Journal of Machine Learning Research 27 (2026) 1-95 Submitted 2/25; Revised 10/25; Published 5/26

Exogenous Randomness Empowering Random Forests

Tianxing Mei TIANXINGMEI@LN.EDU.HK
Hong Kong Institute of Business Studies (HKIBS)
Faculty of Business
Lingnan University
Hong Kong

Yingying Fan FANYINGY@MARSHALL.USC.EDU
Jinchi Lv JINCHILV@MARSHALL.USC.EDU
Data Sciences and Operations Department
University of Southern California
Los Angeles, CA 90089, USA

Editor: Ji Zhu

Abstract

We offer theoretical and empirical insights into the impact of exogenous randomness on the effectiveness of random forests with tree-building rules independent of training data. We formally introduce the concept of exogenous randomness which can come from feature subsampling or tie-breaking in tree-building processes. We develop non-asymptotic expansions for the mean squared error (MSE) for both individual trees and forests and establish sufficient and necessary conditions for their consistency. In the special example of the linear regression model with independent features, our MSE expansions are more explicit, providing more understanding of the random forests' mechanisms. It also allows us to derive an upper bound on the MSE with explicit consistency rates for trees and forests. Guided by our theoretical findings, we conduct simulations to further explore how exogenous randomness enhances random forests performance. Our findings unveil that feature subsampling reduces both the bias and variance of random forests compared to individual trees, serving as an adaptive mechanism to balance bias and variance. Furthermore, our results reveal an intriguing phenomenon: the presence of noise features can act as a "blessing" in enhancing the performance of random forests thanks to feature subsampling.

Keywords: random forests, partitioning rule, feature subsampling, ensemble, exogenous randomness

1. Introduction

Random forests, a type of ensemble estimator, have gained significant attention due to their appealing empirical performance across various applications over the past two decades. At a high level, random forests build individual trees (a type of partitioning estimator) using some recursive tree building rule (a type of domain partitioning rule), and then combine these trees into a single ensemble estimator. In Breiman's random forests algorithm (Breiman, 2001), a key innovation is feature subsampling, where a random subset of features is chosen during each split, enhancing diversity among the trees. The size of the feature subset is

chosen independently of tree construction and is controlled by a parameter $\gamma \in (0, 1]$ with $\gamma = 1$ representing no feature subsampling.

Despite their success, random forests are often regarded as a "black-box" model, as the reasons behind their effectiveness remain incompletely understood. A primary distinction between random forests, individual trees, and bagged trees is feature subsampling; yet its specific impact on performance needs deeper exploration. An important goal of our study is to enhance the understanding of feature subsampling in random forests success. For this purpose, we simplify other aspects of the random forests algorithm by considering deterministic tree-building rules that are independent of the training data and excluding subsampling along the sample dimension (e.g., bootstrapping). We recognize that these simplifications exclude training-data-dependent tree-building rules, such as sample CART (Breiman, 2001), as well as tree bagging estimators which are relatively well understood thanks to the extensive existing literature on bootstrapping. However, they also allow us to peel out the effects of feature subsampling and ensemble. Additionally, the population CART can serve as a good proxy of the sample CART when the sample size is large and the tree depth is not too large, in the sense that with high probability, for all splits in the tree building process, the impurity decrease by the sample CART is asymptotically close to the impurity decrease by the population CART, as has been formally characterized in Chi et al. (2022) (Theorems 3 and 4 therein); see also Klusowski (2020).

We start with considering nonparametric estimation via partitioning estimator ensembles within a general regression framework (see Section 2). Our study focuses on the individual base learners being the partitioning estimators, including tree estimators as a special case, constructed from some training-data-independent partitioning rules. This includes popularly studied methods such as the population CART random forest (Klusowski, 2020; Chi et al., 2022), the Mondrain forest (Mourtada et al., 2020; Cattaneo et al., 2024), the centered forest (Biau, 2012; Klusowski, 2021), and the mean/median forests (Scornet, 2016). To simplify the presentation, we use the term "random forest" (RF) in a broad sense, equating it with the class of partitioning estimator ensemble methods with the feature subsampling component but excluding bootstrap throughout the paper.

In forming the RF estimators, two sources of randomness can arise: the *endogenous* randomness related to training data, and the *exogenous* randomness that is independent of training data. An example of the latter is the randomness introduced by feature subsampling in Breiman's RF. We consider ensemble estimators incorporating feature subsampling when forming their base estimators, and thus, there always exists *exogenous* randomness in our RF. It is worth noting that for some partitioning rules, such as the population CART, there can exist *additional* exogenous randomness beyond the one caused by feature subsampling, such as the randomness related to tie-breaking in choosing the next domain splitting location. We formalize these two types of exogenous randomness in Definition 1. Our study reveals that both types contribute to the success of RF, albeit in *distinct* ways. While our analysis retains tie-breaking randomness for mathematical rigor and completeness, our focus is *primarily* on feature subsampling randomness, as tie-breaking occurs more rarely in practice.

As a general framework, we establish nonasymptotic expansions for the mean squared error (MSE) of the individual tree estimator and the corresponding forest estimator. Our theoretical results in Section 3 show that with exogenous randomness, i) ensemble ensures

that RF has both smaller squared bias and variance than its base estimator; ii) the leading order contributions to both the squared bias and variance are completely different for the forest estimator and the base estimator; and iii) forest estimators can be consistent under weaker conditions than their individual base estimators. We provide sufficient and necessary conditions for the consistency of the partitioning estimator and related forest estimator. These results hold broadly for any partitioning estimator ensembles with training-data-independent partitioning rules and exogenous randomness, including those previously reviewed ones. While similar message as summarized in i) exists in the literature (e.g., Liu and Mazumder (2024); Curth et al. (2024)), those prior studies primarily provide empirical evidence without much theoretical support. In contrast, our paper backs these findings by rigorous theoretical justifications and characterizes the interaction mechanism among individual base estimator; see Theorem 4 and Equation (16). A key step in our analyses of the interaction mechanism is the novel conditional bias-variance decomposition of the MSE, where the randomness in the training data is integrated out first *before* we account for the exogenous randomness. Such a strategy differs from the common approach in the existing literature (Scornet et al., 2015; Scornet, 2016; Curth et al., 2024) in that the exogenous randomness is integrated out first in those works.

