Journal of Machine Learning Research 24 (2023) 1-64 Submitted 1/20; Published 1/23

Convergence Rates of a Class of Multivariate Density Estimation Methods Based on Adaptive Partitioning

Linxi Liu LIL155@PITT.EDU
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260-4601, USA

Dangna Li LDANGNA@ALUMNI.STANFORD.EDU
Institute for Computational & Mathematical Engineering
Stanford University
Stanford, CA 94305-4042, USA

Wing Hung Wong WHWONG@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305-4020, USA

Editor: Erik Sudderth

Abstract

Density estimation is a building block for many other statistical methods, such as classification, nonparametric testing, and data compression. In this paper, we focus on a nonparametric approach to multivariate density estimation, and study its asymptotic properties under both frequentist and Bayesian settings. The estimated density function is obtained by considering a sequence of approximating spaces to the space of densities. These spaces consist of piecewise constant density functions supported by binary partitions with increasing complexity. To obtain an estimate, the partition is learned by maximizing either the likelihood of the corresponding histogram on that partition, or the marginal posterior probability of the partition under a suitable prior. We analyze the convergence rate of the maximum likelihood estimator and the posterior concentration rate of the Bayesian estimator, and conclude that for a relatively rich class of density functions the rate does not directly depend on the dimension. We also show that the Bayesian method can adapt to the unknown smoothness of the density function. The method is applied to several specific function classes and explicit rates are obtained. These include spatially sparse functions, functions of bounded variation, and Hölder continuous functions. We also introduce an ensemble approach, obtained by aggregating multiple density estimates fit under carefully designed perturbations, and show that for density functions lying in a Hölder space ($\mathcal{H}^{1,\beta}, 0 < \beta \leq 1$), the ensemble method can achieve minimax convergence rate up to a logarithmic term, while the corresponding rate of the density estimator based on a single partition is suboptimal for this function class.

Keywords: Multivariate Density Estimation, Sieve Estimates, Adaptive Partitioning, Convergence Rate, Posterior Concentration Rate,

1. Introduction

Density estimation is a fundamental problem in statistics. Once an explicit estimate of the density function is obtained, various kinds of statistical inference can follow, including classification, non-parametric testing, clustering, and data compression. A large collection of parametric and nonparametric methods have been introduced for estimating the density. However, during the past decade, the increasing dimension and data scale have posed great challenges to the existing methods. For instance, a fixed parametric family, such as multivariate Gaussian, may fail to capture the spatial features of the true density function under high dimensions. While the kernel density estimator, as a popular nonparametric method, may suffer from the difficulty of choosing appropriate bandwidths (Jones et al., 1996). The recently introduced generative adversarial networks (GANs, Goodfellow et al. 2014) can be viewed as a scalable and widely applicable method for density estimation. Uppal et al. (2019) have shown that a GAN based on sufficiently large fully-connected neural networks with ReLU activations learns Besov probability distributions at the minimax rate. However, the rate is not adaptive. And in practice, GANs usually require extensive tuning to perform well.

In this paper, we study a nonparametric method for multivariate density estimation. This is a class of estimators which employs simple, but still flexible, binary partitions to adapt to the underlying density function. Under both frequentist and Bayesian settings, we carry out analysis of convergence rate in order to quantify the performance of this method as the dimension increases or the regularity of the true density function varies. The results suggest that the Bayesian estimator is adaptive to the unknown smoothness of the density function, and the method has the potential to overcome the difficulty imposed by the high dimensionality, especially when the dimension is only moderately large (saying 5 to 50) and the density function exhibits certain spatial features that can be leveraged of. A carefully designed ensemble approach is also investigated in the paper. We show that for a class of smooth density functions, the ensemble can achieve faster convergence compared to the estimator constructed on a single binary partition.

1.1 Challenges in Multivariate Density Estimation

Quite a few 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, 1956; Parzen, 1962), 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 estimate becomes very sensitive to the choice of the window size and the shape of the kernel. To obtain 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 the case for the kernel which itself is a multidimensional function. As a result, the performance of current kernel estimators deteriorates rapidly as the dimension increases.

The difficulty caused by high dimensionality is also revealed in a classic result by Stone (1980). In this paper, it was shown that the optimal (uniform) rate of convergence for density in the p-dimensional space, when the density is k times continuously differentiable,

is of the order $n^{-\alpha}$, where $\alpha = k/(2k+p)$. When p is small and the density is smooth (that is k is large), then methods such as kernel density estimator can achieve a convergence rate almost as good as the parametric rate of $n^{-1/2}$. However, when p is large, then even if the density has many bounded derivatives, the best possible rate will still be unacceptably slow. 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 densities 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 controls the complexity (in a sense to be made precise later) 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 as sensitive to p as that of the traditional methods, satisfying the property that $\gamma(r) \uparrow \frac{1}{2}$ as $r \uparrow \infty$. Since this rate is not sensitive to p, 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 1 and Theorem 4, and verified by Example 5 from Section 6.

Another difficulty of multivariate density estimation lies in the detection and characterization of local features. The kernel density estimator may smooth out the local changes, especially when the bandwidth is chosen inappropriately. Estimation based on the basis expansion (Donoho et al., 1996; Tribouley, 1995) allows us to study the density at different scales, but as the dimension increases the number of tensor-product basis functions can be prohibitively large. As opposed to these two types of methods, by learning an adaptive partition of the sample space, the partition based approach can provide an informative summary of the density, and allows the examination and inference at different resolutions (Soriano and Ma, 2017; Li and Ma, forthcoming).

1.2 A Glimpse of Our Contribution

In this paper we study the asymptotic properties of a class of multivariate density estimation methods based on adaptive partitioning and have mainly focused on two types of estimators: the sieve maximum likelihood estimator (sieve MLE) and the Bayesian estimator. As mentioned in Section 1.1, we start by generally formulating a complexity index for a density, denoted as r. A large r implies low complexity in the sense that the density can be approximated at a fast rate by piecewise constant density functions as the underlying partition becomes finer and finer. Later, we calculate the value of r explicitly for three specific density classes, which include the Hölder space, the space of functions with bounded variation, and the density functions with sparsity which is characterized by a weak-ℓ_q constraint on Haar wavelet coefficients.

For the sieve MLE, given the complexity of true density function, a matching partition size is needed in order to achieve a fast convergence. Our analysis shows that up to a factor of $\log n$ the achievable rate is $n^{-\gamma}$, with $\gamma = r/(2r+1)$. Therefore when r is large, our estimate will converge to the true density at a rate close to the parametric rate of $n^{-1/2}$, not directly depending on the dimension p of the sample space. This is in contrast with the achievable convergence rate under smoothness condition (Stone, 1980), which decays rapidly as the dimensional p increases.

We have also studied a class of Bayesian estimators, since it is well known that sieve MLEs are closely related to penalized estimates which are in turn related to Bayesian methods (Wahba, 1978; Shen, 1997; Shen and Wasserman, 2001). We show that, under an appropriate prior distribution, the posterior distribution concentrates on a shrinking Hellinger ball around the true density, where the radius of the ball is $O(n^{-\gamma})$ with $\gamma = r/(2r+1)$ up to a logarithmic term.

Although the concentration rate of the Bayesian method is the same as the convergence rate of the sieve MLE, there is an important difference in terms of adaptivity. The rate is achieved by the Bayesian method without requiring any knowledge of the constant r that characterizes the complexity of the true density function, while the sieve MLE can only achieve this same rate if the size of the sieve grows at a rate that depends on r, specifically, the size of the partition must be of order $n^{1/(2r+1)}$. This implies that the Bayesian estimator is adaptive, thus more preferable when little is known about the true density.

Convergence rate of an ensemble approach is also investigated in this paper. The estimator defined on a binary partition can be viewed as a density tree, if we borrow the terminology from supervised learning. Then it is natural to call the ensemble method as a density forest. For more smooth density functions, such as differentiable ones with Hölder continuous derivatives, the convergence rate of density trees may fall below the minimax rate due to their relatively weak approximation abilities. While density forests, being defined as an aggregation of multiple density trees fit under small perturbations, are generally supported by much finer partitions, thus enjoy better approximation rate for the class. Specifically, for the forest we construct approximating spaces as density functions under shifts along symmetrically spaced directions and define each tree to be the corresponding sieve MLE. We show that for a Hölder space consisting of more smooth density functions, forests can achieve near minimax rate, while the rate of density trees is suboptimal.

For density trees, the explicit rates for several function classes are:

- When the true density function is spatially sparse in the sense that the Haar wavelet coefficients satisfy a weak-ℓ_q ($0 < q < 2$) constraint, the convergence rate of the sieve MLE is $n^{-\frac{1-q/2}{2}}(\log n)^{\frac{1}{2}+\frac{1-q/2}{2}}$, and the posterior concentration rate of the Bayesian method is $n^{-\frac{1-q/2}{2}}(\log n)^{2+\frac{q}{2-q}}$. As the Besov space $B^{\sigma}_{p^*,q^*}$ (for $2 \le p^* \le \infty, 1 \le q^* \le \infty, 0 < \sigma \le 1$) is contained in a weak-ℓ_q ball, the corresponding rate can be immediately obtained, which is minimax up to a logarithmic term.

- For two dimensional density functions of bounded variation, the convergence rate of the sieve MLE is $n^{-1/4}(\log n)^{3/4}$, and the posterior contraction rate of the Bayesian method is $n^{-1/4}(\log n)^3$.

- For Hölder continuous or mixture Hölder continuous (multi-dimensional cases) density functions with regularity parameter β in $(0, 1]$, the convergence rate of the sieve MLE is $n^{-\frac{\beta}{2\beta+p}}(\log n)^{2+\frac{p(p-1)}{2(2\beta+p)}}$, and the posterior concentration rate of the Bayesian method is $n^{-\frac{\beta}{2\beta+p}}(\log n)^{2+\frac{p}{2\beta}}$, whereas the minimax rate for Hölder continuous functions is $(n/\log n)^{-\beta/(2\beta+p)}$.

For density forests, when the underlying true density function lies in the Hölder space $\mathcal{H}^{1,\beta}$ $(0 < \beta \leq 1)$, the convergence rate is $n^{-\frac{1+\beta}{2(1+\beta)+p}}(\log n)^{\frac{1}{2}+\frac{1+\beta}{2(1+\beta)+p}}$, which is minimax up to a logarithmic term for this class.

1.3 Related Work on Density Estimation via Adaptive Partitioning

With univariate (or bivariate) data, the most basic non-parametric method for density estimation is the histogram. With appropriately chosen bin width, the histogram density value within each bin is proportional to the relative frequency of the data points therein. Further developments of the method allow the bins to depend on data, and substantial improvement can be obtained by such "data-adaptive" histograms (Scott, 1979).

This idea has been naturally extended to multivariate cases. Multivariate histograms with data-adaptive partitions have been studied by Shang (1994) and Ooi (2002). The breakthrough work by Lugosi and Nobel (1996) presented general sufficient conditions for the almost sure L_1-consistency based on data-dependent partitions. Later, Barron et al. (1999) constructed a penalized maximum likelihood estimator which achieves asymptotic minimax rates over anisotropic Hölder classes under the Hellinger distance. Another closely-related type of methods is multivariate density estimation based on wavelet expansions (Donoho et al., 1996; Tribouley, 1995). Along this line, Neumann (2000) and Klemelä (2009) showed that estimators based on wavelet expansions achieve minimax convergence rates up to a logarithmic factor over a large scale of anisotropic Besov classes. However, as remarked above, the regularity provided by Besov spaces is not strong enough to yield good rates when p is large.

More recently, A Bayesian approach which can learn a data-adaptive partition in multi-dimensional cases was proposed by Wong and Ma (2010). For a collection of density functions which is obtained by recursively and randomly partitioning the sample space, reweighing the subregions in a partition, and allowing optional stopping, they constructed a prior distribution called the optional Pólya tree (OPT). This prior is shown 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-adaptive partition learning. However, the computational cost of the recursive algorithm is extremely high.

In order to resolve the computational issue, another partition-learning based approach called Bayesian Sequential Partitioning (BSP) was developed by Lu et al. (2013) and Jiang et al. (2016). The major improvement of this new approach is that with slightly modified prior the logarithm of the marginal posterior probability of a fixed partition is asymptotically linear in the estimation error in terms of the Kullback-Leibler divergence. By employing sequential importance sampling (Kong et al., 1994; Liu, 2001), they designed an efficient algorithm to sample from the posterior. The method has been shown to perform well in a range of simulated and real data sets, and also achieves a low error rate when applied to classification problems. Another variation based on maximizing the discrepancy has also been introduced by Li et al. (2016). Comparing to BSP, the prior distribution discussed in

this paper imposes a stronger regularization on the partition size and leads to an adaptive posterior concentration rate. This point will be further clarified in Section 2.

The rest of the paper is organized as follows. In Section 2 we introduce the multivariate density estimation method based on adaptive partitioning (density trees) under both frequentist and Bayesian settings. In Section 3, the main results on the convergence rate of the sieve MLE and the posterior concentration rate of the Bayesian estimator are presented, while the proofs are delayed to Section 7. In Section 4 we study several specific classes of densities and calculate corresponding rates explicitly. We introduce Density forests in Section 5, and derive convergence rate under the frequentist setting, with proofs provided in Section 7. The results are validated in Section 6 by several numerical experiments.

2. Multivariate Density Estimation Based on Adaptive Partitioning

Let $Y_1, Y_2, \cdots, Y_n$ be a sequence of independent random variables distributed according to a density $f_0(y)$ with respect to a σ-finite measure μ on a measurable space $(\Omega, \mathcal{B})$. We are interested in the case when Ω is a bounded region in $\mathbb{R}^p$ and μ is the Lebesgue measure. After translation and scaling, we may assume that the sample space is the unit cube in $\mathbb{R}^p$, that is, $\Omega = \{(y^1, y^2, \cdots, y^p) : y^l \in [0,1]\}$. Let $\mathcal{F} = \{f$ is a nonnegative measurable function on $\Omega : \int_\Omega f d\mu = 1\}$ be the collection of all the density functions on $(\Omega, \mathcal{B}, \mu)$. $\mathcal{F}$ constitutes the parameter space in this problem.

2.1 Densities on Binary Partitions

To address the infinite dimensionality of $\mathcal{F}$, we construct a sequence of finite dimensional approximating spaces $\Theta_1, \Theta_2, \cdots, \Theta_t, \cdots$ based on *binary partitions*. With growing complexity, these spaces provide more and more accurate approximations to the initial parameter space $\mathcal{F}$. Here, we use a recursive procedure to define a binary partition with t subregions of the unit cube in $\mathbb{R}^p$. Let $\Omega = \{(y^1, y^2, \cdots, y^p) : y^l \in [0,1]\}$ be the unit cube in $\mathbb{R}^p$. In the first step, we choose one of the coordinates y^l and cut Ω into two subregions along the midpoint of the range of y^l. That is, $\Omega = \Omega_0^l \cup \Omega_1^l$, where $\Omega_0^l = \{y \in \Omega : y^l \leq 1/2\}$ and $\Omega_1^l = \Omega \backslash \Omega_0^l$. In this way, we get a partition with two subregions. Note that the total number of possible partitions after the first step is equal to the dimension p. Suppose after $t-1$ steps of the recursion, we have obtained a partition $\{\Omega_j\}_{j=1}^t$ with t subregions. In the t-th step, further partitioning of the region is defined as follows:

1. Choose a region from $\Omega_1, \cdots, \Omega_t$. Denote it as Ω_{j_0}.

2. Choose one coordinate y^l and divide Ω_{j_0} into two subregions along the midpoint of the range of y^l.

Such a partition obtained by $t-1$ recursive steps is called a binary partition of size t. Figure 1 displays all possible two dimensional binary partitions when t is 1, 2 and 3.

Now, let

$$\Theta_t = \left\{ f : f = \sum_{j=1}^t \frac{\theta_j}{|\Omega_j|} \mathbb{1}_{\Omega_j}, \sum_{j=1}^t \theta_j = 1, \ \{\Omega_j\}_{j=1}^t \text{ is a binary partition of } \Omega. \right\}$$

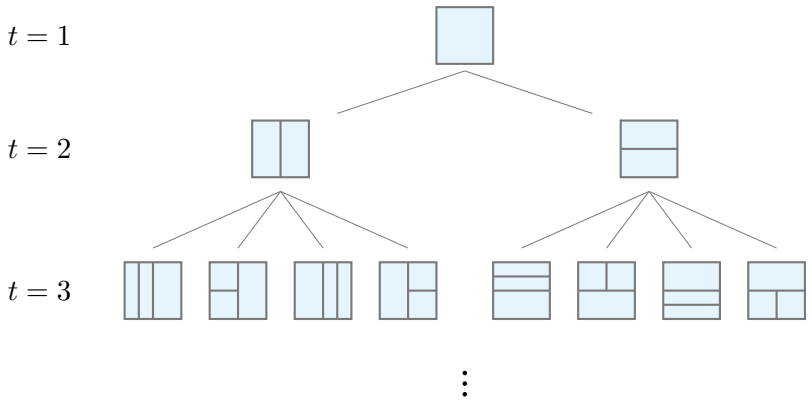

Figure 1: Binary partitions.

where $|\Omega_j|$ is the volume of Ω_j. This is to say, Θ_t is the collection of the density functions supported by the binary partitions of size t. Θ_t's constitute a sequence of approximating spaces (*i.e.* a sieve, see Grenander, 1981; Shen and Wong, 1994 for background on sieve theory), and the approximation accuracy will be characterized later. The density functions within each Θ_t can also be viewed as a special type of multivariate histograms or *density trees*, where the splits are forced to be dyadic.

We take the metric on $\mathcal{F}$ and Θ_t's to be Hellinger distance, which is defined as

$$\rho(f, g) = \Big(\int_\Omega (\sqrt{f(y)} - \sqrt{g(y)})^2 dy \Big)^{1/2}, \ f, g \in \mathcal{F}.$$

We also introduce Kullback-Leibler divergence and the variance of the log-likelihood ratio based on a single observation Y, which are defined to be

$$K(f_0, f) = \mathbb{E}_{f_0}\Big(\log \frac{f_0(Y)}{f(Y)} \Big), \quad \text{and} \quad V(f_0, f) = \mathrm{Var}_{f_0}\Big(\log \frac{f_0(Y)}{f(Y)} \Big).$$

Here we want to point out the Kullback-Leibler divergence is a stronger distance compared to the Hellinger distance, in the sense that for any $f, g \in \mathcal{F}$, $\rho^2(f, g) \le K(f, g)$.

2.2 The Sieve MLE

For any $f \in \Theta_t$, the log-likelihood is defined to be

$$l_n(f) = \sum_{i=1}^n \log f(Y_i) = \sum_{j=1}^t n_j \log \left(\frac{\theta_j}{|\Omega_j|} \right),$$

where n_j is the count of data points in Ω_j, i.e., $n_j = \mathrm{card}\{i : Y_i \in \Omega_j, 1 \le i \le n\}$. The maximum likelihood estimator in Θ_t is defined to be

$$\hat{f}_{n,t} = \arg\max_{f \in \Theta_t} l_n(f). \tag{1}$$

We claim that $\hat{f}_{n,t}$ is well defined. This is the true because given a binary partition $\{\Omega_j\}_{j=1}^t$, our model for data becomes a multinomial one, and $(\theta_1, \cdots, \theta_t)$ can be estimated by the MLE for the multinomial distribution. And within each Θ_t, the number of possible binary partitions is finite. If t grows with n at certain rate, then Θ_t's constitute a sieve to $\mathcal{F}$, and the sequence $\hat{f}_{n,t}$ is a sequence of sieve MLEs (Shen and Wong, 1994).

The key step to obtain the MLE is to learn a partition of the sample space. To better illustrate this point, given the binary partition $\mathcal{A}_t = \{\Omega_j\}_{j=1}^t$, we use $\mathcal{F}(\mathcal{A}_t)$ to denote the collection of piecewise constant density functions defined on this partition,

$$\mathcal{F}(\mathcal{A}_t) = \left\{ f = \sum_{j=1}^t \frac{\theta_j}{|\Omega_j|} \mathbb{1}_{\Omega_j} : \sum_{j=1}^t \theta_j = 1 \text{ and } \theta_j \geq 0, j = 1, \ldots, t. \right\}$$

Then, the maximum log-likelihood can be achieved by the densities defined on $\{\Omega_j\}_{j=1}^t$ is:

$$l_n(\mathcal{F}(\mathcal{A}_t)) \coloneqq \max_{f \in \mathcal{F}(\mathcal{A}_t)} l_n(f) = \sum_{j=1}^t n_j \log\left(\frac{n_j}{n|\Omega_j|}\right). \tag{2}$$

We can treat $-l_n(\mathcal{F}(\mathcal{A}_t))$ as the *deviance* of the partition $\{\Omega_j\}_{j=1}^t$. If we define the "projection" of the true density onto the partition $\mathcal{A}_t$ as

$$f_{\mathcal{A}_t} = \sum_{j=1}^t \frac{\theta_{0,j}}{|\Omega_j|} \mathbb{1}_{\Omega_j} \quad \text{with } \theta_{0,j} = \int_{\Omega_j} f_0,$$

then for a fixed t we can show that

$$\lim_{n\to\infty} (-l_n(\mathcal{F}(\mathcal{A}_t))/n) = K(f_0, f_{\mathcal{A}_t}) - \int f_0 \log(f_0).$$

Therefore, by minimizing the deviance over all binary partitions of a fixed size, we learn a most promising structure in a data-adaptive way. The sieve MLE $\hat{f}_{n,t}$ is simply the histogram defined on that partition.

2.3 Bayesian Multivariate Density Estimation

To introduce the Bayesian approach, the key step is to define an appropriate prior on the parameter space $\Theta = \cup_{t=1}^\infty \Theta_t$. An ideal prior Π is supposed to be capable of balancing the approximation error and the complexity of Θ. The prior distribution studied in this paper penalizes the size of the partition in the sense that the probability mass on each Θ_t is proportional to $\exp(-\lambda t \log t)$. The prior was introduced for density estimation by Liu (2016). The same type of exponentially decaying prior has also been studied under the Bayesian regression tree setting (van der Pas and Ročková, 2017; Ročková and van der Pas, 2020). Given a sample of size n, we restrict our attention to $\Theta_n = \cup_{t=1}^{n/\log n} \Theta_t$, because in practice we need enough samples within each subregion to get a meaningful estimate of the density. This is to say, when $t \leq n/\log n$, $\Pi(\Theta_t) \propto \exp(-\lambda t \log t)$, otherwise $\Pi(\Theta_t) = 0$.

If we use I_t to denote the total number of possible partitions of size t, then it is not hard to see that $\log I_t \leq c^* t \log t$, where c^* is a constant. Within each Θ_t, the prior is uniform

across all binary partitions. Then the prior on $\mathcal{F}\left(\{\Omega_j\}_{j=1}^t\right)$ can be written as

$$\Pi\left(\mathcal{F}\left(\{\Omega_j\}_{j=1}^t\right)\right) \propto \exp(-\lambda t \log t), \quad \lambda > 0. \tag{3}$$

Given a partition $\{\Omega_j\}_{j=1}^t$, the weights θ_j on the subregions follow a Dirichlet distribution with parameters all equal to α ($\alpha < 1$). This is to say, for $x_1, \cdots, x_t \geq 0$ and $\sum_{j=1}^t x_j = 1$,

$$\Pi\left(f = \sum_{j=1}^t \frac{\theta_j}{|\Omega_j|}\mathbb{1}_{\Omega_j} : \theta_1 \in dx_1, \cdots, \theta_t \in dx_t \middle| \mathcal{F}\left(\{\Omega_j\}_{j=1}^t\right)\right) = \frac{1}{D(\alpha, \cdots, \alpha)}\prod_{j=1}^t x_j^{\alpha-1},$$

where $D(\delta_1, \cdots, \delta_t) = \prod_{j=1}^t \Gamma(\delta_j)/\Gamma(\sum_{j=1}^t \delta_j)$.

