MULTIVARIATE DENSITY ESTIMATION BASED ON ADAPTIVE PARTITIONING: CONVERGENCE RATE, VARIABLE SELECTION AND SPATIAL ADAPTATION

By

Linxi Liu
Wing Hung Wong

January 2014

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
Multivariate Density Estimation Based on Adaptive Partitioning: Convergence Rate, Variable Selection and Spatial Adaptation

Linxi Liu* and Wing Hung Wong*†
*Department of Statistics, Stanford University
†Department of Health Research and Policy, Stanford University

January 11, 2014

Abstract
Density estimation is one of the most fundamental problems in statistical inference. Here, we introduce a non-parametric approach to multivariate density estimation. The estimators are piecewise constant density functions supported by binary partitions. The partition of the sample space is learned by maximizing the likelihood of the corresponding histogram on that partition. We analyze the convergence rate of these maximum likelihood estimators and reach a conclusion that for a relatively rich class of density functions the rate does not directly depend on the dimension. This suggests that, under certain conditions, this method is immune to the curse of dimensionality. We also apply this method to variable selection and spatial adaptation cases, and calculate the explicit convergence rates respectively. These results help us further understand under what circumstances, and for which density classes, would this method perform well.

1 Introduction
Density estimation is a fundamental problem of statistics, because once an explicit and precise estimate of the density function is obtained, various kinds of statistical inference become natural and straightforward. Both parametric and nonparametric density estimation methods have been developed. However, currently, increasing dimension and data size impose great difficulty on these traditional methods. For instance, a fixed parametric family, such as multivariate Gaussian, may fail to capture the geometric features of the true density function under high dimensions. On the other hand, a traditional nonparametric method, like the kernel density estimator, may suffer from the difficulty of choosing appropriate bandwidths (Jones et. al. [6]). In this paper, a nonparametric method for multivariate density estimation is studied. This is a sieve
maximum likelihood method that employs simple, but still flexible, binary partitions to adapt to the data distribution. In this paper, thorough analyses of the convergence rate will be carried out to quantify the performance of this method as the dimension increases and the regularity of the true density function varies. These analyses demonstrate the major advantages of this method, especially when dealing with large data set or high dimensional data.

1.1 Challenges in multivariate density estimation

Most of the established methods for density estimation were initially designed for the estimation of univariate or low-dimensional density functions. For example, the popular kernel method (Rosenblatt [12] and Parzen [11]), which approximates the density by the superposition of windowed kernel functions centering on the observed data points, works well for estimating smooth low-dimensional densities. As the dimension increases, the accuracy of the kernel estimates becomes very sensitive to the choice of the window size and the shape of the kernel. To get good performance, both of these choices need to depend on the data. However, the question of how to adapt these parameters to the data has not been adequately addressed. This is especially so for the kernel which is a multidimensional function itself. As a result, the performance of current kernel estimators deteriorates rapidly as dimension increases.

The difficulty caused by high dimensionality is also revealed by a classic result by Charles Stone. In Stone ([15]), it was showed that the optimal rate of convergence for density in $d$-dimensional space, when the density is assumed to have $p$ bounded derivatives, is of the order $n^{-\alpha}$, where $\alpha = p/(2p + d)$. When $d$ is small and the density is smooth (i.e. $p$ is large), then methods such as kernel density estimation can achieve a convergence rate almost as good as the parametric rate of $n^{-1/2}$. However, when $d$ is large, then even if the density has many bounded derivatives, the best possible rate will still be unacceptably slow. For example, in 20-dimensional space, the optimal rate is only $n^{-1/6}$ for densities with bounded derivatives up to the 5th order. Thus standard smoothness assumptions on the density will not protect us from the “curse of dimensionality”. Instead, we must seek alternative conditions on the underlying class of density that are general enough to cover some useful applications under high dimensions, and yet strong enough to enable the construction of density estimators with fast convergence. More specifically, suppose $r$ is a parameter that control the complexity (in a sense to be made precise) of the density class, with large value of $r$ indicating low complexity. We would like to construct density estimators with a convergence rate of the order $n^{-\gamma(r)}$, where $\gamma(\cdot)$ is an increasing function not dependent on $d$, and satisfying the property that $\gamma(r) \uparrow \frac{1}{2}$ as $r \uparrow \infty$. Since this rate does not depend on $d$ in any way, it is possible to obtain fast convergence even in high dimensional cases. For density estimators based on adaptive partitioning, such a result is established in Theorem 3.6 below.
1.2 Adaptive partitioning

Perhaps the most basic method for density estimation is the histogram. With appropriately chosen origin and bin width, the histogram density value within each bin is proportional to the relative frequency of the data points in that bin. Some further developments of the method allow the bin width to vary, so that data adaption can be brought in (Scott [13] and Lugosi and Nobel [8]). Although the estimation made by the histogram is coarse in the univariate case, it sheds light on high dimensional generalizations: if a good partition of the sample space is a concise representation of the geometry of the true density, we can estimate the multivariate density by learning a partition in the first stage and then estimate the density on each subregion based on the number of data points therein. The challenge is how to learn the partition in the multivariate case.

In Wong and Ma [18], a Bayesian formulation was proposed to learn the partition in multi-dimensional space. By recursively and randomly partitioning sample space, reweighing the subregions generated from the partitioning, and allowing optional stopping, a density function can be obtained. They constructed a prior distribution of the collection of these density functions, called the optional Pólya tree (OPT). This prior is proven to have large support in the space of absolutely continuous distributions, or equivalently, in the space of densities, in the sense that it has positive probability in all total variation neighborhoods. The posterior distribution yielded from this prior is also an OPT and the parameters governing the posterior tree can be determined by a recursive algorithm. The optional Pólya tree successfully incorporates the idea of data-driven partition learning. However, the computational cost of the recursive algorithm is extremely high.

In order to resolve the computational issues, another partition-learning based approach called Bayesian Sequential Partitioning (BSP) was developed (Ju, Jiang and Wong [9]). The major improvement of this new approach is that the posterior distribution can be analytically derived, and the log-posterior probability of a fixed partition is asymptotically linear in the estimation error of the corresponding histogram in terms of the Kullback-Leibler divergence. By employing Sequential Importance Sampling, they designed an efficient algorithm to sample from the posterior. Because of its computational feasibility, this approach becomes more powerful when dealing with high dimensional problems. Here is an example cited from their work. The 5-dimensional target density is

\[
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \frac{2}{5} \mathcal{N}\left(\begin{pmatrix} 0.3 \\ 0.6 \end{pmatrix}, 0.05^2 I_{2 \times 2}\right) + \frac{3}{5} \mathcal{N}\left(\begin{pmatrix} 0.7 \\ 0.4 \end{pmatrix}, 0.05^2 I_{2 \times 2}\right),
\]

\[X_3, X_4 \sim \mathcal{N}(0.5, 0.1),\]

\[X_5 \sim \frac{1}{2} \mathcal{N}(0.35, 0.1) + \frac{1}{2} \mathcal{N}(0.6, 0.05).\]

The density is designed to have multi-modality and correlation among some component variables. Both BSP and the kernel density estimator (KDE) were
Table 1: Comparison between the kernel density estimator (KDE) and the Bayesian Sequential Partitioning (BSP). For a fixed sample size, three simulations were carried out.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>KDE</th>
<th>BSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 \times 10^4$</td>
<td>0.32</td>
<td>0.1151</td>
</tr>
<tr>
<td>$5 \times 10^4$</td>
<td>0.26</td>
<td>0.06348</td>
</tr>
</tbody>
</table>

applied. The sample size was varied from $1 \times 10^4$ to $5 \times 10^4$. They compared the Hellinger distances between the true and the estimated density functions. The results are summarized in Table 1.

This comparison illustrates the advantages of the BSP when the unknown dependency structure is the main obstacle to estimating the density. Recall that from the histogram to the OPT, and then to the BSP, the essence of all the three methods is to learn a partition of the sample space from the data, and then to estimate the density within each subregion separately. If the partitions are flexible enough to represent the key geometric features of the true density, then the difficulty imposed by high dimensionality can be overcome. Thus density estimation scheme based on partition learning is of great interest.

In this paper, we will focus on a class of density estimators based on adaptive partitioning. Learning an appropriate partition of the sample space from the data is still the essence of this method. Our interests lie in the following areas: how does this class of estimators perform under high dimensions, and for which classes of density functions can the partitions achieve concise representations of their geometry. The performance will be quantified by the convergence rate of the estimators, and the representation quality will be formulated as the approximation error to the true density by using those densities which are constant on each subregion of the partitions. We will demonstrate that the optimal convergence rate can be achieved by this class of estimators, and that the rate does not directly depend on the dimension. This suggests that the method has the potential to meet the challenge of high dimensionality. In order to further reveal the data-adaptive nature of the method, we will apply it to variable selection and spatial adaptation cases, and will derive explicit convergence rates respectively. From these results, we will gain more insights into how this partition-learning based method can circumvent the obstacle of high dimensionality, and for which density classes it will be more effective.