To gain additional insights into the CART partitioning rule, we further consider a simplified setting of sparse linear regression with independent features. Two different feature distributions, binary and continuous uniform, are studied. Thanks to the simplified model setting, our general results on the MSE expansion take mathematically more specific forms, providing us additional insights into the effects of exogenous randomness and ensemble. We also establish a (conservative) consistency rate for both the tree and the forest estimators, which clearly shows a bias-variance trade-off as the tree depth and γ vary. In the special case of one signal feature, we derive explicit consistency rates for forest and tree estimators, where the dependence on model parameters such as subsampling rate and feature dimensionality are explicit and clearly support the blessing of dimensionality phenomenon. We further conduct theory-guided simulations based on our exact MSE expansions. Our theoretical and simulation results reveal an interesting yet surprising phenomenon of *blessing* of dimensionality and noise features, which is made possible by feature subsampling. We list our additional major findings here: iv) the exogenous randomness from tie-breaking (if present) in partitioning rules, often overlooked in the literature, can improve the performance of ensemble estimators; v) when $\gamma < 1$, the exogenous randomness caused by feature subsampling and the existence of numerous noise features act together to enable early-stage variance reduction during tree building, potentially coexisting with bias reduction for a while as tree depth grows; additionally, feature subsampling may also help with bias reduction compared to the $\gamma = 1$ case, although this effect can be model-specific; vi) for each fixed γ value, as tree grows deeper, RF gradually shifts its focus from bias reduction to variance reduction (although they can coexist at all tree depths), often resulting in a U-shaped MSE curve. Although these additional insights are gained in linear models, we conjecture that they can carry over to broad model settings. We believe that some of the findings also shed light on the sample CART random forests, because of their close connection as revealed in the literature (Chi et al., 2022), especially the role that feature subsampling plays in balancing the bias and variance of forest estimators.

1.1 Related Literature

Existing literature on RFs broadly falls into two categories. The first category focuses on theoretically understanding RF's consistency and predictive performance. Due to the complicated training-data-dependent nature of the sample CART, there only exist limited works addressing the consistency of Breiman's original RF, including works on the one-split stumps (Bühlmann and Yu, 2002), sparse additive regression models (Scornet et al., 2015; Klusowski and Tian, 2024), binary feature models (Syrgkanis and Zampetakis, 2020), and high-dimensional nonlinear models satisfying the sufficient impurity decrease (SID) condition (Chi et al., 2022) and the merged-staircase property (MSP) (Tan et al., 2024). To make the problem more accessible, many studies have instead focused on stylized RF variants, such as purely random forests (Biau, 2012; Klusowski, 2021), median forests (Scornet, 2016), Mondrian forests (Mourtada et al., 2020; Cattaneo et al., 2024), and honest forests (Wager and Athey, 2018; Athey et al., 2019). These studies, through their specific ways of introducing exogenous randomness either in the partitioning process or in the evaluation step, have provided important theoretical insights into the mechanisms of the original CART-based RF and revealed the interplay between parameter tuning and consistency within their respective modeling frameworks. In contrast, our work focuses on the role of *general* exogenous randomness in the partitioning process, including the one caused by feature subsampling, and formulates a framework to understand how such randomness induces interactions across partitions, thereby explaining *why and how* RFs can outperform a single decision tree.

The second category focuses on understanding the role of tuning parameters in the success of RFs, mostly from the empirical perspective accompanied by heuristic insights. For instance, Mentch and Zhou (2020) provided a degree-of-freedom explanation with empirical evidence that in a low signal-to-noise ratio (SNR) setting, the feature subsampling randomization serves as an implicit regularization mechanism like Lasso and ridge regression in improving the performance of RFs. Building upon this, Liu and Mazumder (2024) used a simulation study to show that even in a high SNR setting, RF can achieve both bias and variance reduction, highlighting the role of feature subsampling in helping RFs capture the hidden patterns missed by bagging. Lin and Jeon (2006) and Curth et al. (2024) interpreted RFs as adaptive potential nearest-neighbor estimators and highlighted that the feature subsampling and ensemble serve as smoother in out-of-bag predictions. Additionally, Mentch and Zhou (2022) also investigated how noise features influence the performance of bagging methods including RFs, and proposes a novel AugBagg procedure that introduces an additional set of noise features into the training data for improved prediction performance. The impact of other parameters, such as the tree depth, the forest size, the noise variance, and the maximum number of leaf nodes, as well as the interactions among these parameters, have been extensively explored in Fernández-Delgado et al. (2014); Probst and Boulesteix (2018); Le et al. (2023); Bernard et al. (2009); Zhou (2022); Zhou and Mentch (2023). Our paper contributes and strengthens this research direction by providing *both* theoretical and additional empirical insights into understanding how exogenous randomness contributes to the success of random forests over trees. Further comparisons and connections to the literature will be provided throughout the paper as the context becomes clearer.

2. Ensemble Estimator with Exogenous Randomness

Consider a training data set consisting of independent and identically distributed (i.i.d.) observations $\boldsymbol{Z} = \{(\boldsymbol{X}_i, Y_i)\}_{1 \le i \le n}$ drawn from a generic population $(\boldsymbol{X}, Y)$, where random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ takes values in a d-dimensional product topological space $\mathcal{X}^d = \mathcal{X} \times \cdots \times \mathcal{X}$ and response Y takes values in $\mathcal{Y} = \mathbb{R}$. The relationship between Y and $\boldsymbol{X}$ is expressed through the nonlinear model

$$Y = \mu(\boldsymbol{X}) + \varepsilon, \tag{1}$$

where $\mu : \mathcal{X}^d \to \mathcal{Y}$ represents an unknown function to be estimated, and ε is the random model error satisfying $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$ and $\mathrm{Var}(\varepsilon|\boldsymbol{X}) = \sigma^2(\boldsymbol{X})$ with $\sigma^2 : \mathcal{X}^d \to \mathcal{Y}$ a variance profile function. The model above accommodates heteroscedastic error variance. Our results in this paper are *general* and apply to both fixed and diverging dimensionality d.