The idea of imposing such a prior is not completely new. The prior distribution introduced above is very similar to that in Lu et al. (2013). However, there is a significant difference lying in how we penalize the size of the partition. To further explain this difference, we take a closer look at the marginal posterior probability of a partition. Let $\Pi_n(\cdot|Y_1, \cdots, Y_n)$ denote the posterior distribution. After integrating out the weights θ_j, we can obtain the marginal posterior probability of $\mathcal{F}\left(\{\Omega_j\}_{j=1}^t\right)$:

$$\begin{aligned}
\Pi_n\left(\mathcal{F}(\{\Omega_j\}_{j=1}^t)\middle|Y_1, \cdots, Y_n\right) &\propto \Pi\left(\mathcal{F}(\{\Omega_j\}_{j=1}^t)\right)\int\left(\prod_{j=1}^t (\theta_j/|\Omega_j|)^{n_j}\right) \\
&\quad \times \left(\frac{1}{D(\alpha, \cdots, \alpha)}\prod_{j=1}^t \theta_j^{\alpha-1}\right) d\theta_1 \cdots d\theta_t \\
&\propto \exp(-\lambda t \log t) \cdot \frac{D(\alpha + n_1, \cdots, \alpha + n_t)}{D(\alpha, \cdots, \alpha)}\prod_{j=1}^t \frac{1}{|\Omega_j|^{n_j}},
\end{aligned}$$

where n_j is still the number of observations in Ω_j. While under the prior introduced in Lu et al. (2013), the marginal posterior probability is:

$$\Pi_n^*\left(\mathcal{F}(\{\Omega_j\}_{j=1}^t)\middle|Y_1, \cdots, Y_n\right) \propto \exp(-\lambda t)\frac{D(\alpha + n_1, \cdots, \alpha + n_t)}{D(\alpha, \cdots, \alpha)}\prod_{j=1}^t \frac{1}{|\Omega_j|^{n_j}}.$$

From a model selection perspective, we may treat the densities defined on each binary partition as a model of the data. When $t \ll n$, asymptotically,

$$\log\left(\Pi_n^*\left(\mathcal{F}(\{\Omega_j\}_{j=1}^t)\middle|Y_1, \cdots, Y_n\right)\right) \asymp l_n(\mathcal{F}(\{\Omega_j\}_{j=1}^t)) - \frac{1}{2}(t-1)\log n, \tag{4}$$

where $l_n(\mathcal{F}(\{\Omega_j\}_{j=1}^t))$ is defined in (2). This is to say, under the prior distribution introduced in the paper by Lu et al. (2013), selecting the partition which maximizes the marginal posterior probability is equivalent to performing model selection by using the Bayesian information criterion (BIC). However, if we allow t to increase with n, (4) will not hold any

more. But if we use the prior introduced in this section, when $t = C(n/\log n)^{\alpha}, 0 < \alpha < 1$ and $C > 0$, or even when $t/(n/\log n) \to \zeta \in (0,1)$ as $n \to \infty$, we still have

$$\log\left(\Pi_n\left(\mathcal{F}(\{\Omega_j\}_{j=1}^t)\middle|Y_1,\cdots,Y_n\right)\right) \asymp l_n(\mathcal{F}(\{\Omega_j\}_{j=1}^t)) - \lambda t\log t. \tag{5}$$

More details for deriving (5) will be provided in Appendix. From the model selection perspective, this is closer to the risk inflation criterion (RIC, Foster and George, 1994).

3. Convergence Rates for Density rees

In this section, we present our main results on convergence rate. We further restrict our interest to a subsect $\mathcal{F}_0 \subset \mathcal{F}$ of densities which satisfies the following two conditions: First, $\int_\Omega f^2 < \infty$. Second, for any $f \in \mathcal{F}_0$, there exists a sequence of approximations $f_t \in \Theta_t$ such that

$$\rho(f, f_t) \le At^{-r}, \tag{6}$$

where r is a parameter characterizing the decay rate of the approximation error and A is a constant that may depend on f, or the function class under consideration. In order to demonstrate that $\mathcal{F}_0$ is still a rich class, from Section 4.1 to Section 4.3, we apply the main results to calculate explicit rates for several specific density classes belonging to $\mathcal{F}_0$. These function classes are widely used in statistical modelding.

3.1 Convergence Rate of Sieve MLE

The following theorem provides convergence rate of the sieve MLE, while the proof is delayed to Section 7.2.

Theorem 1 *For any $f_0 \in \mathcal{F}_0$, $\hat{f}_{n,t}$ is the sieve maximum likelihood estimator defined in equation (1). Assume that f_0 can be approximated by Θ_t's at the rate r. When t and n satisfy*

$$M_1\left(\frac{n}{\log n}\right)^{\frac{1}{2r+1}} \le t \le M_2\left(\frac{n}{\log n}\right)^{\frac{1}{2r+1}} \quad \text{for some } M_2 \ge M_1 > 0, \tag{7}$$

the convergence rate of the sieve MLE is $n^{-\frac{r}{2r+1}}(\log n)^{(\frac{1}{2}+\frac{r}{2r+1})}$, in the sense that

$$\mathbb{P}_0^n\left(\rho(\hat{f}_{n,t}, f_0) \ge Dn^{-\frac{r}{2r+1}}(\log n)^{(\frac{1}{2}+\frac{r}{2r+1})}\right) \to 0,$$

where $D > 1$ is a constant and $\mathbb{P}_0^n$ is the probability measure on the product space corresponding to the true density f_0. A possible choice of t in the definition of sieve MLE $\hat{f}_{n,t}$ is $t = \left((2^8A^2r/c_1)(n/\log n)\right)^{\frac{1}{2r+1}}$, where A is the constant introduced in (6) and c_1 can be chosen in $(0,1)$.

Remark 2 *It is possible to partition the sample space in a more flexible way. In particular, if we replace the binary partition at mid-point by one at a point chosen from a fixed sized grid (such as regular equi-spaced grid), the analysis and resulting rate remain the same.*

Remark 3 *For any two distributions with densities f and g, as $\|f-g\|_{L^1(\Omega)} \leq 2\rho(f,g)$, we can immediately obtain an upper bound for the convergence result under the total variation norm based on Theorem 1. The rate is the same as that under the Hellinger distance. Similar argument can be applied to all following results on convergence under the Hellinger distance.*

Further analysis demonstrates that the rate $n^{-\frac{r}{2r+1}}(\log n)^{(\frac{1}{2}+\frac{r}{2r+1})}$ can only be achieved when we appropriately balance the complexity of the approximating spaces with the sample size. On one hand, the complexity of Θ_t affects the convergence rate in a way that, the richer the approximating spaces the lower bias the estimators have. Conversely, 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. The merit of Theorem 1 is that it clarifies how to strike the balance between these two components.

At the same time, Theorem 1 implies that for the sieve MLE, in order to achieve the rate $n^{-\frac{r}{2r+1}}(\log n)^{(\frac{1}{2}+\frac{r}{2r+1})}$, we need to have some prior knowledge of the true density function, that is, the approximation rate r, which is usually inaccessible in real problems. This drawback may limit the applicability and performance of the method in practice, which also becomes our motivation to study the Bayesian estimator. From (5), we know that under the Bayesian setting, imposing a prior distribution can be alternatively viewed as imposing a penalty. We wonder by considering a Bayesian estimator whether the method can adapt to the unknow complexity r of the true density function. This is the focus of the next subsection.

3.2 Posterior Concentration Rate of the Bayesian Estimator

To characterize asymptotic property of Bayesian inference, we study how fast the posterior probability measure concentrates around the true the density f_0. The posterior probability is the random measure given by

$$\Pi(B|Y_1, \cdots, Y_n) = \frac{\int_B \prod_{i=1}^n f(Y_i) d\Pi(f)}{\int_\Theta \prod_{i=1}^n f(Y_i) d\Pi(f)}, \quad B \in \mathcal{B}.$$

A Bayesian inference is said to be *consistent* if the posterior distribution concentrates on arbitrarily small neighborhoods of f_0, with probability tending to 1 under $\mathbb{P}_0^n$. The posterior concentration rate refers to the rate at which these neighborhoods shrink to zero while still possessing most of the posterior mass. More explicitly, a sequence $\epsilon_n \to 0$ is called an upper bound for posterior concentration rate if there exists a constant $M > 0$ so that

$$\Pi(f : \rho(f, f_0) \geq M\epsilon_n | Y_1, \cdots, Y_n) \to 0, \quad \text{in } \mathbb{P}_0^n\text{-probability}.$$

The problem is also studied by Liu et al. (2017). The following theorem gives the posterior concentration rate under the prior distribution specified in Section 2.3.

Theorem 4 *$Y_1, \cdots, Y_n$ is a sequence of independent random variables distributed according to f_0. Θ is the collection of all the p-dimensional density functions supported by the binary partitions as defined in Section 2.1. The prior distribution on Θ is as specified in Section 2.3. If $f_0 \in \mathcal{F}_0$, then $\epsilon_n = n^{-\frac{r}{2r+1}}(\log n)^{2+\frac{1}{2r}}$ is an upper bound of the posterior concentration rate.*

The strategy to prove this theorem is to write the posterior probability measure as

$$\begin{aligned}&\Pi(f:\rho(f,f_0)\geq M\epsilon_n|Y_1,\cdots,Y_n)\\&\quad=\frac{\sum_{t=1}^{\infty}\int_{\{f:\rho(f,f_0)\geq M\epsilon_n\}\cap\Theta_t}\prod_{i=1}^{n}\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f)}{\sum_{t=1}^{\infty}\int_{\Theta_t}\prod_{i=1}^{n}\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f)}.\end{aligned}\tag{8}$$

Using similar techniques as Ghosal et al. (2000) and Shen and Wasserman (2001), we first obtain the upper bounds for the items in the numerator by dividing them into three blocks, each of which accounts for bias, variance, and rapidly decaying prior respectively. Then we provide the prior thickness result in Section 7.3.1, i.e., we bound the prior mass of a ball around the true density from below. In Section 7.3.2, these bounds are combined together, leading to a complete proof of the posterior concentration rate.

We want to point out that, based on the minimaxity of the Bayes estimator, it is necessary to restrict our attention to a subset of $\mathcal{F}$. In the paper by Farrell (1967) and Birgé and Massart (1998), the authors showed that it is impossible to find an estimator which works uniformly well for every f in $\mathcal{F}$. This is the case because for any estimator $\hat{f}$, there always exists an $f\in\mathcal{F}$ for which $\hat{f}$ is inconsistent.

Theorem 4 suggests the following two take-away messages. First, the rate is adaptive to the unknown smoothness of the true density. Second, the posterior contraction rate does not directly depend on the dimension p. Instead, it only depends on how well the true density can be approximated by the sieve. As r increases, up to a $\log n$ term, the rate can get close to the parametric rate of $n^{-1/2}$. Admittedly, for many classes of density functions, r may depend on p. But in several special cases, like when the density function is spatially sparse or when the density function only has variations along a subset of dimensions, we can show that the rate will not be affected by the full dimension of the problem. More details will be provided in Section 4.

3.2.1 Relations with Existing Bayesian Nonparametric Methods.

The adaptivity of Bayesian approaches has drawn a lot of attention in recent years. In terms of density estimation, there are mainly two categories of adaptive Bayesian nonparametric approaches. The first category of work relies on basis expansion of the density function and typically imposes a random series prior (Rivoirard and Rousseau, 2012; Shen and Ghosal, 2015). When the prior on the expansion coefficients is set to be normal, it is also a Gaussian process prior (de Jonge and van Zanten, 2012). In the multivariate case, most existing work (de Jonge and van Zanten, 2012; Shen and Ghosal, 2015) chooses to use the tensor-product basis. Comparing to this line of work, the advantage of the partition based method mainly lies in the spatial adaptivity. In fact, the number of tensor-product basis functions increases exponentially with the dimension, which poses a great challenge in terms of computation. For our method, based on adaptive partitioning, it learns a good approximation to the underlying density function by making more splits in the subregion where the probability mass concentrates while saving computational cost on the rest part, thus it can successfully handle multivariate cases even when the dimension is 30 (see Example 2 in Section 6).

Another line of work considers mixture priors (Rousseau, 2010; Kruijer et al., 2010; Shen et al., 2013). The mixture distributions usually enjoy good approximation properties

and naturally lead to adaptivity to very high level of smoothness. However, they may fail to capture the local features. On the other hand, the method discussed in this paper can provide multi-resolution approximations to the density function, generally leading to a better performance for local features (Ma and Wong, 2011).

4. Applications

In this section, we apply the previous results to calculate explicit convergence rates when the true density functions belong to several special classes. The results in this section help further clarify, besides the dimension, which factors may affect the convergence of the partition based method.

4.1 Spatial Adaptation

First, we assume that the density concentrates spatially. Mathematically, this implies the density function satisfies a type of *spatial sparsity*. In the past two decades, sparsity has become one of the most discussed types of structure under which we are able to overcome the curse of dimensionality. A remarkable example is that it allows us to solve high-dimensional linear models, especially when the system is underdetermined. It would be interesting to study how we can benefit from the sparse structure when performing density estimation. Under current settings, a natural way to characterize the spatial sparsity is by describing the decay rate of the ordered Haar wavelet coefficients.

4.1.1 Multiresolution Analysis and High-Dimensional Haar Basis

Haar basis is one type of simple and widely used wavelet bases. In the one dimensional case, the Haar mother wavelet function is defined as $\psi(y) = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)}$. And its scaling function is $\phi(y) = \mathbb{1}_{[0,1)}(y)$.

When extended to the multivariate case, the scaling function can be defined as $\xi^0(y) = \prod_{l=1}^p \phi(y^l)$, and the wavelet functions are

$$\boldsymbol{\Xi}^{(0)} = \{\xi^\epsilon : \epsilon = (\epsilon^l)_{1\le l\le p} \in \{0,1\}^p \setminus \{(0,\dots,0)\}\}, \quad \text{with } \xi^\epsilon(y) = \prod_{l=1}^p \psi^{\epsilon^l}(y^l)\cdot\phi^{1-\epsilon^l}(y^l).$$

On Ω, let V_0 be the subspace of $L^2(\Omega)$ spanned by ξ^0. For all $j \in \mathbb{N}$, $V_j \subset V_{j+1}$ if V_j denotes the space spanned by $\{\xi^0_{jk}, k \in \mathbb{N}^p, 0 \le k^l < 2^j\}$, where $\xi^0_{jk} = (\sqrt{2})^{jp}\xi^0(2^j y - k)$. Then we define the space W_j by

$$V_{j+1} = V_j \oplus W_j.$$

It can be shown that

- $\boldsymbol{\Xi}^{(0)}$ is an orthonormal basis of W_0;
- For any $j \in \mathbb{N}$, let $\boldsymbol{\Xi}^{(j)} = \{\xi^\epsilon_{jk}, k \in \mathbb{N}^p, 0 \le k^l < 2^j, \epsilon \in \{0,1\}^p \setminus \{(0,\dots,0)\}\}$, where $\xi^\epsilon_{jk}(y) = (\sqrt{2})^{jp}\xi^\epsilon(2^j y - k)$. Then $\cup_{j\in\mathbb{N}}\boldsymbol{\Xi}^{(j)} \cup \{\xi^0\}$ is an orthonormal basis of $L^2(\Omega)$.

We also have the decomposition

$$L^2(\Omega) = V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots .$$

This implies for all $g \in L^2(\Omega)$,

$$g = \sum_{k \in \mathbb{N}^p, 0 \le k^l < 2^{j_0}} \langle g, \xi^0_{j_0 k} \rangle \xi^0_{j_0 k} + \sum_{j \ge j_0} \sum_{\xi^{(j)} \in \Xi^{(j)}} \langle g, \xi^{(j)} \rangle \xi^{(j)}, \tag{9}$$

where the expansion holds under the norm of the space $L^2(\Omega)$.

4.1.2 Spatial Sparsity

Let f be a p dimensional density function. We will work with $g = \sqrt{f}$ to characterize the sparsity. Note that $g \in L^2(\Omega)$, thus we can obtain an expansion of g with respect to the Haar basis in the form (9). We rearrange this summation by the size of wavelet coefficients. In other words, we order the coefficients as the following:

$$|\langle g, \xi_{(1)} \rangle| \ge |\langle g, \xi_{(2)} \rangle| \ge \cdots \ge |\langle g, \xi_{(k)} \rangle| \ge \cdots .$$

Then the sparsity condition imposed on the density functions is that the decay of the wavelet coefficients follows a power law,

$$|\langle g, \xi_{(k)} \rangle| \le C k^{-1/q} \text{ for all } k \in \mathbb{N}^+ \text{ and } 0 < q < 2, \tag{10}$$

where C is a constant.

The condition (10) on the decay of ordered wavelet coefficients is also called a weak-ℓ_q constraint (Abramovich et al., 2006). This condition has been widely used to characterize the sparsity of signals and images (Abramovich et al., 2006; Candès and Tao, 2006; DeVore et al., 1992). In particular, in the work by DeVore et al. (1992), it has been shown that in the two-dimensional cases, when $0 < q < 2$, this condition reasonably captures the sparsity of real world images.

In fact, the weak-ℓ_q ball contains a Besov space. To clarify the relation, we start with a brief description of the Besov space. Assume $j \in \mathbb{N}$ is an index of the resolution. With notations introduced in Section 4.1.1, we define $P^{(j)}$ to the be projection operator onto V_j and $E^{(j)} = P^{(j+1)} - P^{(j)}$. Then for $\sigma \in (0, 1]$ and $p^*, q^* \in [1, \infty]$, g lies in the Besov space $B^\sigma_{p^*,q^*}$ if and only if

$$\|g\|_{B^\sigma_{p^*,q^*}} := \|P^{(j_0)} g\|_{p^*} + \left(\sum_{j \ge j_0} \left(2^{j\sigma} \|E^{(j)} g\|_{p^*} \right)^{q^*} \right)^{1/q^*} < \infty.$$

With the multivariate Haar basis,

$$\begin{aligned} P^{(j_0)} g &= \sum_{k \in \mathbb{N}^p, 0 \le k^l < 2^{j_0}} \langle g, \xi^0_{j_0 k} \rangle \xi^0_{j_0 k}, \\ E^{(j)} g &= \sum_{\xi^{(j)} \in \Xi^{(j)}} \langle g, \xi^{(j)} \rangle \xi^{(j)}. \end{aligned}$$

Therefore, an equivalent norm for the Besov space in terms of wavelet coefficients is

$$\left\|g\right\|_{B^{\sigma}_{p^*,q^*}} \coloneqq \left\|P^{(j_0)}g\right\|_{p^*} + \left(\sum_{j\geq j_0}\left(2^{j(\sigma+p(1/2-1/p^*))}\left\|\{\langle g,\xi^{(j)}\rangle\}_{\xi^{(j)}\in\Xi^{(j)}}\right\|_{p^*}\right)^{q^*}\right)^{1/q^*}.$$

For $\sigma \in (0,1]$, the above definition matches that in the literature (for example, see Nikol'skii 2012), and is equivalent to the definition of Besov spaces based on the *modulus of continuity* or with the aid of best approximations. For $\sigma > 1$, even it is not in line with the standard description of the Besov space, we can still introduce a space of functions with fast decaying Haar wavelet coefficients, called a Besov ball, in the following way. The Besov ball $B^{\sigma}_{p^*,q^*}, \sigma > 1, p^*, q^* \in [1,\infty]$ with radius L is the collection of functions g satisfying

$$\left\|P^{(j_0)}g\right\|_{p^*} + \left(\sum_{j\geq j_0}\left(2^{j(\sigma+p(1/2-1/p^*))}\left\|\{\langle g,\xi^{(j)}\rangle\}_{\xi^{(j)}\in\Xi^{(j)}}\right\|_{p^*}\right)^{q^*}\right)^{1/q^*} < L.$$

An extreme case is that for a piecewise constant function defined on a binary partition $\mathcal{A}_{t_0}$, it lies in a Besov ball with arbitrarily large σ.

If we denote the class of functions on Ω satisfying the weak-ℓ_q condition as $wl^q(\Omega)$, it is easy to show for all $p^*, q^* \in [1,\infty], \sigma > 0$,

$$B^{\sigma}_{p^*,q^*}(\Omega, L) \subset wl^q(\Omega), \quad \text{when } q = \frac{1}{\sigma/p + 1/2} \text{ (recall that } p \text{ is the dimension)}, \tag{11}$$

where $B^{\sigma}_{p^*,q^*}(\Omega, L)$ is the collection of functions in $B^{\sigma}_{p^*,q^*}(\Omega)$ (when $0 < \sigma \leq 1$ it is the standard Besov space and when $\sigma > 1$ it is the Besov ball defined above) with a norm bounded by L, and the constant C in condition (10) can be set as L. The same embedding result has also been discussed by Donoho (1993). Therefore, the condition (10) can be viewed more generally as one that characterizes the regularity or smoothness of density functions. It covers the spatially sparse density functions as a special case.

Next we use an example to illustrate how this condition implies spatial sparsity.

Example 1 *We study the two-dimensional true density function*

$$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} \sim \frac{2}{5}\mathcal{N}\left(\begin{pmatrix}0.25\\0.25\end{pmatrix}, 0.05^2 I_{2\times 2}\right) + \frac{3}{5}\mathcal{N}\left(\begin{pmatrix}0.75\\0.75\end{pmatrix}, 0.05^2 I_{2\times 2}\right).$$

Figure 2 displays the heatmap of the density function and its Haar coefficients. The last panel in the right plot displays the ordered coefficients in log-scale. From this we can clearly see that the power-law decay is satisfied.

4.1.3 Convergence Rate

We first provide the approximation result for the class of density functions satisfying the weak-ℓ_q condition, while the proof is delayed to the Appendix.

Lemma 5 *Let f_0 be a p-dimensional density function satisfying the condition (10). There exists a sequence of $f_t \in \Theta_t$, such that $\rho(f_0, f_t) \lesssim t^{-(1/q-1/2)}$, or equivalently, $\rho(f_0, f_t) \leq ct^{-(1/q-1/2)}$, where a possible choice for c can be $(2C^2)/(2^p(2/q-1))$, where C is the constant in condition (10).*

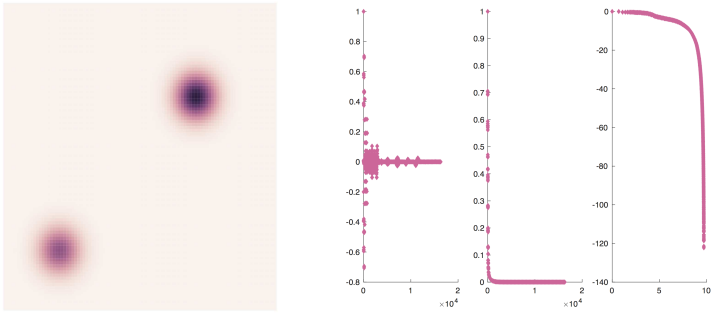

Figure 2: Heatmap of the density and plots of 2-dimensional Haar coefficients. For the plot on the right, the left panel is the plot of Haar coefficients from low resolution to high resolution up to level 6. The middle one is the plot of sorted coefficients according to their absolute values. And the right one is the same as the middle plot but with the abscissa in log scale.

The convergence rate can be immediately obtained by plugging the explicit approximation rate calculated in Lemma 5 into the results of Theorem 1 and Theorem 4, and is summarized in the following corollary.

Corollary 6 *(Spatial Adaptation) Let f_0 be a p-dimensional density function satisfying the weak-ℓ_q (10). If we apply our methods to this type of density functions, the convergence rate of the sieve MLE is $n^{-\frac{1-q/2}{2}}(\log n)^{\frac{1}{2}+\frac{1-q/2}{2}}$, and the posterior concentration rate of the Bayesian estimator is $n^{-\frac{1-q/2}{2}}(\log n)^{2+\frac{q}{2-q}}$.*

From the Corollary 6 we see that the convergence rate only depends on how fast the coefficients decay as opposed to the dimension of the sample space. Therefore, for a relatively small q, the estimator is able to take advantage of the spatial sparsity to achieve fast convergence rate even in high dimensions.

According to the embedding result (11), we can also derive the corresponding convergence rate for the Besov space. As we have implicitly assumed $\sqrt{f_0} \in L^2(\Omega)$, we only study Besov spaces or balls $B^{\sigma}_{p^*,q^*}$ with $p^* \geq 2$ here. In specific, we have the following result:

Corollary 7 *(Besov Spaces) Let f_0 be a p-dimensional density function such that $\sqrt{f_0}$ lies in a Besov space $B^{\sigma}_{p^*,q^*}(\Omega)$ ($p^* \geq 2, q^* \geq 1, 0 < \sigma \leq 1$) or a Besov ball $B^{\sigma}_{p^*,q^*}(\Omega)$ ($p^* \geq 2, q^* \geq 1, \sigma > 1$) with a bounded norm. If we apply our methods to this type of density functions, the convergence rate of the sieve MLE is $n^{-\frac{\sigma}{2\sigma+p}}(\log n)^{\frac{1}{2}+\frac{\sigma}{2\sigma+p}}$, and the posterior concentration rate of the Bayesian estimator is $n^{-\frac{\sigma}{2\sigma+p}}(\log n)^{2+\frac{p}{2\sigma}}$.*

The rate is minimax up to a logarithmic term for the Besov space (Donoho et al., 1996). Again, as what has been pointed out in Section 3.2, the Bayesian method is adaptive in the sense that it can achieve such a rate without the need to specify σ, p^*, q^*, as long as they are within the range stated in Corollary 7. Here, our assumption is imposed on $\sqrt{f}$ instead of f, as the L_2-distance between the square root of two density functions exactly corresponds to the Hellinger distance between two probability mesures. If we assume $f \in B^{\sigma}_{p^*,q^*}$ with

$p^* \geq 2$, then when f takes values close to zero in a region with positive Lebesgue measure, the posterior concentration rate under the Hellinger distance can be slower, as in this case the prior may not put enough probability mass around the true density.

