constants. By Lemma 4.6 in Section 6,

$$| \log \int e^g d\mu | = | \log \int f d\mu - \log \int e^g d\mu | \leq \| \log f - g \|_\infty .$$

Let $\tilde{g} = g - \log \int e^g d\mu$ be the normalized log-density from g. Then

$$\| \log f - \tilde{g} \|_\infty \leq \| \log f - g \|_\infty + \| \log \int e^g d\mu \|_\infty \leq 2 \| \log f - g \|_\infty .$$

Therefore

$$\begin{aligned} \| \tilde{g} \|_\infty &\leq \| \log f \|_\infty + 2 \| \log f - g \|_\infty \\ &\leq \| \log f \|_\infty + \frac{2K'}{(m - s^* + 2)^{s^* - 0.5}} \| \log^{(s^*)} f \|_2. \end{aligned}$$

For the relative entropy approximation error, from Lemma 1 in B & S,

$$\begin{aligned} D(f \| e^{\tilde{g}}) &\leq \frac{1}{2} e^{\| \log f - g \|_\infty} \times \| f \|_\infty \times \| \log f - g \|_2^2 \\ &\leq \frac{K^2 \exp\{ \frac{K'}{(m - s^* + 2)^{r^* - 0.5}} \| \log^{(s^*)} f \|_2 + \| \log f \|_\infty \}}{2(m - s^* + 2)^{2s^*}} \| \log^{(s^*)} f \|_2^2. \end{aligned}$$

Take $L_m = \lceil \| \log f \|_\infty + \frac{2K'}{(m - s^* + 2)^{s^* - 0.5}} \| \log^{(s^*)} f \|_2 \rceil$ (bounded for $\log f \in W_2^{s^*}(U^*)$), then $\lambda^*(s^*, L_m)$ are bounded. Note also that C_{m,s^*,L_m} is asymptotically negligible compared to m. Thus

$$\begin{aligned} R_n(f) &\leq D(f \| e^{\tilde{g}}) + \frac{\lambda^*(s^*, L_m) m}{n} + \frac{9.49 C_{m,s^*,L_m}}{n} \\ &\leq \frac{const_1}{m^{2s^*}} + \frac{const_2 \cdot m}{n}, \end{aligned}$$

where the two constants depend only on s^*, $\| \log f \|_\infty$ and $\| \log^{(s^*)} f \|_2$. Optimizing over m, we obtain the conclusion with the choice of m of order $n^{\frac{1}{2s^* + 1}}$.

General linear spaces

Unlike the localized basis that satisfy (4.7) and (4.8), general basis are not as well handled by the present theory. Here we show a logarithmic factor arises in both the penalty term and in the bound on the convergence rate for polynomial and trigonometric basis.

Let S_j, $j \in J$ be a general linear function spaces on $[0,1]^d$ spanned by a bounded and linearly independent (under L_2 norm) basis 1, $\varphi_{j,1}(x)$, ..., $\varphi_{j,m_j}(x)$. The finite dimensional families we consider are:

$$f_j(x, \theta) = \exp(\sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) - \psi_j(\theta)), \ j \in J,$$

where $\psi_j(\theta) = \log \int \exp(\sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x)) d\mu$ is the normalizing constant.

In B & S, the supreme of the ratio of sup-norm and L_2-norm for functions in S_j plays an important role in the analysis. For general linear spaces, we also consider this ratio.

The linear spaces $S_j, j \in J$ we consider have the property that for each j, there exists a positive constant K_j such that

$$\| h \|_\infty \leq K_j \| h \|_2 \tag{4.10}$$

for all $h \in S_j$. This property follows from the boundness and linear independence (under L_2-norm) assumption on the basis.

For the same reason as in subsection A, break the natural parameter space into a increasing sequence of compact spaces

$$\Theta_{j,L} = \{\theta \in R^{m_j} : \| \log f_j(\cdot, \theta) \|_\infty \leq L\}, \quad L \geq 1,$$

and treat each of them as a model. Then for each j, we have a sequence of models $f_j(x, \theta)$, $\theta \in \Theta_{j,L}, L \geq 1$. We index the new models by $k = (j, L)$ and let Γ be the collection of k.

Lemma 4.3: For each model $k = (j, L)$, Assumption 4.1′ is satisfied with $A_{(j,L)} = 19.28 K_j (1 + L) e^{\frac{L}{2}} + 0.06$ and $m_{(j,L)} = m_j + 1$.

The proof of this lemma is in Section 5.

If an upper bound on $\| \log f \|_\infty$ is known in advance, then for each j, we can consider only $L = \lceil \| \log f \|_\infty \rceil$. Then from the remark to Theorem 4.1, the model complexity can be ignored. However, when $\| \log f \|_\infty$ is unknown, we would like to consider all integer values for L. Then for each model size, we have countably many models. To control the selection bias, we consider the model complexity.

Let C_k, $k \in \Gamma$ be any model complexity satisfying $\sum e^{-C_k} \leq 1$. Let $\lambda^*_{(j,L)} = \Lambda(2A_{(j,L)}) = 42.0 + 4.75 \log \left(K_j (1 + L) e^{\frac{L}{2}} \right)$. Let $\hat{k}$ be the model minimizing

$$-\sum_{i=1}^{n} \log f_k(X_i, \hat{\theta}^{(k)}) + \lambda^*_{(j,L)} m_k + 9.49 C_k.$$

Since the conditions for Theorem 4.1 are satisfied, we have the following result about model selection for a sequence of exponential families with a general linear basis.

Corollary 4.5: For the log-density models with basis satisfying (4.10), for any underline density f,

$$Ed_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) \leq 2657 R_n(f),$$

where

$$R_n(f) = \inf_{L\geq 1} \inf_{j\in J} \{ \inf_{\theta\in\Theta_{j,L}} D(f\|f_{k,\theta}) + \frac{\lambda^*_{(j,L)} m_k}{n} + \frac{9.49 C_k}{n} \}.$$

To apply the corollary for a density class, the approximation error $\inf_{\theta\in\Theta_{j,L}} D(f\|f_{k,\theta})$ should be examined. Then the resolvability will be determined.

Example 4.2: Polynomial case.

Let S_j =span$\{1, x, x^2, ..., x^j\}$, $j \geq 1$. Then $m_j = j$. From Lemma 6 in B & S, $K_j = j+1$. It follows from Lemma 4.3 that $\lambda^*_{(j,L)} = 42.0 + 4.75 \log\left((j+1)(L+1)e^{\frac{L}{2}}\right)$. Take $C_k = \log^* L + \log^* j$. For densities with logarithms in each of the Sobolev spaces $W_2^s(U)$, $s \geq 1$ and $U > 0$, when L is large enough, say $L \geq L^*$ (depending on U and s), the relative entropy approximation error of model (j, L) is bounded by $const_{U,s} \frac{1}{j^{2s}}$ (the examination of relative entropy approximation error is very similar to that in Example 1 in the previous subsection. For details on L_2 and L_∞ error bounds for polynomial approximation, see Section 7 of B & S). Thus $\inf_{\theta\in\Theta_{j,L^*}} D(f\|f_{k,\theta}) + \frac{\lambda^*_{(j,L^*)} j}{n} + \frac{9.49 C_k}{n} \leq const_{U,s}\left(\frac{1}{j^{2s}} + \frac{j \log j}{n}\right)$. Optimizing over j, we obtain that

$$R_n(f) \leq const_{U,s} \times (\log n) n^{-\frac{2s}{2s+1}}.$$

(since the infimum will produce a value at least as small at $j = n^{\frac{1}{2s+1}}$ and $L = L^*$). Therefore, the statistical risks of the density estimators based on the polynomial basis (without knowing the parameters s and U in advance) are within a logarithmic factor $\log n$ of the minimax risks.

Example 4.3: Trigonometric case.

Let S_j =span$\{1, \sqrt{2}\cos(2\pi x), \sqrt{2}\sin(2\pi x), ..., \sqrt{2}\sin(2\pi j x))\}$, $j \geq 1$. Then $m_j = 2j$. From (7.6) in B & S, $K_j = \sqrt{2j+1}$. Again by examining the resolvability (for L_2 and L_∞ error bounds for trigonometric approximation, see Section 7 of B & S), the same convergence rates as those using polynomial bases can be shown for densities with logarithms in the Sobolev spaces and satisfying certain boundary conditions.

The risk bounds derived here using the nonlocalized polynomial or trigonometric basis have an extra $\log n$ factor compared to the minimax risk. The extra factor comes in because the penalty coefficient λ_j^* in the criteria is of order $\log j$ for both cases. Recently, Birgé and Massart (1995) uses a theorem of Talagrand (1994) to show that if $K_j \leq const\sqrt{j}$, then their penalized projection estimator (PPE) with the bias-correction penalty term $const\frac{j}{n}$ converges at the optimal rate. This result is applicable for the trigonometric basis, but not the polynomial basis. Their argument can also be used for log-density estimation using maximum likelihood method with trigonometric basis to derive a criterion giving the optimal convergence rate.

4.4.2 Neural network models

Let $f(x)$ be an unknown density function on $[-\frac{1}{2}, \frac{1}{2}]^d$ with respect to Lebesgue measure. The traditional methods to estimate densities often fail when d is moderately large due to the "curse of dimensionality". Neural network models have been shown to be promising in some statistical applications. Here we consider the estimation of the logarithm of the density $\log f$ by neural nets.

We approximate $g(x) = \log f(x)$ using feedforward neural network models with one layer of sigmoidal nonlinearities, which have the following form:

$$g_k(x, \theta) = \sum_{j=1}^{k} \eta_j \phi(a_j^T x + b_j) + \eta_0.$$

The function is parametrized by θ, consisting of $a_j \in R^d, b_j, \eta_j \in R$, for $j = 1, 2, ...k$. The normalizing constant η_0 is $\eta_0 = -\log \int_{[-\frac{1}{2}, \frac{1}{2}]^d} \exp\{\sum_{j=1}^{k} \eta_j \phi(a_j^T x + b_j) dx$. The integer $k \geq 1$ is the number of nodes (or hidden units). Here ϕ is a given sigmoidal function with $\| \phi \|_\infty \leq 1$, $\lim_{z \to \infty} \phi(z) = 1$ and $\lim_{z \to -\infty} \phi(z) = 0$. Assume also that ϕ satisfies Lipschitz condition $|\phi(z_1) - \phi(z_2)| \leq v_1 |z_1 - z_2|, z_1, z_2 \in R$ for some constant $v_1 > 0$. Let $v = \max(v_1, 1)$. Let

$$f_k(x, \theta) = \exp\{g_k(x, \theta)\} = \exp\{\sum_{j=1}^{k} \eta_j \phi(a_j^T x + b_j) + \eta_0\}$$

be the approximating families. The parameter θ will be estimated and the number of nodes will be automatically selected based on the sample.

