Random Forest Density Estimation

Hongwei Wen [1] **Hanyuan Hang** [1]

Abstract

We propose a density estimation algorithm called *random forest density estimation* (*RFDE*) based on random trees where the split of cell is along the midpoint of the randomly chosen dimension. By combining the efficient random tree density estimation (RTDE) and the ensemble procedure, RFDE can alleviate the problems of boundary discontinuity suffered by partition-based density estimations. From the theoretical perspective, we first prove the fast convergence rates of RFDE if the density function lies in the Hölder space $C^{0,\alpha}$. Moreover, if the density function resides in the subspace $C^{1,\alpha}$, which contains smoother density functions, we for the first time manage to explain the benefits of ensemble learning in density estimation. To be specific, we show that the upper bound of the ensemble estimator RFDE turns out to be strictly smaller than the lower bound of its base estimator RTDE in terms of convergence rates. In the experiments, we verify the theoretical results, and show the promising performance of RFDE on both synthetic and real world datasets. Moreover, we evaluate our RFDE through the problem of anomaly detection as a possible application.

1. Introduction

In the field of machine learning, the leverage of feature density can be found in most tasks. For example, regression and classification problems employ feature density to enhance the estimation of the label density conditioned on features (Maldonado et al., 2019; Tumiran & Sivakumar, 2021; Steininger et al., 2021; Silverman, 2018); clustering and anomaly detection directly use feature density to determine the neighbors or outsiders (Campello et al., 2020; Li et al., 2020; Corizzo et al., 2019; Zhang et al., 2018; Zhao & Shi, 2019; Emadi & Mazinani, 2018); And for adversarial attacks and defenses, algorithms always seek the weakness in feature density to tamper the outcomes (Huang et al., 2021; Hu et al., 2019; Li et al., 2019), etc. Consequently, the study of *density estimation*, which targets on estimating the underlying probability density function of features, has attracted more and more attention (Bodin et al., 2021; Cui et al., 2021).

Density estimation assumes the observations are i.i.d. drawn from the underlying probability density function and constructs an approximated version accordingly. The most popular and widely-used method is called kernel density estimation (KDE) (Rosenblatt, 1956; Parzen, 1962). However, KDE has its own drawbacks. That is, the lack of local adaptivity. In particular, when KDE encounters density functions with different local properties, the performance will be badly affected. Subsequently, partition-based methods have been proposed, which construct appropriate partitions of the input space to better use the local information (Klemelä, 2009; López-Rubio, 2013; Liu & Wong, 2014; Li et al., 2016). The first and most intuitional idea is the histogram density estimation (HDE), which quickly comes into vogue in academy as the basic form of density estimation (Freedman & Diaconis, 1981; Härdle et al., 2012). Although HDE enjoys sound theoretical properties, the histogram partition is of low computational efficiency and even corrupts for high dimension data, which forces researchers to seek tree-based algorithms (Ram & Gray, 2011; Criminisi et al., 2011; Criminisi & Shotton, 2013). Unfortunately, a majority of tree-based methods fail to gain theoretical support from the perspective of the statistical learning theory. Moreover, they suffer from the boundary discontinuity, i.e. the density estimator is discontinuous at the partition boundary.

To overcome the challenges in density estimation problems, we propose a tree-based learning method with learning theory guarantees called *random tree density estimation* (RTDE) based on random tree partition (Breiman, 2004; Biau, 2012). The construction procedure of RTDE starts with partitioning the feature space into non-overlapping cells along the midpoint of the randomly-chosen dimension, which helps separating the local features from area to area. Then we apply a constant estimator to each cell and obtain an RTDE estimator. Since the density estimation at each point is only affected by the samples falling into the cell

[1] Department of Applied Mathematics, University of Twente, Enschede, The Netherlands. Correspondence to: Hongwei Wen <h.wen@utwente.nl>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

containing that point, the RTDE estimator enjoys the nature of being local. However, RTDE still suffers from the boundary discontinuity. To alleviate this problem, we generate several random partitions and the corresponding RTDEs, and then we average these estimators to obtain the *random forest density estimation* (RFDE).

The strengths of RFDE can be summarized as follows: First of all, the tree structure of the random partition enables RFDE to be locally adaptive and the efficient partition rule brings higher computational efficiency than HDE. Moreover, due to the intrinsic randomness of the partition, the probability of a sample point being on the partition boundaries of one tree is zero, and the probability remains zero if it is on the partition boundaries of several trees simultaneously. As a result, with the number of trees increasing, the asymptotic smoothness of RFDE estimators can be achieved, which further leads to the improvement of the estimation accuracy.

The contributions of this paper come from the theoretical and experimental perspectives:

(i) We propose a tree-based density estimation algorithm called *random forest density estimation* (RFDE), which not only alleviates the problem of boundary discontinuity long plaguing partition-based methods, but also enjoys high computational efficiency.

(iii) From a learning theory point of view, we prove the fast convergence rates of RFDE with assumptions that the underlying density functions lie in the Hölder space $C^{0,\alpha}$.

(iii) To our best knowledge, we are the first to explain the strength of ensemble in the density estimation from the theoretical point of view. To be specific, in the space $C^{1,\alpha}$, we show that the lower bound for the excess risk of RTDE turns out to be greater than the upper bound for that of RFDE when $d \geq 2$. That is, when $d \geq 2$, RFDE converges strictly faster than RTDE.

(iv) In experiments, we validate the theoretical results and evaluate our RFDE through comparisons on both synthetic and real data. Moreover, we evaluate our RFDE through the problem of anomaly detection as a possible application.

2. Methodology

We dedicate this section to the methodology of our *random forest density estimation* (RFDE). To this end, we begin by introducing some notations to be used throughout. Then in Section 2.2, we give explicit description of the *random tree partitions*. The formulations of our random tree density estimators and the ensemble version are presented in Section 2.3 and 2.4.

2.1. Notations

Let $\mathcal{X} \subset \mathbb{R}^d$ be a subset, μ be the Lebesgue measure with $\mu(\mathcal{X}) > 0$, and P be a probability measure with support $\mathcal{X}$ which is absolute continuous with respect to μ with density f. Let the training data $D := (X_1, \ldots, X_n)$ be i.i.d observations with the same distribution as X drawn from P. We denote B_r as the centered hypercube of $\mathbb{R}^d$ with side length $2r$, that is $B_r := [-r, r]^d = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_i \in [-r, r], i = 1, \ldots, d\}$, and write $B_r^c := \mathbb{R}^d \setminus [-r, r]^d$ for the complement of B_r. Throughout this paper, we use the notation $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ to denote that there exists positive constants c and c' such that $a_n \leq c b_n$ and $a_n \geq c' b_n$, respectively, for all $n \in \mathbb{N}$.

2.2. Random Tree Partition

In this paper, we investigate the mid-point random tree partitions suggested by Biau (2012) and Breiman (2004). To be specific, let $A_0^1 := B_r \supset \mathcal{X}$ be the initial rectangular cell containing the support $\mathcal{X}$ and $\pi_0 := \{A_0^1\}$ be the initialized cell partition. In addition, let $p \in \mathbb{N}$ be a deterministic parameter, fixed beforehand by the user, and possibly depending on n.

In the first step, we choose one of the coordinates $X = (X_1, \ldots, X_d)$ with the ℓ-th feature X_ℓ having a probability $1/d$ of being selected, and then split B_r into two rectangular cells along the midpoint of the chosen side. In other words, there exist $1 \leq \ell \leq d$ such that $B_r = A_1^1 \cup A_1^2$, where $A_1^1 := \{(x, y) \in B_r : x_\ell \leq 0\}$ and $A_1^2 := B_r \setminus A_1^1$. In this way, we get a partition with two rectangular cells denoted as $\pi_1 := \{A_1^1, A_1^2\}$. Note that the total number of possible partitions after the first step is equal to the dimension d. Suppose after $i - 1$ steps of the recursion, $1 \leq i \leq p$ we have obtained a partition π_{i-1} of B_r with 2^{i-1} rectangular cells. In the i-th step, further partitioning of the region is defined as follows:

(i) For each rectangular cell A_{i-1}^j, $1 \leq j \leq 2^{i-1}$, a coordinate of $X = (X_1, \ldots, X_d)$, namely $Z_{i,j}$ is selected, with the ℓ-th feature having a probability $1/d$ to be chosen, that is,

$$\mathrm{P}(Z_{i,j} = \ell) = 1/d, \qquad \text{for } 1 \leq \ell \leq d. \tag{1}$$

(ii) For each rectangular cell A_{i-1}^j, $1 \leq j \leq 2^{i-1}$, once the coordinate is selected, the split is at the midpoint of the chosen side. As a result, each rectangular cell A_{i-1}^j is divided into two new ones, namely A_i^{2j-1} and A_i^{2j}. We denote the set of all these cells $\{A_i^j, 1 \leq j \leq 2^i\}$ by π_i.

After p recursive steps, we obtain the partition of B_r, i.e.

$$\pi_p := \{A_p^j\}_{j \in \mathcal{I}_p} := \{A_p^j, 1 \leq j \leq 2^p\}. \tag{2}$$

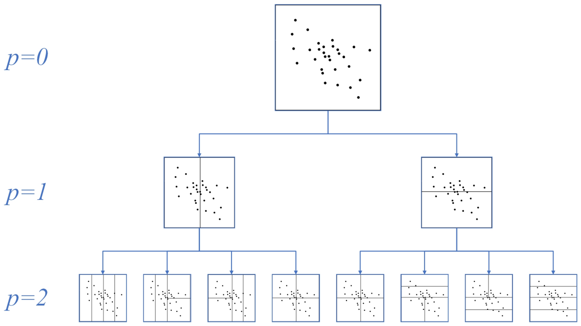


Figure 1. Random tree partitions with $p = 2$ and $d = 2$.

We call it a *random tree partition* with depth p. The complete process is presented in Algorithm 1 and an illustration is shown in Figure 1.

Algorithm 1 Random Tree Partition

```
Input: Depth of the random tree $p$;
       Initial partition $\pi_0 = \{A_0^1 := B_r\}$.
for $i = 1$ to $p$ do
  for $j = 1$ to $2^{i-1}$ do
    For rectangular cell $A_{i-1}^j$, randomly choose one dimension coordinate $Z_{i,j}$ whose probability distribution is given by (1);
    Divide the cell $A_{i-1}^j$ into two subregions, that is, $A_{i-1}^j = A_i^{2j-1} \cup A_i^{2j}$, along the midpoint of the dimension $Z_{i,j}$;
  end for
  Get $\pi_i = \{A_i^j, 1 \leq j \leq 2^i\}$.
end for
Output: Random tree partition $\pi_p$.
```

For any $x \in B_r$, there exists $j \in \mathcal{I}_p$ such that $x \in A_p^j$. Then we denote the cell containing x as $A_p(x) := A_p^j$.

2.3. Random Tree Density Estimation

In this subsection, we introduce the *random tree density estimation* (RTDE) based on the above mentioned random tree partition π_p. According to the random tree partition rule, for all $j \in \mathcal{I}_p$, the Lebesgue measure $\mu(A_p^j) > 0$.

Definition 1 (Random Tree Density Estimation) *Let* Q *be a probability measure on* $\mathbb{R}^d$*. Let* $\pi_p := \{A_p^j\}_{j \in \mathcal{I}_p}$ *be a random tree with depth p as in* (2)*. Then, the function* $f_{\mathrm{Q}}^p : \mathbb{R}^d \to [0, \infty)$ *defined by*

$$f_{\mathrm{Q}}^p(x) := \sum_{j \in \mathcal{I}_p} \frac{\mathrm{Q}(A_p^j) \mathbf{1}_{A_p^j}(x)}{\mu(A_p^j)}.$$

is called a random tree density estimation of Q.

Recalling that P is a probability measure on $\mathbb{R}^d$ with the corresponding density function f, by taking $\mathrm{Q} = \mathrm{P}$ with $d\mathrm{P} = f\, d\mu$, then for $x \in A_p^j$, we have

$$f_{\mathrm{P}}^p(x) = \frac{\mathrm{P}(A_p^j)}{\mu(A_p^j)} = \frac{1}{\mu(A_p^j)} \int_{A_p^j} f(x')\, d\mu(x'). \tag{3}$$

In other words, for $x \in A_p^j$, then $f_{\mathrm{P}}^p(x)$ is the average true density on A_p^j. Since P is inaccessible, in order to obtain the random tree density estimator, we take Q to be the empirical measure $\mathrm{D}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ instead of P, where δ_{x_i} denotes the Dirac distribution. For a set $A \subset \mathbb{R}^d$, $\mathrm{D}_n(A)$ is the expectation of $\mathbf{1}_A$ with respect to D_n, which is $\mathrm{D}_n(A) := \mathbb{E}_{\mathrm{D}_n}\mathbf{1}_A = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_A(x_i)$. For $x \in A_p^j$, the random tree density estimator is

$$f_{\mathrm{D}_n}^p(x) = \frac{\mathrm{D}_n(A_p^j)}{\mu(A_p^j)} = \frac{1}{n\mu(A_p^j)} \sum_{i=1}^n \mathbf{1}_{A_p^j}(x_i) \tag{4}$$

where A_p^j can also be written as $A_p(x)$. The summation on the right-hand side of (4) counts the number of observations falling in A_p^j. From now on, for notational simplicity, we will suppress the subscript n of D_n and denote $\mathrm{D} := \mathrm{D}_n$, e.g., $f_{\mathrm{D}}^p := f_{\mathrm{D}_n}^p$.

