An Asymptotic Property of Model Selection Criteria

Yuhong Yang and Andrew R. Barron, *Member, IEEE*

***Abstract*—Probability models are estimated by use of penalized log-likelihood criteria related to AIC and MDL. The accuracies of the density estimators are shown to be related to the tradeoff between three terms: the accuracy of approximation, the model dimension, and the descriptive complexity of the model classes. The asymptotic risk is determined under conditions on the penalty term, and is shown to be minimax optimal for some cases. As an application, we show that the optimal rate of convergence is simultaneously achieved for log-densities in Sobolev spaces $W_2^s(U)$ without knowing the smoothness parameter s and norm parameter U in advance. Applications to neural network models and sparse density function estimation are also provided.**

***Index Terms*— Complexity penalty, convergence rate, model selection, nonparametric density estimation, resolvability.**

I. Introduction

We consider the estimation of an unknown probability density function $f(x)$ defined on a measurable space S with respect to some σ-finite measure μ. Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed (i.i.d.) sample drawn according to $f(x)$. To estimate f, a sequence of finite-dimensional density families $\{f_k(x, \theta^{(k)}),\ \theta^{(k)} \in \Theta_k\}$ are suggested to approximate the true density. For example, one might approximate the logarithm of the density function by a basis function expansion using polynomial, trigonometric, wavelet, or spline series. For a given model k, we consider the maximum-likelihood estimator $\hat{\theta}^{(k)}$ of $\theta^{(k)}$. Then a model $\hat{k}$ is selected by optimizing a penalized log-likelihood criterion. As we discuss below, there are a number of criteria of this type including those proposed by Akaike [1], Rissanen [34], Schwartz [37], and others, where the penalty involves the parameter dimension and/or the model complexity. Here we are interested in understanding the accuracy of the density estimate $\hat{f}(x) = f_{\hat{k}}(x, \hat{\theta}^{(\hat{k})})$. As in Barron and Cover [4], the accuracy can be related to an index of resolvability expressing the tradeoff between the error of approximation as measured by the relative entropy distance between f and the family f_k and the complexity relative to the sample size. However, the work in [4] is restricted to criteria with a description length interpretation. Here we consider more general penalized likelihood criteria. We relate the risk of the density estimator to an accuracy index (or index of resolvability) expressing the tradeoff between the relative entropy distance and the dimension of the parametric families. The penalty term is proportional to the dimension of the models (in some cases without a logarithmic factor) thereby permitting an improved accuracy index and smaller corresponding statistical risk in some cases.

The present paper presents two types of results. First resolvability bounds on the statistical risk of penalized likelihood density estimates are given that capture the tradeoff discussed above. These bounds are valid for each density f and each sample size. Secondly, we show how such bounds provide a tool for demonstrating nonparametric adaptation properties, specifically minimax optimal convergence simultaneously for multiple function classes. We do not attempt to cover all cases that may be of interest, rather we give representative examples involving adaptation to unknown order of smoothness and norm in L_2 Sobolev spaces by spline selection and adaptation to function classes that represent sparseness by Fourier and neural net methods. In the remainder of this section, we review model selection criteria for function estimation, we discuss the form of the criteria studied here and the separate roles of parameter dimension and model complexity in these criteria, we review issues of log-density estimation and sieve estimation, and we discuss other work on adaptive estimation.

Sequences of exponential models are previously considered for density estimation by Barron and Sheu [3], Cencov [14], Portnoy [33], and many others (for a detailed review on this topic, see [3]). In [3] it is shown that the relative entropy (Kullback–Leibler distance)

$$\int f(x) \log \left(\frac{f(x)}{f_k(x, \hat{\theta}^{(k)})} \right) dx$$

converges to zero at the optimal rate $n^{-2s/(2s+1)}$ for densities whose logarithms have s square-integrable derivatives when the model size $k = n^{1/(2s+1)}$ is chosen according to a presumed degree of smoothness s. Stone [40] obtains similar results for log-spline models. Stone [41] later develops convergence rates for multidimensional function estimation (including density estimation) using tensor products of splines with a given order of interaction. The convergence rates are also obtained with presumed knowledge of the smoothness property of the target function. More recent results in this direction include [12] on minimum-contrast estimators on sieves and [46] on convergence rates of sieve MLE's. 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, it is desirable to have a statistical procedure which can automatically adapt to the true smoothness. That is, we wish to have a single estimator which behaves

Manuscript received June 11, 1995; revised July 1, 1997. This work was supported by NSF under Grant ECS-9410760. The material in this paper was presented in part at the IEEE-IMS Workshop on Information Theory, Alexandria, VA, October 1994.

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

A. R. Barron is with the Department of Statistics, Yale University, New Haven, CT 06520-8290 USA.

Publisher Item Identifier S 0018-9448(98)00079-0.

optimally for each smoothness condition, yet the estimator does not require the knowledge of true smoothness in advance. For density estimation, [24] considered estimating a density having a Fourier representation satisfying a certain smoothness assumption with smoothness parameters unknown. He proposed certain projection estimators and showed that the estimators converge at the optimal rates without knowing the smoothness parameter in advance. In later years, Donoho, Johnstone, Kerkyacharian, Picard, and others advocated the use of wavelet subset selection in both nonparametric regression and density estimation (see, e.g., [22] and [23]). For orthogonal wavelet expansion, subsets are selected by thresholding wavelet coefficients to zero out the coefficients of small magnitude. They showed that the wavelet-threshold estimators converge near optimally simultaneously over the Besov spaces also without the knowledge of the smoothness parameters. We here intend to use a model selection criterion to adaptively chose a suitable model so that the density estimator based on the selected model converge optimally for various unknown smoothness conditions. We will not restrict attention to orthogonal expansion nor even to linear models. We need to mention that at about the same time of this work, related results are obtained by Birgé and Massart [13] and Barron, Birgé, and Massart [8] concerning general penalized minimum-contrast estimators. Unlike these works, we focus here on penalized maximum likelihood with expansions for the log densities.

Next we discuss the forms of model selection criteria. AIC [1] is widely used in many statistical applications. This criterion is derived by Akaike from the consideration of 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 $-\log \text{likelihood}$ 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

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

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 [37] proposed BIC based on some Bayesian analysis; Rissanen [34] suggested the minimum-description length (MDL) criterion from an information-theoretic point of view. Usually the MDL criterion takes the form

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

The term $\frac{m_k}{2}\log n$ is the description length of the parameters with precision of order $1/\sqrt{n}$ for each parameter, and the likelihood is maximized over the parameters represented with this precision (addition terms that appear in refinements of MDL are in [2], [17], [35], [44], and [45]).

Some 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., [25] and [38]). For AIC, however, under the same setting, the probability of selecting a wrong model does not vanish as the sample size approaches ∞. Our interest in this paper is not in determination of a true finite-dimensional model but rather in selection of as accurate a model as permitted in view of the tradeoff with complexity for the given sample size in the case that the true density is not necessarily in any of the finite-dimensional operating models.

In a related nonparametric regression setting, an asymptotic optimality property is shown for AIC with fixed design [28] and [36]. 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 using descriptive length criteria, Barron and Cover [4] show that the Hellinger distance between the true density and the estimated one converges at a rate within a logarithmic factor of the optimal rate. As we mentioned before, some recent results in this direction are in [8] and [13].

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 smoothness and norm parameters in advance.

As opposed to AIC, we allow the bias-correction penalty term $\lambda_k m_k$ to be a multiple of the number of parameters in the model, and the coefficient λ_k will depend on a dimensionality constant of the model related to the metric entropy. In this paper, the coefficients are specified so that the asymptotic results hold. With this consideration, the criteria take the form

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

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

We evaluate the criteria by comparing the Hellinger distance

$$d_H^2(f,\, f_{\hat{k},\hat{\theta}^{(\hat{k})}}) = \int \left(\sqrt{f} - \sqrt{f_{\hat{k},\hat{\theta}^{(\hat{k})}}}\right)^2 d\mu$$

with an index of resolvability. The concept of resolvability was introduced by Barron and Cover [5] in the context of description length criteria. It naturally captures the capability of estimating a function by a sequence of models. In the present context, the index of resolvability can be defined as

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

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 $\lambda_k m_k/n$ 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 tradeoff between the approximation error and the estimation error. It is shown in this work that for the new criteria, when the λ_k's are chosen large enough, the statistical risk $Ed_H^2(f, f_{\hat{k}, \hat{\theta}(\hat{k})})$ is bounded by a multiple of $R_n(f)$.

To apply the above results, we can evaluate $R_n(f)$ for f in various nonparametric classes of functions, then upper bounds of the convergence rates can be easily obtained. Examples will be given to show these bounds for the maximum penalized likelihood estimator correspond to optimal or near-optimal rates of convergence simultaneously for density functions in various nonparametric classes. This provides a tool to show adaptivity of the estimator based on model selection.

In statistical applications, due to the lack of knowledge on the true function, it is often more flexible to consider a large number of models such as the case of subset selection in regression. 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 cannot 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 cannot 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, one can add a model complexity penalty (see, Yang [47]).

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,

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

See [4] and [13] for similar use of a model complexity term. The complexity $L_k = C_k \log_2 e$ can be interpreted as the 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

$$-\sum_{i=1}^{n} \log f_k(X_i, \hat{\theta}^{(k)}) + \lambda_k m_k + \nu C_k \tag{2}$$

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:

$$R_n(f) = \inf_k \left\{ \inf_{\theta^{(k)} \in \Theta_k} D(f || f_{k, \theta^{(k)}}) + \frac{\lambda_k m_k}{n} + \frac{\nu C_k}{n} \right\}. \tag{3}$$

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

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

The condition needed for our results is a metric dimension assumption involving both Hellinger distance and L_∞ distance on the approximating models. This assumption determines how large the penalty constant λ_k should be for the above claim to be valid. For exponential families $f_j(x, \theta^{(j)})$, $\theta^{(j)} \in \Theta_j$, to satisfy the assumption, it is often necessary to break the natural parameter space Θ_j into an increasing sequence of subsets $\Theta_{j,L}$, $L = 1, 2, \cdots$ according to the sup-norm of the corresponding log density. Then the classes of densities corresponding to these subsets of parameters are treated as models indexed by $k = (j, L)$ and the sup-norm of the log density appears as a factor in determining the penalty λ_k. Thus indirectly the criterion chooses parameter estimates for each model j not to maximize the likelihood for the full model but rather to maximize a penalized likelihood. For details of that treatment, see Section III.