The target class we are interested in here was previously studied by Barron (1993, 1994), Modha and Masry (1994). The log-density $g(x)$ is assumed to have a Fourier representation of the form $g(x) = \int_{R^d} e^{i\omega^T x}\tilde{g}(\omega)d\omega$. Let $\sigma_g = \int |\omega|_1|\tilde{g}(\omega)|d\omega$, where $|\omega|_1 = \sum_{j=1}^d |\omega_j|$ is the l_1 norm of ω in R^d. For the target density, we assume $\sigma_g \leq \sigma$. Recent work of Barron (1994) gives nice approximation bounds using the network models for the class of functions with σ_g bounded and the bounds are applied to obtain good convergence rates for nonparametric regression. Modha and Masry prove similar convergence results for density estimation. In these works, the parameter spaces are discretized. We here intend to obtain similar conclusion without discretization.

Consider the parameter space

$$\Theta_{k,\tau_k,\sigma} = \{\theta : \max_{1\leq j\leq k} |a_j|_1 \leq \tau_k, \max_{1\leq j\leq k} |b_j| \leq \tau_k, \sum_{j=1}^k |\eta_j| \leq 2\sigma\}.$$

The constant τ_k is chosen such that

$$\mathrm{dis}(\phi_{\tau_k}, \mathrm{sgn}) =: \inf_{0<\varepsilon\leq\frac{1}{2}} \left(2\varepsilon + \sup_{|z|\geq\varepsilon} |\phi(\tau_k z) - \mathrm{sgn}(z)|\right) \leq \frac{1}{\sqrt{k}}.$$

The compact parameter spaces are used so that the cardinality assumption is satisfied. From Theorem 3 in Barron (1993), for a log-density g with $\sigma_g \leq \sigma$, there exists a $\theta \in \Theta_{k,\tau_k,\sigma}$ such that

$$\| g - g_{k,\theta} \|_{[-\frac{1}{2},\frac{1}{2}]^d} \leq \frac{4\sigma_g}{\sqrt{k}}, \tag{4.11}$$

where $\| \ \|_{[-\frac{1}{2},\frac{1}{2}]^d}$ denote the L_2-norm for functions defined on $[-\frac{1}{2},\frac{1}{2}]^d$.

For simplicity, for the target density class, the upper bound σ on σ_g is assumed to be known (otherwise an increasing sequence of σ values can be considered and let the model selection criterion choose a suitable one).

Now we want to show that Assumption 4.1′ is satisfied for these models. For any $\varepsilon > 0, \sigma \geq 1$, from the proof of Lemma 6 in Barron (1994), there exists a set $\Theta_{k,\varepsilon,\tau_k,\sigma}$ such that for any $\theta \in \Theta_{k,\tau_k,\sigma}$, there is $\tilde{\theta} \in \Theta_{k,\varepsilon,\tau_k,\sigma}$ satisfying $\| g_k(x,\theta) - g_k(x,\tilde{\theta}) \|_\infty \leq 8v\sigma\varepsilon$ with

$$\mathrm{card}\,(\Theta_{k,\varepsilon,\tau,\sigma}) \leq \left(\frac{2e(\tau_k+\varepsilon)}{\varepsilon}\right)^{k(d+1)} \left(\frac{2(1+\varepsilon)}{\varepsilon}\right)^k.$$

Take $\varepsilon = \frac{\delta}{8v\sigma}$. Because $B_k(\theta^*, r) = \{\theta \in \Theta_{k,\tau_k,\sigma} : d_H^2(f_{k,\theta^*}, f_{k,\theta}) \leq r^2\} \subset \Theta_{k,\tau_k,\sigma}$, so for $\delta \leq \rho r$, we have a δ-net in $B_k(\theta^*, r)$ with cardinality bounded by

$$\left(\frac{2e(8v\sigma\tau_k+\rho r)}{\delta}\right)^{k(d+1)} \left(\frac{2(8v\sigma+\rho r)}{\delta}\right)^k$$

$$= \left(\frac{2e(8v\sigma\tau_k+\rho r)}{r}\right)^{k(d+1)} \left(\frac{2(8v\sigma+\rho r)}{r}\right)^k \left(\frac{r}{\delta}\right)^{kd+2k}.$$

Notice that Assumption 4.0 only needs to be checked for $r \geq \sqrt{\frac{4\log 2}{1-4\gamma}\frac{m_k}{n}}$, where $m_k = kd + 2k + 1$ is the number of parameters, for such r, the above quantity is bounded by

$$\left(\frac{16ev\sigma\tau_k}{\sqrt{\frac{4\log 2}{1-4\gamma}\frac{m_k}{n}}} + 2e\rho\right)^{k(d+1)} \left(\frac{16v\sigma}{\sqrt{\frac{4\log 2}{1-4\gamma}\frac{m_k}{n}}} + 2\rho\right)^{k} \left(\frac{r}{\delta}\right)^{kd+2k}.$$

Thus Assumption 4.1′ is satisfied with

$$A_k = A_{k,n} = \left(\frac{16ev\sigma\tau_k}{\sqrt{\frac{4\log 2}{1-4\gamma}\frac{m_k}{n}}} + 2e\rho\right)^{\frac{k(d+1)}{kd+2k}} \left(\frac{16v\sigma}{\sqrt{\frac{4\log 2}{1-4\gamma}\frac{m_k}{n}}} + 2\rho\right)^{\frac{k}{kd+2k}} \leq \text{const} \times \tau_k\sqrt{\frac{n}{m_k}}.$$

As shown in Barron (1993), if $\phi(z)$ approaches its limits at least polynomially fast, then there exist constants β_1 and β_2 such that $\tau_k \leq \beta_1 k^{\beta_2}$. As a consequence, $A_{k,n} \leq \text{const} \times k^{\beta_2 - \frac{1}{2}}\sqrt{n}$. By Theorem 4.1, when we choose the penalty constants $\lambda_k = \lambda_k^* = \Lambda(2A_{k,n})$ and $\nu = 9.49$ in the model selection criterion given in (4.2), for the density estimator $f_{\hat{k},\hat{\theta}(\hat{k})}$, we have

$$Ed_H^2(f, \hat{f}) \leq 2657R_n(f),$$

where $R_n(f) = \inf_{k\geq 1}\{\inf_{\theta\in\Theta_{k,\tau_k,\sigma}} D(f \parallel f_{k,\theta}) + \frac{\lambda_k^* m_k}{n} + 9.49\log^* k\}$.

For the targeted densities, under the assumption $\sigma_g \leq \sigma$, the log-density is uniformly bounded (see Lemma 5.3 in Modha and Masry (1994)). Indeed, because $\| g(x) - g(0) \|_\infty = \| \int_{R^d} \left(e^{i\omega^T x} - 1\right)\tilde{g}(\omega)d\omega \| \leq \int_{R^d} |\omega^T x||\tilde{g}(\omega)|d\omega \leq \frac{1}{2}\int_{R^d} |\omega|_1|\tilde{g}(\omega)|d\omega \leq \frac{1}{2}\sigma_g$, so $|g(0)| = |-\log \int_{[-\frac{1}{2},\frac{1}{2}]^d} e^{g(x)-g(0)}dx| \leq \| g(x) - g(0) \|_\infty \leq \frac{1}{2}\sigma_g$. It follows that $\| g(x) \|_\infty \leq \| g(x) - g(0) \|_\infty + |g(0)| \leq \sigma_g$. Thus by Lemma 1 of B & S, for the target densities, $D(f \parallel f_{k,\theta}) \leq \text{const}_\sigma \| g - g_{k,\theta} \|^2_{[-\frac{1}{2},\frac{1}{2}]^d}$ for $\theta \in \Theta_{k,\tau_k,\sigma}$. So from (4.11), $\inf_{\theta\in\Theta_{k,\tau_k,\sigma}} D(f \parallel f_{k,\theta}) \leq \text{const}_\sigma \frac{1}{k}$. Note λ_k^* is of order $\log A_k = O(\log(nk^{2\beta_2-1}) = O(\log n)$. Therefore

$$R_n(f) = O\left(\inf_{k\geq 1}\left(\frac{\sigma}{k} + \frac{kd+2k}{n}\log n\right)\right) = O\left(\frac{\sigma d\log n}{n}\right)^{\frac{1}{2}}.$$

Note that for the class of functions considered, the rate of convergence $\frac{1}{2}$ is independent of the function dimension as in Barron (1993), Modha and Masry (1994).

4.4.3 Estimating a not strictly positive density

An unpleasant property of the exponential families, log neural network models, or some other log-density estimation methods is that each density is bounded away from 0 on the whole space $[0,1]^d$. If the support of the true density is only a subset of $[0,1]^d$, the resolvability bounds derived in the above sections are still valid. However, for such densities, the

approximation capability of the exponential families may be very poor. Here we present a way to get around this difficulty. We get the optimal rates in L_1 with localized basis while still using the resolvability for upper bounds.

We here use the same idea in Section 2.3 in Chapter 2 to change the original estimation problem to another one which can be handled more easily. In addition to the observed i.i.d. sample $X_1, X_2, ..., X_n$ from f with respect to μ on a compact space $\mathcal{X}$ with $\mu(\mathcal{X}) = 1$, let $Y_1, Y_2, ...Y_n$ be a generated i.i.d. sample (independent of X_i's) from the uniform distribution on $\mathcal{X}$ (with respect to μ). Let Z_i be X_i or Y_i with probability $(\frac{1}{2}, \frac{1}{2})$ using $V_i \sim Bernoulli(\frac{1}{2})$ independently for $i = 1, ..., n$. Then Z_i has density $g(x) = \frac{1}{2}(f + 1)$. Then g is bounded below from 0. We will first use the exponential models $f_k(x, \theta), \theta \in \Theta_k$ to estimate g and then construct a suitable estimator for f.

Let $\hat{g}$ be the density estimator of g based on $Z_1, ...Z_n$ using the criterion in (4.2) from the models in Γ, which satisfy Assumption 4.1'. Then when λ_k and ν are chosen large enough, by Corollary 4.1,

$$E \parallel g - \hat{g} \parallel_{L_1} \leq 104\sqrt{R_n(g)}.$$

Let $\tilde{g}(x) = \hat{g}(x)I_{\{\hat{g}(x) \geq \frac{1}{2}\}} + \frac{1}{2}I_{\{\hat{g}(x) < \frac{1}{2}\}}$. Then because pointwise in x, $|g - \tilde{g}| \leq |g - \hat{g}|$,

$$E \int |g - \tilde{g}| d\mu \leq E \int |g - \hat{g}| d\mu \leq 104\sqrt{R_n(g)} \ .$$

In particular, $E \int \tilde{g} d\mu - 1 \leq 104\sqrt{R_n(g)}$. Let

$$\hat{f}_{rand}(x) = \frac{2\tilde{g}(x) - 1}{2 \int \tilde{g}(x) d\mu - 1} \ .$$

Then $\hat{f}_{rand}(x)$ is a nonnegative and normalized probability density estimate and depend on the $X_1, ..., X_n$ and the auxiliary variables $Y_1, ..., Y_n$, $V_1, ..., V_n$. So it is a randomized estimator. Now

$$\begin{aligned} E \int |f(x) - \hat{f}_{rand}(x)| d\mu &\leq E \int |f(x) - 2\tilde{g}(x) + 1| d\mu + E \int |\hat{f}_{rand}(x) - 2\tilde{g}(x) + 1| d\mu \\ &= 2E \int |g - \tilde{g}| d\mu + 2E(\int \tilde{g}(x) d\mu - 1) \\ &\leq 416\sqrt{R_n(g)}. \end{aligned}$$

Thus, we have the following result.