2.4. Random Forest Density Estimation

In this subsection, we formulate the *random forest density estimation* (RFDE). Ensembles consisting of a combination of different estimators have been highly recognized as an effective technique to significantly improve the performance over a single estimator in the literature, which inspires us to apply them to our random tree density estimation. In our cases, we first train T RTDEs based on different random tree partitions, separately; once this is accomplished, the outputs of the individual estimators are combined to give the ensemble output for new data points. Here, we use the simplest possible combination mechanism by taking a uniform-weighted average.

To be specific, assume that $\{Z_{i,j}^t, 1 \leq i \leq p, 1 \leq j \leq 2^{i-1}, 1 \leq t \leq T\}$ is an i.i.d. sequence of selected coordinate to split drawn from the probability measure P_Z given by (1). For $1 \leq t \leq T$, given select coordinates $Z_t := \{Z_{i,j}^t, 1 \leq i \leq p, 1 \leq j \leq 2^{i-1}\}$, following Algorithm 1, we generate a partition $\pi_p^t := \{A_p^{j,t}\}_{j \in \mathcal{I}_p^t}$. Thus for $1 \leq t \leq T$, RTDE is defined by

$$f_{\mathrm{D},t}^p(x) := \sum_{j \in \mathcal{I}_p^t} \frac{\mathrm{D}(A_p^{j,t}) \mathbf{1}_{A_p^{j,t}}(x)}{\mu(A_p^{j,t})}.$$

Therefore, the random forest density estimation with T base learners can be presented as

$$f_{\mathrm{D,E}}(x) := \frac{1}{T} \sum_{t=1}^T f_{\mathrm{D},t}^p(x). \tag{5}$$

It is worth mentioning that RFDE enjoys three advantages. First, the ensemble procedure alleviates the discontinuity and brings smoothness to tree-based density estimators, thanks to the randomness of base learners. To be specific, RFDE average the random base learners with different partition boundaries. As the number of base learners T grows, RFDE turns out to approximate a smooth density function well. Consequently, it can be more smooth than its base learner, i.e. $T = 1$, which will also be theoretically verified in Section 3 and experimentally validated by numerical simulations in Section 4.2.2.

Second, the random tree partition is more efficient than the histogram partition, therefore random tree density estimators are able to cope with higher dimensional density estimation compared with HDE. In the ordinary histogram partition, the number of cells grows exponentially with both the depth of splits p and the dimension of data d. The total number of cells will be 2^{pd}, which causes a heavy computational burden when d is large. On the contrary, in the random tree partition, the number of cells, 2^p, is significantly smaller than 2^{pd}, even if $d = 2$.

Third, the algorithm can be locally adaptive by applying random partitions. Ordinary density estimators such as KDE adopt uniform bandwidth, regardless of the fact that the local structures of real-world data usually vary from area to area. On the contrary, it is well known that partition-based algorithms take local data structures into consideration, and thus the cells with different shapes exactly catch various local features of the input data. Thus, the combination of random trees with the ensemble procedure can lead to great local adaptivity.

3. Theoretical Results

In this section, we present main results on the convergence rates of our density estimators. We first introduce the fundamental Hölder continuous assumption for the density function f to achieve convergence rates in Section 3.1. Then the results concerning convergence rates of RFDE with $f \in C^{0,\alpha}$ are shown in Section 3.2. Moreover, when $f \in C^{1,\alpha}$, Section 3.3 establishes the upper bound of the excess risk for RFDE and the lower bound of that for RTDE, which theoretically explains the benefits of RFDE over RTDE.

3.1. Fundamental Assumption

Our theoretical analysis concerning convergence rate is built upon the fundamental assumption about the smoothness of the density function. To be more concrete, we assume the underlying density function f resides in the general function space $C^{k,\alpha}$ consisting of (k, α)-Hölder continuous functions. The definition is shown below.

Definition 2 *Let $r \in (0, \infty), k \in \mathbb{N} \cup \{0\}$ and $\alpha \in (0, 1]$. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is (k, α)-Hölder continuous, if there exists a finite constant $c_L > 0$ such that (i) $\|\nabla^\ell f\| \leq c_L$ for all $\ell \in \{1, \ldots, k\}$; (ii) $\|\nabla^k f(x) - \nabla^k f(x')\| \leq c_L \|x - x'\|^\alpha$ for all $x, x' \in \mathbb{R}^d$. The set of such functions is denoted by $C^{k,\alpha}$.*

We remark that k decides the order of smoothness for $f \in C^{k,\alpha}$ and larger k indicates that f enjoys a higher order of smoothness. For the special case $k = 0$, the function space $C^{0,\alpha}$ coincides with the commonly used α-Hölder continuous function space C^α.

3.2. Convergence Rates of RFDE for $f \in C^{0,\alpha}$

In the following theorem, we present the convergence rates of the RFDE estimators with respect to the L_2-norm. For this purpose, we first need to introduce the notation

$$\|f\|_{L_2(\nu)}^2 := \int_{B_r} \mathbb{E}_{\mathrm{P}_Z \otimes \mathrm{P}^n} f(x)^2 \, d\mu(x),$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

Theorem 1 *Let $f_{\mathrm{D,E}}$ be the RFDE estimator with T base learners as in (5). Suppose that the true density $f \in C^{0,\alpha}$ with support $\mathcal{X} \subset B_r$. For any $T \geq 1$, let (p_n) be the sequences defined by $p_n := d(d \log 2 + 1 - 4^{-\alpha})^{-1} \log n$. Then we have*

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 \lesssim n^{-\frac{1-4^{-\alpha}}{d \log 2 + 1 - 4^{-\alpha}}}, \qquad (6)$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

For the special case $T = 1$, RFDE reduces to base learner RTDE and thus Theorem 1 implies that RTDE also enjoys the same convergence rate $n^{-(1-4^{-\alpha})/(d \log 2 + 1 - 4^{-\alpha})}$. Thus we are not able to show the advantage of RFDE over RTDE from the perspective of the convergence rate when $f \in C^{0,\alpha}$.

3.3. Results for $f \in C^{1,\alpha}$

In the previous analysis of $f \in C^{0,\alpha}$, we show the convergence rates of RFDE under L_2-norm. However, we fail to show the discrepancy between tree estimators and forest estimator for $f \in C^{0,\alpha}$ in terms of convergence rates. Therefore, in this part, we turn to consider that the true density f resides in the subspace $C^{1,\alpha}$, which contains smoother functions. For $f \in C^{1,\alpha}$, we manage to show that RFDE exceeds RTDE in the sense of convergence rate.

3.3.1. Convergence Rates of RFDE for $f \in C^{1,\alpha}$

The following theorem gives an upper bound for the convergence rate of RFDE estimator.

Theorem 2 *Let $f_{\mathrm{D,E}}$ be the RFDE estimator with T base learners as in (5). Suppose that the true density $f \in C^{1,\alpha}$ with support $\mathcal{X} \subset B_r$. Let (p_n), (T_n) be the sequences defined by $p_n := d(1 + d\log 2)^{-1}\log n$, $T_n \gtrsim n^{\frac{1}{4+4d\log 2}}$. Then we have*

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 \lesssim n^{-\frac{1}{d\log 2+1}}, \tag{7}$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

Theorem 2 shows that the L_2-error decreases as T_n grows at first, and when T_n achieves a certain level, RFDE achieves the best convergence rate. Moreover, comparing with Theorem 1, when the underlying density function turns more smooth, RFDE achieves a faster convergence rate with $f \in C^{1,\alpha}$ than that with $f \in C^{0,\alpha}$, where a relatively large T_n helps the density estimator to achieve asymptotic smoothness.

3.3.2. Lower bound of RTDE for $f \in C^{1,\alpha}$

The next theorem gives the lower bound of convergence rate for tree estimators.

Theorem 3 *Let the random tree density estimator f_{D}^p be defined as in (4). Moreover, assume that $f \in C^{1,\alpha}$ with the compact support $\mathcal{X} \subset B_r$ and there exists a constant $\underline{c}_f \in (0,\infty)$ such that $\|\nabla f\| \geq \underline{c}_f$. Then we have*

$$\|f_{\mathrm{D}}^p - f\|_{L_2(\nu)}^2 \geq c_0 n^{\frac{\log(1-0.75/d)}{\log 2 - \log(1-0.75/d)}} \vee c_1, \tag{8}$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$, c_0 and c_1 are constants depending on r, d, $\underline{c}_f$ specified in the proof.

Theorem 3 gives a lower bound of the convergence rate for the tree estimator. In particular, when the dimension $d \to \infty$, the lower bound in (8) becomes $n^{-0.75/(0.75+d\log 2)}$. More importantly, by combining Theorem 2 and 3, we find that when $d \geq 2$, the lower bound in (8) for RTDE turns out to be larger than the upper bound in (7) for RFDE in the sense of convergence rate. This indicates that random forest converges faster than trees when the density function is smooth. Therefore, the results demonstrate that ensemble learning can enhance smoothness of tree regressors and thus alleviate the boundary discontinuity problem.

4. Numerical Experiments

4.1. Evaluation Criteria

Mean absolute error (*MAE*). The first criterion of evaluating the accuracy of density estimator is the mean absolute error, defined by *MAE*$(\widehat{f}) = \frac{1}{M}\sum_{j=1}^{M} |\widehat{f}(x_j) - f(x_j)|$, where $x_1, \ldots, x_M$ are test samples. It is used in synthetic data experiments where the true density function is known.

Average negative log-likelihood (*ANLL*). Another effective measure of estimation accuracy, especially when facing real data and the true density function is unknown, is the average negative log-likelihood, defined by *ANLL*$(\widehat{f}) = -\frac{1}{M}\sum_{j=1}^{M} \log \widehat{f}(x_j)$, where $\widehat{f}(x_j)$ represents the estimated probability density for the test sample x_j and M is the size of test samples. Note that the lower the *ANLL* is, the better estimation we obtain.

4.2. Empirical Understandings

In this part, we conduct simulations concerning RFDE for density estimation. Based on several synthetic datasets, we show the power of ensemble procedure through simulations, and we provide a possible explanation for the improvement in accuracy from the perspective of the asymptotic smoothness. Then we study how the hyper-parameter, the depth of split p, affects the estimation accuracy.

4.2.1. Synthetic Data Settings

We conduct the simulations on four different types of synthetic distributions, each with dimension $d \in \{2, 5, 7\}$, respectively. The premise of constructing data sets is that we assume that the components $X_i \sim f_i$, $i = 1 \ldots, d$, of the random vector $X = (X_1, \ldots, X_d)$ are independent of each other. To be specific, Types I and II represent density functions with bounded support and unbounded support, respectively. Finally, Type III represents the case where the marginal distributions of each dimension are not identical. More detailed descriptions and visual illustrations are shown in Section C.1 of the appendix.

In the following experiments, we generate $2,000$ and $10,000$ i.i.d samples as training and testing data respectively from each type of synthetic datasets, and each with dimension $d \in \{2, 5, 7\}$.

4.2.2. The Power of Ensemble

To show the behavior of T, we carry out the experiments with $T \in \{1, 5, 10, 20, 50, 100, 500, 1000\}$, and the hyper-parameter p are chosen by 3-fold cross-validation. We pick the optimal p from 1 to 15. For each T we repeat this procedure 10 times.

As can be seen in Figure 2, regardless of the dimension d, as T grows, the accuracy performance of RFDE (both *MAE* and *ANLL*) first enhances dramatically when T grows, but as T continues to grow, a steady state will be reached. This coincides with Theorem 2, where the convergence rate attains the optimum when T_n is greater than a certain value. A large number of base learners leads to a more accurate model but brings about the additional burden of computation.

