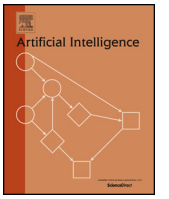
Artificial Intelligence 313 (2022) 103788

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Towards convergence rate analysis of random forests for classification

Wei Gao, Fan Xu, Zhi-Hua Zhou *

National Key Laboratory for Novel Software Technology, Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University, Nanjing 210093, China

ARTICLE INFO

Article history:
Received 2 September 2021
Received in revised form 26 July 2022
Accepted 17 September 2022
Available online 21 September 2022

Keywords:
Machine learning
Classification
Random forests
Convergence
Consistency

ABSTRACT

Random forests have been one of the successful ensemble algorithms in machine learning, and the basic idea is to construct a large number of random trees individually and make predictions based on an average of their predictions. The great successes have attracted much attention on theoretical understandings of random forests, mostly focusing on regression problems. This work takes one step towards the convergence rates of random forests for classification. We present the first finite-sample rate $O(n^{-1/(8d+2)})$ on the convergence of purely random forests for binary classification, which can be improved to be of $O(n^{-1/(3.87d+2)})$ by considering the midpoint splitting mechanism. We introduce another variant of random forests, which follows Breiman's original random forests but with different mechanisms on splitting dimensions and positions. We present the convergence rate $O(n^{-1/(d+2)}(\ln n)^{1/(d+2)})$ for the variant of random forests, which reaches the minimax rate, except for a factor $(\ln n)^{1/(d+2)}$, of the optimal plug-in classifier under the L-Lipschitz assumption. We achieve the tighter convergence rate $O(\sqrt{\ln n/n})$ under some assumptions over structural data. This work also takes one step towards the convergence rate of random forests for multi-class learning, and presents the same convergence rates of random forests for multi-class learning as that of binary classification, yet with different constants. We finally provide empirical studies to support the theoretical analysis.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

From the pioneer work [12], random forests have been regarded as one of the successful ensemble algorithms in machine learning, which construct a large number of random trees individually and then make predictions based on an average of their predictions. This idea is partly motivated from geometric feature selection [2], random subspace [33], random split selection [23], as well as earlier ensemble decision trees [38]. Random forests have achieved good performance empirically [10,12,25,56], and have been involved in diverse applications such as ecology [18], computational biology [48], computer vision [16], objection recognition [55], remote sensing [7], and so on. Numerous variants have been developed to improve performance and reduce computational costs [4,6,19,30,39,40,44,51,60,66]. For an overview of random forests, we refer readers to the works of [10,17,28].

Empirical successes have attracted much attention on theoretical explorations of random forests. Breiman [12] presented the generalization bounds for random forests based on the correlation and strength of individual random trees, followed by

* Corresponding author.
E-mail address: zhouzh@nju.edu.cn (Z.-H. Zhou).

https://doi.org/10.1016/j.artint.2022.103788
0004-3702/© 2022 Elsevier B.V. All rights reserved.

consistency analysis of a simple model of random forests [13]. Lin and Jeon [41] established a connection between random forests and adaptive nearest neighbors, and Meinshausen [43] studied consistency of random forests for regression in the context of conditional quantile predictions. The consistency results place random forests in a favored category of ensemble algorithms [8,9,46,52,53,59]. Denil et al. [20] narrowed the gap between theory and practice of random forests for regression, and Goetz et al. [31] proposed an active learning algorithm for non-parametric regression using random forests. Li et al. [40] derived non-asymptotic bounds on the expected bias of MDI importance for random forests, along with variable importance [35,42]. Tang et al. [58] discussed when random forests fail and examined the influences of parameters over performance. Most previous studies focus on the theoretical understandings of random forests for regression problems.

For classification, Biau et al. [9] took a crucial milestone on the consistency of randomized ensemble classifiers, and Denil et al. [19] showed the first consistency of online random forests. For a full understanding, however, it is necessary to take one further step towards the convergence rates of random forests for classification, which would be beneficial to design better random forests, and comprehend the effects of different splitting mechanisms during the constructions of random forests for classification.

This work takes one step towards convergence rates of random forests for classification, and the main contributions can be summarized as follows:

- We present the first finite-sample rate on the convergence of purely random forests, which were proposed originally by Breiman [11]. Specifically, a convergence rate $O(n^{-1/(8d+2)})$ is derived for binary classification by selecting leaves number $k = O(n^{4d/(4d+1)})$, where n and d denote the size of training data and dimension, respectively. This rate can be further improved to be of $O(n^{-1/(3.87d+2)})$ if we instead split a leaf along the dimension at the midpoint of the chosen side. As a by-product, we present the convergence rates between random forests and individual random trees, and make a better estimate on the height of random trees than was previously known.
- We introduce another simplified variant of random forests, which follows Breiman's original random forests [12] but with different mechanisms on splitting dimensions and positions. We derive a convergence rate $O(n^{-1/(d+2)}(\ln n)^{1/(d+2)})$ for the simplified random forests, which reaches the minimax rate, except for a factor $(\ln n)^{1/(d+2)}$, of the optimal plug-in classifiers under the L-Lipschitz assumption. We finally achieve the tighter convergence rate $O(\sqrt{\ln n/n})$ under some assumptions over structural data, which sheds insights on random forests by correlating randomization process with data-dependent tree structure.
- We further study the convergence analysis of random forests for multi-class learning, and achieve the same convergence rates of random forests for multi-class learning as that of binary classification, yet with different constants. To our knowledge, this presents the first convergence analysis for multi-class learning under only the L-Lipschitz assumption, and the proofs are rather technical. Relevant results may present independent interests on the convergence analysis of other multi-class algorithms and problems.
- We finally provide empirical studies to support our theoretical analysis.

1.1. Related work

A large number of variants of random forests have been developed for different problems and settings during the past decades. Geurts et al. [30] introduced the *extremely randomized trees* and Amaratunga et al. [1] provided the *enriched random forests* for DNA microarray data of huge features. Menze et al. [44] presented the *oblique random forests* for multivariate trees by explicitly learning the optimal split directions with linear discriminative models. Clémençon et al. [14] introduced the *ranking forests* based on aggregation and feature randomization principles for bipartite ranking. Athey et al. [4] developed a flexible and computationally efficient algorithm for the generalized random forests. A general framework is presented in [63] on various splitting criteria for random forests based on loss functions. Zhou and Feng [65,66] proposed *gcForest* with performance highly competitive to deep neural networks. Online random forests have also been developed with strong theoretical guarantees [19,39,46,57].

For regression, much attention has been paid on the $\mathbb{L}_2^2$-consistency of random forests with relevant variants [3,8,20, 27,43,53]. In particular, Scornet et al. [53] proved the first $\mathbb{L}_2^2$-consistency of Breiman's original random forests based on some assumptions such as additive regression functions and uniform distribution over instance space $\mathcal{X}$. The crucial analysis technique is the classical decomposition of variance and bias for random forests regression, whereas it is difficult to make such decomposition for random forests in classification. Moreover, the stopping-splitting criteria are different for random forests classification and regression, as shown in Algorithm 1 and work [53], respectively. We do not directly compare the convergence rates of random forests for regression and classification due to different settings and performance measures.

For classification, Biau et al. [9] made a crucial milestone on the consistency of some randomized ensemble classifiers such as purely random forests. The key technical tool is the general consistency theorem for partition classifiers [22, Theorem 6.1], that is, partition classifiers are consistent if the followings hold in probability (written with our notations),

$$\nu(C(\boldsymbol{x})) \to 0 \ \text{ and } \ |C(\boldsymbol{x}) \cap S_n| \to +\infty \ \text{ as } \ n \to +\infty\,,$$

where $\nu(C(\boldsymbol{x}))$ denotes the diameter of the leaf or rectangular cell $C(\boldsymbol{x})$. Based on this result, some variants of random forests classifiers have been proven to be consistent such as online random forests [19] and Mondrian forests [46]. Our work presents the convergence rates of random forests for classification based on different analysis techniques.

Mourtada et al. [46] presented the consistency of online Mondrian forests classifiers from [22, Theorem 6.1], and derived the minimax rate $O(n^{-1/(d+2)})$ for plug-in classifiers based on the estimation of conditional probability, that is, they took an average of conditional probabilities calculated by individual Mondrian trees. This is different from random forests classifier, which takes a majority over the predictions made by individual random trees. Wang et al. [61,62] proposed the novel Bernoulli random forests with theoretical consistency and empirical supports.

The rest of this work is organized as follows: Section 2 shows the convergence rates between random forests and individual random trees. Section 3 presents the convergence rates of purely random forests with its variants. Section 4 provides the convergence rates of the simplified variant of Breiman's random forests. Section 5 gives the convergence rates of random forests for multi-class learning. Section 6 presents the detailed proofs of our theoretical results. Section 7 conducts empirical studies. Section 8 concludes with future work.

2. Convergence rates between random forests and random trees

We begin with some notations used in this work. Let $\mathcal{B}(p)$ be a Bernoulli distribution with parameter $p \in [0,1]$, and $\mathcal{U}(a,b)$ represents a uniform distribution over the interval $[a,b]$. We denote by $\mathcal{M}(p_1, p_2, \cdots, p_n)$ the multinomial distribution with parameters $p_1, p_2, \cdots, p_n \in [0,1]$ and $p_1 + p_2 + \cdots + p_n = 1$. For positive $f(n)$ and $g(n)$, we write $f(n) = O(g(n))$ if there exist two constants $c, n_0 \in (0, +\infty)$ such that $f(n) \leq cg(n)$ for $n \geq n_0$. For integer $n > 0$ and real r, we introduce $[n] := \{1, 2, \cdots, n\}$ and let $\lfloor r \rfloor$ denote the largest integer which is no more than r. Denote by Euler's constant $e = 2.718\cdots$.

Let $\mathcal{X} \subseteq [0,1]^d$ and $\mathcal{Y} = [\tau]$ denote the instance and label space, respectively, and this work focuses on binary classification ($\tau = 2$) and multi-class learning ($\tau > 2$). Suppose that $\mathcal{D}$ is an underlying (unknown) distribution over the product space $\mathcal{X} \times \mathcal{Y}$, and let $\mathcal{D}_{\mathcal{X}}$ be its marginal distribution over instance space $\mathcal{X}$. Denote by

$$\eta_j(\boldsymbol{x}) = \Pr[y = j | \boldsymbol{x}] \quad \text{for} \quad j \in [\tau]$$

the conditional probability of $y = j$ over instance $\boldsymbol{x}$ w.r.t. distribution $\mathcal{D}$, and $\sum_{j=1}^{\tau} \eta_j(\boldsymbol{x}) = 1$.

Given a hypothesis $h \colon \mathcal{X} \to \mathcal{Y}$, we define the classification error over distribution $\mathcal{D}$ as

$$R_{\mathcal{D}}(h) = \Pr_{(\boldsymbol{x},y)\sim\mathcal{D}}[h(\boldsymbol{x}) \neq y] = E_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathbb{I}[h(\boldsymbol{x}) \neq y]] = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\sum_{j=1}^{\tau} \eta_j(\boldsymbol{x}) \mathbb{I}[h(\boldsymbol{x}) \neq j]\right].$$

Here, $\mathbb{I}[\cdot]$ denotes the indicator function, which returns 1 if the argument is true and 0 otherwise. Hence, the optimal Bayes' error (the minimum of classification error) and the optimal Bayes' classifier can be given, respectively, by

$$R^*_{\mathcal{D}} = E_{\boldsymbol{x}}\left[\min_{j\in[\tau]}\{1 - \eta_j(\boldsymbol{x})\}\right] \quad \text{and} \quad h^*_{\mathcal{D}}(\boldsymbol{x}) = \arg\max_{j\in[\tau]}\{\eta_j(\boldsymbol{x})\},$$

where ties are broken arbitrarily.

Notice that distribution $\mathcal{D}$ is unknown in practice, and what we observe is a training data

$$S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\},$$

where each example is drawn independently and identically (i.i.d.) from distribution $\mathcal{D}$. Our goal is to learn a classifier $\hat{h}_n$ from the training data S_n of smaller classification error. As the training data size n increases, we could obtain a sequence of classifiers $\hat{h}_1, \hat{h}_2, \cdots, \hat{h}_n, \cdots$. A sequence of classifiers $\{\hat{h}_n\}_{n=1}^{\infty}$ is said to be *consistent* if $E_{S_n}[R_{\mathcal{D}}(\hat{h}_n)] \to R^*_{\mathcal{D}}$ as $n \to \infty$.

Random forests classifier $f_m(\boldsymbol{x})$ takes a majority vote over m individual randomized trees $f_{S_n,\Theta_1}(\boldsymbol{x})$, $f_{S_n,\Theta_2}(\boldsymbol{x}), \ldots,$ $f_{S_n,\Theta_m}(\boldsymbol{x})$, that is,

$$f_m(\boldsymbol{x}) = \arg\max_{j\in[\tau]}\left\{\sum_{i=1}^{m} \mathbb{I}\left[f_{S_n,\Theta_i}(\boldsymbol{x}) = j\right]\right\}, \tag{1}$$

where ties are broken arbitrarily. The random vectors $\Theta_1, \Theta_2, \ldots, \Theta_m$ are distributed identically and independently, and characterize the mechanisms of random selections of splitting leaves, dimensions, and positions during the construction of randomized trees. The random vectors $\Theta_1, \Theta_2, \ldots, \Theta_m$ will be specified according to different random forests in the subsequent sections.

We first present the following relationship of convergence rate between random forests classifier and individual random tree classifier, and the detailed proof is presented in Section 6.1.

Lemma 1. *Let $f_m(\boldsymbol{x})$ be the random forests classifier given by Eqn. (1), and $f_{S_n,\Theta}(\boldsymbol{x})$ denotes a classifier of individual tree with respect to random vector Θ. We have*

$$E_{\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m(\boldsymbol{x}))] - R^*_{\mathcal{D}} \leq \tau(E_{\Theta}[R_{\mathcal{D}}(f_{S_n,\Theta}(\boldsymbol{x}))] - R^*_{\mathcal{D}}),$$

where τ is the number of labels in classification.

This lemma shows that the convergence rate of a random forests classifier $f_m(\boldsymbol{x})$ is no more than τ-times that of individual random tree classifier $f_{S_n,\Theta}(\boldsymbol{x})$, and the convergence rate of random forests is obtained from the expectation of convergence rates of individual trees, which can be viewed as an average of convergence rate of all individual random trees.

Lemma 1 recovers the convergence rate for binary classification [26, Lemma 1] when $\tau = 2$. If $E_\Theta[R_\mathcal{D}(f_{S_n,\Theta}(\boldsymbol{x}))] \to R^*_\mathcal{D}$, then we have $E_{\Theta_1,\ldots,\Theta_m}[R_\mathcal{D}(f_m(\boldsymbol{x}))] \to R^*_\mathcal{D}$; therefore, the consistency of random forests follows from the consistency of individual random tree. Such result also recovers the consistency results for binary classification [9, Proposition 1] and for multi-class learning [19, Proposition 1], and what's more, our lemma could present its convergence rate.

Notice that Lemma 1 is almost tight without additional assumptions, which can be shown by the following example:

Example 1. We consider binary classification with label space $\mathcal{Y} = \{0, 1\}$. For $\eta \in (0, 1/4)$ and $\delta \in (0, 1/2)$, we suppose $\eta_1(\boldsymbol{x}) = \Pr[y = 1|\boldsymbol{x}] = 1/2 + \eta$ and $\Pr_\Theta\left[f_{S_n,\Theta}(\boldsymbol{x}) = 1|\boldsymbol{x}\right] = 1/2 - \delta$ for every $\boldsymbol{x} \in \mathcal{D}$. We have $E_\Theta\left[R_\mathcal{D}\left(f_{S_n,\Theta}\right)\right] - R^*_\mathcal{D} = \eta(1 - 2\delta)$ and it also holds that

$$E_{\Theta_1,\ldots,\Theta_m}\left[R_\mathcal{D}\left(f_m\right)\right] - R^*_\mathcal{D} = 2\eta E_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left[\Pr_{\Theta_1,\ldots,\Theta_m}\left[f_m(\boldsymbol{x}) = 0\right]\right]$$
$$= 2\eta E_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left[\Pr_{\Theta_1,\ldots,\Theta_m}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left[f_{S_n,\Theta_j}(\boldsymbol{x}) = 0\right] \geq \frac{1}{2}\right]\right] \geq 2\eta(1 - \exp(-2m\delta^2)) ,$$

where the last inequality holds from the Chernoff bounds [34]. By setting $\delta = 1/m^{1/4}$, we have

$$E_{\Theta_1,\ldots,\Theta_m}\left[R_\mathcal{D}\left(f_m\right)\right] - R^*_\mathcal{D} = 2\eta = \tau(E_\Theta\left[R_\mathcal{D}\left(f_{S_n,\Theta}\right)\right] - R^*_\mathcal{D}) \quad \text{as} \quad m \to +\infty .$$

3. Convergence rates of the purely random forests for binary classification

This section focuses on binary classification with label space $\mathcal{Y} = \{1, 0\}$. Let $\eta_1(\boldsymbol{x})$ be the conditional probability of $y = 1$ w.r.t. distribution $\mathcal{D}$, and it follows that $\Pr[y = 0|\boldsymbol{x}] = 1 - \eta_1(\boldsymbol{x})$. We assume that $\eta_1(\boldsymbol{x})$ is L-Lipschitz for some constant $L > 0$, that is, $|\eta_1(\boldsymbol{x}) - \eta_1(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|$ for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. This assumption has been taken in random forests for regression [8,46] and binary classification [15,54], and an intuitive explanation on such assumption is that two instances are likely to have similar labels for smaller distance.

We begin with the *purely random forests*, which were originally proposed by Breiman [11]. Genuer [27] studied the variance reductions of purely random forests for regression, while Arlot and Genuer [3] presented its bias-variance analysis. For classification, Biau et al. [9] made an important milestone on the consistency of purely random forests. This section takes one further step on the convergence rate of purely random forests for binary classification.

Formally, a purely random tree can be constructed as follows. Each node is associated with a rectangular cell, and all leaves (external nodes) constitute a partition of $[0, 1]^d$ at each iteration of tree construction. The root of random partition is $[0, 1]^d$ itself. The following procedure is repeated $k - 1$ iterations for some pre-defined parameter $k \geq 2$ in advance, and hence the output random tree has k leaves.

- A split leaf is selected at random, uniformly over all leaves at the current iteration.
- Once the leaf is selected, a split dimension is selected at random, uniformly over $[d]$.
- The leaf is split along the split dimension at random, uniformly over the chosen side.

A purely random tree classifier $f_{S_n,\Theta}(\boldsymbol{x})$ takes a majority vote over labels y_i, whose corresponding instances $\boldsymbol{x}_i$ belong to the same cell of random partition as instance $\boldsymbol{x}$. The main difference, between the purely random tree and Breiman's original random tree [12], is that recursive cell splits are irrelevant to label information, and the growth of individual random tree is independent of training sample. Given m individual purely random trees $f_{S_n,\Theta_1}(\boldsymbol{x})$, $f_{S_n,\Theta_2}(\boldsymbol{x}), \ldots, f_{S_n,\Theta_m}(\boldsymbol{x})$, the random forests classifier takes a majority vote over those random trees, that is, the voting classifier $f_m(\boldsymbol{x}) = \mathbb{I}[\sum_{i=1}^{m} f_{S_n,\Theta_i}(\boldsymbol{x}) \geq m/2]$.

We now go into the details of randomness Θ on the construction of purely random forests. Given a purely random tree, we associate k leaves with k disjoint rectangular cells $C_1, C_2, \ldots, C_k$, constituting a partition of instance space $\mathcal{X} = [0, 1]^d$. Let $C(\boldsymbol{x})$ denote the rectangular cell of random tree, that contains the instance $\boldsymbol{x}$.