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 $W_2^s(U)$, $s \in N = \{1, 2, \cdots\}$, $U > 0$. We approximate the logarithm of the density by spline functions. If we knew U and s, then by using suitably predetermined order splines, the optimal rate of convergence is achieved. However, this rate of convergence of $n^{-2s/(2s+1)}$ 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 other words, the density estimator based on the model selection adapts to every class $W_2^s(U)$, $s \in N$, $U > 0$.

The above example suggests that good model-selection criteria can provide us with minimax optimal function estimation strategies simultaneously for many different classes. For related results on adaptive function estimation, see [8], [13], [22], [24], [30], and [32]. As some other applications of our results, neural network models and sparse density function estimation will be considered.

We finally comment on the relationship with MDL criteria. 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 $-\log \text{likelihood}_k + \text{constant} \cdot m_k$, whose penalty term is of the same order as that in AIC. (Some results in that direction were presented by Barron, Yang, and Yu [7] for ellipsoidal constraints on parameters). 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, here the penalty term can take the form $\text{constant} \cdot m_k$

for a larger class of models than considered in [7]. We should also note that our criteria are not necessarily Bayesian.

The paper is organized as follows: in Section II, we state and prove the main theorem; in Section III, we provide some applications of the main results; in Section IV, we give the proofs of the key lemma and some other lemmas; and finally in the appendix, we prove several useful inequalities.

II. 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 Section III-B). Lemma 0 in Section IV will be used to derive the main theorem.

In our analysis, we will consider sup-norm distance between the logarithms of densities. In this paper, 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 is a δ-net if for any density $f \in B$, there exists $\tilde{f} \in F$ such that $||\log \tilde{f} - \log f||_\infty \le \delta$. For convenience, the index set of F might also be called a δ-net.

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

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

Assumption 1: For each $k \in \Gamma$, there exist a positive constant A_k and an integer $m_k \ge 1$ such that for any $\theta_0^{(k)} \in \Theta_k$, any $r > 0$ and $\delta \le 0.0056r$, there exists a δ-net $F_{k,r,\delta,\theta_0^{(k)}}$ for $B_k(\theta_0^{(k)}, r)$ satisfying the following requirement:

$$\operatorname{card}(F_{k,r,\delta,\theta_0^{(k)}}) \le \left(\frac{A_k r}{3\delta}\right)^{m_k}.$$

Here m_k is called the metric dimension of model k.

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.

Let $\hat{k}$ minimize the criterion in (2) over all models in Γ. The final density estimator $\hat{f}$ then is $\hat{f} = f_{\hat{k},\hat{\theta}^{(\hat{k})}}$, i.e., the maximum-likelihood density estimator in the selected model.

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

$$\Lambda(A) = 4.75 \log A + 27.93. \tag{4}$$

Theorem 1: Assume Assumption 1 is satisfied. Take $\lambda_k \ge \Lambda(A_k)$ and $\nu \ge 9.49$ in the model selection criterion given in (2). Then for the density estimator $f_{\hat{k},\hat{\theta}^{(\hat{k})}}$, we have

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

where

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

Note the condition needed for the above conclusion is only on the operating models. This property is essential for demonstrating adaptivity of the estimator based on model selection among many function classes (e.g., Sobolev classes with unknown smoothness and norm parameters).

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 $\lambda_k m_k/n$ 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

$$\inf_{\theta^{(k)}\in\Theta_k} D(f||f_{k,\theta^{(k)}}) + m_k/n$$

have A_{k_n} bounded and if $C_{k_n} = O(m_{k_n})$, then $R_n(f)$ is asymptotically comparable to

$$\inf_{k,\theta^{(k)}} \{D(f||f_{k,\theta^{(k)}}) + m_k/n + C_k/n\}$$

and consequently

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

The best tradeoff between approximation error and estimation error often gives the minimax rate of convergence for full approximation sets of functions (see [48, Sec. 5]). 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 III.
3) One way to assign the complexities for the models is by considering only the number of models for each dimension. Let $N(m) = \operatorname{card}\{k \in \Gamma : m_k = m\}$ 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

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

If N_m grows more slowly than exponential in m, then $[\log N(m_k) + 2\log(m_k+1)]/m_k$ 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 C_k/n is not

negligible compared to m_k/n (for related discussions, see [13] and [47]).

4) From the proof of Lemma 0 in Section IV, it can be seen that the requirement in Assumption 1 only needs to be checked for $r \geq 5.5\sqrt{m_k/n}$ for the conclusion of Theorem 1 to hold. If this weaker requirement is satisfied with $A_{k,n}$ (depending also on the sample size) instead of A_k, and we use $\lambda_{k,n} \geq \Lambda(A_{k,n})$ in the criterion, the conclusion of Theorem 1 is still valid.
5) The various constants (e.g., 2657) involved in Theorem 1 and subsequent results are given to ensure the risk bounds theoretically. As suggested by a referee, the estimator is likely to behave much better in practice.

Proof of Theorem 1: Clearly the working criterion is theoretically equivalent to selecting $\hat{k}$ and $\hat{\theta}^{(\hat{k})}$ to minimize

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

by adding

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) + t/n$$

(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 $V(k, \theta^{(k)})) \geq \gamma d_H^2(f, f_{k,\theta^{(k)}})$ 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

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

Then

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

We now use Lemma 0 in Section IV to give an upper bound on the above probability. For $r > 0$, let $\mathcal{B}_{\Theta_k}(f, r)$ be a Hellinger ball in Θ_k around the true density f (f may not be in the parametric family) with radius r defined by

$$\mathcal{B}_{\Theta_k}(f, r) = \{\theta^{(k)}: \theta^{(k)} \in \Theta_k, d_H(f, f_\theta) \leq r\}.$$

We next show that Assumption 0 for Lemma 0 is satisfied for the family $\{f_k(x, \theta^{(k)}), \theta^{(k)} \in \Theta_k\}$ under Assumption 1. Let $\theta_*^{(k)} \in \Theta_k$ satisfy

$$d_H(f, f_{k,\theta_*^{(k)}}) \leq \inf_{\theta^{(k)} \in \Theta_k} d_H(f, f_{k,\theta^{(k)}}) + r.$$

Then because for any $\theta^{(k)} \in \Theta_k$

$$d_H(f, f_{k,\theta^{(k)}}) \geq \frac{1}{2}(d_H(f, f_{k,\theta_*^{(k)}}) + d_H(f, f_{k,\theta^{(k)}})) - \frac{r}{2}$$
$$\geq \frac{1}{2} d_H(f_{k,\theta^{(k)}}, f_{k,\theta_*^{(k)}}) - \frac{r}{2}$$

we have

$$\mathcal{B}_{\Theta_k}(f, r) \subset \{\theta^{(k)}: \theta^{(k)} \in \Theta_k, d_H(f_{k,\theta_*^{(k)}}, f_{k,\theta^{(k)}}) \leq 3r\} = B_k(\theta_*^{(k)}, 3r).$$

Thus under Assumption 1, Assumption 0 is satisfied with $A = A_k$, $m = m_k$, and $\rho = 0.0056$ (note if

$$\inf_{\theta^{(k)} \in \Theta_k} d_H(f, f_{k,\theta^{(k)}})$$

is achievable, then Assumption 0 is satisfied with a better constant $A = \frac{2}{3}A_k$). Now by Lemma 0 with $\xi = \lambda_k m_k + \nu C_k + t$ $(t > 0)$, if

$$\lambda_k \geq 4/(1-4\gamma) \log\left(15.4 A_k \sqrt{1-4\gamma}/\gamma\right)$$

with $\gamma = 0.039$ (which satisfies $\rho = 0.13\gamma/\sqrt{1-4\gamma}$), then

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

Sum over $k \in \Gamma$

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

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,

$$\int_0^\infty q_n(t)\, dt \leq 85.6/(1-4\gamma).$$

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

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

and let $(k_n, \theta_n^{(k_n)})$ be a choice such that

$$R_n(k_n, \theta_n^{(k_n)}) \le (1+\epsilon)R_n(f)$$

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

$$t \ge t_0 = \frac{4\log 2}{2\log 2 - 1 + 4\gamma}$$

we have

$$\begin{aligned} p_n(t) &=: P\{V(k_n, \theta^{(k_n)}) \ge tR_n(k_n, \theta^{(k_n)})\} \\ &\le P\left\{L_n \ge n\left[tD(f||f_{k_n,\theta^{(k_n)}}) + \frac{(t-1)\lambda_{k_n}m_{k_n}}{n} \right.\right. \\ &\qquad \left.\left. + \frac{(t-1)\nu C_{k_n}}{n} - \frac{t}{n}\right]\right\} \\ &\le P\left\{L_n \ge \frac{nt}{2}\left(D(f||f_{k_n,\theta^{(k_n)}}) + \frac{\lambda_{k_n}m_{k_n}}{n} + \frac{\nu C_{k_n}}{n}\right)\right\} \\ &= P\left\{L_n \ge \frac{nt}{2}R_n(k_n, \theta^{(k_n)})\right\}. \end{aligned}$$

For the last inequality above, we use the fact

$$\lambda_k m_k \ge 4\log 2/(1-4\gamma)$$

for all $k \in \Gamma$. Note also

$$\frac{n}{2}R_n(k_n, \theta^{(k_n)}) \ge \frac{\lambda_{k_n}m_{k_n}}{2} \ge \frac{2\log 2}{1-4\gamma}.$$