4.2 Density Functions of Bounded Variation

In image analysis, the denoised image is usually assumed to be a function of bounded variation. For a number of image analysis techniques, such as image segmentation, object recognition, and motion detection, an indispensable building block is to obtain an approximation to the denoised image. Nonlinear approximation, such as, tree-structured vector quantization (Poggi and Olshen, 1995), wavelet compression and wavelet shrinkage (Donoho et al., 1995), neural networks (LeCun et al., 1989; Le et al., 2012), and GANs (Goodfellow et al., 2014), is widely used in this field. For two-dimensional gray scale images embedded in $\mathbb{R}^3$ with z-axis representing intensities, if we view our data as the xy-coordinates of points uniformly sampled from the region below the intensity, then estimating the denoised image is equivalent to a density estimation problem up to a normalizing constant and our method can be applied to obtain an approximation. We evaluate the performance of the method by calculating the convergence rate when the density function is of bounded variation. The posterior concentration rate of our method indicates that the approach achieves the same minimax rate (up to a logarithmic term) as that of wavelet thresholding.

To begin with, we briefly introduce the space BV. Let $\Omega = [0,1)^2$. For a vector $\nu \in \mathbb{R}^2$, the difference operator Δ_ν along the direction ν is defined by

$$\Delta_\nu(f, y) := f(y + \nu) - f(y).$$

For functions f defined on Ω, $\Delta_\nu(f, y)$ is defined whenever $y \in \Omega(\nu)$, where $\Omega(\nu) := \{y : [y, y+\nu] \subset \Omega\}$ and $[y, y+\nu]$ is the line segment connecting y and $y+\nu$. Denote by $e^l, l = 1, 2$ the two coordinate vectors in $\mathbb{R}^2$. We say that a function $f \in L^1(\Omega)$ is in $BV(\Omega)$ iff

$$V_\Omega(f) := \sup_{h>0} h^{-1} \sum_{l=1}^{2} \|\Delta_{he^l}(f, \cdot)\|_{L^1(\Omega(he^l))} = \lim_{h \to 0} h^{-1} \sum_{l=1}^{2} \|\Delta_{he^l}(f, \cdot)\|_{L^1(\Omega(he^l))} < \infty.$$

The quantity $V_\Omega(f)$ is the *variation* of f over Ω. The rate for this class is provided in the next corollary.

Corollary 8 *Assume that $f_0 \in BV(\Omega)$. If we apply the multivariate density estimator based on adaptive partitioning to estimate f_0, the convergence rate of the sieve MLE is $n^{-1/4}(\log n)^{3/4}$, and the posterior concentration rate of the Bayes estimator is $n^{-1/4}(\log n)^3$.*

The rate is minimax up to a logarithmic term. The proof is provided in the Appendix.

4.3 Hölder Space

The class of Hölder continuous functions $\mathcal{H}^{k,\beta}(\Omega)$, where $k \in \mathbb{N}$, $0 < \beta \leq 1$, is defined as the collection of functions on Ω for which all partial derivatives up to order k exist, and there exists a constant $L > 0$, such that

$$\|\nabla^l f\|_2 \leq L \text{ for all } l \in [k], \tag{12}$$

$$\|\nabla^k f(y) - \nabla^k f(y')\|_2 \leq L\|y - y'\|_2^\beta \text{ for all } y, y' \in \Omega. \tag{13}$$

In multi-dimensional cases, we also introduce the mixed-Hölder continuity. In order to simplify the notation, we give the definition when the dimension is two. It can be generalized to high-dimensional cases in the same way. A real-valued function f on $\mathbb{R}^2$ is called mixed-Hölder continuous for some nonnegative constant C and $\beta \in (0, 1]$, if for any $(x_1, y_1), (x_1, y_2) \in \mathbb{R}^2$,

$$|f(x_2, y_2) - f(x_2, y_1) - f(x_1, y_2) + f(x_1, y_1)| \leq L|x_1 - x_2|^\beta |y_1 - y_2|^\beta.$$

For Hölder continuous functions, we have the following result:

Corollary 9 *Let f_0 be the p-dimensional density function. If $\sqrt{f_0}$ is Hölder continuous $\sqrt{f_0} \in \mathcal{H}^{0,\beta}(\Omega)$ or mixed-Hölder continuous (when $p \geq 2$) with regularity parameter $\beta \in (0, 1]$, then the convergence rate of the sieve MLE is $n^{-\frac{\beta}{2\beta+p}}(\log n)^{1+\frac{p(p-1)}{2(2\beta+p)}}$, and the posterior concentration rate of the Bayesian estimator is $n^{-\frac{\beta}{2\beta+p}}(\log n)^{2+\frac{p}{2\beta}}$.*

The proof of the corollary is based on the following approximation result for the Hölder space.

Lemma 10 *If $\sqrt{f_0}$ is p-dimensional Hölder continuous with regularity parameter $\beta \in (0, 1]$, then there exists a sequence of $f_t \in \Theta_t$, such that $\rho(f_0, f_t) \leq 2pLt^{-\beta/p}$. For mixed-Hölder continuous (when $p \geq 2$) functions, the approximation rate is $2^p Lt^{-\beta/p}(\log t)^{p/2}$.*

A detailed proof will be provided in the Appendix.

Remark 11 *This result also has the following useful implications: if the true density f_0 only depends on $\tilde{p}$ variable with $\tilde{p} < p$, but we do not know in advance which $\tilde{p}$ variables, then the rate of the partition-based method enjoys an "oracle property". That is, the rate is determined by the effective dimension $\tilde{p}$, since the density function can be viewed as a $\tilde{p}$-dimensional Hölder or mixed-Hölder continuous one and the smoothness parameter r is only a function of $\tilde{p}$ according to Lemma 10. In Section 6, we will use a simulated data set (see Example 5) to illustrate this point.*

4.4 Summary

To better understand the asymptotic properties of density trees, we would like to provide the following embedding results between different function classes we have studied so far. The Besov space is a rich class, in the sense that both the space of bounded variations and the Hölder class can be embedded into certain Besov spaces. In specific, Donoho (1993) pointed out that in the one-dimensional case,

$$B^1_{1,1}(\Omega) \subset \mathrm{BV}(\Omega) \subset B^1_{1,\infty}(\Omega),$$

and the embedding result for the Hölder space is discussed by Nikol'skii (2012),

$$\mathcal{H}^{0,\beta}(\Omega) = B^\beta_{\infty,\infty}(\Omega), \quad \beta \in (0, 1].$$

For the Besov space $B^\sigma_{p^*,q^*}$, the parameter p^* defines the norm $\|\cdot\|_{p^*}$ that is applied to characterize the regularity of the function. $p^* = \infty$ indicates a strong metric is used and

consequently the density function within the corresponding Besov space is uniformly continuous (indeed Hölder continuous). As p^* decreases, the metric becomes weaker, implying the density function may have sharper local variations.

The appealing feature of the partition based method is that it can achieve minimax rate for large collection of function spaces. For the Bayesian approach, the rate is also adaptive from the following two perspectives. First, for Besov spaces the minimax rate can be achieved without knowing parameters σ, p^*, q^* of the underlying function class. Second, for a Hölder continuous density function, if it only depends on $\tilde{p} < p$ variables, exhibiting no variation along the other $p - \tilde{p}$ dimensions, then the rate is determined by the effective dimension $\tilde{p}$. This setting can be true when the random vector is independent among the remaining $p - \tilde{p}$ components and a copula transformation is applied to each marginal distribution. The second property distinguishes the partition based method from those based on wavelet thresholding, as for the latter ones the rate is determined by the full dimension p without any pre-screening of variables.

To derive the convergence rate for density trees, our assumptions on the true density function are quite mild. We only assume it can be approximated by density trees at a reasonable rate, and it has finite second moment. Besides these two we do not further impose any boundary conditions or lower or upper bound of the true density. This implies that the method can even be applied to estimate some density with probability mass concentrating around the boundary, such as a Dirichlet distribution with parameters smaller than one.

Our method requires the density is supported by a bounded region in $\mathbb{R}^p$. It can also be applied to estimate unbounded distributions under certain approximation, such as one based on the range of data. However, if all observations lie on a low-dimensional manifold, implying a density on $\mathbb{R}^p$ does not exist, our method will fail.

5. Convergence Rates for Density Forests

The estimates obtained by using density trees are piece-wise constant functions. The approximation ability of these functions is limited by their non-continuous nature. This is especially the case when the underlying true density function enjoys good smoothness properties.

We take the Hölder class $\mathcal{H}^{k,\beta}(\Omega)$ as an example. Piece-wise constant functions can only approximate the function class $\mathcal{H}^{0,\beta}(\Omega)$ reasonably well. For more smooth functions, the unsatisfying approximation result leads to a suboptimal convergence rate. For instance, for the Hölder class $\mathcal{H}^{1,\beta}(\Omega)$, the minimax rate is $(n/\log n)^{-\frac{1+\beta}{2(1+\beta)+p}}$, where p is the dimension of data, while by simply applying density trees, the fastest rate is only $n^{-\frac{1}{(2+p)}}$ (up to a logarithmic factor).

To improve the approximation ability, in this section we introduce a new ensemble approach—*density forests*. A density forest is obtained by fitting a number of density trees and aggregating the estimates according to certain scheme. Similar to the popular ensemble methods for supervised learning, such as bagging and random forest, we hope trees in the ensemble to be slight different from each other. The diversity of trees is achieved by adding a small perturbation. For the proposed method, a shift will be added to the binary partition. We are inspired by Scott (1985), where the author has pointed out, using average shifted histograms is asymptotically equivalent to using a linear kernel if the shift

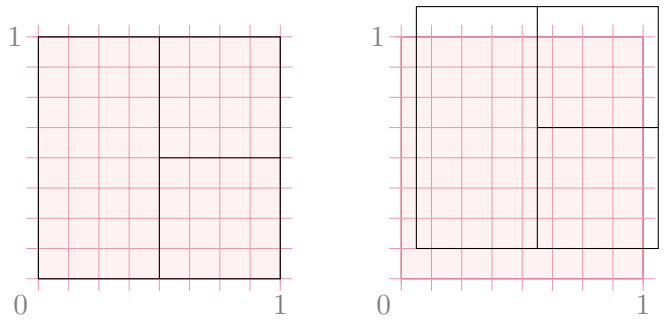

Figure 3: An example of shifted binary partitions in $\mathbb{R}^2$. Left: a binary partition of $[0,1]^2$. Right: shift of the partition along $h = (\frac{1}{16}, \frac{1}{8})$.

is carefully designed. However, in the paper this intuition has only been characterized in detail for twice differentiable density functions and the analysis is mainly for the univariate or bivariate case. Cui et al. (2021) studied a class of transformed histograms. But the transformation is too flexible so that the variance component is slightly out of control. Therefore they only obtained a suboptimal rate. Methods based on shifted histograms have also been examined by Randrianarisoa (2022), where the author has focused on the estimation of one-dimensional density functions. While in this paper, we will propose an ensemble approach for multivariate distributions and study its asymptotic properties.

5.1 Density Forests

We add small perturbations to trees by shifting the underlying binary partitions. This way, the split of a region can be more flexible.

5.1.1 Shifted Binary Partitions

We use Θ_t^h to denote the collection of piecewise constant density functions supported by shifted binary partitions of size t along the vector h ($h \in \mathbb{R}^p$). More precisely, $\{\Omega_j^h\}_{j=1}^t$ is called a shifted binary partition of size t along h if $\{\Omega_j^h - h\}_{j=1}^t$ is a binary partion defined before, where for any set B, $B - h$ is defined as the set $\{y : y + h \in B\}$. In Figure 3, we provide an example of a shifted binary partition in the two dimensional case. We define

$$\Theta_t^h = \Bigg\{ f : f = \sum_{j=1}^t \frac{\theta_j}{\mu(\Omega_j^h \cap \Omega)} \mathbb{1}_{\Omega_j^h \cap \Omega}, \sum_{j=1}^t \theta_j = 1,\ \theta_j \geq 0, \\ \{\Omega_j^h\}_{j=1}^t \text{ is a shifted binary partition.} \Bigg\}$$

One may raise the concern that for a density in Θ_t^h, the support of the distribution is only a subset of Ω due to the shift. Later we will introduce a boundary condition of the underlying distribution to circumvent this issue.

5.1.2 Equally Spaced and Balanced Partitions

By considering all binary partitions of size t, we have essentially employed a very flexible type of spatial approximation scheme, which allows reasonably well approximation to both Hölder classes and Besov ones simultaneously (see results from Section 4). When it is combined with the ensemble approach, the parameter space might be over-complex, leading to a large variance and consequently potential suboptimal rate. Therefore, we need to make certain restrictions to the shape of binary partitions. A type of balanced binary partitions is of particular interest. For a binary partition $\mathcal{A}_t = \{\Omega_j\}_{j=1}^t$, we use $e_j^l(\mathcal{A}_t)$ to denote the edge length of Ω_j along the dimension l. $E_{\min}^l(\mathcal{A}_t) \coloneqq \min_{1\le j\le t} e_j^l(\mathcal{A}_t)$ and $E_{\max}^l(\mathcal{A}_t) \coloneqq \max_{1\le j\le t} e_j^l(\mathcal{A}_t)$ denote the smallest and largest edge lengths along the l-th dimension over all rectangles within the partition. We say a partition is *balanced* across different dimensions if there exist positive constants c_1^e and c_2^e independent of t, such that

$$c_1^e \le \frac{\min_{1\le l\le p} E_{\min}^l(\mathcal{A}_t)}{\max_{1\le l\le p} E_{\max}^l(\mathcal{A}_t)} \le c_2^e. \quad \text{(balanced across all dimensions)} \tag{14}$$

Note that if property (14) holds for a binary partition $\mathcal{A}_t = \{\Omega_j\}_{j=1}^t$, then the same property also applies to the shifted one $\mathcal{A}_t^h = \{\Omega_j^h\}_{j=1}^t$.

5.1.3 Density Forests with Small Perturbations for the Hölder Class

When taking a closer look at binary partitions that are balanced across different dimensions, we notice that the number of different partitions of the same size meeting the criterion is finte, and moreover they are quite "similar" to each other. In particular, it is easy to show that for a partition of size t satisfying condition (14), there exists constants $c_1^{\text{space}}, c_2^{\text{space}} > 0$ not relying on t, such that

$$\mathcal{E}_t^{\min} \coloneqq \min_{\mathcal{A}_t:\ \mathcal{A}_t \text{ satisfies (14)}} \min_l E_{\min}^l(\mathcal{A}_t) \ge c_1^{\text{space}} t^{-1/p}, \tag{15}$$

$$\mathcal{E}_t^{\max} \coloneqq \max_{\mathcal{A}_t:\ \mathcal{A}_t \text{ satisfies (14)}} \max_l E_{\max}^l(\mathcal{A}_t) \le c_2^{\text{space}} t^{-1/p}. \tag{16}$$

Otherwise, the total volume of t subregions cannot be 1.

For each t, let $\mathcal{E}_t^{\max}$ be the largest possible edge length defined in (16). Then we design the shifts in the following way: $h = (i_l\delta)_{1\le l\le p}$, where $\delta = (\mathcal{E}_t^{\max})^2$ and $i_l \in [0, 1/\mathcal{E}_t^{\max})$ is an integer. This is to say, the possible shift along dimension l is $0, \delta, 2\delta, \ldots, \mathcal{E}_t^{\max} - \delta$. We enumerate all possible shifts as a sequence, denoted as $\{h_m\}_{1\le m\le M}$. It is worthy to note that when condition (14) is satisfied, the number of possible shifts M is bounded by t. With shape constraints, the parameter space under a shift h_m is defined as

$$\Theta_t^{(m)} \coloneqq \Bigg\{ f : f = \sum_{j=1}^t \frac{\theta_j}{\mu(\Omega_j^h \cap \Omega)} \mathbb{1}_{\Omega_j^h \cap \Omega}, \sum_{j=1}^t \theta_j = 1,\ \theta_j \ge 0, \\ \{\Omega_j^h\}_{j=1}^t \text{ is a shifted binary partition along } h_m \text{ satisfying (14).} \Bigg\}$$

Clearly, $\Theta_t^{(m)} \subset \Theta_t^{h_m}$. For the density forest, each single tree estimate is obtained by searching for the MLE over $\Theta_t^{(m)}$, and is denoted as $\hat{f}_{n,t}^{(m)}$. Then the density forest is defined

as

$$\hat{f}_{n,t}^{\text{forest}} = \frac{1}{M}\sum_{m=1}^{M} \hat{f}_{n,t}^{(m)}. \tag{17}$$

5.2 Convergence Rates

To derive the convergence rate, we first need to introduce a boundary condition.

H1. The density f_0 vanishes close to the boundary. More explicitly, let $\partial\Omega = \text{closure}(\Omega) \cap \text{closure}(\mathbb{R}^p \setminus \Omega)$ denote the boundary of Ω. For $\tau \geq 0$, let $\overline{\Omega}_\tau$ denote the set of points in Ω with distance no more than τ to the boundary

$$\overline{\Omega}_\tau = \{y \in \Omega : d(y, \partial\Omega) \leq \tau\},$$

where $d(y, \partial\Omega)$ denotes the infimum Euclidean distance between y and any point in $\partial\Omega$. Then there exists a $\tau_0 > 0$, such that $f_0 = 0$ a.e. on $\overline{\Omega}_{\tau_0}$.

The boundary condition H1 is mild in the sense that it can be easily satisfied after rescaling f_0 in combination with a smooth and fast decaying interpolation at the boundary. The convergence rate of density forests for the Hölder class is summarized in the following theorem.

Theorem 12 *For any $f_0 \in \mathcal{H}^{1,\beta}(\Omega), 0 < \beta \leq 1$, assume the condition H1 holds and let $\hat{f}_{n,t}^{forest}$ be the density forest defined in (17). When t and n satisfy*

$$M_1\left(\frac{n}{\log n}\right)^{\frac{p}{2(1+\beta)+p}} \leq t \leq M_2\left(\frac{n}{\log n}\right)^{\frac{p}{2(1+\beta)+p}}, \quad \textit{for some } 0 < M_1 \leq M_2,$$

the convergence rate of the density forest is $n^{-\frac{1+\beta}{2(1+\beta)+p}}(\log n)^{\frac{1}{2}+\frac{1+\beta}{2(1+\beta)+p}}$, in the sense that

$$\mathbb{P}_0^n\left(\rho(\hat{f}_{n,t}^{forest}, f_0) \geq Dn^{-\frac{1+\beta}{2(1+\beta)+p}}(\log n)^{\frac{1}{2}+\frac{1+\beta}{2(1+\beta)+p}}\right) \to 0,$$

where $D > 1$ is a constant and $\mathbb{P}_0^n$ is the probability measure on the product space corresponding to the true density f_0. A possible choice of t is

$$t = \left(((c_2^{space})^2 pL^2/\log(\mathcal{E}_t^{\max}/\mathcal{E}_t^{\min}))\frac{n}{\log n}\right)^{\frac{p}{2(1+\beta)+p}}. \tag{18}$$

The proof for the theorem will be provided in Section 7.4. In contrast, for the Hölder space $\mathcal{H}^{1,\beta}(\Omega)$ the fastest rate can be achieved by density trees is $n^{-\frac{1}{2+p}}$ (up to a logarithmic term). This indicates the gain of applying the ensemble approach. It is mainly due to the better approximation ability of density forests, as the aggregation of shifted binary partitions of a fixed size leads to a much finer partition, while the variance of the forest estimator is still under control.

6. Simulations

In this section, we use a range of numerical examples to illustrate convergence properties of density trees and density forests. For density trees, the MLE is obtained by a greedy search. The point estimate of the Bayesian method is defined to be the posterior mean. To draw samples from the posterior distribution, we employ a sequential importance sampling scheme, which will be introduced in the following subsection.

6.1 Sequential Importance Sampling

Each partition $\mathcal{A}_t = \{\Omega_j\}_{j=1}^t$ is obtained by recursively partitioning the sample space. We can use a sequence of partitions $\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_t$ to keep track of the path leading to $\mathcal{A}_t$. Let $\Pi_n(\cdot)$ denote the posterior distribution $\Pi_n(\cdot|Y_1, \cdots, Y_n)$ for simplicity, and Π_n^t be the posterior distribution conditioning on Θ_t. Then $\Pi_n^t(\mathcal{A}_t)$ can be decomposed as

$$\Pi_n^t(\mathcal{A}_t) = \Pi_n^t(\mathcal{A}_1)\Pi_n^t(\mathcal{A}_2|\mathcal{A}_1)\cdots\Pi_n^t(A_t|A_{t-1}).$$

The conditional distribution $\Pi_n^t(\mathcal{A}_{k+1}|\mathcal{A}_k)$ can be calculated by $\Pi_n^t(\mathcal{A}_{k+1})/\Pi_n^t(\mathcal{A}_k)$. However, the computation of the marginal distribution $\Pi_n^t(\mathcal{A}_k)$ is sometimes infeasible, especially when both t and $t-k$ are large, because we need to sum the marginal posterior probability over all binary partitions of size t for which the first k steps in the partition generating path are the same as those of $\mathcal{A}_k$. Therefore, we adopt the sequential importance sampling algorithm proposed in Lu et al. (2013). In order to build a sequence of binary partitions, at each step, the conditional distribution $\Pi_n^t(\mathcal{A}_{k+1}|\mathcal{A}_k)$ is approximated by $\Pi_n^{k+1}(\mathcal{A}_{k+1}|\mathcal{A}_k)$. Note that this approximation can become less accurate when k is much smaller than t, leading to an sequential importance sampling algorithm with high variance. However, from the computational perspective, it is much more efficient in the sense that the cost for calculating $\Pi_n^{k+1}(\mathcal{A}_{k+1}|\mathcal{A}_k)$ is $O(np)$ while that for $\Pi_n^t(\mathcal{A}_{k+1}|\mathcal{A}_k)$ is about $O(np^{t-k}t!/((k+1)!))$. The obtained partition is assigned a weight to compensate the approximation, where the weight is

$$w_t(\mathcal{A}_t) = \frac{\Pi_n^t(\mathcal{A}_t)}{\Pi_n^1(\mathcal{A}_1)\Pi_n^2(\mathcal{A}_2|\mathcal{A}_1)\cdots\Pi_n^t(\mathcal{A}_t|\mathcal{A}_{t-1})}.$$

When implementing the algorithm, we also employ a resampling and pruning scheme. We refer to the paper by Lu et al. (2013) for more details.

By applying such an algorithm, we can draw samples from the posterior distribution. Even in simulation studies both posterior mean and posterior mode converge under the KL divergence as expected, from the theoretical perspective, the convergence properties of the algorithm has not been fully explored.

As the number of binary partitions of size t increases very fast in t (at the rate $t!$), searching for the MLE over each Θ_t is an NP-hard problem. However, according to equation (5), after a logarithmic transformation, asymptotically the marginal posterior probability of a partition can be equivalently viewed as the penalized log-likelihood. Therefore, to obtain the MLE we may consider maximizing the penalized log-likehood with the penalty term matching that in (5), and developing a greedy algorithm which looks one step ahead at each level. In the following simulation studies, the MLE and the tree estimator in a forest is

obtained in this way if not otherwise specified. As observed in the following numerical examples, the greedy algorithm can provide a reasonably good approximation for low-dimensional distributions, but suffers from high variance for moderately high-dimensional cases.

In order to make the data points as uniform as possible, we apply a copula transformation to each variable in advance whenever the dimension exceeds 3. More specifically, we estimate the marginal distribution of each variable X_j by our approach, denoted as $\hat{f}_j$ (we use $\hat{F}_j$ to denote the cdf of X_j), and transform each point $(y^1, \cdots, y^p)$ to $(\hat{F}_1(y^1), \cdots, \hat{F}_p(y^p))$. Another advantage of this transformation is that after the transformation the sample space naturally becomes $[0,1]^p$.