Theorem 4.2: Let $\hat{f}_{rand}$ be constructed in the above way with a choice of the penalty constants satisfying $\lambda_k \geq \lambda_k^*$, $\nu \geq 9.49$, $k \in \Gamma$, then

$$E \int |f(x) - \hat{f}_{rand}(x)| d\mu \leq 416\sqrt{R_n(g)}.$$

Because of convexity, a nonrandomized estimator can be obtained with no bigger L_1 risk.

Because g is bounded below from 0, g can be better approximated by the exponential families. Then $\sqrt{R_n(g)}$ can yield a much faster rate of convergence compared to $\sqrt{R_n(f)}$. We next give an example to show that for some classes of densities, with the modifications, the modified estimator achieves the optimal rate of convergence.

Example 4.1: (continued): We now assume that $\int (f^{(s^*)}(x))^2 dx < \infty$ for some unknown integer s^*. Note that the densities considered here are not necessarily strictly positive on [0,1].

Let $\hat{f}$ be the estimator constructed according to the above procedure. Then we have

$$E \int |f(x) - \hat{f}(x)| dx \leq 416 \sqrt{R_n(g)} \ .$$

From $\int (f^{(s^*)}(x))^2 dx < \infty$, it can be shown that $\int \left((\log g)^{(s^*)} \right)^2 dx < \infty$. Then from previous result, $R_n(g) = O(n^{-\frac{2s^*}{2s^*+1}})$. Thus

$$E \int |f(x) - \hat{f}(x)| dx \leq \zeta n^{-\frac{s^*}{2s^*+1}} \ ,$$

where the constant ζ depends only on s^* and $\int (f^{(s^*)}(x))^2 dx$. Therefore, the density estimator converges in L_1-norm to the true density at the optimal rate simultaneously for the classes of densities $G(s, U), s \geq 1, U > 0$, where $G(r, U)$ is defined to be the collection of densities with square-integral of the s-th derivative bounded by U.

4.4.4 Complete models versus sparse subset models

As in section 4.1, we consider the estimation of the log-density $\log f(x)$ on $[0,1]^d$ using a sequence of linear spaces. Traditionally, the linear spaces are chosen by spanning the basis functions in a series expansion using polynomial, or trigonometric, or splines, etc., up to certain orders. Then use a model selection criterion to select the order for good statistical estimation. When the true function is sparse in the sense that only a small fraction of the basis functions in the linear spaces are needed to provide a nearly as good approximation as that using all the basis functions, then a subset model might dramatically outperform the

complete models, because excluding many (nearly) unnecessary terms significantly reduces the variability of the function estimate.

For simplicity, assume the linear spaces are nested, i.e., $S_i \subset S_j$ for $i < j$. Let S_j be spanned by a bounded and linearly independent (under L_2 norm) basis 1, $\varphi_{j,1}(x), \varphi_{j,2}(x)$, ...,$\varphi_{j,L_j}(x)$. Let

$$f_j(x,\theta) = \exp(\sum_{i=1}^{L_j} \theta_i \varphi_{j,i}(x) - \psi_j(\theta)), \quad \theta = (\theta_1, ...\theta_{L_j}) \in \Theta_j,$$

where $\psi_j(\theta) = \log \int \exp(\sum_{i=1}^{L_j} \theta_i \varphi_{j,i}(x)) d\mu$ is the normalizing constant. Including all of the L_j terms, we have dimension $m_j = L_j$. We call such a model a complete one (with respect to the given linear spaces) because it uses all the L_j basis functions in S_j. On the other hand, we can also consider the subset models

$$f_{I_j}(x,\theta) = \exp(\sum_{i \in I_j} \theta_i \varphi_{j,i}(x) - \psi_{I_j}(\theta)), \theta \in \Theta_{I_j},$$

where $\psi_{I_j}(\theta) = \log \int \exp(\sum_{i \in I_j} \theta_i \varphi_{j,i}(x)) d\mu$ and $I_j \subset \{1, 2, ..., L_j\}$ is a subset. We next show the possible advantage of considering these subset models through the comparison of the resolvability for the complete models with that for the subset models for some classes of densities.

Suppose that Assumption 4.1 is satisfied with dimensionality constant A_j and dimension L_j for the complete models and with A_{I_j} and m_{I_j} for the subset models, where $m_{I_j} = |I_j|$ is the number of parameters in model I_j. We also assume that there exist two positive constants β_1 and β_2 such that $A_{I_j} \leq \beta_1 L_j^{\beta_2}$ for all the subset models. To satisfy this requirement, we may need to restrict the parameters to compact spaces $\Theta_{I_j,L} = \{\theta \in R^{m_{I_j}} : \| \log f_{I_j}(\cdot,\theta) \|_\infty \leq L\}$ for a fixed value L. Then from Lemma 4.3, this condition is satisfied if K_j in (4.10) is bounded by a polynomial of L_j, which is satisfied by polynomial, spline, and trigonometric basis. (When $\| \log f \|_\infty < \infty$ but no upper bound on $\| \log f \|_\infty$ is known, increasing sequences of compact parameter spaces could be considered and the condition could be replaced by $A_{I_j,L} \leq \beta_{1,L} L_j^{\beta_2}$, where $\beta_{1,L}$ is allowed to grow in L. Then similar asymptotic results hold.)

For a sequence of positive integers $N_n \uparrow \infty$, let $\Gamma_n = \{j : L_j \leq N_n\}$ and $\tilde{\Gamma}_n = \{(j, I_j) : L_j \leq N_n$ and $I_j \subset \{1, 2, ...L_j\}\}$. For each sample size n, the list of the models we consider is either Γ_n (complete models) or $\tilde{\Gamma}_n$ (subset models). In our analysis, we need the condition that N_n grows no faster than polynomially in n to have a good control of the model

complexities for the subset modes. This restriction is quite reasonable because usually a model with the number of parameters bigger than the number of observations can not be estimated well.

For the complete models, the model complexity C_j can be taken as $C_j = \log^* j$. Let $\lambda_j^* = \Lambda(A_j)$. Let $\hat{j}$ be the model minimizing the following criterion value

$$-\sum_{i=1}^{n} \log f_j(X_i, \hat{\theta}^{(j)}) + \lambda_j^* L_j + 9.49 C_k$$

over $j \in \Gamma_n$. Then from Theorem 4.1, the statistical risk of the density estimator $f_{\hat{j},\hat{\theta}^{(\hat{j})}}$ from the selected model $\hat{j}$ under the squared Hellinger loss is bounded by a multiple of the following index of resolvability

$$R_n(f) = \inf_{j \in \Gamma_n} \{ \inf_{\theta^{(j)} \in \Theta_j} D(f \| f_{j,\theta^{(j)}}) + \frac{\lambda_j^* L_j}{n} + \frac{9.49 \log^* j}{n} \}.$$

Let j_n be the optimal model which minimizes $R_n(f)$.

Now consider the subset models. We have exponentially many (2^{L_j} to be exact) subset models from the complete model j. To apply the model selection results, we consider choosing an appropriate model complexity. A natural way to describe a subset model is that first describe j, then describe the number of terms m_{I_j} in the model, and finally describe which one the model is among $\binom{L_j}{m_{I_j}}$ possibilities. This strategy suggests the following choice of complexity:

$$C_{I_j} = \log^* j + \log L_j + \log \binom{L_j}{m_{I_j}}.$$

Take $\lambda_{I_j}^* = \Lambda(A_{I_j})$. Let $\bar{j}$ and $\bar{I} = \bar{I}_{\bar{j}}$ be the minimizer of the following criterion value

$$-\sum_{i=1}^{n} \log f_{I_j}(X_i, \hat{\theta}^{(I_j)}) + \lambda_{I_j}^* m_{I_j} + 9.49 C_{I_j}$$

over $(j, I_j) \in \tilde{\Gamma}_n$. Again from Theorem 4.1, the risk of the density estimator $f_{\bar{I},\hat{\theta}^{(\bar{I})}}$ resulting from model selection among the subset models is bounded by a multiple of the following index of resolvability for the subset models

$$\tilde{R}_n(f) = \inf_{j \in \Gamma_n} \{ \inf_{I_j} \{ \inf_{\theta^{(I_j)} \in \Theta_{I_j}} D(f \| f_{I_j,\theta^{(I_j)}}) + \frac{\lambda_{I_j}^* m_{I_j}}{n} + \frac{9.49 C_{I_j}}{n} \} \}.$$

For the subset models, another quantity similar to the above resolvability is of interest. Let

$$r_n(f) = \inf_{j \in \Gamma_n} \inf_{I_j} \max \left(\inf_{\theta^{(I_j)} \in \Theta_{I_j}} D(f \| f_{I_j,\theta^{(I_j)}}), \frac{m_{I_j}}{n} \right).$$

Then $r_n(f)$ is roughly the ideal best trade-off between the approximation error and the estimation error among all the subset models. Let $\tilde{j}_n$, $I^* = I^*_{\tilde{j}_n}$ and $\theta_* = \theta^{(I^*)}_*$ be the minimizer of $r_n(f)$. Ideally, we wish the density estimator $f_{\overline{I},\hat{\theta}(\overline{I})}$ converges at the same rate as $r_n(f)$. But this may not be possible because so many models are present that it is too much to hope that the likelihood processes behave well uniformly for all the models. In the next proposition, we compare $R_n(f)$, $\tilde{R}_n(f)$ and $r_n(f)$.