Let

$$\tilde{L}_n = L_n I_{\left\{L_n \ge \frac{t_0 n}{2}R_n(k_n, \theta^{(k_n)})\right\}}$$

and

$$S_t = \left\{L_n \ge \frac{nt}{2}R_n(k_n, \theta^{(k_n)})\right\}.$$

Then for $t \ge t_0$

$$S_t = \left\{\tilde{L}_n \ge \frac{nt}{2}R_n(k_n, \theta^{(k_n)})\right\}.$$

Now

$$\begin{aligned} \int_{t_0}^{\infty} p_n(t)\,dt &\le \int_{t_0}^{\infty} EI_{S_t}\,dt \\ &= E\left(\int_{t_0}^{\infty} I_{S_t}\,dt\right) \\ &= E\,\frac{\tilde{L}_n}{\frac{n}{2}R_n(k_n, \theta^{(k_n)})} - t_0 \\ &= \frac{1}{\frac{n}{2}R_n(k_n, \theta^{(k_n)})}\int_{\{L_n \ge t_0 n R_n(k_n, \theta^{(k_n)})\}} L_n \\ &\qquad \cdot \prod_{i=1}^{n} f(x_i)\,d\mu - t_0. \end{aligned}$$

Here

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

To bound the last integral involving the tail of an expected log-likelihood ratio, we apply Lemma 3 in the appendix with $\alpha^* = \alpha(e^{16}) = 1.07$ and obtain

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

Now, from the analysis above

$$\begin{aligned} \gamma d_H^2(f, f_{\hat{k},\hat{\theta}(\hat{k})}) &\le V(\hat{k}, \hat{\theta}(\hat{k})) \le V(k_n, \theta_n^{(k_n)}) \\ &\le tR_n(k_n, \theta^{(k_n)}) \le (1+\epsilon)tR_n(f) \end{aligned}$$

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

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

Let

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

then

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

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

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

The choice $\gamma = 0.039$ minimizes

$$\frac{4}{1-4\gamma}\log\left(\frac{15.4A_k\sqrt{1-4\gamma}}{\gamma}\right)$$

at $A_k = 1$. The corresponding value of ρ is 0.0056, which is used in Assumption 1. This completes the proof of Theorem 1.

Remark: In the proof of the theorem, for the (nearly) best models k_n, we just use the fact that $D(f||f_{k_n,\theta^{(k_n)}})$ is finite. If

$$||\log\frac{f}{f_{k_n,\theta_n^{(k_n)}}}||_{\infty}$$

is bounded (which is satisfied, for instance, if $\log f$ is in a Sobolev ball, and the approximating models are truncated polynomial, or trigonometric, or spline series, see [3]), we can use Hoeffding's inequality to obtain exponential bound on

the tail probability for these models. Then one 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.,

$$Ed_H^{2j}(f, f_{\hat{k}, \hat{\theta}^{(\hat{k})}}) = O(R_n^j(f)), \qquad \text{for all } j > 0.$$

The criteria in (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

$$\sum_{q \in \Gamma_n} e^{-L(q)} \le 1.$$

Treat each density in Γ_n as a model, then Assumption 1 is satisfied with $A_k = 1$ and $m_k = 1$. Thus $\lambda_k m_k = 4/(1 - 4\gamma) \log 2$ is a constant independent of k. Therefore, when taking $\nu = 9.49$, the criterion in (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 (3) is essentially the same as the resolvability

$$\inf_{q \in \Gamma_n} \left\{ D(f\|q) + \frac{L(q)}{n} \right\}$$

considered by Barron and Cover [4].

III. Applications

A. Sequences of Exponential Families

As an application of the theorem we develop in Section II, 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.

1) 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 $\varphi_{j,1}(x), \varphi_{j,2}(x), \cdots, \varphi_{j,m_j}(x)$ for S_j satisfying the following two conditions with constants T_1 and T_2 not depending on j:

$$\left\| \sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) \right\|_\infty \le T_1 \max_i |\theta_i| \tag{5}$$

$$\left\| \sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) \right\|_2 \ge \frac{T_2}{\sqrt{m_j}} \|\theta\|_2. \tag{6}$$

Here $\|\ \|_\infty$ and $\|\ \|_2$ denote the sup-norm and L_2-norm, respectively. This type of conditions was previously used by Birgé and Massart [12]. The first condition is satisfied with a localized basis. The second one is part of the requirement that $\varphi_{j,1}(x), \varphi_{j,2}(x), \cdots, \varphi_{j,m_j}(x)$ forms a frame (see, e.g., [16, ch. 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:

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

where

$$\psi_j(\theta) = \log \int \exp\left(\sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x) \right) d\mu$$

is the normalizing constant. (If there is no restriction on the parameters $(\theta_1, \cdots, \theta_{m_j})$, the above parameterization 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 II, the models need to satisfy the metric dimension assumption. For that purpose, we cannot directly use the natural parameter space R^{m_j}. Instead, we consider a sequence of compact parameter spaces

$$\Theta_{j,L} = \{\theta \in R^{m_j} : -\underline{L} \le \log f_j(\cdot, \theta) \le \overline{L}\}$$

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

Lemma 1: There exists a constant

$$A(L, T_1, T_2) = 28.92 \frac{T_1}{T_2} (2 + \underline{L} + \overline{L}) e^{\underline{L}/2} + 0.18$$

such that Assumption 1 is satisfied with $A_{(j,L)} = A(L, T_1, T_2)$ and $m_{(j,L)} = m_j$.

Note in Lemma 1, $A(L, T_1, T_2)$ 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 1 is provided in Section V.

In practice, we might consider many different "types" of localized basis which satisfy (5) and (6) for each type of basis. For example, different order splines are useful when the smoothness condition of the true function is unknown. If q is the order of the splines, the constants $T_{q,1}$ and $T_{q,2}$ may not be bounded over all choices of q, 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 automatically based on data.

Assume for each q in an index set Q, we have a collection of models J_q satisfying (5) and (6) 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 $\lambda(q, L) = \Lambda(A(L, T_{q,1}, T_{q,2}))$. It depends on $\overline{L}$ logarithmically, but basically linearly on $\underline{L}$, indicating quicker increase of penalty when density values may get closer to zero. Let $\hat{k}$ be the model minimizing

$$-\sum_{i=1}^{n} \log f_k(X_i, \hat{\theta}^{(k)}) + \lambda(q, L) m_j + 9.49 C_k. \tag{7}$$

Then from Theorem 1, we have the following conclusion.

Corollary 1: For localized basis models for the log density satisfying (5) and (6), for any underlying density f

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

where

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

Corollary 1 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

$$R_n(f) = O(\inf_{j \in J_{q^*}} \{ \inf_{\theta \in \Theta_{j,q^*,L}} D(f||f_{j,\theta}) + \frac{m_j}{n} \})$$

which usually gives the minimax optimal rate of convergence for the density in the class.

Example 1: Univariate Log-spline models.

Let $S_{m,q} (m \ge q)$ 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

$$\varphi_{m,q,1}(x), \varphi_{m,q,2}(x), \cdots, \varphi_{m,q,m}(x)$$

be the B-spline basis. Let

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

where

$$\psi_{m,q}(\theta) = \log \int \exp \left(\sum_{i=1}^{m} \theta_i \varphi_{m,q,i}(x) \right) d\mu.$$

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

$$\Theta_{m,q,L} = \{ \theta \in R^m : -\underline{L} \le \log f_{m,q}(\cdot, \theta) \le \overline{L} \}$$

where $L = (\underline{L}, \overline{L})$, $q \ge 1$, $m \ge q$ all take positive integer values. Each parameter space $\Theta_{m,q,L}$ corresponds to a model.

The B-spline basis is known to satisfy the two conditions (5) and (6). In fact, the sup-norm of spline expressed by B-splines is bounded by the sup-norm of the coefficients (see, [20, p. 155]), that is,

$$\left\| \sum_{i=1}^{m} \theta_i \varphi_{m,q,i}(x) \right\|_\infty \le \max_{1 \le i \le m} |\theta_i|.$$

The second requirement follows from the frame property of the B-splines. From Stone [39, eq. (12)]

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

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 1 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 $L = (\underline{L}, \overline{L})$ using $\log_2^* \underline{L} + \log_2^* \overline{L}$ bits
2) describe q using $\log_2^* q$ bits
3) describe m using $\log_2^* m$ bits

where the function $\log^*$ is defined by

$$\log^* i = \log(i+1) + 2 \log \log (i+1), \qquad \text{for } i > 0.$$

Then the total number of bits needed to describe k is

$$\log_2^* \underline{L} + \log_2^* \overline{L} + \log_2^* q + \log_2^* m.$$

Thus a natural choice of C_k is

$$C_k = \log_2^* \underline{L} + \log_2^* \overline{L} + \log^* q + \log^* m.$$

Assume the logarithm of the target density belongs to $W_2^{s^*}(U^*)$ for some $s^* \ge 1$ 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 $\int (g^{(s)}(x))^2 dx \le U$. The parameters s^* and U^* are not known.

Corollary 2: Let $\hat{f} = f_{\hat{k}, \hat{\theta}(\hat{k})}$ be the density estimator with $\hat{k}$ selected by the criterion in (7) with

$$\lambda(q, L) = 43.92 + 4.75 \log \left(\frac{1}{\gamma_q} (2 + \underline{L} + \overline{L}) e^{\underline{L}/2} \right).$$

Then for any f with $\log f \in W_2^{s^*}(U^*)$

$$Ed_H^2(f, f_{\hat{k}, \hat{\theta}(\hat{k})}) \le M \cdot n^{-[2s^*/(2s^*+1)]}$$

where the constant M depends 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 and norm parameters. This theorem demonstrates an example of success of a completely data-driven strategy for nonparametric density estimation. For similar adaptation results, see the references cited in Section I.