6.2 Computational Complexity

For the Bayesian approach, assume that we sample L density trees from the posterior distribution. The computational cost of the importance sampling algorithm is mainly determined by the calculation of proposal distributions $\Pi_n^{k+1}(\mathcal{A}_{k+1}|\mathcal{A}_k)$ for $k = 1, \ldots t$. Without of loss generality, we may assume $\mathcal{A}_{k+1}$ is obtained by dividing the subregion Ω_j in $\mathcal{A}_k$ and Ω_j is further divided into $\Omega_j^{(1)}$ and $\Omega_j^{(2)}$, with number of observations $n_j^{(1)}$ and $n_j^{(2)}$ respectively. By the definition of the proposal distribution,

$$\Pi_n^{k+1}(\mathcal{A}_{k+1}|\mathcal{A}_k) = C(\mathcal{A}_k)2^{n_j}\frac{\Gamma(n_j^{(1)})\Gamma(n_j^{(2)})}{\Gamma(n_j)},$$

where $C(\mathcal{A}_k)$ is a constant that has been calculated for the partition $\mathcal{A}_k$. Therefore, the calculation of the proposal distribution involves the allocation of n_j points into two new subregions. By examining all subregions and all possible splits, at step k, the total complexity is at the order $O(npL)$.

For the MLE, the computational cost of the greedy algorithm for obtaining a single tree estimator can be estimated similarly, which is at the order $O(np)$ at each step.

For density forests, if there are M trees in a forest and each of them is an MLE, then the total cost is $O(npM)$.

In the following simulation studies, for the density forest, instead of aggregating over all possible shifts as defined in Section 5.1.3, we randomly sample the shift from a uniform distribution.

6.3 Low-Dimensional Distributions

For low-dimensional distributions, we compare density trees and forests with the kernel density estimator (KDE). For the KDE, the bandwidth is selected by using a cross-validation type approach. We use the first example to illustrate the performance of density trees.

Example 2 *This is Example 1 studied in Section 4.1. The density function is smooth in the sense that it belongs to a Hölder space $\mathcal{H}^{1,\beta}(\Omega)$ and consequently satisfies the power law decay condition (10).*

We apply the Bayesian density tree to this example, and allow the sample size to increase from 1×10^2 to 1×10^5. In Figure 4, the left plot is a visualization of the fitted density tree

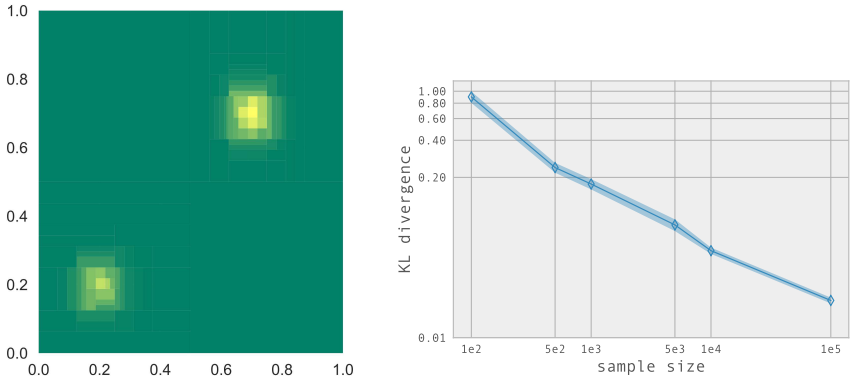

Figure 4: Plot of the estimated density and KL divergence against sample size. We use the posterior mean as the estimate. The right plot is on log-log scale, while the labels of x and y axes still represent the sample size and the KL divergence before we take the logarithm.

based on 10,000 samples. The right one is the plot of the Kullback-Leibler (KL) divergence from the estimated density to f_0 against sample size in log-scale. The linear trend in the plot validates the posterior concentration rates calculated in Section 3 and Section 4 . The reason why we use the KL divergence instead of the Hellinger distance is that for any $f_0 \in \mathcal{F}_0$ and $\hat{f} \in \Theta$, the squared Hellinger distance is bounded by the KL divergence. While the latter one is relatively easier to compute in our setting. For each fixed sample size, we run the experiment 10 times and estimate the standard error, which is shown by the lighter blue part in the plot.

We compare density trees and forests with the standard nonparametric density estimation method in the following two examples.

Example 3 *(Mixture of Beta distribution) Assume the distribution is*

$$Y \sim 0.7Beta(3, 12) + 0.3Beta(300, 10).$$

This is the a one-dimensional distribution with smooth density function. Therefore it lies in the Hölder space $\mathcal{H}^{1,\beta}(\Omega)$ and satisfies the power law decay condition (10). Given the sample size, for each method we run 50 repetitions to estimate the KL divergence and corresponding variance. For density forests, the ensemble consists of 200, 800, and 2000 trees, and the smallest shift size δ (see Section 5 for more details of this parameter) is 2^{-8}, 2^{-10} and 2^{-12} respectively. The comparison result is summarized in Table 1. Actually, the setting is very favorable for the KDE, but we see the density forest can achieve a even better performance. A possible explanation is that for this example the density function has sharp local variations.

Example 4 *(Trigonometric distribution) Assume the density is*

$$f(x) = 1 + \sin(2\pi x - \pi/2). \tag{19}$$

n	KDE	Density tree (MLE)	Density tree (Bayesian)	Density forest
5×10^2	0.13 (0.12)	0.056 (0.022)	0.053 (0.010)	**0.042** (0.014)
1×10^3	0.088 (0.0775)	0.0405 (0.0078)	0.0335 (0.0063)	**0.018** (0.0045)
1×10^4	0.015 (0.010)	0.0093 (0.00166)	0.0080 (0.0010)	**0.0053** (0.0010)

Table 1: The mean KL divergence for the mixture of Beta distribution example: for each sample size n, 50 replicates are performed. The standard deviation is reported in the parentheses. For each case, the best result is highlighted in bold.

n	KDE	Density tree (MLE)	Density tree (Bayesian)	Density forest
5×10^2	**0.0091** (0.0036)	0.023 (0.0073)	0.035 (0.0086)	0.014 (0.0054)
1×10^3	**0.0050** (0.0019)	0.017 (0.0055)	0.0206 (0.0061)	0.0094 (0.0034)
1×10^4	**0.0010** (0.00033)	0.0037 (0.00071)	0.0052 (0.00070)	0.0017 (0.00041)

Table 2: The mean KL divergence for the trigonometric distribution example: given the sample size n, 50 replicates are performed for each approach. The standard deviation is reported in the parentheses. For each case, the best result is highlighted in bold.

Compared to Example 3, the density function (19) has milder local variations while still satisfying the same type of regularity conditions. For density forests, the number of trees is 50, 200, and 200 with smallest shift size δ being 2^{-6}, 2^{-8}, and 2^{-8} respectively. We run 50 replicates for each case. As shown in Table 2, KDE can outperform tree based methods for this example.

6.4 Moderately High-Dimensional Examples

Example 5 *(Mixture of Gaussian distribution—5-30 dimensions) In the third example we work with a density function of moderately high dimensions. Assume that data are generated from the following location mixture of Gaussian distribution:*

$$\begin{aligned}
\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} &\sim \frac{1}{2}\mathcal{N}\left(\begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \begin{pmatrix} 0.05^2 & 0.03^2 & 0 \\ 0.03^2 & 0.05^2 & 0 \\ 0 & 0 & 0.05^2 \end{pmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix} 0.75 \\ 0.75 \\ 0.75 \end{pmatrix}, 0.05^2 I_{3\times 3}\right), \\
Y_4, Y_5 &\overset{i.i.d.}{\sim} \mathcal{N}(0.5, 0.01), \\
Y_6, \ldots, Y_p &\overset{i.i.d.}{\sim} \frac{1}{2}\mathcal{N}(0.35, 0.01) + \frac{1}{2}\mathcal{N}(0.6, 0.05^2), \quad \textit{for } p > 5.
\end{aligned}$$

The density function is continuous differentiable. Therefore, it lies in the Hölder space $\mathcal{H}^{1,1}(\Omega)$. We run experiments for $p = 5, 10$, and 30. For a fixed p, we generate $n = 500, 1 \times 10^3, 5 \times 10^3, 1 \times 10^4$, and 1×10^5 data points. For each pair of p and n, we repeat the experiment 10 times and calculate the standard error. As a multivariate density function after a marginal copula transformation may not satisfy the boundary condition H1, for density forests, we first implement forests to estimate marginal distributions, then a density tree is applied for the joint distribution after the copula transformation.

p	KDE	Density tree (MLE)	Density tree (Bayesian)	Density forest
5	0.583 (0.0010)	1.292 (0.62)	0.0226 (0.031)	**0.0117** (0.0027)
10	1.456 (0.00081)	1.350 (0.676)	0.03895 (0.0046)	**0.0191** (0.0045)
30	7.724 (0.010)	1.275 (0.82)	0.08495 (0.0063)	**0.0351** (0.0049)

Table 3: The mean KL divergence for the mixture Gaussian distribution example: for each case, the sample size is 5×10^4 and 10 replicates are performed. The standard deviation is reported in the parentheses. For each case, the best result is highlighted in bold.

We compare the performance of density trees and forests with KDE, and the results are summarized in Table 3. For the three cases listed in the table, the sample size is 5×10^4, and the number of trees in the forest is 800, 800, and 400 respectively with smallest shift size δ all equal to 2^{-10}. We can see that the performance of the frequentist tree obtained by the greedy algorithm is much worse than the Bayesian one due to the high variance. Therefore, for density forests, we aggregate the posterior mode of the Bayesian method instead of the MLE. When the dimension is moderately high, the cross-validation type bandwidth selection for the KDE becomes more computationally intensive. Therefore, bandwidths are chosen by Silverman's "rule of thumb" (Silverman, 1986) and "Maximal Smoothing Principle" (Terrell, 1990). We can see the performance of KDE deteriorates fast as the dimension increases, as it is more difficult to select the bandwidth.

Figure 5: KL divergence vs. sample size. The red, green and blue curves correspond to the cases when $p = 5, 10$ and 30 respectively. The slopes of the three lines are almost the same, implying that the concentration rate only depends on the effective dimension of the problem (which is 2 in this case).

Then we check the convergence of density trees by plotting the KL divergence against the sample size in log-log scale. After a copula transformation, the effective dimension of this example is $\tilde{p} = 2$, and this is reflected in the visualization of convergence rate for density trees in Figure 5: slopes of the three lines, which correspond to the concentration

rates under different dimensions, almost remain the same no matter how the full dimension of the problem varies. The observation verifies the implication of Corollary 9, which is summarized in Remark 11.

6.5 Summary of Simulation Studies

From these examples, we have learned that for low dimensional distributions, partition based methods can better adapt to sharp local changes, while when the density only has mild variations, a method that can take advantage of the smoothness, such as the kernel density estimator, performs better. However, for moderately high-dimensional cases, especially when the density is a mixture of several low dimensional components, partition based methods can achieve a significantly lower error. This is particularly the case for estimates obtained by applying the sequential importance sampling algorithm. Greedy search of the MLE suffers from high variance as the dimension increases, although theoretically the frequentist method should perform similarly compared to the Bayesian one. As in our simulation studies the true density function is quite smooth, the ensemble method always outperforms the density tree. This is mainly due to the lower bias of density forests.

7. Proofs

In this section, we provide detailed proofs for the main results. Our proofs rely on the previous results from studies of empirical processes indexed by log-likelihood ratios. However, while the results in the landmark work by Shen and Wong (1994); Wong and Shen (1995); Ghosal et al. (2000); Shen and Wasserman (2001) are the most applicable ones to our study, they must be modified to adapt to the current settings. In Section 7.1, we first derive some useful tools.

7.1 Preliminaries

In Section 7.1.1 we briefly discuss *metric entropy with bracketing*, which measures the complexity of the approximating spaces by "counting" how many pairs of functions in an ϵ-net are needed to provide simultaneous upper and lower bounds of all the elements. An important result of this section is an upper bound for the bracketing metric entropy of Θ_t. Properties of lower-truncated likelihood ratios are summarized in Section 7.1.2. In Section 7.1.3, a large-deviation type inequality for the likelihood ratio surface are provided.

Previous results on the convergence rate of sieve MLE assume that the true parameter can be approximated by the sieve under Kullback-Leibler divergence. Here we impose a weaker assumption in terms of the Hellinger distance. This is because under the current settings, we can obtain an explicit bound for the Kullback-Leibler divergence in terms of the Hellinger distance. This result is stated in Section 7.1.4.

7.1.1 Calculation of the Metric Entropy with Bracketing

A general discussion of *metric entropy* can be found in the paper by Kolmogorov and Tikhomirov (1992). In this section, we introduce a form of metric entropy with bracketing corresponding to the parameter space under consideration, and provide an upper bound for the bracketing metric entropy of the approximating spaces defined in Section 2.1.

Definition 13 *Let (Θ, ρ) be a seperable pseudo-metric space. $\Theta(\epsilon)$ is a finite set of pairs of functions $\{(f_j^L, f_j^U), j = 1, \cdots, N\}$ satisfying*

$$\rho(f_j^L, f_j^U) \leq \epsilon \text{ for } j = 1, \cdots, N, \tag{20}$$

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

$$f_j^L \leq f \leq f_j^U. \tag{21}$$

Let

$$N^{[]}(\epsilon, \Theta, \rho) = min\{|\Theta(\epsilon)| : \Theta(\epsilon) \text{ is a set satisfying (20) and (21)}\}.$$

The metric entropy with bracketing of Θ is defined to be

$$H^{[]}(\epsilon, \Theta, \rho) = \log N^{[]}(\epsilon, \Theta, \rho).$$

Recall that $\Theta_1, \cdots, \Theta_t, \cdots$ are the approximating spaces defined in section 2.1. The next two lemmas are devoted to an upper bound for the bracketing metric entropy of Θ_t.

Lemma 14 *Take ρ to be the Hellinger distance. Let $\Theta_t^{\mathcal{A},d} = \{f \in \Theta_t : f \text{ is supported by the binary partition } \mathcal{A} = \{\Omega_j\}_{j=1}^t, \text{ and } \rho(f, f_0) \leq d\}$. Then,*

$$H^{[]}(u, \Theta_t^{\mathcal{A},d}, \rho) \leq \frac{t}{2} \log t + t \log \frac{d}{u} + b',$$

where b' is a constant not dependent on the binary partition.

Proof Assume $f = \sum_{j=1}^t \beta_j \mathbb{1}_{\Omega_j}$. When the binary partitions $\{\Omega_j\}_{j=1}^t$ are fixed, there exits a one-to-one correspondance between any $f \in \Theta_t^{\mathcal{A},d}$ and an t-dimensional vector $\left(\sqrt{\beta_1 \mu(\Omega_1)}, \cdots, \sqrt{\beta_t \mu(\Omega_t)}\right)$. As a consequence of Cauchy-Schwartz inequality,

$$\begin{aligned}
\rho(f, f_0)^2 &= \sum_{j=1}^t \int_{\Omega_j} \left(\sqrt{\beta_j} - \sqrt{f_0(x)}\right)^2 \mu(dx) \\
&\geq \sum_{j=1}^t \mu(\Omega_j) \left(\int_{\Omega_j} (\sqrt{\beta_j} - \sqrt{f_0(x)}) \frac{\mu(dx)}{\mu(\Omega_j)} \right)^2 \\
&= \sum_{j=1}^t \mu(\Omega_j) \left(\sqrt{\beta_j} - \frac{\int_{\Omega_j} \sqrt{f_0(x)} \mu(dx)}{\mu(\Omega_j)} \right)^2.
\end{aligned}$$

Then, we have,

$$\begin{aligned}
&\left\{ \left(\sqrt{\beta_1 \mu(\Omega_1)}, \cdots, \sqrt{\beta_t \mu(\Omega_t)}\right) : \sum_{j=1}^t \int_{\Omega_j} \left(\sqrt{\beta_j} - \sqrt{f_0(x)}\right)^2 \mu(dx) \leq d^2 \right\} \\
&\subset \left\{ \left(\sqrt{\beta_1 \mu(\Omega_1)}, \cdots, \sqrt{\beta_t \mu(\Omega_t)}\right) : \sum_{j=1}^t \left(\sqrt{\beta_j \mu(\Omega_j)} - \frac{\int_{\Omega_j} \sqrt{f_0(x)} \mu(dx)}{\sqrt{\mu(\Omega_j)}} \right)^2 \leq d^2 \right\} \\
&=: B_t^{\mathcal{A},d}.
\end{aligned}$$

If we treat the element in $\Theta_t^{\mathcal{A},d}$ as the t-dimensional vector $\left(\sqrt{\beta_1\mu(\Omega_1)},\cdots,\sqrt{\beta_t\mu(\Omega_t)}\right)$, then from the above inclusion relation we learn that, $\Theta_t^{\mathcal{A},d} \subset B_t^{\mathcal{A},d}$. We also note that the Hellinger distance on $B_t^{\mathcal{A},d}$ is equivalent to the L_2 norm on the t-dimensional Euclidean space. Thus,

$$N^{[]}(u,\Theta_t^{\mathcal{A},d},\rho) \leq N^{[]}(u,B_t^{\mathcal{A},d},\|\cdot\|_2).$$

Because the metric entropy is invariant under translation, calculating the bracketing metric entropy of $B_t^{\mathcal{A},d}$ is equivalent to calculating that of

$$\tilde{B}_t^{\mathcal{A},d} := \left\{ \left(\sqrt{\beta_1\mu(\Omega_1)},\cdots,\sqrt{\beta_t\mu(\Omega_t)}\right) : \sum_{j=1}^{t}\left(\sqrt{\beta_j\mu(\Omega_j)}\right)^2 \leq d^2 \right\}.$$

The unit sphere under L_2-norm is

$$S = \left\{ \left(\sqrt{\beta_1\mu(\Omega_1)},\cdots,\sqrt{\beta_t\mu(\Omega_t)}\right) : \sum_{j=1}^{t}(\beta_j\mu(\Omega_j)) \leq 1 \right\}.$$

The unit sphere under L_∞-norm is

$$S_\infty = \left\{ \left(\sqrt{\beta_1\mu(\Omega_1)},\cdots,\sqrt{\beta_t\mu(\Omega_t)}\right) : \max_{1\leq j\leq t}\sqrt{\beta_j\mu(\Omega_j)} \leq 1 \right\}.$$

Note that $\max_{1\leq j\leq t}\sqrt{\beta_j\mu(\Omega_j)} \leq 1/\sqrt{t}$ implies that $\sum_{j=1}^{t}\beta_j\mu(\Omega_j) \leq 1$, and $\sum_{j=1}^{t}\beta_j\mu(\Omega_j) \leq d^2$ implies that $\max_{1\leq j\leq t}\sqrt{\beta_j\mu(\Omega_j)} \leq d$, we have

$$S \subset \tilde{B}_t^{\mathcal{A},d} \subset dS_\infty \quad \text{and} \quad S_\infty \subset \sqrt{t}S.$$

Note that the bracketing metric entropy of $d\sqrt{t}S$ under $\|\cdot\|_2$ is bounded by its metric entropy (without bracketing) under $\|\cdot\|_\infty$. Therefore,

$$N^{[]}(u,\Theta_t^{\mathcal{A},d},\rho) \leq N^{[]}(u,\tilde{B}_t^{\mathcal{A},d},\|\cdot\|_2) \leq \left(\frac{d\sqrt{t}}{u}+2\right)^t \leq b't^{t/2}(d/u)^t,$$

where b' is a constant not dependent on the partiton. The desired result follows. ■

Lemma 15 *Under the same assumptions as in Lemma 14, let* $\Theta_t^d = \{f \in \Theta_t : \rho(f,f_0) \leq d\}$. *Then,*

$$H^{[]}(u,\Theta_t^d,\rho) \leq t\log p + (t+1)\log(t+1) + \frac{t}{2}\log t + t\log\frac{d}{u} + b,$$

where b *is a constant not dependent on* t *or* d.

Proof According to the construction of the sieve, given the size t, the number of possible binary partitions is upper bounded by $p^t t!$ (p is the dimension of the Euclidean space). Therefore,

$$N^{[]}(u,\Theta_t^d,\rho) \leq p^t t! N^{[]}(u,\Theta_t^{\mathcal{A},d},\rho) \leq b'p^t t! t^{t/2}(d/u)^t,$$

and the result is obtained by taking logarithm on both sides. ■

7.1.2 Lower Truncation of Log-Likelihood Ratios

A key step to derive the convergence rates for partition-based methods is to obtain a concentration inequality for the likelihood ratio $\Pi_{i=1}^n f(Y_i)/f_0(Y_i)$, or equivalently, the log-likelihood ratio,

$$l_n(f) - l_n(f_0) = \sum_{i=1}^n Z_f(Y_i),$$

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

$$\tilde{f} = \begin{cases} f, & \text{if } f > \exp(-\tau) f_0, \\ \exp(-\tau) f_0, & \text{if } f \le \exp(-\tau) f_0. \end{cases} \qquad \widetilde{Z}_f = Z_{\tilde{f}} = \begin{cases} Z_f, & \text{if } Z_f > -\tau, \\ -\tau, & \text{if } Z_f \le -\tau. \end{cases}$$

We will cite two results by Wong and Shen (1995) to demonstrate that after the truncation, $\widetilde{Z}_f$ still maintains some key properties of the log-likelihood ratio. This guarantees that the behavior of the empirical 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 process 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 16 *Let $\gamma = 2\exp(-\tau/2)/(1-\exp(-\tau/2))^2$. Then*

$$\mathbb{E}\widetilde{Z}_f \le -(1-\gamma)\|f^{1/2} - f_0^{1/2}\|_2^2.$$

Proof See the paper by Wong and Shen (1995), Lemma 4 in Section 2. ■

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

Lemma 17 *(One-Sided Large Deviation Inequality) Let $\nu_n(\widetilde{Z}_f) = n^{-1/2}\sum_{i=1}^n (\widetilde{Z}_f(Y_i) - \mathbb{E}\widetilde{Z}_f(Y_i))$. Let $\mathcal{G}$ be a class of densities with bracketing Hellinger metric entropy $H^{[]}(u, \mathcal{G}, \rho)$. For $d > 0$, consider the empirical process*

$$\{\nu_n(\widetilde{Z}_f) : f \in \mathcal{G}, \rho(f, f_0) \le d\}$$

induced by the truncated log-likelihood ratios for $f \in \mathcal{G}$ inside a Hellinger ball around f_0. For any $d > 0$, $0 < b < 1$ and $L > 0$, let

$$\varphi(L, d^2, n) = L^2/8(8c_0 d^2 + L/n^{1/2}),$$

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

$$L \le bn^{1/2}d^2/4, \tag{22}$$

and

$$\int_{bL/(32n^{1/2})}^{d} \Big(H^{[]}(u/(2\exp(\tau/2)),\mathcal{G},\rho)\Big)^{1/2} du \leq Lb^{3/2}/(2^{10}(c_0+1/8)). \tag{23}$$

Then

$$\mathbb{P}_{f_0}\left(\sup_{\{\|f^{1/2}-f_0^{1/2}\|_2\leq d, f\in\mathcal{G}\}} \nu_n(\widetilde{Z}_f) \geq L\right) \leq 3\exp(-(1-b)\varphi(L,d^2,n)),$$

where $\mathbb{P}_{f_0}$ is understood to be the outer probability measure under f_0.

Proof See the paper by Wong and Shen (1995), Lemma 7 in Section 2. ■

7.1.3 An inequality for the likelihood ratio surface

In this section, we focus on bounding the tail of the likelihood ratio. First, we cite a theorem by Wong and Shen (1995), which provides a uniform exponential bound for likelihood ratios, when the metric entropy of the parameter space is under control. The result can be shown by applying Lemma 17.

Lemma 18 *Let ρ be the Hellinger distance and $\mathcal{P}_n$ be a space of densities. There exist positive constants $a>0$, c, c_1 and c_2, such that, for any $\epsilon>0$, if*

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \Big(H^{[]}(u/a,\mathcal{P}_n,\rho)\Big)^{1/2} du \leq cn^{1/2}\epsilon^2, \tag{24}$$

then

$$\mathbb{P}_{f_0}\left(\sup_{\{\rho(f,f_0)\geq\epsilon, f\in\mathcal{P}_n\}} \prod_{i=1}^{n}\frac{f(Y_i)}{f_0(Y_i)} \geq \exp(-c_1n\epsilon^2)\right) \leq 4\exp(-c_2n\epsilon^2),$$

where $\mathbb{P}_{f_0}$ is understood to be the outer probability mesure under f_0. The constants c_1 and c_2 can be chosen in $(0,1)$ and c can be set as $(2/3)^{5/2}/512$.

Proof See the paper by Wong and Shen (1995), Theorem 1 in Section 3. ■

In our case, we need to strike a balance between the complexity of the parameter space indexed by t and the sample size n, such that the condition (24) is satisfied. The balance is achieved by considering the space Θ_t for which the metric entropy of a $\delta_{n,t}$-net can be controlled, where $\delta_{n,t}$ is at the order $(\frac{t\log t}{n/\log n})^{1/2}$. Based on Lemma 18 and the entropy bounds provided in Section 7.1.1, we obtain the following uniform exponential bound for the likelihood ratio.