Proposition 4.1:

1. The resolvability for the subset models is at least as good as that for the complete models asymptotically. That is,

$$\overline{\lim}_{n\to\infty} \frac{\tilde{R}_n(f)}{R_n(f)} \leq 1. \tag{4.12}$$

2. Let $N_n \leq n^\kappa$ for some positive constant κ. Then the resolvability for the subset models is within a $\log n$ factor of the ideal convergence rate $r_n(f)$. That is,

$$\tilde{R}_n(f) = O(r_n(f) \log n). \tag{4.13}$$

3. With the above choice of N_n, the improvement of the subset models over the complete models in terms of resolvability is characterized by how small the optimal subset model size is compared to the optimal complete model size as suggested by the following inequality:

$$\frac{\tilde{R}_n(f)}{R_n(f)} = O\left(\frac{m_{I^*}}{L_{j_n}} \log n\right). \tag{4.14}$$

The results in the proposition can be easily proved. The inequality (4.12) is suggested by that the complete model is included as one of the subset models. Indeed, $\tilde{R}_n(f) \leq \inf_{j\in\Gamma_n}\{\inf_{\theta(j)\in\Theta_j} D(f\|f_{j,\theta(j)}) + \frac{\lambda^*_j L_j}{n} + \frac{9.49(\log^* j + \log^* L_j)}{n}\}$, and since the logarithmic terms in this case are of smaller order than the $\frac{L_j}{n}$ term, it follows that $\overline{\lim}_{n\to\infty} \frac{\tilde{R}_n(f)}{R_n(f)} \leq 1$. When the true density is sparse, we have a good chance of obtaining a much more accurate estimate. For (4.13), Because $\log\binom{L_j}{m} \leq m \log L_j$ and $L_j \leq n^\kappa$, we have that $C_{I_j} = O(m_{I_j} \log n)$. Since $A_{I_j} \leq \beta_1 {L_j}^{\beta_2}$, $\lambda^*_{I_j} = O(\log n)$. It follows that $\tilde{R}_n(f) = O(r_n(f) \log n)$. Finally, from (4.13) , $\frac{\tilde{R}_n(f)}{R_n(f)} = O\left(\frac{\max(D(f\|f_{I^*,\theta_*}), \frac{m_{I^*}}{n})}{\frac{L_{j_n}}{n}} \log n\right)$. For the best trade-off, $D(f\|f_{I^*,\theta_*})$ and $\frac{m_{I^*}}{n}$

are of the same order, so $\frac{\tilde{R}_n(f)}{R_n(f)} = O\left(\frac{m_{I^*}}{L_{j_n}} \log n\right)$. This bound is useful when $\frac{m_{I^*}}{L_{j_n}}$ is of smaller order than $\frac{1}{\log n}$.

The ratio $\alpha_n = \frac{m_{I^*}}{L_{j_n}}$ describes how small the (ideally) optimal (in the sense that it gives the resolvability) subset model size is compared to the optimal size of the complete models. We call it a sparsity index for sample size n. The obtained inequality $\frac{\tilde{R}_n(f)}{R_n(f)} \leq O(\alpha_n \log n)$ shows that ignoring the logarithmic factor $\log n$, the sparsity index characterizes the improvement of the index of resolvability bound using the subset models over the complete models.

Example 4.4: Sparse trigonometric series.

Consider the trigonometric expansion on [0,1]. Let

$$F_2^s = \{f : \log f \in W_2^1, \log f = \theta_0 + \sum_{j=1}^{\infty} \theta_j \sin(2j^2\pi x) \text{ for some } \theta\}.$$

Note that some functions in F_2^1 have no more than 1 derivatives. But the functions in F_2^1 are sparse. It can be shown that the resolvability resulting from the complete models is of order $\left(\frac{\log n}{n}\right)^{\frac{2}{3}}$, while the subset models give a resolvability of order $\log n \cdot \left(\frac{1}{n}\right)^{\frac{4}{5}}$. The sparsity index is of order $\left(\frac{\log n}{n}\right)^{\frac{2}{15}}$.

The above example is somewhat artificial. If we knew that only the frequencies of the square numbers are useful before hand, then the good models $\{\sin(2\pi x), \sin(8\pi x), ..., \sin(2^{m+1}\pi x)\}$, $m \geq 1$ could be described much more simply than most subsets of size m out of m^2 terms. However, in realistic situations, the knowledge of the best subset models is not available, so we have to search over the subset models.

Even for one dimensional function estimation, the sparse subset models also turn out to be advantageous in several related settings such as estimating a function with bounded variation using histogram, and estimating a function in the Besov spaces using wavelets. For high dimensional function estimation, there are even more advantages in considering the sparse subset models. When the input dimension is large, the sparse models such as additive models, low order interaction models might give good estimates if the true function can be well approximated by these sparse models. The complete models, on the other hand, often fail with moderate sample size due to the curse of dimensionality. The following ex-

ample demonstrates the advantage of the sparse subset models for high dimensional density estimation.

Example 4.5: Sparse tensor product series.

Let $\{\varphi_0(x), \varphi_1(x), \varphi_2(x), ...\}$ be a bounded orthonormal basis for $L_2[0,1]$. Then the tensor products

$$\{\varphi_{\underline{i}}(x) = \Pi_{l=1}^{d} \varphi_{i_l}(x_l) : \underline{i} = (i_1, ..., i_d) \in \{0, 1, 2, ...\}^d\}$$

provides an orthonormal basis for $L_2[0,1]^d$. Let $|\underline{i}| = \max_{l \le d} i_l$. The complete models are

$$f_j(x, \theta) = \exp(\sum_{|\underline{i}| \le j} \theta_{\underline{i}} \varphi_{\underline{i}}(x) - \psi_j(\theta)),$$

where $\psi_j(\theta) = \log \int \exp(\sum_{|\underline{i}| \le j} \theta_{\underline{i}} \varphi_{\underline{i}}(x)) d\mu$ and the model dimension is $L_j = j^d$. These models often encounter a great difficulty when the function dimension d is large because exponentially many coefficients need to be estimated even if j is small. However, when the true function is sparse, then good estimates are possible by considering the sparse subset models. The subset models are

$$f_{I_j}(x, \theta) = \exp(\sum_{\underline{i} \in I_j} \theta_{\underline{i}} \varphi_{\underline{i}}(x) - \psi_{I_j}(\theta)),$$

where $\psi_{I_j}(\theta) = \log \int \exp(\sum_{\underline{i} \in I_j} \theta_{\underline{i}} \varphi_{\underline{i}}(x)) d\mu$ and $I_j \subset \{\underline{i} : |\underline{i}| \le j\}$. Assume Assumption 4.1 is satisfied with $A_j \le \beta_1 \left(j^d\right)^{\beta_2}$ and dimension $m_j = j^d$ for the complete models and with $A_{I_j} \le \beta_1 \left(j^d\right)^{\beta_2}$ and dimension $m_{I_j} = |I_j|$ for the subset models for some positive constants β_1 and β_2 (as stated before, satisfaction of this condition may require suitable compactification of nature parameter spaces).

Assume $\| \log f \|_\infty \le M_1$, $\log f(x) = \sum_{\underline{i}} \theta^*_{\underline{i}} \varphi_{\underline{i}}(x)$ and the coefficients satisfy the following two conditions for some positive constants M_2, M_3, and s:

$$\sum_{\underline{i}} |\theta^*_{\underline{i}}| \le M_2, \tag{4.15}$$

$$\sum_{\underline{i}} (i_1^2 + ... + i_d^2)^s |\theta^*_{\underline{i}}|^2 \le M_3. \tag{4.16}$$

Let $F(M_1, M_2, M_3, s)$ be the collection of the densities satisfying the above conditions. The hyper-parameters M_1, M_2, M_3, and s are not necessarily known. In the following evaluations of the resolvabilities, these parameters are fixed.

Let $g_j(x) = \sum_{\underline{i}:|\underline{i}|\le j} \theta^*_{\underline{i}}\varphi_{\underline{i}}(x)$ be the best approximator of $\log f$ in the model j in L_2 sense. Then the complete model j has an approximation error

$$\| \log f - g_j \|_2^2 = \sum_{\underline{i}:|\underline{i}|>j} \left(\theta^*_{\underline{i}}\right)^2 \le \frac{1}{(j+1)^{2s}} \sum_{\underline{i}:|\underline{i}|>j} \left(i_1^2 + ... + i_d^2\right)^s \left(\theta^*_{\underline{i}}\right)^2 \le \frac{M_3}{(j+1)^{2s}}.$$

Then using the same technique used in Subsection 4.4.1 (Lemma 1 in B & S is still applicable because $\| g_j \|_\infty$ is bounded, which follows from boundedness of the basis functions and $\sum_{\underline{i}} |\theta^*_{\underline{i}}| \le M_2$), it can be shown that the resolvability for the complete models is of order $\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d}}$.

Now consider the approximation error for the subset models from the complete model j. Let $g_{m,j}(x)$ be the sum using the m largest $|\theta^*_{\underline{i}}|$ among the j^d terms. Let $|\theta_{(1)}| \ge |\theta_{(2)}| \ge ... \ge |\theta_{(j^d)}|$ be the ordered coefficients of the first j^d terms. Then the approximation error of $g_{m,j}$ is $\| \log f - g_{m,j} \|_2^2 = \| \log f - g_j \|_2^2 + \| g_{m,j} - g_j \|_2^2 \le \frac{M_3}{(j+1)^{2s}} + \| g_j - g_{m,j} \|_2^2$. But $\| g_j - g_{m,j} \|_2^2 = \sum_{k\ge m+1} |\theta^*_{(k)}|^2 \le \sum_{k\ge m+1} |\theta^*_{(m+1)}||\theta^*_{(k)}| \le |\theta^*_{(m+1)}| \sum_{k\ge m+1} |\theta^*_{(k)}| \le \frac{M_2^2}{m}$. Thus $\| \log f - g_{m,j} \|_2^2 \le \frac{M_3}{(j+1)^{2s}} + \frac{M_2^2}{m}$. For $s \ge \frac{1}{2}$, to achieve the approximation error of rate $\frac{1}{m}$, j can be taken as m. The corresponding complexity is $\log^* j + \log\left(j^d\right) + \log\binom{j^d}{m} = O(md\log m)$. Again, with the technique used in Subsection 4.4.1, the resolvability for the sparse subset models is seen to be within a multiple (depending only on M_1, M_2, M_3, s, β_1 and β_2) of $\sqrt{\frac{d\log n}{n}}$. The resulting rate of convergence is independent of the function dimension d and is better than that from the complete models for $F(M_1, M_2, M_3, s)$ with $2s < d$. For $s < \frac{1}{2}$, in order to achieve the approximation error of rate $\frac{1}{m}$, j needs to be at least of order $m^{\frac{1}{2s}}$. Then the model complexity is of order $\frac{dm\log m}{s}$ and the resolvability for the sparse subset models is again within a multiple (depending only on M_1, M_2, M_3, s, β_1 and β_2) of $\sqrt{\frac{d\log n}{sn}}$. For these cases, the subset models give a much better resolvability than the complete models.

To achieve the rate $O\left(\sqrt{\frac{d\log n}{n}}\right)$ suggested by the resolvability of the sparse subset models, we use the following criterion to select a suitable subset. Choose the model $(\hat{j}, \hat{I}_{\hat{j}})$ minimizing

$$-\sum_{i=1}^n \log f_{I_j}\left(X_i, \hat{\theta}^{(I_j)}\right) + \frac{\lambda^*_{I_j} m_{I_j}}{n} + \frac{9.49\left(\log^* j + \log\left(j^d\right) + \log\left(\binom{j^d}{m_{I_j}}\right)\right)}{n},$$

where $\hat{\theta}^{(\hat{I}_j)}$ is the maximum likelihood estimator and $\lambda^*_{I_j} = \Lambda(A_{I_j})$. Denote $\hat{I}_{\hat{j}}$ by $\hat{I}$ and $\hat{\theta}^{(\hat{I}_j)}$ by $\hat{\theta}$ for short. The density estimator is then $f_{\hat{I},\hat{\theta}}$. By Theorem 4.1, we have the following conclusion.