Proof of Corollary 2: 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 [19, Theorems 5.2 and 2.1], for $\log f \in W_2^{s^*}(U^*)$ and for each $m \ge s^*$, there exists

$$g(x, \beta) = \sum_{i=1}^{m} \beta_i \varphi_{m,s^*,i}(x)$$

such that

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

where K and K' are absolute constants. By Lemma 5 in the appendix

$$\left|\log \int e^g \, d\mu\right| = \left|\log \int f \, d\mu - \log \int e^g \, d\mu\right| \le ||\log f - g||_\infty.$$

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

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

Therefore,

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

For the relative entropy approximation error, from [3, Lemma 1] (as shown at the bottom of this page). Take

$$L_{1,m} = L_{2,m} = \lceil ||\log f||_\infty + 2K'/(m - s^* + 2)^{s^* - 0.5} ||\log^{(s^*)} f||_2 \rceil$$

(bounded for $\log f \in W_2^{s^*}(U^*)$) and $L_m = (L_{1,m}, L_{2,m})$, then $\lambda(s^*, L_m)$ are bounded. Note also that C_{m, s^*, L_m} is asymptotically negligible compared to m. Thus

$$R_n(f) \le D(f||e^{\tilde{g}}) + \frac{\lambda(s^*, L_m) m}{n} + \frac{9.49 C_{m, s^*, L_m}}{n} \le \frac{\text{const}_1}{m^{2s^*}} + \frac{\text{const}_2 \cdot \text{m}}{n}$$

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 $n^{1/(2s^*+1)}$.

2) General Linear Spaces: Unlike the localized basis that satisfy (5) and (6), 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, $\varphi_{j,1}(x)$, $\cdots$, $\varphi_{j,m_j}(x)$. The finite dimensional families we consider are

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

where

$$\psi_j(\theta) = \log \int \exp\left(\sum_{i=1}^{m_j} \theta_i \varphi_{j,i}(x)\right) d\mu$$

is the normalizing constant.

In [3] (also in [12] and [13]), 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

$$||h||_\infty \le K_j ||h||_2 \tag{8}$$

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 1), break the natural parameter space into an increasing sequence of compact spaces

$$\Theta_{j,L} = \{\theta \in R^{m_j} : -\underline{L} \le \log f_j(\cdot, \theta) \le \overline{L}\}, \quad L = (\underline{L}, \overline{L})$$

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

Lemma 2: For each model $k = (j, L)$, Assumption 1 is satisfied with

$$A_{(j,L)} = 28.92 K_j (2 + \underline{L} + \overline{L}) e^{\underline{L}/2} + 0.18$$

and $m_{(j,L)} = m_j + 1$.

The proof of this lemma is in Section VI.

If an upper bound on $||\log f||_\infty$ is known in advance, then for each j, we can consider only $\underline{L} = \overline{L} = \lceil ||\log f||_\infty \rceil$. Then from Remark 3 to Theorem 1, the model complexity can be ignored. However, when $||\log f||_\infty$ is unknown, we would like to consider all integer values for $\underline{L}$ and $\overline{L}$. 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 $\sum e^{-C_k} \le 1$. Let

$$\lambda_{(j,L)} = \Lambda(A_{(j,L)}) = 43.92 + 4.75 \log(K_j(2 + \underline{L} + \overline{L}) e^{\underline{L}/2}).$$

Let $\hat{k}$ be the model minimizing

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

$$D(f||e^{\tilde{g}}) \le \frac{1}{2} e^{||\log f - g||_\infty} \times ||f||_\infty \times ||\log f - g||_2^2 \le \frac{K^2 \exp\left\{\frac{K'}{(m - s^* + 2)^{s^* - 0.5}} ||\log^{(s^*)} f||_2 + ||\log f||_\infty\right\}}{2(m - s^* + 2)^{2s^*}} ||\log^{(s^*)} f||_2^2$$

Since the conditions for Theorem 1 are satisfied, we have the following result about model selection for a sequence of exponential families with a general linear basis.

Corollary 3: For the log-density models with basis satisfying (8), for any underlying density f

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

where

$$R_n(f) = \inf_L \inf_{j \in J} \left\{ \inf_{\theta \in \Theta_{j,L}} D(f \| f_{k,\theta}) + \frac{\lambda_{(j,L)} m_k}{n} + \frac{9.49 C_k}{n} \right\}.$$

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 2: Polynomial case.

Let $S_j = \text{span}\{1, x, x^2, \cdots, x^j\}$, $j \ge 1$. Then $m_j = j$. From [3, Lemma 6], $K_j = j+1$. It follows from Lemma 2 that

$$\lambda_{(j,L)} = 43.92 + 4.75 \log((j+1)(2 + \underline{L} + \overline{L}) e^{\underline{L}/2}).$$

Take $C_k = \log^* \underline{L} + \log^* \overline{L} + \log^* j$. For densities with logarithms in each of the Sobolev spaces $W_2^s(U)$, $s \ge 1$, and $U > 0$, when $\overline{L}$ and $\underline{L}$ are large enough, say $\overline{L}, \underline{L} \ge l^*$ (depending on U and s), the relative entropy approximation error of model (j, L) is bounded by $\text{const}_{U,s}(1/j^{2s})$ (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 [3, Sec. 7]). Thus

$$\inf_{\theta \in \Theta_{j,(l^*,l^*)}} D(f \| f_{k,\theta}) + \frac{\lambda_{(j,(l^*,l^*))} j}{n} + \frac{9.49 C_k}{n} \le \text{const}_{U,s} \left(\frac{1}{j^{2s}} + \frac{j \log j}{n} \right).$$

Optimizing over j, we obtain that

$$R_n(f) \le \text{const}_{U,s} \times (\log n / n)^{2s/(2s+1)}$$

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

Example 3: Trigonometric case.

Let

$$S_j = \text{span}\left\{1, \sqrt{2}\cos(2\pi x), \sqrt{2}\sin(2\pi x), \cdots, \sqrt{2}\sin(2\pi j x)\right\}, \qquad j \ge 1.$$

Then $m_j = 2j$. From [3, eq. (7.6)], $K_j = \sqrt{2j+1}$. Again by examining the resolvability (for L_2 and L_∞ error bounds for trigonometric approximation, see [3, Sec. 7]), 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 [13] have used a theorem of Talagrand [42] to show that if $K_j \le \text{const}\sqrt{j}$, then their penalized projection estimator with the bias-correction penalty term $\text{const}\,(j/n)$ 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 a maximum-likelihood method with trigonometric basis to derive a criterion giving the optimal convergence rate.

B. 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 log density by neural nets.

We approximate $g = \log f$ 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 parameterized by θ, consisting of $a_j \in R^d$, $b_j, \eta_j \in R$, for $j = 1, 2, \cdots k$. The normalizing constant η_0, is

$$\eta_0 = -\log \int_{[-\frac{1}{2}, \frac{1}{2}]^d} \exp\left\{ \sum_{j=1}^{k} \eta_j \phi(a_j^T x + b_j) \right\} dx.$$

The integer $k \ge 1$ is the number of nodes (or hidden units). Here ϕ is a given sigmoidal function with $\|\phi\|_\infty \le 1$, $\lim_{z \to \infty} \phi(z) = 1$, and $\lim_{z \to -\infty} \phi(z) = 0$. Assume also that ϕ satisfies Lipschitz condition

$$|\phi(z_1) - \phi(z_2)| \le v_1 |z_1 - z_2|, \qquad z_1, z_2 \in R$$

for some constant $v_1 > 0$. Let $v = \max(v_1, 1)$. Let

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

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 [5], [6], and Modha and Masry [29]. 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

$$\varsigma_g = \int |\omega|_1 |\tilde{g}(\omega)|\, d\omega$$

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 $\varsigma_g \le \varsigma$. Recent work of Barron [5] 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. Barron, Birgé, and Massart [8] has similar conclusions with neural net modeling of the density rather than the log density.

Consider the parameter space

$$\Theta_{k,\tau_k,\varsigma} = \Bigg\{\theta\colon \max_{1\le j\le k} |a_j|_1 \le \tau_k, \max_{1\le j\le k} |b_j| \le \tau_k, \sum_{j=1}^{k} |\eta_j| \le 2\varsigma\Bigg\}.$$

The constant τ_k is chosen such that

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

The compact parameter spaces are used so that the cardinality assumption is satisfied. From [5, Theorem 3], for a log density g with $\varsigma_g \le \varsigma$, there exists a $\theta \in \Theta_{k,\tau_k,\varsigma}$ such that

$$||g - g_{k,\theta}||_{[-(1/2),1/2]^d} \le \frac{4\varsigma_g}{\sqrt{k}} \tag{9}$$

where $||\cdot||_{[-(1/2),1/2]^d}$ 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 1 is satisfied for these models. For any $\varepsilon > 0$, $\varsigma \ge 1$, from [6, Proof of Lemma 2], there exists a set $\Theta_{k,\varepsilon,\tau_k,\varsigma}$ such that for any $\theta \in \Theta_{k,\tau_k,\varsigma}$, there is $\tilde\theta \in \Theta_{k,\varepsilon,\tau_k,\varsigma}$ satisfying $||g_k(x,\theta) - g_k(x,\tilde\theta)||_\infty \le 8v\varsigma\varepsilon$ with

$$\operatorname{card}(\Theta_{k,\varepsilon,\tau,\varsigma}) \le \left(\frac{2e(\tau_k+\varepsilon)}{\varepsilon}\right)^{k(d+1)} \left(\frac{2(1+\varepsilon)}{\varepsilon}\right)^{k}.$$

Take $\varepsilon = \delta/8v\varsigma$. Because

$$B_k(\theta^*, r) = \{\theta \in \Theta_{k,\tau_k,\varsigma} : d_H(f_{k,\theta^*}, f_{k,\theta}) \le r\} \subset \Theta_{k,\tau_k,\varsigma}$$

so for $\delta \le \rho r$, we have a δ-net in $B_k(\theta^*, r)$ with cardinality bounded by

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

For $r \ge 5.5\sqrt{m_k/n}$, where $m_k = kd + 2k + 1$ is the number of parameters, the above quantity is bounded by

$$\left(\frac{16ev\varsigma\tau_k}{5.5\sqrt{\frac{m_k}{n}}} + 2e\rho\right)^{k(d+1)} \left(\frac{16v\varsigma}{5.5\sqrt{\frac{m_k}{n}}} + 2\rho\right)^{k} \left(\frac{r}{\delta}\right)^{kd+2k}.$$