Lemma 19 *Let $\delta_{n,t}=(\frac{t\log t}{n/\log n})^{1/2}$, where $t=O(n^{\frac{1}{2r+1}})$ for some $r>0$. When n and t are sufficiently large, we have*

$$\mathbb{P}_{f_0}\Big(\sup_{\{\rho(f,f_0)\geq\delta_{n,t}, f\in\Theta_t\}} \prod_{i=1}^{n}\frac{f(Y_i)}{f_0(Y_i)} \geq \exp(-c_1n\delta_{n,t}^2)\Big) \leq 4\exp(-c_2n\delta_{n,t}^2).$$

Proof The key to the proof is to check condition (24) so that we can apply Lemma 18 to the approximating space Θ_t. By Lemma 15,

$$\begin{aligned}
&\int_{\delta_{n,t}^2/2^8}^{\sqrt{2}\delta_{n,t}} \left(H^{[]}(u/a, \Theta_t, \rho)\right)^{1/2} du \\
&\quad \leq \int_{\delta_{n,t}^2/2^8}^{\sqrt{2}\delta_{n,t}} \left(t\log(4pa) + 2(t+1)\log(t+1) - t\log u\right)^{1/2} du \\
&\quad \asymp t^{1/2} \int_{\delta_{n,t}^2/2^8}^{\sqrt{2}\delta_{n,t}} \left(\log \frac{t^2}{u}\right)^{1/2} du.
\end{aligned}$$

We calculate the integral and obtain

$$\begin{aligned}
&\int_{\delta_{n,t}^2/2^8}^{\sqrt{2}\delta_{n,t}} \left(H^{[]}(u/a, \Theta_t, \rho)\right)^{1/2} du \\
&\leq t^{1/2}\left(u\sqrt{\log(t^2/u)} - \frac{\sqrt{\pi}}{2}t^2\mathrm{erf}\left(\sqrt{\log(t^2/u)}\right)\right)\Bigg|_{\delta_{n,t}^2/2^8}^{\sqrt{2}\delta_{n,t}} \\
&\leq t^{1/2}\Bigg(\sqrt{2}\delta_{n,t}\sqrt{\log\left(t^2/(\sqrt{2}\delta_{n,t})\right)} - \frac{\sqrt{\pi}}{2}t^2\mathrm{erf}\left(\sqrt{\log\left(t^2/(\sqrt{2}\delta_{n,t})\right)}\right) \\
&\quad - \frac{\delta_{n,t}^2}{2^8}\sqrt{\log\left(2^8t^2/\delta_{n,t}^2\right)} + \frac{\sqrt{\pi}}{2}t^2\mathrm{erf}\left(\sqrt{\log\left(2^8t^2/\delta_{n,t}^2\right)}\right)\Bigg) \\
&\asymp t^{1/2}\left(\delta_{n,t}\sqrt{\log\left(t^2/\delta_{n,t}\right)} + \frac{\sqrt{\pi}}{2}t^2\left(\mathrm{erf}\left(\sqrt{\log\left(2^8t^2/\delta_{n,t}^2\right)}\right) - \mathrm{erf}\left(\sqrt{\log\left(t^2/(\sqrt{2}\delta_{n,t})\right)}\right)\right)\right) \\
&\asymp t^{1/2}\left(\delta_{n,t}\sqrt{\log\left(t^2/\delta_{n,t}\right)} + \frac{\sqrt{\pi}}{2}t^2\left(1 - \frac{\delta_{n,t}^2/(2^8t^2)}{\sqrt{\pi\log\left(2^8t^2/\delta_{n,t}^2\right)}} - 1 + \frac{\sqrt{2}\delta_{n,t}/t^2}{\sqrt{\pi\log\left(t^2/(\sqrt{2}\delta_{n,t})\right)}}\right)\right) \\
&\asymp t^{1/2}\delta_{n,t}\sqrt{\log\left(t^2/\delta_{n,t}\right)} \\
&\leq c\sqrt{n}\delta_{n,t}^2.
\end{aligned}$$

Therefore, condition (24) is satisfied. The desired result follows from Lemma 18. ■

Remark 20 *Since the metric entropy decreases as the width of the net ϵ increases, this lemma also holds for any $\epsilon \geq \delta_{n,t}$. This property is quite useful for deriving the posterior concentration rate.*

7.1.4 An Inequality for the Kullback-Leibler Divergence

It is well known that the Hellinger distance can be bounded by the Kullback-Leibler divergence. In the paper by Wong and Shen (1995), the authors showed that the other direction also holds under certain conditions. This type of result becomes quite useful in this paper because it would allow us to impose the assumption on approximation rate under the weaker Hellinger distance. We first cite the result from their paper, and then derive a more

explicit bound for density functions in $\mathcal{F}_0$ that can be approximated by spaces Θ_t at a rate t^{-r} under the Hellinger distance.

Lemma 21 *Let f, f_0 be two densities, $\rho^2(f_0, f) \leq \epsilon^2$. Suppose that $M_\lambda^2 = \int_{\{f_0/f \geq e^{1/\lambda}\}} f_0 (f_0/f)^\lambda < \infty$ for some $\lambda \in (0, 1]$. Then for all $\epsilon^2 \leq \frac{1}{2}(1 - e^{-1})^2$, we have*

$$\int f_0 \log(f_0/f) \leq \left(6 + \frac{2\log 2}{(1-e^{-1})^2} + (8/\lambda)\max\{1, \log(M_\lambda/\epsilon)\}\right)\epsilon^2,$$

$$\int f_0 (\log(f_0/f))^2 \leq 5\epsilon^2 \left(\frac{1}{\lambda}\max\{1, \log(M_\lambda/\epsilon)\}\right)^2.$$

Proof See the paper by Wong and Shen (1995), Theorem 5 in Section 6. ■

Assume that f_0 is the true density function in $\mathcal{F}_0$, and f_t is an approximation to f_0 in Θ_t. We can obtain an explicit bound of the approximation error to f_0 by Θ_t under the Kullback-Leibler divergence in terms of the Hellinger distance between f_0 and f_t. The result is summarized in the lemma below.

Lemma 22 *f_0 is a density function defined on Ω. If $f_0 \in \mathcal{F}_0$, then we can find an approximation $g_t \in \Theta_t$, such that*

$$\int f_0 \log(f_0/g_t) \leq 128 A^2 r t^{-2r} \log t,$$

$$\int f_0 (\log(f_0/g_t))^2 \leq 320 A^2 r^2 t^{-2r} (\log t)^2.$$

Proof Assume that $f_t = \sum_{j=1}^t \beta_j \mathbb{1}_{\Omega_j}$ is an approximation to f_0, where $\{\Omega_j\}_{j=1}^t$ is a binary partition of Ω, and $\rho(f_0, f_t) \leq A t^{-r}$. Based on the property of L_2-projection, we have

$$\begin{aligned}
\rho^2(f_0, f_t) &= \sum_{j=1}^t \int_{\Omega_j} (\sqrt{f_0(x)} - \sqrt{\beta_j})^2 \mu(dx) \\
&\geq \sum_{j=1}^t \int_{\Omega_j} \left(\sqrt{f_0} - \beta_j^0\right)^2, \quad \text{where } \beta_j^0 = \int_{\Omega_j} \sqrt{f_0}/\mu(\Omega_j).
\end{aligned}$$

Following this property, we can construct an alternative approximation to f_0 defined on the same binary partition, such that the new approximation also achieve an error at the order of t^{-r}. Let $h_t = \sum_{j=1}^t \left(\beta_j^0\right)^2 \cdot \mathbb{1}_{\Omega_j}$, then

$$\int_\Omega h_t = \sum_{j=1}^t \frac{(\int_{\Omega_j} \sqrt{f_0})^2}{\mu(\Omega_j)} \leq \sum_{j=1}^t \frac{(\int_{\Omega_j} f_0)\mu(\Omega_j)}{\mu(\Omega_j)} = 1.$$

Define a density function g_t as $g_t = h_t / \int_\Omega h_t$. Then we have

$$\begin{aligned}
\rho^2(f_0, g_t) &= \left\| \sqrt{f_0} - \sqrt{h_t} + \sqrt{h_t} - \left(\sqrt{h_t}/\|\sqrt{h_t}\|_2\right)\right\|_2^2 \\
&\leq 2\left\|\sqrt{f_0} - \sqrt{h_t}\right\|_2^2 + 2\left(1 - 1/\|\sqrt{h_t}\|_2\right)^2 \|\sqrt{h_t}\|_2^2 \leq 4\rho^2(f_0, f_t).
\end{aligned}$$

The result implies that, if there exists a function $f_t \in \Theta_t$, such that $\rho(f_0, f_t) \leq At^{-r}$, then we can define a density function g_t supported by the same binary partition as that of f_t in the above way, such that $g_t \in \Theta_t$ and $\rho(f_0, g_t) \leq 2At^{-r}$. Next, for this specific approximation g_t, we check whether the conditions of Lemma 21 are satisfied. First, as $\|\sqrt{h_t}\|_2 \leq 1$,

$$M_{1/4}^2 = \sum_{j=1}^{t} \int_{\Omega_j \cap \{f_0/g_t > e^4\}} f_0 \left(f_0 / \left(\beta_j^0 / \|\sqrt{h_t}\|_2 \right)^2 \right)^{1/4} \leq \sum_{j=1}^{t} \left(\int_{\Omega_j} f_0^{1+1/4} \right) \Big/ \left(\beta_j^0 \right)^{1/2}.$$

By applying the Cauchy-Schwarz inequality, we have $M_{1/4}^2 \leq \left(\int_\Omega f_0^2 \right)^{1/2}$. Therefore, if we set $\lambda = 1/4$, when t is large enough

$$\begin{aligned} \int f_0 \log(f_0/g_t) &\leq \left(6 + \frac{2\log 2}{(1-e^{-1})^2} + 32 \max\left\{ 1, \log \frac{(\int f_0^2)^{1/4}}{2At^{-r}} \right\} \right) \cdot 4A^2 t^{-2r} \\ &\leq 128 A^2 r t^{-2r} \log t. \end{aligned}$$

Similarly,

$$\int f_0 (\log(f_0/g_t))^2 \leq 320 A^2 r^2 t^{-2r} (\log t)^2.$$

■

7.2 Proof of Theorem 1

In this section, we apply the previous uniform bound for the likelihood ratio together with the bound for the Kullback-Leibler divergence to derive convergence rate of the sieve MLE.

Proof [Proof of Theorem 1] For any $g \in \Theta_t$ and $D > 1$, we have

$$\mathbb{P}_0^n \left(\rho(f_0, \hat{f}_{n,t}) \geq D\delta_{n,t} \right) \leq \mathbb{P}_{f_0} \left(\sup_{\{\rho(f_0,f) \geq D\delta_{n,t}, f \in \Theta_t\}} \prod_{i=1}^{n} f(Y_i)/g(Y_i) \geq 1 \right), \tag{25}$$

where $\mathbb{P}_{f_0}$ is understood to be the outer probability measure under f_0. Let $C = \{f \in \Theta_t : \rho(f_0, f) \geq D\delta_{n,t}\}$. Then for $g \in \Theta_t$, the right hand of (25) can be bounded by

$$\mathbb{P}_{f_0} \left(\sup_{f \in C} \prod_{i=1}^{n} f(Y_i)/g(Y_i) \geq 1 \right) \leq P_1 + P_2,$$

where

$$P_1 = \mathbb{P}_{f_0} \left(\sup_{f \in C} \prod_{i=1}^{n} f(Y_i)/f_0(Y_i) \geq \exp\left(-c_1 n (D\delta_{n,t})^2 \right) \right),$$

$$P_2 = \mathbb{P}_0^n \left(\prod_{i=1}^{n} f_0(Y_i)/g(Y_i) \geq \exp\left(c_1 n (D\delta_{n,t})^2 \right) \right).$$

In order to bound P_1, we can still apply Lemma 19 here with $\delta_{n,t}$ replaced by $D\delta_{n,t}$. Therefore, $P_1 \leq 4\exp(-c_2 n D^2 \delta_{n,t}^2)$.

To bound P_2, we can write it as

$$\begin{aligned} P_2 &= \mathbb{P}_0^n \left(\sum_{i=1}^n \log\left(f_0/g\right)(Y_i) \geq c_1 n (D\delta_{n,t})^2 \right) \\ &= \mathbb{P}_0^n \left(\sum_{i=1}^n \left(\log\left(f_0/g\right)(Y_i) - \int f_0 \log\left(f_0/g\right) \right) \geq c_1 n (D\delta_{n,t})^2 - n \int f_0 \log\left(f_0/g\right) \right). \end{aligned}$$

If $\int f_0 \log(f_0/g) < c_1 D^2 \delta_{n,t}^2$, then

$$P_2 \leq \frac{n \int f_0 \log\left(f_0/g\right)^2}{n^2 \left(c_1 D^2 \delta_{n,t}^2 - \int f_0 \log\left(f_0/g\right)\right)^2}.$$

Based on our assumption, there exists $f_t \in \Theta_t$, such that $\rho(f_0, f_t) \leq A t^{-r}$. Then by applying Lemma 22, we have

$$\int f_0 \log(f_0/f_t) \leq 128 A^2 r t^{-2r} \log t,$$
$$\int f_0 (\log(f_0/f_t))^2 \leq 320 A^2 r^2 t^{-2r} (\log t)^2.$$

Therefore,

$$\inf_{g \in \Theta_t} P_2 \leq \frac{320 A^2 r^2 t^{-2r} (\log t)^2}{n \left(c_1 D^2 \delta_{n,t}^2 - 128 A^2 r t^{-2r} \log t\right)^2}.$$

If we take $t = \left((2^8 A^2 r / c_1) \frac{n}{\log n} \right)^{\frac{1}{2r+1}}$, then the condition $\int f_0 \log(f_0/f_t) < c_1 D^2 \delta_{n,t}^2$ is satisfied. If n and t are matched in this way, the order of $\delta_{n,t}$ determines the final convergence rate, which is $n^{-\frac{r}{2r+1}} (\log n)^{(\frac{1}{2} + \frac{r}{2r+1})}$. This finishes the proof. ■

7.3 Proof of Theorem 4

The posterior concentration rate of the Bayesian method is obtained by bounding the numerator and denominator of (8) simultaneously. For the upper bound of the numerator, we apply both Lemma 17 and Lemma 19. To complete the proof, we first focus on the concentration property of the prior in Section 7.3.1, and then combine these two bounds to derive the posterior concentration rate in Section 7.3.2.

7.3.1 Lower Bound of the Denominator (Prior Thickness)

In this section, we study how the prior distribution concentrates on the shrinking neighborhoods around the true density function. We develop our result through a series of lemmas. The connection between lower bounds of the items in the denominator of (8) and the concentration property of the prior distribution is first revealed by Lemma 23. By employing a property of the Dirichlet distribution (summarized in Lemma 24) and inequalities bounding the Kullback-Leibler divergence by the Hellinger distance (Lemma 19), we obtain lower bounds of the items in the denominator of (8) in Lemma 25.

To begin with, we cite a result by Shen and Wasserman (2001). In this lemma, it is shown that with probability close to 1, the denominator is bounded from below by the prior probability mass concentrating on a ball around f_0 multiplied by a coefficient depending on the radius of the ball. Recall that,

$$K(f_0, f) = \mathbb{E}_{f_0}\Big(\log \frac{f_0(Y)}{f(Y)}\Big), \quad V(f_0, f) = \mathrm{Var}_{f_0}\Big(\log \frac{f_0(Y)}{f(Y)}\Big).$$

Lemma 23 *Let $K(\cdot,\cdot)$ and $V(\cdot,\cdot)$ be quantities defined above, and let $S(d) = \{f \in \Omega : K(f_0, f) \le d, V(f_0, f) \le d\}$. Set $S_n = S(d_n)$. When d_n is a sequence of positive numbers satisfying $nd_n \to \infty$,*

$$\mathbb{P}_0^n \left(\int_\Omega \prod_{i=1}^n \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f) \le \frac{1}{2}\Pi(S_n) e^{-2nd_n} \right) \le \frac{2}{nd_n}.$$

Proof See the paper by Shen and Wasserman (2001), Lemma 1 in Section 3. ■

More explicitly, from this lemma we learn that, given the condition $nd_n \to \infty$, $\int_\Omega \prod_{i=1}^n \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f) \ge \frac{1}{2}\Pi(S_n)e^{-2nd_n}$ with probability close to 1.

Lemma 21 tells us that, although the Kullback-Leibler divergence can not be bounded by the Hellinger distance, under mild conditions it only inflates by a logarithmic factor. Based on this lemma, we have for any $f \in \cup_{t=1}^\infty \Theta_t$, if $\rho^2(f, f_0) \le \epsilon^2$, then

$$\max\Big\{K(f_0, f), \mathbb{E}_{f_0}\Big(\Big(\log \frac{f_0(Y)}{f(Y)}\Big)^2\Big)\Big\} = O\Big(\epsilon^2\Big(\log \frac{M_\delta}{\epsilon}\Big)^2\Big),$$

where the constant M_δ should be appropriately chosen. This further implies that, there exists a constant L, such that

$$\Big\{f : \rho(f, f_0) \le \frac{L\epsilon}{\log \frac{M_\delta}{\epsilon}}\Big\} \subset \Big\{f : K(f_0, f) \le \epsilon^2, \mathbb{E}_{f_0}\Big(\Big(\log \frac{f_0(Y)}{f(Y)}\Big)^2\Big) \le \epsilon^2\Big\}. \tag{26}$$

The result allows us to work with a Hellinger ball instead of a Kullback-Leibler one. The transition is necessary because it is more straightforward to apply a property of the Dirichlet distribution to estimate the probability mass of a Hellinger ball around the true density function. In the lemma below, this specific property of the Dirichlet distribution is stated in terms of L_1-distance, which is equivalent to the Hellinger distance. We would like to point out that this lemma is a variation of Lemma 6.1 in the paper by Ghosal et al. (2000) and the proof is adapted from their paper.

Lemma 24 *$(X_1, \cdots, X_t)$ is distributed according to the Dirichlet distribution. Let $(x_{10}, \cdots, x_{t0})$ be any point on the t-simplex. Take $\epsilon < 1/t$. With $\tau < \epsilon^2$, we have*

$$P\Big(\sum_{j=1}^t |X_j - x_{j0}| \le 2\epsilon, X_j \ge \tau \text{ for all } j\Big) \ge \frac{\Gamma(\alpha t)}{(\Gamma(\alpha))^t}(\epsilon^2 - \tau)^t. \tag{27}$$

Proof We can find an index j such that $x_{j0} > 1/t$. By relabeling, we can assume that $j = t$. if $|x_j - x_{j0}| \leq \epsilon^2$ for $j = 1, \cdots, t-1$, then

$$\sum_{j=1}^{t-1} x_j \leq 1 - x_{t0} + (t-1)\epsilon^2 \leq (t-1)(\epsilon^2 + 1/t) \leq 1 - \epsilon^2 < 1.$$

Therefore, there exists $x = (x_1, \cdots, x_t)$ in the simplex with these first $t-1$ coordinates. And

$$\sum_{j=1}^{t} |x_j - x_{j0}| \leq 2\sum_{j=1}^{t-1} |x_j - x_{j0}| \leq 2\epsilon^2(t-1) \leq 2\epsilon.$$

Therefore, the probability on the left hand side of (27) is bounded below by

$$P(|X_j - x_{j0}| \leq \epsilon^2, X_j \geq \tau, j = 1, \cdots, t-1) \geq \frac{\Gamma(\alpha t)}{(\Gamma(\alpha))^t} \prod_{j=1}^{t-1} \int_{\max((x_{j0}-\epsilon^2),\tau)}^{\min((x_{j0}+\epsilon^2),1)} x_j^{\alpha-1} dx_j.$$

Since $\alpha < 1$, we can lower bound the integrand by 1 and the interval of integration contains at least an interval of length $\epsilon^2 - \tau$. Therefore, the result above can be further lower bounded by

$$\frac{\Gamma(\alpha t)}{(\Gamma(\alpha))^t}(\epsilon^2 - \tau)^{t-1} \geq \frac{\Gamma(\alpha t)}{(\Gamma(\alpha))^t}(\epsilon^2 - \tau)^t.$$

This finishes the proof. ■

Now, we are ready to derive lower bounds for the prior probability mass on Θ_t's when t varies within a certain range. Before stating the result, we want to briefly review the assumptions we made in Section 2 and Section 3. First, in terms of approximation error, we assume that for any $f_0 \in \mathcal{F}_0$, there exists a sequence of $f_t \in \Theta_t$, such that $A_1 t^{-r} \leq \min_{g \in \Theta_t} \rho(g, f_0) \leq \rho(f_t, f_0) \leq A_2 t^{-r}$ for some positive constants A_1 and A_2 (If the lower bound does not hold, we can always obtain a faster concentration rate). Second, we impose a moment condition on $\mathcal{F}_0$. For any $f_0 \in \mathcal{F}_0$, we assume that $\int f_0^2 < \infty$. Under these two assumptions, we provide the lower bound in the lemma below.

Lemma 25 *Assume that $f_0 \in \mathcal{F}_0$. Π is the prior probability specified in 2. Let $d_{n,t} = \epsilon_{n,t}^2 = \frac{t \log t}{n/\log n}$. Take $t = n^{\frac{1}{2r+1}}$, we have*

$$\mathbb{P}_0^n\Big(\int_{\Theta_t} \prod_{i=1}^{n} \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f) \\ \leq \frac{1}{2}\Pi(\Theta_t) \exp(-2nd_{n,t} - c^* t \log t - 4\omega t \log n - t \log \Gamma(\alpha)) \Big) \leq \frac{2}{nd_{n,t}},$$

where $\omega = \max(1, 1/2r)$, and c^* is the constant introduced in Section 2.3.

Proof Let $S_{n,t} = \{f \in \Theta_t : K(f_0, f) \leq d_{n,t}, V(f_0, f) \leq d_{n,t}\}$. By applying lemma 23, we have the bound

$$\mathbb{P}_0^n\Big(\int_{\Theta_t} \prod_{i=1}^{n} \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f) \leq \frac{1}{2}\Pi(S_{n,t}) e^{-2nd_{n,t}} \Big) \leq \frac{2}{nd_{n,t}}. \tag{28}$$

To prove the lemma, it suffices to provide a lower bound for $\Pi(S_{n,t})$. The way to approach this is to find a subset of $S_{n,t}$ for which Lemma 24 can be applied. Our argument is as the following.