Theorem 4.3: For the density estimator $\hat{f} = f_{\hat{I},\hat{\theta}}$, for any M_1, M_2, M_3, s, the density estimator $\hat{f}$ converges in squared Hellinger distance at a rate bounded above by $\sqrt{\frac{d \log n}{n}}$ uniformly for $f \in F(M_1, M_2, M_3, s)$. That is

$$\sup_{f \in F(M_1, M_2, M_3, s)} E d_H^2(f, \hat{f}) \leq \zeta(M_1, M_2, M_3, s) \cdot \sqrt{\frac{d \log n}{n}},$$

where the constant $\zeta(M_1, M_2, M_3, s)$ depend only on M_1, M_2, M_3, s, β_1 and β_2.

Note the model selection criterion does not depend on M_1, M_2, M_3, s. Therefore, the procedure is adaptive for the families $F(M_1, M_2, M_3, s)$, $M_1 > 0$, $M_2 > 0$, $M_3 > 0$, $s > 0$.

Remarks:

1. If we use the usual trigonometric basis, the condition (4.16) corresponds to the familiar smoothness condition on $\log f$ when s is an integer, namely, $\sum_{s_1+s_2+\ldots+s_d=s} \| \frac{\partial^s \log f}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}} \|_\mu^2 \leq \frac{M_3}{(2\pi)^{2s}}$, where μ is the Lebesgue measure on $[0,1]^d$. For the class of densities satisfying this condition, it is known that the optimal rate of convergence under the squared Hellinger distance is $\left(\frac{1}{n}\right)^{\frac{2s}{2s+d}}$. The condition (4.15) also controls the high frequency components of the Fourier representation of $\log f$, but it is not directly connected with the derivative condition on $\log f$.

2. The above analysis does not depend on any special properties of the tensor product basis. Therefore, the result applies to any multi-indexed orthonormal basis satisfying the two conditions (4.15) and (4.16).

The subset models considered here naturally correspond to the choices of the basis functions in the linear spaces to include in the models. The problem of estimating nonlinear parameters can also be changed into the problem of subset selection. In Subsection 4.4.2, we estimate linear and nonlinear parameters in the neural network models by the maximum likelihood principle. A different treatment is as follows. First suitably discretize the parameter spaces for the nonlinear parameters a and b. Treat $\phi(a^T x + b)$ as a basis function for all the discretized values of a and b. Then selecting the number of hidden layers and estimating the discretized values of the nonlinear parameters is equivalent to selecting the basis functions among exponentially many possibilities.

4.5 Proofs of the main Lemmas

Proof of Lemma 4.0: We use a "chaining" argument similar to that used in Birgé and Massart (1993, 1995).

We consider dividing the parameter space into rings as following:

$$\Theta_0 = \{\theta \in \Theta : d_H^2(f, f_\theta) \leq \frac{\xi}{n}\},$$

$$\Theta_i = \{\theta \in \Theta : \frac{2^{i-1}\xi}{n} \leq d_H^2(f, f_\theta) \leq \frac{2^i\xi}{n}\} \quad i = 1, 2, \ldots .$$

Then Θ_i is a Hellinger ring with inner radius r_{i-1}, outer radius r_i, where $r_i = 2^{\frac{i}{2}} r_0$ for $i \geq 0$, $r_{-1} = 0$, and $r_0 = \sqrt{\frac{\xi}{n}}$. We first concentrate on Θ_i.

Let a sequence $\delta_j \downarrow 0$ be given with $\delta_0 \leq \rho r_0$, then by the assumption, there is a sequence $F_0, F_1, F_2, \ldots$ of $\frac{\delta_0}{2}, \delta_1, \ldots$ nets in Θ_i satisfying the cardinality bounds. For each $\theta \in \Theta_i$, let $\tau_j(\theta) = \arg\min_{\theta' \in F_j} \| \log \frac{f_\theta}{f_{\theta'}} \|_\infty$ be the nearest representor of θ in net F_j. Denote

$$\ell_0(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \tau_0(\theta))}{f(X_i)},$$

$$\ell_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \tau_j(\theta))}{f(X_i, \tau_{j-1}(\theta))} .$$

Then because $\lim_{j\to\infty} f(x, \tau_j(\theta)) = f(x, \theta)$, so

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \theta)}{f(X_i)} = \ell_0(\theta) + \sum_{j=1}^{\infty} \ell_j(\theta).$$

Let $q_i = P^*\{$ for some $\theta \in \Theta_i, \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i)} \geq -\gamma d_H^2(f, f_\theta) + \frac{\xi}{n}\}$, then because $\sum_{j=1}^{\infty} E\ell_j(\theta) = -E \log \frac{f(X_1,\tau_0(\theta))}{f(X_1,\theta)}$, we have

$$\begin{aligned} q_i &= P^*\{\text{for some } \theta \in \Theta_i, \ell_0(\theta) + \textstyle\sum_{j=1}^{\infty}(\ell_j(\theta) - E\ell_j(\theta)) \\ &\quad \geq E \log \tfrac{f(X_1,\tau_0(\theta))}{f(X_1,\theta)} - \gamma d_H^2(f, f_\theta) + \tfrac{\xi}{n}\}. \end{aligned}$$

For $\theta_0 \in F_0$, consider $B_{\theta_0} =: \{\theta : \theta \in \Theta_i, \tau_0(\theta) = \theta_0\}$. For an arbitrary $\epsilon > 0$, choose $\tilde{\theta}_0 \in B_{\theta_0}$ satisfying

$$E \log \frac{f(X_1, \theta_0)}{f(X_1, \tilde{\theta}_0)} \leq \inf_{\theta \in B_{\theta_0}} E \log \frac{f(X_1, \theta_0)}{f(X_1, \theta)} + \epsilon .$$

Then let $\tilde{F}_0 = \{\tilde{\theta}_0 : \theta_0 \in F_0\}$. By triangle inequality, $\tilde{F}_0$ is a δ_0 net in Θ_i. Now replace F_0 by $\tilde{F}_0$ and accordingly replace τ_0 by $\tilde{\tau}_0$. For convenience, we will not distinguish $\tilde{\tau}_0$ from τ_0. Now notice for $\theta \in B_{\theta_0}$,

$$\begin{aligned} E \log \tfrac{f(X_1,\tau_0(\theta))}{f(X_1,\theta)} &= E \log \tfrac{f(X_1,\tau_0(\theta))}{f(X_1,\theta_0)} + E \log \tfrac{f(X_1,\theta_0)}{f(X_1,\theta)} \\ &\geq -\inf_{\theta' \in B_{\theta_0}} E \log \tfrac{f(X_1,\theta_0)}{f(X_1,\theta')} - \epsilon + E \log \tfrac{f(X_1,\theta_0)}{f(X_1,\theta)} \\ &\geq -\epsilon , \end{aligned}$$

so we have

$$\begin{aligned} q_i &\leq P^*\{\text{for some } \theta \in \Theta_i, \ell_0(\theta) + \textstyle\sum_{j=1}^{\infty}(\ell_j(\theta) - E\ell_j(\theta)) \\ &\quad \geq -\gamma d_H^2(f, f_0) + \tfrac{\xi}{n} - \epsilon\} \\ &\leq P^*\{\text{for some } \theta \in \Theta_i, \ell_0(\theta) \geq -2\gamma r_i^2 + \tfrac{\xi}{n} - \epsilon\} \\ &\quad + \textstyle\sum_{j=1}^{\infty} P\{\text{for some } \theta \in \Theta_i, \ \ell_j(\theta) - E\ell_j(\theta) \geq \eta_j\} \\ &=: q_i^{(1)} + \textstyle\sum_{j=1}^{\infty} q_{i,j}^{(2)}, \end{aligned}$$

where $\eta_j, j \geq 1$ are positive numbers satisfying

$$\sum_{j=1}^{\infty} \eta_j \leq \gamma r_i^2. \tag{4.17}$$

To bound $q_i^{(1)}$, we use a familiar exponential inequality as follows (see, e.g., Barron and Cover (1991), Chernoff (1952).

Fact: Let g_1 and g_2 be two probability density functions with respect to some σ-finite measure, then if $X_1, X_2, ..., X_n$ is an i.i.d. sample from g_2, we have that for every $t \in R$,

$$P\{\frac{1}{n}\sum_{i=1}^{n} \log \frac{g_1(X_i)}{g_2(X_i)} \geq t\} \leq e^{-\frac{n}{2}(d_H^2(g_1,g_2)+t)}.$$

From the above fact, we have that for each $\theta_0 \in F_0$,

$$\begin{aligned} & P\{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log \frac{f(X_i,\theta_0)}{f(X_i)} \geq -2\gamma r_i^2 + \tfrac{\xi}{n} - \epsilon\} \\ \leq\ & \exp(-\tfrac{n}{2}(d_H^2(f, f_{\theta_0}) - 2\gamma r_i^2 + \tfrac{\xi}{n} - \epsilon\} \\ \leq\ & \exp(-\tfrac{n}{2}(r_{i-1}^2 - 2\gamma r_i^2 + \tfrac{\xi}{n} - \epsilon\} \ . \end{aligned}$$

Note that for every $\theta_0 \in F_0$, $\ell_0(\theta)$ is the same for all $\theta \in B_{\theta_0}$. Thus by the union bound,

$$\begin{aligned} q_i^{(1)} &\leq P\left(\cup_{\theta_0 \in F_0}\{\ell_0(\theta_0) \geq -2\gamma r_i^2 + \tfrac{\xi}{n} - \epsilon\}\right) \\ &\leq \text{card}(F_0) \exp\left(-\tfrac{n}{2}(r_{i-1}^2 - 2\gamma r_i^2 + \tfrac{\xi}{n} - \epsilon)\right) \ . \end{aligned}$$

Because $\epsilon > 0$ is arbitrary, we know

$$q_i \leq \text{card}(F_0) \exp\left(-\frac{n}{2}(r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n})\right) + \sum_{j=1}^{\infty} q_{i,j}^{(2)} \ .$$

Note for $i \geq 1$,

$$r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} \geq \frac{2^{i-1}\xi}{n} - \frac{2\gamma 2^i \xi}{n} + \frac{\xi}{n} \geq (i+1)(1-4\gamma)\frac{\xi}{n} \ ,$$

and for $i = 0$,

$$r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} = (1-2\gamma)\frac{\xi}{n} \geq (1-4\gamma)\frac{\xi}{n},$$

so

$$q_i \leq \text{card}(F_0) \exp\left(-\frac{(i+1)(1-4\gamma)\xi}{2}\right) + \sum_{j=1}^{\infty} q_{i,j}^{(2)} \ .$$