1.3 Organization of the paper

The rest of the paper is organized in the following way. In Section 2, we will discuss the partition scheme and define the maximum likelihood estimator of the
true density. Main results on the convergence rate will be obtained in Section 3. Section 4 and Section 5 will be devoted to the applications to the variable section case and the spatial adaptation one respectively.

2 The maximum likelihood estimators based on adaptive partitioning

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of independent random variables distributed according to a density $f_0(y)$ with respect to a $\sigma$-finite measure $\mu$ on a measurable space $(\mathcal{Y}, \mathcal{B})$. We are interested in the case that $\mathcal{Y}$ is a bounded rectangle in $\mathbb{R}^p$ and $\mu$ is the Lebesgue measure. After scaling, we may assume that the support of the random variables is the unit cube in $\mathbb{R}^p$, that is, $\mathcal{Y} = \{(y^1, y^2, \ldots, y^p) : y^l \in [0, 1]\}$. Let $\Theta = \{f \in m\mathcal{Y}^+ : \int_{\mathcal{Y}} f d\mu = 1\}$ be the collection of all the density functions on $(\mathcal{Y}, \mathcal{B}, \mu)$. $\Theta$ is the parameter space we are going to deal with.

2.1 Densities on binary partitions

In this paper, the partitions are restricted to binary ones. To address the infinite dimensionality of $\Theta$, we construct a sequence of finite dimensional approximating spaces $\Theta_1, \Theta_2, \ldots, \Theta_I, \ldots$ based on binary partitions. With growing complexity, these spaces are approximations to the parameter space $\Theta$ with increasing precision. The idea of partition learning is integrated in the operations on these approximating spaces, such as searching over the spaces. We use a recursive procedure to define a binary partition with $I$ subregions of the unit cube in $\mathbb{R}^p$.

Let $A = \{(y^1, y^2, \ldots, y^p) : y^l \in [0, 1]\}$ be the unit cube in $\mathbb{R}^p$. In the first step, we choose one of the coordinates $y^l$ and cut $A$ into two subregions along the midpoint of the range of $y^l$. That is, $A = A^0_0 \cup A^1_0$, where $A^0_0 = \{y \in A : y^l \leq 1/2\}$ and $A^1_0 = A \setminus A^0_0$. In this way, we get a partition with two subregions. Note that the total number of all the possible partitions after the first step is equal to the dimension $p$. Suppose after $I - 1$ steps of the recursion, we have obtained a partition $A = \bigcup_{i=1}^I A_i$ with $I$ subregions. In the $I$-th step, further partitioning of the region is defined as follows:

1. Choose a region from $A_1, \ldots, A_I$. Denote it as $A_{i_0}$.

2. Choose one coordinate $y^j$ and divide $A_{i_0}$ into two subregions along the midpoint of the range of $y^j$.

Such a partition obtained by $I$ recursive steps is called a binary partition of size $I + 1$. Figure 2.1 displays all the possible two dimensional binary partitions when $I$ is 1, 2 and 3.
Now, let

$$\Theta_I = \{ f \in \Theta : f = \sum_{i=1}^{I} \beta_i I_{A_i}, \sum_{i=1}^{I} \beta_i \mu(A_i) = 1, \cup_{i=1}^{I} A_i \text{ is a binary partition of } A \text{ of size } I \}. $$

Then, $\Theta_I$ is the collection of the density functions supported by the binary partitions of size $I$. They constitute a sequence of approximating spaces (i.e. a sieve, see Shen and Wong [14] for background on sieve theory).

The topology on $\Theta$ and $\Theta_I$ is induced by the Hellinger distance, which is defined to be

$$\rho(f, g) = \left( \int_Y (\sqrt{f(y)} - \sqrt{g(y)})^2 dy \right)^{1/2}, \ f, g \in \Theta. \quad (2.1)$$

For $f^m_I \in \Theta_I$, $m = 1, 2$, let $f^m_I = \sum_{i=1}^{I} \beta_{i}^m A_i^m$, where $\cup_{i=1}^{I} A_i^m$ is a binary partition of $A$. Then the Hellinger distance between $f^1_I$ and $f^2_I$ can be written as

$$\rho^2(f^1_I, f^2_I) = \sum_{i=1}^{I} \sum_{j=1}^{I} (\sqrt{\beta_i^1} - \sqrt{\beta_j^2})^2 \mu(A_i^1 \cap A_j^2). \quad (2.2)$$

The quality of the approximation to the true density by the elements in $\Theta_I$ is formulated in the following way. A density function $f \in \Theta$ is said to be well approximated by those density functions which are constant within
each subregion of the binary partitions, if there exits a sequence of \( \pi f \in \Theta_I \), satisfying that \( \rho(\pi f, f) = O(I^{-r})(r > 0) \). Let \( \Theta \) be a collection of these density functions. We will derive convergence rate results under this condition. And we will also see that \( \Theta \) is still a rich class in Section 4 and Section 5, and will introduce more straightforward replacements of this condition there. In the rest of the paper, our attention is restricted to \( \Theta \), but we still denote \( \Theta \) by \( \Theta \).

2.2 The MLEs

For any \( f \in \Theta_I \), the log-likelihood is defined to be

\[
L_n(f) = \sum_{j=1}^{n} \log f(Y_j) = \sum_{i=1}^{I} N_i \log \beta_i,
\]

(2.3)

where \( N_i \) is the count of data points in \( A_i \), i.e., \( N_i = \text{card}\{ j: Y_j \in A_i, 1 \leq j \leq n \} \). The maximum likelihood estimator on \( \Theta_I \) is defined to be

\[
\hat{f}_{n,I} = \arg \max_{f \in \Theta_I} L_n(f).
\]

(2.4)

We claim that \( \hat{f}_{n,I} \) is well defined. Because given the binary partition \( A = \cup_{i=1}^{I} A_i \), the underlying distribution becomes a multinomial one and \((\beta_1, \cdots, \beta_I)\) can be determined by maximizing the log-likelihood. Within \( \Theta_I \), the number of possible binary partitions is finite.

From the above argument, we can also see how the idea of partition learning is incorporated. Indeed, in the above procedure, we scored each partition by the maximum likelihood it can achieve, and then selected the most promising one according to these scores.

3 Main results on convergence rate

We have defined a sequence of maximum likelihood estimators \( \hat{f}_{n,I}s \). Now we are ready to study the rate at which \( \hat{f}_{n,I} \) converges to \( f_0 \).

To define the convergence rate of \( \hat{f}_{n,I} \), let \( B(f_0) \) be a ball centering at \( f_0 \) with radius \( \delta \). The radius can be adjusted so that the estimator \( \hat{f}_{n,I} \) lies within the ball with high probability. In order to cover the estimator, the size of the ball should depend on how large the sample size is and how precise the true density \( f_0 \) can be approximated by the elements in \( \Theta_I \). In fact, the radius decreases to zero at a certain rate as \( n \) and \( I \) go to infinity. The optimal decay rate of \( \delta \) is the convergence rate we will focus on.

The analyses in this paper will demonstrate that the optimal convergence rate can be achieved by balancing the sample size with the complexity of the approximating spaces. On one hand, the complexity of the \( \Theta_I \) affects the convergence rate in a way that, the richer the approximating spaces, the more accurate the approximation of the true density. On the other hand, given a sample \( Y_1, Y_2, \cdots, Y_n \) of fixed size, there is a point beyond which the limited amount
of information conveyed in the data may be overwhelmed by the overly-complex approximating spaces. More information about how to strike the balance will be another major result of this paper. However, before we move to the core issues, an appropriate measure of the complexity must be chosen. In Section 3.1 we will briefly discuss metric entropy with bracketing, which measures the complexity by “counting” how many pairs of functions in an \( \epsilon \)-net are needed to provide simultaneous upper and lower bounds of all the elements. An important result of this section will be an upper bound for the bracketing metric entropy of the \( \Theta_I \). In the proof, we will cite several results from the studies of empirical process. Here, the process is indexed by the log-likelihood ratios. A difficulty is that most of these results require boundedness or finite moment generating functions of the function class indexing the process while the log-likelihood ratios do not always satisfy these conditions. The way we address this difficulty is by introducing the lower-truncated log-likelihood ratios, which have been proven to meet the above requirements while still maintaining the properties of the log-likelihood ratios needed to study the behavior of the maximum likelihood estimator (Wong and Shen devoted to this topic).

Finally in Section 3.3 we will establish the main result of the convergence rate. The key element of the proof is a change-of-measure argument. Although our goal is to bound the tail probability under \( P_{f_0} \), we will bound it under \( P_{\pi, f_0} \) first. Because under \( P_{\pi, f_0} \), the distribution turns out to be multinomial and the simple structure of the distribution makes the explicit calculation feasible.