Notice that in order for the conclusion of Theorem 1 to hold, the cardinality requirement in Assumption 1 only needs to be checked for $r \ge 5.5\sqrt{m_k/n}$ (see the remarks to Theorem 1 and after the proof of Lemma 0), and the weaker assumption is satisfied with

$$A_{k,n} = 3\left(\frac{3.0ev\varsigma\tau_k}{\sqrt{\frac{m_k}{n}}} + 2e\rho\right)^{k(d+1)/(kd+2k)} \cdot \left(\frac{3.0v\varsigma}{\sqrt{\frac{m_k}{n}}} + 2\rho\right)^{k/(kd+2k)} \le \text{const} \times \left(\tau_k\sqrt{\frac{n}{m_k}} + 1\right)$$

where the constant depends only on v and ς. As shown in [5], if $\phi(z)$ approaches its limits at least polynomially fast, then there exist constants β_1 and β_2 such that $\tau_k \le \beta_1 k^{\beta_2}$. As a consequence,

$$A_{k,n} \le \text{const} \times k^{\beta_2 - 1/2}\sqrt{n}$$

with the constant depending on v, ς, and β_1. By Theorem 1, when we choose the penalty $\lambda_k = \Lambda(A_{k,n})$ and $\nu_k = 9.49$ in the model selection criterion given in (2), for the density estimator $f_{\hat k, \hat\theta(\hat k)}$, we have $Ed_H^2(f, \hat f) \le 2657R_n(f)$, where

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

For the targeted densities, under the assumption $\varsigma_g \le \varsigma$, the log density is uniformly bounded (see [18, Lemma 5.3]). Indeed, because

$$|g(x) - g(0)| = \left|\int_{R^d} (e^{i\omega^T x} - 1)\tilde g(\omega)\, d\omega\right| \le \int_{R^d} |\omega^T x||\tilde g(\omega)|\, d\omega \le \frac{1}{2}\int_{R^d} |\omega|_1 |\tilde g(\omega)|\, d\omega \le \frac{1}{2}\varsigma_g$$

it follows that

$$|g(0)| = \left| -\log \int_{[-\frac{1}{2},\frac{1}{2}]^d} e^{g(x)-g(0)}\,dx \right| \le |g(x)-g(0)| \le \frac{1}{2}\varsigma_g.$$

Thus we have

$$|g(x)| \le |g(x)-g(0)| + |g(0)| \le \varsigma_g.$$

Then by [3, Lemma 1], for the target densities,

$$D(f||f_{k,\theta}) \le \text{const}_\varsigma ||g-g_{k,\theta}||^2_{[-\frac{1}{2},\frac{1}{2}]^d}$$

for $\theta \in \Theta_{k,\tau_k,\varsigma}$. So from (9)

$$\inf_{\theta\in\Theta_{k,\tau_k,\varsigma}} D(f||f_{k,\theta}) \le \text{const}_\varsigma \frac{1}{k}.$$

Note λ_k is of order $\log A_k = O(\log(nk^{2\beta_2-1}) = O(\log n)$ for $k \le n$. Therefore,

$$R_n(f) \le \text{const}\left(\inf_{k\ge 1}\left(\frac{1}{k} + \frac{kd+2k}{n}\log n\right)\right) \le \text{const}'\left(\frac{d\log n}{n}\right)^{1/2}$$

where the constants depend on v, ς, and the choice of ϕ. Together with Theorem 1, we have

$$Ed_H^2(f,\hat{f}) \le \text{const}_{v,\varsigma,\phi}(d\log n/n)^{1/2}.$$

Note that for the class of functions considered, the rate of convergence $\frac{1}{2}$ is independent of the function dimension as in [5], [27], and [29].

C. 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 $[0,1]^d$ (or $[-\frac{1}{2},\frac{1}{2}]^d$). 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 bound.

In addition to the observed i.i.d. sample $X_1, X_2, \cdots, X_n$ from f, let $Y_1, Y_2, \cdots, Y_n$ be a generated i.i.d. sample (independent of X_i's) from $U[0,1]^d$. Let Z_i be X_i or Y_i with probability $(\frac{1}{2},\frac{1}{2})$ using $V_i \sim \text{Bernoulli}(\frac{1}{2})$ independently for $i=1,\cdots,n$. Then Z_i has density $g(x) = \frac{1}{2}(f+1)$. Clearly, 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 $Z_1,\cdots,Z_n$ using the criterion in (2) from the models in Γ, which satisfy Assumption 1. Then when λ_k and ν are chosen large enough, by Theorem 1 and the familiar relationship between Hellinger distance and L_1 distance, namely, $||f_1-f_2||_1 \le 2d_H(f_1,f_2)$ for any two densities, we have

$$E||g-\hat{g}||_1 \le 2Ed_H(g,\hat{g}) \le 2\sqrt{Ed_H^2(g,\hat{g})} \le 104\sqrt{R_n(g)}.$$

Let $\tilde{g}(x) = \hat{g}(x)I_{\{\hat{g}(x)\ge 1/2\}} + \frac{1}{2}I_{\{\hat{g}(x)<1/2\}}$. Then pointwise in x, $|g-\tilde{g}| \le |g-\hat{g}|$. Indeed, they are the same for x with $\hat{g}(x)\ge\frac{1}{2}$, and the inequality holds when $\hat{g}(x)<\frac{1}{2}$ since then $\tilde{g}(x)=\frac{1}{2}$ and $g(x)\ge\frac{1}{2}$. Consequently,

$$E\int |g-\tilde{g}|\,dx \le E\int |g-\hat{g}|\,dx \le 104\sqrt{R_n(g)}.$$

In particular,

$$E\int \tilde{g}\,dx - 1 \le E\int |g-\tilde{g}|\,dx \le 104\sqrt{R_n(g)}.$$

Now we construct a density estimator for f. Note that $f(x) = 2g(x)-1$, let

$$\hat{f}(x) = \frac{2\tilde{g}(x)-1}{2\int\tilde{g}(x)\,dx - 1}.$$

Then $\hat{f}(x)$ is a nonnegative and randomized probability density estimate (depending on $(X_i,Y_i,Z_i)_{i=1}^n$) and

$$\begin{aligned} E\int |f(x)-\hat{f}(x)|\,dx &\le E\int |f(x)-2\tilde{g}(x)+1|\,dx \\ &\quad + E\int |\hat{f}(x)-2\tilde{g}(x)+1|\,dx \\ &= 2E\int |g-\tilde{g}|\,dx + 2E\left(\int\tilde{g}(x)\,dx - 1\right) \\ &\le 416\sqrt{R_n(g)} \end{aligned}$$

where the equality uses the fact that $\tilde{g}\ge\frac{1}{2}$. To avoid randomization, one could take conditional expectation of $\hat{f}$ with respect to the randomness of $Y_1, Y_2, \cdots, Y_n$ and $Z_1,\cdots,Z_n$. Then by convexity, the new estimator has no bigger L_1 risk than $\hat{f}$. From above, we have the following result.

Theorem 2: Let $\hat{f}$ be constructed in the above way with a choice of the penalty constants satisfying $\lambda_k \ge \Lambda(A_k)$, $\nu \ge 9.49$, $k\in\Gamma$, then

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

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 rates of convergence in L_1.

Example 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

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

From

$$\int (f^{(s^*)}(x))^2\,dx < \infty$$

it can be shown that

$$\int ((\log g)^{(s^*)})^2\,dx < \infty.$$

Then from previous result, $R_n(g) = O(n^{-(2s^*/2s^*+1)})$. Thus

$$E\int |f(x)-\hat{f}(x)|\,dx \le \zeta n^{-[s^*/(2s^*+1)]}$$

where the constant ζ depends only on s^*, $||f||_\infty$ 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 $G(s, U)$, $s \ge 1$, $U > 0$, where $G(s, U)$ is defined to be the collection of densities with $||f||_\infty$ and the square integral of the sth derivative bounded by U.

D. Complete Models versus Sparse Subset Models

As in Section III-A, we consider the estimation of $\log f$ 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 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.

We first illustrate heuristically possible advantages of sparse subset models. Formal results are given afterwards.

Let $\Phi = \{\phi_1, \cdots, \phi_k, \cdots\}$ be a chosen collection of uniformly bounded basis functions of d variables. Assume that

$$\log f = \sum_{i\ge 1} \theta_i^* \phi_i$$

can be linearly approximated polynomially well in the sense that there exist constants $s > 0$ and c_1 such that

$$\inf_\theta ||\log f - \sum_{i=1}^m \theta_i\phi_i||_2 \le c_1 m^{-s/d}$$

for all $m \ge 1$. Here s is possibly very small compared to d, which implies that the linear approximation error may decay very slowly.

Using linear approximation with the first m basis functions, the tradeoff between the (squared) approximation error of order $m^{-2s/d}$ and estimation error (say, under squared L_2 norm) of order m/n due to estimating m parameters is optimized when m is of order $n^{d/(2s+d)}$. It results in an order of $n^{-[2s/(2s+d)]}$ on the total error. On the other hand, nonlinear approximations can be used. Here we consider the use of sparse-terms approximation. Let

$$g_I(x, \theta) = \sum_{i\in I} \theta_i\phi_i$$

where I is a finite subset of integers indicating which basis functions are used. Assume the coefficients in

$$\log f = \sum_{i=1}^\infty \theta_i\phi_i$$

satisfy

$$\sum_{i=1}^\infty |\theta_i| \le c_2 < \infty.$$

Then from [5], one can show that for each $m \ge 1$, there exists a subset $I(m, g)$ of size m such that

$$\inf_{\theta_{I(m,g)}} ||\log f - g_I(\cdot, \theta_{I(m,g)})||_2 \le \frac{\beta}{\sqrt{m}}$$

for some constant β depending only on c_2 and a uniform bound on the basis functions. If one knew $I(m, g)$ and use the corresponding basis functions to estimate $\log f$, again by balancing the approximation error of order $1/m$ and the estimation error of order m/n, one seems to get order $1/\sqrt{n}$ on the total error. This rate is much smaller than $n^{-[2s/(2s+d)]}$ when s is small compared with d. For applications, of course, the question is how to find $I(m, g)$ or a nearly good subset. It is probably realistic to expect that a price should be paid for selecting such a sparse subset. As will be seen later in the analysis, when $\log f$ can be linearly approximated polynomially well by the system Φ, searching for a good sparse subset causes only an extra logarithmic factor in the total risk bound compared to that could be obtained with the knowledge of a best subset in advance.