Since random forests belong to the family of partitioning estimator ensemble methods, we introduce the concept of partitioning estimators and their ensemble in this section. Before proceeding, we need some necessary notation. Let g be a generic integrable function on $\mathcal{X}^d$ and A a generic measurable subset of $\mathcal{X}^d$. We use $\mathbb{E}_{\boldsymbol{X}}(g)$ and $\mathbb{P}_{\boldsymbol{X}}(A)$ to stand for $\mathbb{E}(g(\boldsymbol{X}))$ and $\mathbb{P}(\boldsymbol{X} \in A)$, respectively, and $I_A(\boldsymbol{x}) = I\{\boldsymbol{x} \in A\}$ to stand for the indicator function of set A. Denote by $\mathbb{P}_{\boldsymbol{X}}(A|B) = \mathbb{P}(\boldsymbol{X} \in A|\boldsymbol{X} \in B)$ for any $A \subset B$, $\mathbb{E}_{\boldsymbol{X}}(g|A) = \mathbb{E}(g(\boldsymbol{X})|\boldsymbol{X} \in A)$, and $\mathrm{Var}_{\boldsymbol{X}}(g|A) = \mathrm{Var}(g(\boldsymbol{X})|\boldsymbol{X} \in A)$. By $f(n) \lesssim g(n)$, we mean that there exists a constant C independent of n such that $|f(n)| \le Cg(n)$; if $f(n)$ depends on additional quantities d, μ, σ^2, and so on, the implicit constant C is assumed to be independent of those parameters unless specified otherwise. Throughout, we use $a_n \asymp b_n$ to denote that there exist positive constants c and C such that $cb_n \le a_n \le Cb_n$. For any real number x, the smallest integer greater than or equal to x is written as $\lceil x \rceil$.

2.1 Partitioning Estimator

A (finite) partition P of space $\mathcal{X}^d$ consists of a finite number of disjoint subsets with their union being the whole space; that is, $P = \left\{P_j \subset \mathcal{X}^d : \bigcup_{j=1}^d P_j = \mathcal{X}^d,\ P_j \cap P_j = \varnothing \text{ for } i \ne j\right\}$. We use $|P|$ to represent the cardinality of P. Denote by $\mathcal{F}_P = \sigma(P)$ the σ-algebra generated by partition P. For any square-integrable function μ, we call the conditional expectation $\mu_P = \mathbb{E}_{\boldsymbol{X}}(\mu|\mathcal{F}_P) = \sum_{P_j \in P} \mathbb{E}_{\boldsymbol{X}}(\mu|P_j) I_{P_j}$ the projection of μ on $\mathcal{F}_P$. It is well known that this projection minimizes the mean squared error (MSE) $\mathbb{E}_{\boldsymbol{X}}[(\mu - f)^2]$ over the class of $\mathcal{F}_P$-measurable and square-integrable functions.

In practice, μ_P is inaccessible to us because the ground truth μ is generally unknown. With observed data $\boldsymbol{Z}$, we can form the partitioning estimator below based on partition P by locally estimating the population means in μ_P using sample observations

$$\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}; \boldsymbol{Z}, P) := \frac{1}{N_{\boldsymbol{x}}} \sum_{i=1}^n Y_i I\{\boldsymbol{X}_i \in P_{\boldsymbol{x}}\}, \tag{2}$$

where $\boldsymbol{x} \in \mathcal{X}^d$ is a target test point, $P_{\boldsymbol{x}}$ stands for the unique region in P that contains $\boldsymbol{x}$, and $N_{\boldsymbol{x}} = \sum_{i=1}^n I\{\boldsymbol{X}_i \in P_{\boldsymbol{x}}\}$ represents the sample size in $P_{\boldsymbol{x}}$. By convention, we define $0/0 = 0$ to avoid ambiguity.

2.2 Ensemble Estimator with Exogenous Randomness

Various methods can be used to form partitions P. A *partitioning rule* is a set of procedural steps, typically implemented algorithmically, that defines how a finite partition P is formed for the feature space $\mathcal{X}^d$. These rules determine partition boundaries based on some decision factors, which can carry inherent randomness. As a result, the corresponding partitions are typically random as well. Since we focus on *data-independent* partitioning rules, the resulting randomness in partitioning is determined only by *exogenous* factors introduced by the user or inherently by the algorithm, independently of the training data.

An arguably most well-known example is the CART partitioning rule for building trees in RF. We provide details on the population CART partitioning rule, which will be the main focus of Section 4. First, we define the *impurity decrement* when splitting a cell (hyperrectangle) $\mathbf{t} = \prod_{j=1}^{d} t_j \subset \mathcal{X}^d$ along direction j at location $c \in t_j \subset \mathcal{X}$ into two daughter cells $\mathbf{t}_{j,c;L} = \mathbf{t} \cap \{X_j \leq c\}$ and $\mathbf{t}_{j,c;R} = \mathbf{t} \cap \{X_j > c\}$ as

$$\Delta_{j,c}(\mathbf{t}) := \mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}) - \mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}_{j,c;L})\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}_{j,c;L}|\mathbf{t}) - \mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}_{j,c;R})\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}_{j,c;R}|\mathbf{t}). \quad (3)$$

For a parent cell $\mathbf{t}$, a subset of features $\Xi \subset \{1, \ldots, d\}$ is randomly selected with $|\Xi| = \lceil \gamma d \rceil$ for a given $\gamma \in (0, 1]$, and the population CART finds the optimal pair (j^*, c^*) as

$$(j^*, c^*) = \underset{j \in \Xi, c \in t_j}{\arg\max}\, \Delta_{j,c}(\mathbf{t}). \quad (4)$$

The parent cell $\mathbf{t}$ is then split into two daughter cells $\mathbf{t}_{j^*,c^*;L}$ and $\mathbf{t}_{j^*,c^*;R}$, each of which becomes new parent cell for the next round of splits. Ties are broken randomly in the optimization problem (4). The entire process starts from the root cell $\mathcal{X}^d$, proceeds recursively, and stops when a pre-given tree depth l is attained. This results in a so-called (population) CART decision tree whose terminal cells (the cells at depth l) form a partition P of space $\mathcal{X}^d$. A partitioning estimator in the form of (2) can be defined, which we will refer to as the CART tree estimator. The RF algorithm in Breiman (2001) forms tree estimators in a similar fashion to the process discussed above, with the difference that the sample version of the CART partitioning rule is used.