### 3.1 Calculation of the metric entropy with bracketing

A general discussion of metric entropy can be found in Kolmogorov and Tikhomirov [7]. In this section, we will introduce a form of metric entropy with bracketing corresponding to the parameter space we are dealing with, and will provide an upper bound for the bracketing metric entropy of the approximating spaces defined in Section 2.1.

**Definition 3.1.** Let \((\Theta, \rho)\) be a separable pseudo-metric space. \( \Theta(\epsilon) \) is a finite set of pairs of functions \( \{(f_j^L, f_j^U), j = 1, \ldots, N\} \) satisfying

\[
\rho(f_j^L, f_j^U) \leq \epsilon \quad \text{for} \quad j = 1, \ldots, N, \tag{3.5}
\]

and for any \( f \in \Theta \), there is a \( j \) such that

\[
f_j^L \leq f \leq f_j^U. \tag{3.6}
\]

Let

\[
N(\epsilon, \Theta, \rho) = \min \{ \text{card} \Theta(\epsilon) : (3.5) \text{ and } (3.6) \text{ are satisfied} \}. \tag{3.7}
\]

Then, we define the metric entropy with bracketing of \( \Theta \) to be

\[
H(\epsilon, \Theta, \rho) = \log N(\epsilon, \Theta, \rho) \tag{3.8}
\]
Recall that $\Theta_1, \ldots, \Theta_I, \ldots$ are the approximating spaces defined in section 2.1. The next two lemmas will be devoted to an upper bound for the bracketing metric entropy of $\Theta_I$.

**Lemma 3.2.** Take $\rho$ to be the Hellinger distance. Let $\Theta^{\cup A_i,d}_I = \{ f \in \Theta_I : \text{the partition of } f \text{ is } \cup_{i=1}^I A_i \text{ and } \rho(f, \pi_I f_0) \leq d \}$. Then,

$$H(u, \Theta^{\cup A_i,d}_I, \rho) \leq \frac{I}{2} \log I + I \log \frac{d}{u} + c,$$

where $c$ is a constant not dependent on the binary partition.

**Proof.** Assume $\pi_I f_0 = \sum_{i=1}^I \beta_i^0 I_{A_i}$, and $f = \sum_{i=1}^I \beta_i I_{A_i}$. Because the binary partitions $\cup_{i=1}^I A_i$ and $\cup_{i=1}^I A_i^0$, and the $\beta_i^0$s are all fixed, there exits a one-to-one correspondence between any $f \in \Theta^{\cup A_i,d}_I$ and an $I$-dimensional vector $(\sqrt{\beta_i} \mu(A_i), \ldots, \sqrt{\beta_i} \mu(A_I))$. Note that,

$$\rho(f, \pi_I f_0)^2 = \sum_{i=1}^I \sum_{j=1}^I (\sqrt{\beta_i} - \sqrt{\beta_j^0})^2 \mu(A_i \cap A_j^0) \geq \sum_{i=1}^I \mu(A_i) (\sum_{j=1}^I (\sqrt{\beta_i} - \sqrt{\beta_j^0})^2 \frac{\mu(A_i \cap A_j^0)}{\mu(A_i)})^2 = \sum_{i=1}^I \mu(A_i) (\sqrt{\beta_i} - \sqrt{\sum_{j=1}^I \beta_j^0 \frac{\mu(A_i \cap A_j^0)}{\mu(A_i)})^2}.$$

We have,

$$\{(\sqrt{\beta_i} \mu(A_i), \ldots, \sqrt{\beta_i} \mu(A_I)) : \sum_{i=1}^I \sum_{j=1}^I (\sqrt{\beta_i} - \sqrt{\beta_j^0})^2 \mu(A_i \cap A_j^0) \leq d^2 \} \subset \{(\sqrt{\beta_i} \mu(A_i), \ldots, \sqrt{\beta_i} \mu(A_I)) : \sum_{i=1}^I (\sqrt{\beta_i} \mu(A_i)) - \sum_{j=1}^I \beta_j^0 \mu(A_i \cap A_j^0) \leq d^2 \} =: B^{\cup A_i,d}_I.$$  

From this we learn that,

$$N(u, \Theta^{\cup A_i,d}_I, \rho) \leq N(u, B^{\cup A_i,d}_I, \| \cdot \|_2).$$

(3.10)

Because metric entropy is invariant under translation, calculating the bracketing metric entropy of $B^{\cup A_i,d}_I$ is equivalent to calculating that of

$$\bar{B}^{\cup A_i,d}_I := \{(\sqrt{\beta_i} \mu(A_i), \ldots, \sqrt{\beta_i} \mu(A_I)) : \sum_{i=1}^I (\sqrt{\beta_i} \mu(A_i))^2 \leq d^2 \}.$$  

The unit sphere under $L_2$-norm is

$$S = \{(\sqrt{\beta_i} \mu(A_i), \ldots, \sqrt{\beta_i} \mu(A_I)) : \sum_{i=1}^I (\beta_i \mu(A_i)) \leq 1 \}.$$
The unit sphere under $L_\infty$-norm is
\[ S_\infty = \{(\sqrt{\beta_1 \mu(A_1)}, \cdots, \sqrt{\beta_I \mu(A_I)}): \max_{1 \leq i \leq I} \sqrt{\beta_i \mu(A_i)} \leq 1\}. \]

Note that $\max_{1 \leq i \leq I} \sqrt{\beta_i \mu(A_i)} \leq 1/\sqrt{T}$ implies that $\sum_{i=1}^I \sqrt{\beta_i \mu(A_i)} \leq 1$ and $\sum_{i=1}^I \beta_i \mu(A_i) \leq d^2$ implies that $\max_{1 \leq i \leq I} \sqrt{\beta_i \mu(A_i)} \leq d$, we have
\[ \frac{1}{\sqrt{T}} S_\infty \subset S \subset \tilde{B}_I^{u,d} \subset dS_\infty. \] (3.11)

Therefore,
\[ N(u, \Theta_{I}^{\cup A_i,d}, \rho) \leq N(u, \tilde{B}_I^{u,d}, \|\cdot\|_2) \leq (\frac{d\sqrt{T}}{u} + 2)^I \leq c'I^{1/2}(\frac{d}{u})^I, \]
where $c'$ is a constant not dependent on the partition. The desired result follows. \hfill \Box

**Lemma 3.3.** Let $\Theta_I^{d} = \{f \in \Theta_I: \rho(f, \pi_I f_0) \leq d\}$. Then,
\[ H(u, \Theta_I^{d}, \rho) \leq I \log p + (I + 1) \log(I + 1) + \frac{I}{2} \log I + I \log \frac{d}{u} + c, \] (3.12)
where $c$ is a constant not dependent on $I$ or $d$.

**Proof.** According to the construction of the sieve, given the size $I$, the number of possible binary partitions is upper bounded by $p^I I!$ ($p$ is the dimension of the Euclidean space). Therefore,
\[ N(u, \Theta_I^{d}, \rho) \leq p^I I! N(u, \Theta_{I}^{\cup A_i,d}, \rho) \leq c' p^I I^{1/2}(\frac{d}{u})^I, \]
and,
\[ H(u, \Theta_I^{d}, \rho) \leq I \log p + (I + 1) \log(I + 1) + \frac{I}{2} \log I + I \log \frac{d}{u} + c. \] (3.13) \hfill \Box
3.2 Lower truncation of log-likelihood ratios

In order to analyze the convergence rate of the maximum likelihood estimators, we need to deal with an empirical process indexed by the likelihood ratios. One of our focuses will lie on the likelihood ratio surface $\Pi_{j=1}^n f(Y_j)/f_0(Y_j)$, or equivalently, the log-likelihood ratio

$$L_n(f) - L_n(f_0) = \sum_{j=1}^n Z_f(Y_j), \quad (3.14)$$

where $Z_f(Y_j) = \log f(Y_j)/f_0(Y_j)$. An obstacle here is that the negative part of the log-likelihood ratio is not always bounded or has absolute moment generating functions. To handle this difficulty, we will study lower-truncated versions of $Z_f(\cdot)$ instead. Let $\tau$ be a truncation constant. The lower-truncated versions of $f$ and $Z_f$ are defined as:

$$\tilde{f} = \begin{cases} f, & \text{if } f > \exp(-\tau)f_0, \\ \exp(-\tau)f_0, & \text{if } f \leq \exp(-\tau)f_0. \end{cases} \quad (3.15)$$

$$\tilde{Z}_f = Z_{\tilde{f}} = \begin{cases} Z_f, & \text{if } Z_f > -\tau, \\ -\tau, & \text{if } Z_f \leq -\tau. \end{cases} \quad (3.16)$$

Next, we will cite several results from Wong and Shen [19] to demonstrate that after truncation, $\tilde{Z}_f$ still maintains some key properties of the log-likelihood ratio. This guarantees that the behavior of the process indexed by the truncated log-likelihood ratios does not differ much from that of the original one, and at the same time some existing techniques which fail for the original one now can be applied after the truncation. The proofs are omitted here.