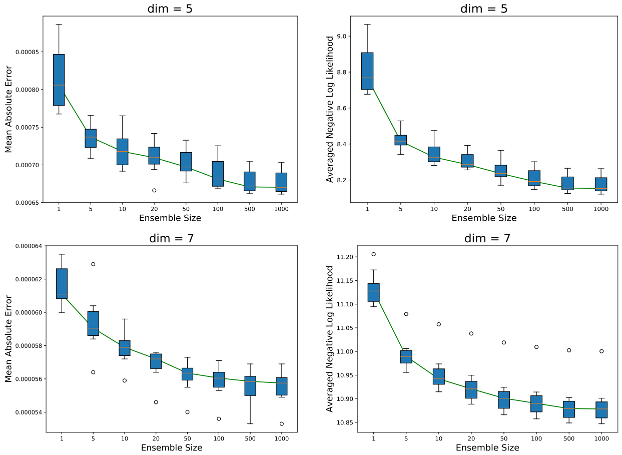
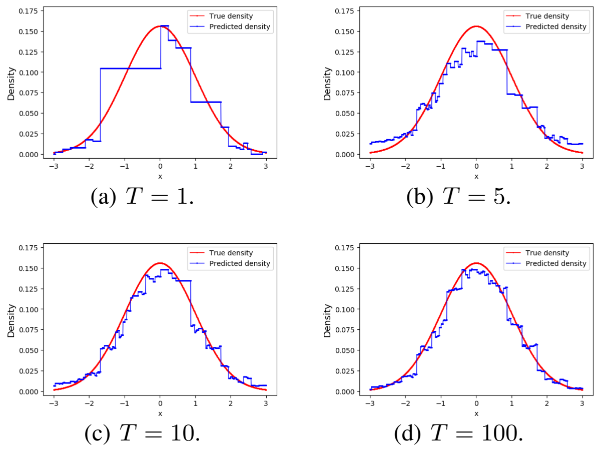


Figure 2. The study of parameter T on RFDE of Type II synthetic distribution, where the first and the second row respectively illustrate the results with dimension $d = 5$ and $d = 7$. The left column indicates how *MAE* varies along parameters T, and the right column shows the variation of *ANLL*.

To give a possible explanation of the improvement in accuracy with the ensemble procedure, we conduct simulations to show that RFDE achieves asymptotic smoothness with T increases. For the sake of clearer visualization, we utilize a toy example with $2,000$ samples i.i.d. generated from the two-dimensional standard normal distribution, and use RFDE to conduct density estimation, where the number of trees T is set to $1, 5, 10, 100$, respectively. To visualize the estimation of the 2-dimension density function, we fix the first coordinate $x_1 = 0.2$ and plot it with x_2 from -3 to 3.



(a) $T = 1$.

(b) $T = 5$.

(c) $T = 10$.

(d) $T = 100$.

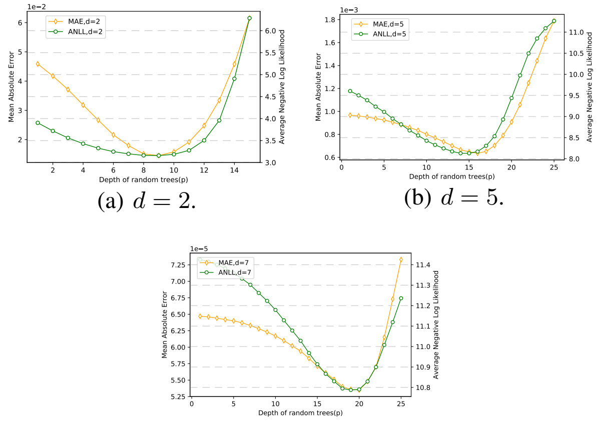
Figure 3. The study of parameter T on RFDE on a 2-dimension Standard Normal distribution. The red line represents the underlying density on the intersecting surface where $x_1 = 0.2$, and the blue line represents the density estimator returned by RFDE.

From Figure 3 we see that with $T = 1$, the base estimator turns out to be a step function with discontinuous boundaries, and the estimation is far from satisfactory. Nevertheless, as the number of base learners T increases from 1 to 10, the forest estimator becomes more continuous and smooth with the corresponding accuracy enhancing greatly. With $T = 10$, our RFDE is able to approximate the smooth density function well and achieve high estimation accuracy. If we continue to add more base learners to $T = 100$, there is no more significant improvement on the accuracy, which coincides with Theorem 2.

4.2.3. Parameter Analysis

Here we mainly conduct experiments concerning the parameter p of RFDE. To this end, we consider the Type II synthetic dataset of three different dimensions to see how the parameter p affects the performance of RFDE.

Recall that the larger p, the smaller and more cells, which is helpful to learn the local structure of the density function. However, if p is too large, few samples will be contained in each cell, which may lead to large variance. We conduct experiments over $p \in \{1, 2, \dots, 25\}$. We select $T = 500$ to make the density estimator converge with the sufficient base learners.



(a) $d = 2$.

(b) $d = 5$.

(c) $d = 7$.

Figure 4. The study of parameter p on RFDE of the Type II synthetic distribution. The green line and orange line represents the MAE and ANLL of RFDE, respectively.

As is shown in Figure 4, regardless of the dimension d, there exists an optimal value of p which minimizes both ANLL and MAE at the same time. In addition, the three subfigures in Figure 4 demonstrate that the optimal value of p becomes higher as d increases, which coincides the optimal order of p in Theorem 2. Therefore, it is of great importance to choose p properly.

Table 1. Average *ANLL* and *MAE* over simulated datasets

d	Method	Type I		Type II		Type III	
		ANLL	*MAE*	*ANLL*	*MAE*	*ANLL*	*MAE*
2	RFDE (Ours)	$\mathbf{-0.57}*$	$\mathbf{0.65}$	$\mathbf{3.14}*$	**1.64e-2**$*$	$\mathbf{1.97}*$	**3.29e-2**$*$
	KDE	-0.37	1.06	3.27	2.31e-2	2.14	5.32e-2
	HDE	-0.52	0.66	3.21	1.81e-2	2.01	3.82e-2
5	RFDE (Ours)	$\mathbf{-1.18}*$	$\mathbf{7.77}*$	$\mathbf{8.17}*$	**6.78e-4**$*$	$\mathbf{3.12}*$	$\mathbf{0.09}*$
	KDE	-0.32	12.40	8.65	8.27e-4	3.86	0.15
	HDE	10.17	19.70	10.77	1.33e-3	6.09	0.17
7	RFDE (Ours)	$\mathbf{-1.48}*$	$\mathbf{30.60}*$	$\mathbf{10.89}*$	**5.54e-5**$*$	$\mathbf{3.96}*$	$\mathbf{0.13}*$
	KDE	0.03	40.74	12.48	6.05e-5	5.16	0.18
	HDE	11.48	73.97	11.49	1.05e-4	9.88	0.20

* The best results are marked in **bold**. We use $*$ to represent that the best method is significantly better than the other compared methods.

4.3. Performance Comparisons

In this section, we conduct performance comparisons on both synthetic and real datasets. Recall that both our theoretical results (shown in Theorems 2 and 3) and empirical illustrations (shown in Figure 3) demonstrate that ensemble improves the performance of partition-based methods by enhancing the smoothness of the estimator. Therefore, we compare our RFDE with the kernel density estimator (KDE) which enjoys high order of smoothness. We also compare our RFDE with the histogram density estimator (HDE). We run HDE with the bin width of histogram chosen by Sturges' rule (Sturges, 1926).

4.3.1. Synthetic Data Comparisons

Following the experimental settings in Section 4.2, we conduct empirical comparisons between RFDE and the prevailing KDE and HDE to further demonstrate the desirable performance of RFDE on synthetic datasets. We apply the Wilcoxon signed-rank test (Wilcoxon, 1992) at the significance level $\alpha = 0.05$. Table 1 records average *ANLL* and *MAE* over simulation data sets for KDE, HDE and RFDE with $T = 500$. For higher dimensions $d = 5$ and $d = 7$, our RFDE always outperforms KDE and HDE in terms of *ANLL* and *MAE*.

4.3.2. Real Data Comparisons

We conduct numerical comparisons on real datasets from the UCI repository (Dua & Graff, 2017). We put the detailed description of datasets in Section C.2 of the appendix.

Experimental Settings. In order to evaluate the performance of density estimators on datasets with various dimensions, we apply the following data preprocessing pipeline. Firstly, we remove duplicate observations as well as those with missing values. Then each dimension of the datasets is scaled to $[0, 1]$ and each dataset is reduced to lower dimensions d' through PCA, e.g. to 10%, 30%, 50% and 70% of the original dimension d, respectively. Finally, in each dataset, we randomly select 70% of the samples for training and the remaining 30% for testing.

The number of iterations T is set to be 100 and the two hyper-parameters p are chosen from $\{1, 2, \ldots, 15\}$, respectively, by 3-fold cross-validation. We repeat this procedure 10 times to evaluate the standard deviation for *ANLL*. The average *ANLL* on test sets are recorded in Table 2.

Since real density often resides in a low-dimensional manifold instead of filling the whole high-dimensional space, it is reasonable to study the density estimation problem after dimensionality reduction. Therefore, in data preprocessing, all data sets are reduced to various lower dimensions through PCA. However, we need to take the to-be-reduced dimension as a hyper-parameter, since in general, the dimension of the manifold is unknown.

Experimental Results. In Table 2, we summarize the comparisons with the widely used density estimator KDE and HDE on six real datasets. For most of the redacted datasets, RFDE shows its superiority on the accuracy, whereas the standard deviation of RFDE is slightly larger than that of KDE due to the randomness of random tree partitions. Compared with HDE which corrupts when the redacted dimension $d' > 7$, our RFDE achieves the satisfying performance benefiting from its high computational efficiency.

4.4. RFDE for Anomaly Detection

To showcase a potential application of RFDE, we propose a density-based method for anomaly detection. Given a density level ρ, we regard the sample points with low density estimation $\{x_i \in D \mid f_{\mathrm{D,E}}(x_i) \leq \rho\}$ as anomaly points. Then we are able to use RFDE for anomaly detection as

Table 2. Average *ANLL* over real data sets

Datasets	d'	RFDE	KDE	HDE	Datasets	d'	RFDE	KDE	HDE
Adult	2	$-\mathbf{1.5226}$	-0.7402	-0.9838	Diabetes	1	$-\mathbf{0.8073}$	-0.2627	-0.6067
		(0.0113)	(0.0027)	(0.0143)			(0.0576)	(0.0111)	(0.0676)
	4	$-\mathbf{1.8374}$	-0.3075	-0.7789		3	$-\mathbf{1.5378}$	-0.4042	-0.3142
		(0.0141)	(0.0032)	(0.0303)			(0.0953)	(0.0403)	(0.3422)
	8	$-\mathbf{5.7832}$	-2.2970	–		4	$-\mathbf{1.8387}$	-0.8353	2.9933
		(0.0557)	(0.0108)	–			(0.1433)	(0.0773)	(0.6034)
	10	$-\mathbf{6.6704}$	-3.4372	–		6	$-\mathbf{2.3838}$	-1.9693	9.1732
		(0.0475)	(0.0110)	–			(0.1912)	(0.1550)	(0.3902)
Australian	2	$-\mathbf{0.5836}$	1.3155	0.3898	Credit	2	$\mathbf{1.2659}$	1.5435	1.6649
		(0.1796)	(0.0234)	(0.1494)			(0.1142)	(0.0183)	(0.1968)
	4	$-\mathbf{5.2131}$	0.8518	-2.2163		5	$-\mathbf{1.3479}$	1.4844	1.3455
		(0.3508)	(0.0291)	(0.2507)			(0.2889)	(0.0516)	(0.5457)
	8	$-\mathbf{3.6821}$	0.6879	–		8	$\mathbf{2.1191}$	3.0453	–
		(0.3678)	(0.1056)	–			(0.2905)	(0.1067)	–
	10	$-\mathbf{1.8187}$	0.4995	–		11	$\mathbf{3.1343}$	3.5221	–
		(0.3474)	(0.1748)	–			(0.3182)	(0.2292)	–
Breast-cancer	1	$-\mathbf{0.0323}$	0.6907	0.3697	Abalone	1	0.5664	$\mathbf{0.5458}$	0.5609
		(0.2059)	(0.0394)	(0.1011)			(0.0144)	(0.0103)	(0.0140)
	3	$-\mathbf{3.3262}$	0.1743	1.3773		3	$-\mathbf{2.6793}$	-0.9493	-1.2716
		(0.5219)	(0.1268)	(0.3432)			(0.0818)	(0.0282)	(0.0594)
	6	$-\mathbf{7.5657}$	-1.1397	1.8392		4	$-\mathbf{4.0743}$	-2.6572	-2.2145
		(0.9746)	(0.2788)	(0.5542)			(0.0619)	(0.0309)	(0.1534)
	8	$-\mathbf{5.1952}$	-2.1110	–		6	$-\mathbf{7.1922}$	-6.4804	0.3270
		(1.2260)	(0.3906)	–			(0.0722)	(0.0445)	(0.3553)

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis. The results of HDE with $d' > 7$ is corrupted due to numerical problems.

shown in Algorithm 2.

We conduct real-data experiments to compare our RFDE with several popular anomaly detection algorithms such as the forest-based Isolation Forest (iForest) (Liu et al., 2008), the distance-based k-Nearest Neighbor (k-NN) (Ramaswamy et al., 2000) and Local Outlier Factor (LOF) (Breunig et al., 2000), the kernel-based one-class SVM (OCSVM) (Schölkopf et al., 2001), the boosting-based Lump (Ridgeway, 2002), HDBSCAN (Campello et al., 2015) and the ensemble-based AOM+VR (Aggarwal & Sathe, 2015) on 20 real-world benchmark outlier detection datasets from the ODDS library. We perform ranking according to the best AUC when parameters go through their parameter grids. Detailed experimental settings and comparison results are shown in Section C.3.