Definition 1 (Exogenous randomness in population CART) *There are two types of exogenous randomness in forming a population CART tree estimator: i) (Type I) The randomness introduced by feature subsampling; ii) (Type II) The random tie-breaking in the optimization problem* (4)*.*

We remark that although the above definition is for the population CART partitioning rule, the general concept is broadly applicable to other partitioning rules. In the absence of exogenous randomness, RF consists of one tree and thus becomes a single tree estimator. The main text of our paper will focus *primarily* on Type I exogenous randomness, as tie-breaking happens more rarely in practice and thus the results specific to Type II exogenous randomness are relegated to the Supplementary Material.

To facilitate a rigorous analysis, we formalize the concept of a partitioning rule in mathematical terms. Let $\mathcal{P}$ denote the collection of all finite partitions of $\mathcal{X}^d$. A partitioning rule $\boldsymbol{P} : \mathcal{D} \to \mathcal{P}$ is a mapping from space $\mathcal{D}$ to collection $\mathcal{P}$ of finite partitions, where $\mathcal{D}$ summarizes all potential decision factors to determine the partition boundaries. Let $\Theta : \Omega \to \mathcal{D}$ be

a random element mapping from certain probability space Ω to the decision space $\mathcal{D}$, and is assumed to be independent of the training data. Here, Θ is a $\mathcal{D}$-valued random variable. So $\boldsymbol{P}(\Theta) = \boldsymbol{P} \circ \Theta$, a composition of the deterministic rule $\boldsymbol{P}$ with the random seed Θ, yields a random partition of feature space $\mathcal{X}^d$ driven by exogenous randomness.

Let $\boldsymbol{P}(\Theta_{1:B}) = \{\boldsymbol{P}(\Theta_1), \ldots, \boldsymbol{P}(\Theta_B)\}$ be a set of independent copies of random partition $\boldsymbol{P}(\Theta)$, where $\Theta_{1:B} = \{\Theta_1, \ldots, \Theta_B\}$ is an i.i.d. sample from a population Θ of the exogenous random element. The *ensemble estimator* averages the outputs of B partitioning estimators (2) formed on the *same* input training sample $\boldsymbol{Z}$; that is,

$$\widehat{\mu}_{\text{ens}}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}(\Theta_{1:B})) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}_{\text{part}}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}(\Theta_b)). \tag{5}$$

3. General Results on Partitioning Estimator Ensemble

In this section, we establish the non-asympototic expansion of the MSE for general partitioning estimator ensemble (5) under the setting in Section 2. We will work under the following assumptions.

Assumption 1 (On $\boldsymbol{X}$) *The random vector $\boldsymbol{X}$ can take either continuous or discrete values with bounded joint density function (or mass function) from the above and below.*

Assumption 2 (On μ) *The ground truth function $\mu(\cdot)$ belongs to the class $\mathcal{M}(\mathcal{X}^d, \mathcal{Y})$, the collection of measurable functions mapping from $\mathcal{X}^d$ to $\mathcal{Y}$. Continuity is not required.*

Assumption 3 (On σ^2) *The function $\sigma^2(\cdot)$ is bounded from both above and below.*

Additionally, we introduce the concept of *indistinguishablity*, which plays a fundamentally role in our theoretical development.

Definition 2 (Indistinguishability) *Two subsets A and B of feature space $\mathcal{X}^d$ are indistinguishable under probability $\mathbb{P}_{\boldsymbol{X}}$ if $\mathbb{P}_{\boldsymbol{X}}(A \cap B) = \mathbb{P}_{\boldsymbol{X}}(A) = \mathbb{P}_{\boldsymbol{X}}(B)$. Two partitions P and P' are indistinguishable if for any pair of $P_i \in P$ and $P'_j \in P'$ with $P_i \cap P'_j \neq \varnothing$, sets P_i and P'_j are indistinguishable.*

We treat indistinguishable sets as equivalent (equal), which allows us to simplify the technical derivations. Recall the definition of partitioning rule $\boldsymbol{P}$, space $\mathcal{D}$ of decision factors, and exogenous random element Θ introduced right after Definition 1. We add superscripts to $\boldsymbol{P}$ and $\mathcal{D}$ in the assumption below to emphasize their dependence on n.

Assumption 4 (On $\boldsymbol{P}^n$) *Let $\mathcal{D}_\Theta^n \subset \mathcal{D}^n$ be the subset of decision parameters that can be taken by the exogenous random seed Θ. Accordingly, we define $\mathcal{P}_0^n = \{\boldsymbol{P}^n(D) : D \in \mathcal{D}_\Theta^n\}$, which represents the collection of all possible deterministic partitions that may arise from the random partitioning rule $\boldsymbol{P}^n(\Theta)$. The following regularity conditions hold:*

(1) Each cell in a partition $P \in \mathcal{P}_0^n$ has a positive probability under $\mathbb{P}_{\boldsymbol{X}}$.

(2) For cells $P_1 \in P$ and $P_2 \in P'$ with partitions P and P' two distinct elements in $\mathcal{P}_0^n$, probability $\mathbb{P}_{\boldsymbol{X}}(P_1 \cap P_2) > 0$ if $P_1 \cap P_2 \neq \varnothing$.

(3) Any distinct partitions P and P' in $\mathcal{P}_0^n$ are distinguishable.