Define $\tilde{S}_{n,t} = \{f \in \Theta_t : K(f_0, f) \leq d_{n,t}, \mathbb{E}_{f_0}\big((\log \frac{f_0(Y)}{f(Y)})^2\big) \leq d_{n,t}, f \geq \tau\}$, where τ is a truncation parameter. Note that $\mathbb{E}_{f_0}\big((\log \frac{f_0(Y)}{f(Y)})^2\big) \geq V(f_0, f)$, we have $\tilde{S}_{n,t} \subset S_{n,t}$. From (26), we know that

$$W_{n,t} := \{f \in \Theta_t : \rho(f_0, f) \leq \frac{L\epsilon_{n,t}}{\log \frac{M_\delta}{\epsilon_{n,t}}}, f \geq \tau\} \subset \tilde{S}_{n,t}.$$

Set τ to be $Dt^{-\eta}$ with $\eta > \max\{2, 4r\}$, then $M_\delta = O(t^{\delta\eta} \int f_0^{(1+\delta)})$. Furthermore,

$$\begin{aligned} \frac{\epsilon_{n,t}}{\log \frac{M_\delta}{\epsilon_{n,t}}} &= O\left(\left(\frac{t\log t}{n/\log n}\right)^{1/2} \Big/ \log\left(t^{\delta\eta}\int f_0^{(1+\delta)}(\frac{n/\log n}{t\log t})^{1/2}\right)\right) \\ &= O\left(\left(\frac{t\log t}{n\log n}\right)^{1/2}\right). \end{aligned}$$

Under the assumptions that $t = n^{\frac{1}{1+2r}}$, there exists $f_t \in \Theta_t$, such that $\rho(f_0, f_t) < \frac{L\epsilon_{n,t}}{\log \frac{M_\delta}{\epsilon_{n,t}}}$. If we define

$$\tilde{W}_{n,t} := \Big\{f \in \Theta_t : \rho(f, f_t) \leq \frac{L\epsilon_{n,t}}{\log \frac{M_\delta}{\epsilon_{n,t}}} - \rho(f_0, f_t), f \geq \tau\Big\},$$

by triangle inequality, we know that $\tilde{W}_{n,t} \subset W_{n,t}$. Together with the previous result, we claim that there exists a constant L', such that

$$\tilde{B}_{n,t} := \Big\{f \in \Theta_t : \rho(f, f_t) \leq L'\left(\frac{t\log t}{n\log n}\right)^{1/2}, f \geq \tau\Big\} \subset \tilde{W}_{n,t}.$$

Next, based on the fact $\rho^2(f, g) \leq \|f - g\|_{L^1(\Omega)}$, we have

$$B_{n,t} := \Big\{f \in \Theta_t : \|f_t - f\|_{L^1(\Omega)} \leq \frac{L'^2 t\log t}{n\log n}, f \geq \tau\Big\} \subset \tilde{B}_{n,t}.$$

Note that $\Pi(B_{n,t}) = \Pi(\Theta_t)\Pi(B_{n,t}|\Theta_t)$. Assume that f_t is supported by the binary partition $\{\Omega_{j0}\}_{j=1}^t$. Let $F_0 = \{f \in \Theta_t : f = \sum_{j=1}^t \frac{\theta_j}{|\Omega_{j0}|}\mathbb{1}_{\Omega_{j0}}, \theta_j \geq 0, \sum_{j=1}^t \theta_j = 1\}$ be the collection of all the density functions in Θ_t which are supported by the same binary partition as f_t. Then

$$\Pi(B_{n,t}|\Theta_t) \geq \Pi(B_{n,t}|F_0)\Pi(F_0|\Theta_t) \geq \exp(-c^* t\log t)\Pi(B_{n,t}|F_0). \tag{29}$$

Now we apply Lemma 24 to bound $\Pi(B_{n,t}|F_0)$ from below. We work with an L_1-ball with radius $(\frac{L'^2 t\log t}{n\log n})^\omega$, where ω is chosen to be $\max(1, 1/2r)$. We can always assume that $L' < 1$, otherwise we can work with a ball shringking to zero at a faster rate instead. Obviously, this ball is contained in $B_{n,t}$. When $t = n^{\frac{1}{2r+1}}$, we have $(\frac{L'^2 t\log t}{n\log n})^\omega < \frac{1}{t}$. Under the assumptions

$\eta > \max(2, 4r)$, we know that when $t/n^{\gamma_1} = o(1)$ with $\gamma_1 = \frac{2\omega}{2\omega+\eta}$, $Dt^{-\eta} = o((\frac{t\log t}{n\log n})^{2\omega})$. By setting x_{j0} in the lemma to be the probability mass on Ω_{j0} under f_t, we have

$$\begin{aligned}\Pi(B_{n,t}|F_0) &\geq \frac{\Gamma(\alpha t)}{(\Gamma(\alpha))^t}((\frac{L'^2 t\log t}{2n\log n})^{2\omega} - Dt^{-\eta})^t \\ &\geq \exp(-t\log\Gamma(\alpha) - 4\omega t\log n). \qquad (30)\end{aligned}$$

Combining (28), (29) and (30) together, we get the desired result. ■

7.3.2 Proof of Theorem 4

In this section, we calculate the posterior concentration rate based on Lemma 19 in Section 7.1.3, Lemma 17 in Section 7.1.2, and the lower bound derived in Section 7.3.1.

Proof [Proof of Theorem 4] Let $\epsilon_n = n^{-\frac{r}{2r+1}}(\log n)^{2+\frac{1}{2r}}$ and $\eta_{n,t} = \left(\frac{t(\log t)^{1/r+1}}{n/\log n}\right)^{1/2}$. First, we divide the items in numerator of (8) into three blocks. We define

$$\begin{aligned}A_{\text{Num}} &= \sum_{t=1}^{N_1-1}\int_{\{f:\rho(f,f_0)\geq M\epsilon_n\}\cap\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f),\\ B_{\text{Num}} &= \sum_{t=N_1}^{N_2}\int_{\{f:\rho(f,f_0)\geq M\epsilon_n\}\cap\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f),\\ C_{\text{Num}} &= \sum_{t=N_2+1}^{n/\log n}\int_{\{f:\rho(f,f_0)\geq M\epsilon_n\}\cap\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f),\end{aligned}$$

where $N_1 = D_1 n^{\frac{1}{2r+1}}(\log n)^{-\frac{1}{r}}$ and $N_2 = D_2 n^{\frac{1}{2r+1}}(\log n)^2$.

We deal with each block in the numerator separately. Roughly speaking, when t is small, the approximation error to f_0 dominates, and these items can be bounded by the Hellinger distance between f and f_0. The items in the middle range can be bounded by controlling the metric entropy of Θ_t. The items in the last block are negligible because the prior probability decays to zero fast.

An upper bound for A_{Num}. We assume that there exists a sequence of $f_t \in \Theta_t$, such that $A_1 t^{-r} \leq \min_{g\in\Theta_t}\rho(g,f) \leq \rho(f_t,f) \leq A_2 t^{-r}$ for some positive constants A_1 and A_2. Let $N_3 = D_3 n^{\frac{1}{2r+1}}(\log n)^{-\frac{2}{r}-\frac{1}{2r^2}}$. With an appropriately chosen D_3, when $t < N_3$, $A_1 t^{-r}$ is greater than $M\epsilon_n$. Therefore,

$$\sum_{t=1}^{N_3-1}\int_{\{f:\rho(f,f_0)\geq M\epsilon_n\}\cap\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f) = \sum_{t=1}^{N_3-1}\int_{\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f). \qquad (31)$$

When $N_3 \leq t < N_1$, given that $A_1 t^{-r} < M\epsilon_n$, we have

$$\sum_{t=N_3}^{N_1-1}\int_{\{f:\rho(f,f_0)\geq M\epsilon_n\}\cap\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f) \leq \sum_{t=N_3}^{N_1-1}\int_{\{f:\rho(f,f_0)\geq A_1 t^{-r}\}\cap\Theta_t}\prod_{i=1}^n\frac{f(Y_i)}{f_0(Y_i)}d\Pi(f). \qquad (32)$$

Combining (31) and (32) together and applying Lemma 19 by setting $\delta_{n,t}$ to be $A_1 t^{-r}$, we obtain

$$\begin{aligned} A_{\text{Num}} &\leq \sum_{t=1}^{N_1-1} \int_{\{f:\rho(f,f_0)\geq A_1 t^{-r}\}\cap\Theta_t} \prod_{i=1}^{n} \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f) \\ &\leq \sum_{t=1}^{N_1-1} \Pi(\Theta_t) \exp(-A_1 n t^{-2r}) \\ &\leq \left(\sum_{t=1}^{N_1-1} \exp(-2A_1 n t^{-2r}) \right)^{1/2}, \quad \text{with probability tending to 1 under } \mathbb{P}_0^n. \end{aligned}$$

The last line is based on the Cauchy-Schwarz inequality. Now, we will estimate the order of the summation in the last line. In order to simplify the notation, we will discuss the order of $\sum_{t=1}^{N_1-1} \exp(-\frac{2A_1 n}{t^{2r}})$ in detail.

We know that the mass is centered around $t = N_1 - 1$. Power series expansion around that point gives

$$\sum_{t=1}^{(1-\epsilon)N_1} \leq (1-\epsilon) N_1 \exp\left(-\frac{2A_1 n}{((1-\epsilon)N_1)^{2r}} \right),$$

which is a lower order term compared to the last term in the summation, thus it does not contribute significantly to the summation. Let $1 - \delta = \frac{t}{N_1}$, and expand

$$(1-\delta)^{-2r} = 1 + 2r\delta + \binom{-2r}{2}\delta^2 + o(\delta^2).$$

Then,

$$\begin{aligned} \sum_{t=(1-\epsilon)N_1}^{N_1-1} \exp(-\frac{2A_1 n}{t^{2r}}) &\leq \int_{(1-\epsilon)N_1}^{N_1} \exp(-\frac{2A_1 n}{x^{2r}}) dx \\ &\asymp \int_0^{\epsilon} \exp\left(-2\frac{A_1}{D_1^{2r}} n^{\frac{1}{2r+1}} (\log n)^2 (1-\delta)^{-2r} \right) N_1 d\delta \\ &\asymp \int_0^{\epsilon} \exp\left(-2\frac{A_1}{D_1^{2r}} n^{\frac{1}{2r+1}} (\log n)^2 (1 + 2r\delta + o(\delta)) \right) N_1 d\delta \\ &\asymp \frac{1}{(\log n)^{1/r+2}} \exp\left(-2\frac{A_1}{D_1^{2r}} n^{\frac{1}{2r+1}} (\log n)^2 \right). \end{aligned}$$

Therefore, with probability tending to 1 under $\mathbb{P}_0^n$,

$$A_{\text{Num}} \leq (\log n)^{-1-\frac{1}{2r}} \exp(-\frac{A_1}{D_1^{2r}} n^{\frac{1}{2r+1}} (\log n)^2).$$

An upper bound for $B_{\mathbf{Num}}$. From Lemma 19 and Remark 20, we know that if the result holds for $\delta_{n,t}$, then it also applies to $M\eta_{n,t} > \delta_{n,t}$. When $N_1 \le t \le N_2$,

$$\begin{aligned} B_{\mathrm{Num}} &\le \sum_{t=N_1}^{N_2} \int_{\{f:\rho(f,f_0)\ge M\eta_{n,t}\}\cap\Theta_t} \prod_{i=1}^{n} \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f) \\ &\le \sum_{t=N_1}^{N_2} \exp(-\lambda t \log t) \exp(-M^2 t (\log t)^{1+\frac{1}{r}} \log n) \\ &\le \left(\sum_{t=N_1}^{N_2} \exp(-2\lambda t \log t) \right)^{1/2} \left(\sum_{t=N_1}^{N_2} \exp\left(-2M^2 t (\log t)^{1+\frac{1}{r}} \log n\right) \right)^{1/2} \\ &\asymp \exp\left(-M^2 n^{\frac{1}{2r+1}} (\log n)^2\right), \quad \text{with probability tending to 1 under } \mathbb{P}_0^n, \end{aligned}$$

where the last line is obtained by integration by part.

An upper bound for $C_{\mathbf{Num}}$. For the last block C_{Num}, we have

$$C_{\mathrm{Num}} \le \sum_{t=N_2+1}^{n/\log n} \int_{\Theta_t} \prod_{i=1}^{n} \frac{f(Y_i)}{f_0(Y_i)} d\Pi(f).$$

Next, we want to apply Lemma 17 to show that there exist constants $\eta, c > 0$, such that

$$\mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t} \prod_{i=1}^{n} \frac{f(Y_i)}{f_0(Y_i)} \ge \eta t \log t \right) \le \exp\left(-ct/\log t\right).$$

Recall that $\widetilde{Z}_f$ is the truncated log-likelihood ratio, and τ is the truncation parameter. When τ is large, based on Lemma 16, we know that

$$-\mathbb{E}\widetilde{Z}_f \ge (1-\gamma)\|f^{1/2} - f_0^{1/2}\|_2^2 \ge 0,$$

where $\gamma \in (0,1)$ is the constant defined in the statement of Lemma 16. The bound is uniform for all densities f under consideration, therefore

$$\begin{aligned} \mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t} \sum_{i=1}^{n} Z_f(Y_i) \ge \eta t \log t \right) &\le \mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t} \sum_{i=1}^{n} \widetilde{Z}_f(Y_i) \ge \eta t \log t \right) \\ &\le \mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t} \sum_{i=1}^{n} \left(\widetilde{Z}_f(Y_i) - \mathbb{E}\widetilde{Z}_f(Y_i)\right) \ge \eta t \log t \right) \\ &= \mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t} \nu_n(\widetilde{Z}_f) \ge n^{1/2} \eta t \log t / n \right). \end{aligned}$$

We will show the last probability is exponentially small by applying Lemma 17.

Recall that in Lemma 19, we define $\delta_{n,t} = (\frac{t \log t}{n/\log n})^{1/2}$. Then

$$\begin{aligned}
&\mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t} \nu_n(\widetilde{Z}_f) \geq n^{1/2}\eta t \log t/n\right) \\
&\quad \leq \mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t, \rho(f,f_0)\geq\delta_{n,t}} \nu_n(\widetilde{Z}_f) \geq n^{1/2}\eta t \log t/n\right) \\
&\qquad + \mathbb{P}_{f_0}\left(\sup_{f\in\Theta_t, \rho(f,f_0)\leq\delta_{n,t}} \nu_n(\widetilde{Z}_f) \geq n^{1/2}\eta t \log t/n\right) \\
&\quad = P_1 + P_2.
\end{aligned}$$

For P_1, based on Lemma 19, we have

$$P_1 \leq \exp(-c_2 t(\log t)^2).$$

Next, we apply Lemma 17 to bound P_2. We may set d in Lemma 17 to be $\delta_{n,t}$ here. L in Lemma 17 is $n^{1/2}(\eta t \log t/n)$ here. In the first place, (22) is satisfied because

$$L = n^{1/2}(\eta t \log t/n) = o(bn^{1/2}d^2/4), \quad \text{for } N_2 < t \leq n/\log n.$$

By Lemma 15,

$$\begin{aligned}
&\int_{bL/(32n^{1/2})}^{d} \left(H^{[]}(u/(2\exp\tau/2), \Theta_t, \rho)\right)^{1/2} du \\
&\quad \leq \int_{bL/(32n^{1/2})}^{d} (t\log(4p\exp(\tau/2)) + 2(t+1)\log(t+1) - t\log u)^{1/2} du. \qquad (33)
\end{aligned}$$

The order of the right hand side of (33) can be estimated in a similar way as that used in the proof of Lemma 19. Specifically, we have the following result

$$\int_{bL/(32n^{1/2})}^{d} \left(H^{[]}(u/(2\exp\tau/2), \Theta_t, \rho)\right)^{1/2} du \sim t^{1/2} d\sqrt{\log(t^2/d)}.$$

Therefore, (23) is also satisfied when $N_2 < t \leq n/\log n$. Note that

$$\begin{aligned}
\varphi(L, d^2, n) &= \frac{n(\eta t \log t/n)^2}{8(8c_0\delta_{n,t}^2 + \eta t \log t/n)} \\
&\geq \frac{n(\eta t \log t/n)^2}{8(8c_0\delta_{n,t}^2 + \delta_{n,t}^2)} \\
&\geq \frac{n\eta^2(t\log t/n)^2}{(2^6c_0+8)(\frac{t\log t}{n/\log n})^2} \\
&\geq \frac{\eta^2}{2^6c_0+8} t/\log t.
\end{aligned}$$

Applying Lemma 17, we have

$$P_2 \leq 3\exp\left(-\frac{\eta^2(1-b)}{2^6c_0+8} t/\log t\right).$$

The result implies that, with probability tending to 1,

$$C_{\text{Num}} \lesssim \sum_{t=N_2}^{n/\log n} 4\exp(-\lambda t\log t)\cdot\exp(\eta t\log t).$$

When $\lambda \geq \eta + 1$,

$$C_{\text{Num}} \lesssim \exp(-D_2 n^{\frac{1}{2r+1}}(\log n)^2).$$

Posterior contraction rate. Combining the bounds for $A_{\text{Num}}, B_{\text{Num}}$, and C_{Num} together, we have that with probability tending to 1,

$$\begin{aligned}(8) \lesssim & \frac{(\log n)^{-1-\frac{1}{2r}}\exp(-(A_1/D_1^{2r})n^{\frac{1}{2r+1}}(\log n)^2) + \exp(-M^2 n^{\frac{1}{2r+1}}(\log n)^2 + \exp(-D_2 n^{\frac{1}{2r+1}}(\log n)^2))}{\sum_{t=1}^{\infty}\int_{\Theta_t}\prod_{j=1}^{n}\frac{f(Y_j)}{f_0(Y_j)}d\Pi(f)} \\ \leq & \frac{(\log n)^{-1-\frac{1}{2r}}\exp(-(A_1/D_1^{2r})n^{\frac{1}{2r+1}}(\log n)^2) + \exp(-M^2 n^{\frac{1}{2r+1}}(\log n)^2) + \exp(-D_2 n^{\frac{1}{2r+1}}(\log n)^2)}{\frac{1}{2}\exp\left(-\frac{2}{2r+1}n^{\frac{1}{2r+1}}(\log n)^2 - (\frac{c^*}{2r+1}+4\omega)n^{\frac{1}{2r+1}}\log n - n^{\frac{1}{2r+1}}(\log\Gamma(\alpha)+1)\right)},\end{aligned}$$

where the last inequality is obtained by applying Lemma 25 to the space Θ_t with $t = n^{\frac{1}{2r+1}}$. The last line goes to zero when A_1/D_1^{2r}, M^2 and D_2 are all larger than $\frac{2}{2r+1}$. Therefore, we have

$$\Pi\left(f : \rho(f, f_0) \geq M\epsilon_n | Y_1, \cdots, Y_n\right) \leq \exp\left(-bn^{\frac{1}{2r+1}}(\log n)^2\right),$$

with probability tending to 1, where b is a positive constant. This concludes the proof. ■

7.4 Proof of Theorem 12

In this subsection, we provide the proof for Theorem 12. We first outline the proof and decompose the error into bias and variance. Then we provide bound for each component respectively.

7.4.1 Outline of the Proof

Proof [Proof of Theorem 12] As discussed before, if we check binary partitions of density functions in $\Theta_t^{(m)}$'s, their shapes are similar to each other in the sense that (15) and (16) hold with $\mathcal{E}_t^{\min}$ and $\mathcal{E}_t^{\max}$ defined therein. Given these two quantities, we can find an equally spaced and balanced partition with all subregions in the partition being a cube and edge length exactly equal to $\mathcal{E}_t^{\min}$ or $\mathcal{E}_t^{\max}$, denoted as $\mathcal{A}_{t,\min}$ and $\mathcal{A}_{t,\max}$ respectively. Note that the size of these two partitions may not be exactly t. If we use $\gamma_1(t)$ and $\gamma_2(t)$ to denote corresponding sizes, then for all t large enough, $(c_2^{\text{space}})^{-p} \leq \gamma_2(t)/t \leq \gamma_1(t)/t \leq (c_1^{\text{space}})^{-p}$. With respect to densities supported by the partition $\mathcal{A}_{t,\max}$ and its shifts, we have the following bias-variance decomposition:

$$\rho(\hat{f}_{n,t}^{\text{forest}}, f_0) \leq \underbrace{\rho(\hat{f}_{n,t}^{\text{forest}}, f_{t,\max}^{\text{forest}})}_{\text{estimation error or variance}} + \underbrace{\rho(f_{t,\max}^{\text{forest}}, f_0)}_{\text{bias}}.$$

$f^{\text{forest}}_{t,\max}$ is defined to be a density function independent of data taking the form $\frac{1}{M}\sum_{m=1}^{M} f^{(m)}_{t,\max}$, where $f^{(m)}_{t,\max}$ is supported by $\mathcal{A}^{h_m}_{t,\max}$, and

$$f^{(m)}_{t,\max} = \sum_{j=1}^{\gamma_2(t)} \frac{\int_{\Omega_j^{h_m}} f_0}{\mu(\Omega_j^{h_m} \cap \Omega)} \cdot \mathbb{1}_{\Omega_j^{h_m} \cap \Omega}.$$

The boundary condition H1 guarantees that for t large enough and shift sizes bounded by $(\mathcal{E}^{\max}_t)^2$, each $f^{(m)}_{t,\max}$ is an appropriately defined density and so is $f^{\text{forest}}_{t,\max}$.

Bias. For the bias part, the small perturbations in combination with the ensemble method in some sense is equivalent to partitioning the sample spaces into a much finer grid. Specifically, according to Lemma 26,

$$\|f^{\text{forest}}_{t,\max} - f_0\|_\infty \le A t^{-(1+\beta)/p},$$

where $A > 0$ is a constant specified in the lemma.

Variance. To bound the estimation error, based on properties of the Hellinger distance,

$$\rho^2\left(\hat{f}^{\text{forest}}_{n,t}, f^{\text{forest}}_{t,\max}\right) \le \frac{1}{M}\sum_{m=1}^{M} \rho^2\left(\hat{f}^{(m)}_{n,t}, f^{(m)}_{t,\max}\right) \le \frac{1}{M}\sum_{m=1}^{M} K\left(\hat{f}^{(m)}_{n,t}, f^{(m)}_{t,\max}\right).$$

Given a partition $\mathcal{A}_t$, with or without shift, we define $f_{\mathcal{A}_t}$ as

$$f_{\mathcal{A}_t} = \sum_{j=1}^{t} \frac{\theta_j}{\mu(\Omega_j \cap \Omega)} \cdot \mathbb{1}_{\Omega_j \cap \Omega}, \quad \text{with } \theta_j = \int_{\Omega_j} f_0.$$

Under the boundary condition, $f_{\mathcal{A}_t}$ is also an appropriately defined density. Within each $\Theta^{(m)}_t$, we denote the partition of the MLE $\hat{f}^{(m)}_{n,t}$ as $\mathcal{A}^{h_m}_t$. To simplify the notation, we view the partition as one without the shift, denoted as $\mathcal{A}_t$, as the following argument can be easily verified under a shift. Then $\mathcal{A}_t$ is always a finer partition compared to $\mathcal{A}_{t,\max}$. This is to say, to obtain $\mathcal{A}_t$, we can start from the partition $\mathcal{A}_{t,\max} = \{\Omega_j\}_{j=1}^{\gamma_2(t)}$, and then split each Ω_j into k_j subregions $\{\Omega_{j,l}\}_{l=1}^{k_j}$. Let $n_{j,l}$ be the number of observation in $\Omega_{j,l}$

$$K\left(\hat{f}^{(m)}_{n,t}, f^{(m)}_{t,\max}\right) = \int \hat{f}^{(m)}_{n,t} \log \frac{\hat{f}^{(m)}_{n,t}}{f^{(m)}_{t,\max}} = \sum_{j=1}^{\gamma_2(t)}\sum_{l=1}^{k_j} \frac{n_{j,l}}{n} \log \frac{(n_{j,l}/n)/\mu(\Omega_{j,l}\cap\Omega)}{(\int_{\Omega_j\cap\Omega} f_0)/\mu(\Omega_j\cap\Omega)}.$$

As $\int_{\Omega_j\cap\Omega} f_0 \ge \int_{\Omega_{j,l}\cap\Omega} f_0$ and $\mu(\Omega_j\cap\Omega)/\mu(\Omega_{j,l}\cap\Omega) \le (\mathcal{E}^{\max}_t/\mathcal{E}^{\min}_t)^p$, the quantity above can be further bounded by

$$\begin{aligned} K\left(\hat{f}^{(m)}_{n,t}, f^{(m)}_{t,\max}\right) &\le p\log(\mathcal{E}^{\max}_t/\mathcal{E}^{\min}_t)\sum_{j=1}^{\gamma_2(t)}\sum_{l=1}^{k_j}\frac{n_{j,l}}{n}\log\frac{n_{j,l}/n}{\int_{\Omega_{j,l}\cap\Omega} f_0} \\ &= p\log(\mathcal{E}^{\max}_t/\mathcal{E}^{\min}_t) K\left(\hat{f}^{(m)}_{n,t}, f_{\mathcal{A}^{h_m}_t}\right) \\ &\le p\log(\mathcal{E}^{\max}_t/\mathcal{E}^{\min}_t)\cdot\frac{1}{n}\sup_{f\in\mathcal{F}_{\mathcal{A}^{h_m}_t}}\sum_{i=1}^{n}\log\left(\frac{f(Y_i)}{f_{\mathcal{A}_t}(Y_i)}\right). \end{aligned}$$

While for each single tree, by Theorem 27 we have

$$\mathbb{P}_0^n\left(K\left(\hat{f}_{n,t}^{(m)}, f_{\mathcal{A}_t^{h_m}}\right) \geq p\log(\mathcal{E}_t^{\max}/\mathcal{E}_t^{\min})\eta^2 t\log t/(n/\log n)\right) \\ \leq 8\exp(-c_2\eta^2 t(\log t)(\log n)) \quad \text{for some } \eta > 1, c_2 > 0.$$

Note that the same bound applies to all partitions $\mathcal{A}_t^{h_m}$ satisfying condition 14 and the corresponding MLE $\hat{f}_{n,t}^{(m)}$. Thus for the density forest,

$$\mathbb{P}_0^n\left(\frac{1}{M}\sum_{m=1}^{M} K\left(\hat{f}_{n,t}^{(m)}, f_{t,\max}^{(m)}\right) \geq \eta^2 p\log(\mathcal{E}_t^{\max}/\mathcal{E}_t^{\min})t(\log t)/(n/\log n)\right) \\ \leq 8M\exp(-c_2\eta^2 t(\log t)(\log n)) \quad \text{for some } \eta > 1, c_2 > 0,$$

where M is bounded by a linear function of t. To simplify the notation, we define $(\eta')^2 = \eta^2 p\log(\mathcal{E}_t^{\max}/\mathcal{E}_t^{\min})$.