The first lemma shows that after the truncation, the log-likelihood ratio still has negative expected value.

**Lemma 3.4.** Let $\gamma = 2\exp(-\tau/2)/(1 - \exp(-\tau/2))^2$. Then

$$\mathbb{E}\tilde{Z}_f \leq - (1 - \gamma) \|f^{1/2} - \pi f_0^{1/2}\|_2^2. (3.17)$$

**Proof.** See Wong and Shen [19], Lemma 4 in Section 2.

The second lemma gives a one-side large deviation inequality for the empirical process indexed by the lower-truncated log-likelihood ratios.

**Lemma 3.5.** (One-side large deviation inequality)

Let $\nu_n(\tilde{Z}_f) = n^{-1/2} \sum_{j=1}^n (\tilde{Z}_f(Y_j) - E\tilde{Z}_f(Y_j))$. Let $\mathcal{G}$ be a class of densities with bracketing Hellinger metric entropy $H(u, \mathcal{G}, \rho)$. For $t > 0$, consider the empirical process

$$\{\nu_n(\tilde{Z}_f) : f \in \mathcal{G}, \rho(f, f_0) \leq t\}$$
induced by the truncated log-likelihood ratios for \( f \in G \) inside a Hellinger ball around \( f_0 \).

For any \( t > 0, 0 < \alpha < 1 \) and \( M > 0 \), let

\[
\varphi(M, t^2, n) = M^2/8(8c_0t^2 + M/n^{1/2}),
\]

where \( c_0 \) is set to be \((\exp(\tau/2) - 1 - \tau/2)/(1 - \exp(-\tau/2))^2\). Assume that

\[
M \leq \alpha n^{1/2}t^2/4,
\]

and

\[
\int_{\alpha M/(32n^{1/2})}^t H^{1/2}(u/(2\exp(\tau/2)), G, \rho)du \leq M\alpha^{3/2}/(2^{10}(c_0 + 1/8)).
\]

Then

\[
P_{f_0}\left(\sup_{\{\|f/t^2 - f_0/t^2\|_2 \leq t, f \in G\}} \nu_n(\tilde{Z}_f) \geq M\right) \leq 3\exp(-(1 - \alpha)\varphi(M, t^2, n)),
\]

where \( P_{f_0} \) is understood to be the outer probability measure under \( f_0 \).

Proof. See Wong and Shen [19], Lemma 7 in Section 2.

\[ \square \]

### 3.3 Convergence rate

In this section, we will apply the above large deviation inequality and a change of measure argument to derive the convergence rate of \( \hat{f}_{n,I} \). Recall that in Section 2.1 we made an assumption that, for any \( f_0 \in \Theta \) there exists \( \pi_I f_0 \in \Theta_I \) such that \( \rho(f_0, \pi_I f_0) = O(I^{-r}) \). In this section, We will obtain a convergence rate relying on this assumption. Later, driven by the purposes of the applications, we will focus on more specific classes of density functions accordingly, and will further explain how this condition is satisfied.

**Theorem 3.6.** \( \hat{f}_{n,I} \) is the maximum likelihood estimator defined as the above on \( \Theta_I \). When \( r > 1/2 \), assume that \( n \) and \( I \) satisfy

\[
nI^{-2r} = o(1),
\]

and

\[
\frac{I \log I}{n/\log n} = o(1).
\]

Then the convergence rate of the \( \hat{f}_{n,I} \) is \( \max\{\left(\frac{I \log I}{n/\log n}\right)^{1/2}, I^{-r}\} \).

Proof. We will explain the change-of-measure argument first. Under \( P_{\pi_I f_0} \), assume that there exits a ball covering the maximum likelihood estimator \( \hat{f}_{n,I} \) with high probability. The ball centers at \( \pi_I f_0 \) and has radius \( \delta(n, I) \). As mentioned in the beginning of this section, we are interested in the rate at which \( \delta(n, I) \) decreases to zero as \( n \) and \( I \) increase.
In particular, if we can establish the inequality
\[
P_{\pi, f_0}(\rho(\hat{f}_{n,I}, f_0) > 2D\delta(n, I)) \leq \sigma(n, I) \tag{3.22}
\]
where \(\sigma(n, I)\) converges to zero and \(D > 1\). Then, by the triangle inequality
\[
P_{\pi, f_0}(\rho(\hat{f}_{n,I}, f_0) > AI^{-r} + 2D\delta(n, I)) \leq \sigma(n, I). \tag{3.23}
\]
For any probability density functions \(f\) and \(g\) on \((\mathcal{Y}, \mathcal{B}, \mu)\), denote the corresponding probability measures as \(P_f\) and \(P_g\). Then,
\[
\frac{1}{4}\|P_f^{(n)} - P_g^{(n)}\|_{L_1}^2 \leq \rho(P_f^{(n)}, P_g^{(n)})^2
\]
\[
= \int_{\mathcal{Y}} \left( \sqrt{\prod_{j=1}^{n} f(y_j)} - \sqrt{\prod_{j=1}^{n} g(y_j)} \right)^2 dy_1 \cdots dy_n
\]
\[
= 2 - 2\int_{\mathcal{Y}} \sqrt{f(y_1)g(y_1)} dy_1 \cdots dy_n
\]
\[
= 2 - 2(\int_{\mathcal{Y}} \sqrt{f(y_1)} dy_1 \cdots dy_n)^n
\]
\[
= 2 - 2(1 - \frac{\rho^2(f,g)}{2})^n
\]
\[
= np^2(f,g) + O((np^2(f,g))^2). \tag{3.24}
\]
In our case, \(\rho(\pi_f f_0, f_0) = O(I^{-r})\). Under the assumptions of the theorem, it is easy to check that \(np^2(\pi_f f_0, f_0) = o(1)\). Therefore,
\[
P(\rho(\hat{f}_{n,I}, f_0) > AI^{-r} + 2D\delta(n, I)) \lesssim \sigma(n, I) + 2A\sqrt{n}I^{-r}. \tag{3.25}
\]
Notice that the right hand side of the inequality (3.25) converges to zero under the assumption of the theorem. From the left hand side we know that under \(f_0\), the convergence rate is determined by \(\max\{I^{-r}, \delta(n, I)\}\).

Now, we return to the proof of (3.22) to derive an explicit expression of \(\delta(n, I)\). Lemma 3.5 is iteratively applied here to achieve the optimal rate.

Let \(C_1 := \{f \in \Theta_I : \rho(f, \pi f_0) \geq D\delta_1(n, I)\}\) and \(C_k := \{f \in \Theta_I : D\delta_k(n, I) \leq \rho(f, \pi f_0) < D\delta_{k-1}(n, I)\}\) for \(k \geq 2\).
When $k = 1$, let $\delta_1(n, I) = (\frac{I \log L}{n})^{1/4}$. Then,

\[
\mathbb{P}_{\pi, f_0}(\rho(\hat{f}_{n,I}, \pi f_0) \geq D\delta_1(n, I)) \\
\leq \mathbb{P}_{\pi, f_0}(\sup_{f \in \Theta, \rho(g, \pi f_0) \geq D\delta_1(n, I)} L_n(f) \geq L_n(\pi f_0) - \eta) \\
(\eta \text{ is assumed to be zero in our definition}) \\
\leq \mathbb{P}_{\pi, f_0}(\sup_{f \in \mathcal{C}_1} L_n(f) \geq L_n(\pi f_0)) \\
\leq \mathbb{P}_{\pi, f_0}(\sup_{f \in \mathcal{C}_1} \sum_{j=1}^n \tilde{Z}_f(Y_j) \geq 0) \\
\leq \mathbb{P}_{\pi, f_0}(\sup_{f \in \mathcal{C}_1} n^{-1/2} \sum_{j=1}^n (\tilde{Z}_f(Y_j) - \mathbb{E}\tilde{Z}_f(Y_j)) \geq n^{-1/2} \sum_{j=1}^n (-\mathbb{E}\tilde{Z}_f(Y_j)) \\
\leq \mathbb{P}_{\pi, f_0}(\sup_{f \in \mathcal{C}_1} n^{-1/2} \sum_{j=1}^n (\tilde{Z}_f(Y_j) - \mathbb{E}\tilde{Z}_f(Y_j)) \geq n^{1/2}(1 - \gamma)D^2\delta_1^2(n, I)).
\]

Here, we applied Lemma 3.4 to obtain the last inequality.

In order to apply Lemma 3.5 to bound the probability above, we need to check the conditions (3.18) and (3.19).