Moreover, $\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}}^n(\Theta))$, $\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}_{\boldsymbol{x}}^n(\Theta))$, and $I_{\boldsymbol{P}_{\boldsymbol{x}}^n(\Theta)}$ are bivariate measurable functions on $\mathcal{X}^d \times \Omega$, where $\boldsymbol{P}_{\boldsymbol{x}}^n(\Theta)$ is the cell in $\boldsymbol{P}^n(\Theta)$ that contains $\boldsymbol{x}$.

Let $\widehat{\mu}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}(\Theta))$ be either $\widehat{\mu}_{\text{part}}$ or $\widehat{\mu}_{\text{ens}}$ as in (2) and (5). We use the following global mean squared error (GMSE) to evaluate the performance of $\widehat{\mu}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}(\Theta))$

$$\text{MSE}(\widehat{\mu}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z},\Theta}\left[(\mu(\boldsymbol{X}) - \widehat{\mu}(\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{P}(\Theta))^2\right], \tag{6}$$

where $\boldsymbol{X}$ is an independent test point that is identically distributed as the training feature. A key step in our analysis is the following decomposition of GMSE

$$\text{MSE}(\widehat{\mu}) = \underbrace{\mathbb{E}_{\boldsymbol{X},\Theta}[\widetilde{\mu}(\boldsymbol{X}; \boldsymbol{P}^n(\Theta)) - \mu(\boldsymbol{X})]^2}_{\text{squared bias}} + \underbrace{\mathbb{E}_{\boldsymbol{X},\Theta}[\text{Var}_{\boldsymbol{Z}}(\widehat{\mu}(\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{P}^n(\Theta)))]}_{\text{variance}}, \tag{7}$$

where $\widetilde{\mu}(\boldsymbol{x}; \boldsymbol{P}^n(\Theta)) = \mathbb{E}_{\boldsymbol{Z}}[\widehat{\mu}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}^n(\Theta))]$ with the expectation taken with respect to training sample $\boldsymbol{Z}$.

Decomposition (7) above *differs* from the the existing literature (e.g., Scornet et al. (2015); Scornet (2016); Curth et al. (2024)) in that we first integrate out the effect of $\boldsymbol{Z}$, whereas most existing work first integrates out the effect of exogenous randomness Θ. In our decomposition, $\widetilde{\mu}(\boldsymbol{x}; \boldsymbol{P}^n(\Theta)) - \mu(\boldsymbol{x})$ is precisely the bias of a single partitioning estimator conditional on $\boldsymbol{P}^n(\Theta)$. Similarly, $\text{Var}_{\boldsymbol{Z}}(\widehat{\mu}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}^n(\Theta)))$ is the variance of individual tree estimator conditional on $\boldsymbol{P}^n(\Theta)$. In this sense, (7) averages over individual partitioning estimators' squared bias and variance. Our new way to decompose MSE effectively supports our goal of analyzing the interactions among individual partition estimators and characterizing their effect on the estimation accuracy of RFs.

The facilitate future presentation, we define two kernel functions K_μ and $K_{\sigma^2} : \mathcal{P}_0^n \times \mathcal{P}_0^n \to \mathbb{R}$ as

$$\begin{aligned} K_\mu(P, P') &= \sum_{\substack{P_i \in P, \\ P_j' \in P'}} \mathbb{E}_{\boldsymbol{X}}\left((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))(\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_j')|P_i \cap P_j'\right) \frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P_j')^2}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P_j')}, \\ K_{\sigma^2}(P, P') &= \sum_{\substack{P_i \in P, \\ P_j' \in P'}} \mathbb{E}_{\boldsymbol{X}}\left(\sigma^2|P_i \cap P_j'\right) \frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P_j')^2}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P_j')}, \end{aligned} \tag{8}$$

respectively, where $P = \{P_i\}$ and $P' = \{P_j'\}$ are elements in $\mathcal{P}_0^n$. In particular, when $P = P'$, we denote by

$$\begin{aligned} Q_\mu(P) &= K_\mu(P, P) = \sum_{P_i \in P} \text{Var}_{\boldsymbol{X}}(\mu|P_i), \\ Q_{\sigma^2}(P) &= K_{\sigma^2}(P, P) = \sum_{P_i \in P} \mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_i). \end{aligned} \tag{9}$$

The two kernel functions introduced in (8) above arise naturally in the non-asymptotic error analysis of random partitioning ensembles, capturing the interaction mechanism of

random partitions in achieving the variance reduction. Furthermore, the proposition below shows that they are indeed *reproducing kernels* on certain *Hilbert spaces* and thus satisfy the Cauchy–Schwarz inequality. The proof is provided in Section C.1.

Proposition 3 *Let P and P' be two components in $\mathcal{P}_0^n$, and $(\boldsymbol{X}', \varepsilon')$ an independent copy of $(\boldsymbol{X}, \varepsilon)$. Define the feature mapping $\Phi_\mu : \mathcal{P}_0^n \to L^2(\boldsymbol{X}, \boldsymbol{X}')$ as*

$$\Phi_\mu(P; \boldsymbol{X}, \boldsymbol{X}') = \sum_{P_i \in P} \frac{(\mu(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))}{\mathbb{P}_{\boldsymbol{X}}(P_i)} I_{P_i}(\boldsymbol{X}) I_{P_i}(\boldsymbol{X}'), \tag{10}$$

and another feature mapping $\Phi_\sigma^2 : \mathcal{P}_0^n \to L^2(\boldsymbol{X}, \varepsilon, \boldsymbol{X}')$ as

$$\Phi_{\sigma^2}(P; \boldsymbol{X}, \varepsilon, \boldsymbol{X}') = \sum_{P_i \in P} \frac{\varepsilon}{\mathbb{P}_{\boldsymbol{X}}(P_i)} I_{P_i}(\boldsymbol{X}) I_{P_i}(\boldsymbol{X}'). \tag{11}$$