Given an instance $\boldsymbol{x} \in \mathcal{X}$, we introduce $k - 1$ Bernoulli random variables $X_1, X_2, \cdots, X_{k-1}$ to characterize the random events that the node, containing instance $\boldsymbol{x}$, was selected for splitting in the construction of random tree. Specially, the event $X_i = 1$ implies that the node containing $\boldsymbol{x}$ is selected for splitting in the i-th iteration of random tree construction; otherwise, $X_i = 0$. It follows that $X_i \sim \mathcal{B}(1/i)$, since there are i leaves for selection with identical probability during the i-th iteration of random tree construction.

Let $h(C(\boldsymbol{x}))$ denote the height of the rectangular cell $C(\boldsymbol{x})$, i.e., the splitting times of $C(\boldsymbol{x})$ during the construction of random tree. It is easy to obtain

$$h(C(\boldsymbol{x})) = \sum_{i=1}^{k-1} X_i .$$

We present upper and lower bounds on $h(C(\boldsymbol{x}))$ in expectation and in probability as follows:

Lemma 2. *Let $X_1, X_2, \ldots, X_{k-1}$ be $k-1$ random variables such that $X_i \sim \mathcal{B}(1/i)$ for $i \in [k-1]$. Given an instance $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\ln(k) \leq E_{X_1,X_2,\ldots,X_{k-1}}[h(C(\boldsymbol{x}))] \leq 1 + \ln(k-1) .$$

For any $\epsilon \in (0,1)$, we also have

$$\Pr_{X_1,X_2,\ldots,X_{k-1}}[h(C(\boldsymbol{x})) \leq (1-\epsilon)\ln k] \leq k^{-\epsilon^2/2} ,$$
$$\Pr_{X_1,X_2,\ldots,X_{k-1}}[h(C(\boldsymbol{x})) \geq (1+\epsilon)(1+\ln(k-1))] \leq k^{-\epsilon^2/2} .$$

We have $h(C(\boldsymbol{x})) = O(\log k)$ with high probability, especially for large k. Lemma 2 improves the previous work [9] on the bounds of $h(C(\boldsymbol{x}))$, where the saturation level is considered in random binary search tree [21,49], and their bounds can be rewritten (with our notation) as follows:

$$\Pr[h(C(\boldsymbol{x})) < (c^* - \epsilon)\ln k] \leq O(\log(k) k^{(c^*-\epsilon)\ln(2e/(c^*-\epsilon))-1}) .$$

Here, $c^* = 0.3733\ldots$ is the unique solution of $c\ln(2e/c) = 1$ $(c < 1)$ and $\epsilon < c^*$. As can be seen, Lemma 2 makes better estimations of $h(C(\boldsymbol{x}))$ with larger probability. The detailed proof of Lemma 2 is presented in Section 6.2.

Given cell $C(\boldsymbol{x})$, we define its diameter as $\nu(C(\boldsymbol{x})) = \max_{\boldsymbol{x},\boldsymbol{x}' \in C(\boldsymbol{x})}\{\|\boldsymbol{x} - \boldsymbol{x}'\|\}$, and bound $\nu(C(\boldsymbol{x}))$ in probability as follows:

Lemma 3. *For integer $k \geq 2$, real $\epsilon > -1$ and instance $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\Pr\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\frac{\sqrt{d}}{k^{1/8d}}\right] \leq \frac{ed}{(1+\epsilon)k^{1/8d}} ,$$

where the probability takes over the random selections of splitting leaves, dimensions and positions.

This lemma shows that, for every instance $\boldsymbol{x} \in \mathcal{X}$, the diameter of rectangle cell of $C(\boldsymbol{x})$ can be upper bounded by $(1+\epsilon)\sqrt{d}/k^{1/8d}$ with probability at least $1 - ed/(1+\epsilon)k^{1/8d}$. We also have $\nu(C(\boldsymbol{x})) \to 0$ in probability as $k \to +\infty$. For simplicity, we do not formalize the random selections of splitting leaves, dimensions and positions in Lemma 3, while the detailed formalization and proof are presented in Section 6.3.

Recall that there are k disjoint rectangular cells $C_1, C_2, \ldots, C_k$ during the construction of purely random tree with $k-1$ iterations. We could bound the classification error over each rectangular cell, and the detailed proof is given in Section 6.4.

Lemma 4. *Let $C_1, C_2, \ldots, C_k$ be the k disjoint rectangular cells associating with the leaves of randomized tree, and $f_{\Theta,S_n}(\boldsymbol{x})$ denotes the classifier generated by random tree. For L-Lipschitz conditional probability $\eta_1(\boldsymbol{x})$ and for every $i \in [k]$, we have*

$$\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i\right]\Pr[\boldsymbol{x} \in C_i]$$
$$\leq 2L\nu(C_i)\Pr[\boldsymbol{x} \in C_i] + E_{\boldsymbol{x}}[\min\{\eta_1(\boldsymbol{x}), 1-\eta_1(\boldsymbol{x})\} | \boldsymbol{x} \in C_i]\Pr[\boldsymbol{x} \in C_i] + \sqrt{\Pr[\boldsymbol{x} \in C_i]/n} + 3/n .$$

Based on Lemmas 2-4, we could present the convergence rates of purely random forests for binary classification, and the detailed proof is presented in Section 6.5.

Theorem 1. *Let $f_m(\boldsymbol{x})$ be the random forests classifier by applying purely random tree to training data S_n of k leaves $(k \geq 2)$. Under the L-Lipschitz assumption over conditional probability $\eta_1(\boldsymbol{x})$, we have*

$$R^*_{\mathcal{D}} \leq E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq R^*_{\mathcal{D}} + \frac{4\sqrt{2eLd^{3/2}}}{k^{1/8d}} + 2\sqrt{\frac{k}{n}} + \frac{6k}{n} .$$

From this theorem, we obtain a convergence rate $O(n^{-1/(8d+2)})$ of purely random forests for binary classification, by selecting leaves number $k = O(n^{4d/(4d+1)})$. To the best of our knowledge, this presents the first finite-sample converge rate of purely random forests for binary classification. Also, it is easy to observe that

$$E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \to R^*_{\mathcal{D}} \quad \text{as} \quad k \to +\infty \quad \text{and} \quad k/n \to 0,$$

which recovers the consistency result of random forests for classification [9, Theorem 2].

We further study the effects of different splitting mechanisms during the construction of random forests. For example, how about the convergence rates for different selections of splitting leaves, dimensions and positions? Here, we consider

Algorithm 1 A simplified variant of Breiman's original random tree [12].

Input: Training sample S_n and leaves number k.
Output: A random tree
Initialize: Set $\mathcal{P} = \{[0,1]^d\}$ and $n_{\text{leaf}} = 1$.
while $n_{\text{leaf}} < k$ **and** $\mathcal{P}$ is not empty **do**
  Let C be the first rectangle cell in $\mathcal{P}$, and remove it from $\mathcal{P}$.
  if All training examples in C have the same label (including less than one example) **then**
    Do nothing and the cell C will not be split any more.
  else
    Select a dimension Y at random, uniformly over dimensions along which the side length is maximal in the cell C.
    Split cell C along Y at the midpoint of the chosen side, called C_L, C_R two resulting cells.
    Update $\mathcal{P}$ by appending C_L and C_R, and $n_{\text{leaf}} \leftarrow n_{\text{leaf}} + 1$.
  end if
end while

purely random forests with midpoint splits, where midpoint splits have been well-studied for random forests in regression [3,8,36]. Formally, a purely random tree with midpoint splits can be constructed as follows. The root of random partition is $[0,1]^d$ itself. The following procedure is repeated $k-1$ iterations for some pre-defined parameter $k \geq 2$ in advance.

- A split leaf is selected at random, uniformly over all leaves at the current iteration.
- Once the leaf is selected, a split dimension is selected at random, uniformly over $[d]$.
- The leaf is split along the split dimension at the midpoint of the chosen side.

Given individual random tree classifiers $f_{S_n,\Theta_1}(\boldsymbol{x})$, $f_{S_n,\Theta_2}(\boldsymbol{x}), \ldots, f_{S_n,\Theta_m}(\boldsymbol{x})$, the random forests classifier takes a majority vote over m random trees. We present a convergence rate of purely random forests with midpoint splits for binary classification as follows:

Theorem 2. *Let $f_m(\boldsymbol{x})$ be the random forests classifier by applying purely random tree with midpoint splits to training data S_n of k leaves ($k \geq 2$). Under the L-Lipschitz assumption over conditional probability $\eta_1(\boldsymbol{x})$, we have*

$$R^*_{\mathcal{D}} \leq E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq R^*_{\mathcal{D}} + \frac{8L^{3/5}d^{7/10}}{k^{1/3.87d}} + 2\sqrt{\frac{k}{n}} + \frac{6k}{n}\,.$$

Based on this theorem, we get a convergence rate $O(n^{-1/(3.87d+2)})$ of purely random forests with midpoint splits for binary classification, by selecting leaves number $k = O(n^{3.87d/(3.87d+2)})$. As can be seen, we achieve a better convergence rate by considering the midpoint splitting mechanism during the construction of purely random forests, and an intuitive explanation is that midpoint splits yield smaller rectangle cells. The detailed proof of Theorem 2 is presented in Section 6.6.

4. Convergence rates of the simplified random forests for binary classification

This section also focuses on binary classification and presents the convergence analysis towards Breiman's original random forests [12]. We follow the procedures of Breiman's random forests, but with different mechanisms on the selections of splitting dimensions and positions due to technical analysis challenges. Algorithm 1 presents a detailed description of the simplified variant of Breiman's random forests.

We introduce a structural list $\mathcal{P}$ to store leaves for further splitting, which aims to keep the leaves split in successive layers. Such mechanism is essentially the same as that of random forests for regression [53]. At each iteration, the first leaf is selected and removed from $\mathcal{P}$, and it will not be split if all training examples have the same label in the leaf (including less than one example in the leaf). For a split leaf, we select a dimension at random, uniformly over dimensions along which the side length is maximal in the leaf, and then split the leaf along the dimension at the midpoint of the chosen side. We finally update list $\mathcal{P}$ by appending two resulting leaves.

A leaf (rectangle cell) will not be split in Algorithm 1 if all training examples have the same label in this leaf. Such stopping-splitting criterion is different from purely random forests [11] and Mondrian forests [39,46], where the growth of individual random tree is independent of training sample. In addition, it is also different from random forests regression [53], where a leaf will not be split only when the leaf has exactly one training example.

Let $f_{S_n,\Theta_1}(\boldsymbol{x})$, $f_{S_n,\Theta_2}(\boldsymbol{x}), \ldots, f_{S_n,\Theta_m}(\boldsymbol{x})$ denote m individual random tree classifiers according to Algorithm 1. Then, random forests classifier takes a majority vote over m random trees, that is, $f_m(\boldsymbol{x}) = \mathbb{I}[\sum_{i=1}^m f_{S_n,\Theta_i}(\boldsymbol{x}) \geq m/2]$. We present a convergence rate of the simplified variant of random forests for binary classification as follows:

Theorem 3. *Let $f_m(\boldsymbol{x})$ be the random forests classifier by applying Algorithm 1 to training data S_n of k leaves ($k \geq 2$ and $n \geq 4$). Under the L-Lipschitz assumption over conditional probability $\eta_1(\boldsymbol{x})$, we have*

$$R^*_{\mathcal{D}} \leq E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq R^*_{\mathcal{D}} + 4\sqrt{\frac{k\ln n}{n}} + 2\sqrt[4]{\frac{4k^3\ln n}{n^3}} + \frac{12k}{n} + 4\sqrt{\frac{k}{n}} + \frac{32L\sqrt{d}}{k^{1/d}}\,.$$

We get a convergence rate $O(n^{-1/(d+2)}(\ln n)^{1/(d+2)})$ for random forests based on Algorithm 1, by selecting leaves number $k = O((n/\ln n)^{2d/(d+2)})$. This presents a significantly better convergence rate than that of purely random forests due to different splitting mechanisms and stopping-splitting criteria. The detailed proof of Theorem 3 is given in Section 6.7.

Under the L-Lipschitz assumption, it is well-known [5,64] that the minimax rate is of $O(n^{-1/(d+2)})$ for the optimal plug-in classifiers $f(\boldsymbol{x}) = \mathbb{I}[\hat{\eta}_1(\boldsymbol{x}) \geq 1/2]$, where $\hat{\eta}_1(\boldsymbol{x})$ is the estimated conditional probability. Hence, our simplified variant of random forests reaches the minimax convergence rate, except for a factor $(\ln n)^{1/(1+d)}$, as that of the optimal plug-in classifiers, despite random forests are not plug-in classifiers. This is because random forests take a majority vote over the predictions of individual random trees, rather than the estimation of conditional probability.

Essentially, the stopping criteria in Algorithm 1 can be viewed as a pre-pruning of decision trees, which stops the tree-building process early when all samples have the same label in a cell, and avoids producing leaves with smaller samples. Such mechanism has been well-known as a regularization to prevent overfitting for decision trees, while our work further shows a better convergence rate for random forests. This is because we could reduce the splittings of leaves and make use of exponential inequality for the convergence analysis of random forests when all samples have the same label. In addition, the stopping criterion makes our Algorithm 1 different from ordinary histograms, where our algorithm could dynamically adjust the partition according to training data, while more partitions as in ordinary histograms may lead to overfitting or inconsistency for random forests.

Breiman's original random forests [12] took some splitting criteria, such as information gain and entropy, to select the best-split dimension and position, which correlates the randomization process with data-dependent tree structure. This is the main difference from our simplified variant of random forests in Algorithm 1. Intuitively, such correlation could yield tighter convergence rates of random forests for classification, whereas it remains a big challenge to present theoretical analysis from a technical view. One possible solution is to introduce two samples for splitting and predicting of random forests, respectively; for example, Wager and Athey [59] introduced the honesty and regularity assumptions based on two samples to analyze random forest regression. This also remains some technical difficulties on how to merge two samples into one sample for Breiman's original random forests based on majority voting in classification; therefore, it is still a long way to theoretically understanding the mechanism of Breiman's original random forests.

We now make some assumptions over structural data, which could yield tighter convergence rate for the simplified variant of random forests. Suppose that there is a constant $k_0 \geq 2$, such that the output random trees from Algorithm 1 have at most k_0 leaves with all training examples in each leaf having the same label. Based on such assumption, we present a convergence rate of the simplified variant of random forests for binary classification as follows.

Theorem 4. *Suppose that there is a constant $k_0 \geq 2$, such that the output random trees from Algorithm 1 have at most k_0 leaves with all training examples having the same label in each leaf. Let $f_m(\boldsymbol{x})$ be the random forests classifier by applying Algorithm 1 to training data S_n. We have*

$$E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq 4\sqrt{\frac{k_0 \ln n}{n}} + 2\sqrt[4]{\frac{4k_0^3 \ln n}{n^3}} + 2\sqrt{\frac{k_0}{n \ln n}} + \frac{6k_0}{n} .$$

From this theorem, we could achieve the tighter convergence rate $O(\sqrt{\ln n/n})$ of the simplified variant of random forests for classification, which is independent of dimension d. The main improvements are owing to the data assumption and stopping criteria in Algorithm 1 from a technical view, where we could make use of exponential inequality for all cells to derive the tighter convergence rate when all samples have the same label for each cell.

Theorem 4 sheds lights on Breiman's original random forests [12] of tighter convergence rates via correlating randomization process and data-dependent tree structure. The assumption in Theorem 4 is relevant to algorithm, while it still holds for some irrelevant cases; for example, Algorithm 1 satisfies such assumption when the data is separable and the separable hyperplane is parallel to axis. The detailed proof of Theorem 4 is presented in Section 6.8.

5. Convergence rates of random forests for multi-class learning

This section focuses on multi-class learning with label space $\mathcal{Y} = [\tau]$ for $\tau \geq 3$. Recall that $\eta_j(\boldsymbol{x})$ denotes the conditional probability of $y = j$ over instance $\boldsymbol{x}$ w.r.t. distribution $\mathcal{D}$ for $j \in [\tau]$, and it holds that $\eta_1(\boldsymbol{x}) + \eta_2(\boldsymbol{x}) + \cdots + \eta_\tau(\boldsymbol{x}) = 1$. We also make L-Lipschitz assumption for multi-class learning, that is, for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, there exists a constant $L > 0$ such that

$$|\eta_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\| \quad \text{for every } j \in [\tau] .$$

Such assumption has been studied for multi-class learning [24,32,37,47], which can be viewed as a direct extension of Lipschitz assumption of binary classification to multi-class learning. To our knowledge, it still remains an open problem on the convergence rate of multi-class algorithms under only the L-Lipschitz assumption.

Given m individual purely random trees $f_{S_n,\Theta_1}(\boldsymbol{x})$, $f_{S_n,\Theta_2}(\boldsymbol{x}), \ldots, f_{S_n,\Theta_m}(\boldsymbol{x})$, the random forests classifier takes a majority vote over those random trees for multi-class learning, that is,

$$f_m(\boldsymbol{x}) = \arg\max_{j\in[\tau]} \left\{ \sum_{i=1}^{m} \mathbb{I}\left[f_{S_n,\Theta_i}(\boldsymbol{x}) = j \right] \right\}, \tag{2}$$

where ties are broken arbitrarily.

We begin with some helpful lemmas on the convergence analysis of random forests for multi-class learning. Given a rectangle cell C, we assume that there are n' training examples falling in cell C, and denote by $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_{n'}, y_{n'})$ without loss of generality. Recall that the conditional probabilities $\eta_i(\boldsymbol{x}_j) = \Pr[y_j = i|\boldsymbol{x}_j]$ ($j \in [n']$ and $i \in [\tau]$) according to distribution $\mathcal{D}$. For instance $\boldsymbol{x} \in C$, let $\eta_i(\boldsymbol{x}) = \Pr[y = i|\boldsymbol{x}]$ denote the conditional probabilities for $i \in [\tau]$.

Conditioned on instances $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n'}$, their labels y and y_j ($j \in [n']$) can be viewed from the following multi-nominal distributions, respectively,

$$y \sim \mathcal{M}\left(\eta_1(\boldsymbol{x}), \eta_2(\boldsymbol{x}), \cdots, \eta_\tau(\boldsymbol{x})\right) \quad \text{and} \quad y_j \sim \mathcal{M}\left(\eta_1(\boldsymbol{x}_j), \eta_2(\boldsymbol{x}_j), \cdots, \eta_\tau(\boldsymbol{x}_j)\right) \quad \text{for} \quad j \in [n'] .$$

For simplicity, we denote by

$$\vartheta_i = \sum_{j=1}^{n'} I[y_j = i] \quad \text{and} \quad \rho_i = \sum_{j=1}^{n'} \frac{\eta_i(\boldsymbol{x}_j)}{n'} \quad \text{for} \quad i \in [\tau] . \tag{3}$$

We present the following lemma to decompose a multi-class learning problem into a series of individual 3-class learning problems, and the detailed proof is presented in Section 6.9.

Lemma 5. *For integer $\tau \geq 3$, let $\vartheta_1, \vartheta_2, \cdots, \vartheta_\tau$ and $\rho_1, \rho_2, \cdots, \rho_\tau$ be defined by Eqn. (3). If $\rho_1 \geq \max(\rho_2, \ldots, \rho_\tau)$, then we have*

$$\rho_1 \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[\tau]\setminus\{1\}}\{\vartheta_j\}\right] - \sum_{i=2}^{\tau} \rho_i \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_i = \max_{j\in[\tau]}\{\vartheta_j\}\right] \leq \sum_{i=2}^{\tau} (\rho_1 - \rho_i) \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_i] .$$

We further present the following lemma to bound $(\rho_1 - \rho_i)\Pr_{y_1,\ldots,y_m}[\vartheta_1 < \vartheta_i]$ for $2 \leq i \leq \tau$, and the detailed proof is presented in Section 6.10.