Note that $\|f^{1/2} - \pi f_0^{1/2}\|_2^2 = \int_Y (\sqrt{f(y)} - \sqrt{\pi f_0(y)})^2 dy = 2 - 2 \int_Y \sqrt{f(y)\pi f_0(y)} dy \leq 2$. We may set $t$ in Lemma 3.5 to be 2 here. Given the assumptions of the Theorem 3.6, $\delta_1(n, I) = (\frac{I \log L}{n})^{1/4}$ converges to zero as $n, I \to \infty$. $M$ in Lemma 3.5 is $n^{1/2}(1 - \gamma)D^2\delta_1^2(n, I)$ here. In the first place, (3.18) is satisfied because

\[
M = n^{1/2}(1 - \gamma)D^2\delta_1^2(n, I) = o(n^{1/2}t^2/4).
\]
By Lemma 3.3

\[
\int_{\frac{1}{2}(1-\gamma)D^2 n}^{t} H^{1/2}\left( \frac{1}{u/(2 \exp(\tau/2), \Theta t, \rho)du} \right) \\
\leq \int_{\frac{1}{2}(1-\gamma)D^2 n}^{t} (I \log(4p\exp(\tau/2)) + 2(I + 1) \log(I + 1) - I \log u)^{1/2} du \\
(\text{Let } \beta := \frac{\alpha(1-\gamma)D^2}{32}) \\
\sim I^{1/2} \int_{\beta \delta^2(n,t)}^{t} (\log \frac{I^2}{u})^{1/2} du \\
\leq I^{1/2}(u \sqrt{\log \frac{I^2}{u}} - \frac{\sqrt{\pi}}{2} I^2 \text{erf}(\sqrt{\log \frac{I^2}{u}})\beta^2(n,t)) \\
\leq I^{1/2}(2 \sqrt{\log \frac{I^2}{2}} - \frac{\sqrt{\pi}}{2} I^2 \text{erf}(\sqrt{\log \frac{I^2}{2}}) - \beta^2(n,t) \sqrt{\frac{I^2}{\beta^2(n,t)}} \\
+ \frac{\sqrt{\pi}}{2} I^2 \text{erf}(\sqrt{\log \frac{I^2}{\beta^2(n,t)}}) \\
\sim I^{1/2}(2 \sqrt{\log \frac{I^2}{2}} + \frac{\sqrt{\pi}}{2} I^2 \text{erf}(\sqrt{\log \frac{I^2}{\beta^2(n,t)}} - \text{erf}(\sqrt{\log \frac{I^2}{2}})) \\
\sim I^{1/2}(2 \sqrt{\log \frac{I^2}{2}} + \frac{\sqrt{\pi}}{2} I^2 (1 - \frac{\beta^2(n,t)}{\sqrt{\frac{I^2}{\beta^2(n,t)}}} - 1 + \frac{2}{\sqrt{I^2}})) \\
\sim \sqrt{I \log I}.
\]

Therefore, (3.19) is also satisfied when \( n \) and \( I \) are large enough.

Apply Lemma 3.5, we have

\[
\mathbb{P}_{\pi t f_0}(\rho(\hat{f}_{n,t}, \pi t f_0) > D \delta_1(n, I)) \\
\leq 3 \exp(- (1-\alpha) \varphi(M, I^2, n) \\
\leq 3 \exp(- (1-\alpha)(1-\gamma)^2 D^2 \frac{I \log I (\log 2)}{2^8 c_0 + 8}).
\]

When \( k \geq 2 \), we still want to apply Lemma 3.3 and establish the inequality

\[
\mathbb{P}_{\pi t f_0}(D \delta_k(n, I) \leq \rho(\hat{f}_{n,t}, \pi t f_0) < D \delta_{k-1}(n, I)) \\
\leq 3 \exp(- (1-\alpha)(1-\gamma)^2 D^2 \frac{I \log I \log n}{2^8 c_0 + 8}),
\]

where \( \delta_k = (\frac{I \log I}{n \log n})^{\omega_k} \) and \( \omega_k \) is a sequence defined by

\[
\omega_1 = 1/4 \text{ and } \omega_{k+1} = \frac{1}{2} \omega_k + \frac{1}{4}.
\]
By a similar argument as before, we have

$$\mathbb{P}_{\pi f_0}(D\delta_k(n, I) \leq \rho(f_n, I, \pi f_0) < D\delta_{k-1}(n, I)) \leq \mathbb{P}_{\pi f_0}(\sup_{f \in C_\pi} n^{-1/2} \sum_{j=1}^{n} \{\tilde{Z}_f(Y_j) - \mathbb{E}\tilde{Z}_f(Y_j)\} \geq n^{1/2}(1 - \gamma)D^2\delta^2_k(n, I)).$$

(3.29)

By our assumption, \(\delta_k(n, I) = o(\delta_{k-1}(n, I))\) for \(k \geq 2\). Then, condition (3.18) is satisfied automatically. Also by a similar argument as in the case \(k = 1\),

$$\int_{\frac{n(I-1)}{2}}^{\frac{n(I-1)}{2}} \frac{D\delta_k(n, I)}{\beta \delta^2_k(n, I)} H^{1/2}(u/(2 \exp(\tau/2)), \Theta_1, \rho)du \sim I^{1/2}(D\delta_k(n, I))/\sqrt{\log \log \frac{I^2}{D\delta_k(n, I)} - \beta \delta^2_k(n, I)/\beta \delta^2_k(n, I)}$$

$$+ \frac{\sqrt{\pi} I^2}{2} \left( \frac{D\delta_k(n, I)}{\beta \delta^2_k(n, I)} - \frac{\beta \delta^2_k(n, I)/I^2}{\beta \delta^2_k(n, I)} \right)$$

$$\sim I^{1/2}\delta_k(n, I) \sqrt{\log \frac{I^2}{\beta \delta^2_k(n, I)}},$$

Replace \(\delta_k(n, I)\) and \(\delta_k(n, I)\) by \((\frac{1}{n/\log n})^{\omega_{k-1}}\) and \((\frac{1}{n/\log n})^{\omega_k}\) respectively, after some calculation, it can be checked that condition (3.19) is also satisfied when \(n\) and \(I\) are large enough.

Note that

$$\phi(n^{1/2}(1 - \gamma)D^2\delta^2_k(n, I), (D\delta_k(n, I))^2, n)$$

$$= \frac{n(1 - \gamma)^2 D^4\delta^4_k(n, I)}{8(8c_0 D^2\delta^2_k(n, I) + (1 - \gamma) D^2\delta^2_k(n, I))}$$

$$\geq \frac{n(1 - \gamma)^2 D^2\delta^2_k(n, I)}{8(8c_0 \delta^2_k(n, I) + (1 - \gamma) \delta^2_k(n, I))}$$

$$\geq \frac{n(1 - \gamma)^2 D^2(\frac{I log I}{n/\log n})^{2\omega_{k-1}+1}}{(2^c c_0 + 8(1 - \gamma))(\frac{I log I}{n/\log n})^{2\omega_{k-1}}}$$

$$\geq \frac{(1 - \gamma)^2 D^2}{2^{c_0}c_0 + 8} I \log I \log n.$$
then $D\delta_{K(n,I)} \leq 2D\delta(n,I)$. We also have

$$\mathbb{P}_{\pi,f_0}(\rho(\hat{f}_{n,I}, \pi_I f_0) \geq 2D\delta(n,I)) \leq \mathbb{P}_{\pi,f_0}(\rho(\hat{f}_{n,I}, \pi_I f_0) \geq D\delta_1(n,I)) + \sum_{k=2}^{K(n,I)} \mathbb{P}_{\pi,f_0}(D\delta_k(n,I) \leq \rho(\hat{f}_{n,I}, \pi_I f_0) < D\delta_{k-1}(n,I)) \leq 3K(n,I) \exp(-\frac{(1-\alpha)(1-\gamma)^2}{2^{8c_0+8}} I \log I \log n).$$

If $\sigma(n,I)$ is set to be $3K(n,I) \exp(-\frac{(1-\alpha)(1-\gamma)^2}{2^{8c_0+8}} I \log I \log n)$, then this is exactly (3.22). Therefore,

$$\mathbb{P}(\rho(\hat{f}_{n,I}, f_0) > AI^{-r} + 2D(\frac{I \log I}{n/\log n})^{1/2} \leq \sigma(n,I) + 2A\sqrt{nI^{-r}}, (3.30)$$

and the convergence rate of $\hat{f}_{n,I}$ is $\max\{(\frac{I \log I}{n/\log n})^{1/2}, I^{-r}\}$.

**Corollary 3.7.** $\hat{f}_{n,I}$ is the maximum likelihood estimator defined before. Assume that $r > 1/2$. When $I$ is of the order $(\frac{n}{\log n})^{1/2+2\xi}$ ($\xi$ is positive and can be arbitrarily small), an optimal convergence rate can be achieved, which is $(\frac{\log n}{n})^{\frac{1}{2} - \frac{1}{4} + \xi}$.