Convergence rates. Combining the analysis for bias and variance, we have

$$\rho(\hat{f}_{n,t}^{\text{forest}}, f_0) \leq At^{-(1+\beta)/p} + \eta'(t(\log t)/(n/\log n))^{1/2} \quad \text{with probability tending to 1 under } \mathbb{P}_0^n.$$

The "optimal" rate is obtained by making a trade-off between the bias and the variance. If we set $\eta = 2$ and

$$t = \left(((c_2^{\text{space}})^2 pL^2/\log(\mathcal{E}_t^{\max}/\mathcal{E}_t^{\min}))\frac{n}{\log n}\right)^{\frac{p}{2(1+\beta)+p}},$$

the corresponding convergence rate is $n^{-\frac{1+\beta}{2(1+\beta)+p}}(\log n)^{\frac{1}{2}+\frac{1+\beta}{2(1+\beta)+p}}$. ■

7.4.2 Approximation Error

Lemma 26 *Assume $f_0 \in \mathcal{H}^{1,\beta}(\Omega)$the boundary condition H1 satisfied. For each t, the parameter spaces under shifts $\{\Theta_t^{(m)}\}_{1\leq m\leq M}$ are defined as that in Section 5.1.3. Then there exists a density f_t^{forest} of the form*

$$f_t^{forest} = \frac{1}{M}\sum_{m=1}^{M} f_t^{(m)}, \quad f_t^{(m)} \in \Theta_t^{(m)},$$

such that

$$\|f_0 - f_t^{forest}\|_\infty \leq 2c_2^{space} pLt^{-(1+\beta)/p},$$

where c_2^{space} is the constant in condition (16), and L is the constant in (12) and (13) for the Hölder space.

Proof For the ensemble estimate, it is sufficient to study the approximation error of densities supported by $\mathcal{A}_{t,\max}$ and its shifts, as the size of $\mathcal{A}_{t,\max}$ is no larger than t. Assume

$\mathcal{A}_{t,\max} = \{\Omega_j\}_{j=1}^{\gamma_2(t)}$. Note that all Ω_j's in $\mathcal{A}_{t,\max}$ are p-dimensional cubes of equal sizes with edge length $\mathcal{E}_t^{\max}$. To simplify the notation, in the proof, we use ϵ to denote it. Recall that along each dimension, the smallest nonzero value of h_m^l is $\delta = \epsilon^2$. The approximation within $\Theta_t^{(m)}$ is set to be

$$f_t^{(m)}(x) = \sum_{j=1}^{\gamma_2(t)} \frac{1}{\mu(\Omega_j^{h_m} \cap \Omega)} \left(\int_{\Omega_j^{h_m}} f_0(x')dx' \right) \mathbb{1}_{\Omega_j^{h_m} \cap \Omega}(x),$$

where $h_m = (h_m^1, \ldots, h_m^p)^\top$ is the vector representing the shift, and $f_t^{\text{forest}} = (\sum_{m=1}^M f_t^{(m)})/M$. We would like to show f_t^{forest} can achieve the desired approximation rate.

We divide Ω into an interior region and a boundary region as

$$\Omega = A_\epsilon \cup B_\epsilon,$$

where $A_\epsilon = \{x \in \Omega : \min_{1\le l\le p} x^l > \epsilon \text{ and } \min_{1\le l\le p}(1-x^l) > \epsilon\}$ and $B_\epsilon = \Omega \setminus A_\epsilon$.

For any $x_0 \in A_\epsilon$,

$$f_0(x_0) - f_t^{\text{forest}}(x_0) = f(x_0) - \frac{1}{M}\sum_{m=1}^M \sum_{j:x_0\in\Omega_j^{h_m}} \frac{1}{\mu(\Omega_j^{h_m})} \left(\int_{\Omega_j^{h_m}} f_0(x)dx \right), \quad 0 \le h_m^l < \epsilon.$$

Assume $x_0 \in \Omega_{j_0}$ for some j_0. If we allow h_m^l to take negative values, then the approximation at a point x_0 can be rewritten as

$$f_t^{\text{forest}}(x_0) = \frac{1}{M} \sum_{m:x_0\in\Omega_{j_0}^{h_m}} \frac{1}{\mu(\Omega_{j_0}^{h_m})} \left(\int_{\Omega_{j_0}^{h_m}} f_0(x)dx \right), \quad -\epsilon < h_m^l < \epsilon. \tag{34}$$

Let $\Omega_{j_0} = \otimes_{l=1}^p [b_{L,j_0}^l, b_{U,j_0}^l)$. The condition $x_0 \in \Omega_{j_0}^{h_m}$ is satisfied when $b_{L,j_0}^l + h_m^l \le x_0^l < b_{U,j_0}^l + h_m^l$ with $|h_m| < \epsilon$. Although for the definition of density forests we require $h_m^l \in [0,\epsilon)$, here an h_m with negative entries can be understood as the partition is shifted along direction $\tilde{h}_m$, where $\tilde{h}_m^l = h_m^l$ for $h_m^l \ge 0$ and $\tilde{h}_m^l = \epsilon + h_m^l$ for $h_m^l < 0$. After the shift, we can find an $\Omega_{j'}$ such that $x_0 \in \Omega_{j'}^{\tilde{h}_m}$, and the set $\Omega_{j'}^{\tilde{h}_m}$ can be equivalently viewed as $\Omega_{j_0}^{h_m}$ for an interior point.

As f_0 lies in the Hölder space $\mathcal{H}^{1,\beta}(\Omega)$ $(0 < \beta \le 1)$, we can find the following expansion based on the mean value theorem, for any x, x_0 lying in the interior region of Ω, :

$$f_0(x) - f_0(x_0) = \nabla f_0\left(x_0 + t(x-x_0)\right)^\top (x - x_0) \quad \text{for some } t \in (0,1).$$

Then

$$\begin{aligned}
&\left| f_0(x) - f_0(x_0) - \nabla f_0(x_0)^\top (x - x_0) \right| \\
&\qquad = \left| \left(\nabla f_0\left(x_0 + t(x-x_0)\right) - \nabla f_0(x_0)\right)^\top (x - x_0) \right| \\
&\qquad \le \left\| \nabla f_0\left(x_0 + t(x-x_0)\right) - \nabla f_0(x_0) \right\|_2 \cdot \|x - x_0\|_2 \\
&\qquad \le \left(\sum_{l=1}^p L^2 \|x - x_0\|_2^{2\beta} \right)^{1/2} \cdot \|x - x_0\|_2 \\
&\qquad \le Lp\epsilon^{(1+\beta)}.
\end{aligned}$$

Therefore, for $x_0 \in A_\epsilon$, by taking integral over $\Omega_{j_0}^{h_m}$'s in (34), we have

$$\begin{aligned}
&|f_t^{\text{forest}}(x_0) - f_0(x_0)| \\
&\quad \le \left| \frac{1}{M} \sum_{m: x_0 \in \Omega_{j_0}^{h_m}} \frac{1}{\mu(\Omega_{j_0})} \left(\int_{\Omega_{j_0}^{h_m}} \nabla f_0(x_0)^\top (x - x_0) dx \right) \right| + Lp\epsilon^{(1+\beta)} \\
&\quad = \left| \nabla f_0(x_0)^\top \left(\frac{1}{M} \sum_{m: x_0 \in \Omega_{j_0}^{h_m}} \frac{1}{\mu(\Omega_{j_0})} \int_{\Omega_{j_0}} (x + h_m - x_0) dx \right) \right| + Lp\epsilon^{(1+\beta)} \\
&\quad = \left| \nabla f_0(x_0)^\top \left(\frac{1}{\mu(\Omega_{j_0})} \int_{\Omega_{j_0}} (x - x_0) dx + \frac{1}{M} \sum_{m: x_0 \in \Omega_{j_0}^{h_m}} h_m \right) \right| + Lp\epsilon^{(1+\beta)},
\end{aligned}$$

where the integral of a p-dimensional vector is understood as the integral coordinate-wise. We can calculate the integral along the l-th dimension as the following:

$$\frac{1}{\mu(\Omega_{j_0})} \int_{\Omega_{j_0}} (x^l - x_0^l) dx = \frac{1}{2}(b_{j_0,L}^l + b_{j_0,U}^l) - x_0^l.$$

At the same time, all h_m's with $x_0 \in \Omega_{j_0}^{h_m}$ satisfy $b_{L,j_0}^l + h_m^l \le x_0^l < b_{U,j_0}^l + h_m^l$ and are equally spaced along each dimension. If $x_0 \in [b_{L,j_0}^l + m_0\delta, b_{L,j_0}^l + (m_0+1)\delta)$ for some integer $0 \le m_0 < \lceil \epsilon/\delta \rceil$, then it is easy to show that

$$\left| \frac{1}{\mu(\Omega_{j_0})} \int_{\Omega_{j_0}} (x^l - x_0^l) dx + \frac{1}{M} \sum_{m: x_0 \in \Omega_{j_0}^{h_m}} h_m^l \right| \le \delta/2.$$

Together with boundedness of $\nabla f_0(x)$, we have

$$|f_0(x_0) - f_t^{\text{forest}}(x_0)| \le Lp\delta + Lp\epsilon^{(1+\beta)} \quad \text{for all } x_0 \in A_\epsilon.$$

Meanwhile, when ϵ is small enough, for any $x_0 \in B_\epsilon$, we have $f_0(x_0) = f_t^{\text{forest}}(x_0) = 0$. Combining the approximation results for the interior part and for the boundary, we have

$$\|f_0 - f_t^{\text{forest}}\|_\infty \le 2pL\epsilon^{(1+\beta)} \le 2pLc_2^{\text{space}} t^{-(1+\beta)/p},$$

where the last inequality is obtained by (16). This finishes the proof. ■

7.4.3 Estimation Error

Theorem 27 *Given a partition $\mathcal{A}_t = \{\Omega_j\}_{j=1}^t$ of size t, either with or without the shift, we define the "projection" of the true density function onto the partition as*

$$f_{\mathcal{A}_t} = \sum_{j=1}^t \frac{\theta_j}{\mu(\Omega_j)} \cdot \mathbb{1}_{\Omega_j}, \quad \textit{with } \theta_j = \int_{\Omega_j} f_0.$$

and the maximum likelihood estimator supported by the partition $\mathcal{A}_t$ is denoted as $\hat{f}_{n,\mathcal{A}_t}$. Let $\delta_{n,t} = \left(\frac{t\log t}{n/\log n}\right)^{1/2}$, where $t = O(n^{\frac{1}{2r+1}})$ for some $r > 0$. For $\eta > 1$ and some $c_2 \in (0,1)$, we have

$$\mathbb{P}_0^n\left(\rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) \geq \eta\delta_{n,t}\right) \leq 4\exp(-c_2\eta^2 n\delta_{n,t}^2), \tag{35}$$

and

$$\mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{F}_{\mathcal{A}_t}}\prod_{i=1}^n \frac{f(Y_i)}{f_{\mathcal{A}_t}(Y_i)} \geq \exp(\eta^2 n\delta_{n,t}^2)\right) \leq 8\exp(-c_2\eta^2 n\delta_{n,t}^2). \tag{36}$$

The result can be obtained by applying the one-sided large deviation inequality iteratively. Before providing more details of the proof, we would like to introduce a truncated version of the likehood ratio, when the true density f_0 is replaced by the "best approximation" $f_{\mathcal{A}_t}$ within $\mathcal{F}_{\mathcal{A}_t}$. Let $Z_{\mathcal{A}_t,f}(Y_i) = \log(f(Y_i)/f_{\mathcal{A}_t}(Y_i))$. The lower-truncated versions of f and $Z_{\mathcal{A}_t,f}$ are defined similarly as before,

$$\tilde{f} = \begin{cases} f, & \text{if } f > \exp(-\tau)f_{\mathcal{A}_t}, \\ \exp(-\tau)f_{\mathcal{A}_t}, & \text{if } f \leq \exp(-\tau)f_{\mathcal{A}_t}. \end{cases} \quad \widetilde{Z}_{\mathcal{A}_t,f} = Z_{\mathcal{A}_t,\tilde{f}} = \begin{cases} Z_{\mathcal{A}_t,f}, & \text{if } Z_{\mathcal{A}_t,f} > -\tau, \\ -\tau, & \text{if } Z_{\mathcal{A}_t,f} \leq -\tau. \end{cases}$$

We can show that, the likelihood ratio with respect to $f_{\mathcal{A}_t}$ still enjoys most desired properties. In particular, Lemma 16 and Lemma 17 still holds if we replace f_0 by $f_{\mathcal{A}_t}$. Without checking the proofs for those two lemmas, we can verify this point by the following "change-of-measure" type argument: given the partition, the approximating space can be viewed as a collection of multinomial distributions. The sufficient statistics for this model is the vector of counts $(\sum_{i=1}^n \mathbb{1}_{X_i\in\Omega_j})_{1\leq j\leq t}$. The distribution of the random vector remains the same if we view the data as they are generated from $f_{\mathcal{A}_t}$. This implies that, in Lemma 16 and Lemma 17, the probability of the events envolved in the proof is the same as that measured under $f_{\mathcal{A}_t}$. Therefore, results can be obtained by treating $f_{\mathcal{A}_t}$ as the true distribution.

Proof In the proof, Lemma 17 will be applied iteratively to obtain a "sharp" bound of $\rho\left(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}\right)$. Let $\delta_k(n,t) = \left(\frac{t\log t}{n/\log n}\right)^{\omega_k}$, where ω_k is a sequence defined by

$$\omega_1 = 1/4 \quad \text{and} \quad \omega_{k+1} = \frac{1}{2}\omega_k + \frac{1}{4}. \tag{37}$$

We also define $\mathcal{C}_1 := \{f \in \mathcal{F}_{\mathcal{A}_t} : \rho(f, f_{\mathcal{A}_t}) \geq \eta\delta_1(n,t)\}$ and $\mathcal{C}_k := \{f \in \mathcal{F}_{\mathcal{A}_t} : \eta\delta_k(n,t) \leq \rho(f, f_{\mathcal{A}_t}) < \eta\delta_{k-1}(n,t)\}$ for $k \geq 2$.

When $k = 1$, by definition of the MLE, we have

$$\begin{aligned} \mathbb{P}_0^n\left(\rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) \geq \eta\delta_1(n,t)\right) &\leq \mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{C}_1} L_n(f) \geq \sup_{f\in\mathcal{F}_{\mathcal{A}_t}} L_n(f)\right) \\ &\leq \mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{C}_1} L_n(f) \geq L_n(f_{\mathcal{A}_t})\right). \end{aligned} \tag{38}$$

By applying the lower truncation of the log-likelihood ratio, the above probability can be further bounded by

$$\mathbb{P}_0^n\left(\rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) \geq \eta\delta_1(n,t)\right) \leq \mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{C}_1}\sum_{i=1}^n \widetilde{Z}_{\mathcal{A}_t,f}(Y_i) \geq 0\right). \tag{39}$$

Note that the modified version of Lemma 16 can still be applied to $\widetilde{Z}_{\mathcal{A}_t,f}$, which implies that

$$\sup_{f\in\mathcal{C}_1}\left(-\mathbb{E}\widetilde{Z}_{\mathcal{A}_t,f}(Y_1)\right) \geq (1-\gamma)\rho^2(f, f_{\mathcal{A}_t}),$$

Together with previous upper bound (39), we obtain

$$\begin{aligned}
&\mathbb{P}_0^n\left(\rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) \geq \eta\delta_1(n,t)\right) \\
&\quad\leq \mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{C}_1} n^{-1/2}\sum_{i=1}^n\left(\widetilde{Z}_{\mathcal{A}_t,f}(Y_i) - \mathbb{E}\widetilde{Z}_{\mathcal{A}_t,f}(Y_i)\right) \geq n^{-1/2}\sum_{i=1}^n \sup_{f\in\mathcal{C}_1}\left(-\mathbb{E}\widetilde{Z}_{\mathcal{A}_t,f}(Y_i)\right)\right) \\
&\quad\leq \mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{C}_1} n^{-1/2}\sum_{i=1}^n\left(\widetilde{Z}_{\mathcal{A}_t,f}(Y_i) - \mathbb{E}\widetilde{Z}_{\mathcal{A}_t,f}(Y_i)\right) \geq n^{1/2}(1-\gamma)\eta^2\delta_1^2(n,t)\right). \qquad (40)
\end{aligned}$$

In order to apply Lemma 17 to bound the probability on the right hand side of inequality (40), we need to check the conditions (22) and (23). Note that $\|f^{1/2} - f_{\mathcal{A}_t}^{1/2}\|_2^2 = 2 - 2\int_\Omega \sqrt{f(y)f_{\mathcal{A}_t}(y)}dy \leq 2$. We may set d in Lemma 17 to be 2 here. When $t \sim n^{\frac{1}{2r+1}}$, $\delta_1(n,t) = (\frac{t\log t}{n/\log n})^{1/4}$ converges to zero as $n, t \to \infty$. L in Lemma 17 is $n^{1/2}(1-\gamma)\eta^2\delta_1^2(n,t)$ here. First (22) is satisfied, since

$$L = n^{1/2}(1-\gamma)\eta^2\delta_1^2(n,t) = o\left(bn^{1/2}d^2/4\right).$$

By observing $\mathcal{F}_{\mathcal{A}_t} \subset \mathcal{F}_t$, for the function class $\mathcal{F}_{\mathcal{A}_t}$ we can use the same bound for the bracketing entropy as that in Lemma 15. Even it seems to be a slightly loose upper bound, it is still sufficient for our analysis of the estimation error. By carrying out similar calculation as that for Lemma 19,

$$\int_{\frac{b(1-\gamma)}{32}\eta^2\delta_1^2(n,t)}^d \left(H^{[]}(u/(2\exp\tau/2), \mathcal{F}_{\mathcal{A}_t}, \rho)\right)^{1/2} du \lesssim \sqrt{t\log t}.$$

Therefore, (23) is also satisfied when n and t are large enough. With Lemma 17, we have

$$\begin{aligned}
\mathbb{P}_0^n\left(\rho(\hat{f}_{n,t}, f_{\mathcal{A}_t}) > \eta\delta_1(n,t)\right) &\leq 3\exp\left(-(1-b)\varphi(L, d^2, n)\right) \\
&\leq 3\exp\left(-\frac{(1-b)(1-\gamma)^2\eta^2}{2^8c_0+8}\cdot t(\log t)(\log n)\right). \qquad (41)
\end{aligned}$$

When $k \geq 2$, we still want to apply Lemma 17 to establish the inequality

$$\begin{aligned}&\mathbb{P}_0^n \left(\eta\delta_k(n,t) \leq \rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) < \eta\delta_{k-1}(n,t)\right) \\ &\qquad \leq 3\exp\left(-\frac{(1-b)(1-\gamma)^2\eta^2}{2^8c_0+8}\cdot t(\log t)(\log n)\right). \end{aligned} \tag{42}$$

To do so, at the first step, by a similar argument as that for inequalities (38), (39) and (40), we have

$$\begin{aligned}&\mathbb{P}_0^n \left(\eta\delta_k(n,t) \leq \rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) < \eta\delta_{k-1}(n,t)\right) \\ &\qquad \leq \mathbb{P}_{f_0}\left(\sup_{f\in\mathcal{C}_k} n^{-1/2}\sum_{i=1}^n \left(\widetilde{Z}_{\mathcal{A}_t,f}(Y_i) - \mathbb{E}\widetilde{Z}_{\mathcal{A}_t,f}(Y_j)\right) \geq n^{1/2}(1-\gamma)\eta^2\delta_k^2(n,t)\right).\end{aligned}$$

Now L in Lemma 17 is $n^{1/2}(1-\gamma)\eta^2\delta_k^2(n,t)$ here and d can be set as $\eta\delta_{k-1}(n,t)$. By our assumptions, $\delta_k(n,t) = o(\delta_{k-1}(n,t))$ for $k \geq 2$. Then condition (22) is satisfied. Based on the same type of calculation,

$$\begin{aligned}&\int_{\frac{b(1-\gamma)}{32}\eta^2\delta_k^2(n,t)}^{\eta\delta_{k-1}(n,t)} \left(H^{[]}(u/(2\exp\tau/2), \mathcal{F}_{\mathcal{A}_t}, \rho)\right)^{1/2} du \\ &\qquad \lesssim t^{1/2}(\eta\delta_{k-1}(n,t)\sqrt{\log\frac{t^2}{\eta\delta_{k-1}(n,t)}} - \beta\delta_k^2(n,t)\sqrt{\log\frac{t^2}{\beta\delta_k^2(n,t)}} \\ &\qquad\qquad + \frac{\sqrt{\pi}}{2}t^2\left(\frac{\frac{\eta\delta_{k-1}(n,t)}{t^2}}{\sqrt{\pi\log\frac{t^2}{\eta\delta_{k-1}(n,t)}}} - \frac{\frac{\beta\delta_k^2(n,t)}{t^2}}{\sqrt{\pi\log\frac{t^2}{\beta\delta_k^2(n,t)}}}\right)) \\ &\qquad \asymp t^{1/2}\delta_{k-1}(n,t)\sqrt{\log\frac{t^2}{\delta_{k-1}(n,t)}},\end{aligned}$$

where $\beta = \frac{b(1-\gamma)}{32}\eta^2$. Plugging in $\delta_{k-1}(n,t) = (\frac{t\log t}{n/\log n})^{\omega_{k-1}}$, after some calculation, we can verify that condition (23) is also satisfied when n and t are large enough. Note that

$$\begin{aligned}&\varphi(n^{1/2}(1-\gamma)\eta^2\delta_k^2(n,t), (\eta\delta_{k-1}(n,t))^2, n) \\ &\quad = \frac{n(1-\gamma)^2\eta^4\delta_k^4(n,t)}{8(8c_0\eta^2\delta_{k-1}^2(n,t) + (1-\gamma)\eta^2\delta_k^2(n,t))} \\ &\quad \geq \frac{n(1-\gamma)^2\eta^2\delta_k^4(n,t)}{8(8c_0\delta_{k-1}^2(n,t) + (1-\gamma)\delta_{k-1}^2(n,t))} \\ &\quad \geq \frac{n(1-\gamma)^2\eta^2(\frac{t\log t}{n/\log n})^{2\omega_{k-1}+1}}{(2^8c_0+8(1-\gamma))(\frac{t\log t}{n/\log n})^{2\omega_{k-1}}} \\ &\quad \geq \frac{(1-\gamma)^2\eta^2}{2^8c_0+8}\cdot t(\log t)(\log n).\end{aligned}$$

Applying Lemma 17, we get the inequality (42).

Based on (37), it is easy to see $\omega_k = \frac{1}{2} - \left(\frac{1}{2}\right)^{k+1}$. If we set $K(n,t)$ to be $\frac{\log\log \frac{n/\log n}{t\log t} - \log\log c}{\log c}$, then $\eta\delta_{K(n,t)} \le c\eta\delta_{n,t}$. It implies that

$$\begin{aligned}
&\mathbb{P}_0^n\left(\rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) \ge c\eta\delta_{n,t}\right) \\
&\quad\le \mathbb{P}\left(\rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) \ge \eta\delta_1(n,I)\right) \\
&\qquad + \sum_{k=2}^{K(n,t)} \mathbb{P}_0^n\left(\eta\delta_k(n,t) \le \rho(\hat{f}_{n,\mathcal{A}_t}, f_{\mathcal{A}_t}) < \eta\delta_{k-1}(n,t)\right) \\
&\quad\le 3K(n,t)\exp\left(-\frac{(1-b)(1-\gamma)^2\eta^2}{2^8c_0+8}\cdot t(\log t)(\log n)\right).
\end{aligned}$$

Therefore, for some $\eta > 1$ and appropriately chosen c_2, the inequality (35) holds.

The proof for (36) is similar. ■

8. Discussion

In this paper, we study the asymptotic properties of a class of multivariate density estimation methods based on adaptive partitioning, including density trees and density forests. For the former ones, under both frequentist and Bayesian settings, explicit convergence rates are obtained, while a significant difference between these two types of methods is that the posterior concentration rate is adaptive to the smoothness of the underlying density function. We also obtain explicit rates when the density function is spatially sparse, belongs to the space of bounded variation, or is Hölder continuous. For density functions lying in the Hölder or Besov space, the rate is minimax up to a logarithmic term. Another advantage of the partition based method is that, when the density function only show variations with respect to a subset of variables or is sparse, the rate will not be affected by the full dimension of the problem. Instead, it is determined by the effective dimension or complexity of the density.