Lemma 6. *Let $X_1, X_2, \cdots, X_{n'}$ be n' independent random variables with $X_i \in \{-1, 0, +1\}$. Let $\eta_i^+ = \Pr[X_i = +1]$ and $\eta_i^- = \Pr[X_1 = -1]$, and we have $\Pr[X_i = 0] = 1 - \eta_i^+ - \eta_i^-$. Write $\rho^+ = \sum_{i=1}^{n'} \eta_i^+/n'$ and $\rho^- = \sum_{i=1}^{n'} \eta_i^-/n'$. If $\rho^+ > \rho^-$, then we have*

$$(\rho^+ - \rho^-) \Pr_{X_1,\ldots,X_{n'}}\left[\sum_{i=1}^{n'} X_i < 0\right] \leq \frac{1}{\sqrt{en'}} .$$

Based on Lemmas 5 and 6, we have

Lemma 7. *Let $C_1, C_2, \ldots, C_k$ be the k disjoint rectangular cells associating with the leaves of randomized tree, and $f_{\Theta,S_n}(\boldsymbol{x})$ denotes the classifier generated by random tree for multi-class learning. Under the L-Lipschitz assumption over $\eta_j(\boldsymbol{x})$ ($j \in [\tau]$), we have, for every C_i ($i \in [k]$),*

$$\begin{aligned}
&\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i\right] \Pr[\boldsymbol{x} \in C_i] \\
&\leq (\tau+2) L \nu(C_i) \Pr[C_i] + E_{\boldsymbol{x}}\left[\min_{j\in[\tau]}\{1 - \eta_j(\boldsymbol{x})\} \middle| \boldsymbol{x} \in C_i\right] \Pr[C_i] + \tau\sqrt{\frac{2}{en}\Pr[C_i]} + \frac{3}{n} .
\end{aligned}$$

This lemma presents a key ingredient to exploit the convergence rates of random forests for multi-class learning, and it is also helpful to analyze the convergence rates of other multi-class algorithms such as nearest neighbor, which may present independent interests in the machine learning community. The detailed proof is given in Section 6.11.

Based on Lemmas 5-7, we could study the convergence rates of different variants of random forests for multi-class learning, and all detailed proofs are presented in Section 6.12. We first present the convergence rates of purely random forests for multi-class learning as follows:

Theorem 5. *Let $f_m(\boldsymbol{x})$ be the random forests classifier for multi-class learning, by applying the purely random tree to training data S_n of k leaves ($k \geq 2$). Under the L-Lipschitz assumption over conditional probability $\eta_j(\boldsymbol{x})$ ($j \in [\tau]$), we have*

$$R_{\mathcal{D}}^* \leq E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq R_{\mathcal{D}}^* + \frac{2\tau\sqrt{(\tau+2)eLd^{3/2}}}{k^{1/8d}} + \tau^2\sqrt{\frac{2k}{en}} + \frac{3\tau k}{n} .$$

This theorem presents a convergence rate $O(n^{-1/(8d+2)})$ of purely random forests for multi-class learning, by selecting leaves number $k = O(n^{4d/(4d+1)})$. As can be seen, we achieve the same convergence of random forests for multi-class learning as that of binary classification (Theorem 1), but with different constants.

We present the convergence analysis of purely random forests with midpoint splitting for multi-class learning as follows:

Theorem 6. *Let $f_m(\boldsymbol{x})$ be the random forests classifier for multi-class learning, by applying the purely random tree with midpoint splits to training data S_n of k leaves ($k \geq 2$). Under the L-Lipschitz assumption over conditional probability $\eta_j(\boldsymbol{x})$ ($j \in [\tau]$), we have*

$$R^*_{\mathcal{D}} \leq E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq R^*_{\mathcal{D}} + \frac{3\tau(\tau+2)^{3/5}L^{3/5}d^{7/10}}{k^{1/3.87d}} + \tau^2\sqrt{\frac{2k}{en}} + \frac{3\tau k}{n}.$$

Based on this theorem, we get a convergence rate $O(n^{-1/(3.87d+2)})$ of purely random forests with midpoint splits for multi-class learning, by selecting leaves number $k = O(n^{3.87d/(3.87d+2)})$. As can be seen, we achieve the same convergence rate of random forests for multi-class learning as that of binary classification (Theorem 2), but with different constants.

We further study the simplified variant of Breiman's random tree for multi-class learning, as shown by Algorithm 1 in Section 4, and take the final prediction by majority vote for multiple classes. We present convergence analysis as follows:

Theorem 7. *Let $f_m(\boldsymbol{x})$ be the random forests classifier for multi-class learning, by applying Algorithm 1 to training data S_n of k leaves ($k \geq 2$ and $n \geq 4$). Under the L-Lipschitz assumption over conditional probability $\eta_j(\boldsymbol{x})$ ($j \in [\tau]$), we have*

$$E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq R^*_{\mathcal{D}} + 2\tau\sqrt{\frac{k\ln n}{n}} + \tau\sqrt[4]{\frac{4k^3\ln n}{n^3}} + \frac{6\tau k}{n} + \tau^2\sqrt{\frac{2k}{en}} + \tau\sqrt{\frac{k}{n}} + \frac{8(\tau^2+2\tau)L\sqrt{d}}{k^{1/d}}.$$

This theorem presents a convergence rate $O(n^{-1/(d+2)}(\ln n)^{1/(d+2)})$ for random forests based on Algorithm 1 for multi-class learning, by selecting leaves number $k = O((n/\ln n)^{2d/(d+2)})$. As can be seen, we achieve the same convergence rate of random forests for multi-class learning as that of binary classification (Theorem 3), but with different constants.

We finally study the convergence rate of random forests under some assumptions over structural data for multi-class learning. Suppose that there is a constant $k_0 \geq 2$, such that the output random trees from Algorithm 1 have at most k_0 leaves with all training examples in each leaf having the same label. Based on such assumption, we present a convergence rate of the simplified variant of random forests for multi-class learning.

Theorem 8. *Suppose that there is a constant $k_0 \geq 2$, such that the output random trees from Algorithm 1 have at most k_0 leaves with all training examples having the same label in each leaf. Let $f_m(\boldsymbol{x})$ be the random forests classifier for multi-class learning, by applying Algorithm 1 to training data S_n. We have*

$$E_{S_n,\Theta_1,\ldots,\Theta_m}[R_{\mathcal{D}}(f_m)] \leq 2\tau\sqrt{\frac{k_0\ln n}{n}} + \tau\sqrt[4]{\frac{4k_0^3\ln n}{n^3}} + \tau\sqrt{\frac{k_0}{n\ln n}} + \frac{3\tau k_0}{n}.$$

This theorem presents a convergence rate $O(\sqrt{\ln n/n})$ of the simplified variant of random forests in multi-class learning, which achieves the same convergence rate as that of binary classification (Theorem 4), but with different constants.

6. Proofs

We begin with a series of useful lemmas before the detailed proofs of our main results. The following lemma is a variant of Chernoff bounds from [45].

Lemma 8. *Let $X_1, X_2, \ldots, X_m$ be m independent random variables with $X_i \in [0,1]$. Denote by $X = \sum_{i=1}^m X_i$ and $p = \sum_{i=1}^m E[X_i]$. We have*

$$\Pr_{X_1,\ldots,X_m}[X < (1-\delta)p] \leq \exp(-p\delta^2/2).$$

Lemma 9. *For any integer $k \geq 2$, we have*

$$\ln k \leq \sum_{i=1}^{k-1}\frac{1}{i} \leq 1 + \ln(k-1).$$

Proof. For any integer $i > 0$, we have

$$\frac{1}{t} \le \frac{1}{i} \text{ for } t \in [i, i+1] \qquad \text{and} \qquad \frac{1}{i} \le \frac{1}{t} \text{ for } t \in [i-1, i].$$

It follows that

$$\ln k = \int_1^k \frac{1}{t} dt \le \sum_{i=1}^{k-1} \frac{1}{i} \le 1 + \int_1^{k-1} \frac{1}{t} dt = 1 + \ln(k-1) ,$$

which completes the proof. □

Lemma 10. *For integers $k \ge 2$ and $d \ge 2$, we have*

$$\sum_{i=1}^{k-1} \ln\left(1 - \frac{3}{4id}\right) \ge -\frac{9 + 3\ln(k-1)}{4d} , \tag{4}$$

$$\sum_{i=1}^{k-1} \ln\left(1 - \frac{1}{2id}\right) \ge -\frac{3 + \ln(k-1)}{2d} . \tag{5}$$

Proof. We first have

$$\begin{aligned}
\sum_{i=1}^{k-1} \ln\left(1 - \frac{3}{4id}\right) &\ge \ln\left(1 - \frac{3}{4d}\right) + \int_1^{k-1} \ln\left(1 - \frac{3}{4dt}\right) dt \\
&= \ln\left(1 - \frac{3}{4d}\right) + \left[t \ln\left(1 - \frac{3}{4dt}\right)\right]_1^{k-1} - \int_1^{k-1} \frac{3}{4dt - 3} dt \\
&= (k-1) \ln\left(1 - \frac{3}{4d(k-1)}\right) - \frac{3}{4d} \ln\left(k - 1 - \frac{3}{4d}\right) + \frac{3}{4d} \ln\left(1 - \frac{3}{4d}\right) .
\end{aligned}$$

It is easy to observe $\ln(k - 1 - 3/4d) \le \ln(k-1)$, and

$$(k-1) \ln\left(1 - \frac{3}{4d(k-1)}\right) + \frac{3}{4d} \ln\left(1 - \frac{3}{4d}\right) \ge -\frac{3}{2d} - \frac{9}{8d^2} \ge \frac{-33}{16d} \ge \frac{-9}{4d} ,$$

by using $\ln(1-x) \ge -2x$ for $x \in [0, 1/2]$ and $d \ge 2$. Eqn. (4) holds by simple calculations.

For Eqn. (5), we similarly have

$$\begin{aligned}
\sum_{i=1}^{k-1} \ln\left(1 - \frac{1}{2id}\right) &\ge \ln\left(1 - \frac{1}{2d}\right) + \int_1^{k-1} \ln\left(1 - \frac{1}{2dt}\right) dt \\
&= \ln\left(1 - \frac{1}{2d}\right) + \left[t \ln\left(1 - \frac{1}{2dt}\right)\right]_1^{k-1} - \int_1^{k-1} \frac{1}{2dt - 1} dt \\
&= (k-1) \ln\left(1 - \frac{1}{2d(k-1)}\right) - \frac{1}{2d} \ln\left(k - 1 - \frac{1}{2d}\right) + \frac{1}{2d} \ln\left(1 - \frac{1}{2d}\right) .
\end{aligned}$$

It follows that, by using $\ln(1-x) > -2x$ for $x \in [0, 1/2]$ and for $d \ge 2$,

$$(k-1) \ln\left(1 - \frac{1}{2d(k-1)}\right) + \frac{1}{2d} \ln\left(1 - \frac{1}{2d}\right) \ge -\frac{1}{d} - \frac{1}{2d^2} \ge -\frac{5}{4d} \ge -\frac{3}{2d} ,$$

which completes the proof of Eqn. (5) by simple calculations. □

6.1. Proof of Lemma 1

For every $\boldsymbol{x} \in \mathcal{X}$, we denote by $\eta^*(\boldsymbol{x}) = \max_{j\in[\tau]}\{\eta_j(\boldsymbol{x})\}$. This follows that, from the definitions of $R_{\mathcal{D}}(h)$ and $R^*_{\mathcal{D}}$,

$$E_{\Theta}\left[R_{\mathcal{D}}\left(f_{S_n,\Theta}\right)\right] - R^*_{\mathcal{D}} = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\sum_{j=1}^{\tau}\Pr_{\Theta}\left[f_{S_n,\Theta}(\boldsymbol{x}) = j\right]\left(\eta^*(\boldsymbol{x}) - \eta_j(\boldsymbol{x})\right)\right].$$

In a similar manner, we have

$$E_{\Theta_1,\ldots,\Theta_m}\left[R_{\mathcal{D}}\left(f_m\right)\right] - R^*_{\mathcal{D}} = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\sum_{j=1}^{\tau}\Pr_{\Theta_1,\ldots,\Theta_m}\left[f_m(\boldsymbol{x}) = j\right]\left(\eta^*(\boldsymbol{x}) - \eta_j(\boldsymbol{x})\right)\right].$$

For $j \in [\tau]$, we have, by Markov's inequality,

$$\begin{aligned}\Pr_{\Theta_1,\ldots,\Theta_m}\left[f_m(\boldsymbol{x}) = j\right] &\le \Pr_{\Theta_1,\ldots,\Theta_m}\left[\sum_{i=1}^{m}\mathbb{I}\left[f_{S_n,\Theta_j}(\boldsymbol{x}) = j\right] \ge \frac{m}{\tau}\right]\\ &\le \frac{\tau}{m}\sum_{i=1}^{m}E_{\Theta_i}\left[\mathbb{I}\left[f_{S_n,\Theta_i}(\boldsymbol{x}) = j\right]\right] = \frac{\tau}{m}\sum_{i=1}^{m}\Pr_{\Theta_i}\left[f_{S_n,\Theta_i}(\boldsymbol{x}) = j\right] = \tau\Pr_{\Theta}\left[f_{S_n,\Theta}(\boldsymbol{x}) = j\right],\end{aligned}$$

which completes the proof. □

6.2. Proof of Lemma 2

For $k \ge 2$, let $X_1, X_2, \cdots, X_{k-1}$ denote $k-1$ independent Bernoulli random variables with $X_i \sim \mathcal{B}(1/i)$ for $i \in [k-1]$. For any instance $\boldsymbol{x} \in \mathcal{X}$, we have

$$h(C(\boldsymbol{x})) = \sum_{i=1}^{k-1}X_i \quad \text{and} \quad E_{X_1,X_2,\cdots,X_k}[h(C(\boldsymbol{x}))] = \sum_{i=1}^{k-1}\frac{1}{i}.$$

Based on Lemma 9, we have

$$\ln k \le E[h(C_i)] \le 1 + \ln(k-1).$$

It follows that, for any $\lambda < 0$ and by Markov's inequality,

$$\Pr\left[\sum_{i=1}^{k-1}X_i - E[X_i] \le -\epsilon\right] \le \exp\left(\lambda\epsilon - \lambda\sum_{i=1}^{k-1}\frac{1}{i}\right)E\left[\exp\left(\sum_{i=1}^{k-1}\lambda X_i\right)\right]. \tag{6}$$

We have, from the independence of random variables $X_1, X_2, \ldots, X_{k-1}$ with $X_i \sim \mathcal{B}(1/i)$,

$$E\left[\exp\left(\sum_{i=1}^{k-1}\lambda X_i\right)\right] = \prod_{i=1}^{k-1}E\left[\exp\left(\lambda X_i\right)\right] = \exp\left(\sum_{i=1}^{k-1}\ln\left(1 - \frac{1}{i} + \frac{1}{i}e^{\lambda}\right)\right). \tag{7}$$

Denote by

$$g_i(\lambda) = \ln\left(1 - \frac{1}{i} + \frac{1}{i}e^{\lambda}\right),$$

and we have

$$\begin{aligned}g_i'(\lambda) &= \frac{e^{\lambda}}{i - 1 + e^{\lambda}},\\ g_i''(\lambda) &= \frac{e^{\lambda}}{i - 1 + e^{\lambda}} - \frac{e^{2\lambda}}{(i - 1 + e^{\lambda})^2} \le \frac{e^{\lambda}}{i - 1 + e^{\lambda}} < \frac{1}{i} \quad \text{for} \quad \lambda < 0.\end{aligned}$$

Based on the Taylor expansions, we have

$$g_i(\lambda) \le g_i(0) + \lambda g_i'(0) + \frac{\lambda^2}{2i} = \frac{\lambda}{i} + \frac{\lambda^2}{2i}.$$

Combining with Eqns. (6) and (7), this yields

$$\Pr\left[\sum_{i=1}^{k-1} X_i - \sum_{i=1}^{k-1}\frac{1}{i} \le -\epsilon\right] \le \exp\left(\lambda\epsilon + \frac{\lambda^2}{2}\sum_{i=1}^{k-1}\frac{1}{i}\right).$$

By setting $\lambda = -\epsilon/\sum_{i=1}^{k-1} 1/i$, we have

$$\Pr\left[\sum_{i=1}^{k-1} X_i - \sum_{i=1}^{k-1}\frac{1}{i} \le -\epsilon\right] \le \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^{k-1} 1/i}\right).$$

It follows that, by setting $\epsilon = \epsilon\sum_{i=1}^{k-1} 1/i$ and from Lemma 9,

$$\Pr_{X_1, X_2, \ldots, X_k}[h(C(\boldsymbol{x})) < (1-\epsilon)\ln k] \le k^{-\epsilon^2/2}.$$

In a similar manner, we have, for any $\lambda > 0$,

$$\Pr\left[\sum_{i=1}^{k-1}(X_i - E[X_i]) \ge \epsilon\right] \le \exp\left(-\lambda\epsilon - \lambda\sum_{i=1}^{k-1}\frac{1}{i}\right) E\left[\exp\left(\sum_{i=1}^{k-1}\lambda X_i\right)\right] \le \exp\left(-\lambda\epsilon + \frac{\lambda^2}{2}\sum_{i=1}^{k-1}\frac{1}{i}\right).$$

By setting $\lambda = \epsilon/\sum_{i=1}^{k-1} 1/i$, we have

$$\Pr\left[\sum_{i=1}^{k-1} X_i - \sum_{i=1}^{k-1}\frac{1}{i} \ge \epsilon\right] \le \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^{k-1} 1/i}\right).$$

We further set $\epsilon = \epsilon\sum_{i=1}^{k-1} 1/i$ in the above, and this follows that, from Lemma 9,

$$\Pr_{X_1, X_2, \ldots, X_{k-1}}[h(C(\boldsymbol{x})) \ge (1+\epsilon)(1+\ln(k-1))] \le k^{-\epsilon^2/2},$$

which completes the proof. □

6.3. Proof of Lemma 3

Given any instance $\boldsymbol{x} \in \mathcal{X}$, recall that $C(\boldsymbol{x})$ denotes the rectangular cell containing instance $\boldsymbol{x}$, and $X_1, X_2, \cdots, X_{k-1}$ characterize the random events that the node containing instance $\boldsymbol{x}$ was selected for splitting in the construction of random tree, where $X_i \sim \mathcal{B}(1/i)$.

For $j \in [d]$, let $\ell_j(C(\boldsymbol{x}))$ denote the length of the j-th dimension of rectangular cell $C(\boldsymbol{x})$, and it is necessary to introduce the following random variables to analyze $\ell_j(C(\boldsymbol{x}))$.

- Let $Y_{1,j}, Y_{2,j}, \cdots, Y_{k-1,j}$ denote $k-1$ Bernoulli random variables such that $Y_{i,j} \sim \mathcal{B}(1/d)$ for $i \in [k-1]$. Here, $Y_{i,j} = 1$ denotes the random event that the j-th coordinate of the node, that contains the instance $\boldsymbol{x}$, is selected for random partition under the condition $X_i = 1$. We use $Y_{i,j}$ to illustrate the selection of coordinates with identical probability.
- Let $U_{1,j}, U_{2,j}, \cdots, U_{k-1,j}$ denote $k-1$ random variables with uniform distribution over $[0,1]$, i.e., $U_{i,j} \sim \mathcal{U}[0,1]$ for $i \in [k-1]$. Here, we use random variable $U_{i,j}$ to characterize the uniform and random splitting of the j-th coordinate of the node containing $\boldsymbol{x}$ under the condition $X_i Y_{i,j} = 1$ during the i-th construction of random tree.