**Proof.** It is easy to check that both condition (3.20) and condition (3.21) are satisfied. The final rate depends on the orders of $I^{-r}$ and $(\frac{I \log I}{n/\log n})^{1/2}$. Under our assumptions, $I^{-r} = o((\frac{I \log I}{n/\log n})^{1/2})$. The best achievable rate is $(\frac{\log n}{n})^{\frac{1}{2} - \frac{1}{4} + \xi}$.

**4 An application to variable selection**

For high dimensional data analysis, selecting significant variables greatly contributes to simplifying the model, improving model interpretability, and reducing overfitting. Here, the variable selection problem is formulated as follows. Assume $f_0$ is a $p$-dimensional density function and it only depends on $\tilde{p}$ variables, but we do not know which $\tilde{p}$ variables in advance. We apply the multivariate density estimation method here to estimate the $p$-dimensional density. The essence of our method is to learn how to partition the support of the true density function, and then to estimate the density on each subregion separately. Because the support of the true density lies in a $\tilde{p}$-dimensional space, we may conjecture that the corresponding convergence rate only depends on the effective dimension. In Section 4.3, an exact calculation will be carried out to reveal that this is indeed the case. In Section 4.4, we will also compare the rate with existing results of the convergence rate by using the kernel density estimator.
The theorem regarding convergence rate in the previous section relies on the rate at which the true density function can be approximated. More specifically, we make the assumption that for any $f_0 \in \Theta$, there exists $\pi_I f_0 \in \Theta_I$ such that $\rho(f_0, \pi_I f_0) = O(I^{-r})$. Questions raised here are which class of density functions satisfies this assumption, and what is the corresponding approximation rate $r$.

The Haar transform of a density function will be introduced in Section 4.1. The approximation of the true density will be obtained by selecting the terms in the Haar expansion according to certain criteria, and it will be shown that the approximation rate only depends on the effective dimension. This scheme is motivated by Strömberg [17], in which tensor Haar basis with large support is shown to be efficient in representing functions on $[0, 1]^p$. More precisely, if the function satisfies the bounded mixed variation condition, then it can be approximated to error $\epsilon > 0$ using no more than $\frac{1}{\epsilon} (\log(\frac{1}{\epsilon}))^{p-1}$ terms, and the volume of the supporting rectangle of each wavelet basis function involved in the approximation is greater than $\epsilon$. However, the result is restrictive in application since the mixed derivative is not rotationally invariant. In Gavish and Coifman [5], a similar framework is extended to matrix approximation. A significant improvement is that, in their paper, the bounded mixed variation condition is replaced by a Mixed-Hölder condition, which is more natural and accessible. For two-dimensional discrete analyses of matrices, they also show that the Mixed-Hölder condition is connected to the decay rate of tensor product wavelet coefficient.

The idea of controlling wavelet coefficient decay will be further developed in Section 4.2. It will lead to the characterization of the density class that can be well approximated by the densities on binary partitions. Before providing an accurate description, we will start with the notations.

### 4.1 Tensor Haar basis

Haar basis is the simplest but widely used wavelet basis. In one dimension, the Haar wavelet’s mother wavelet function is

$$
\psi(y) = \begin{cases} 
1 & \text{if } 0 \leq y < 1/2, \\
-1 & \text{if } 1/2 \leq y < 1, \\
0 & \text{otherwise.}
\end{cases}
$$

And its scaling function is

$$
\phi(y) = \begin{cases} 
1 & \text{if } 0 \leq y < 1, \\
0 & \text{otherwise.}
\end{cases}
$$

For any $l \in \mathbb{N}$ and $0 \leq k < 2^l$, the Haar function is defined to be

$$
\psi_{l,k}(y) = 2^{l/2} \psi(2^l y - k).
$$

Then Haar basis $\Psi$ is the collection of all Haar functions together with the scale function. Namely,

$$
\Psi = \{\phi\} \cup \{\psi_{l,k}, l \in \mathbb{N}, 0 \leq k < 2^l\}.
$$
It is an orthonormal basis for Hilbert space $L^2[0, 1]$.

Turning to the high dimensional settings, we will obtain an orthonormal basis for $L^2[0, 1]^p$ by using the fact that the Hilbert space $L^2[0, 1]^p$ is isomorphic to the tensor product of $p$ one-dimensional spaces. In detail, if $X_1, \cdots, X_p$ are $p$ copies of $L^2[0, 1]$ and $\Psi_1, \cdots, \Psi_p$ are Haar basis of these spaces respectively, then $L^2[0, 1]^p$ is isomorphic to $\bigotimes_{i=1}^p X_i$. Define tensor Haar basis $\Psi$ by

$$\Psi = \{\psi : \psi = \prod_{i=1}^p \psi_i, \psi_i \in \Psi_i\}.$$  

From the property of tensor product of Hilbert spaces, we know that $\Psi$ is an orthonormal basis for $L^2[0, 1]^p$. In the rest of this section, our calculation will heavily rely on the expansion of a $p$-dimensional density function with respect to the tensor Haar basis.

### 4.2 Characterization of a special density class

For any $f \in \Theta$, $\sqrt{f} \in L^2[0, 1]^p$. Because the topology on the parameter space is induced by Hellinger distance (the calculation of Hellinger distance involves $\sqrt{f}$ instead of $f$), we will expand $\sqrt{f}$ with respect to the tensor Haar basis. Let $g = \sqrt{f}$. Then

$$g = \sum_{\psi} <g, \psi > \psi,$$  

where $<g, \psi > = \int g(y)\psi(y)dy$.

For each tensor Haar function $\psi$, let $R(\psi)$ denote its supporting rectangle. The special density class is defined to be

$$\Theta^* = \{f \in \Theta : |<\sqrt{f}, \psi>| \leq C|R(\psi)|^{\lambda+1/2} \text{ for all } \psi\},$$

where $C$ is a constant, $|R(\psi)|$ denotes the volume of the rectangle and $\lambda$ is a positive constant which is closely related to final approximation rate.

### 4.3 Convergence rate

A result on the rate at which $p$-dimensional density functions can be approximated will be provided first.

**Theorem 4.1.** $\Theta^*$ and $\Theta_I$ are defined as above. The topology on these spaces are induced by Hellinger distance. For any $f_0 \in \Theta^*$, there exists $\pi f_0 \in \Theta_I$, such that $p(f_0, \pi f_0) = O(1^{-\lambda/p})$, where $\lambda$ is the positive constant defined in the Section 4.2 and $p$ is the dimension of the Euclidean space.

**Proof.** Let $g_0 = \sqrt{f_0}$. we can expand $g_0$ with respect to tensor Haar basis. The expansion can be written as $g_0 = \sum_{\psi} <g_0, \psi > \psi$.

Let $g_\epsilon = \sum_{\psi : |R(\psi)| < \epsilon} <g_0, \psi > \psi$. Then $g_\epsilon$ is an approximation of $g_0$ obtained by requiring that the volumes of the supporting rectangles of the involved
wavelet basis functions are greater than $\epsilon$. We will derive an approximation rate in terms of $\epsilon$ first. Then the lower bound of the volume is transformed into an upper bound for the size of the partition, yielding an approximation rate in terms of the size of the partition. Note that $g_\epsilon$ is not a density function, but it is easier to work with. Let $\tilde{g}_\epsilon = g_\epsilon / \|g_\epsilon\|_2$ be the normalization of $g_\epsilon$. The upper bounds for the approximation errors $\rho(f_0, g_\epsilon^2)$ and $\rho(f_0, \tilde{g}_\epsilon^2)$ will be derived successively.

Before delving into the proof, we will introduce some notations. For each supporting rectangle $|R(\psi)|$, the lengths of its edges should be powers of $1/2$. We may assume that $\psi = \prod_{i=1}^{p} \psi_i$ and for each $\psi_i$ the length of its supporting interval is $(1/2)^{l_i}$. Let $R^{l_1, \ldots, l_p}$ denote the collection of the rectangles for which the lengths of the edges are $(1/2)^{l_1}, \cdots, (1/2)^{l_p}$, respectively.