Algorithm 2 RFDE for Anomaly Detection

```
Input: Training data $D := \{x_1, \dots, x_n\}$;
       Density threshold parameters $\rho$.
Compute RFDE $f_{\mathrm{D,E}}$ (5).
Output: Recognize anomalies as
        $\{x_i \in D \mid f_{\mathrm{D,E}}(x_i) \leq \rho\}$.
```

From the perspective of best performance, our method RFDE wins in 5 out of 20 datasets, while the iForest and OCSVM win both 4 out of 20 datasets, respectively. Moreover, in the aspect of the average performance of benchmark datasets, our RFDE has the lowest rank-sum 55 whereas the iForest has the second lowest rank-sum 72. Overall, our experiments on benchmark datasets show that our method has satisfying performance among competitive anomaly detection algorithms.

5. Conclusion

In this study, we propose the *random forest density estimator* (RFDE), constructed by generating random tree partitions, building tree estimators, and finally ensembling trees together to obtain the forest. We verify that RFDE alleviates the problem of boundary discontinuity from both the theoretical and experimental perspective. From the theoretical perspective, we prove fast convergence rates of RFDE under the assumption that the true density function is Hölder continuous. Moreover, to explain the benefits of ensemble learning in density estimation, we turn to consider more smooth density functions. We establish the upper bound of the excess risk for RFDE, which is strictly smaller than the lower bound of that for the tree base learners. In the aspect of experiment, we demonstrate that RFDE turns out to be more continuous as T grows and thus it achieves the asymptotic smoothness. Moreover, we conduct the experimental comparisons both on synthetic and real-world datasets. Last but not least, we carry out an application of anomaly detection compared with other widely used methods to show the promising performance of RFDE.

References

Aggarwal, C. C. and Sathe, S. Theoretical foundations and algorithms for outlier ensembles. *Acm sigkdd explorations newsletter*, 17(1):24–47, 2015.

Biau, G. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

Bodin, E., Dai, Z., Campbell, N., and Ek, C. H. Black-box density function estimation using recursive partitioning. In *International Conference on Machine Learning*, pp. 1015–1025. PMLR, 2021.

Breiman, L. Consistency for a simple model of random forests. 2004.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, pp. 93–104, 2000.

Campello, R. J., Moulavi, D., Zimek, A., and Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–51, 2015.

Campello, R. J., Kröger, P., Sander, J., and Zimek, A. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2): e1343, 2020.

Corizzo, R., Pio, G., Ceci, M., and Malerba, D. Dencast: Distributed density-based clustering for multi-target regression. *Journal of Big Data*, 6(1):1–27, 2019.

Criminisi, A. and Shotton, J. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Science & Business Media, 2013.

Criminisi, A., Shotton, J., and Konukoglu, E. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Technical Report 2011–114*, 2011.

Cui, J., Hang, H., Wang, Y., and Lin, Z. GBHT: Gradient boosting histogram transform for density estimation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2233–2243, 2021.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Emadi, H. S. and Mazinani, S. M. A novel anomaly detection algorithm using dbscan and svm in wireless sensor networks. *Wireless Personal Communications*, 98(2): 2025–2035, 2018.

Freedman, D. and Diaconis, P. On the histogram as a density estimator: L_2 theory. *Probability Theory and Related Fields*, 57(4):453–476, 1981.

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. *Nonparametric and Semiparametric Models*. Springer Science & Business Media, 2012.

Hu, S., Yu, T., Guo, C., Chao, W.-L., and Weinberger, K. Q. A new defense against adversarial images: Turning a weakness into a strength. *Advances in Neural Information Processing Systems*, 32, 2019.

Huang, Y.-T., Liao, W.-H., and Huang, C.-W. Defense mechanism against adversarial attacks using density-based representation of images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3499–3504. IEEE, 2021.

Klemelä, J. Multivariate histograms with data-dependent partitions. *Statistica Sinica*, 19(1):159–176, 2009.

Li, D., Yang, K., and Wong, W. H. Density estimation via discrepancy based adaptive sequential partition. In *Advances in neural information processing systems*, pp. 1091–1099, 2016.

Li, H., Liu, X., Li, T., and Gan, R. A novel density-based clustering algorithm using nearest neighbor graph. *Pattern Recognition*, 102:107206, 2020.

Li, W., Yongbo, L., and Xiangyang, X. Coda: Counting objects via scale-aware adversarial density adaption. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 193–198. IEEE, 2019.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 413–422, 2008.

Liu, L. and Wong, W. H. Multivariate density estimation via adaptive partitioning (I): sieve MLE. *arXiv preprint arXiv:1401.2597*, 2014.

López-Rubio, E. A histogram transform for probability density function estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):644–656, 2013.

Maldonado, S., Merigó, J., and Miranda, J. Iowa-svm: A density-based weighting strategy for svm classification via owa operators. *IEEE Transactions on Fuzzy Systems*, 28(9):2143–2150, 2019.

Parzen, E. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

Ram, P. and Gray, A. G. Density estimation trees. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 627–635. ACM, 2011.

Ramaswamy, S., Rastogi, R., and Shim, K. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 427–438, 2000.

Ridgeway, G. Looking for lumps: Boosting and bagging for density estimation. *Computational Statistics & Data Analysis*, 38(4):379–392, 2002.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pp. 832–837, 1956.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7): 1443–1471, 2001.

Silverman, B. W. *Density estimation for statistics and data analysis*. Routledge, 2018.

Steininger, M., Kobs, K., Davidson, P., Krause, A., and Hotho, A. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Sturges, H. A. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

Tumiran, S. A. and Sivakumar, B. Community structure concept for catchment classification: A modularity density-based edge betweenness (mdeb) method. *Ecological Indicators*, 124:107346, 2021.

Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pp. 196–202. Springer, 1992.

Zhang, L., Lin, J., and Karim, R. Adaptive kernel density-based anomaly detection for nonlinear systems. *Knowledge-Based Systems*, 139:50–63, 2018.

Zhao, L. and Shi, G. Maritime anomaly detection using density-based clustering and recurrent neural network. *The Journal of Navigation*, 72(4):894–916, 2019.

This appendix consists of supplementaries for both theoretical analysis and experiments. In Section A, we prove the approximation error and estimation error term for the underlying density function residing in space $C^{0,\alpha}$ and $C^{1,\alpha}$, respectively. The corresponding proofs of Section A and Section 3 are shown in Section B. In Section C we show the supplementaries for numerical experiments including the results of anomaly detection.

A. Error Analysis

In this subsection, we present the proofs when the true density $f \in C^{1,\alpha}$. To make the bias-variance decomposition, we introduce some notations. This section provides a more comprehensive error analysis for the theoretical results in Section 3. To be specific, we first present the approximation error and sample error of RFDE under the assumption that the density function resides in the Hölder spaces $C^{0,\alpha}$ in Section A.1. Then for $f \in C^{1,\alpha}$, we gives the upper bound of the two error terms for RFDE based on the new bias-variance error decomposition in Section A.2.1 and the lower bound of them for RTDE in Section A.2.2.

A.1. Error Analysis for $f \in C^{0,\alpha}$

To make the bias-variance decomposition, we first introduce the population version of $f_{\mathrm{D,E}}$ in (5). For fixed $p \in \mathbb{N}_+$, let $\{A_{p,t}\}_{t=1}^T$ be random tree partitions with depth p and split coordinates Z_t, $t = 1, \ldots, T$. Moreover, let $\{f_{\mathrm{P},t}^p\}_{t=1}^T$ be defined as (3), then we define the population version of the RFDE by

$$f_{\mathrm{P,E}}(x) := \frac{1}{T} \sum_{t=1}^{T} f_{\mathrm{P},t}^p(x) \tag{A.1}$$

A.1.1. Bounding the Approximation Error Term

Proposition 1 *Let $f_{\mathrm{P,E}}$ be defined by (A.1). Moreover, let the density function $f \in C^{0,\alpha}$ and P_X has the bounded support $\mathcal{X} \subset B_r$. Then we have*

$$\|f_{\mathrm{P,E}} - f\|_{L_2(\nu)}^2 \leq c_L^2 (2r)^{2\alpha} d \exp\bigl((2^{-2\alpha} - 1)p/d\bigr),$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

A.1.2. Bounding the Estimation Error Term

We firstly present a fundamental lemma, which shows that the $\|\cdot\|_2$-distance between $f_{\mathrm{D,E}}$ and $f_{\mathrm{P,E}}$.

Proposition 2 *Let $f_{\mathrm{D,E}}$ and $f_{\mathrm{P,E}}$ be defined by (4) and (A.1) respectively. Moreover, let P_X has the bounded support $\mathcal{X} \subset B_r$. Then we have*

$$\|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\nu)}^2 \leq \|f\|_\infty \cdot 2^p/n,$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

A.2. Error Analysis for $f \in C^{1,\alpha}$

In this subsection, we present the proofs when the true density $f \in C^{1,\alpha}$. A drawback to the analysis in $C^{0,\alpha}$ is that the usual Taylor expansion involved techniques for error estimation may not apply directly. As a result, we fail to prove the exact benefits of the ensemble procedure. Therefore, in this subsection, we turn to the function space $C^{1,\alpha}$ consisting of smoother functions. To be specific, we study the convergence rates of $f_{\mathrm{D},B}$ to the density function $f \in C^{1,\alpha}$. To this end, there is a point in introducing some notations.

Then it is clear to have the following error decomposition,

$$\begin{aligned}\mathbb{E}_{\mathrm{P}_Z}\left(f_{\mathrm{P,E}}(x) - f(x)\right)^2 &= \mathrm{Var}_{\mathrm{P}_Z}\left(f_{\mathrm{P,E}}(x)\right) + \left(\mathbb{E}_{\mathrm{P}_Z}\left(f_{\mathrm{P,E}}(x)\right) - f(x)\right)^2 \\ &= \frac{1}{T} \cdot \mathrm{Var}_{\mathrm{P}_Z}\bigl(f_{\mathrm{P},1}^p(x)\bigr) + \bigl(\mathbb{E}_{\mathrm{P}_Z}\bigl(f_{\mathrm{P},1}^p(x)\bigr) - f(x)\bigr)^2.\end{aligned}$$

In particular, for the random tree density estimator, we are concerned with the lower bound for f_{D}^p. We make the error decomposition

$$\begin{aligned}\mathbb{E}_{\mathrm{P}^n \otimes \mathrm{P}_Z} \|f_{\mathrm{D}}^p - f\|_{L_2(\mu)}^2 &= \mathbb{E}_{\mathrm{P}_Z \otimes \mathrm{P}^n} \big\|f_{\mathrm{D}}^p - f_{\mathrm{P}}^p + f_{\mathrm{P}}^p - f\big\|_{L_2(\mu)}^2 \\ &= \mathbb{E}_{\mathrm{P}_Z \otimes \mathrm{P}^n} \big\|f_{\mathrm{D}}^p - f_{\mathrm{P}}^p\big\|_{L_2(\mu)}^2 + \mathbb{E}_{\mathrm{P}_Z \otimes \mathrm{P}^n} \big\|f_{\mathrm{P}}^p - f\big\|_{L_2(\mu)}^2 + \mathbb{E}_{\mathrm{P}_Z \otimes \mathrm{P}^n} \int_{B_r} 2\big(f_{\mathrm{D}}^p - f_{\mathrm{P}}^p\big)\big(f_{\mathrm{P}}^p - f\big)\, d\mu.\end{aligned}$$

The last term equal to 0 by exchanging the order of integration. Consequently, we get

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 = \|f_{\mathrm{P,E}} - f\|_{L_2(\nu)}^2 + \|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\nu)}^2. \tag{A.2}$$

It is important to note that both of the two terms on the right-hand side are data- and partition-independent due to the expectation with respect to P^n and P_Z respectively. Loosely speaking, the first error term corresponds to the expected estimation error of the estimator f_{D}^p, while the second one demonstrates the expected approximation error.

A.2.1. Upper Bound for Convergence Rate of RFDE

Proposition 3 and Proposition 4 gives upper bounds for the approximation and estimation error terms of the forest estimator, respectively.

Proposition 3 *Let $f_{\mathrm{P,E}}$ be defined by (A.1). Moreover, let the density function $f \in C^{1,\alpha}$ and P_X has the bounded support $\mathcal{X} \subset B_r$. Then we have*

$$\|f_{\mathrm{P,E}} - f(x)\|_{L_2(\nu)}^2 \leq c_L^2 (2r)^4 d^2 T^{-1} \exp\big(-0.75p/d\big) + 4c_L^2 (2r)^{2d+2} d^2 \exp(-p/d),$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

Proposition 4 *Let $f_{\mathrm{D,E}}$ and $f_{\mathrm{P,E}}$ be defined by (5) and (A.1) respectively. Moreover, let P_X has the bounded support $\mathcal{X} \subset B_r$. Then we have*

$$\|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\nu)}^2 \leq \|f\|_\infty \cdot 2^p/n,$$

where $\nu := \mu \otimes \mathrm{P}_Z \otimes \mathrm{P}^n$.

A.2.2. Lower Bound for Convergence Rate of RTDE

Proposition 5 and 6 gives lower bounds for the approximation and estimation error terms of the tree estimator, respectively.