It is easy to upper and lower bound $\ell_j(C(\boldsymbol{x}))$ as follows:

$$\prod_{i=1}^{k-1}\min(1-U_{i,j}, U_{i,j})^{X_i Y_{i,j}} \le \ell_j(C(\boldsymbol{x})) \le \prod_{i=1}^{k-1}\max(1-U_{i,j}, U_{i,j})^{X_i Y_{i,j}}. \tag{8}$$

Lemma 11. *For $k \ge 2$ and $j \in [d]$, we have*

$$E\left[\prod_{i=1}^{k-1}(\max(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}}\right] = \prod_{i=1}^{k-1}\left(1 - \frac{1}{4id}\right) \le \exp\left(-\frac{\ln k}{4d}\right), \tag{9}$$

$$E\left[\prod_{i=1}^{k-1}(\min(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}}\right] = \prod_{i=1}^{k-1}\left(1 - \frac{3}{4id}\right) \ge \exp\left(-\frac{9+3\ln(k-1)}{4d}\right), \tag{10}$$

and we also have, for any instance $\boldsymbol{x} \in \mathcal{X}$,

$$\exp\left(-\frac{9+3\ln(k-1)}{4d}\right) \le E[\ell_j(C(\boldsymbol{x}))] \le \exp\left(-\frac{\ln k}{4d}\right) .$$

Here, all expectations take over independent random variables $X_1, \ldots, X_{k-1}$, $Y_{1,j}, \ldots, Y_{k-1,j}$ and $U_{1,j}, \ldots, U_{k-1,j}$ with $X_i \sim \mathcal{B}(1/i)$, $Y_{i,j} \sim \mathcal{B}(1/d)$ and $U_{i,j} \sim \mathcal{U}(0,1)$ for $i \in [k-1]$.

Proof. For Eqn. (9), we first write $Z_{i,j} = (\max(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}}$, and it follows that

$$E_{X_i,Y_{i,j},U_{i,j}}[Z_{i,j}] = 1 - \frac{1}{id} + \frac{1}{id} E_{U_{i,j}}[\max(U_{i,j}, 1-U_{i,j})] = 1 - \frac{1}{4id} ,$$

by using the fact

$$E_{U_{i,j}}[\max(U_{i,j}, 1-U_{i,j})] = \int_0^1 \max(U_{i,j}, 1-U_{i,j}) dU_{i,j} = \int_0^{1/2} (1-U_{i,j}) dU_{i,j} + \int_{1/2}^1 U_{i,j} dU_{i,j} = \frac{3}{4} .$$

It holds that, from Lemma 9 and by using the fact $1 - x \le e^{-x}$,

$$E\left[\prod_{i=1}^{k-1} (\max(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}}\right] = \prod_{i=1}^{k-1} \left(1 - \frac{1}{4id}\right) \le \exp\left(-\frac{1}{4d}\sum_{i=1}^{k-1}\frac{1}{i}\right) \le \exp\left(-\frac{\ln k}{4d}\right) .$$

In a similar manner, we have

$$E_{X_i,Y_{i,j},U_{i,j}}\left[(\min(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}}\right] = 1 - \frac{1}{id} + \frac{1}{id} E_{U_{i,j}}[\min(U_{i,j}, 1-U_{i,j})] = 1 - \frac{3}{4id} ,$$

by using the fact

$$E_{U_{i,j}}[\min(U_{i,j}, 1-U_{i,j})] = \int_0^1 \min(U_{i,j}, 1-U_{i,j}) dU_{i,j} = \int_0^{1/2} U_{i,j} dU_{i,j} + \int_{1/2}^1 (1-U_{i,j}) dU_{i,j} = \frac{1}{4} .$$

It follows that

$$E\left[\prod_{i=1}^{k-1} (\min(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}}\right] = \prod_{i=1}^{k-1} \left(1 - \frac{3}{4id}\right) = \exp\left(\sum_{i=1}^{k-1} \ln\left(1 - \frac{3}{4id}\right)\right) ,$$

which completes the proof of Eqn. (10) by combining with Lemma 10. □

Lemma 12. *For integer $k \ge 2$, $j \in [d]$ and real $\epsilon > -1$, we have*

$$\Pr\left[\prod_{i=1}^{k-1} \left(\max(U_{i,j}, 1-U_{i,j})\right)^{X_i Y_{i,j}} \ge (1+\epsilon)\exp\left(-\frac{\ln k}{8d}\right)\right] \le \frac{e}{(1+\epsilon)k^{1/8d}} ,$$

where the probability takes over random variables $X_1, \ldots, X_{k-1}$, $Y_{1,j}, \ldots, Y_{k-1,j}$ and $U_{1,j}, \ldots, U_{k-1,j}$ with $X_i \sim \mathcal{B}(1/i)$, $Y_{i,j} \sim \mathcal{B}(1/d)$ and $U_{i,j} \sim \mathcal{U}(0,1)$ for $i \in [k-1]$.

Proof. Based on the Markov's inequality and Lemma 11, we have, for any $\lambda > 0$,

$$\begin{aligned}
&\Pr\left[\prod_{i=1}^{k-1} (\max(U_{i,j}, 1-U_{i,j}))^{X_i Y_{i,j}} \ge (1+\epsilon)\exp\left(-\frac{\ln k}{8d}\right)\right] \\
&= \Pr\left[\prod_{i=1}^{k-1} (\max(U_{i,j}, 1-U_{i,j}))^{\lambda X_i Y_{i,j}} \ge (1+\epsilon)^{\lambda}\left(\exp\left(-\frac{\ln k}{8d}\right)\right)^{\lambda}\right] \\
&\le (1+\epsilon)^{-\lambda}\exp\left(\frac{\lambda \ln k}{8d}\right) \times E\left[\prod_{i=1}^{k-1} (\max(U_{i,j}, 1-U_{i,j}))^{\lambda X_i Y_{i,j}}\right] .
\end{aligned}$$

Let $Z_{i,j} = (\max(U_{i,j}, 1-U_{i,j}))^{\lambda X_i Y_{i,j}}$, and we have

$$E_{X_i\sim\mathcal{B}(1/i),Y_{i,j}\sim\mathcal{B}(1/d),U_{i,j}\sim\mathcal{U}(0,1)}[Z_{i,j}]$$
$$= 1-\frac{1}{id}+\frac{1}{id}E_{U_{i,j}\sim\mathcal{U}(0,1)}[(\max(U_{i,j},1-U_{i,j}))^{\lambda}] \le 1-\frac{1}{id}+\frac{2-1/2^{\lambda}}{id(\lambda+1)},$$

where the last equation holds from

$$E_{U_{i,j}\sim\mathcal{U}(0,1)}[(\max(U_{i,j},1-U_{i,j}))^{\lambda}] = \int_0^{1/2}(1-U_{i,j})^{\lambda}dU_{i,j}+\int_{1/2}^{1}U_{i,j}^{\lambda}dU_{i,j} = \frac{2-1/2^{\lambda}}{\lambda+1}.$$

It follows that, by using $1+x\le e^x$,

$$E\left[\prod_{i=1}^{k-1}(\max(U_{i,j},1-U_{i,j}))^{\lambda X_i Y_{i,j}}\right] \le \exp\left(-\sum_{i=1}^{k-1}\frac{1}{id}+\sum_{i=1}^{k-1}\frac{2-1/2^{\lambda}}{(\lambda+1)id}\right).$$

Based on Lemma 9, we have

$$E\left[\prod_{i=1}^{k-1}(\max(U_{i,j},1-U_{i,j}))^{\lambda X_i Y_{i,j}}\right] \le \exp\left(-\frac{\ln k}{d}+\frac{(2-1/2^{\lambda})(1+\ln(k-1))}{(\lambda+1)d}\right).$$

In summary, we have

$$\Pr\left[\prod_{i=1}^{k-1}(\max(U_{i,j},1-U_{i,j}))^{X_i Y_{i,j}} \ge (1+\epsilon)\exp\left(\frac{\ln k}{8d}\right)\right]$$
$$\le \exp\left(-\lambda\ln(1+\epsilon)-\frac{\ln k}{d}+\frac{\lambda\ln k}{8d}+\frac{(2-1/2^{\lambda})(1+\ln(k-1))}{(\lambda+1)d}\right).$$

By setting $\lambda=1$, we have

$$\Pr\left[\prod_{i=1}^{k-1}(\max(U_{i,j},1-U_{i,j}))^{X_i Y_{i,j}} \ge (1+\epsilon)\exp\left(\frac{\ln k}{8d}\right)\right]$$
$$\le \exp\left(-\ln(1+\epsilon)-\frac{7\ln k}{8d}+\frac{3(1+\ln(k-1))}{4d}\right) \le \frac{e^{3/4d}}{(1+\epsilon)^{1/8d}},$$

which completes the proof for dimension $d\ge 1$. □

Proof of Lemma 3. Based on the union bounds, we have

$$\Pr\left[\nu[C(\boldsymbol{x})]\ge\frac{(1+\epsilon)\sqrt{d}}{k^{1/8d}}\right] = \Pr\left[\nu[C(\boldsymbol{x})]\ge(1+\epsilon)\sqrt{d}\exp\left(-\frac{\ln k}{8d}\right)\right]$$
$$\le \Pr\left[\exists j\in[d]:\ell_j(C(\boldsymbol{x}))\ge(1+\epsilon)\exp\left(-\frac{\ln k}{8d}\right)\right]$$
$$\le d\Pr\left[\ell_1(C(\boldsymbol{x}))\ge(1+\epsilon)\exp\left(-\frac{\ln k}{8d}\right)\right]$$
$$\le d\Pr\left[\prod_{i=1}^{k-1}\left(\max(U_{i,1},1-U_{i,1})\right)^{X_i Y_{i,1}}\ge(1+\epsilon)\exp\left(-\frac{\ln k}{8d}\right)\right] \le \frac{ed}{(1+\epsilon)k^{1/8d}},$$

where the last inequality holds from Lemma 12. This completes the proof. □

6.4. Proof of Lemma 4

It is necessary to introduce two lemmas as follows:

Lemma 13. *For any rectangular cell $C_i\subseteq\mathcal{X}$, we have*

$$\Pr[\boldsymbol{x}\in C_i]\Pr[|C_i\cap S_n|<n\Pr[\boldsymbol{x}\in C_i]/2]\le 3/n.$$

Proof. From Lemma 8, we have

$$\Pr\left[|C_i \cap S_n| < n \Pr[\boldsymbol{x} \in C_i]/2\right] \leq \exp(-n \Pr[\boldsymbol{x} \in C_i]/8) ,$$

and it holds that

$$\Pr[\boldsymbol{x} \in C_i] \Pr\left[|C_i \cap S_n| < n \Pr[\boldsymbol{x} \in C_i]/2\right] \leq \frac{8}{ne} \leq \frac{3}{n}$$

by using $\max_x xe^{-ax} \leq 1/ae$. This completes the proof. □

Lemma 14. *Let $X_1, X_2, \ldots, X_m$ be m independent random variables with $X_i \sim \mathcal{B}(\eta_i)$ for $i \in [m]$, and set $\rho = \sum_{i=1}^m \eta_i/m$. For $\rho \in [0, 1/2)$, we have*

$$(1-2\rho) \Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^m X_i \geq \frac{m}{2}\right] \leq \frac{1}{\sqrt{2m}} . \tag{11}$$

For $\rho \in [1/2, 1]$, we also have

$$(2\rho-1) \Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^m X_i < \frac{m}{2}\right] \leq \frac{1}{\sqrt{2m}} . \tag{12}$$

Proof. For any $\lambda > 0$, we have, from the Markov's inequality,

$$\begin{aligned} &\Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^m X_i \geq \frac{m}{2}\right] \\ &\leq e^{-m\lambda/2} \mathop{E}_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\exp\left(\lambda \sum_{i=1}^m X_i\right)\right] = e^{-m\lambda/2} \prod_{i=1}^m \mathop{E}_{X_i \sim \mathcal{B}(\eta_i)} [\exp(\lambda X_i)] . \end{aligned}$$

From $X_i \sim \mathcal{B}(\eta_i)$, we have

$$E[\exp(\lambda X_i)] = 1 - \eta_i e^0 + \eta_i e^\lambda \leq \exp(\eta_i(e^\lambda - 1)) .$$

Write $\rho = \sum_{i=1}^m \eta_i/m$, and it holds that

$$\Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^m X_i \geq \frac{m}{2}\right] \leq \exp(-m\lambda/2 + m\rho(e^\lambda - 1)) .$$

By setting $\lambda = -\ln(2\rho)$, we have

$$\Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^m X_i \geq \frac{m}{2}\right] \leq \exp(m/2 + m\ln(2\rho)/2 - m\rho) . \tag{13}$$

We introduce another function

$$g_1(\rho) = (1-2\rho)\exp(m/2 + m\ln(2\rho)/2 - m\rho) , \tag{14}$$

and the derivative is given by

$$g_1'(\rho) = \exp(m/2 + m\ln(2\rho)/2 - m\rho)(2\rho m - 2m - 2 + m/2\rho) .$$

Solving $g_1'(\rho) = 0$ gives the optimal solution

$$\rho^* = \frac{1}{2} - \frac{1}{1+\sqrt{2m+1}} .$$

It is easy to find that, for continuous function $g(\rho)$ with $\rho \in [0, 1/2)$

$$g_1(\rho) \leq \max_{\rho \in [0,1/2)} g_1(\rho) = \max\{g_1(0), g_1(1/2), g_1(\rho^*)\} = g_1(\rho^*) , \tag{15}$$

and we further have

$$g_1(\rho^*) = \frac{1}{1+\sqrt{1+2m}} \exp\left(\frac{m}{1+\sqrt{1+2m}} + \frac{m}{2}\ln\left(1 - \frac{2}{1+\sqrt{1+2m}}\right)\right) \leq \frac{1}{1+\sqrt{1+2m}} \leq \frac{1}{\sqrt{2m}} ,$$

where the first inequality holds from $\ln(1-x) \leq -x$. Hence, Eqn. (11) holds from Eqns. (13)-(15).

For Eqn. (12), we similarly have, by using Markov's inequality,

$$\Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^{m} X_i < \frac{m}{2} \right] \leq \exp(-m\lambda/2 + m\rho(e^{\lambda} - 1))$$

for $\lambda \leq 0$. By setting $\lambda = -\ln(2\rho)$ for $\rho \in [1/2, 1]$, we have

$$\Pr_{X_1 \sim \mathcal{B}(\eta_1), \ldots, X_m \sim \mathcal{B}(\eta_m)} \left[\sum_{i=1}^{m} X_i < \frac{m}{2} \right] \leq \exp(m/2 + m\ln(2\rho)/2 - m\rho) \,. \tag{16}$$

We also introduce another function

$$g_2(\rho) = (2\rho - 1)\exp(m/2 + m\ln(2\rho)/2 - m\rho) \,, \tag{17}$$

and solving $g_2'(\rho) = 0$ gives the optimal solution

$$\rho^* = \frac{1}{2} + \frac{1}{1 + \sqrt{2m+1}} \,.$$

It is easy to find that, for continuous function $g(\rho)$ with $\rho \in [1/2, 1]$,

$$g_2(\rho) \leq \max_{\rho \in [0, 1/2)} g(\rho) = \max\{g_2(1/2), g_2(1), g(\rho^*)\} = g_2(\rho^*) \leq 1/\sqrt{2m} \,.$$

This proves Eqn. (12) by combining with Eqns. (16) and (17). □

Proof of Lemma 4. This lemma holds obviously when $\Pr[\boldsymbol{x} \in C_i] = 0$, and it suffices to consider $\Pr[\boldsymbol{x} \in C_i] > 0$. We introduce the random events

$$\Gamma_1 = \{|C_i \cap S_n| \geq n \Pr[\boldsymbol{x} \in C_i]/2\} \quad \text{and} \quad \Gamma_2 = \{|C_i \cap S_n| < n \Pr[\boldsymbol{x} \in C_i]/2\} \,.$$

Based on the law of total probability, we have

$$\Pr_{S_n, (\boldsymbol{x}, y)} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i \right] = \Pr_{S_n, (\boldsymbol{x}, y)} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1 \right] \Pr[\Gamma_1] + \Pr_{S_n, (\boldsymbol{x}, y)} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_2 \right] \Pr[\Gamma_2] \,.$$

It follows that, from Lemma 13,

$$\Pr_{S_n, (\boldsymbol{x}, y)} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i \right] \Pr[\boldsymbol{x} \in C_i] \leq \Pr_{S_n, (\boldsymbol{x}, y)} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1 \right] \Pr[\boldsymbol{x} \in C_i] \Pr[\Gamma_1] + 3/n \,. \tag{18}$$

To bound the term $\Pr_{S_n, (\boldsymbol{x}, y)} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1 \right]$, we further introduce the set S_n^i of training examples falling into the cell C_i, that is, $S_n^i = \left\{ (\boldsymbol{x}_j, y_j) \colon (\boldsymbol{x}_j, y_j) \in S_n \text{ and } \boldsymbol{x}_j \in C_i \right\}$. Under the condition Γ_1, we have

$$m := |S_n^i| = |S_n \cap C_i| \geq n \Pr[C_i]/2 \,. \tag{19}$$

Without loss of generality, we denote by $S_n^i = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$. For instance $\boldsymbol{x} \in C_i$, its label can be predicted by random forests classifier as

$$f_{\Theta, S_n}(\boldsymbol{x}) = I \left[\sum_{j=1}^{m} y_j \geq m/2 \right] \,.$$

Conditioned on $\boldsymbol{x}, \boldsymbol{x}_1, x_2, \ldots, \boldsymbol{x}_m$, we can observe that $y \sim \mathcal{B}(\eta_1(\boldsymbol{x}))$ and $y_j \sim \mathcal{B}(\eta_1(\boldsymbol{x}_j))$ for $j \in [m]$, and set $\rho = \sum_{j=1}^{m} \eta_1(\boldsymbol{x}_j)/m$. It follows that

$$\Pr_{y_1, \ldots, y_m, y} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \right] = \eta_1(\boldsymbol{x}) \Pr_{y_1, \ldots, y_m} \left[\sum_{j=1}^{m} y_j < \frac{m}{2} \right] + (1 - \eta_1(\boldsymbol{x})) \Pr_{y_1, \ldots, y_m} \left[\sum_{j=1}^{m} y_j \geq \frac{m}{2} \right] . \tag{20}$$

If $\rho \in [0, 1/2)$, then we have, from Eqn. (20)

$$\Pr_{y_1, \ldots, y_m, y} \left[f_{\Theta, S_n}(\boldsymbol{x}) \neq y \right] = \eta_1(\boldsymbol{x}) + (1 - 2\eta_1(\boldsymbol{x})) \Pr_{y_1, \ldots, y_m} \left[\sum_{j=1}^{m} y_j \geq \frac{m}{2} \right] ,$$

and it follows that:

- If $\eta_1(x) = 1/2$, then we have $1 - 2\eta_1(x) = 0$ and

$$\Pr_{y_1,\ldots,y_m,y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y\right] = \eta_1(\boldsymbol{x}) = \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} .$$

- If $\eta_1(x) > 1/2$, then we have $1 - 2\eta_1(x) < 0$, and for $\rho \in [0, 1/2)$, we also have

$$\begin{aligned}\Pr_{y_1,\ldots,y_m,y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y\right] < \eta_1(\boldsymbol{x}) &= \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2\eta_1(\boldsymbol{x}) - 1 \\ &\leq \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2|\eta_1(\boldsymbol{x}) - \rho| .\end{aligned}$$

- If $\eta_1(x) < 1/2$, then we have

$$\begin{aligned}&\Pr_{y_1,\ldots,y_m,y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y\right] \\ &\leq \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2|\eta_1(\boldsymbol{x}) - \rho| + (1 - 2\rho)\Pr_{y_1,\ldots,y_m}\left[\sum_{j=1}^{m} y_j \geq \frac{m}{2}\right] \\ &\leq \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2|\eta_1(\boldsymbol{x}) - \rho| + 1/\sqrt{2m} ,\end{aligned}$$

  where the last inequality holds from $\rho \in [0, 1/2)$ and Eqn. (11) in Lemma 14.