Recall that $f_0$ satisfies the condition

$$| < \sqrt{f_0}, \psi > | \leq C |R(\psi)|^{\lambda + 1/2}$$

for all $\psi$. (4.31)

Then,

$$\rho^2(f_0, g_\epsilon^2) = \|g_0 - g_\epsilon\|^2 = \| \psi |R(\psi)| < \epsilon \psi < g_0, \psi \|^2$$

$$= \sum_{|R(\psi)| < \epsilon} < g_0, \psi >^2$$

$$\leq C^2 \sum_{|R(\psi)| < \epsilon} |R(\psi)|^{2\lambda + 1}$$

$$\leq 2p C^2 \sum_{l_1, \ldots, l_p \in R^{l_1, \ldots, l_p}, |R| < \epsilon} |R|^{2\lambda + 1}.$$  (4.32)

The last inequality follows from the fact that, given a supporting rectangle, there are at most $2^p$ basis functions defined on it. Let $N = \lceil \log_2 \epsilon \rceil$,

$$= 2p C^2 \sum_{l_1, \ldots, l_p \geq N} \sum_{R \in R^{l_1, \ldots, l_p}} |R|^{2\lambda + 1}$$

$$= 2p C^2 \sum_{l_1, \ldots, l_p \geq N} (\frac{1}{2})^{2\lambda(l_1 + \cdots + l_p)} \sum_{R \in R^{l_1, \ldots, l_p}} |R|$$

$$= 2p C^2 \sum_{l_1, \ldots, l_p \geq N} (\frac{1}{2})^{2\lambda(l_1 + \cdots + l_p)}.$$  (4.33)
The last equality is obtained by plugging in \( \sum_{R \in R^1, \ldots, p} |R| = 1 \). Note that

\[
\sum_{l_1 + \cdots + l_p \geq N} \left( \frac{1}{2} \right)^{2\lambda(l_1 + \cdots + l_p)} \leq \sum_{l_1=0}^{N} \sum_{l_2=0}^{N-l_1} \cdots \sum_{l_p=N-(l_1+\cdots+l_{p-1})}^{+\infty} \left( \frac{1}{2} \right)^{2\lambda(l_1 + \cdots + l_p)}
\]

\[
+ \sum_{l_1=0}^{+\infty} \sum_{l_2=0}^{N-l_1} \cdots \sum_{l_{p-1}=N-(l_1+\cdots+l_{p-2})}^{+\infty} \sum_{l_p=0}^{+\infty} \left( \frac{1}{2} \right)^{2\lambda(l_1 + \cdots + l_p)}
\]

\[
+ \cdots
\]

\[
+ \sum_{l_1=0}^{+\infty} \sum_{l_2=0}^{+\infty} \cdots \sum_{l_{p-1}=0}^{+\infty} \sum_{l_p=0}^{+\infty} \left( \frac{1}{2} \right)^{2\lambda(l_1 + \cdots + l_p)}
\]

\[
\leq (N + 1)^{p-1} \frac{\left( \frac{1}{2} \right)^{2\lambda N}}{1 - 2^{-2\lambda}} + (N + 1)^{p-2} \frac{\left( \frac{1}{2} \right)^{2\lambda N}}{(1 - 2^{-2\lambda})^2} + \cdots + \frac{\left( \frac{1}{2} \right)^{2\lambda N}}{(1 - 2^{-2\lambda})^p}
\]

\[
= \frac{\left( \frac{1}{2} \right)^{2\lambda N} (N + 1)^p}{(N + 1) (1 - 2^{-2\lambda}) - 1}
\]

\[
\leq C'' \epsilon^{2\lambda (\log_2 \epsilon)^p}.
\]

From this, we know that

\[
\rho^2(f_0, g^2_\epsilon) = \|g_0 - g_\epsilon\|_2^2 \leq 2^p C'' \epsilon^{2\lambda (\log_2 \epsilon)^p}.
\]

We normalize \( g_\epsilon \) to \( \tilde{g}_\epsilon \), then

\[
\rho^2(f_0, \tilde{g}^2_\epsilon) = \|g_0 - \tilde{g}_\epsilon\|_2^2
\]

\[
= \|g_0 - g_\epsilon\|_2^2 + (1 - \frac{1}{\|g_\epsilon\|_2})^2 \|g_\epsilon\|_2^2
\]

\[
\leq \|g_0 - g_\epsilon\|_2^2 + 1 - \|g_\epsilon\|_2^2
\]

\[
= 2\|g_0 - g_\epsilon\|_2^2.
\]

The last equality is obtained by using \( \|g_0 - g_\epsilon\|_2^2 + \|g_\epsilon\|_2^2 = \|g_0\|_2^2 = 1 \). Therefore,

\[
\rho^2(f_0, \tilde{g}^2_\epsilon) = \|g_0 - \tilde{g}_\epsilon\|_2^2 \leq 2^p C'' \epsilon^{2\lambda (\log_2 \epsilon)^p},
\]

where \( C'' \) is a constant.

Next, we will transform the lower bound on the volume of the supporting rectangles to an upper bound on the size of the partition and derive the approximation rate in terms of the latter one.

If we require the volumes of the supporting rectangles be greater than \( \epsilon \), then the length of the edge can not be smaller than \( 2^{-\lfloor \log_2 \epsilon \rfloor} \). The size of the partition supporting \( \tilde{g}_\epsilon \) can be bounded by \( 2^p 2^{\log_2 \epsilon} = 2^p \epsilon^{-p} \). There is a coefficient \( 2^p \) in front. This is the case because given a supporting rectangle, the
positive and negative parts of the tensor Haar basis defined on it will further divide the original rectangle into smaller subregions and the number of such subregions is at most $2^p$.

Given the size of the partition $I$, we can determine $\epsilon$ by solving $2^p \epsilon^{-p} = I$ and define $\tilde{g}_\epsilon \in \Theta_I$ as above. Then from (4.36) we reach a conclusion that $\tilde{g}_\epsilon$ is an approximation satisfying $\rho(f_0, \tilde{g}_\epsilon^2) = O(I^{-\lambda/p})$. This finishes the proof. \( \square \)

**Theorem 4.2. (An application to variable selection)** Assume that $f_0$ is a $p$-dimensional density function. It only depends on $\tilde{p}$ arguments which are not specified in advance. $f_0$ satisfies the condition that for each tensor Haar function, $| < f_0, \psi > | \leq C |R(\psi)|^{k+1/2}$. If we apply the multivariate density estimation method to this problem, an achievable and optimal convergence rate will be $(\log n/n)^{\frac{1}{2} - \frac{\tilde{p}}{2\lambda}}$, where $\xi$ is positive and can be arbitrarily small.

**Proof.** This follows Corollary 3.7 and Theorem 4.1 directly. \( \square \)

### 4.4 Discussion

If we estimate the density function by a traditional nonparametric method such as a kernel density estimator, then an achievable and optimal convergence rate is obtained by in Stone [16]. Let $\Gamma = (\Gamma_1, \ldots, \Gamma_p)$ denote a $p$-tuple of nonnegative integers and set $[\Gamma] = \Gamma_1 + \cdots + \Gamma_p$. $D^\Gamma$ denotes the differential operator defined by

$$D^\Gamma = \frac{\partial^{[\Gamma]}}{\partial^{\Gamma_1}y_1 \cdots \partial^{\Gamma_p}y_p}.$$  

Let $\alpha$, $\zeta$, and $L$ be real constants such that $\alpha$ is a nonnegative integer, $0 < \zeta \leq 1$ and $L > 0$. If $f_0$ is an $\alpha$-times continuously differentiable probability density on $\mathbb{R}^p$ such that

$$|D^\Gamma f_0(y_2) - D^\Gamma f_0(y_1)| \leq L |y_2 - y_1|^{\zeta},$$

where $[\Gamma] = \alpha$. Set $\kappa = \alpha + \zeta$, then the achievable and optimal convergence rate is $(\log n/n)^{\frac{\kappa}{\kappa + \tilde{p}}}$. 

If we compare this convergence rate to the one we obtain in Theorem 4.2, a major difference is that the rate of the kernel density estimator depends on the full dimension, while only the effective dimension affects the rate of our method. This implies that in an extreme case of $\tilde{p} \ll p$, if we fix the effective dimension and increase the full one, the convergence rate of the traditional nonparametric density estimator decreases quickly. However, our method can still achieve stable performances. This theoretical analysis is in line with the example cited in Section 1.2. More straightforwardly, the advantage of our method is demonstrated by the following facts: in the partition learning stage, it can quickly restrict our attention to those determining variables. Ideally, it can estimate the density as a function of the effective variables alone, although they are not specified in advance.
5 An application to spatial adaptation

This section is devoted to the case in which the density concentrates spatially. More explicitly, we assume that the density highly aggregates in several locations while dropping quickly away from these locations. Mathematically, this is to say, the density function satisfies a type of spatial sparsity. In Section [5.2] we will provide a more rigorous characterization of this class of density functions, and will illustrate this characterization by several examples. In Section [5.3] we will apply the previous multivariate density estimation method to adapt to the unknown sparsity, and will calculate the corresponding convergence rate. As a preliminary, in Section [5.1] we first discuss the construction of another system of high-dimensional Haar basis.