For density forests, we have focused on ensembles for which each tree estimator is obtained by maximizing the likelihood. We demonstrate for the Hölder space $\mathcal{H}^{1,\beta}, 0 < \beta \le 1$, minimax rate can be achieved by density forests while the rate for density trees is suboptimal. However, the result is not adaptive, in the sense that to achieve fast convergence the size of binary partitions should match to the sample size and the matching depends on the parameter β of the Hölder space. It would be interesting to study whether a penalized estimator or a Bayesian approach under an appropriated prior can achieve minimax convergence adaptively.

Another limit of the current theoretical result for density forests is that the rate is determined by the full dimension of the problem, as we have narrowed the parameter space to a collection of balanced partitions across all dimensions, instead of searching over all possible shapes, in order to control the variance. An interesting further direction is to investigate whether it is affordable to search over a larger parameter space to retain the variable screening property.

Acknowledgments

We would like to acknowledge support for this research project from the National Science Foundation (NSF grant DMS1407557, DMS1811920 and DMS1952386) and the University of Pittsburgh Center for Research Computing.

Appendix A Proofs for Corollaries in Section 4

The key to calculating the convergence rate is to derive an explicit approximation rate. That is, we need to know the value of r, such that for any true density function f_0 belonging to the function class under consideration, there exists a sequence of approximating density functions $f_t \in \Theta_t$ satisfying that $\rho(f_0, f_t) \leq At^{-r}$, where $A > 0$ is a constant. In this section, we provide proofs for approximation rates for the three specific classes of density functions discussed in Section 4.

A.1 Spatial Sparsity

We prove Lemma 5 from Section 4.1 in this section.

Proof Let $g_K = \sum_{k=1}^{K} \langle g_0, \xi_{(k)} \rangle \xi_{(k)}$. From condition (10) we have

$$\begin{aligned} \|g_0 - g_K\|_2^2 &= \Big\| \sum_{k=K+1}^{+\infty} \langle g_0, \xi_{(k)} \rangle \xi_{(k)} \Big\|_2^2 \\ &= \sum_{k=K+1}^{+\infty} \langle g_0, \xi_{(k)} \rangle^2 \\ &\leq C^2 \sum_{k=K+1}^{+\infty} k^{-2/q} \leq \frac{C^2}{2/q-1} K^{-(2/q-1)}. \end{aligned}$$

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

$$\begin{aligned} \rho^2(f_0, \tilde{g}_K^2) &= \|g_0 - \tilde{g}_K\|_2^2 \\ &= \|g_0 - g_K\|_2^2 + \Big(1 - \frac{1}{\|g_K\|_2}\Big)^2 \|g_K\|_2^2 \\ &\leq \|g_0 - g_K\|_2^2 + 1 - \|g_K\|_2^2 \\ &= 2\|g_0 - g_K\|_2^2 \\ &\leq \frac{2C^2}{2/q-1} K^{-(2/q-1)}. \end{aligned} \tag{43}$$

Note that 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, $2^p K$ is the largest possible sized binary partition on which the density function $\tilde{g}_K$ is piecewise constant. Replacing K in (43) by $t/2^p$, we get the desired result of the approximation rate. ■

A.2 Density Functions of Bounded Variation

Let Λ be the set of indices for the wavelet basis. Each element in Λ is a pair of scale and location parameters. We will denote by Σ_N the spaces consisting of N-term approximation in the Haar system, in other words,

$$\Sigma_N := \left\{ \sum_{\lambda \in E} c_\lambda \xi_\lambda : E \subset \Lambda, |E| \leq N \right\},$$

where $|E|$ denotes the cardinality of the discrete set E.

First, we cite a result by Cohen et al. (1999). It provides a bound for the approximation rate to a function of bounded variation by Σ_N.

Lemma 28 *If $f \in BV(\Omega)$ has mean value zero on Ω, we have*

$$\inf_{g \in \Sigma_N} \|f - g\|_2 \leq C N^{-1/2} V_\Omega(f), \tag{44}$$

with $C = 2592(3\sqrt{5} + \sqrt{3})$.

Assume f_0 is a density function on Ω of bounded variation. By subtracting the mean, we can always assume that $\sqrt{f_0}$ has mean value zero over Ω. For the square root of f_0, applying the lemma above, we can find an N-term approximation g in the Haar system, such that $\|\sqrt{f_0} - g\|_2 \lesssim N^{-1/2}$. Translating this inequality into the size of partition, we reach the conclusion that for any density function in $BV(\Omega)$, we can find a sequence of approximations in Θ_t, such that $\rho(f_0, f_t) \lesssim t^{-1/2}$. Corollary 8 follows.

A.3 Hölder Space

For the Hölder space, the approximation result is derived based on an alternative construction of multivariate Haar basis. We first introduce this construction in the following subsection.

A.3.1 Tensor Haar Basis

In the one-dimensional case, the Haar wavelet's mother wavelet function ψ and its scaling function ϕ are the same as those defined in Section 4.1.1. For any $j \in \mathbb{N}$ and $0 \leq k < 2^j$, define

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

Then the Haar basis $\boldsymbol{\Xi}$

$$\boldsymbol{\Xi} = \{\phi\} \cup \{\psi_{jk}, j \in \mathbb{N}, 0 \leq k < 2^j\},$$

forms an orthonormal basis for Hilbert space $L^2([0,1])$.

Under multivariate settings, we can 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 specific, if $\mathcal{X}_1, \ldots, \mathcal{X}_p$ are p copies of $L^2([0,1])$ and $\boldsymbol{\Xi}_1, \cdots, \boldsymbol{\Xi}_p$ are Haar bases of these spaces respectively, then $L^2([0,1]^p)$ is isomorphic to $\otimes_{l=1}^p \mathcal{X}_l$. Define tensor Haar basis $\boldsymbol{\Xi}$ by

$$\boldsymbol{\Xi} = \{\xi : \xi = \prod_{l=1}^p \xi_l, \xi_l \in \boldsymbol{\Xi}_l\}.$$

Based on the property of tensor product of Hilbert spaces, we know that $\boldsymbol{\Xi}$ is an orthonormal basis for $L^2([0,1]^p)$.

Next, we derive the approximation rate for the Hölder space by using the tensor Haar basis.

A.3.2 Approximation Rate

Here, we provide a proof for Lemma 10.

Proof The proof for Hölder continuous functions is relatively simple. We can consider a balanced partition $\mathcal{A}_t$ of size t, with both $E^l_{\max}(\mathcal{A}_t)$ and $E^l_{\min}(\mathcal{A}_t)$ at the order $t^{-1/p}$ for all $1 \le l \le p$. Then the approximation

$$f_t = \sum_{i=1}^{t} \frac{\int_{\Omega_j} f_0}{\mu(\Omega_j)} \mathbb{1}_{\Omega_j}$$

can achieve the desired approximation rate. The analysis is similar to that for the bias term of density forests.

For each tensor Haar function ξ, we use $R(\xi)$ to denote its supporting rectangle. First we claim that, for any density function f_0 which is mixed-Hölder continuous,

$$|\langle \sqrt{f_0}, \xi \rangle| \le L |R(\xi)|^{\beta+1/2} \text{ for all } \xi,$$

where $L > 0$ is a constant, and $|R(\xi)|$ is the volume of the rectangle. In the two-dimensional case, for any point $(x_1', x_2') \in R(\xi)$, we have

$$\begin{aligned}
&|\langle \sqrt{f_0}, \xi \rangle|^2 \\
&= \left(\int_R \sqrt{f_0(x_1, x_2)} \xi(x_1, x_2) dx_1 dx_2 \right)^2 \\
&= \left(\int_R \left(\sqrt{f_0(x_1, x_2)} - \sqrt{f_0(x_1, x_2')} - \sqrt{f_0(x_1', x_2)} + \sqrt{f_0(x_1', x_2')} \right) \xi(x_1, x_2) dx_1 dx_2 \right)^2 \\
&\le \int_R \left(\sqrt{f_0(x_1, x_2)} - \sqrt{f_0(x_1, x_2')} - \sqrt{f_0(x_1', x_2)} + \sqrt{f_0(x_1', x_2')} \right)^2 dx_1 dx_2 \cdot \int_R \xi^2 \\
&\le L^2 \int_R |x_1 - x_1'|^{2\beta} |x_2 - x_2'|^{2\beta} dx_1 dx_2 \\
&= L^2 |R|^{2\beta+1}.
\end{aligned}$$

The claim follows. In the multi-dimensional case, the claim can be shown in a similar way.

Now let $g_0 = \sqrt{f_0}$. We can expand g_0 with respect to the tensor Haar basis. The expansion can be written as $g_0 = \sum_\xi \langle g_0, \xi \rangle \xi$. Define

$$g_\epsilon = \sum_{\xi : |R(\xi)| > \epsilon} \langle g_0, \xi \rangle \xi.$$

g_ϵ is an approximation to g_0 obtained by requiring that the volumes of the supporting rectangles of the involved wavelet basis functions are greater than ϵ. We will derive an approximation rate as a function of ϵ first, and then convert the lower bound on the volume to an upper bound on the size of the partition, which yields an approximation rate as a function of the size of the partition. Note that g_ϵ 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_ϵ. The upper bounds for the approximation errors $\|g_0 - g_\epsilon\|^2$ and $\rho^2(f_0, \tilde{g}_\epsilon)$ will be derived next.

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

$$\begin{aligned}
\|g_0 - g_\epsilon\|^2 &= \Big\| \sum_{\xi:|R(\xi)|<\epsilon} \langle g_0, \xi\rangle \xi \Big\|_2^2 \\
&= \sum_{\xi:|R(\xi)|<\epsilon} \langle g_0, \xi\rangle^2 \\
&\le L^2 \sum_{\xi:|R(\xi)|<\epsilon} |R(\xi)|^{2\beta+1} \\
&\le 2^p L^2 \sum_{l_1,\cdots,l_p} \sum_{R\in\mathcal{R}^{l_1,\cdots,l_p},|R|<\epsilon} |R|^{2\beta+1}. \qquad (45)
\end{aligned}$$

The last inequality follows from the fact that, given a supporting rectangle, there are at most 2^p basis functions defined on it. Let $N = \lceil \log_{\frac{1}{2}} \epsilon \rceil$,

$$\begin{aligned}
(45) &= 2^p L^2 \sum_{l_1+\cdots+l_p\ge N} \sum_{R\in\mathcal{R}^{l_1,\cdots,l_p}} |R|^{2\beta+1} \\
&= 2^p L^2 \sum_{l_1+\cdots+l_p\ge N} (\frac{1}{2})^{2\beta(l_1+\cdots+l_p)} \sum_{R\in\mathcal{R}^{l_1,\cdots,l_p}} |R| \\
&= 2^p L^2 \sum_{l_1+\cdots+l_p\ge N} (\frac{1}{2})^{2\beta(l_1+\cdots+l_p)}.
\end{aligned}$$

The last equality is obtained by plugging in $\sum_{R\in\mathcal{R}^{l_1,\cdots,l_p}} |R| = 1$. Note that

$$\begin{aligned}
&\sum_{l_1+\cdots+l_p\ge N} (\frac{1}{2})^{2\beta(l_1+\cdots+l_p)} \\
&\le \sum_{l_1=0}^{N} \sum_{l_2=0}^{N-l_1} \cdots \sum_{l_p=N-(l_1+\cdots+l_{p-1})}^{+\infty} (\frac{1}{2})^{2\beta(l_1+\cdots+l_p)} \\
&\quad + \sum_{l_1=0}^{N} \sum_{l_2=0}^{N-l_1} \cdots \sum_{l_{p-1}=N-(l_1+\cdots+l_{p-2})}^{+\infty} \sum_{l_p=0}^{+\infty} (\frac{1}{2})^{2\beta(l_1+\cdots+l_p)} \\
&\quad + \cdots + \sum_{l_1=N}^{+\infty} \sum_{l_2=0}^{+\infty} \cdots \sum_{l_p=0}^{+\infty} (\frac{1}{2})^{2\beta(l_1+\cdots+l_p)} \\
&\le (N+1)^{p-1} \frac{(\frac{1}{2})^{2\beta N}}{1-2^{-2\beta}} + (N+1)^{p-2} \frac{(\frac{1}{2})^{2\beta N}}{(1-2^{-2\beta})^2} + \cdots + \frac{(\frac{1}{2})^{2\beta N}}{(1-2^{-2\beta})^p} \\
&= (\frac{1}{2})^{2\beta N} \frac{(N+1)^p - (1-2^{-2\beta})^{-p}}{(N+1)(1-2^{-2\beta}) - 1} \\
&\le 2\epsilon^{2\beta} (\log_{\frac{1}{2}} \epsilon)^p.
\end{aligned}$$

From this, we know that

$$\|g_0 - g_\epsilon\|_2^2 \leq 2^{p+1} L^2 \epsilon^{2\beta} (\log_{\frac{1}{2}} \epsilon)^p.$$

After we normalize g_ϵ to $\tilde{g}_\epsilon$,

$$\begin{aligned} \rho^2(f_0, \tilde{g}_\epsilon^2) &= \|g_0 - \tilde{g}_\epsilon\|_2^2 \\ &= \|g_0 - g_\epsilon\|_2^2 + \left(1 - \frac{1}{\|g_\epsilon\|_2}\right)^2 \|g_\epsilon\|_2^2 \\ &\leq \|g_0 - g_\epsilon\|_2^2 + 1 - \|g_\epsilon\|_2^2 \\ &= 2\|g_0 - g_\epsilon\|_2^2. \end{aligned}$$

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

$$\rho^2(f_0, \tilde{g}_\epsilon^2) = \|g_0 - \tilde{g}_\epsilon\|_2^2 \leq 2^{p+2} L^2 \epsilon^{2\beta} (\log_{\frac{1}{2}} \epsilon)^p,$$

Next, we will convert 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 ϵ, then the lengths of the edges can not be smaller than $2^{-\lfloor \log_{\frac{1}{2}} \epsilon \rfloor}$. The size of the partition supporting $\tilde{g}_\epsilon$ can be bounded by $2^p 2^{p \log_{\frac{1}{2}} \epsilon} = 2^p \epsilon^{-p}$. 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 t, we can choose the value of ϵ by solving $2^p \epsilon^{-p} = t$ and define $\tilde{g}_\epsilon \in \Theta_t$ as above. Then based on the upper bound in terms of ϵ we reach a conclusion that $\tilde{g}_\epsilon$ is an approximation satisfying $\rho^2(f_0, \tilde{g}_\epsilon) \leq 2^p L t^{-\beta/p} (\log t)^{p/2}$. This finishes the proof. ■

Appendix B Analysis of Marginal Posterior Probability of a Partition

In Section 2.3, we have a discussion about the connection between the Bayesian estimator and the penalized MLE. Here, we provide more details about how to derive (4) and (5).

Under the prior distribution introduced by Lu et al. (2013), the marginal posterior probability of a partition after a logarithmic trasformation can be written as

$$\log\left(\Pi_n^*\left(\mathcal{F}(\{\Omega_j\}_{j=1}^t) \middle| Y_1, \cdots, Y_n\right)\right) = -\lambda t + \log\left(\frac{D(\alpha + n_1, \cdots, \alpha + n_t)}{D(\alpha, \cdots, \alpha)} \prod_{j=1}^t \frac{1}{|\Omega_j|^{n_j}}\right). \quad (46)$$

We replace the multivariate Beta function by Gamma functions

$$\begin{aligned} (46) &= -\lambda t + \sum_{j=1}^n n_j \log\left(\frac{1}{\mu(\Omega_j)}\right) \\ &\quad + \log\Gamma(\alpha t) - t\log\Gamma(\alpha) + \sum_{j=1}^t \log\Gamma(\alpha + n_j) - \log\Gamma(\alpha t + n) \end{aligned}$$

By applying the Stirling's formula, we obtain

$$\begin{aligned}
(46) &= -\lambda t + \sum_{j=1}^{n} n_j \log\left(\frac{1}{\mu(\Omega_j)}\right) \\
&\quad + \frac{1}{2}\log(\alpha t) + \alpha t \log(\alpha t) - \alpha t - t\log\Gamma(\alpha) + O(t) \\
&\quad + \sum_{j=1}^{t}\left(\frac{1}{2}\log(\alpha + n_j - 1) + (\alpha + n_j - 1)\log(\alpha + n_j - 1) - (\alpha + n_j - 1)\right) \\
&\quad - \frac{1}{2}\log(\alpha t + n) - (\alpha t + n - 1)\log(\alpha t + n) + (\alpha t + n) \\
&= -\lambda t + \sum_{j=1}^{t} n_j \log\left(\frac{n_j}{n\mu(\Omega_j)}\right) \\
&\quad + \frac{1}{2}\log(\alpha t) - \alpha t - t\log\Gamma(\alpha) + \alpha t\log(\alpha t) - (\alpha t - \frac{1}{2})\log(\alpha t + n) + O(t) \\
&\quad + \sum_{j=1}^{t}\left((\alpha - \frac{1}{2})\log(\alpha + n_j - 1) + n_j \log((\alpha + n_j - 1)/n_j)\right) - n\log((\alpha t + n)/n).
\end{aligned}$$

After applying the Taylor's expansion of $\log(1+x)$, we have

$$\begin{aligned}
(46) &= -\lambda t + \sum_{j=1}^{t} n_j \log\left(\frac{n_j}{n\mu(\Omega_j)}\right) \\
&\quad + \frac{1}{2}\log(\alpha t) - \alpha t - t\log\Gamma(\alpha) + \alpha t\log(\alpha t) - (\alpha t - \frac{1}{2})\log(\alpha t + n) + O(t) \\
&\quad + \sum_{j=1}^{t}\left((\alpha - \frac{1}{2})\log(\alpha + n_j - 1) + n_j((\alpha - 1)/n_j + O((\alpha/n_j)^2))\right) - n(\alpha t/n + O((\alpha t/n)^2)) \\
&= -\lambda t + \sum_{j=1}^{t} n_j \log\left(\frac{n_j}{n\mu(\Omega_j)}\right) \\
&\quad + \frac{1}{2}\log(\alpha t) - (\alpha + \log\Gamma(\alpha))t + \alpha t\log(\alpha t) + O(t) \\
&\quad - (\alpha t - \frac{1}{2})\log(\alpha t + n) + \sum_{j=1}^{t}(\alpha - \frac{1}{2})\log(\alpha + n_j - 1).
\end{aligned}$$

For the last part, when t is viewed as fixed or $t \ll n$,

$$\begin{aligned}
&-(\alpha t - \frac{1}{2})\log(\alpha t + n) + \sum_{j=1}^{t}(\alpha - \frac{1}{2})\log(\alpha + n_j - 1) \\
&\asymp -(\alpha t - \frac{1}{2})\log(\alpha t + n) + (\alpha - \frac{1}{2})t\left(\log(\alpha t - t + n) - \log t\right) \\
&= -\frac{1}{2}t\log(\alpha t + n) + \frac{1}{2}\log(\alpha t + n) + (\alpha - \frac{1}{2})t\log t.
\end{aligned}$$

This way, we derive (4).

As t grows with n, the analysis of last part is different. It is easy to see the quantity is at the order of $t \log t$ when $t = C(n/\log n)^r, 0 < r < 1$ and $C > 0$, or even when $t/n \to \zeta \in (0, 1)$. Therefore, under the prior distribution introduced in this paper, we have asymptotic result (5).

References

Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2): 584–653, 2006.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

Emmanuel. J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52 (12):5406–5425, 2006.

Albert Cohen, Ronald DeVore, Pencho Petrushev, and Hong Xu. Nonliner approximation and the space $bv(\mathbb{R}^2)$. *American Journal of Mathematics*, 121(3):587–628, 1999.

Jingyi Cui, Hanyuan Hang, Yisen Wang, and Zhouchen Lin. GBHT: Gradient boosting histogram transform for density estimation. In *Proceedings of the Thirty-Eighth International Conference on Machine Learning*, pages 2233–2243. 2021.

R. de Jonge and J.H. van Zanten. Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics*, 6: 1984–2001, 2012.

Ronald A. DeVore, B.D. Jawerth, and B.J. Lucier. Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38(2):719–746, 1992.

David L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):100–115, 1993.

David L. Donoho, Iain M. Johnstone, Gerard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 57(2):301–369, 1995.

David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

R. H. Farrell. On the lack of a uniformly consistent sequence of estimators of a density function in certain cases. *The Annals of Mathematical Statistics*, 38(2):471–474, 1967.

Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Montréal, Canada, 2014.

U. Grenander. *Abstract Inference*. Probability and Statistics Series. John Wiley & Sons, 1981.

Hui Jiang, John Chong Mu, Kun Yang, Chao Du, Luo Lu, and Wing Hung Wong. Computational aspects of optional Pólya tree. *Journal of Computational and Graphical Statistics*, 25(1):301–320, 2016.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

Jussi Klemelä. Multivariate histograms with data-dependent partitions. *Statistica Sinica*, 19:159–176, 2009.

Andrey N. Kolmogorov and Vladimir M. Tikhomirov. *Selected Works of A.N. Kolmogorov*. Number v.2 in Mathematics and Its Applications: Soviet Series. Kluwer Academic Publishers, 1992.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

Willem Kruijer, Judith Rousseau, and Aad van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning. In *Proceedings of the Twenty-Ninth International Coference on International Conference on Machine Learning*, pages 507–514. Edinburgh, Scotland, 2012.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Dangna Li, Kun Yang, and Wing Hung Wong. Density estimation via discrepancy based adaptive sequential partition. In *Conference on Neural Information Processing Systems 29*, pages 1091–1099. Barcelona, Spain, 2016.

Meng Li and Li Ma. Learning asymmetric and local features in multi-dimensional data through wavelets with adaptive recursive partitioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, forthcoming.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.

Linxi Liu. *Convergence Rates of a Class of Multivariate Density Estimators Based on Adaptive Partitioning*. PhD thesis, Stanford University, Stanford, California, 2016.

Linxi Liu, Dangna Li, and Wing Hung Wong. Convergence rates of a partition based Bayesian multivariate density estimation method. In *Advances in Neural Information Processing Systems 30*, pages 4738–4746. Long Beach, California, 2017.

Luo Lu, Hui Jiang, and Wing H. Wong. Multivariate density estimation by Bayesian sequential partitioning. *Journal of the American Statistical Association*, 108(504):1402–1410, 2013.

Gábor Lugosi and Andrew Nobel. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

Li Ma and Wing Hung Wong. Coupling optional Pólya trees and the two sample problem. *Journal of the American Statistical Association*, 106(496):1553–1565, 2011.

Michael H. Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statistica Sinica*, 10:399–431, 2000.

Sergei Mihailovic Nikol'skii. *Approximation of Functions of Several Variables and Imbedding Theorems*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 2012.

Hong Ooi. Density visualization and mode hunting using trees. *Journal of Computational and Graphical Statistics*, 11(2):328–347, 2002.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

G. Poggi and R. A. Olshen. Pruned tree-structured vector quantization of medical images with segmentation and improved prediction. *IEEE Transactions on Image Processing*, 4 (6):734–742, 1995.

Thibault Randrianarisoa. Smoothing and adaptation of shifted Pólya tree ensembles. *Bernoulli*, 28(4):2492 – 2517, 2022.

Vincent Rivoirard and Judith Rousseau. Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7(2):311–334, 2012.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38(1): 146–180, 2010.

Veronika Ročková and Stéphanie van der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108 – 2131, 2020.

David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

David W. Scott. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics*, 13(3):1024 – 1040, 1985.

Nong Shang. Tree-structured density estimation and dimensionality reduction. In *Proceedings of the Twenty-Sixth Symposium on the Interface*, pages 172–176. Research Triangle Park, North Carolina, 1994.

Weining Shen and Subhashis Ghosal. Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213, 2015.

Weining Shen, Surya T. Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6): 2555–2591, 1997.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615, 1994.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall London, New York, 1986.

Jacopo Soriano and Li Ma. Probabilistic multi-resolution scanning for two-sample differences. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2): 547–572, 2017.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

George R. Terrell. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410):470–477, 1990.

Karine Tribouley. Practical estimation of multivariate densities using wavelet methods. *Statistica Neerlandica*, 49(1):41–62, 1995.

Ananya Uppal, Shashank Singh, and Barnabás Poczós. Nonparametric density estimation and convergence rates for GANs under Besov IPM losses. In *Advances in Neural Information Processing Systems 32*, pages 9086–9097. Vancouver, Canada, 2019.

Stéphanie van der Pas and Veronika Ročková. Bayesian dyadic trees and histograms for regression. In *Advances in Neural Information Processing Systems 30*, pages 2089–2099. Long Beach, California, 2017.

Grace Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 40(3):364–372, 1978.

Wing Hung Wong and Li Ma. Optional Pólya tree and Bayesian inference. *The Annals of Statistics*, 38(3):1433–1459, 2010.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.