Now because

$$\begin{aligned} \|\log \tfrac{f(\cdot,\tau_j(\theta))}{f(\cdot,\tau_{j-1}(\theta))}\|_\infty &\leq \|\log \tfrac{f(\cdot,\tau_j(\theta))}{f(\cdot,\theta)}\|_\infty + \|\log \tfrac{f(\cdot,\theta)}{f(\cdot,\tau_{j-1}(\theta))}\|_\infty \\ &\leq \delta_{j-1} + \delta_j \\ &\leq 2\delta_{j-1} \ , \end{aligned}$$

we have

$$\|\log \frac{f(\cdot,\tau_j(\theta))}{f(\cdot,\tau_{j-1}(\theta))} - E\log \frac{f(\cdot,\tau_j(\theta)}{f(\cdot,\tau_{j-1}(\theta))}\|_\infty \leq 4\delta_{j-1} \ .$$

Observe that $\ell_j(\theta)$ is the same for all θ such that $(\tau_{j-1}(\theta),\tau_j(\theta)) = (\theta_{j-1},\theta_j)$, for any pair $(\theta_{j-1},\theta_j) \in F_{j-1} \times F_j$, together with Hoeffding's inequality (see, e.g., Pollard (1984, pp. 191-192)), we get

$$\begin{aligned} \textstyle\sum_{j=1}^\infty q_{i,j}^{(2)} &\leq \textstyle\sum_{j=1}^\infty \mathrm{card}(F_j)\cdot\mathrm{card}(F_{j-1})\exp(-\frac{n\eta_j^2}{8\delta_{j-1}^2}) \\ &\leq \left(\tfrac{Ar_i}{\delta_0/2}\right)^m \left(\tfrac{Ar_i}{\delta_1}\right)^m \exp(-\tfrac{n\eta_1^2}{8\delta_0^2}) + \textstyle\sum_{j=2}^\infty \left(\tfrac{Ar_i}{\delta_j}\right)^m \left(\tfrac{Ar_i}{\delta_{j-1}}\right)^m \exp(-\tfrac{n\eta_j^2}{8\delta_{j-1}^2}) \ . \end{aligned}$$

Given ξ,γ,A,m,n, we choose the sequence δ_j,η_j as follows. First, δ_0 is chosen such that

$$\log\left(\frac{Ar_0}{\delta_0/2}\right)^m = \frac{(1-4\gamma)\xi}{4} \ .$$

Similarly each $\delta_j, j \geq 1$ is chosen such that

$$\log\left(\frac{Ar_0}{\delta_j}\right)^m = \frac{(j+1)(1-4\gamma)\xi}{4} \ ,$$

and $\eta_j, j \geq 1$ is defined such that

$$\frac{n\eta_j^2}{8\delta_{j-1}^2} = (\log 2)mi + \frac{(2j+1)(1-4\gamma)\xi}{4} + \frac{(i+1)j(1-4\gamma)\xi}{8} \ .$$

With these choices, the bound on q_i becomes

$$\begin{aligned} q_i &\leq \exp\left(m\log\tfrac{A2^{\frac{i}{2}}r_0}{\delta_0/2} - \tfrac{(i+1)(1-4\gamma)\xi}{2}\right) + \exp\left(m\log\tfrac{A2^{\frac{i}{2}}r_0}{\delta_0/2} + m\log\tfrac{A2^{\frac{i}{2}}r_0}{\delta_1} - \tfrac{n\eta_1^2}{8\delta_0^2}\right) \\ &\quad + \textstyle\sum_{j=2}^\infty \exp\left(m\log\tfrac{A2^{\frac{i}{2}}r_0}{\delta_j} + m\log\tfrac{A2^{\frac{i}{2}}r_0}{\delta_{j-1}} - \tfrac{n\eta_j^2}{8\delta_{j-1}^2}\right) \\ &\leq \exp\left(\tfrac{\log 2}{2}mi - \tfrac{(i+1)(1-4\gamma)\xi}{4}\right) + \exp\left(-\tfrac{(i+1)(1-4\gamma)\xi}{8}\right) \\ &\quad + \textstyle\sum_{j=2}^\infty \exp\left(-\tfrac{(i+1)j(1-4\gamma)\xi}{8}\right) \\ &\leq \left(1 + \tfrac{1}{1-\exp(-\frac{(1-4\gamma)\xi}{8})}\right)\exp\left(-\tfrac{(i+1)(1-4\gamma)\xi}{8}\right) \\ &\leq \left(1+\tfrac{\sqrt{2}}{\sqrt{2}-1}\right)\exp\left(-\tfrac{(i+1)(1-4\gamma)\xi}{8}\right). \end{aligned}$$

For the third inequality, we need $\left(\frac{\log 2}{2}\right)mi \leq \frac{(i+1)(1-4\gamma)\xi}{8}$, which is satisfied if $\frac{\xi}{m} \geq \frac{4}{1-4\gamma}\log\frac{2A}{\rho}$ with $\rho \leq A$. The last inequality follows from $\frac{(i+1)(1-4\gamma)\xi}{8} \geq \frac{\log 2}{2}$.

From our choices of $\delta_0, \delta_j, \eta_j$, it follows that

$$\delta_0 = 2Ar_0 \exp(-\frac{(1-4\gamma)\xi}{4m}), \tag{4.18}$$

$$\delta_j = Ar_0 \exp(-\frac{(j+1)(1-4\gamma)\xi}{4m}) \quad \text{for} \quad j \geq 1,$$

$$\eta_1 = 2A\sqrt{1-4\gamma}\sqrt{3i+9}\frac{\xi}{n}\exp(-\frac{(1-4\gamma)\xi}{4m}),$$

and for $j \geq 2$,

$$\begin{aligned}
\eta_j &= \delta_{j-1}\sqrt{8(\log 2)\tfrac{mi}{n} + \tfrac{2(2j+1)(1-4\gamma)\xi}{n} + \tfrac{(i+1)j(1-4\gamma)\xi}{n}} \\
&\leq \delta_{j-1}\sqrt{\tfrac{2(i+1)(1-4\gamma)\xi}{n} + \tfrac{2(2j+1)(1-4\gamma)\xi}{n} + \tfrac{(i+1)j(1-4\gamma)\xi}{n}} \\
&\leq A\sqrt{1-4\gamma}\sqrt{2i+5j+ij+4}\tfrac{\xi}{n}\exp(-\tfrac{j(1-4\gamma)\xi}{4m}) \\
&\leq A\sqrt{1-4\gamma}\sqrt{i+5}\sqrt{j+2}\tfrac{\xi}{n}\exp(-\tfrac{j(1-4\gamma)\xi}{4m}) \\
&\leq A\sqrt{1-4\gamma}\sqrt{i+5}\tfrac{\xi}{n}\exp(\tfrac{1}{2}(j+1) - \tfrac{j(1-4\gamma)\xi}{4m}) \\
&\leq A\sqrt{1-4\gamma}\sqrt{i+5}\tfrac{\xi}{n}\exp(-\tfrac{j(1-4\gamma)\xi}{8m}) \ .
\end{aligned}$$

It remains to check whether $\delta_0 \leq \rho r_0$ and whether $\sum_{j=1}^{\infty}\eta_j \leq \gamma r_i^2$ as required in (4.17). Indeed,

$$\begin{aligned}
\textstyle\sum_{j=1}^{\infty}\eta_j &\leq 2A\sqrt{1-4\gamma}\sqrt{3i+9}\tfrac{\xi}{n}\exp(-\tfrac{(1-4\gamma)\xi}{4m}) \\
&\quad + A\sqrt{1-4\gamma}\sqrt{i+5}\tfrac{\xi}{n}\tfrac{\exp(-\frac{2(1-4\gamma)\xi}{8m})}{1-\exp(-\frac{(1-4\gamma)\xi}{8m})} \\
&\leq A\sqrt{1-4\gamma}\sqrt{i+5}\tfrac{\xi}{n}\exp(-\tfrac{(1-4\gamma)\xi}{4m})\left(2\sqrt{3} + \tfrac{1}{1-\exp(-\frac{(1-4\gamma)\xi}{8m})}\right) \\
&\leq \left(2\sqrt{3} + \tfrac{\sqrt{2}}{\sqrt{2}-1}\right)A\sqrt{1-4\gamma}\sqrt{i+5}\tfrac{\xi}{n}\exp(-\tfrac{(1-4\gamma)\xi}{4m}) \\
&\leq 6.88A\sqrt{1-4\gamma}\sqrt{i+5}\tfrac{\xi}{n}\exp(-\tfrac{(1-4\gamma)\xi}{4m}) \ .
\end{aligned}$$

Thus, for $\sum_{j=1}^{\infty}\eta_j \leq \gamma r_i^2$ to hold, it suffices to have

$$6.88A\sqrt{1-4\gamma}\sqrt{i+5}\frac{\xi}{n}\exp(-\frac{(1-4\gamma)\xi}{4m}) \leq \gamma r_i^2 = \gamma 2^i\frac{\xi}{n} \ .$$

Using $\frac{2^i}{\sqrt{i+5}} \geq \frac{1}{\sqrt{5}}$ for $i \geq 0$, it is enough to require

$$\frac{\xi}{m} \geq \frac{4}{1-4\gamma}\log\left(\frac{15.4A}{\gamma}\sqrt{1-4\gamma}\right) = \frac{4}{1-4\gamma}\log\frac{2A}{\rho}, \tag{4.19}$$

where $\rho = \frac{2\gamma}{15.4\sqrt{1-4\gamma}}$.

Finally we sum over the rings indexed by i,

$$\begin{aligned}
&P^*\{\text{for some } \theta \in \Theta, \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log\frac{f(X_i,\theta)}{f(X_i)} \geq -\gamma d_H^2(f,f_0) + \frac{\xi}{n}\} \\
&\leq \textstyle\sum_{i=0}^{\infty} P^*\{\text{for some } \theta \in \Theta_i, \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i,\theta)}{f(X_i)} \geq -\gamma d_H^2(f,f_0) + \frac{\xi}{n}\} \\
&\leq \textstyle\sum_{i=0}^{\infty}(1 + \frac{\sqrt{2}}{\sqrt{2}-1})\exp\left(-\frac{(i+1)(1-4\gamma)\xi}{8}\right) \\
&\leq (1 + \tfrac{\sqrt{2}}{\sqrt{2}-1})\tfrac{\exp(-\frac{(1-4\gamma)\xi}{8})}{1-\exp(-\frac{(1-4\gamma)\xi}{8})} \\
&\leq 15.1\exp(-\tfrac{(1-4\gamma)\xi}{8}) \ .
\end{aligned}$$

From (4.18) and (4.19), $\frac{\delta_0}{r_0} = 2Ae^{-\frac{(1-4\gamma)\xi}{4m}} \leq 2A \times \frac{\rho}{2A} = \rho$ as required. This completes the proof of the lemma.

Proof of Lemma 4.2: We first show the Hellinger ball is contained in some square-norm ball. Then for the square-norm ball, we provide a suitable δ-net satisfying the cardinality bound.