Proposition 5 *Let the random tree partition A_p be defined as in Algorithm 1. Moreover, let the density function $f \in C^{1,\alpha}$ with support $\mathcal{X} \subset B_r$. Furthermore, suppose that there exists a fixed constant $\underline{c}_f \in (0, \infty)$ such that $\|\nabla f\| \geq \underline{c}_f$. Then we have*

$$\|f_{\mathrm{P}}^p - f\|_{L_2(\mathrm{P}_Z \otimes \mu)}^2 \geq 0.75 \underline{c}_f^2 r^2 d (2r)^d (-0.75/d)^p.$$

Proposition 6 *Let the random tree partition A_p be defined as in Algorithm 1 with depth $p \geq \ln(\|f\|_\infty 2^{d+1} r^d)/\log 2$. Moreover, suppose that P_X has the compact support $\mathcal{X} \subset B_r$. Then we have*

$$\|f_{\mathrm{D}}^p - f_{\mathrm{P}}^p\|_{L_2(\mathrm{P}_Z \otimes \mu)}^2 \geq 2^p/(2^{d+1} r^d n).$$

B. Proofs

This section consists of four parts, with the first sections concerning fundamental lemmas on properties of random tree and the following two sections showing the proof related to the results for the space $C^{0,\alpha}$ and $C^{1,\alpha}$ respectively. The last one presents the proof related to the main theoretical results. To be specific, Section B.1 presents the properties related to the mid-point splitting rule. Section B.2 and B.3 present all proofs related to the space $C^{0,\alpha}$ and $C^{1,\alpha}$, respectively. The proofs related to Section 3 are presented in Section B.4.

B.1. Properties of Random Tree

Throughout the proof of this paper, we will make repeated use of the following two facts proposed by (Biau, 2012).

Fact A.1 *For $x \in B_r$, let $A_p(x)$ defined by (2) be the rectangular cell of the random tree containing x and $S_p^j(x)$ be the number of times that $A_p(x)$ is split on the j-th coordinate ($j = 1, \ldots, d$). Then $S_p(x) := (S_p^1(x), \ldots, S_p^j(x))$ has multi-nomial distribution with parameters p and probability vector $(1/d, \ldots, 1/d)$ and satisfies $\sum_{j=1}^d S_p^j(x) = p$. Moreover, let $A_p^j(x)$ be the size of the j-th dimension of $A_p(x)$. Then we have*

$$A_p^j(x)|R \overset{\mathcal{D}}{=} 2r \cdot 2^{-S_p^j(x)}, \tag{A.3}$$

where $\cdot|R$ denotes the probability distribution and $\overset{\mathcal{D}}{=}$ indicates that variables in the two sides of the equation have the same distribution.

Fact A.2 *Let μ be the Lebesgue measure. For $x \in B_r$, let $N_p(x)$ be the number of samples falling in the same cell as x, that is, $N_p(x) = \sum_{i=1}^n \mathbf{1}_{\{X_i \in A_p(x)\}}$. By construction, we have*

$$\mu(A_p(x)) = (2r)^d \cdot 2^{-p}. \tag{A.4}$$

Before we proceed, we present the following lemma, which helps to bound the diameter of the rectangular cell $A_p(x)$.

Lemma A.1 *Suppose that $x_i > 0$, $1 \leq i \leq d$ and $0 < \alpha \leq 1$. Then we have*

$$\left(\sum_{i=1}^d x_i\right)^\alpha \leq \sum_{i=1}^d x_i^\alpha. \tag{A.5}$$

Proof A.1 (Proof of Lemma A.1) *Obviously, for any $1 \leq i \leq n$, we have $0 < x_i / \sum_{i=1}^d x_i < 1$. Since $0 < \alpha \leq 1$, we have*

$$\frac{\sum_{i=1}^d x_i^\alpha}{(\sum_{i=1}^d x_i)^\alpha} = \sum_{i=1}^d \left(\frac{x_i}{\sum_{i=1}^d x_i}\right)^\alpha \geq \sum_{i=1}^d \frac{x_i}{\sum_{i=1}^d x_i} = \frac{\sum_{i=1}^d x_i}{\sum_{i=1}^d x_i} = 1.$$

Consequently, we get $\left(\sum_{i=1}^d x_i\right)^\alpha \leq \sum_{i=1}^d x_i^\alpha$, which finishes the proof.

Combining Lemma A.1 with Fact A.2, it is easy to derive the following lemma which plays an important role to bound the approximation error of the estimator.

Lemma A.2 *Let the diameter of the set $A \subset \mathbb{R}^d$ be defined by $\mathrm{diam}(A) := \sup_{x,x' \in A} \|x - x'\|_2$. Then for any $x \in \mathcal{X}$ and $0 < \beta \leq 2$, there holds*

$$\mathbb{E}_{\mathrm{P}_Z}\big(\mathrm{diam}(A_p(x))^\beta\big) \leq (2r)^\beta d \exp\big((2^{-\beta} - 1)p/d\big).$$

Proof A.2 (Proof of Lemma A.2) *By definition, we have $\mathrm{diam}(A_p(x)) := \big(\sum_{j=1}^d A_p^j(x)^2\big)^{1/2}$. Consequently, (A.3) in Fact A.1 implies $\mathrm{diam}(A_p(x))^\beta = (2r)^\beta \big(\sum_{j=1}^d 2^{-2S_p^j(x)}\big)^{\beta/2}$. Applying Lemma A.1, we get*

$$\mathrm{diam}(A_p(x))^\beta \leq (2r)^\beta \sum_{j=1}^d 2^{-\beta S_p^j(x)}. \tag{A.6}$$

Consequently, we have

$$\begin{aligned}\mathbb{E}_{\mathrm{P}_Z}\big(\mathrm{diam}(A_p(x))^\beta\big) \leq \mathbb{E}_{\mathrm{P}_Z}\left((2r)^\beta \sum_{j=1}^d 2^{-\beta S_p^j(x)}\right) &= (2r)^\beta \sum_{j=1}^d \mathbb{E}_{\mathrm{P}_Z}\big(2^{-\beta S_p^j(x)}\big) \\ &= (2r)^\beta d\big(1 - (1 - 2^{-\beta})/d\big)^p \leq (2r)^{2\alpha} d \exp\big((2^{-\beta} - 1)p/d\big).\end{aligned}$$

Taking expectation with respect to P_R, we prove the desired assertion.

For any $x \in B_r$, let $\underline{a}_p^j(x)$ and $\overline{a}_p^j(x)$ be the minimum and maximum values of the j-th entries of points in $A_p(x)$. Then, by the construction of random tree, we have $A_p(x) = [\underline{a}_p^1(x), \overline{a}_p^1(x)] \times \cdots \times [\underline{a}_p^d(x), \overline{a}_p^d(x)]$.

The next theorem gives an explicit form of the distance between x_i and the center of the interval $[\underline{a}_p^j(x), \overline{a}_p^j(x)]$, which is used to derive the lower bound for the error of single random tree density estimation.

Lemma A.3 *Let the random tree A_p be defined as in Algorithm 1. Moreover, let $A_p(x)$ be the rectangular cell containing x and $S_p^j(x)$ be the number of times that $A_p(x)$ is split on the j-th coordinate ($j = 1, \ldots, d$). For any $x \in B_r$, let x_j be the j-th entry of x. If $S_p^j(x) = k$, $0 \leq k \leq q$, then we have*

$$\bigl|x_j - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr| = \min_{q \in Q_k} |x_j - q|,$$

where $Q_k := \bigl\{r(2i-1)/2^k \bigm| -2^{k-1} + 1 \leq i \leq 2^{k-1}\bigr\}$.

Proof A.3 (Proof of Lemma A.3) *If $S_p^j(x) = k$, by the construction of random tree partition, we have*

$$\bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2 \in Q_k. \tag{A.7}$$

By the definition of Q_k, for any $q^* \in Q_k$, there holds $\bigl|q^* - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr| \geq r/2^{k-1}$. Since $x \in A_p(x)$, we have

$$\bigl|x_j - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr| \leq r/2^k. \tag{A.8}$$

Therefore, using the triangular inequality, we obtain

$$|x_j - q^*| \geq \bigl|\bigl|x_j - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr| - \bigl|q^* - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr|\bigr| \geq r/2^k.$$

This together with (A.8) implies that $|x_j - q^*| \geq \bigl|x_j - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr|$ holds for any $q^* \in Q_k$. Combining this with (A.7), we get

$$\bigl|x_j - \bigl(\underline{a}_p^j(x) + \overline{a}_p^j(x)\bigr)/2\bigr| = \min_{q \in Q_k} |x_j - q|,$$

which leads to the desired assertion.

B.2. Proof of Results for $f \in C^{0,\alpha}$

In this subsection, we present the proofs related to Section 3.2 and Section A.1, where the true density $f \in C^{0,\alpha}$.

B.2.1. Proof Related to Section A.1.1

Proof A.4 (Proof of Proposition 1) *By Cauchy-Schwarz inequality and the fact $f_{\mathrm{P},t}^p(x)$ are i.i.d, there holds*

$$\|f_{\mathrm{P,E}} - f\|_{L_2(\nu)}^2 = \biggl\|\frac{1}{T}\sum_{t=1}^T \bigl(f_{\mathrm{P},t}^p - f\bigr)\biggr\|_{L_2(\nu)}^2 \leq \frac{1}{T}\sum_{t=1}^T \|f_{\mathrm{P},t}^p - f\|_{L_2(\nu)}^2 = \|f_{\mathrm{P}}^p - f\|_{L_2(\nu)}^2. \tag{A.9}$$

The assumption $f \in C^{0,\alpha}$ implies that for any $x \in B_r$, there holds

$$\begin{aligned}\mathbb{E}_{\mathrm{P}_Z}\bigl(f_{\mathrm{P}}^p(x) - f(x)\bigr)^2 &= \mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{1}{\mu(A_p(x))}\int_{A_p(x)} f(x')\,dx' - f(x)\biggr)^2 = \mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{1}{\mu(A_p(x))}\int_{A_p(x)} \bigl(f(x') - f(x)\bigr)\,dx'\biggr)^2 \\ &\leq \mathbb{E}_{\mathrm{P}_Z}\bigl(c_L \mathrm{diam}\bigl(A_p(x)\bigr)^\alpha\bigr)^2 = c_L^2 \mathbb{E}_{\mathrm{P}_Z} \mathrm{diam}\bigl(A_p(x)\bigr)^{2\alpha}.\end{aligned} \tag{A.10}$$

According to Lemma A.2, (A.10) can be further bounded by

$$\mathbb{E}_{\mathrm{P}_Z}\bigl(f_{\mathrm{P}}^p(x) - f(x)\bigr)^2 \leq c_L^2 (2r)^{2\alpha} d \exp\bigl((2^{-2\alpha} - 1)p/d\bigr).$$

Taking expectation with respect to P_X, we have

$$\|f_{\mathrm{P}}^p - f\|_{L_2(\nu)}^2 \leq c_L^2 (2r)^{2\alpha} d \exp\bigl((2^{-2\alpha} - 1)p/d\bigr).$$

This together with (A.9) yields the assertion.

B.2.2. Proof Related to Section A.1.2

Proof A.5 (Proof of Proposition 2) *By Cauchy-Schwarz inequality and the fact that $f^p_{\mathrm{D},t}(x) - f^p_{\mathrm{P},t}(x)$ are i.i.d, there holds*

$$\|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|^2_{L_2(\nu)} = \Big\|\frac{1}{T}\sum_{t=1}^{T}\big(f^p_{\mathrm{D},t} - f^p_{\mathrm{P},t}\big)\Big\|^2_{L_2(\nu)} \leq \frac{1}{T}\sum_{t=1}^{T}\big\|f^p_{\mathrm{D},t} - f^p_{\mathrm{P},t}\big\|^2_{L_2(\nu)} = \big\|f^p_{\mathrm{D}} - f^p_{\mathrm{P}}\big\|^2_{L_2(\nu)}. \tag{A.11}$$

It is clear to see that

$$\mathbb{E}_{\mathrm{P}_Z}\mathbb{E}_{\mathrm{P}^n}\big(f^p_{\mathrm{D}}(x) - f^p_{\mathrm{P}}(x)\big)^2 = \mathbb{E}_{\mathrm{P}_Z}\frac{\mathrm{P}\big(A^j_p(x)\big)\big(1 - \mathrm{P}(A^j_p(x))\big)}{n\mu^2(A^j_p(x))} \leq \mathbb{E}_{\mathrm{P}_Z}\frac{\mathrm{P}\big(A^j_p(x)\big)}{n\mu^2(A^j_p(x))} = \mathbb{E}_{\mathrm{P}_Z}\sum_{j=0}^{2^p}\frac{\mathrm{P}\big(A^j_p\big)}{n\mu^2(A^j_p)}\cdot\mathbf{1}_{A^j_p}(x).$$

Fubini's theorem implies

$$\int_{B_r}\mathbb{E}_{\mathrm{P}_{R,Z}\otimes\mathrm{P}^n}\big(f^p_{\mathrm{D}}(x) - f^p_{\mathrm{P}}(x)\big)^2\,d\mu(x) = \mathbb{E}_{\mathrm{P}_Z}\int_{B_r}\sum_{j=0}^{p}\frac{\mathrm{P}\big(A^j_p\big)}{n\mu^2(A^j_p)}\mathbf{1}_{A^j_p}(x)\,d\mu(x) = \mathbb{E}_{\mathrm{P}_Z}\sum_{j=0}^{2^p}\frac{\mathrm{P}\big(A^j_p\big)}{n\mu(A^j_p)} \leq \|f\|_\infty\cdot\frac{2^p}{n}.$$

In other words, we have $\|f^p_{\mathrm{D}} - f^p_{\mathrm{P}}\|^2_{L_2(\nu)} \leq \|f\|_\infty\cdot 2^p/n$, where $\nu := \mu\otimes\mathrm{P}_Z\otimes\mathrm{P}^n$. This together with (A.11) yields the assertion.