Then it holds that

$$\begin{aligned} K_\mu(P, P') &= \langle \Phi_{\sigma^2}(P; \boldsymbol{X}, \boldsymbol{X}'), \Phi_{\sigma^2}(P'; \boldsymbol{X}, \boldsymbol{X}') \rangle_{L^2(\boldsymbol{X}, \boldsymbol{X}')}, \\ K_{\sigma^2}(P, P') &= \langle \Phi_{\sigma^2}(P; \boldsymbol{X}, \varepsilon, \boldsymbol{X}'), \Phi_{\sigma^2}(P'; \boldsymbol{X}, \varepsilon, \boldsymbol{X}') \rangle_{L^2(\boldsymbol{X}, \varepsilon, \boldsymbol{X}')}, \end{aligned} \tag{12}$$

where $\langle \cdot, \cdot \rangle_{L^2(\boldsymbol{X}, \boldsymbol{X}')}$ and $\langle \cdot, \cdot \rangle_{L^2(\boldsymbol{X}, \varepsilon, \boldsymbol{X}')}$ represent the standard inner products in the L^2-spaces. Consequently, we have that

$$K_\mu(P, P') \le \sqrt{Q_\mu(P) Q_\mu(P')}, \quad K_{\sigma^2}(P, P') \le \sqrt{Q_{\sigma^2}(P) Q_{\sigma^2}(P')}, \tag{13}$$

where the first equality holds if and only if for any $P_i \in P$, $P_j' \in P'$ with $P_i \cap P_j' \neq \varnothing$, either P_i and P_j' are indistinguishable or μ is constant on $P_i \cup P_j'$, while the second equality holds if and only if P and P' are indistinguishable. Therefore, Assumption 4(3) ensures that the second result in (13) *is a strict inequality, and that when μ is a non-constant function, the first result in* (13) *is a strict inequality.*

Theorem 4 *Assume that*

$$\sup_n \mathbb{E}_\Theta \left[\frac{1}{n} |\boldsymbol{P}^n(\Theta)| \right] < \infty, \tag{14}$$

where $|\boldsymbol{P}^n(\Theta)|$ is the total number of cells in $\boldsymbol{P}^n(\Theta)$. Then the MSE of $\widehat{\mu}_{ens}$ satisfies

$$\begin{aligned} MSE(\widehat{\mu}_{ens}) =& \frac{B-1}{B} \underbrace{\mathbb{E}_{\boldsymbol{X}} \left[(\mu - \mathbb{E}_\Theta(\mu_{\boldsymbol{P}^n(\Theta)}))^2 \right]}_{\text{squared bias of ensemble}} + \frac{1}{B} \underbrace{\mathbb{E}_{\boldsymbol{X}, \Theta} \left[(\mu - \mu_{\boldsymbol{P}^n(\Theta)})^2 \right]}_{\text{squared bias of a single partition}} \\ &+ \frac{B-1}{B} \underbrace{\mathbb{E}_{\Theta, \Theta'} \left\{ \frac{1}{n} K_\mu(\boldsymbol{P}^n(\Theta'), \boldsymbol{P}^n(\Theta')) + \frac{1}{n} K_{\sigma^2}(\boldsymbol{P}^n(\Theta'), \boldsymbol{P}^n(\Theta')) \right\}}_{\text{cross-partition covariance}} \\ &+ \frac{1}{B} \underbrace{\mathbb{E}_\Theta \left\{ \frac{1}{n} Q_\mu(\boldsymbol{P}^n(\Theta)) + \frac{1}{n} Q_{\sigma^2}(\boldsymbol{P}^n(\Theta)) \right\}}_{\text{single-partition variance}} + \mathcal{R}_{ens}(\mu), \end{aligned} \tag{15}$$

where Θ and Θ' are independent copies of the exogenous random elements, and

$$\mathcal{R}_{ens}(\mu) \lesssim (\|\mu\|_\infty + \|\sigma^2\|_\infty + 1)^2 \times \mathbb{E}_\Theta \left(\frac{|\boldsymbol{P}^n(\Theta)|}{n} \frac{1}{\sqrt{1 + \min_{P_i \in \boldsymbol{P}^n(\Theta)} (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)}} \right)$$

with the constant in $\lesssim$ independent of μ, σ^2, $\boldsymbol{P}^n(\Theta)$, and n. The MSE of a single partitioning estimator $\widehat{\mu}_{part}$ corresponds to the special case of $B = 1$ in (15) with

$$\mathcal{R}_{part}(\mu) \lesssim (\|\mu\|_\infty + \|\sigma^2\|_\infty + 1)^2 \times \mathbb{E}_\Theta \left(\frac{|\boldsymbol{P}^n(\Theta)|}{n} \frac{1}{1 + \min_{P_i \in \boldsymbol{P}^n(\Theta)} (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)} \right).$$

Theorem 4 above is built on a local MSE bound for ensemble estimators detailed in Section B, and its proof is provided in Section C.2. Condition (14) indicates that the partitioning rule is allowed to vary as sample size n increases, as we need at least one training data point in forming prediction in each terminal cell. Under Assumption 4(3) and when μ is non-constant, the results in Proposition 3 become strict inequalities. Thus, the leading terms of both squared bias and variance for the ensemble estimator are strictly less than those of a single partitioning estimator, showing the advantage of the ensemble estimator. As $B \to \infty$, the leading terms in MSE decomposition for ensemble estimator are driven by the pairwise interactions among partitions resulted from exogenous randomness, and the effect from single partitioning estimators vanishes completely. These theoretical results support the new insights discussed in the Introduction.