In summary, we have, for $\rho \in [0, 1/2)$,

$$\Pr_{y_1,\ldots,y_m,y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y\right] \leq \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2|\eta_1(\boldsymbol{x}) - \rho| + 1/\sqrt{2m} .$$

In a similar manner, we have, for $\rho \in [1/2, 1]$,

$$\begin{aligned}\Pr_{y_1,\ldots,y_m,y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y\right] &= 1 - \eta_1(\boldsymbol{x}) + (2\eta_1(\boldsymbol{x}) - 1)\Pr_{y_1,\ldots,y_m}\left[\sum_{j=1}^{m} y_j \geq \frac{m}{2}\right] \\ &\leq \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2|\eta_1(\boldsymbol{x}) - \rho| + 1/\sqrt{2m} .\end{aligned}$$

From the L-Lipschtiz assumption, we have, for $\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in C_i$,

$$|\eta_1(\boldsymbol{x}) - \rho| = \left|\eta_1(\boldsymbol{x}) - \sum_{j=1}^{m}\frac{\eta_1(\boldsymbol{x}_j)}{m}\right| \leq \sum_{j=1}^{m}\left|\eta_1(\boldsymbol{x}) - \eta_1(\boldsymbol{x}_j)\right|/m \leq L\nu(C_i) .$$

It follows that

$$\Pr_{y_1,\ldots,y_m,y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y\right] \leq \min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} + 2L\nu(C_i) + 1/\sqrt{2m} .$$

Hence, we have, from Eqn. (19)

$$\begin{aligned}&\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1\right]\Pr[C_i]\Pr[\Gamma_1] \\ &\leq E_{\boldsymbol{x}}[\min\{\eta_1(\boldsymbol{x}), 1 - \eta_1(\boldsymbol{x})\} | \boldsymbol{x} \in C_i]\Pr[C_i] + 2L\nu(C_i)\Pr[C_i] + \sqrt{\Pr[C_i]/n} ,\end{aligned}$$

which completes the proof by combining with Eqn. (18). □

6.5. Proof of Theorem 1

It suffices to derive the convergence rate of individual random tree classifier $f_{S_n,\Theta}(\boldsymbol{x})$, and we complete the proof by combining with Lemma 1. We first have

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = \Pr_{(\boldsymbol{x},y)\sim\mathcal{D}}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y] = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y]\right] .$$

For random tree classifier $f_{\Theta,S_n}(\boldsymbol{x})$, we associate a set as follows:

$$\Lambda = \left\{\boldsymbol{x} \in \mathcal{X} : \nu(C(\boldsymbol{x})) \geq (1 + \epsilon)\sqrt{d}/k^{1/8d}\right\} , \tag{21}$$

where $\nu(C(\boldsymbol{x}))$ denotes the diameter of rectangle cell $C(\boldsymbol{x})$. It follows that

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x})\neq y]\left(\mathbb{I}[\boldsymbol{x}\in\Lambda]+\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right)\right]$$
$$\leq E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}[\mathbb{I}[\boldsymbol{x}\in\Lambda]]+E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x})\neq y]\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right]. \tag{22}$$

Notice that $C_1, C_2, \ldots, C_k$ is a partition of the instance space $\mathcal{X}$ from the construction of random tree. Based on the law of total probability, we have

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x})\neq y]\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right]=\sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\not\subseteq\Lambda],$$

by using the fact $C(\boldsymbol{x})=C_i$ for every $\boldsymbol{x}\in C_i$. Combining with Eqns. (21) and (22), we have

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right]\leq E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})]\geq(1+\epsilon)\sqrt{d}/k^{1/8d}\right]\right] \tag{23}$$
$$+E_{\Theta}\left[\sum_{i=1}^{k}E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\not\subseteq\Lambda]\right]. \tag{24}$$

From Lemma 3, Eqn. (23) can be further upper bounded by

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})]\geq(1+\epsilon)\sqrt{d}/k^{1/8d}\right]\right]\leq\frac{ed}{(1+\epsilon)k^{1/8d}}. \tag{25}$$

Based on Lemma 4 and Eqn. (21), we can bound Eqn. (24) as follows

$$\sum_{i=1}^{k}E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\not\subseteq\Lambda]\leq R^*_{\mathcal{D}}+\frac{2(1+\epsilon)L\sqrt{d}}{k^{1/8d}}+\sum_{i=1}^{k}\sqrt{\frac{\Pr[C_i]}{n}}+\frac{3k}{n}, \tag{26}$$

where we use the law of total expectation and $R^*_{\mathcal{D}}=E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}[\min\{\eta_1(\boldsymbol{x}),1-\eta_1(\boldsymbol{x})\}]$. By using the Jensen's inequality, we have $(EX)^2\leq E[X^2]$, which gives

$$\left(\frac{1}{k}\sum_{i=1}^{k}\sqrt{\Pr[C_i]}\right)^2\leq\frac{1}{k}\sum_{i=1}^{k}\Pr[C_i]=\frac{1}{k}.$$

It follows that, by combining with Eqns. (23)-(26),

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right]\leq R^*_{\mathcal{D}}+\frac{ed}{(1+\epsilon)k^{1/8d}}+\frac{2(1+\epsilon)L\sqrt{d}}{k^{1/8d}}+\sqrt{\frac{k}{n}}+\frac{3k}{n}.$$

We have, by setting $\epsilon=\sqrt{e\sqrt{d}/2L}-1$ and algebra calculations,

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right]\leq R^*_{\mathcal{D}}+\frac{2\sqrt{2eLd^{3/2}}}{k^{1/8d}}+\sqrt{\frac{k}{n}}+\frac{3k}{n},$$

which completes the proof by combining with Lemma 1. □

6.6. Proof of Theorem 2

We first introduce some lemmas before the proof of Theorem 2.

Lemma 15. *For integer $k\geq 2$, $d\geq 2$, $j\in[d]$ and real $\epsilon>-1$, we have*

$$\Pr\left[\prod_{i=1}^{k-1}\left(\frac{1}{2}\right)^{X_iY_{i,j}}\geq(1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right]\leq\frac{3/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}},$$

where the probability takes over random variables $X_1,\ldots,X_{k-1},Y_{1,j},\ldots,Y_{k-1,j}$ with $X_i\sim\mathcal{B}(1/i)$ and $Y_{i,j}\sim\mathcal{B}(1/d)$ for $i\in[k-1]$.

Proof. For any $\lambda > 0$, we have, based on the Markov's inequality,

$$\begin{aligned}
&\Pr\left[\prod_{i=1}^{k-1}\left(\frac{1}{2}\right)^{X_i Y_{i,j}} \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \\
&= \Pr\left[\prod_{i=1}^{k-1}\left(\frac{1}{2}\right)^{\lambda X_i Y_{i,j}} \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \\
&\leq (1+\epsilon)^{-\lambda}\exp\left(\frac{\lambda \ln k}{4d}\right) \times E\left[\prod_{i=1}^{k-1}\left(\frac{1}{2}\right)^{\lambda X_i Y_{i,j}}\right].
\end{aligned}$$

From $X_i \sim \mathcal{B}(1/i)$ and $Y_{i,j} \sim \mathcal{B}(1/d)$ $(i \in [k-1])$, we have, by using $1 + x \leq e^x$,

$$E\left[\prod_{i=1}^{k-1}\left(\frac{1}{2}\right)^{\lambda X_i Y_{i,j}}\right] = \prod_{i=1}^{k-1}\left(1 - \frac{1}{id} + \frac{1}{id2^{\lambda}}\right) \leq \exp\left(-\sum_{i=1}^{k-1}\frac{1}{id} + \sum_{i=1}^{k-1}\frac{1}{id2^{\lambda}}\right),$$

which yields that

$$\Pr\left[\prod_{i=1}^{k-1}\left(\frac{1}{2}\right)^{X_i Y_{i,j}} \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \leq \exp\left(-\lambda\ln(1+\epsilon) + \frac{\lambda\ln k}{4d} - \sum_{i=1}^{k-1}\frac{1}{id} + \sum_{i=1}^{k-1}\frac{1}{id2^{\lambda}}\right).$$

By setting $\lambda = 3/2$, we have

$$\begin{aligned}
&\Pr\left[\prod_{i=1}^{k-1}(\max(U_i, 1-U_i))^{X_i Y_{i,j}} \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \\
&\leq \exp\left(-\frac{3}{2}\ln(1+\epsilon) - \frac{5\ln k}{8d} + \frac{1+\ln(k-1)}{2\sqrt{2}d}\right) \leq \frac{e^{1/(2\sqrt{2}d)}}{(1+\epsilon)^{3/2}k^{1/3.6846d}},
\end{aligned}$$

which completes the proof by using $e^{1/(2\sqrt{2}d)} \leq 3/2$. □

Based on Lemma 15, we can bound the diameter $\nu(C(\boldsymbol{x}))$ as follows:

Lemma 16. *For real $\epsilon > -1$ and instance $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\Pr\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\sqrt{d}/k^{1/4d}\right] \leq \frac{3d/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}},$$

where the probability takes over random selection of splitting leaves and dimensions.

Proof. For $j \in [d]$, recall that $\ell_j(C(\boldsymbol{x}))$ denotes the length of the j-th coordinate of $C(\boldsymbol{x})$ for $j \in [d]$. Let $X_1, \ldots, X_{k-1}$, $Y_{1,j}, \ldots, Y_{k-1,j}$ be random variables with $X_i \sim \mathcal{B}(1/i)$ and $Y_{i,j} \sim \mathcal{B}(1/d)$ for $i \in [k-1]$. According to the construction of purely random tree with midpoint split, we have

$$\ell_j(C(\boldsymbol{x})) = 1/2^{X_i Y_{i,j}}.$$

Based on Lemma 15, we have

$$\Pr\left[\ell_j(C(\boldsymbol{x})) \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \leq \frac{3/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}}. \tag{27}$$

Based on union bounds, we have

$$\begin{aligned}
&\Pr\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\sqrt{d}/k^{1/4d}\right] = \Pr\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\sqrt{d}\exp\left(-\frac{\ln k}{4d}\right)\right] \\
&\leq \Pr\left[\exists j \in [d] : \ell_j(C(\boldsymbol{x})) \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \\
&\leq d\Pr\left[\ell_1(C(\boldsymbol{x})) \geq (1+\epsilon)\exp\left(-\frac{\ln k}{4d}\right)\right] \leq \frac{3d/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}},
\end{aligned}$$

where the last inequality holds from Eqn. (27). □

Proof of Theorem 2. Similarly to Theorem 1, we first study the convergence rate of individual random tree classifier $f_{S_n,\Theta}(\boldsymbol{x})$. Based on the law of total probability, we have

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = \Pr_{(\boldsymbol{x},y)\sim\mathcal{D}}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y] = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y]\right].$$

For random forests classifier $f_{\Theta,S_n}(\boldsymbol{x})$, we associate a set as follows

$$\Lambda_2 = \left\{\boldsymbol{x}\in\mathcal{X}\colon \nu(C(\boldsymbol{x})) \geq (1+\epsilon)\sqrt{d}/k^{1/4d}\right\}, \tag{28}$$

and it follows that

$$R_{\mathcal{D}}(f_{S_n,\Theta}) \leq E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}[\mathbb{I}[\boldsymbol{x}\in\Lambda_2]] + E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y|\boldsymbol{x}]\mathbb{I}[\boldsymbol{x}\notin\Lambda_2]\right]. \tag{29}$$

Notice that $C_1, C_2, \ldots, C_k$ is a partition of the instance space $\mathcal{X}$, and we have

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{B}(\eta_1(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y]\mathbb{I}[\boldsymbol{x}\notin\Lambda_2]\right] = \sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y|\boldsymbol{x}\in C_i]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\nsubseteq\Lambda_2],$$

where we use the fact $C(\boldsymbol{x}) = C_i$ for every $\boldsymbol{x}\in C_i$. It follows that, from Eqns. (28) and (29),

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \leq E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})] \geq \frac{(1+\epsilon)\sqrt{d}}{(1+k)^{1/4d}}\right]\right] \tag{30}$$

$$+ E_{\Theta}\left[\sum_{i=1}^{k} E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y|\boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\nsubseteq\Lambda_2]\right]. \tag{31}$$

From Lemma 16, Eqn. (30) can be further upper bounded by

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\sqrt{d}/k^{1/4d}\right]\right] \leq \frac{3d/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}}. \tag{32}$$

Based on Lemma 4 and Eqn. (28), we can bound the term in Eqn. (31) as

$$\sum_{i=1}^{k} E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y|\boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\nsubseteq\Lambda_2] \leq R^*_{\mathcal{D}} + \frac{2(1+\epsilon)L\sqrt{d}}{k^{1/4d}} + \sqrt{\frac{k}{n}} + \frac{3k}{n}. \tag{33}$$

It follows that, by combining with Eqns. (30)-(33),

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \leq R^*_{\mathcal{D}} + \frac{3d/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}} + \frac{2(1+\epsilon)L\sqrt{d}}{k^{1/4d}} + \sqrt{\frac{k}{n}} + \frac{3k}{n}.$$

By setting

$$\epsilon = \left(\frac{9\sqrt{d}}{8L}k^{\frac{1}{4d}-\frac{1}{3.6848d}}\right)^{2/5} - 1,$$

we have, by simple algebraic calculations,

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \leq R^*_{\mathcal{D}} + \frac{4L^{3/5}d^{7/10}}{k^{1/3.87d}} + \sqrt{\frac{k}{n}} + \frac{3k}{n},$$

which completes the proof by combining with Lemma 1. □

6.7. Proof of Theorem 3

We begin with a lemma as follows:

Lemma 17. *Let S_n be a training data drawn i.i.d. from distribution $\mathcal{D}$. For any rectangle cell $C\subseteq\mathcal{X}$ and integer $\kappa\geq 2$, we have*

$$\Pr[\boldsymbol{x}\in C]\Pr[|C\cap S_n| \leq \kappa] \leq \frac{\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right).$$

Proof. For any $\delta \in (0,1)$, if $\kappa \leq (1-\delta)n\Pr[\boldsymbol{x} \in C]$, then we have, based on Lemma 8,

$$\Pr[|C \cap S_n| \leq \kappa] \;\leq\; \Pr[|C \cap S_n| \leq (1-\delta)n\Pr[\boldsymbol{x} \in C]] \;\leq\; \exp(-n\Pr[\boldsymbol{x} \in C]\delta^2/2)\,.$$

It follows that, by using $\max_x xe^{-ax} = 1/ae$,

$$\Pr[\boldsymbol{x} \in C]\Pr[|C \cap S_n| \leq \kappa] \leq \Pr[\boldsymbol{x} \in C]\exp(-n\Pr[\boldsymbol{x} \in C]\delta^2/2) \leq \frac{2}{ne\delta^2}\,.$$

If $\kappa \geq (1-\delta)n\Pr[\boldsymbol{x} \in C]$, then we have

$$\Pr[\boldsymbol{x} \in C]\Pr[|C \cap S_n| \leq \kappa] \leq \Pr[\boldsymbol{x} \in C] \leq \frac{\kappa}{n(1-\delta)}\,.$$

By setting $\delta = (\sqrt{1+2\kappa e}-1)/\kappa e$, we have

$$\frac{\kappa}{n(1-\delta)} = \frac{2}{ne\delta^2} = \frac{\kappa}{n} \times \frac{\kappa e + 1 + \sqrt{2\kappa e+1}}{\kappa e} \leq \frac{\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right) \text{ for } \kappa \geq 2,$$

which completes the proof. □

Proof of Theorem 3. Similarly to the proof of Theorem 1, we first present the convergence rate of individual random tree classifier $f_{S_n,\Theta}(\boldsymbol{x})$ according to Algorithm 1. Let $C_1, C_2, \ldots, C_k$ be a partition of instance space $\mathcal{X}$, which are associated with k leaves of random tree. Based on the law of total probability, we have the classification error of random forests classifier $f_{\Theta,S_n}(\boldsymbol{x})$ with respect to distribution $\mathcal{D}$

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = \Pr_{(\boldsymbol{x},y)\sim\mathcal{D}}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y] = \sum_{i=1}^{k} \Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i]\Pr[\boldsymbol{x} \in C_i]\,.$$

We introduce a set

$$\Lambda_3 = \{C_i : \text{ all training examples in } C_i \text{ have the same label}\}.$$

It follows that

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = \sum_{i=1}^{k} \Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i]\Pr[\boldsymbol{x} \in C_i]\mathbb{I}[C_i \in \Lambda_3] \tag{34}$$

$$+ \sum_{i=1}^{k} \Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i]\Pr[\boldsymbol{x} \in C_i]\mathbb{I}[C_i \notin \Lambda_3]\,. \tag{35}$$

If $C_i \in \Lambda_3$, then we have, for $\kappa \geq 2$,

$$\begin{aligned}
&\Pr[C_i]\Pr\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \middle| \boldsymbol{x} \in C_i\right] \\
&= \Pr[C_i]\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| \leq \kappa | \boldsymbol{x} \in C_i] + \Pr[C_i]\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| > \kappa | \boldsymbol{x} \in C_i] \\
&\leq \Pr[C_i]\Pr[|C_i \cap S_n| \leq \kappa] + \Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \big| |C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i]\Pr[|C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i]\,.
\end{aligned}$$

From $C_i \in \Lambda_3$, all training examples in C_i have the same label, and we assume positive training examples in C_i without loss of generality. Then, we have $f_{\Theta,S_n}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in C_i$. Denote by the expected conditional probability over cell C_i

$$\bar{\eta}_1(C_i) = E[\eta_1(\boldsymbol{x}) | \boldsymbol{x} \in C_i]\,.$$

If $\bar{\eta}_1(C_i) \geq 1-\epsilon$, then we have

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \big| |C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i]\Pr[|C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i] \leq \epsilon\,;$$

If $\bar{\eta}_1(C_i) < 1-\epsilon$ and $C_i \in \Lambda_3$, then we have

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \big| |C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i]\Pr[|C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i] \leq \Pr[|C_i \cap S_n| > \kappa] \leq \exp(-\kappa\epsilon)\,.$$

This follows that, for $C_i \in \Lambda_3$,

$$\Pr[C_i]\Pr\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \middle| \boldsymbol{x} \in C_i\right] \leq \Pr[C_i]\Pr[|C_i \cap S_n| \leq \kappa] + \Pr[C_i](\epsilon + \exp(-\kappa\epsilon))\,.$$

By setting $\epsilon = (\ln\kappa)/\kappa$, we have

$$\Pr[C_i]\Pr\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \middle| \boldsymbol{x} \in C_i\right] \leq \Pr[C_i]\Pr[|C_i \cap S_n| \leq \kappa] + \Pr[C_i]\frac{1+\ln\kappa}{\kappa}\,.$$

It follows that, by combining with Lemma 17 and Eqns. (34)-(35)

$$E_{S_n,\Theta}[R_{\mathcal{D}}(f_{S_n,\Theta})] \leq \frac{k\kappa\left(1+\sqrt{2/\kappa}\right)}{n} + \frac{1+\ln\kappa}{\kappa} + \sum_{i=1}^{k} E_{S_n,\Theta}\left[\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y|\boldsymbol{x} \in C_i]\Pr[\boldsymbol{x} \in C_i]\mathbb{I}[C_i \notin \Lambda_3]\right]\,. \quad (36)$$

For $C_i \notin \Lambda_3$, we have different labels in C_i. It follows the height $h(C_i) \geq \log_2 k - 2$ and the splitting times for each dimension are more than $(\log_2 k - 2)/d - 1$ from the construction of random tree in Algorithm 1. Hence, we upper bound the diameter of rectangle cell C_i as follows:

$$\nu(C_i) \leq \sqrt{d}\left(\frac{1}{2}\right)^{(\log_2 k-2)/d-1} = \frac{2^{1+2/d}\sqrt{d}}{k^{1/d}} \leq \frac{8\sqrt{d}}{k^{1/d}}\,.$$

It follows that, from Lemma 4 and Eqn. (36),

$$\begin{aligned} E_{S_n,\Theta}[R_{\mathcal{D}}(f_{S_n,\Theta})] &\leq R^*_{\mathcal{D}} + \frac{k\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right) + \frac{1}{\kappa}(1+\ln\kappa) + \frac{16L\sqrt{d}}{k^{1/d}} + \frac{3k}{n} + \sum_{i=1}^{k}\sqrt{\frac{\Pr[C_i]}{n}} \\ &\leq R^*_{\mathcal{D}} + \frac{k\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right) + \frac{1}{\kappa}(1+\ln\kappa) + \frac{16L\sqrt{d}}{k^{1/d}} + \frac{3k}{n} + \sqrt{\frac{k}{n}}\,. \end{aligned}$$