B.3. Proof of Results for $f \in C^{1,\alpha}$

B.3.1. Proof Related to Section A.2.1

The next proposition presents the upper bound of the L_2-distance between the random forest density estimation $f_{\mathrm{P,E}}$ and the density function f in the Hölder space $C^{1,\alpha}$.

Proof A.6 (Proof of Proposition 3) *According to the random tree splitting rule, the split coordinates $\{Z_t\}_{t=1}^T$ are i.i.d. Therefore, for any $x \in B_r$, the expected approximation error term can be decomposed as follows:*

$$\begin{aligned}&\mathbb{E}_{\mathrm{P}_Z}\big(f_{\mathrm{P,E}}(x) - f(x)\big)^2 = \mathbb{E}_{\mathrm{P}_Z}\big((f_{\mathrm{P,E}}(x) - \mathbb{E}_{\mathrm{P}_Z}(f_{\mathrm{P,E}}(x))) + \mathbb{E}_{\mathrm{P}_Z}(f_{\mathrm{P,E}}(x)) - f(x))\big)^2\\ &= \mathrm{Var}_{\mathrm{P}_Z}(f_{\mathrm{P,E}}(x)) + (\mathbb{E}_{\mathrm{P}_Z}(f_{\mathrm{P,E}}(x)) - f(x))^2 = T^{-1}\cdot\mathrm{Var}_{\mathrm{P}_Z}(f^p_{\mathrm{P},1}(x)) + \big(\mathbb{E}_{\mathrm{P}_Z}(f^p_{\mathrm{P},1}(x)) - f(x)\big)^2.\end{aligned} \tag{A.12}$$

For the first term in (A.12), we have

$$\begin{aligned}\mathrm{Var}_{\mathrm{P}_Z}\big(f^p_{\mathrm{P}}(x)\big) &= \mathbb{E}_{\mathrm{P}_Z}\big(f^p_{\mathrm{P}}(x) - \mathbb{E}_{\mathrm{P}_Z}(f^p_{\mathrm{P}}(x))\big)^2 \leq \mathbb{E}_{\mathrm{P}_Z}\big(f^p_{\mathrm{P}}(x) - f(x)\big)^2 = \mathbb{E}_{\mathrm{P}_Z}\bigg(\frac{1}{\mu(A_p(x))}\int_{A_p(x)}f(x')\,dx' - f(x)\bigg)^2\\ &= \mathbb{E}_{\mathrm{P}_Z}\bigg(\frac{1}{\mu(A_p(x))}\int_{A_p(x)}\big(f(x') - f(x)\big)\,dx'\bigg)^2 \leq \mathbb{E}_{\mathrm{P}_Z}\big(c_L\mathrm{diam}\big(A_p(x)\big)\big)^2.\end{aligned} \tag{A.13}$$

According to Lemma A.2, the first term is further bounded by

$$\mathrm{Var}_{\mathrm{P}_Z}\big(f^p_{\mathrm{P}}(x)\big) \leq c_L^2(2r)^4 d\exp(-0.75p/d). \tag{A.14}$$

We now consider the second term in (A.12). Taking the first-order Taylor expansion of $f(x')$ at x, we get

$$f(x') - f(x) = \int_0^1\big(\nabla f(x + t(x' - x))\big)^\top(x' - x)\,dt. \tag{A.15}$$

Therefore, the assumption $f \in C^{1,\alpha}$ implies

$$\begin{aligned}\big|f(x') - f(x) - \nabla f(x)^\top(x' - x)\big| &= \bigg|\int_0^1\big(\nabla f(x + t(x' - x)) - \nabla f(x)\big)^\top(x' - x)\,dt\bigg|\\ &\leq \int_0^1 c_L(t\|x' - x\|_2)^\alpha\|x' - x\|_2\,dt \leq c_L\|x' - x\|^{1+\alpha}.\end{aligned}$$

Now, by the triangle inequality, we have

$$\begin{aligned}\biggl|\mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{1}{\mu(A_p(x))}\int_{A_p(x)}(f(x')-f(x))dx'\biggr)\biggr| - \biggl|\mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{1}{\mu(A_p(x))}\int_{A_p(x)}\nabla f(x)^\top(x'-x)dx'\biggr)\biggr|& \\ \leq \biggl|\mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{1}{\mu(A_p(x))}\int_{A_p(x)}(f(x')-f(x)-\nabla f(x)^\top(x'-x))dx'\biggr)\biggr|& \\ \leq \mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{c_L}{\mu(A_p(x))}\int_{A_p(x)}\|x'-x\|^{1+\alpha}dx'\biggr) \leq c_L\mathbb{E}_{\mathrm{P}_Z}(\mathrm{diam}(A_p(x))^{1+\alpha}).&\end{aligned}$$

Then we get

$$\bigl|\mathbb{E}_{\mathrm{P}_Z}(f_{\mathrm{P},1}^{p}(x))-f(x)\bigr| \leq \biggl|\mathbb{E}_{\mathrm{P}_Z}\biggl(\frac{1}{\mu(A_p(x))}\int_{A_p(x)}\nabla f(x)^\top(x'-x)dx'\biggr)\biggr| + c_L\mathbb{E}_{\mathrm{P}_Z}(\mathrm{diam}(A_p(x))^{1+\alpha}). \tag{A.16}$$

Since $\|\nabla f\| \leq c_L$, we find

$$\biggl|\int_{A_p(x)}\nabla f(x)^\top(x'-x)dx'\biggr| \leq c_L\sum_{j=1}^{d}\biggl|\int_{A_p(x)}(\tilde{x}'_j-\tilde{x}_j)d\tilde{x}'\biggr|, \tag{A.17}$$

where $\tilde{x}_j$ and $\tilde{x}'_j$ denote the j-th entries of the vectors $\tilde{x}$ and $\tilde{x}'$ respectively.

Recall that $\underline{a}_p^j(x)$ and $\overline{a}_p^j(x)$ denotes the minimum and maximum values of the j-th entries of points in $A_p(x)$. Moreover, by the construction of the random tree, there holds $A_p(x)=[\underline{a}_p^1(x),\overline{a}_p^1(x)]\times\cdots\times[\underline{a}_p^d(x),\overline{a}_p^d(x)]$. Since $|\tilde{x}'_j-\tilde{x}_j| \leq \overline{a}_p^j(x)-\underline{a}_p^j(x)$ for any $\tilde{x}',\tilde{x}\in A_p(x)$ and $1\leq j\leq d$. Consequently, we get

$$\begin{aligned}\biggl|\int_{A_p(x)}(\tilde{x}'_j-\tilde{x}_j)\,d\tilde{x}'\biggr| \leq \int_{A_p(x)}|\tilde{x}'_j-\tilde{x}_j|\,d\tilde{x}' \leq (\overline{a}_p^j(x)-\underline{a}_p^j(x))\int_{A_p(x)}d\tilde{x}'& \\ = \mu(A_p(x))(\overline{a}_p^j(x)-\underline{a}_p^j(x)) = \mu(A_p(x))A_p^j(x).&\end{aligned} \tag{A.18}$$

Combining (A.17) with (A.18), we obtain

$$\biggl|\int_{A_p(x)}\nabla f(x)^\top(x'-x)\,dx'\biggr| \leq c_L\mu(A_p(x))\sum_{j=1}^{d}A_p^j(x). \tag{A.19}$$

Combining this with (A.16), we obtain

$$\bigl|\mathbb{E}_{\mathrm{P}_Z}(f_{\mathrm{P}}^{p,1}(x))-f(x)\bigr| \leq c_L\mathbb{E}_{\mathrm{P}_Z}\biggl(\sum_{j=1}^{d}A_p^j(x)\biggr) + c_L\mathbb{E}_{\mathrm{P}_Z}(\mathrm{diam}(A_p(x))^{1+\alpha}). \tag{A.20}$$

By (A.3), we have

$$\mathbb{E}_{\mathrm{P}_Z}\biggl(\sum_{j=1}^{d}A_p^j(x)\biggr) = \sum_{j=1}^{d}\mathbb{E}_{\mathrm{P}_R}\mathbb{E}_{\mathrm{P}_Z}(A_p^j(x)|R) = 2rd(1-0.5/d)^p \leq 2rd\exp(-0.5p/d). \tag{A.21}$$

Moreover, Lemma A.2 implies

$$\mathbb{E}_{\mathrm{P}_Z}(\mathrm{diam}(A_p(x))^{1+\alpha}) \leq (2r)^{1+\alpha}d\exp\bigl((2^{-\alpha-1}-1)p/d\bigr). \tag{A.22}$$

Combining (A.21), (A.22) with (A.20), we get

$$\begin{aligned}\bigl|\mathbb{E}_{\mathrm{P}_Z}(f_{\mathrm{P}}^{p,1}(x))-f(x)\bigr| &\leq 2c_Ldr\exp(-0.5p/d)+c_L(2r)^{1+\alpha}d\exp\bigl((2^{-\alpha-1}-1)p/d\bigr) \\ &\leq c_Ld(2r)^2\exp(-0.5p/d).\end{aligned} \tag{A.23}$$

Combining (A.14), (A.23) *with* (A.12), *we find*

$$\mathbb{E}_{\mathrm{P}_Z}\big(f_{\mathrm{P,E}}(x) - f(x)\big)^2 \leq c_L^2 (2r)^4 d^2 T^{-1} \exp(-0.75p/d) + 4c_L^2 (2r)^{2d+2} d^2 \exp(-p/d).$$

Taking expectation with respect to the Lebesgue measure μ, we get

$$\|f_{\mathrm{P,E}}(x) - f(x)\|_{L_2(\nu)}^2 \leq c_L^2 (2r)^4 d^2 T^{-1} \exp(-0.75p/d) + 4c_L^2 (2r)^{2d+2} d^2 \exp(-p/d),$$

which leads to the desired assertion by exchanging the order of integration.

Proof A.7 (Proof of Proposition 4) *According to the random tree splitting rule, split coordinates $\{Z_t\}_{t=1}^T$ are i.i.d. Thus we have*

$$\begin{aligned}\mathbb{E}_{\mathrm{P}_Z}\big(f_{\mathrm{D,E}}(x) - f_{\mathrm{P,E}}(x)\big)^2 &= \mathbb{E}_{\mathrm{P}_Z}\Big(\frac{1}{T}\sum_{t=1}^T \big(f_{\mathrm{D},t}^p(x) - f_{\mathrm{P},t}^p(x)\big)\Big)^2 \\ &\leq \mathbb{E}_{\mathrm{P}_Z}\frac{1}{T}\sum_{t=1}^T \big(f_{\mathrm{D},t}^p(x) - f_{\mathrm{P},t}^p(x)\big)^2 = \mathbb{E}_{\mathrm{P}_Z}\big(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x)\big)^2,\end{aligned}$$

where the inequality holds due to the Cauchy-Schwarz inequality. Let $\mathcal{A}_p = (A_p^j)_{j\in\mathcal{I}_p}$ be the random tree partition with p splits. Then we have

$$\begin{aligned}\mathbb{E}_{\mathrm{P}_{R,Z}\otimes\mathrm{P}^n}\big(f_{\mathrm{D,E}}(x) - f_{\mathrm{P,E}}(x)\big)^2 \leq \mathbb{E}_{\mathrm{P}_Z}\mathbb{E}_{\mathrm{P}^n}\big(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x)\big)^2 &= \mathbb{E}_{\mathrm{P}_Z}\frac{\mathrm{P}\big(A_p^j(x)\big)\big(1-\mathrm{P}(A_p^j(x))\big)}{n\mu^2(A_p^j(x))} \\ &\leq \mathbb{E}_{\mathrm{P}_Z}\frac{\mathrm{P}\big(A_p^j(x)\big)}{n\mu^2(A_p^j(x))} = \mathbb{E}_{\mathrm{P}_Z}\sum_{j=0}^{2^p}\frac{\mathrm{P}\big(A_p^j\big)}{n\mu^2(A_p^j)}\cdot \boldsymbol{1}_{A_p^j}(x).\end{aligned}$$

Fubini's theorem implies

$$\int_{B_r}\mathbb{E}_{\mathrm{P}_{R,Z}\otimes\mathrm{P}^n}\big(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x)\big)^2\,d\mu(x) = \mathbb{E}_{\mathrm{P}_Z}\int_{B_r}\sum_{j=0}^{p}\frac{\mathrm{P}\big(A_p^j\big)}{n\mu^2(A_p^j)}\boldsymbol{1}_{A_p^j}(x)\,d\mu(x) = \mathbb{E}_{\mathrm{P}_Z}\sum_{j=0}^{2^p}\frac{\mathrm{P}\big(A_p^j\big)}{n\mu(A_p^j)} \leq \|f\|_\infty\cdot\frac{2^p}{n}.$$

In other words, we have $\|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\nu)}^2 \leq \|f\|_\infty\cdot 2^p/n$, where $\nu := \mu\otimes\mathrm{P}_Z\otimes\mathrm{P}^n$. This proves the assertion.