Because $1 \in S_j$, so $1 = \sum_{i=1}^{m_j} \eta_i \varphi_{j,i}(x)$ for some $\eta \in R^{m_j}$. Then the log-density may be written as

$$\log f_j(x,\theta) = \sum_{i=1}^{m_j} \beta_i \varphi_{j,i}(x) \ ,$$

where $\beta_i = \theta_i - \psi_j(\theta)\eta_i$. Because for $\theta \in \Theta_{j,L}$, $\| \log f_j(x,\theta) \|_\infty \leq L$, it follows that for any $\theta, \theta^* \in \Theta_{j,L}$, $\frac{f_j(x,\theta)}{f_j(x,\theta^*)} \leq e^{2L}$. Let $M_L = \frac{\phi_1(e^{2L})}{\phi_2(e^{2L})} e^{-L}$, from Lemma 4.5 in Section 6,

$$\begin{aligned} d_H^2(f_{\theta^*}, f_\theta) &\geq \tfrac{\phi_1(e^{2L})}{\phi_2(e^{2L})} \int f_{\theta^*}(\log f_\theta - \log f_{\theta^*})^2 d\mu \\ &\geq M_L \int (\log f_\theta - \log f_{\theta^*})^2 dx \\ &\geq \tfrac{M_L T_2^2}{m_j} \textstyle\sum_{i=1}^{m_j} (\beta_i - \beta_i^*)^2. \end{aligned}$$

For the last inequality, we use the frame assumption in (4.8). Therefore, for any $\theta^* \in \Theta_{j,L}$,

$$B_j(\theta^*, r) \subset \tilde{B}_j\left(\beta^*, \sqrt{\frac{m_j r}{M_L T_2^2}}\right) = \{\beta : \beta \in R^{m_j}, \| \beta - \beta^* \|^2 \leq \frac{m_j r^2}{M_L T_2^2}\}.$$

The inclusion above refers to the functions represented by the parameters θ and β. Now we want to find a suitable δ-net on $\tilde{B}(\beta^*, \sqrt{\frac{m_j r^2}{M_L T_2^2}})$. We consider a rectangular grid spaced at width $\varepsilon > 0$ for each coordinate. If β belongs to a cube with at least one element $\tilde{\beta}$ corresponding to $\tilde{\theta} \in B_j(\theta^*, r)$, then

$$\| \beta - \beta^* \|^2 \leq 2 \| \tilde{\beta} - \beta^* \|^2 + 2 \| \beta - \tilde{\beta} \|^2 \leq \frac{2m_j r^2}{M_L T_2^2} + 2m_j \varepsilon^2 \ .$$

Thus, all the cubes with at least one element in $B_j(\theta^*, r)$ are included in $\tilde{B}_j(\beta^*, \overline{r})$ where $\overline{r} = \sqrt{\frac{2m_j r^2}{M_L T_2^2} + 2m_j \varepsilon^2}$. Therefore, the number of these cubes is bounded by

$$\frac{\mathrm{Vol}(\tilde{B}(\beta^*, \overline{r}))}{\varepsilon^{m_j}} = \frac{(\sqrt{\pi})^{m_j}\, \overline{r}^{m_j}}{\Gamma(\frac{m_j}{2}+1)\varepsilon^{m_j}} \leq \frac{1}{\sqrt{m_j \pi}} \left(\frac{\sqrt{2\pi e}\,\overline{r}}{\sqrt{m_j}\delta}\right)^{m_j} .$$

From (4.7), for any β and $\tilde{\beta}$ corresponding to θ and $\tilde{\theta}$ respectively in the same cube, we have

$$\| \log f_\theta - \log f_{\tilde{\theta}} \|_\infty \leq \max_{1 \leq i \leq m_j} |\beta_i - \tilde{\beta}_i| \leq T_1 \varepsilon.$$

Take $\varepsilon = \frac{\delta}{T_1}$, then $\| \log f_\theta - \log f_{\tilde{\theta}} \|_\infty \leq \delta$. For $\delta < \rho r$, $\bar{r} \leq r\sqrt{\frac{2m_jT_1^2}{M_LT_2^2} + 2\rho^2}$.

Now, for each cube that intersects with $\tilde{B}_j(\beta^*, \bar{r})$, choose a parameter β that corresponds to a probability density function and let F_δ be the collection of the corresponding densities. Then

$$|F_\delta| \leq \frac{1}{\sqrt{m_j\pi}} \left(\frac{\sqrt{2\pi e^2(\frac{2T_1^2}{M_LT_2^2} + 2\rho^2)}r}{\delta} \right)^{m_j} .$$

Clearly, F_δ is a δ-net for densities in $B_j(\theta^*, r)$. Thus Assumption 4.1$'$ is satisfied with $A_k = \sqrt{2\pi e^2(\frac{2T_1^2}{M_LT_2^2} + 2\rho^2)}$. From Lemma 4.5, $M_L \geq \frac{1}{4(1+L)^2e^L}$, so $A_k \leq \sqrt{2\pi e^2 \cdot \frac{2T_1^2}{M_LT_2^2}} + \sqrt{2\pi e^2 \cdot 2\rho^2} \leq 19.28\frac{T_1}{T_2}(1+L)e^{\frac{L}{2}} + 0.06$.

Proof of Lemma 4.3: We consider an orthonormal basis 1, $\varphi_{j,1}(x), \varphi_{j,2}(x)$, ..., $\varphi_{j,m_j}(x)$ in S_j. Let $\beta = (\theta_1, \theta_2, ..., \theta_{m_j}, \psi_j(\theta))$. From the proof of Lemma 4.2, we know that for any $\theta, \theta^* \in \Theta_{j,L}$,

$$d_H^2(f_{\theta^*}, f_\theta) \geq M_L \int (\log f_\theta - \log f_{\theta^*})^2 dx = M_L \| \beta \|^2 .$$

Therefore

$$B_j(\theta^*, r) \subset \tilde{B}_j(\beta^*, \sqrt{\frac{1}{M_L}}r) = \{\beta : \beta \in R^{m_j+1}, \| \beta - \beta^* \|^2 \leq \frac{1}{M_L}r^2\}.$$

The inclusion above is meant for the functions that the parameters represent. Similarly to the counting argument in the proof of Lemma 4.2, a rectangular grid spaced at width $\frac{\delta}{K_j\sqrt{m_j+1}}$ for each coordinate provides the desired δ-net. The cardinality constant $A_{(j,L)} = \sqrt{2\pi e^2(\frac{2K_j^2}{M_L} + 2\rho^2)} \leq 19.28K_j(1+L)e^{\frac{L}{2}} + 0.06$ for $\rho = 0.0056$. This completes the proof Lemma 4.3.

4.6 Some simple inequalities used for main results

Lemma 4.4: Assume f and g are two probability density functions with respect to some σ-finite measure μ. Let $s > 1$ be any constant, then

$$\int_{\{\frac{f}{g} \geq s\}} f \log \frac{f}{g} d\mu \leq \alpha(s) D(f \| g)$$

where $\alpha(s) = \frac{\log s}{\log s + \frac{1}{s} - 1}$. Also $\alpha(s)$ is decreasing in s for $s > 1$.

Remark: The best available bound with $s = 1$ is

$$\int_{\{\frac{f}{g} \geq 1\}} f \log \frac{f}{g} d\mu \leq D(f \parallel g) + \sqrt{2D(f \parallel g)}.$$

Here we avoid the square root with $s > 1$. Note $\alpha(s) \to 1$ as $s \to \infty$. Improved bounds of the form $O(\frac{c}{s^2} D(f \parallel g))$ are possible under the condition $\mathrm{var}\left(\log \frac{f}{g}\right) \leq cD(f \parallel g)$. Here we have chosen to avoid higher order moment conditions on the logarithm of the density ratio. Hence no uniform tail rate of convergence to zero exists.

Proof of Lemma 4.4: We consider a familiar expression of the relative entropy:

$$\begin{aligned} D(f \parallel g) &= \int f \log \tfrac{f}{g} d\mu \\ &= \int f (\log \tfrac{f}{g} + \tfrac{g}{f} - 1) d\mu \\ &= \int_{\{\frac{f}{g} \geq s\}} f (\log \tfrac{f}{g} + \tfrac{g}{f} - 1) d\mu + \int_{\{\frac{f}{g} < s\}} f (\log \tfrac{f}{g} + \tfrac{g}{f} - 1) d\mu \ . \end{aligned}$$

Because $(\log \frac{f}{g} + \frac{g}{f} - 1) \geq 0$, to proof the lemma, it suffices to show

$$\log \frac{f}{g} \leq \alpha(s)(\log \frac{f}{g} + \frac{g}{f} - 1) \quad \text{for } \frac{f}{g} \geq s \ .$$

This follows from the monotonicity of $\alpha(s)$, which can be shown from simple calculation. This completes the proof of the lemma.

Lemma 4.5: Let p and q be two probability density functions with respect to some σ-finite measure μ. If $\frac{p(x)}{q(x)} \leq V$ 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 $\phi_1(V) = \frac{\log V + \frac{1}{V} - 1}{\log^2 V} \geq \frac{1}{2 + \log V}$ and $\phi_2(V) = \frac{V \log V + 1 - V}{(\sqrt{V} - 1)^2} \leq (2 + \log V)$.

The above upper bound on the relative entropy is given in Birgé & Massart (1994, Lemma 4).

Proof of Lemma 4.5: We note $D(p \parallel q) = \int p \left(\log \frac{p}{q} + \frac{q}{p} - 1\right) d\mu$. It can be shown from calculus that $\phi_1(x) = \frac{\log x + \frac{1}{x} - 1}{\log^2 x}$ is decreasing on $(0, \infty)$, which implies

$$\frac{\log V + \frac{1}{V} - 1}{\log^2 V} \int p \left(\log \frac{p}{q}\right)^2 d\mu \leq D(p \parallel q) \ .$$

To prove the other inequality, we consider the following parts of $D(p \parallel q)$ and $d_H^2(p,q)$.

$$D(p \parallel q) = \int_{\{q>p\}} p\left(\log\frac{p}{q} + \frac{q}{p} - 1\right) d\mu + \int_{\{q<p\}} q\left(\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}\right) d\mu$$

$$d_H^2(p,q) = \int_{\{q>p\}} p\left(\sqrt{\frac{q}{p}} - 1\right)^2 d\mu + \int_{\{q<p\}} q\left(\sqrt{\frac{p}{q}} - 1\right)^2 d\mu \ .$$

For $p < q$, $\log\frac{p}{q} + \frac{q}{p} - 1 \le 2\left(\sqrt{\frac{q}{p}} - 1\right)^2$, so

$$\int_{\{q>p\}} p\left(\log\frac{p}{q} + \frac{q}{p} - 1\right) d\mu \le 2\int_{\{q>p\}} p\left(\sqrt{\frac{q}{p}} - 1\right)^2 d\mu \ .$$

For $p > q$, $\phi_2(\frac{p}{q}) = \frac{\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}}{\left(\sqrt{\frac{p}{q}} - 1\right)^2}$ is increasing in $\frac{p}{q}$. It follows that

$$\int_{\{q<p\}} q\left(\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}\right) d\mu \le \frac{\log V + \frac{1}{V} - 1}{\log^2 V}\int_{\{q<p\}} q\left(\sqrt{\frac{p}{q}} - 1\right)^2 d\mu \ .$$

Combining the integrals together, we conclude

$$D(p \parallel q) \le \frac{V\log V + 1 - V}{(\sqrt{V} - 1)^2} d_H^2(p,q) \ ,$$

which completes the proof of Lemma 4.5.