Analogous conclusions are in [21] using the idea of unconditional bases and sparsity indices. However, unlike the present analysis, Donoho's treatment requires the basis providing sparse approximations to be orthonormal. Relaxing orthonormaility permits consideration of multivariate B-splines, trigonometric expansions with fractional frequencies, and neural net models.

Now let us give a formal analysis. 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, $\varphi_{j,1}(x)$, $\varphi_{j,2}(x)$, $\cdots$, $\varphi_{j,L_j}(x)$. Let

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

where

$$\psi_j(\theta) = \log \int \exp\left(\sum_{i=1}^{L_j} \theta_i \varphi_{j,i}(x)\right) d\mu$$

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

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

where

$$\psi_{I_j}(\theta) = \log \int \exp\left(\sum_{i \in I_j} \theta_i \varphi_{j,i}(x)\right) d\mu$$

and $I_j \subset \{1, 2, \cdots, L_j\}$ 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.

Suppose that Assumption 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 $A_{I_j} \le \beta_1 {L_j}^{\beta_2}$ for all the subset models. To satisfy this requirement, we may need to restrict the parameters to compact spaces

$$\Theta_{I_j, L} = \{\theta \in R^{m_{I_j}} \colon \|\log f_{I_j}(\cdot, \theta)\|_\infty \le L\}$$

for a fixed value L. Then from Lemma 2, this condition is satisfied if K_j in (8) is bounded by a polynomial of L_j, which is the case for 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 $A_{I_j, L} \le \beta_{1,L} {L_j}^{\beta_2}$, 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 $\Gamma_n = \{j : L_j \le N_n\}$ and $\tilde{\Gamma}_n = \{(j, I_j)\colon L_j \le N_n$ and $I_j \subset \{1, 2, \cdots, L_j\}\}$. 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 models. This restriction is also reasonable for the complete models because usually a model with the number of parameters bigger than the number of observations cannot 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

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

over $j \in \Gamma_n$.Then from Theorem 1, the squared Hellinger risk of the density estimator $f_{\hat{j}, \hat{\theta}^{(\hat{j})}}$ from the selected model $\hat{j}$ is bounded by a multiple of

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

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 it is among $\binom{L_j}{m_{I_j}}$ possibilities. This strategy suggests the 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 $\overline{j}$ and $\overline{I} = \overline{I}_{\overline{j}}$ be the minimizer of

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

over $(j, I_j) \in \tilde{\Gamma}_n$. Again from Theorem 1, the risk of the density estimator $f_{\overline{I}, \hat{\theta}^{(\overline{I})}}$ from subset selection is bounded by a multiple of

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

A related quantity is

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

which is roughly the ideal best tradeoff 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. We compare $R_n(f)$, $\tilde{R}_n(f)$, and $r_n(f)$ in the next proposition.

Proposition:

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

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

2) Let $N_n \le n^\kappa$ 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, $\tilde{R}_n(f) = O(r_n(f) \log n)$.

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 inequality

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

The proof of the proposition is straightforward and is omitted here.

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

$$\frac{\tilde{R}_n(f)}{R_n(f)} \le O(\alpha_n \log n)$$

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.

Even for one-dimensional function estimation, the sparse subset models also turn out to be advantageous in several related settings such as the estimation of a function with bounded variation using variable bin histograms, and the estimation of a function in some Besov spaces using wavelets (see [8] and [23]). For high-dimensional function estimation, there are even more advantages in considering the sparse subset models.

Example 4: Sparse multi-index series.

Let $\underline{i} = (i_1, \cdots, i_d)$ be a vector of integers. Consider a multi-indexed basis

$$\{\varphi_{\underline{i}}(x): \underline{i} = (i_1, \cdots, i_d) \in \{0, 1, 2, \cdots\}^d\}$$

on $[0,1]^d$ with $\varphi_{\underline{i}}(x)$ uniformly bounded. Here the basis could be a tensor product basis

$$\varphi_{\underline{i}}(x) = \prod_{l=1}^{d} \varphi_{i_l}(x_l)$$

produced from a one-dimensional basis

$$\{\varphi_0(x), \varphi_1(x), \varphi_2(x), \cdots\}.$$

Another multi-indexed basis is

$$\{\sin(2\pi(\underline{i} \cdot x)), \cos(2\pi(\underline{i} \cdot x)), \underline{i} \in \{0, 1, 2, \cdots\}^d\}.$$

Let $|\underline{i}| = \max_{l \le d} i_l$. The complete models are

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

where

$$\psi_j(\theta) = \log \int \exp\left(\sum_{|\underline{i}| \le j} \theta_{\underline{i}} \varphi_{\underline{i}}(x)\right) d\mu$$

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. The subset models are

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

where

$$\psi_{I_j}(\theta) = \log \int \exp\left(\sum_{\underline{i} \in I_j} \theta_{\underline{i}} \varphi_{\underline{i}}(x)\right) d\mu$$

and $I_j \subset \{\underline{i} : |\underline{i}| \le j\}$. Assume Assumption 1 is satisfied with $A_j \le \beta_1 (j^d)^{\beta_2}$ and dimension $m_j = j^d$ for the complete models and with $A_{I_j} \le \beta_1 (j^d)^{\beta_2}$ 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 natural 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:

$$\sum_{\underline{i}} |\theta^*_{\underline{i}}| \le M_2 \tag{10}$$

$$\left\| \sum_{|\underline{i}| > j} \theta^*_{\underline{i}} \varphi_{\underline{i}}(x) \right\|_2 \le \sqrt{M_3} j^{-s}, \qquad \text{for all } j \ge 1. \tag{11}$$

If the basis is orthonormal, then (11) is

$$\sum_{|\underline{i}| > j} (\theta^*_{\underline{i}})^2 \le M_3 j^{-2s}.$$

Let $F(M_1, M_2, M_3, s)$ be the collection of the densities satisfying the above conditions.

Let

$$g_j(x) = \sum_{|\underline{i}| \le j} \theta^*_{\underline{i}} \varphi_{\underline{i}}(x)$$

be a good approximator of $\log f$ in the model j. Then the complete model j has an approximation error $\|\log f - g_j\|_2^2 \le M_3 j^{-2s}$. Using the same technique used in Section III-A ([3, Lemma 1] is still applicable because $\|g_j\|_\infty$ is bounded), it can be shown that the resolvability for the complete models is of order $(\log n/n)^{2s/(2s+d)}$.

Now consider the approximation error for the subset models from the complete model j. From [5, Lemma 1] we know that there exists a constant $\beta_3 > 0$ (depending only on M_1 and M_2) such that for any $1 \le m \le j-1$, there is a subset I^* (depending on f) of size m and some parameter values $\tilde{\theta}_{\underline{i}}$ such that

$$\|g_j - \sum_{\underline{i} \in I^*} \tilde{\theta}_{\underline{i}} \varphi_{\underline{i}}\|_2^2 \le \frac{\beta_3}{m}.$$

Then the approximation error of $g_{I^*} = \sum_{\underline{i} \in I^*} \tilde{\theta}_{\underline{i}} \varphi_{\underline{i}}$ is

$$\|\log f - g_{I^*}\|_2^2 \le 2\|\log f - g_j\|_2^2 + 2\|g_{I^*} - g_j\|_2^2 \le \frac{2M_3}{j^{2s}} + \frac{2\beta}{m}.$$

Take j of order $n^{1/4s}$ and m of order $\min(n^{1/4s}, n^{1/2})$, we have $\|\log f - g_{I^*}\|_2^2 \le \text{const} \cdot n^{-1/2}$, where the constant depends only on M_1, M_2, and M_3. The corresponding model complexity is

$$\log^* j + \log(j^d) + \log \binom{j^d}{m} = O\left(\sqrt{n} d \log n\right).$$

Again, with the technique used in Section III-A, 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, β_2) of $\sqrt{d \log n/n}$. The resulting rate of convergence of the resolvability bound is independent of the function dimension d and is better than $(\log n/n)^{2s/(2s+d)}$ from the complete models when $2s < d$ (when $2s \ge d$, the resolvabilities of complete models and subset models are of the same order).

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

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

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 $\hat{\theta}^{(\hat{I}_{\hat{j}})}$ by $\hat{\theta}$ for short. The density estimator is then $f_{\hat{I}, \hat{\theta}}$. By Theorem 1, we have the following conclusion.

Theorem 3: For $f \in F(M_1, M_2, M_3, s)$, the density estimator $\hat{f} = f_{\hat{I}, \hat{\theta}}$ converges in squared Hellinger distance at a rate bounded above by $\sqrt{d \log n/n}$ uniformly. That is,

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

where the constant ζ 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 $F(M_1, M_2, M_3, s)$, $M_1 > 0$, $M_2 > 0$, $M_3 > 0$, $s > 0$.

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 Section III-B, we estimated linear and nonlinear parameters in the neural-network models. 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.

IV. Proofs of the Main Lemmas

We now state and prove Lemma 0, which is used to prove Theorem 1 in the main section.

Let f be the true density function, and $f(x, \theta)$, $\theta \in \Theta$ be a parametric family of densities. For $r > 0$, let $\mathcal{B}_\Theta(f, r)$ be a Hellinger ball in Θ around f (f may not be in the parametric family) with radius r defined by

$$\mathcal{B}_\Theta(f, r) = \{\theta: \theta \in \Theta, d_H(f, f_\theta) \le r\}.$$

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

Lemma 0 gives an exponential inequality which is used 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 [26], Birgé [10], and others.

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

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

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

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

then

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

Proof of Lemma 0: We use a "chaining" argument similar to that used in Birgé and Massart [11], [12]. For related techniques, see [43] and [46].