B.3.2. Proof Related to Section A.2.2

We first show proofs for lower bound of approximation error for random tree density estimation.

Proof A.8 (Proof of Proposition 5) *For any $j\in\mathcal{I}_p$ and $x\in A_j$, we have*

$$f_{\mathrm{P}}^p(x) = \frac{\mathrm{P}(A_j)}{\mu(A_j)} = \frac{1}{\mu(A_j)}\int_{A_j} f(x')\,dx'.$$

Since $f(x)\in C^{1,\alpha}$, f is differentiable. Then according to the mean-value theorem, there exists $x^j\in A_j$ such that

$$f(x^j) = \frac{1}{\mu(A_j)}\int_{A_j} f(x')\,dx' = f_{\mathrm{P}}^p(x).$$

Consequently, we have

$$\int_{B^r}(f_{\mathrm{P}}^p(x) - f(x))^2 dx = \sum_{j\in\mathcal{I}_p}\int_{A_j}(f_{\mathrm{P}}^p(x) - f(x))^2\,dx = \sum_{j\in\mathcal{I}_p}\int_{A_j}(f(x^j) - f(x))^2\,dx. \tag{A.24}$$

Let $g(t) := f(x^j + t(x - x^j)) - f(x^j)$, $0\leq t\leq 1$. Since $f(x)\in C^{1,\alpha}$, $g(t)$ is differentiable at every $t\in(0,1)$. According to Lagrange's mean value theorem, there exists $t^*\in(0,1)$ such that

$$g(1) - g(0) = g'(t^*) = \nabla f(x^j + t^*(x - x^j))^\top(x - x^j).$$

Let $\xi^*_{j,x} := x^j + t^*(x - x^j)$*. Then we have*

$$(f(x^j) - f(x))^2 = (\nabla f(\xi^*_{j,x})(x - x^j))^\top \nabla f(\xi^*_{j,x})(x - x^j) = \|\nabla f(\xi^*_{j,x})\|^2 \|x - x^j\|^2.$$

Since $\|\nabla f\| \geq \underline{c}_f$*, we have* $(f(x^j) - f(x))^2 \geq \underline{c}_f^2 \|x - x^j\|^2$*. This together with* (A.24) *yields*

$$\begin{aligned}\int_{B_r} (f_{\mathrm{P}}^p(x) - f(x))^2 dx &\geq \underline{c}_f^2 \sum_{j \in \mathcal{I}_p} \int_{A_j} \|x - x^j\|^2 \, dx \\ &= \underline{c}_f^2 \sum_{j \in \mathcal{I}_p} \sum_{i=1}^d \int_{A_j} |x_i - x_i^j|^2 \, dx = \underline{c}_f^2 \sum_{i=1}^d \sum_{j \in \mathcal{I}_p} \int_{A_j} (x_i - x_i^j)^2 \, dx, \end{aligned} \tag{A.25}$$

where x_i^j *denotes the* i*-th entry of the vector* x^j*. Let* $\underline{a}_j^i$ *and* $\overline{a}_j^i$ *be the minimum and maximum values of the* i*-th coordinates of points in* A_j*. Then by the construction of the random tree partition, we have* $A_j = [\underline{a}_j^1, \overline{a}_j^1] \times \cdots \times [\underline{a}_j^d, \overline{a}_j^d]$*. Moreover, let* $h(t) := \int_{A_j} (x_i - t)^2 \, dx$*. Then by the iterated integral rule, we have*

$$h(t) = \prod_{s \neq i} (\overline{a}_j^s - \underline{a}_j^s) \int_{\underline{a}_j^i}^{\overline{a}_j^i} (x_i - t)^2 \, dx_i = \prod_{s \neq i} (\overline{a}_j^s - \underline{a}_j^s) \biggl((\overline{a}_j^i - \underline{a}_j^i) t^2 - 2t \int_{\underline{a}_j^i}^{\overline{a}_j^i} x_i \, dx_i + \int_{\underline{a}_j^i}^{\overline{a}_j^i} x_i^2 \, dx_i \biggr) \geq h\bigl((\underline{a}_j^i + \overline{a}_j^i)/2\bigr).$$

Consequently, we get

$$\int_{A_j} (x_i - x_i^j)^2 \, dx = h(x_i^j) \geq h\bigl((\underline{a}_j^i + \overline{a}_j^i)/2\bigr) = \int_{A_j} \bigl(x_i - (\underline{a}_j^i + \overline{a}_j^i)/2\bigr)^2 \, dx.$$

This together with (A.25) *implies*

$$\begin{aligned}\int_{B_r} (f_{\mathrm{P}}^p(x) - f(x))^2 dx &\geq \underline{c}_f^2 \sum_{i=1}^d \sum_{j \in \mathcal{I}_p} \int_{A_j} \bigl(x_i - (\underline{a}_j^i + \overline{a}_j^i)/2\bigr)^2 \, dx \\ &= \underline{c}_f^2 \sum_{i=1}^d \sum_{j \in \mathcal{I}_p} \int_{A_j} \bigl(x_i - (\underline{a}_p^i(x) + \overline{a}_p^i(x))/2\bigr)^2 \, dx = \underline{c}_f^2 \int_{B_r} \sum_{i=1}^d \bigl(x_i - (\underline{a}_p^i(x) + \overline{a}_p^i(x))/2\bigr)^2 \, dx, \end{aligned}$$

where $\underline{a}_p^i(x)$ *and* $\overline{a}_p^i(x)$ *are the minimum and maximum values of the* i*-th coordinates of points in* $A_p(x)$*. Therefore, we obtain*

$$\mathbb{E}_{\mathrm{P}_Z} \int_{B_r} (f_{\mathrm{P}}^p(x) - f(x))^2 dx \geq \underline{c}_f^2 \int_{B_r} \sum_{i=1}^d \mathbb{E}_{\mathrm{P}_Z} \bigl(x_i - (\underline{a}_p^i(x) + \overline{a}_p^i(x))/2\bigr)^2 dx. \tag{A.26}$$

Let $S_p^i(x)$ *be the number of times that* $A_p(x)$ *is split on the* i*-th coordinate. According to Lemma* A.3*, if* $S_p^i(x) = k$*,* $0 \leq k \leq q$*, then we have*

$$\bigl|x_i - (\underline{a}_p^i(x) + \overline{a}_p^i(x))/2\bigr| = \min_{q \in Q_k} |x_i - q|,$$

where $Q_k = \bigl\{ r(2j-1)/2^k \bigm| -2^{k-1} + 1 \leq j \leq 2^{k-1} \bigr\}$*. Therefore, we obtain*

$$\mathbb{E}_{\mathrm{P}_Z} \bigl(x_i - (\underline{a}_p^i(x) + \overline{a}_p^i(x))/2\bigr)^2 = \sum_{k=0}^p \mathrm{P}_Z(S_p^i(x) = k) \min_{q \in Q_k} (x_i - q)^2.$$

This together with (A.26) *implies*

$$\mathbb{E}_{\mathrm{P}_Z} \int_{B_r} (f_{\mathrm{P}}^p(x) - f(x))^2 \, dx \geq \underline{c}_f^2 \int_{B_r} \sum_{i=1}^d \sum_{k=0}^p \mathrm{P}_Z(S_p^i(x) = k) \min_{q \in Q_k} (x_i - q)^2 \, dx$$

$$= \underline{c}_f^2 \sum_{i=1}^d \biggl(\sum_{k=0}^p f(k,p,1/d) \int_{B_r} \min_{q \in Q_k} (x_i - q)^2 \, dx \biggr), \tag{A.27}$$

where $f(k,p,1/d) = \binom{p}{k} (\frac{1}{d})^k (1 - \frac{1}{d})^{n-k}$*. By the definition of* Q_k*, we have*

$$\begin{aligned}\int_{B_r} \min_{q \in Q_k} (x_i - q)^2 \, dx &= (2r)^{d-1} \int_{-r}^{r} \min_{q \in Q_k} (x_i - q)^2 \, dx_i \\ &= (2r)^{d-1} \cdot 2^{k+1} \int_{r-r/2^k}^{r} (x_i - (r - r/2^k))^2 \, dx_i = \frac{3(2r)^{d-1}}{2} \cdot \frac{r^3}{2^{2k}}.\end{aligned}$$

This together with (A.27) *implies*

$$\mathbb{E}_{\mathrm{P}_Z \otimes \mu} (f_{\mathrm{P}}^p(x) - f(x))^2 \geq \underline{c}_f^2 \sum_{i=1}^d \biggl(\sum_{k=0}^p 2^{-2k} \cdot f(k,p,1/d) \biggr) = 0.75 \underline{c}_f^2 r^2 d (2r)^d (1 - 0.75/d)^p,$$

which completes the proof.

Then we present proofs for lower bound of sample error for random tree density estimation.

Proof A.9 (Proof of Proposition 6) *Since* $H(x) = x$ *is a identity map, for any fixed split coordinates* $Z = \{Z_{i,j}, 1 \leq i \leq p, 1 \leq j \leq 2^{i-1}\}$, $\{A_j\}_{j \in \mathcal{I}_p}$ *forms a partition of* B_r*, then for* $j \in \mathcal{I}_p$ *we define the random variable* N_j *by* $N_j := \sum_{i=1}^n \mathbf{1}_{A_j}(X_i)$*. Since the random variables* $\{\mathbf{1}_{A_j}(X_i)\}_{i=1}^n$ *are i.i.d. Bernoulli distributed with parameter* $\mathrm{P}_X(x \in A_j)$*, it is clear to see that the random variable* N_j *is Binomial distributed with parameters* n *and* $\mathrm{P}_X(x \in A_j)$*. Therefore, for any* $j \in \mathcal{I}_p$*, we have* $\mathbb{E}(N_j) = n \cdot \mathrm{P}_X(x \in A_j)$*. Moreover, the random tree density estimator* f_{D}^p *can be defined by*

$$f_{\mathrm{D}}^p(x) = \begin{cases} \frac{N_j}{n\mu(A_j)} \cdot \mathbf{1}_{A_j}(x) & \text{if } N_j > 0, \\ 0 & \text{if } N_j = 0. \end{cases}$$

Then we have

$$\begin{aligned}\mathbb{E}_{D \sim \mathrm{P}^n} \int_{B_r} \bigl(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x) \bigr)^2 \, d\mu &= \mathbb{E}_{D \sim \mathrm{P}^n} \biggl(\sum_{j \in \mathcal{I}_p} \int_{A_j} \bigl(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x) \bigr)^2 \biggr) d\mu \\ &= \sum_{j \in \mathcal{I}_p} \frac{1}{\mu(A_j)^2} \int_{A_j} \mathbb{E}_{\mathrm{P}^n} \biggl(\frac{N_j}{n} - \mathrm{P}(A_j) \biggr)^2 \biggr) d\mu. \end{aligned} \tag{A.28}$$

Since for a fixed $j \in \mathcal{I}_p$*, there holds*

$$\begin{aligned}\mathbb{E}_{\mathrm{P}^n} \biggl(\frac{N_j}{n} - \mathrm{P}(A_j) \biggr)^2 &= \frac{1}{n^2} \mathbb{E}_{\mathrm{P}^n} N_j^2 - \frac{2\mathrm{P}(A_j)}{n} \mathbb{E}_{\mathrm{P}^n} N_j + \mathrm{P}(A_j)^2 \\ &= \frac{n\mathrm{P}(A_j)(1 - \mathrm{P}(A_j)) + n^2 \mathrm{P}(A_j)^2}{n^2} - \frac{2n\mathrm{P}(A_j)^2}{n} + \mathrm{P}(A_j)^2 = \frac{\mathrm{P}(A_j)(1 - \mathrm{P}(A_j))}{n}\end{aligned}$$

Therefore, together with (A.28)*, we have*

$$\mathbb{E}_{D \sim \mathrm{P}^n} \int_{B_r} \bigl(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x) \bigr)^2 \, d\mu = \sum_{j \in \mathcal{I}_p} \frac{1}{\mu(A_j)^2} \int_{A_j} \frac{\mathrm{P}(A_j)(1 - \mathrm{P}(A_j))}{n} \, d\mu = \sum_{j \in \mathcal{I}_p} \frac{\mathrm{P}(A_j)(1 - \mathrm{P}(A_j))}{n\mu(A_j)}.$$

By the assumption $p \geq \ln(\|f\|_\infty 2^{d+1} r^d) / \log 2$*, we have* $\mathrm{P}(A_j) \leq \|f\|_\infty \mu(A_j) = \|f\|_\infty (2r)^d / 2^p \leq 1/2$*. Consequently, we get*

$$\mathbb{E}_{D \sim \mathrm{P}^n} \int_{B_r} \bigl(f_{\mathrm{D}}^p(x) - f_{\mathrm{P}}^p(x) \bigr)^2 \, d\mu \geq \frac{1}{2} \cdot \sum_{j \in \mathcal{I}_p} \frac{\mathrm{P}(A_j)}{n\mu(A_j)} = \frac{2^p}{2^{d+1} r^d n}. \tag{A.29}$$

Hence we prove the desired assertion.