We have, by setting $\kappa = \left\lceil\sqrt{n\ln n/k}\right\rceil$ and algebra calculations,

$$\begin{aligned} &E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \\ &\leq R^*_{\mathcal{D}} + \sqrt{\frac{k\ln n}{n}} + \sqrt[4]{\frac{4k^3\ln n}{n^3}} + \sqrt{\frac{k}{n\ln n}}\left(1+\frac{1}{2}\ln\frac{n\ln n}{k}\right) + \frac{6k}{n} + \sqrt{\frac{k}{n}} + \frac{16L\sqrt{d}}{k^{1/d}} \\ &\leq R^*_{\mathcal{D}} + 2\sqrt{\frac{k\ln n}{n}} + \sqrt[4]{\frac{4k^3\ln n}{n^3}} + \frac{6k}{n} + 2\sqrt{\frac{k}{n}} + \frac{16L\sqrt{d}}{k^{1/d}} \quad (n \geq 4, k \geq 2), \end{aligned}$$

which completes the proof by combining with Lemma 1. □

6.8. Proof of Theorem 4

The proof is essentially similar to that of Theorem 3. Given a random tree classifier $f_{\Theta,S_n}(\boldsymbol{x})$ with k leaves ($k \leq k_0$), let $C_1, C_2, \ldots, C_k$ be a partition of the instance space $\mathcal{X}$. Based on the law of total probability, we have the classification error of random forests classifier $f_{\Theta,S_n}(\boldsymbol{x})$ over distribution $\mathcal{D}$

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = \Pr_{(\boldsymbol{x},y)\sim\mathcal{D}}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y] = \sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y|\boldsymbol{x} \in C_i]\Pr[\boldsymbol{x} \in C_i]\,. \quad (37)$$

For any $i \in [k]$ and $\kappa \geq 1$, we have

$$\begin{aligned} &\Pr[C_i]\Pr\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \middle| \boldsymbol{x} \in C_i\right] \\ &\qquad = \Pr[C_i]\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| \leq \kappa|\boldsymbol{x} \in C_i] + \Pr[C_i]\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| > \kappa|\boldsymbol{x} \in C_i]\,. \end{aligned} \quad (38)$$

From Lemma 17, we have

$$E_{S_n}\left[\Pr[C_i]\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| \leq \kappa|\boldsymbol{x} \in C_i]\right] \leq \frac{\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right)\,. \quad (39)$$

From the assumption in Theorem 4, we see that all training examples in each C_i have the same label, and we assume positive training examples in C_i without loss of generality. It follows that $f_{\Theta,S_n}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in C_i$. Denote by

$$\bar{\eta}_1(C_i) = E[\eta_1(\boldsymbol{x})|\boldsymbol{x} \in C_i]$$

the expected conditional probability over the rectangle cell C_i. If $\bar{\eta}_1(C_i) \geq 1-\epsilon$, then we have

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| > \kappa \,|\boldsymbol{x} \in C_i] \leq \epsilon \ ;$$

If $\bar{\eta}_1(C_i) < 1-\epsilon$ and $C_i \in \Lambda_3$, then we have

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| > \kappa \,|\boldsymbol{x} \in C_i] \leq (1-\epsilon)^{\kappa} \leq \exp(-\kappa\epsilon) \ .$$

It follows that, by setting $\epsilon = (\ln\kappa)/\kappa$,

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y, |C_i \cap S_n| > \kappa \,|\boldsymbol{x} \in C_i] \leq \epsilon + \exp(-\kappa\epsilon) \leq (1+\ln\kappa)/\kappa \ . \tag{40}$$

Combining with Eqns. (37)-(40), we have

$$E_{S_n,\Theta}[R_{\mathcal{D}}(f_{S_n,\Theta})] \leq \frac{k\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right) + \frac{1}{\kappa}(1+\ln\kappa) \leq \frac{k_0\kappa}{n}\left(1+\sqrt{\frac{2}{\kappa}}\right) + \frac{1}{\kappa}(1+\ln\kappa) \ .$$

By setting $\kappa = \left\lceil \sqrt{n\ln n/k_0} \right\rceil$ and simple algebraic calculations, we have

$$E_{S_n,\Theta}[R_{\mathcal{D}}(f_{S_n,\Theta})] \leq 2\sqrt{\frac{k_0\ln n}{n}} + \sqrt[4]{\frac{4k_0^3\ln n}{n^3}} + \frac{3k_0}{n} + \sqrt{\frac{k_0}{n\ln n}} \ ,$$

which completes the proof by combining with Lemma 1. □

6.9. Proof of Lemma 5

We begin with two lemmas before the proof of Lemma 5.

Lemma 18. *For $\tau = 3$, let $\vartheta_1, \vartheta_2, \vartheta_3$ and ρ_1, ρ_2, ρ_3 be defined by Eqn. (3). If $\rho_1 \geq \max(\rho_2, \rho_3)$, then we have*

$$\begin{aligned}
&\rho_1 \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \max\{\vartheta_2,\vartheta_3\}] - \sum_{i=2}^{3}\rho_i \Pr_{y_1,\ldots,y_{n'}}[\vartheta_i = \max\{\vartheta_1,\vartheta_2,\vartheta_3\}] \\
&\leq \sum_{i=2}^{3}(\rho_1-\rho_i)\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_i] - \rho_1 \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 = \vartheta_3] \\
&\quad -\rho_2 \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 = \vartheta_2 = \max(\vartheta_2,\vartheta_3)] - \rho_3 \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 = \vartheta_3 = \max(\vartheta_2,\vartheta_3)] \ .
\end{aligned}$$

Proof. We first have, by using $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$,

$$\begin{aligned}
\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \max(\vartheta_2,\vartheta_3)] &= \Pr_{y_1,\ldots,y_{n'}}[\{\vartheta_1 < \vartheta_2\} \cup \{\vartheta_1 < \vartheta_3\}] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_3] - \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2, \vartheta_1 < \vartheta_3] \ ,
\end{aligned}$$

and we further have, based on the law of total probability,

$$\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2, \vartheta_1 < \vartheta_3] = \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 < \vartheta_3] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_3 < \vartheta_2] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 = \vartheta_3] \ .$$

This follows that

$$\begin{aligned}
\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \max(\vartheta_2,\vartheta_3)] \ \leq \ &\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_3] \\
&- \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 = \vartheta_3] - \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 < \vartheta_3] - \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_3 < \vartheta_2] \ .
\end{aligned} \tag{41}$$

By using the law of total probability again, we have

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}[\vartheta_2 = \max(\vartheta_1,\vartheta_2,\vartheta_3)] = \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 \leq \vartheta_2 = \max(\vartheta_2,\vartheta_3)] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 = \vartheta_2 = \max(\vartheta_2,\vartheta_3)] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 = \max(\vartheta_2,\vartheta_3)] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 = \vartheta_2 = \max(\vartheta_2,\vartheta_3)] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2] - \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_2 < \vartheta_3] \ ,
\end{aligned} \tag{42}$$

and we similarly have

$$\Pr_{y_1,\ldots,y_{n'}}[\vartheta_3 = \max(\vartheta_1,\vartheta_2,\vartheta_3)] = \Pr_{y_1,y_2,y_3}[\vartheta_1 = \vartheta_3 = \max(\vartheta_2,\vartheta_3)] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_3] - \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_3 < \vartheta_2],$$

which completes the proof by combining with Eqns. (41)-(42) and $\rho_1 \geq \max(\rho_2,\rho_3)$. □

Lemma 19. *For integer $l \geq 4$, let $\vartheta_1, \vartheta_2, \cdots, \vartheta_l$ and $\rho_1, \rho_2, \cdots, \rho_l$ be defined by Eqn. (3). If $\rho_1 \geq \max(\rho_2, \ldots, \rho_l)$, then we have*

$$\begin{aligned}
&\rho_1 \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l]\setminus\{1\}}\{\vartheta_i\}\right] - \sum_{i=2}^{l}\rho_i \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_i = \max_{j\in[l]}\{\vartheta_j\}\right] \\
&\leq (\rho_1-\rho_l)\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_l] + \rho_1 \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_i\}\right] - \sum_{i=2}^{l-1}\rho_i \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_i = \max_{j\in[l-1]}\{\vartheta_i\}\right] \\
&\quad + I_1 + I_2 - \rho_1 \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_i\}\right] - \rho_l \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_l = \max_{j\in[l]\setminus\{1\}}\{\vartheta_i\}\right],
\end{aligned}$$

where I_1 and I_2 are given, respectively, by

$$I_1 = \sum_{i=2}^{l-1}\rho_i \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l)\right] \text{ and } I_2 = \sum_{i=3}^{l-1}\rho_i \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\} < \vartheta_l)\right].$$

Proof. By using $\Pr[A\cup B] = \Pr[A] + \Pr[B] - \Pr[A\cap B]$, we first have

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l]\setminus\{1\}}\{\vartheta_j\}\right] = \Pr_{y_1,\ldots,y_{n'}}[\{\vartheta_1 < \max\{\vartheta_2,\cdots,\vartheta_{l-1}\}\}\cup\{\vartheta_1 < \vartheta_l\}] \\
&= \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_l] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l, \vartheta_1 < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right],
\end{aligned}$$

and we also have, based on the law of total probability,

$$\begin{aligned}
\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l, \vartheta_1 < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] &= \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] \\
&\quad + \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] + \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right].
\end{aligned}$$

We further bound

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right] \\
&\qquad \geq \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_2 = \max_{j\in[l-1]\setminus\{1,2\}}\{\vartheta_j\} < \vartheta_l\right] + \sum_{i=2}^{l-1}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i < \vartheta_l, \vartheta_i > \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\}\right],
\end{aligned}$$

and this follows that

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[l]\setminus\{1\}}\{\vartheta_j\}\right] \\
&\leq \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \max(\vartheta_2,\cdots,\vartheta_{l-1})] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] \\
&\quad - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_2 = \max_{j\in[l-1]\setminus\{1,2\}}\{\vartheta_j\} < \vartheta_l\right] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] \\
&\quad + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_l] - \sum_{i=2}^{l-1}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i < \vartheta_l, \vartheta_i > \max_{j\in[l-1]\setminus\{1,i\}}(\vartheta_j)\right]. \qquad (43)
\end{aligned}$$

For $2 \leq i \leq l-1$, we have, by the law of total probability again,

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}[\vartheta_i = \max(\vartheta_1,\vartheta_2,\cdots,\vartheta_l)] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_i = \max(\vartheta_1,\vartheta_2,\cdots,\vartheta_{l-1}), \vartheta_i \geq \vartheta_l] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_i = \max(\vartheta_1,\vartheta_2,\cdots,\vartheta_{l-1})] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_i = \max_{j\in[l-1]}\{\vartheta_j\} < \vartheta_l\right].
\end{aligned}$$

We further have

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_i = \max_{j\in[l-1]}\{\vartheta_j\} < \vartheta_l\right] \\
&= \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 \le \vartheta_i, \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right] \\
&= \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right] + \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right]
\end{aligned}$$

and

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right] \\
&= \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\} < \vartheta_l\right] + \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i < \vartheta_l, \vartheta_i > \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\}\right].
\end{aligned}$$

This follows that

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}[\vartheta_i = \max(\vartheta_1,\cdots,\vartheta_l)] \\
&= -\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\} < \vartheta_l\right] \\
&\quad + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_i = \max(\vartheta_1,\cdots,\vartheta_{l-1})] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i < \vartheta_l, \vartheta_i > \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\}\right].
\end{aligned} \tag{44}$$

We finally have, by the law of total probability again,

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}[\vartheta_l = \max\{\vartheta_1,\vartheta_2,\cdots,\vartheta_l\}] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_l = \max\{\vartheta_2,\cdots,\vartheta_l\}] + \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 = \vartheta_l = \max\{\vartheta_2,\cdots,\vartheta_l\}] \\
&= \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_l] - \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_l < \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}\right] + \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_l = \max_{j\in[l]\setminus\{1\}}\{\vartheta_j\}\right]
\end{aligned}$$

which completes the proof by combining with Eqns. (43)-(44), $\rho_1 \ge \max\{\rho_2,\ldots,\rho_l\}$ and simple algebraic calculations. □

Proof of Lemma 5. This lemma holds obviously when $\tau = 3$ from Lemma 18, and it suffices to prove the case $\tau \ge 4$. From Lemma 18 again and by setting $l = 4,\ldots,\tau$ in Lemma 19, we have

$$\begin{aligned}
&\rho_1 \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[\tau]\setminus\{1\}}\{\vartheta_j\}\right] - \sum_{i=2}^{\tau}\rho_i \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_i = \max_{j\in[\tau]}\{\vartheta_j\}\right] \\
&\le \sum_{i=2}^{\tau}(\rho_1-\rho_i)\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_i] + \Delta_1 + \Delta_2 - \rho_1\sum_{i=3}^{\tau}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[i-1]\setminus\{1\}}\{\vartheta_j\}\right] \\
&\quad -\rho_2\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 = \vartheta_2 = \max\{\vartheta_2,\vartheta_3\}] - \sum_{i=3}^{\tau}\rho_i\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_i = \max_{j\in[i]\setminus\{1\}}\{\vartheta_j\}\right],
\end{aligned} \tag{45}$$

where Δ_1 and Δ_2 are given, respectively, by,

$$\begin{aligned}
\Delta_1 &= \sum_{l=4}^{\tau}\sum_{i=2}^{l-1}\rho_i\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 = \vartheta_i = \max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\} < \vartheta_l\right], \\
\Delta_2 &= \sum_{l=4}^{\tau}\sum_{i=3}^{l-1}\rho_i\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \vartheta_i = \max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\} < \vartheta_l\right].
\end{aligned}$$

It remains to bound Δ_1 and Δ_2. Based on the law of total probability, we have

$$
\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1=\vartheta_2=\max\{\vartheta_2,\vartheta_3\}]\\
&=\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1=\vartheta_2=\max\{\vartheta_2,\vartheta_3\}<\vartheta_4]+\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1=\vartheta_2=\max\{\vartheta_2,\vartheta_3,\vartheta_4\}]\\
&=\cdots\\
&=\sum_{l=4}^{\tau}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_2=\max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}<\vartheta_l\right]+\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_2=\max_{j\in[\tau]}\{\vartheta_j\}\right]\\
&\geq\sum_{l=4}^{\tau}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_2=\max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}<\vartheta_l\right],
\end{aligned}
\tag{46}
$$

from $\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_2=\max_{j\in[\tau]}\{\vartheta_j\}\right]\geq 0$, and we similarly have, for $3\leq i\leq\tau-1$,

$$
\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_i=\max_{j\in[i]\setminus\{1\}}\{\vartheta_j\}\right]\geq\sum_{l=i+1}^{\tau}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_i=\max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}<\vartheta_l\right].
$$

This follows that, by combining with Eqn. (46),

$$
\begin{aligned}
\Delta_1&=\sum_{i=2}^{\tau-1}\rho_i\sum_{l=\max\{4,i+1\}}^{\tau}\left[\vartheta_1=\vartheta_i=\max_{j\in[l-1]\setminus\{1\}}\{\vartheta_j\}<\vartheta_l\right]\\
&\leq\rho_2\Pr_{y_1,\ldots,y_{n'}}[\vartheta_1=\vartheta_2=\max\{\vartheta_2,\vartheta_3\}]+\sum_{i=3}^{\tau-1}\rho_i\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1=\vartheta_i=\max_{j\in[i]\setminus\{1\}}\{\vartheta_j\}\right].
\end{aligned}
\tag{47}
$$

Based on the law of total probability again, we similarly have, for $3\leq i\leq\tau-1$,

$$
\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1<\vartheta_i=\max_{j\in[i-1]\setminus\{1\}}\{\vartheta_j\}\right]\geq\sum_{l=i+1}^{\tau}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1<\vartheta_i=\max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\}<\vartheta_l\right],
$$

which yields that

$$
\Delta_2=\sum_{i=3}^{\tau-1}\rho_i\sum_{l=i+1}^{\tau}\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1<\vartheta_i=\max_{j\in[l-1]\setminus\{1,i\}}\{\vartheta_j\}<\vartheta_l\right]\leq\sum_{i=3}^{\tau-1}\rho_i\Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1<\vartheta_i=\max_{j\in[i-1]\setminus\{1\}}\{\vartheta_j\}\right].
\tag{48}
$$

We complete the proof by combining with Eqns. (45)-(48) and $\rho_1\geq\max(\rho_2,\ldots,\rho_\tau)$. □

6.10. Proof of Lemma 6

For any $\lambda<0$, we have, from the Markov's inequality

$$
\Pr_{X_1,\ldots,X_{n'}}\left[\sum_{i=1}^{n'}X_i<0\right]=\Pr_{X_1,\ldots,X_{n'}}\left[\sum_{i=1}^{n'}\lambda X_i>0\right]\leq E_{X_1,\ldots,X_{n'}}\left[\exp\left(\sum_{i=1}^{n'}\lambda X_i\right)\right]=\prod_{i=1}^{n'}E_{X_i}\left[\exp(\lambda X_i)\right].
$$

From $\eta_i^+=\Pr[X_i=+1]$, $\eta_i^-=\Pr[X_1=-1]$ and $\Pr[X_i=0]=1-\eta_i^+-\eta_i^-$, we have

$$
E_{X_i}\left[\exp(\lambda X_i)\right]=(1-\eta_i^+-\eta_i^-)e^0+\eta_i^+e^{\lambda}+\eta_i^-e^{-\lambda}\leq\exp(\eta_i^+(e^{\lambda}-1)+\eta_i^-(e^{-\lambda}-1)),
$$

which yields that, from $\rho^+=\sum_{i=1}^{n'}\eta_i^+/n'$ and $\rho^-=\sum_{i=1}^{n'}\eta_i^-/n'$,

$$
\prod_{i=1}^{n'}E_{X_i}\left[\exp(\lambda X_i)\right]=\exp\left(\sum_{i=1}^{n'}\eta_i^+(e^{\lambda}-1)+\sum_{i=1}^{n'}\eta_i^-(e^{-\lambda}-1)\right)=\exp\left(n'\rho^+(e^{\lambda}-1)+n'\rho^-(e^{-\lambda}-1)\right).
$$

By setting $\lambda=\ln(\rho^-/\rho^+)/2<0$, we have

$$
\begin{aligned}
(\rho^+-\rho^-)\Pr_{X_1,\ldots,X_{n'}}\left[\sum_{i=1}^{n'}X_i<0\right]&\leq(\rho^+-\rho^-)\exp\left(-n'\left(\sqrt{\rho^+}-\sqrt{\rho^-}\right)^2\right)\\
&\leq(\rho^+-\rho^-)\exp\left(-n'\left(\rho^+-\rho^-\right)^2/2\right),
\end{aligned}
$$

where we use $(\sqrt{\rho^+}+\sqrt{\rho^-})^2\leq 2$ from $\rho^+,\rho^-\in[0,1]$ and $\rho^++\rho^-\leq 1$. We finally complete the proof by using $\max_{t\geq 0}\{te^{-n't^2/2}\}=1/\sqrt{en'}$. □

6.11. Proof of Lemma 7

This lemma holds obviously when $\Pr[\boldsymbol{x} \in C_i] = 0$, and it suffices to consider $\Pr[\boldsymbol{x} \in C_i] > 0$. We introduce the random events

$$\Gamma_1 = \{|C_i \cap S_n| \geq n \Pr[\boldsymbol{x} \in C_i]/2\} \quad \text{and} \quad \Gamma_2 = \{|C_i \cap S_n| < n \Pr[\boldsymbol{x} \in C_i]/2\} .$$