We consider dividing the parameter space into rings as follows:

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

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

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

Let a sequence $\delta_j \downarrow 0$ be given with $\delta_0 \le \rho r_0$, then by the assumption, there is a sequence $F_0, F_1, F_2, \cdots$ of $\delta_0/2, \delta_1, \cdots$ nets in Θ_i satisfying the cardinality bounds. For each $\theta \in \Theta_i$, let

$$\tau_j(\theta) = \arg \min_{\theta' \in F_j} ||\log (f_\theta / f_{\theta'})||_\infty$$

be the nearest representor of θ in the 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 $\lim_{j\to\infty} f(x, \tau_j(\theta)) = f(x, \theta)$, it follows that

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

Let

$$q_i = P^* \Bigg\{ \text{for some } \theta \in \Theta_i, (1/n) \sum_{i=1}^n \log \frac{f(X_i, \theta)}{f(X_i)} \ge -\gamma d_H^2(f, f_\theta) + \frac{\xi}{n} \Bigg\}$$

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

$$q_i = P^* \Bigg\{ \text{for some } \theta \in \Theta_i, \ell_0(\theta) + \sum_{j=1}^\infty (\ell_j(\theta) - E\ell_j(\theta)) \ge E \log \frac{f(X_1, \tau_0(\theta))}{f(X_1, \theta)} - \gamma d_H^2(f, f_\theta) + \frac{\xi}{n} \Bigg\}.$$

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

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

Then let $\tilde{F}_0 = \{\tilde{\theta}_0 \colon \theta_0 \in F_0\}$. 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. Notice that for $\theta \in B_{\theta_0}$

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

so we have

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

where η_j, $j \ge 1$ are positive numbers satisfying

$$\sum_{j=1}^\infty \eta_j \le \gamma r_i^2. \tag{12}$$

To bound $q_i^{(1)}$, we use a familiar exponential inequality as follows (see, e.g., [4] and [15]).

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

$$P\left\{ \frac{1}{n} \sum_{i=1}^n \log \frac{g_1(X_i)}{g_2(X_i)} \ge t \right\} \le e^{-(n/2)(d_H^2(g_1, g_2) + t)}.$$

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

$$\begin{aligned} &P\left\{ \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i, \theta_0)}{f(X_i)} \ge -2\gamma r_i^2 + \frac{\xi}{n} - \epsilon \right\} \\ &\qquad \le \exp\left(-\frac{n}{2} (d_H^2(f, f_{\theta_0}) - 2\gamma r_i^2 + \frac{\xi}{n} - \epsilon \right\} \\ &\qquad \le \exp\left(-\frac{n}{2} (r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} - \epsilon \right\}. \end{aligned}$$

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

$$\begin{aligned} q_i^{(1)} &\le P\left(\bigcup_{\theta_0 \in F_0} \left\{ \ell_0(\theta_0) \ge -2\gamma r_i^2 + \frac{\xi}{n} - \epsilon \right\} \right) \\ &\le \text{card}\,(F_0) \exp\left(-\frac{n}{2} \left(r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} - \epsilon \right) \right). \end{aligned}$$

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

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

Note for $i \ge 1$

$$\begin{aligned} r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} &\ge \frac{2^{i-1}\xi}{n} - \frac{2\gamma 2^i \xi}{n} + \frac{\xi}{n} \\ &\ge (i+1)(1 - 4\gamma) \frac{\xi}{n} \end{aligned}$$

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

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

Now because

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

we have

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

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., [31, pp. 191–192]), we get

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

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 δ_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 η_j, $j \geq 1$ is defined such that

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

With these choices, the bound on q_i becomes

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

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 δ_0, δ_j,η_j, it follows that

$$\begin{aligned}\delta_0 &= 2Ar_0\exp\left(-\frac{(1-4\gamma)\xi}{4m}\right) \\ \delta_j &= Ar_0\exp\left(-\frac{(j+1)(1-4\gamma)\xi}{4m}\right), \qquad \text{for } j \geq 1 \\ \eta_1 &= 2A\sqrt{1-4\gamma}\sqrt{3i+9}\,\frac{\xi}{n}\exp\left(-\frac{(1-4\gamma)\xi}{4m}\right) \qquad (13)\end{aligned}$$

and for $j \geq 2$ see the top of the following page. 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 (12). Indeed,

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

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

$$\begin{aligned}&\cdot\left(2\sqrt{3} + \frac{1}{1-\exp\left(-\frac{(1-4\gamma)\xi}{8m}\right)}\right) \\ &\le \left(2\sqrt{3} + \frac{\sqrt{2}}{\sqrt{2}-1}\right)A\sqrt{1-4\gamma}\,\sqrt{i+5}\,\frac{\xi}{n} \\ &\quad\cdot\exp\left(-\frac{(1-4\gamma)\xi}{4m}\right) \\ &\le 6.88A\sqrt{1-4\gamma}\,\sqrt{i+5}\,\frac{\xi}{n}\,\exp\left(-\frac{(1-4\gamma)\xi}{4m}\right).\end{aligned}$$

Thus for

$$\sum_{j=1}^{\infty}\eta_j \le \gamma r_i^2$$

to hold, it suffices to have

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

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

$$\begin{aligned}\frac{\xi}{m} &\ge \frac{4}{1-4\gamma}\log\left(\frac{15.4A}{\gamma}\sqrt{1-4\gamma}\right) \\ &= \frac{4}{1-4\gamma}\log\frac{2A}{\rho} \qquad (14)\end{aligned}$$

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

Finally, we sum over the rings indexed by i

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

From (13) and (14)

$$\frac{\delta_0}{r_0} = 2Ae^{-\frac{(1-4\gamma)\xi}{4m}} \le 2A\times\frac{\rho}{2A} = \rho$$

as required. This completes the proof of the lemma.

Remark: From the proof of Lemma 0, it is seen that the requirement in Assumption 0 needs only to be checked for

$$r \ge \sqrt{[4\log 2/(1-4\gamma)](m/n)}$$

for the exponential inequality to be valid.

Proof of Lemma 1: We first show the Hellinger ball is contained in some l_2 ball. Then for the l_2 ball, we provide a suitable δ-net satisfying the cardinality bound. A similar calculation is in [12].

Because $1 \in S_j$, we have

$$1 = \sum_{i=1}^{m_j}\eta_i\varphi_{j,i}(x)$$

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 $\beta_i = \theta_i - \psi_j(\theta)\eta_i$. For any $\theta, \theta^* \in \Theta_{j,L}$

$$\frac{f_j(x,\theta)}{f_j(x,\theta^*)} \le e^{\underline{L}+\overline{L}}$$

so from Lemma 4 in the appendix

$$d_H^2(f_{\theta^*}, f_\theta) \ge \frac{\phi_1(e^{\underline{L}+\overline{L}})}{\phi_2(e^{\underline{L}+\overline{L}})}\int f_{\theta^*}(\log f_\theta - \log f_{\theta^*})^2\,dx$$

$$\geq M_L \int (\log f_\theta - \log f_{\theta^*})^2\, dx$$
$$\geq \frac{M_L T_2^2}{m_j} \sum_{i=1}^{m_j} (\beta_i - \beta_i^*)^2$$

where

$$M_L = \frac{\phi_1(e^{\underline{L}+\overline{L}})}{\phi_2(e^{\underline{L}+\overline{L}})} e^{-\underline{L}}$$

and for the last inequality, we use the frame assumption in (6). Therefore, for any $\theta^* \in \Theta_{j,L}$

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

The inclusion above refers to the functions represented by the parameters θ and β. Now we want to find a suitable δ-net on $\tilde{B}\left(\beta^*, \sqrt{m_j r^2/M_L T_2^2}\right)$. 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

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

From (5), 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 $\varepsilon = \delta/T_1$, then $||\log f_\theta - \log f_{\tilde{\theta}}||_\infty \leq \delta$. For $\delta < \rho r$

$$\overline{r} \leq r\sqrt{\frac{2m_j T_1^2}{M_L T_2^2} + 2\rho^2}.$$

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

$$|F_\delta| \leq \frac{1}{\sqrt{m_j \pi}} \left(\frac{\sqrt{2\pi e^2\left(\frac{2T_1^2}{M_L T_2^2} + 2\rho^2\right)} r}{\delta}\right)^{m_j}.$$

Clearly, F_δ is a δ-net for densities in $B_j(\theta^*, r)$. Thus Assumption 1 is satisfied with

$$A_k = 3\sqrt{2\pi e^2\left(\frac{2T_1^2}{M_L T_2^2} + 2\rho^2\right)}.$$

From Lemma 5

$$M_L \geq \frac{1}{(2 + \overline{L} + \underline{L})^2 e^{\underline{L}}}$$

so

$$A_k \leq 3\sqrt{2\pi e^2 \times \frac{2T_1^2}{M_L T_2^2}} + \sqrt{2\pi e^2 \times 2\rho^2}$$
$$\leq 28.92 \frac{T_1}{T_2}(2 + \overline{L} + \underline{L}) e^{\frac{\underline{L}}{2}} + 0.18$$

with $\rho = 0.0056$.

Proof of Lemma 2: We consider an orthonormal basis 1, $\varphi_{j,1}(x)$, $\varphi_{j,2}(x)$, $\cdots$, $\varphi_{j,m_j}(x)$ in S_j. Let $\beta = (\theta_1, \theta_2, \cdots, \theta_{m_j}, \psi_j(\theta))$. From the proof of Lemma 1, 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,

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

The inclusion above is meant for the functions that the parameters represent. Similarly to the counting argument in the proof of Lemma 1, a rectangular grid spaced at width $\delta/K_j\sqrt{m_j + 1}$ for each coordinate provides the desired δ-net. The cardinality constant

$$A_{(j,L)} = 3\sqrt{2\pi e^2\left(\frac{2K_j^2}{M_L} + 2\rho^2\right)}$$
$$\leq 28.92 K_j (2 + \underline{L} + \overline{L}) e^{\underline{L}/2} + 0.18$$

for $\rho = 0.0056$. This completes the proof Lemma 2.