Lemma 4.6: Suppose h_1 and h_2 are two functions on [0,1] satisfying $\int e^{h_1} d\mu < \infty$, $\int e^{h_2} d\mu < \infty$, where μ is the Lebesgue measure. Then

$$|\log\int e^{h_1} d\mu - \log\int e^{h_2} d\mu| \le \| h_1 - h_2 \|_\infty \ .$$

Proof: By Jensen's inequality.

$$\begin{aligned}
\log\int e^{h_1} d\mu - \log\int e^{h_2} d\mu &= \log\int \frac{e^{(h_1-h_2)+h_2}}{\int e^{h_2} d\mu} d\mu \\
&\ge \int \log(e^{h_1-h_2})\frac{e^{h_2}}{\int e^{h_2} d\mu} d\mu \\
&\ge -\| h_1 - h_2 \|_\infty \ .
\end{aligned}$$

Similarly,

$$\log\int e^{h_1} d\mu - \log\int e^{h_2} d\mu \le \| h_1 - h_2 \|_\infty,$$

which completes the proof.

Bibliography

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Info. Theory*, pp. 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest, 1973.

[2] A.R. Barron, "Are Bayes rules consistent in information?" *Open Problems in Communication and Computation*, pp. 85-91. T. M. Cover and B. Gopinath editors, Spinger-Verlag, 1987.

[3] A.R. Barron and T.M. Cover, "Minimum complexity density estimation, " *IEEE. Trans. on Information Theory*, vol. 37, no. 4, pp. 1034-1054, 1991.

[4] A.R. Barron and C.-H Sheu, "Approximation of density function by sequences of exponential families," *Ann. Statistics*, vol. 19, pp. 1347-1369, 1991.

[5] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE, Trans. on Information Theory*, vol. 39, no. 3, pp. 930-945, 1993.

[6] A.R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115-133, 1994.

[7] A.R. Barron, Y. Yang and B. Yu, "Asymptotically optimal function estimation by minimum complexity criteria," in *Proc. 1994 Int. Symp. Info. Theory*, p. 38, Trondheim, Norway, 1994.

[8] A.R. Barron and N. Hengartner, "Information theory and superefficiency," preprint, 1995.

[9] J.M. Bernardo, "Reference prior distributions for Bayesian inference," *J. Roy. Statist. Soc.* Ser. B, vol. 41, pp. 113-147, 1979.

[10] P.J. Bickel and Y. Ritov, "Estimating integrated squared density derivatives: sharp best order of convergence estimates," *Sankhyā: Indian J. Statist.* Ser. A, vol. 50, pp. 381-393, 1988.

[11] L. Birgé, "Approximation dans les espaces metriques et theorie de l'estimation," *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, vol. 65, pp. 181-237, 1983.

[12] L. Birgé, "On estimating a density using Hellinger distance and some other strange facts," *Probab. Th. Rel. Fields,* vol. 71 pp. 271-291, 1986.

[13] L. Birgé and P. Massart, "Estimation of integral functionals of a density," Technical Report 024-92, Mathematical Sciences Research Institute, Berkeley, 1992.

[14] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimators," *Probability Theory and Related Fields*, vol. 97, pp. 113-150, 1993.

[15] L. Birgé and P. Massart, "Minimum contrast estimators on sieves," Technical report, Universite Paris-Sud, 1994.

[16] L. Birgé and P. Massart, "From model selection to adaptive estimation," *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam*, 1995.

[17] J. Bretagnolle and C. Huber, "Estimation des densites: risque minimax," *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, vol. 47, pp. 119-137, 1979.

[18] N.N. Cencov, *Statistical Decision Rules and Optimal Inference*, Amer. Math. Soc. Transl., vol 53, Providence, RI, 1982.

[19] C.K. Chui, *An Introduction to Wavelets*, Academic Press, Inc., 1991.

[20] B. Clarke and A.R. Barron, "Jeffrey's prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference*, vol. 41, pp. 37-40, 1994.

[21] B. Clarke and A.R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Info. Theory*, vol. 36, pp. 453-471, 1990.

[22] G.F. Clements, "Entropy of several sets of real valued functions," *Pacific J. Math.*, vol. 13, pp. 1085-1095, 1963.

[23] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.

[24] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493-507, 1952.

[25] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, pp. 377-403, 1979.

[26] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[27] L. Davisson, "Universal noiseless coding," *IEEE Trans. Info. Theory*, vol. 19, pp. 783-795, 1973.

[28] C. de Boor, *A practical Guide to Splines*, Springer-Verlag New York, 1978.

[29] C. de Boor and G.J. Fix, "Spline approximation by quasiinterpolents," *J. Approx. Theory*, vol. 8, pp. 19-45, 1973.

[30] L. Devroye, *A Course in Density Estimation*, Birkauser, Boston, 1987.

[31] D.L. Donoho and R.C. Liu, "Geometrizing rates of convergence, II," *Ann. Statistics*, vol. 19, pp. 633-667, 1991.

[32] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," preprint, 1993.

[33] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage, Asymptopia?," *J. R. Statist. Soc. B*, vol. 57, pp. 301-369, 1995.

[34] S.Yu. Efroimovich and M.S. Pinsker, "Estimation of square-integrable probability density of a random variable," *Problemy Peredachi Informatsii*, vol. 18, pp. 19-38, 1982.

[35] S.Yu. Efroimovich, "Nonparametric estimation of a density of unknown smoothness," *Theory probab. Appl.*, vol. 30, pp. 557-568, 1985.

[36] W. Härdle, P. Hall, and S. Marron, "How far are automatically chosen regression smoothing parameters from their optimum," Mimeo Services #1589, Dept. Stat., North Carolina State Univer.-Chapel Hill, 1985.

[37] R.Z. Hasminskii, "A lower bound on the risks of nonparametric estimates of densities in the uniform metric," *Theory probab. Appl.*, vol. 23, pp. 794-796, 1978.

[38] R.Z. Hasminskii and I.A. Ibragimov, "On density estimation in the view of Kolmogorov's ideas in approximation theory," *Ann. Statist.*, vol. 18, pp. 999-1010, 1990.

[39] D. Haughton, "Size of the error in the choice of a model to fit data from an exponential family," *Sankhyā: Indian J. Statist.* Ser. A, vol. 51, pp. 45-58, 1989.

[40] D. Haussler and M. Opper, "General bounds on the mutual information between a parameter and n conditionally independent observations," preprint, 1995.

[41] I.A. Ibragimov and R.Z. Hasminskii, "Estimation of distribution density," *Zap. Nauchn. Semin. LOMI* vol. 98, pp. 61-85, 1980.

[42] I.A. Ibragimov and R.Z. Hasminskii, "Bounds for the risks of non-parametric regression estimates," *Theory probab. Appl.* vol. 27, pp. 84-99, 1982.

[43] A.N. Kolmogorov and V.M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Uspehi Mat. Nauk* vol. 14, pp. 3-86, 1959; English transl., *Amer. Math. Soc. Thansl.* vol. 17, pp. 277-364, 1961.

[44] L.M. Le Cam, "Convergence of estimates under dimensionality restrictions," *Ann. Statististics*, vol. 1, pp. 38-53, 1973.

[45] K.C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set," *Ann. Statistics*, vol. 15, no. 3, pp. 958-975, 1987.

[46] G.G. Lorentz, "Metric entroy and approximation," *Bull. Amer. Math. Soci.* vol. 72, pp. 903-937, 1966.

[47] D.S. Modha and E. Masry, "Rates of convergence in density estimation using neural networks," manuscript, 1994.

[48] A. Nemirovskii, "Nonparametric estimation of smooth regression functions," *J. Comput. Syst. Sci.* vol. 23, no. 6, pp. 1-11, 1986.

[49] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.

[50] D. Pollard, "Hypercubes and minimax rates of convergence," preprint, 1993.

[51] S. Portnoy, "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity," *Ann. Stat.*, vol. 16, pp. 356-366, 1988.

[52] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. on Information Theory*, vol. 30, pp. 629-636, 1984.

[53] J. Rissanen, "Stochastic complexity and modeling," *Ann. Stat.*, vol. 14, pp. 1080-1100, 1986.

[54] J. Rissanen, T. Speed, and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. Info. Theory,* vol. 38, pp. 315-323, 1992.

[55] X. Shen and W.H. Wong, "Convergence rates of sieve estimates," *Ann. Statistics*, vol. 22, No. 2, pp. 580-615, 1994.

[56] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, pp. 45-54, 1981.

[57] G. Shwartz, "Estimating the dimension of a model," *Ann. Statistics*, vol. 6, pp. 461-464, 1978.

[58] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[59] T.P. Speed and B. Yu, "Model selection and prediction: Normal regression," *Ann. Inst. Stat. Math.*, vol. 45, pp. 35-54, 1993.

[60] C.J. Stone, "Optimal global rates of convergence for nonparametric regression," *Ann. Statistics*, vol. 10, No. 4, pp. 1040-1053, 1982.

[61] C.J. Stone, "The dimensionality reduction principle for generalized additive models," *Ann. Statistics*, vol. 14, No. 2, pp. 590-606, 1986.

[62] C.J. Stone, " Large-sample inference for log-spline models," Ann. Statistics, vol. 18, pp. 717-741, 1990.

[63] C.J. Stone, "The use of polynomial splines and their tensor products in multivariate function estimation," *Ann. Statistics*, vol. 22, No. 1, pp. 118-184, 1994.

[64] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Probability*, vol. 22, pp. 28-76, 1994.

[65] S. van de Geer, "Estimating a regression function," *Ann. Statistics*, vol. 18, No. 2, pp. 907-924, 1990.

[66] W.H. Wong and X. Shen, "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs," *Ann. Statistics*, vol. 23, No. 2, pp. 339-362, 1995.

[67] Y. Yang, "Complexity-based model selection, " prospectus submitted to Department of Statistics, Yale University, 1993.

[68] Y. Yang, "An asymptotic property of model selection criteria, " in *Proc. 1994 IEEE-IMS Workshop on Info. Theory and Stat.*, p. 103, Alexandria, Virginia, 1994.

[69] B. Yu, "Assouad, Fano, and Le Cam," To appear in *Festschrift in honor of L. Le Cam on his 70th birthday*, 1995.

[70] B. Yu, "Lower bounds on expected redundancy for non-parametric classes," *IEEE Trans. Info. Theory*, vol. 42, 1996.

[71] B. Yu and T. Speed, "Data compression and histogram," *Probability Theory and Related Fields*, vol. 92, pp. 195-229, 1992.