Based on the law of total probability, we have

$$\begin{aligned}
&\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i\right] \\
&= \Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1\right] \Pr[\Gamma_1] + \Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_2\right] \Pr[\Gamma_2] .
\end{aligned}$$

It follows that, from Lemma 13,

$$\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i\right] \Pr[\boldsymbol{x} \in C_i] \leq \Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1\right] \Pr[\boldsymbol{x} \in C_i] \Pr[\Gamma_1] + 3/n . \tag{49}$$

To bound $\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1\right]$, we introduce the set S_n^i of training examples falling into the cell C_i, i.e., $S_n^i = \{(\boldsymbol{x}_j, y_j) \colon (\boldsymbol{x}_j, y_j) \in S_n \text{ and } \boldsymbol{x}_j \in C_i\}$. Under the condition Γ_1, we have

$$n' := |S_n^i| = |S_n \cap C_i| \geq n \Pr[C_i]/2 . \tag{50}$$

Without loss of generality, we denote by $S_n^i = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{n'}, y_{n'})\}$. For any instance $\boldsymbol{x} \in C_i$, the predicted label by random forests can be given by

$$f_{\Theta,S_n}(\boldsymbol{x}) = \arg\max_{l \in [\tau]} \left\{ \sum_{j=1}^{n'} I[y_j = l] \right\} .$$

Conditioned on $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n'}$, it is easy to observe that $\eta_l(\boldsymbol{x}) = \Pr[y = l | \boldsymbol{x}]$ for $l \in [\tau]$, and $\eta_l(\boldsymbol{x}_j) = \Pr[y_j = l | \boldsymbol{x}_j]$ for $j \in [n']$ and $l \in [\tau]$. Write $\rho_l = \sum_{j=1}^{n'} \eta_l(\boldsymbol{x}_j)/n'$ for $l \in [\tau]$, and we have $\rho_1 + \rho_2 + \cdots + \rho_\tau = 1$ and

$$|\rho_l - \eta_l(\boldsymbol{x})| = \frac{1}{n'} \sum_{j=1}^{n'} |\eta_l(\boldsymbol{x}_j) - \eta_l(\boldsymbol{x})| \leq L \|\boldsymbol{x}_j - \boldsymbol{x}\| \leq L\nu(C_i) \quad \text{for} \quad l \in [\tau] . \tag{51}$$

For simplicity, we denote by

$$\vartheta_l = \sum_{j=1}^{n'} I[y_j = l] \quad \text{for} \quad l \in [\tau] ,$$

and thus $f_{\Theta,S_n}(\boldsymbol{x}) = \arg\max_{l \in [\tau]} \{\vartheta_l\}$. In the following, we assume $\rho_1 = \max\{\rho_1, \rho_2, \cdots, \rho_\tau\}$ without loss of generality, and make similar considerations for $\rho_l = \max\{\rho_1, \rho_2, \cdots, \rho_\tau\}$ as $l \geq 2$. This follows that

$$\begin{aligned}
&\Pr_{y_1,\ldots,y_{n'},y}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n'}\right] \\
&= \sum_{l=1}^{\tau} \eta_l(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}[\vartheta_l < \max(\vartheta_1, \vartheta_2, \cdots, \vartheta_\tau)] \\
&= 1 - \eta_1(\boldsymbol{x}) + \eta_1(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j \in [\tau]}\{\vartheta_j\}\right] - \sum_{l=2}^{\tau} \eta_l(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_l = \max_{j \in [\tau]}\{\vartheta_j\}\right]
\end{aligned} \tag{52}$$

where the last equality holds from $\eta_1(\boldsymbol{x}) + \eta_2(\boldsymbol{x}) + \cdots + \eta_\tau(\boldsymbol{x}) = 1$. Let $\eta_{l^*}(\boldsymbol{x}) = \max_{l \in [\tau]}\{\eta_l(\boldsymbol{x})\}$. Then, we have

$$\begin{aligned}
1 - \eta_1(\boldsymbol{x}) &= \min_{l \in [\tau]}\{1 - \eta_l(\boldsymbol{x})\} + \eta_{l^*}(\boldsymbol{x}) - \eta_1(\boldsymbol{x}) \\
&= \min_{l \in [\tau]}\{1 - \eta_l(\boldsymbol{x})\} + (\eta_{l^*}(\boldsymbol{x}) - \rho_{l^*}) + (\rho_{l^*} - \rho_1) + (\rho_1 - \eta_1(\boldsymbol{x})) \\
&\leq \min_{l \in [\tau]}\{1 - \eta_l(\boldsymbol{x})\} + |\eta_{l^*}(\boldsymbol{x}) - \rho_{l^*}| + |\eta_1(\boldsymbol{x}) - \rho_1| \\
&\leq \min_{l \in [\tau]}\{1 - \eta_l(\boldsymbol{x})\} + 2L\nu(C_i)
\end{aligned} \tag{53}$$

where the first and second inequalities hold from $\rho_1 = \max\{\rho_1, \rho_2, \cdots, \rho_\tau\} \geq \rho_{i^*}$ and Eqn. (51), respectively. We also have

$$\begin{aligned}
&\eta_1(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[\tau]}\{\vartheta_j\}\right] - \sum_{l=2}^{\tau}\eta_l(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_l = \max_{j\in[\tau]}\{\vartheta_j\}\right] \\
&\leq \sum_{l=1}^{\tau}\|\eta_l(\boldsymbol{x}) - \rho_l\| + \rho_1 \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[\tau]}\{\vartheta_j\}\right] - \sum_{l=2}^{\tau}\rho_l \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_l = \max_{j\in[\tau]}\{\vartheta_j\}\right] \\
&\leq \tau L\nu(C_i) + \sum_{l=2}^{\tau}(\rho_1 - \rho_l) \Pr_{y_1,\ldots,y_{n'}}[\vartheta_1 < \vartheta_l] \qquad (54)
\end{aligned}$$

where the last inequality holds from Eqn. (51) and Lemma 5. This follows that, from Lemma 6,

$$\eta_1(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_1 < \max_{j\in[\tau]}\{\vartheta_j\}\right] - \sum_{l=2}^{\tau}\eta_l(\boldsymbol{x}) \Pr_{y_1,\ldots,y_{n'}}\left[\vartheta_l = \max_{j\in[\tau]}\{\vartheta_j\}\right] \leq \tau L\nu(C_i) + \frac{\tau}{\sqrt{en'}}\,.$$

Hence, we have, from Eqn. (50)

$$\begin{aligned}
&\Pr_{S_n,(\boldsymbol{x},y)}\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i, \Gamma_1\right]\Pr[C_i]\Pr[\Gamma_1] \\
&\leq E_{\boldsymbol{x}}\left[\min_{l\in[\tau]}\{1-\eta_l(\boldsymbol{x})\} | \boldsymbol{x} \in C_i\right]\Pr[C_i] + (\tau+2)L\nu(C_i)\Pr[C_i] + \tau\sqrt{2\Pr[C_i]/en}\,,
\end{aligned}$$

which completes the proof by combining with Eqn. (49). □

6.12. Proofs of Theorems 5-8

It is observable that the proof of Theorem 8 is exactly the same as that of Theorem 4, and we will present the detailed proofs for Theorems 5-7.

Proof of Theorem 5. We follow the proof of Theorem 1 and utilize Lemma 7 for multi-class learning. We first have

$$R_{\mathcal{D}}(f_{S_n,\Theta}) \leq E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}[\mathbb{I}[\boldsymbol{x}\in\Lambda]] + E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{M}(\eta_1(\boldsymbol{x}),\ldots,\eta_\tau(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y]\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right] \qquad (55)$$

where $\Lambda = \{\boldsymbol{x}\in\mathcal{X}\colon \nu(C(\boldsymbol{x})) \geq (1+\epsilon)\sqrt{d}/k^{1/8d}\}$. Notice that $C_1, C_2, \ldots, C_k$ is a partition of the instance space $\mathcal{X}$ from the construction of random tree. By the law of total probability, we have

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{M}(\eta_1(\boldsymbol{x}),\ldots,\eta_\tau(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x}) \neq y]\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right] = \sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x}\in C_i]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i \nsubseteq \Lambda]\,,$$

where we use $C(\boldsymbol{x}) = C_i$ for every $\boldsymbol{x}\in C_i$. From Eqn. (55), we have

$$\begin{aligned}
E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \leq\ & E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\sqrt{d}/k^{1/8d}\right]\right] \\
&+ E_{\Theta}\left[\sum_{i=1}^{k}E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i \nsubseteq \Lambda]\right]. \qquad (56)
\end{aligned}$$

From Lemma 3, we have

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})] \geq (1+\epsilon)\sqrt{d}/k^{1/8d}\right]\right] \leq \frac{ed}{(1+\epsilon)k^{1/8d}}\,. \qquad (57)$$

From Lemma 7 and the definition of set Λ, we can bound Eqn. (56) as follows

$$\sum_{i=1}^{k}E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i \nsubseteq \Lambda] \leq R^*_{\mathcal{D}} + \frac{(\tau+2)(1+\epsilon)L\sqrt{d}}{k^{1/8d}} + \sum_{i=1}^{k}\tau\sqrt{\frac{2\Pr[C_i]}{en}} + \frac{3k}{n}\,, \qquad (58)$$

where we use the law of total expectation and $R^*_{\mathcal{D}} = E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}[\min_{j\in[\tau]}\{1-\eta_j(\boldsymbol{x})\}]$. By Jensen's inequality, we have $\sum_{i=1}^{k}\sqrt{\Pr[C_i]} \leq \sqrt{k}$, and this follows that, from Eqns. (56)-(58),

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \le R^*_{\mathcal{D}} + \frac{ed}{(1+\epsilon)k^{1/8d}} + \frac{(\tau+2)(1+\epsilon)L\sqrt{d}}{k^{1/8d}} + \tau\sqrt{\frac{2k}{en}} + \frac{3k}{n}\,.$$

We complete the proof by setting $\epsilon = \sqrt{e\sqrt{d}/(\tau+2)L} - 1$ and combining with Lemma 1 and simple algebra calculations. □

Proof of Theorem 6. We follow the proof of Theorem 2 and utilize Lemma 7 for multi-class learning. We first have

$$R_{\mathcal{D}}(f_{S_n,\Theta}) \le E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{I}[\boldsymbol{x}\in\Lambda]\right] + E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{M}(\eta_1(\boldsymbol{x}),\ldots,\eta_\tau(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}]\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right] \tag{59}$$

with $\Lambda = \{\boldsymbol{x}\in\mathcal{X}\colon \nu(C(\boldsymbol{x})) \ge (1+\epsilon)\sqrt{d}/k^{1/4d}\}$. Notice that $C_1, C_2, \ldots, C_k$ is a partition of the instance space $\mathcal{X}$, and we have

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{y\sim\mathcal{M}(\eta_1(\boldsymbol{x}),\ldots,\eta_\tau(\boldsymbol{x}))}[f_{\Theta,S_n}(\boldsymbol{x})\neq y]\mathbb{I}[\boldsymbol{x}\notin\Lambda]\right] = \sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\nsubseteq\Lambda]\,.$$

This follows that, from (59),

$$\begin{aligned} E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \le{}& E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})] \ge (1+\epsilon)\sqrt{d}/k^{1/4d}\right]\right] \\ &+ E_{\Theta}\left[\sum_{i=1}^{k} E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\nsubseteq\Lambda]\right]. \end{aligned} \tag{60}$$

From Lemma 16, we have

$$E_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\Pr_{S_n,\Theta}\left[\nu[C(\boldsymbol{x})] \ge (1+\epsilon)\sqrt{d}/k^{1/4d}\right]\right] \le \frac{3d/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}}\,. \tag{61}$$

From Lemma 7 and the definition of Λ, we can upper bound Eqn. (60) by

$$\sum_{i=1}^{k} E_{S_n}[\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\nsubseteq\Lambda_2] \le R^*_{\mathcal{D}} + \frac{(\tau+2)(1+\epsilon)L\sqrt{d}}{k^{1/4d}} + \tau\sqrt{\frac{2k}{en}} + \frac{3k}{n}\,.$$

This follows that, from Eqns. (60)-(61),

$$E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta})\right] \le R^*_{\mathcal{D}} + \frac{3d/2}{(1+\epsilon)^{3/2}k^{1/3.6846d}} + \frac{(\tau+2)(1+\epsilon)L\sqrt{d}}{k^{1/4d}} + \tau\sqrt{\frac{2k}{en}} + \frac{3k}{n}\,.$$

We complete the proof by setting $\epsilon = \left(9\sqrt{d}k^{\frac{1}{4d}-\frac{1}{3.6848d}}/4(\tau+2)L\right)^{2/5} - 1$ and combining with Lemma 1 and some simple algebraic calculations. □

Proof of Theorem 7. We follow the proof of Theorem 3 and utilize Lemma 7 for multi-class learning. Let $C_1, C_2, \ldots, C_k$ be a partition of instance space $\mathcal{X}$, which are associated with k leaves of random tree. Based on the law of total probability, we have

$$R_{\mathcal{D}}(f_{S_n,\Theta}) = \sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\in\Lambda] \tag{62}$$

$$+ \sum_{i=1}^{k}\Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y|\boldsymbol{x}\in C_i]\Pr[\boldsymbol{x}\in C_i]\mathbb{I}[C_i\notin\Lambda]\,, \tag{63}$$

where $\Lambda = \{C_i\colon$ all training examples in C_i have the same label$\}$.

We first study the case $C_i\in\Lambda$ $(i\in[k])$, and it holds that have, for $\kappa\ge 2$,

$$\begin{aligned} &\Pr[C_i]\Pr\left[f_{\Theta,S_n}(\boldsymbol{x})\neq y\middle|\boldsymbol{x}\in C_i\right] \\ &\quad\le \Pr[C_i]\Pr[|C_i\cap S_n|\le\kappa] + \Pr[f_{\Theta,S_n}(\boldsymbol{x})\neq y\big||C_i\cap S_n|>\kappa,\boldsymbol{x}\in C_i]\Pr[|C_i\cap S_n|>\kappa,\boldsymbol{x}\in C_i]\,. \end{aligned}$$

From $C_i\in\Lambda$, all training examples in C_i have the same label, and assume that the labels of training examples in C_i are all 1 without loss of generality. We have $f_{\Theta,S_n}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}\in C_i$. Let $\bar{\eta}_1(C_i) = E[\eta_1(\boldsymbol{x})|\boldsymbol{x}\in C_i]$ denote the expected conditional probability over cell C_i. If $\bar{\eta}_1(C_i)\ge 1-\epsilon$, then we have

Table 1
Benchmark datasets.

datasets	# instance	# feature	# label	datasets	# instance	# feature	# label
obesity	2,111	16	7	drybean	13,611	16	7
shill	6,321	9	2	eggeye	14,980	14	2
mfcc	7,195	22	4	magic04	19,020	10	2
firmteacher	10,800	16	4	letter	20,000	16	26
mapping	10,845	28	6	occupancy	20,560	5	3
pendigits	10,992	16	10	firewall	65,532	11	4

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \big| |C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i] \Pr[|C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i] \leq \epsilon \ ;$$

and if $\bar{\eta}_1(C_i) < 1 - \epsilon$ and $C_i \in \Lambda$, then we have

$$\Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \big| |C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i] \Pr[|C_i \cap S_n| > \kappa, \boldsymbol{x} \in C_i] \leq \Pr[|C_i \cap S_n| > \kappa] \leq \exp(-\kappa\epsilon) \ .$$

This follows that, for $C_i \in \Lambda$,

$$\Pr[C_i] \Pr\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \Big| \boldsymbol{x} \in C_i \right] \leq \Pr[C_i] \Pr[|C_i \cap S_n| \leq \kappa] + \Pr[C_i](\epsilon + \exp(-\kappa\epsilon)) \ .$$

By setting $\epsilon = (\ln \kappa)/\kappa$, we have

$$\Pr[C_i] \Pr\left[f_{\Theta,S_n}(\boldsymbol{x}) \neq y \Big| \boldsymbol{x} \in C_i \right] \leq \Pr[C_i] \Pr[|C_i \cap S_n| \leq \kappa] + \Pr[C_i] \frac{1 + \ln \kappa}{\kappa} \ .$$

Combining with Lemma 17, we can upper bound the Eqn. (62) as follows:

$$\sum_{i=1}^{k} \Pr[f_{\Theta,S_n}(\boldsymbol{x}) \neq y | \boldsymbol{x} \in C_i] \Pr[\boldsymbol{x} \in C_i] \mathbb{I}[C_i \in \Lambda] \leq \frac{k\kappa}{n} \left(1 + \sqrt{\frac{2}{\kappa}} \right) + \frac{1 + \ln \kappa}{\kappa} \ . \tag{64}$$

We now consider the case $C_i \notin \Lambda$, i.e., the instances in C_i have different labels. It is easy to get the height $h(C_i) \geq \log_2 k - 2$ and the splitting times for each dimension are more than $(\log_2 k - 2)/d - 1$ from the construction of random tree in Algorithm 1. Hence, we upper bound the diameter of rectangle cell C_i as follows:

$$\nu(C_i) \leq \sqrt{d} \left(\frac{1}{2} \right)^{(\log_2 k - 2)/d - 1} = \frac{2^{1+2/d} \sqrt{d}}{k^{1/d}} \leq \frac{8\sqrt{d}}{k^{1/d}} \ .$$

This follows that, from Lemma 7 and Eqns. (62)-(64),

$$\begin{aligned}
&E_{S_n,\Theta}[R_{\mathcal{D}}(f_{S_n,\Theta})] \\
&\leq R^*_{\mathcal{D}} + \frac{k\kappa}{n} \left(1 + \sqrt{\frac{2}{\kappa}} \right) + \frac{1}{\kappa}(1 + \ln \kappa) + \frac{8(\tau + 2)L\sqrt{d}}{k^{1/d}} + \frac{3k}{n} + \tau \sum_{i=1}^{k} \sqrt{\frac{2 \Pr[C_i]}{en}} \\
&\leq R^*_{\mathcal{D}} + \frac{k\kappa}{n} \left(1 + \sqrt{\frac{2}{\kappa}} \right) + \frac{1}{\kappa}(1 + \ln \kappa) + \frac{8(\tau + 2)L\sqrt{d}}{k^{1/d}} + \frac{3k}{n} + \tau \sqrt{\frac{2k}{en}} \ .
\end{aligned}$$