APPENDIX

Lemma 3: Assume f and g are two probability density functions with respect to some σ-finite measure μ. Let $s > 1$ be any constant, then

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

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

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

Here we avoid the square root with $s > 1$. Note $\alpha(s) \to 1$ as $s \to \infty$. Improved bounds of the form $O((c/s^2)D(f||g))$ are possible under the condition $\mathrm{var}\,(\log(f/g)) \leq cD(f||g)$. 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 3: We consider a familiar expression of the relative entropy

$$\begin{aligned} D(f||g) &= \int f\log\frac{f}{g}\,d\mu \\ &= \int f\left(\log\frac{f}{g} + \frac{g}{f} - 1\right)d\mu \\ &= \int_{\{f/g\geq s\}} f\left(\log\frac{f}{g} + \frac{g}{f} - 1\right)d\mu \\ &\quad + \int_{\{f/g<s\}} f\left(\log\frac{f}{g} + \frac{g}{f} - 1\right)d\mu. \end{aligned}$$

Because $(\log(f/g) + (g/f) - 1) \geq 0$, to prove the lemma, it suffices to show

$$\log\frac{f}{g} \leq \alpha(s)\left(\log\frac{f}{g} + \frac{g}{f} - 1\right), \qquad \text{for } \frac{f}{g} \geq s.$$

This follows from the monotonicity of $\alpha(s)$, which can be shown from simple calculation. This completes the proof of the lemma.

Lemma 4: Let p and q be two probability density functions with respect to some σ-finite measure μ. If $p(x)/q(x) \leq V$ for all x, then

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

where

$$\phi_1(V) = \frac{\log V + (1/V) - 1}{\log^2 V} \geq \frac{1}{2+\log V}$$

and

$$\phi_2(V) = \frac{V\log V + 1 - V}{\left(\sqrt{V} - 1\right)^2} \leq (2 + \log V).$$

The above upper bound on the relative entropy is in [12, Lemma 5].

Proof of Lemma 4: We note

$$D(p||q) = \int p(\log\frac{p}{q} + \frac{q}{p} - 1)\,d\mu.$$

It can be shown from calculus that

$$\phi_1(x) = \frac{\log x + (1/x) - 1}{\log^2 x}$$

is decreasing on $(0, \infty)$, which implies

$$\frac{\log V + \frac{1}{V} - 1}{\log^2 V}\int p\left(\log\frac{p}{q}\right)^2 d\mu \leq D(p||q).$$

To prove the other inequality, we consider the following parts of $D(p||q)$ and $d_H^2(p, q)$:

$$\begin{aligned} D(p||q) &= \int_{\{q>p\}} p\left(\log\frac{p}{q} + \frac{q}{p} - 1\right)d\mu \\ &\quad + \int_{\{q<p\}} q\left(\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}\right)d\mu \\ d_H^2(p,q) &= \int_{\{q>p\}} p\left(\sqrt{\frac{q}{p}} - 1\right)^2 d\mu \\ &\quad + \int_{\{q<p\}} q\left(\sqrt{\frac{p}{q}} - 1\right)^2 d\mu. \end{aligned}$$

For $p < q$

$$\log\frac{p}{q} + \frac{q}{p} - 1 \leq 2\left(\sqrt{\frac{q}{p}} - 1\right)^2$$

so

$$\int_{\{q>p\}} p\left(\log\frac{p}{q} + \frac{q}{p} - 1\right)d\mu \leq 2\int_{\{q>p\}} p\left(\sqrt{\frac{q}{p}} - 1\right)^2 d\mu.$$

For $p > q$

$$\phi_2\left(\frac{p}{q}\right) = \frac{\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}}{\left(\sqrt{p/q} - 1\right)^2}$$

is increasing in p/q. It follows that

$$\begin{aligned} &\int_{\{q<p\}} q\left(\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}\right)d\mu \\ &\qquad \leq \frac{\log V + \frac{1}{V} - 1}{\log^2 V}\int_{\{q<p\}} q\left(\sqrt{\frac{p}{q}} - 1\right)^2 d\mu. \end{aligned}$$

Combining the integrals together, we conclude

$$D(p||q) \leq \frac{V\log V + 1 - V}{\left(\sqrt{V} - 1\right)^2}\,d_H^2(p,q),$$

which completes the proof of Lemma 4.

Lemma 5: 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

$$\left|\log\int e^{h_1}\,d\mu - \log\int e^{h_2}\,d\mu\right| \leq ||h_1 - h_2||_\infty.$$

Proof of Lemma 5:

$$\begin{aligned} \log\int e^{h_1}\,d\mu - \log\int e^{h_2}\,d\mu &= \log\int\frac{(e^{(h_1-h_2)+h_2}}{\int e^{h_2}d\mu}d\mu \\ &\geq \int\log(e^{h_1-h_2})\frac{e^{h_2}}{\int e^{h_2}d\mu}d\mu \\ &\geq -||h_1 - h_2||_\infty \end{aligned}$$

by Jensen's inequality. Similarly,

$$\log\int e^{h_1}\,d\mu - \log\int e^{h_2}\,d\mu \leq ||h_1 - h_2||_\infty$$

which completes the proof.

ACKNOWLEDGMENT

The authors wish to thank the referees for their many valuable suggestions, which led to a significant improvement on presentation of the results.

REFERENCES

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. on Information Theory*, B. N. Petrov and F. Csaki, Eds. Budapest, Hungary: Akademia Kiado, 1973, pp. 267–281.

[2] A. R. Barron, "Logistically smooth density estimation," Ph.D. dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA, 1985.

[3] A. R. Barron and C.-H Sheu, "Approximation of density function by sequences of exponential families," *Ann. Statist.*, vol. 19, pp. 1347–1369, 1991.

[4] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.

[5] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–945, 1993.

[6] ———, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, 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. on Information Theory* (Trondheim, Norway, 1994), p. 38.

[8] A. R. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Prob. Theory Related Fields*, 1996, to be published.

[9] J. M. Bernardo, "Reference prior distributions for Bayesian inference," *J. Roy. Statist. Soc.*, vol. 41, ser. B, pp. 113–147, 1979.

[10] L. Birgé, "Approximation dans les espaces metriques et theorie de l'estimation," *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, vol. 65, pp. 181–237, 1983.

[11] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimators," *Prob. Theory Related Fields*, vol. 97, pp. 113–150, 1993.

[12] ———, "Minimum contrast estimators on sieves: Exponential bounds and rates of convergence," Tech. Rep., Universite Paris-Sud, 1996.

[13] ———, "From model selection to adaptive estimation," in *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam*, D. Pollard and G. Yang, Eds. New York: Springer, 1996.

[14] N. N. Cencov, *Statistical Decision Rules and Optimal Inference.* Providence, RI: Amer. Math. Soc. Transl., 1982, vol. 53.

[15] 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.

[16] C. K. Chui, *An Introduction to Wavelets.* New York: Academic, 1991.

[17] B. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, 1990.

[18] ———, "Jeffrey's prior is asymptotically least favorable under entropy risk," *J. Statist. Planning Infer.*, vol. 41, pp. 37–40, 1994.

[19] C. de Boor and G. J. Fix, "Spline approximation by quasiinterpolents," *J. Approx. Theory*, vol. 8, pp. 19–45, 1973.

[20] C. de Boor, *A Practical Guide to Splines.* New York: Springer-Verlag, 1978.

[21] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Appl. and Comput. Harmonic Anal.*, vol. 1, pp. 100–115, 1993.

[22] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Statist.*, vol. 24, pp. 508–539, 1996.

[23] D. L. Donoho and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," 1994, unpublished manuscript.

[24] S. Yu. Efroimovich, "Nonparametric estimation of a density of unknown smoothness," *Theory Probab. Appl.*, vol. 30, pp. 557–568, 1985.

[25] 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.

[26] L. M. Le Cam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, pp. 38–53, 1973.

[27] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "Efficient agnostic leaning of neural networks with bounded fan-in," *IEEE Trans. Inform. Theory*, to be published.

[28] K. C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *Ann. Statist.*, vol. 15, pp. 958–975, 1987.

[29] D. S. Modha and E. Masry, "Rates of convergence in density estimation using neural networks," 1994, manuscript.

[30] A. Nemirovskii, "Nonparametric estimation of smooth regression functions," *J. Comput. Syst. Sci.*, vol. 23, pp. 1–11, 1986.

[31] D. Pollard, *Convergence of Stochastic Processes.* New York: Springer-Verlag, 1984.

[32] B. T. Polyak and A. B. Tsybakov, "Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression," *Theory Probab. Appl.*, vol. 35, pp. 293–306, 1990.

[33] S. Portnoy, "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tend to infinity," *Ann. Statist.*, vol. 16, pp. 356–366, 1988.

[34] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.

[35] ———, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 40–47, 1996.

[36] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, pp. 45–54, 1981.

[37] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[38] T. P. Speed and B. Yu, "Model selection and prediction: Normal regression," *Ann. Inst. Stat. Math.*, vol. 45, pp. 35–54, 1993.

[39] C. J. Stone, "The dimensionality reduction principle for generalized additive models," *Ann. Statist.*, vol. 14, no. 2, pp. 590–606, 1986.

[40] ———, " Large-sample inference for log-spline models," *Ann. Statist.*, vol. 18, pp. 717–741, 1990.

[41] ———, "The use of polynomial splines and their tensor products in multivariate function estimation," *Ann. Statist.*, vol. 22, pp. 118–184, 1994.

[42] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Probab.*, vol. 22, pp. 28–76, 1994.

[43] S. Van de Geer, "Hellinger consistency of certain nonparametric maximum likelihood estimates," *Ann. Statist.*, vol. 21, pp. 14–44, 1993.

[44] C. S. Wallace and D. M. Boulton, "An invariant Bayes method for point estimation," *Classification Soc. Bull.*, vol. 3, pp. 11–34, 1975.

[45] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Roy. Statist. Soc. B*, vol. 49, pp. 240–265, 1987.

[46] W. H. Wong and X. Shen, "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs," *Ann. Statist.*, vol. 23, pp. 339–362, 1995.

[47] Y. Yang, "Complexity-based model selection," prospectus submitted to Department of Statistics, Yale University, New Haven, CT, 1993.

[48] Y. Yang and A. R. Barron, "Information-theoretic determination of minimax rates of convergence," submitted to *Ann. Statist.*, 1996.