We next present the consistency results and additional conditions under which the remainder terms $\mathcal{R}_{part}$ and $\mathcal{R}_{ens}$ become negligible compared to their corresponding leading order terms. We define, for any partitions $P = \{P_j\}$ and $P' = \{P'_j\}$ with positive cell probabilities, a *model-free* kernel function

$$K(P, P') = \sum_{P_i \in P} \sum_{P'_j \in P'} \mathbb{P}_{\boldsymbol{X}}(P_i | P'_j) \mathbb{P}_{\boldsymbol{X}}(P'_j | P_i). \tag{16}$$

When $P = P'$, it reduces to

$$Q(P) := K(P, P) = \sum_{P_i \in P} 1 = |P|. \tag{17}$$

In addition, we also define a feature mapping

$$\Phi(P; \boldsymbol{X}, \boldsymbol{X'}) = \sum_{P_i \in P} \frac{I_{P_i}(\boldsymbol{X}) I_{P_i}(\boldsymbol{X'})}{\mathbb{P}_{\boldsymbol{X}}(P_i)}. \tag{18}$$

It is clear that $K(P, P') = \langle \Phi(P; \boldsymbol{X}, \boldsymbol{X'}), \Phi(P'; \boldsymbol{X}, \boldsymbol{X'}) \rangle_{L^2(\boldsymbol{X}, \boldsymbol{X'})}$ and thus $0 \le K(P, P') \le \sqrt{Q(P)Q(P')}$, where the last equality holds if and only if P and P' are indistinguishable (see Definition 2). Similar to $K_\mu(P, P')$ and $K_{\sigma^2}(P, P')$, the quantity $K(P, P')$ also characterizes the interaction between partitions P and P' but it is not model specific (i.e., independent of μ and σ^2). Under Assumptions 2 and 3, it serves as a uniform upper bound for both $K_\mu(P, P')$ and $K_{\sigma^2}(P, P')$, and a uniform lower bound for $K_{\sigma^2}(P, P')$, and thus controls the convergence rate of the variance terms in the MSE decomposition.

Furthermore, we can also define the *cross-partition correlation* as

$$\mathrm{Corr}(P, P') = \frac{K(P, P')}{\sqrt{Q(P)Q(P')}} \in [0, 1], \tag{19}$$

which quantifies the correlation between two partitions and plays a key role in understanding the difference between a single partitioning estimator and the corresponding ensemble estimator.

Theorem 5 *Assume that Assumptions 1–4 hold, condition (14) is satisfied for each $n \geq 1$, and with probability one with respect to $\mathbb{P}_\Theta$,*

$$\min_{P_i^n \in \boldsymbol{P}^n(\Theta)} n\mathbb{P}_{\boldsymbol{X}}(P_i^n) \to +\infty. \tag{20}$$

Then for the partitioning estimator $\widehat{\mu}_{part,n}$, the sufficient and necessary condition for the weak consistency of $\widehat{\mu}_{part,n}$ is

$$\mathbb{E}_{\boldsymbol{X},\Theta}\left[(\mu - \mu_{\boldsymbol{P}^n(\Theta)})^2\right] \to 0 \quad \textit{and} \quad \frac{\mathbb{E}_\Theta[|\boldsymbol{P}^n(\Theta)|]}{n} \to 0. \tag{21}$$

Meanwhile, as $B \to \infty$, the ensemble estimator $\widehat{\mu}_{ens,n}$ is weakly consistent if and only if

$$\mathbb{E}_{\boldsymbol{X}}\left[(\mu - \mathbb{E}_\Theta(\mu_{\boldsymbol{P}^n(\Theta)}))^2\right] \to 0 \quad \textit{and} \quad \frac{1}{n}\mathbb{E}_{\Theta,\Theta'}\left[K(\boldsymbol{P}^n(\Theta), \boldsymbol{P}^n(\Theta'))\right] \to 0. \tag{22}$$

In particular, condition (21) implies condition (22).

Under Condition (20), the remainder terms $\mathcal{R}_{part}$ and $\mathcal{R}_{ens}$ in Theorem 4 are negligible compared to $\mathbb{E}[|\boldsymbol{P}^n(\Theta)|/n]$, the order of the leading terms in the MSE decomposition (15).

Remark 6 *It is worth noting that both Theorems 4 and 5 do not require any continuity or smoothness condition on the ground truth μ as in Assumption 2. This is because the analysis is conducted under the global L^2-risk in $L^2(\mathcal{X}^d)$, rather than on local approximation at a specific point. Such a phenomenon is well known in the theory of partitioning estimators; for example, histogram regression (Györfi et al., 2002, Theorem 4.2), Mondrian trees (Mourtada et al., 2020, Theorem 1), and CART random forests (Chi et al., 2022, Theorem 1) achieve the global L^2-consistency without assuming continuity of μ. The continuity or smoothness conditions may become necessary when one seeks local consistency at a given target point as discussed in Györfi et al. (2002)[Chapter 4] and Mourtada et al. (2020).*

The proof of Theorem 5 above is provided in Section C.3. For the single-partition estimator, the first condition in (21) (governing the squared bias) requires that for almost all realizations of $\boldsymbol{P}(\Theta)$, the induced partition is fine enough to capture the structure of the target function μ, while the second condition (controlling the variance) requires that each cell contains sufficiently many samples. For the ensemble estimator, the first condition in (21) relaxes that in (22): an individual partition may be coarse, but their aggregation over Θ (i.e., $\mathbb{E}_\Theta(\mu_{\boldsymbol{P}^n(\Theta)})$) should provide rich enough structure to approximate μ well; and the second condition in (21) for controlling variance requires that partitions be diverse and weakly correlated, a core principle of ensemble methods.Since condition (21) is stronger than condition (22), we obtain immediately that the ensemble estimator may be consistent under *weaker* conditions than its individual partitioning estimator.

4. The CART Random Forests

In this section, we deviate from our general model in (1) and consider the following sparse linear model

$$y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_s X_s + \varepsilon, \tag{23}$$

where $\beta_1, \ldots, \beta_s \in \mathbb{R}\backslash\{0\}$ are nonzero regression coefficients, $X_1, \ldots, X_d$ are i.i.d. random variables, positive integer $s \leq d$ is the sparsity parameter, and the model error ε is independent of covariates with zero mean and a constant variance σ_0^2. Variables $X_1, \ldots, X_s$ are called *informative variables* as they contribute to response y; when $d > s$, the remaining $d - s$ ones are *non-informative* (or noise) variables since they are independent of y. We set $\beta_j = 0$ for $j = s+1, \ldots, d$. We focus on the simpler model here because it allows us to derive mathematically *more explicit* results and thus gain more insights.

In Section 4.1, we will theoretically analyze the binary and continuous feature scenarios. We consider both types of feature distributions because the CART partitioning rule can be significantly simplified in the binary case and is more complicated in the continuous case. Our theoretical results show an intriguing fact that the MSE expansions of tree and forest estimators are strikingly similar for both binary and continuous features, an evidence supporting that the insights gained in this section may be broadly applicable. In Section 4.2, we use theory-guided simulations to gain further insights.