5.1 An alternative construction of high-dimensional Haar basis

In the previous section, with the aid of the high-dimensional Haar basis constructed via tensor product, we analyzed the approximation properties of the binary partitions. Here, an alternative construction will be introduced. This is necessary because the tensor product one mixes the resolutions under high dimension, and the overlaps between the basis functions could exponentially increase the size of the corresponding binary partition. We take two dimensional case as an illustration to show how the alternative system is built. This construction can be extended to the high dimensional cases as well.

Let the Haar wavelet’s mother wavelet function $\psi$ and its scaling function $\phi$ be the same as defined in Section [4.1]. The two-dimensional scaling function is defined to be

$$\phi \phi(x, y) := \phi(x) \phi(y),$$

and three wavelet functions are

$$\phi \psi(x, y) := \phi(x) \psi(y),$$

$$\psi \phi(x, y) := \psi(x) \phi(y),$$

$$\psi \psi(x, y) := \psi(x) \psi(y).$$

If we use a superscript $l$ to index the scaling level of the wavelet function and subscripts $i$ and $j$ (i and j can be 0 or 1) to denote the horizontal and vertical translations respectively, then the scales and translates of the three wavelet functions $\phi \psi$, $\psi \phi$ and $\psi \psi$ are defined to be

$$\phi \psi_{ij}^l(x, y) := 2^l \phi \psi(2^l x - i, 2^l y - j),$$

$$\psi \phi_{ij}^l(x, y) := 2^l \psi \phi(2^l x - i, 2^l y - j),$$

$$\psi \psi_{ij}^l(x, y) := 2^l \psi \psi(2^l x - i, 2^l y - j).$$

These functions together with the single scaling function $\phi \phi$ define the two-dimensional Haar wavelet basis $\Psi$.  

23
5.2 Spatial sparsity of the density functions

Let \( f \) be a \( p \)-dimensional density function and \( \Psi \) the \( p \)-dimensional Haar basis constructed as above. As that in Section 4.2, we will work with \( g = \sqrt{f} \) first. Note that \( g \in L^2(0, 1)^p \), thus we can expand \( g \) with respect to \( \Psi \) as \( g = \sum_{\psi \in \Psi} <g, \psi > \psi \) (note that here \( \psi \) is a basis function instead of the Haar wavelet’s mother wavelet function defined above). We rearrange this summation by the size of wavelet coefficient. This is to say, we will order the coefficients as the following

\[
| < g, \psi^{(1)} > | \geq | < g, \psi^{(2)} > | \geq \cdots \geq | < g, \psi^{(k)} > | \geq \cdots,
\]

then the sparsity condition imposed on the density functions is that the wavelet coefficients can be controlled by a power-law decay

\[
| < g, \psi^{(k)} > | \leq Ck^{-\beta} \text{ for all } k \in \mathbb{N} \text{ and } \beta > 1/2, \tag{5.37}
\]

where \( C \) is a constant.

This condition has been widely used to characterize the sparsity of signals and images (Abramovich et. al. [1] and Candes and Tao [3]). In particular, in the paper by DeVore, Jawerth and Lucier [4], it was shown that for two-dimensional cases, when \( \beta > 1/2 \), this condition reasonably captures the sparsity of real world images.

Next we want to use several examples to illustrate how this condition implies the spatial sparsity of the density function.

**Example 0**

Assume that the we are studying a three dimensional density function. All the mass is localized in a dyadic cube. Without loss of generality, we assume that \( f_0 = 64I_{\{0 \leq x, y, z < 1/4\}} \). In three dimensions, the single scaling function is \( \phi \phi \phi \), and the seven wavelet functions which are defined as

\[
\begin{align*}
\chi_1 &= \psi \phi \phi, \\
\chi_2 &= \phi \psi \phi, \\
\chi_3 &= \phi \phi \psi, \\
\chi_4 &= \psi \psi \phi, \\
\chi_5 &= \psi \phi \psi, \\
\chi_6 &= \phi \psi \psi, \\
\chi_7 &= \psi \psi \psi.
\end{align*}
\]

If we still use the superscript to denote the scaling level and the subscripts to denote the spatial translations, then the expansion of \( f_0 \) with respect to the nonstandard Haar basis is

\[
f_0 = \phi \phi \phi + \sum_{k=1}^{7} \chi_{k,000}^{(0)} + 2\sqrt{2} \sum_{k=1}^{7} \chi_{k,000}^{(1)}.
\]

The coefficients display a decaying trend, although the number of them is finite. More generally, any density function with finite Haar coefficients will satisfy the condition \[5.37\].
Example 1

Assume that the two-dimensional true density function is

\[
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \frac{2}{5} \mathcal{N} \left( \begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix} , 0.05^2 I_{2 \times 2} \right) + \frac{3}{5} \mathcal{N} \left( \begin{pmatrix} 0.75 \\ 0.75 \end{pmatrix} , 0.05^2 I_{2 \times 2} \right).
\]

We perform the Haar transform to this density. The heatmap of the density function is displayed in Figure 2 and the plot of the Haar coefficient is shown in Figure 3. The left panel in Figure 3 is the plot of all the coefficients to level ten from low resolution to high resolution. The middle one is the sorted coefficients according to their absolute value. And the right one is the plot on log-scale. From this we can clearly see that the power-decay law is satisfied by the Haar coefficients, and an empirical estimation of the corresponding \( \beta \) can be obtained in this case.

Example 2

Let the three-dimensional density function be

\[
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \frac{2}{5} \mathcal{N} \left( \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} , \begin{pmatrix} 0.05^2 & 0.03^2 & 0 \\ 0.03^2 & 0.05^2 & 0 \\ 0 & 0 & 0.05^2 \end{pmatrix} \right) + \frac{3}{5} \mathcal{N} \left( \begin{pmatrix} 0.75 \\ 0.75 \\ 0.75 \end{pmatrix} , 0.05^2 I_{3 \times 3} \right).
\]

In this example, we impose some correlation at one local mode. Haar transform is performed and the behavior of the Haar coefficients is summarized in Figure 4. The arrangement of the plots is the same as that in the previous example. For this three-dimensional example, the power-decay law is still satisfied.
Figure 3: Plots of the 2-dimensional Haar coefficients

Figure 4: Plots of the 3-dimensional Haar coefficients
5.3 Convergence rate

Assume that $f_0$ is the $p$-dimensional density function we are trying to estimate. Its mass spatially concentrate in several small regions. We apply the maximum likelihood density estimator based on adaptive partitioning to detect this unknown structure. A lemma on the rate at which the true density can be approximated by the binary partitions is given at first.

**Lemma 5.1.** Suppose $f_0$ is a $p$-dimensional density function. $g_0 = \sqrt{f_0}$ satisfies the condition (5.37). Then there exists a sequence of $f_I \in \Theta_I$, such that $\rho(f_0, f_I) = O(I^{-\beta/2})$.

**Proof.** Let $g_K = \sum_{k=1}^{K} < g_0, \psi(k) > \psi(k)$. From condition (5.37) we have

$$
\rho^2(f_0, g_K^2) = \|g_0 - g_K\|^2_2 = \| \sum_{k=K+1}^{+\infty} < g_0, \psi(k) > \psi(k) \|^2_2
\leq \sum_{k=K+1}^{+\infty} < g_0, \psi(k) >^2
\leq C^2 \sum_{k=K+1}^{+\infty} k^{-2\beta} \leq \frac{C^2}{2\beta - 1} K^{-(2\beta - 1)}
$$

(5.38)

Then we normalize $g_K$ to $\tilde{g}_K$, and get

$$
\rho^2(f_0, \tilde{g}_K^2) \leq 2\|g_0 - g_K\|^2_2 \leq \frac{2C^2}{2\beta - 1} K^{-(2\beta - 1)}
$$

(5.39)

Note that as we mentioned before, given a supporting rectangle, the positive and negative parts of the Haar basis function defined on it can further divide the original rectangle into smaller subregions, and the total number of such subregions is upper bounded by $2^p$. Therefore the density function $\tilde{g}_K$ can be supported by a binary partition of size $2^p K$. Replace $K$ in (5.39) by $1/2^p$, we reach the desired result of approximation rate.

Next theorem calculates the convergence rate for this case.

**Theorem 5.2.** (An application to spatial adaptation) Assume $f_0$ is the same as defined in Lemma (5.1). If the maximum likelihood density estimator based on adaptive partitioning is applied here to estimate the true density function, then the best achievable rate is $(\log n)^{\frac{1}{2}} - \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{2\beta - 1}} - \xi$, where $\xi$ is positive and can be arbitrarily small.

**Proof.** This follows Corollary 3.7 and Lemma 5.1 directly.  

From the theorem we see that the convergence rate only depends on how fast the coefficients decay as opposed to the dimension of the sample space. This agrees with the data-adaptive nature of this multivariate density estimation method.
Acknowledgements

The authors would like to thank Xiaotong Shen and Matan Gavish for helpful discussions.

References