We have, by setting $\kappa = \left\lceil \sqrt{n \ln n / k} \right\rceil$ with algebra calculations,

$$\begin{aligned}
&E_{S_n,\Theta}\left[R_{\mathcal{D}}(f_{S_n,\Theta}) \right] \\
&\leq R^*_{\mathcal{D}} + \sqrt{\frac{k \ln n}{n}} + \sqrt[4]{\frac{4k^3 \ln n}{n^3}} + \sqrt{\frac{k}{n \ln n}} \left(1 + \frac{1}{2} \ln \frac{n \ln n}{k} \right) + \frac{6k}{n} + \tau \sqrt{\frac{2k}{en}} + \frac{8(\tau + 2)L\sqrt{d}}{k^{1/d}} \\
&\leq R^*_{\mathcal{D}} + 2\sqrt{\frac{k \ln n}{n}} + \sqrt[4]{\frac{4k^3 \ln n}{n^3}} + \frac{6k}{n} + \left(\tau \sqrt{\frac{2}{e}} + 1 \right) \sqrt{\frac{k}{n}} + \frac{8(\tau + 2)L\sqrt{d}}{k^{1/d}} \quad (n \geq 4, k \geq 2) \ ,
\end{aligned}$$

which completes the proof by combining with Lemma 1. □

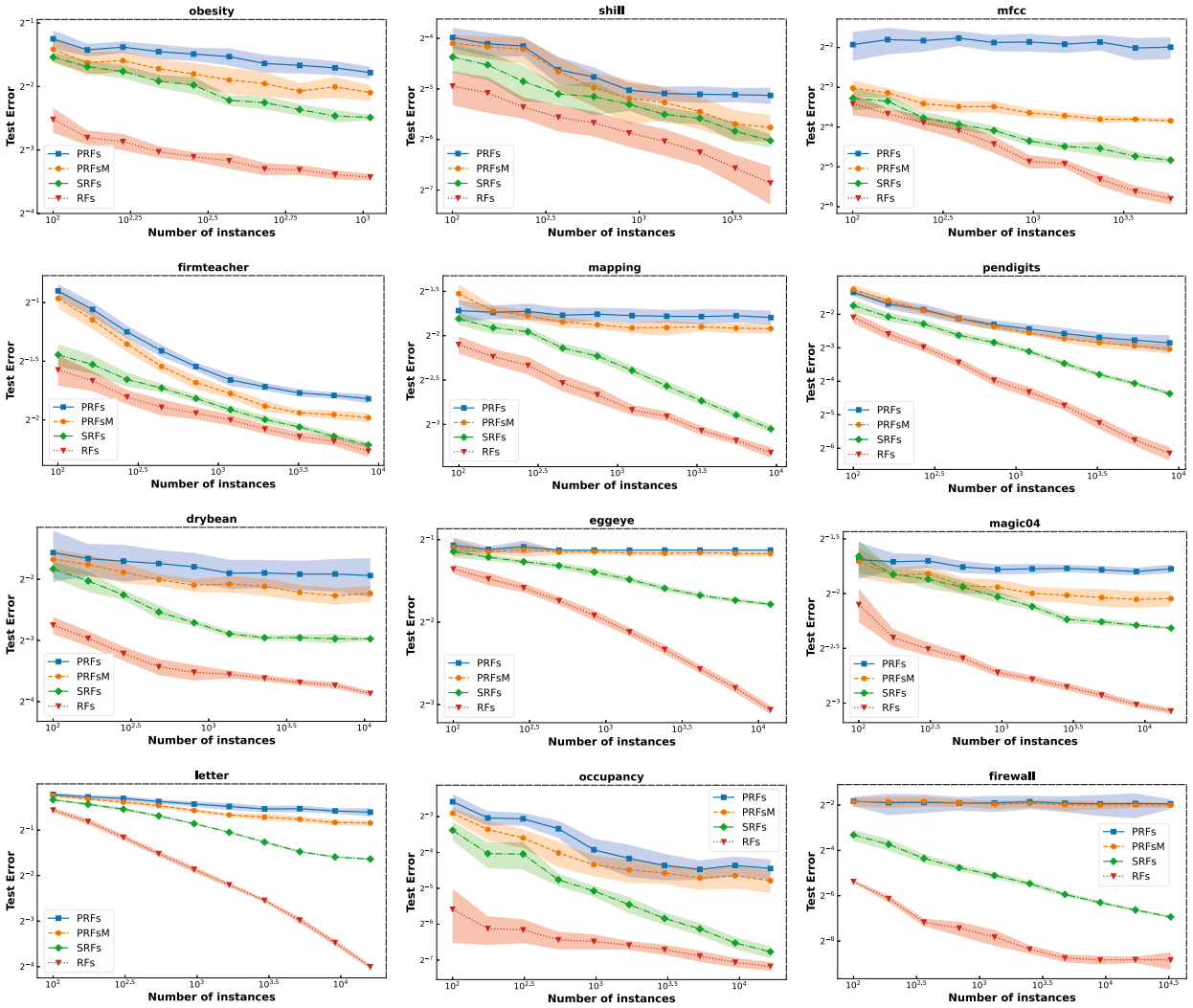


Fig. 1. The convergence curves of four random forests (PRFs, RFs, PRFsM and SRFs) on benchmark datasets.

7. Experiments

This section tries to present empirical studies to support our theoretical analysis. We conduct experiments on twelve benchmark datasets,[1] and the details are summarized in Table 1. Most datasets have been used in previous studies of random forests, and the features have been scaled to $[0, 1]$ for all datasets.

We compare with four random forests, which have been studied theoretically in this work.

- PRFs: Purely random forests [11];
- RFs: Breiman's original random forests [12];
- PRFsM: Purely random forests with midpoint splits as shown in Section 3;
- SRFs: The simplified variant of Breiman's original random forests as shown in Section 4.

All experiments are performed with Python 3 on an Intel Core i9-10900X processor under Ubuntu 20.04 with 128GB RAM. For each datasets, five-fold cross-validation is executed to select the parameters of random trees, that is, number of random trees $m \in \{10, 20, \cdots, 200\}$ for ensemble and leaves number $k \in \{500, 1000, \cdots, 10000\}$. For Breiman's original random forests, we adopt the Gini index as the splitting criterion and randomly select $\lfloor \sqrt{d} \rfloor$ candidate features for splitting.

[1] https://archive.ics.uci.edu/ml/datasets.php.

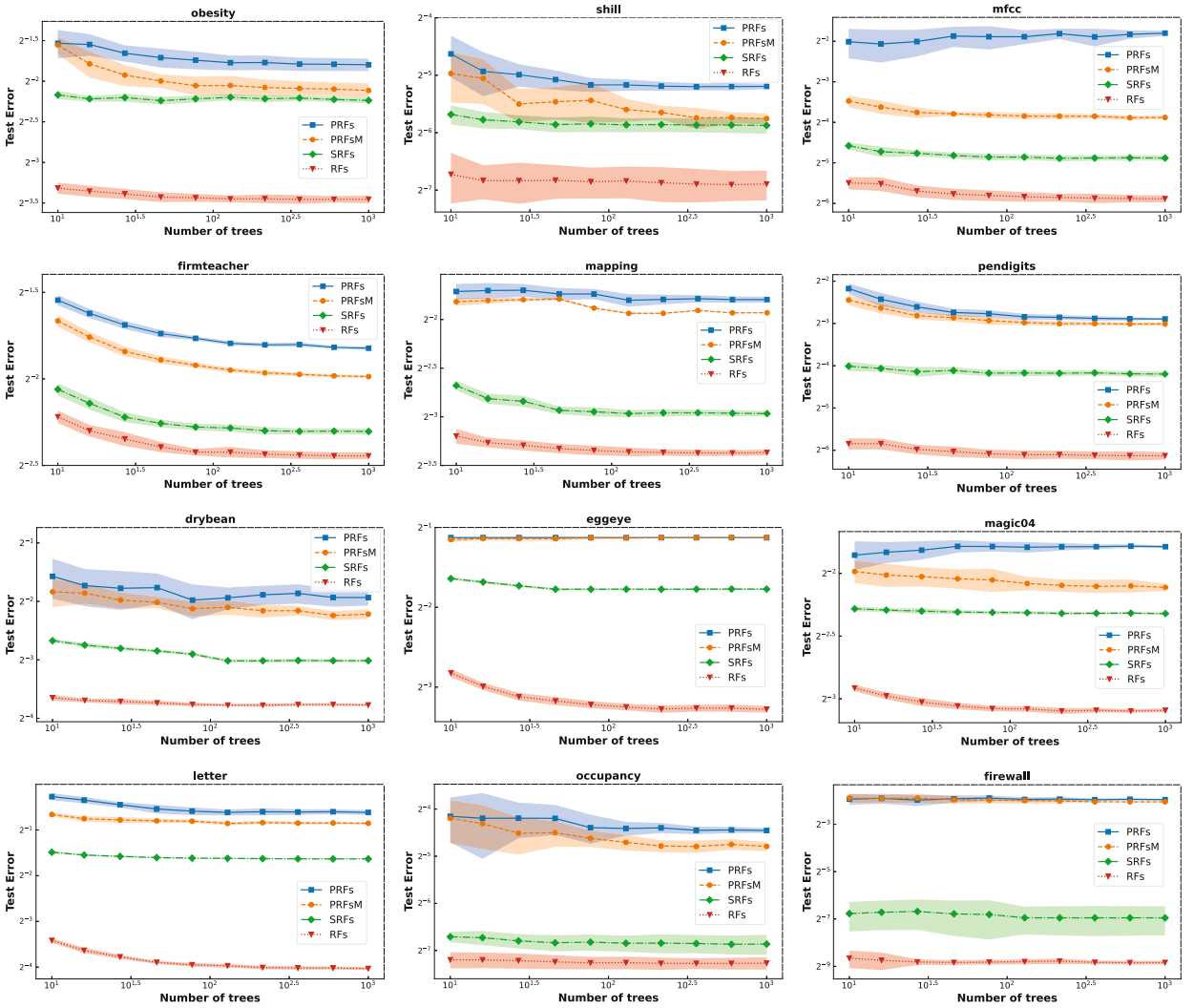


Fig. 2. The influence of number of randomized trees on four random forests: PRFs, RFs, PRFsM and SRFs.

We take the test error as our performance measure for classification, which could directly support our theoretical results on the convergence rates of different random forests to Bayes' risk. This is different from random forests regression, which generally takes mean squared error (MSE) as performance measure. For all datasets, we take five trials of 5-fold cross validation, and the final test error is obtained by averaging over these 25 runs. We perform the log-log plots of test error versus number of instances for all datasets, as shown in Fig. 1, which could make more clear empirical comparisons for different random forests on their convergence rates.

As can be seen from Fig. 1, purely random forests (PRFs) show the slowest convergence curves for all datasets, which are in coordination with Theorems 1 and 5 of the lowest convergence rates. In contrast, purely random forests with mid-point split (PRFsM) take faster convergence than purely random forest empirically, which well supports our theoretical results (Theorems 2 and 6). The simplified variant of Breiman's random forests (SRFs) take faster convergence curves than PRFs and PRFsM, because Theorems 3 and 7 show faster convergence rates theoretically.

It is also observable that, from Fig. 1, Breiman's original random forests (RFs) take the fastest convergence curves for all datasets, and an intuitive explanation is that random forests correlate the randomization process with data-dependent tree structure based on the gini index criterion. However, theoretical understanding remains big challenges from a technical view, and we leave it for future work.

We further exploit the influence of number of randomized trees (in voting) on the convergence of different random forests, and Fig. 2 presents the log-log plots of test error versus number of randomized trees for all datasets. As we can see, four different random forests (PRFs, RFs, PRFsM and SRFs) achieve relatively stable performance for all datasets when we take more than 100 randomized trees (in voting). This evidence empirically supports Lemma 1 on the convergence rate of

random forests from the average of randomized trees, and it is also in accordance with previous empirical studies on the selection of 100 randomized trees in experiments for random forests [12,29,41,50].

8. Conclusion

Random forests have been recognized as one of the successful algorithms for classification and regression, and most previous studies focus on the convergence analysis of random forests for regression. This work takes one step towards the convergence analysis of random forests for classification. Specifically, we present the finite-sample convergence rates of purely random forests, as well as the simplified variant of Breiman's original random forests. We also achieve the same convergence rates of random forests for multi-class learning as that of binary classification, yet with different constants. It is still a long way to fully understand random forests with relevant mechanisms such as bootstrap sampling, data-dependence tree structure, tree pruning, etc., and we leave those to future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors want to thank the associate editor and anonymous reviewers for their helpful comments and suggestions. This research was supported by National Key R&D Program of China (2021ZD0112802) and NSFC (61921006, 61876078), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] D. Amaratunga, J. Cabrera, Y.-S. Lee, Enriched random forests, Bioinformatics 24 (18) (2008) 2010–2014.
[2] Y. Amit, D. Geman, Shape quantization and recognition with randomized trees, Neural Comput. 9 (7) (1997) 1545–1588.
[3] S. Arlot, R. Genuer, Analysis of purely random forests bias, CoRR/Abstract, arXiv:1407.3939, 2014.
[4] S. Athey, J. Tibshirani, S. Wager, Generalized random forests, Ann. Stat. 47 (2) (2019) 1148–1178.
[5] J.-Y. Audibert, A. Tsybakov, Fast learning rates for plug-in classifiers, Ann. Stat. 35 (2) (2007) 608–633.
[6] S. Basu, K. Kumbier, J. Brown, B. Yu, Iterative random forests to discover predictive and stable high-order interactions, Proc. Natl. Acad. Sci. 115 (8) (2018) 1943–1948.
[7] M. Belgiu, L. Dragut, Random forest in remote sensing: a review of applications and future directions, ISPRS J. Photogramm. Remote Sens. 114 (2016) 24–31.
[8] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[9] G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, J. Mach. Learn. Res. 9 (2008) 2015–2033.
[10] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.
[11] L. Breiman, Some infinity theory for predictor ensembles, Technical Report 579, Statistics Department, UC Berkeley, Berkeley, CA, 2000.
[12] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[13] L. Breiman, Consistency for a simple model of random forests, Technical Report 670, Statistics Department, UC Berkeley, Berkeley, CA, 2004.
[14] S. Clémençon, M. Depecker, N. Vayatis, Ranking forests, J. Mach. Learn. Res. 14 (2013) 39–73.
[15] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27.
[16] A. Criminisi, J. Shotton, Decision Forests for Computer Vision and Medical Image Analysis, Springer Science & Business Media, 2013.
[17] A. Criminisi, J. Shotton, E. Konukoglu, Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning, Found. Trends Comput. Graph. Vis. 7 (2–3) (2012) 81–227.
[18] D. Cutler, T. Edwards Jr, K. Beard, A. Cutler, K. Hess, J. Gibson, J. Lawler, Random forests for classification in ecology, Ecology 88 (11) (2007) 2783–2792.
[19] M. Denil, D. Matheson, N. Freitas, Consistency of online random forests, in: Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013, pp. 1256–1264.
[20] M. Denil, D. Matheson, N. De Freitas, Narrowing the gap: random forests in theory and in practice, in: Proceedings of the 31th International Conference on Machine Learning, Beijing, China, 2014, pp. 665–673.
[21] L. Devroye, A note on the height of binary search trees, J. ACM 33 (3) (1986) 489–498.
[22] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[23] T.G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Mach. Learn. 40 (2) (2000) 139–157.
[24] V. Dinh, L. Ho, N. Cuong, D. Nguyen, B. Nguyen, Learning from non-iid data: fast rates for the one-vs-all multiclass plug-in classifiers, in: Proceeding of the 12rd International Conference on Theory and Applications of Models of Computation, Singapore, 2015, pp. 375–387.
[25] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, J. Mach. Learn. Res. 15 (1) (2014) 3133–3181.
[26] W. Gao, Z.-H. Zhou, Towards convergence rate analysis of random forests for classification, in: Advances in Neural Information Processing Systems 33, MIT Press, Cambridge, MA, 2020, pp. 9300–9311.
[27] R. Genuer, Variance reduction in purely random forests, J. Nonparametr. Stat. 24 (3) (2012) 543–562.
[28] R. Genuer, J. Poggi, C. Tuleau, Random forests: some methodological insights, CoRR/Abstract, arXiv:0811.3619, 2008.
[29] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, N. Villa-Vialaneix, Random forests for big data, Big Data Res. 9 (2017) 28–46.
[30] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.
[31] J. Goetz, A. Tewari, P. Zimmerman, Active learning for non-parametric regression using purely random trees, in: Advances in Neural Information Processing Systems 31, MIT Press, Cambridge, MA, 2018, pp. 2537–2546.
[32] L. Györfi, R. Weiss, Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces, CoRR/Abstract, arXiv:2010.00636, 2020.

[33] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[34] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 301 (58) (1963) 13–30.
[35] J. Kazemitabar, A. Amini, A. Bloniarz, A. Talwalkar, Mondrian forests: efficient online random forests, in: Advances in Neural Information Processing Systems 30, MIT Press, Cambridge, MA, 2017, pp. 426–435.
[36] J. Klusowski, Sharp analysis of a simple model for random forests, CoRR/Abstract, arXiv:1805.02587, 2018.
[37] A. Kontorovich, R. Weiss, Maximum margin multiclass nearest neighbors, in: Proceeding of the 31st International Conference on Machine Learning, Beijing, China, 2014, pp. 892–900.
[38] S. Kwok, C. Carter, Multiple decision trees, in: Proceedings of the 4th Annual Conference on Uncertainty in Artificial Intelligence, Minneapolis, MN, 1988, pp. 327–338.
[39] B. Lakshminarayanan, D. Roy, Y. Teh, Mondrian forests: efficient online random forests, in: Advances in Neural Information Processing Systems 27, MIT Press, Cambridge, MA, 2014, pp. 3140–3148.
[40] X. Li, Y. Wang, S. Basu, K. Kumbier, B. Yu, A debiased MDI feature importance measure for random forests, in: Advances in Neural Information Processing Systems 32, MIT Press, Cambridge, MA, 2019, pp. 8047–8057.
[41] Y. Lin, Y. Jeon, Random forests and adaptive nearest neighbors, J. Am. Stat. Assoc. 101 (474) (2006) 578–590.
[42] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, in: Advances in Neural Information Processing Systems 26, MIT Press, Cambridge, MA, 2013, pp. 431–439.
[43] N. Meinshausen, Quantile regression forests, J. Mach. Learn. Res. 7 (2006) 983–999.
[44] B. Menze, M. Kelm, D. Splitthoff, U. Koethe, F. Hamprecht, On oblique random forests, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 2011, pp. 453–469.
[45] M. Mitzenmacher, E. Upfal, Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Cambridge University Press, 2005.
[46] J. Mourtada, S. Gaïffas, E. Scornet, Universal consistency and minimax rates for online Mondrian forests, in: Advances in Neural Information Processing Systems 30, MIT Press, Cambridge, MA, 2017, pp. 3758–3767.
[47] N. Puchkin, V. Spokoiny, An adaptive multiclass nearest neighbor classifier, ESAIM Probab. Stat. 24 (2020) 69–99.
[48] Y. Qi, Random forest for bioinformatics, in: Ensemble Machine Learning, Springer, 2012, pp. 307–323.
[49] B. Reed, The height of a random binary search tree, J. ACM 50 (3) (2003) 306–332.
[50] M. Robnik-Šikonja, Improving random forests, in: Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy, 2004, pp. 359–370.
[51] J. Rodriguez, L. Kuncheva, C. Alonso, Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1619–1630.
[52] E. Scornet, On the asymptotics of random forests, J. Multivar. Anal. 146 (2016) 72–83.
[53] E. Scornet, G. Biau, J. Vert, Consistency of random forests, Ann. Stat. 43 (4) (2015) 1716–1741.
[54] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, Cambridge, 2014.
[55] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.
[56] V. Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan, B. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 1947–1958.
[57] M. Taddy, R. Gramacy, N. Polson, Dynamic trees for learning and design, J. Am. Stat. Assoc. 106 (493) (2011) 109–123.
[58] C. Tang, D. Garreau, U. von Luxburg, When do random forests fail?, in: Advances in Neural Information Processing Systems 31, MIT Press, Cambridge, MA, 2018, pp. 2983–2993.
[59] S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests, J. Am. Stat. Assoc. 113 (523) (2018) 1228–1242.
[60] S. Wager, T. Hastie, B. Efron, Confidence intervals for random forests: the jackknife and the infinitesimal jackknife, J. Mach. Learn. Res. 15 (1) (2014) 1625–1651.
[61] Y. Wang, S.-T. Xia, Q. Tang, J. Wu, X. Zhu, Bernoulli random forests: closing the gap between theoretical consistency and empirical soundness, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, NY, 2016, pp. 2167–2173.
[62] Y. Wang, S.-T. Xia, Q. Tang, J. Wu, X. Zhu, A novel consistent random forest framework: Bernoulli random forests, IEEE Trans. Neural Netw. Learn. Syst. 29 (8) (2017) 3510–3523.
[63] B.-B. Yang, W. Gao, M. Li, On the robust splitting criterion of random forest, in: Proceedings of the 19th IEEE International Conference on Data Mining, Beijing, China, 2019, pp. 1420–1425.
[64] Y. Yang, Minimax nonparametric classification - part I: rates of convergence, IEEE Trans. Inf. Theory 45 (7) (1999) 2271–2284.
[65] Z.-H. Zhou, J. Feng, Deep forest: towards an alternative to deep neural networks, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 2017, pp. 3553–3559.
[66] Z.-H. Zhou, J. Feng, Deep forest, Nat. Sci. Rev. 6 (1) (2019) 74–86.