B.4. Proof Related to Section 3

Proof A.10 (Proof of Theorem 1) *Proposition 1 and 2 yield that*

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 = \|f_{\mathrm{P,E}} - f\|_{L_2(\nu)}^2 + \|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\nu)}^2 \leq c_L^2 (2r)^{2\alpha} d \exp\biggl(\frac{(2^{-2\alpha}-1)p}{d}\biggr) + \|f\|_\infty \cdot \frac{2^p}{n}, \tag{A.30}$$

By choosing $p_n := d \log n/(d\log 2 + 1 - 4^{-\alpha})$, we then obtain

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 \lesssim n^{-\frac{1-4^{-\alpha}}{d\log 2 + 1 - 4^{-\alpha}}},$$

which proves the assertion.

Proof A.11 (Proof of Theorem 2) *Proposition 3 and 4 yield*

$$\begin{aligned}\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 &= \|f_{\mathrm{P,E}} - f\|_{L_2(\nu)}^2 + \|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\nu)}^2 \\ &\leq c_L^2 (2r)^4 d^2 T^{-1} \exp(-0.75p/d) + 4c_L^2(2r)^{2d+2} d^2 \exp(-p/d) + \|f\|_\infty \cdot 2^p/n.\end{aligned} \tag{A.31}$$

By choosing $p_n := d\log n/(1 + d\log 2)$, $T_n := n^{1/(4+4d\log 2)}$, we then obtain

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 \lesssim n^{-\frac{1}{d\log 2 + 1}},$$

which proves the assertion.

Let us consider the case $T = 1$ where RFDE reduce to the single base learner RTDE, the following theorem presents an upper bound for the rate of RTDE.

Theorem A.12 *Let $(A_p^j)_{j\in\mathcal{I}_p}$ be a random tree partition with depth T induced by splitting variable Z. Moreover, let f_{D}^p be the RTDE estimator and assume that the true density $f \in C^{1,\alpha}$ with support $\mathcal{X} \subset B_r$. Let (p_n) be the sequence defined by $p_n := d(0.75 + d\log 2)^{-1}\log n$. Then we have*

$$\|f_{\mathrm{D}}^p - f\|_{L_2(\nu)}^2 \lesssim n^{-\frac{0.75}{d\log 2 + 0.75}}. \tag{A.32}$$

Proof A.13 (Proof of Theorem A.12) *The excess risk bound (A.31) with $T = 1$ and $p_n := d\log n/(0.75 + d\log 2)$ yields*

$$\|f_{\mathrm{D}}^p - f\|_{L_2(\nu)}^2 \lesssim n^{-\frac{0.75}{d\log 2 + 0.75}},$$

which proves the assertion.

Proof A.14 (Proof of Theorem 3) *Recall the error decomposition (A.2). Applying Propositions 5 and 6, we get*

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 \geq \frac{3\underline{c}_f^2 r^2 d (2r)^d}{4d^{d/2}} \biggl(1 - \frac{3}{4d}\biggr)^p + \frac{2^p d^{d/2}}{2^{d+1} r^d n} \geq c_0 n^{\frac{\log(1-0.75/d)}{\log 2 - \log(1-0.75/d)}}. \tag{A.33}$$

if $p \geq p_0 := \lceil \ln(\|f\|_\infty 2^{d+1} r^d)/\log 2\rceil$. On the other hand, if $p \leq p_0$, again by (A.2), we get

$$\|f_{\mathrm{D,E}} - f\|_{L_2(\nu)}^2 \geq \|f_{\mathrm{P,E}} - f\|_{L_2(\nu)}^2 \geq \frac{3\underline{c}_f^2 r^2 d(2r)^d}{4d^{d/2}}\biggl(1 - \frac{3}{4d}\biggr)^{p_0} := c_1. \tag{A.34}$$

Combining (A.33) with (A.34), we find

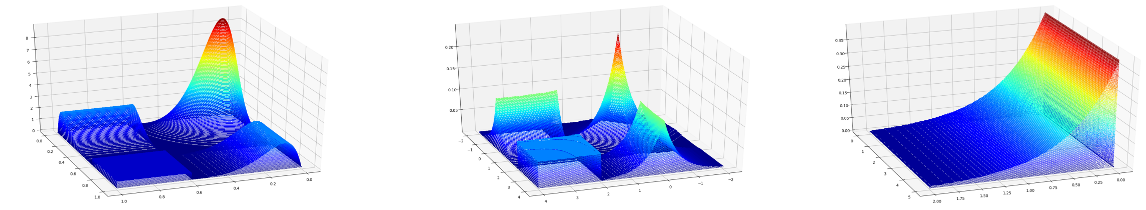
$$\mathbb{E}_{\mathrm{P}^n \otimes \mathrm{P}_Z}\bigl(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}}) - R_{L,\mathrm{P}}^*\bigr) \geq c_0 n^{\frac{\log(1-0.75/d)}{\log 2 - \log(1-0.75/d)}} \vee c_1,$$

which leads to the desired assertion.

Table 3. Descriptions of synthetic datasets.

Type	True (Marginal) Distribution
I	$f_i := 0.7 \cdot \mathrm{Beta}(2, 10) + 0.3 \cdot \mathrm{Unif}(0.6, 1.0)$
II	$f_i := 0.5 \cdot \mathrm{Laplace}(0, 0.5) + 0.5 \cdot \mathrm{Unif}(2, 4)$
III	$f_i := \mathrm{Exp}(0.5)$ for $1 = 1, \dots, d-1$ and $f_d := \mathrm{Unif}(0, 5)$

* Let f_i as the marginal probability distribution of the i-th dimension. For Types I, II and III, the marginal distributions of the true density are independent, and the marginal distributions are identical for Types II and III.



(a) Type I

(b) Type II

(c) Type III

Figure 5. 3D plots of the synthetic distributions with $d = 2$.

C. Supplementary for Experiments

C.1. Descriptions of Synthetic Datasets

The detailed descriptions are shown in Table 3.

In order to give clear visualization of the distributions, we take $d = 2$ for instance, and give the 3D visualization of the above four types of distributions in Figure 5, where x-axis and y-axis represent the 2-dimensional feature space and z-axis represents the value of the density function.

C.2. Descriptions of Real Datasets

As follows are the datasets alphabetically listed, with the number of instances and features reported after preprocessing.

- `Abalone`: contains $4,177$ instances and 9 features with no missing values. The features are physical measurements of abalone, which are originally designed for age predicting.
- `Adult` is also known as "Census Income" dataset. It contains $48,842$ instances with 6 countinuous and 8 discrete attributes. Prediction task is to determine whether a person makes over 50K a year.
- `Australian` is an interesting dataset with a good mix of attributes, which contains continuous, nominal with both small and large numbers of values. The dataset contains 690 instances with 6 numerical and 9 categorical attributes, mainly concerning credit card applications.
- `Breast-cancer` is originally for predicting whether a cancer is recurrence event. It contains 675 instances of dimension 11, describing the status of the tumors and the patients.
- `Credit`: the *Credit Approval* dataset, is a dataset of credit card applications, with 653 instances of dimension 16.
- `Diabetes` dataset comprises 768 samples and 9 features. The attributes concern about the medical records of patients, consisting of 8 numerical features and 1 categorical feature.

C.3. Random Forest Density Estimation (RFDE) for Anomaly Detection

We conduct numerical experiments to make a comparison between our RFDE and several popular anomaly detection algorithms such as the forest-based Isolation Forest (iForest) (Liu et al., 2008), the distance-based k-Nearest Neighbor

(k-NN) (Ramaswamy et al., 2000) and Local Outlier Factor (LOF) (Breunig et al., 2000), and the kernel-based one-class SVM (OCSVM) (Schölkopf et al., 2001), on 20 real-world benchmark outlier detection datasets from the ODDS library. The detailed descriptions of these datasets can be found in Table 4. The measure for the performance evaluation is the area under the ROC curve (*AUC*). For each method, we choose the best *AUC* performance when parameters go though their parameter grids.

The implementation details are below: For our method, the grid of depth p is $\{1, 2, 3, 5, 10, 15, 20, 25, 30\}$. The number of base learners T is set as 100. For iForest, LOF and OCSVM, we utilized the implementation of scikit-learn. For k-NN and LOF, the parameter grid of number of neighbors k is $\{5, 10, 15, \cdots, 45, 50\}$. As for iForest, we set the grid of the number of trees to be $\{100, 500\}$ and sub-sampling size to be 256. For OCSVM, we use RBF kernel with gamma grid $\{0.001, 0.01, \cdots, 1, 10\}$. The experimental results are reported in Table 5.

Table 4. Descriptions of Benchmark Datasets

Datasets	n	d	#outliers(%)	Datasets	n	d	#outliers(%)
annthyroid	7200	6	534(7.42%)	breastw	683	9	239(34.99%)
cardio	1,831	21	176(9.61%)	forestcover	286,048	10	2747(0.96%)
glass	214	9	9(4.2%)	http	567,498	3	2211(0.39%)
ionosphere	351	33	126(35.90%)	letter	1,600	32	100(6.25%)
mammo.	11,183	6	260(2.32%)	mulcross	262,144	4	26214(10.00%)
musk	3,062	166	97(3.2%)	pendigits	6,870	16	156(2.27%)
pima	768	8	268(34.90%)	satellite	6,435	36	2036(32%)
shuttle	49,097	9	3511(7.15%)	smpt	95156	3	30(0.03%)
speech	3686	400	61(1.65%)	thyroid	3772	6	93(2.5%)
vowels	1,456	12	50(3.43%)	wbc	129	13	10(7.7%)

Table 5. *AUC* performance on benchmark datasets

Datasets	RFDE (Ours)	k-NN	iForest	LOF	OCSVM	Lump	HDBSCN	AOM+VR
annthyroid	0.7646	0.7511	0.8209	0.7386	0.6749	**0.8767**	0.7119	0.6655
breastw	**0.9938**	0.9881	0.9884	0.4676	0.9789	0.9882	0.9882	0.9265
cardio	0.8360	0.8744	0.9297	0.6790	**0.9473**	0.8922	0.8775	0.8682
forestcover	0.9168	0.8950	0.8792	0.5778	0.6565	0.8258	0.7668	**0.9232**
glass	0.8599	0.8683	0.7041	0.8385	**0.8748**	0.6049	0.7396	0.6763
http	0.9947	0.2309	**0.9999**	0.3675	0.9953	0.9964	0.3724	0.4219
ionosphere	**0.9398**	0.9294	0.8520	0.9023	0.9382	0.7431	0.9255	0.8451
letter	0.8384	0.9071	0.6258	**0.9120**	0.6860	0.3480	0.7735	0.806
mammo.	0.8501	0.8527	0.8631	0.7568	**0.8721**	0.8615	0.709	0.7992
mulcross	0.9474	0.0013	0.9642	0.5848	0.9778	**0.9989**	0.7868	0.3434
musk	1.0000	0.9367	**1.0000**	0.5476	0.5281	0.9632	0.3815	0.8651
pendigits	0.9558	0.8607	0.9538	0.5437	**0.9607**	0.8971	0.6115	0.8342
pima	0.6927	0.6437	0.6796	0.6162	0.5842	0.6600	**0.7283**	0.6376
satellite	0.6850	**0.7374**	0.7041	0.5701	0.7064	0.6395	0.5717	0.6602
shuttle	0.9806	0.8004	**0.9974**	0.6035	0.9918	0.9861	0.5289	0.6786
smtp	**0.9378**	0.9338	0.9076	0.9299	0.7752	0.8156	0.8817	0.8343
speech	**0.6255**	0.4862	0.4782	0.6247	0.5564	0.5083	0.4832	0.4839
thyroid	0.9582	0.9510	**0.9785**	0.9464	0.9491	0.9714	0.9482	0.9353
vowels	0.9548	**0.9749**	0.7588	0.9467	0.9153	0.6968	0.9523	0.9419
wbc	**0.9689**	0.9501	0.9412	0.9460	0.9469	0.9292	0.949	0.9406
Rank Sum	**55**	78	72	116	82	91	108	118

* The best results are marked in **bold**, the second best results are marked in underline.
** The last row shows the summation of ranks for each method, which is the lower the better.