4.1 Theoretical Analysis

4.1.1 Model with Binary Features

In this section, we assume that $X_1, \ldots, X_d$ are i.i.d. Bernoulli random variables $\mathcal{B}(1, 0.5)$, and that the tree depth $l < \lceil \gamma d \rceil$. The splitting rule is now uniquely determined once the direction for split is selected because any splitting position along that direction results in the same outcome. Hence, for any cell $\mathbf{t} = \prod_{j=1}^{d} t_j \subset \{0,1\}^d$, the impurity decrement (3) is independent of split location and takes the explicit form

$$\Delta_j(\mathbf{t}) = \begin{cases} \beta_j^2/4 & \text{if } j \in \{1, \ldots, s\} \text{ and } t_j = \{0,1\}; \\ 0 & \text{if } j \in \{s+1, \ldots, d\} \text{ or } t_j = \{0\}, \{1\}. \end{cases}$$

Thus, in view of the description around (3), for each parent node $\mathbf{t}$ and a randomly subsampled feature set $\Xi \subset \{1, \ldots, d\}$ with size $\lceil \gamma d \rceil$, if there are indices in $\{1, \ldots, s\} \cap \Xi$ that have not been split yet, we *randomly* choose one with the *maximal* β_j^2 as the optimal j^*; and if all indices in $\{1, \ldots, s\} \cap \Xi$ have been split, then we *randomly* pick one from Ξ that have not been split and set it as j^*. The resulting daughter cells are then treated as new parent cells and split independently according to the same rule. The splitting process starts from root cell $\mathbf{t}_0 = \{0,1\}^d$ and stops after *tree depth* reaches l (which may be tuned according to n to ensure that each terminal cell has at least one data point with high probability for forming meaningful prediction). Throughout the paper, we choose the depth $l < \lceil \gamma d \rceil$ so that there are always indices that have not been split yet in the feature subsample Ξ, and consequently, we will have exactly 2^l terminal cells.

It is seen that each terminal cell $\mathbf{t}$ at depth l can be uniquely determined by the states of whether each coordinate is split or not. We thus define a *stochastic process*, named the *binary CART process*, to describe the states of coordinates in the tree-growing process.

Algorithm 1 Binary CART process

Initialize $I_0 = (0, \ldots, 0)$ of length d.
for $k = 1$ to l **do**
    Randomly select a subset Ξ_k of size $\lceil \gamma d \rceil$ from $\{1, \ldots, d\}$.
    Filter Ξ_k to obtain unsplit features $\Xi_{0k} = \{j : I_{k-1,j} = 0\}$.
    if $\Xi_{0k} \neq \varnothing$ **then**
        Choose randomly an index j^* such that $j^* = \arg\max_{j \in \Xi_{0k}} \beta_j^2$.
    end if
    Update I_k by setting $I_{k,j^*} = 1$ for j^* and $I_{k,j} = I_{k-1,j}$ for the rest, if $\Xi_{0k} \neq \varnothing$; otherwise, set $I_k = I_{k-1}$.
end for
return I_l

Definition 7 (Binary CART process) *For a terminal cell* $\mathbf{t}$ *in a depth-*l *tree, the binary CART process* $I_k = (I_{k1}, \ldots, I_{kd})$ *records the states of coordinates by tree depth* k *for* $k = 0, 1, \ldots, l$*, where* $I_{kj} = 1$ *indicates that* X_j *has been split by depth* k*, and* $I_{kj} = 0$ *otherwise.*

Algorithm 1 outlines the binary CART process as tree depth increases. For a point $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{0d})^\top$ and a tree with depth l, let $\mathbf{t}_{\boldsymbol{x}_0}$ be the terminal cell containing $\boldsymbol{x}_0$. Then we have $\mathbf{t}_{\boldsymbol{x}_0} = \prod_{j:I_{lj}=1}\{x_{0j}\} \times \prod_{j:I_{lj}=0}\{0,1\}$, which is completely characterized by I_l.

Using Definition 1, it is seen that Type II exogenous randomness can occur when multiple informative features have identical β_j^2's, or when a random non-informative feature is chosen as j^*. Type I exogenous randomness is controlled by parameter γ, and Type II exogenous randomness in this example is controlled by multiple parameters s, β_j's, d, and γ. It will be made clear that both types can contribute to the success of random forests.

Observe that given I_k, the distribution of I_{k+1} depends only on the exogenous randomness discussed above. Further, given model (23), the binary CART process I_l is *Markovian* and parameterized by feature subsampling rate γ. It is also important to note that I_l and I_l' from two separate depth l trees are independent and identically distributed.

Denote by $\widehat{\mu}_{\text{tree}}$ and $\widehat{\mu}_{\text{RF}}$ the tree and forest estimates formed with the CART partitioning rule described above, respectively. We present below the explicit expansion for the MSE in (6) of these two estimators using the binary CART process.

Theorem 8 *Let I_l and I_l' be two independent binary CART processes. It holds that*

$$\begin{aligned} MSE(\widehat{\mu}_{RF}) = \frac{B-1}{B}\underbrace{\mathbb{E}\left[\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-\max\{I_{lj},I_{lj}'\})\right]}_{\text{squared bias from tree ensemble}} + \frac{1}{B}\underbrace{\mathbb{E}\left[\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-I_{lj})\right]}_{\text{single-tree squared bias}} \\ + \frac{B-1}{B}\underbrace{\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-\max\{I_{lj},I_{lj}'\})\right)\frac{2^{\sum_{j=1}^{d}\min\{I_{lj},I_{lj}'\}}}{n}\right]}_{\text{cross-tree covariance}} \\ + \frac{1}{B}\underbrace{\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-I_{lj})\right)\frac{2^{\sum_{j=1}^{d}I_{lj}}}{n}\right]}_{\text{single-tree variance}} + \mathcal{R}_{RF}, \end{aligned} \tag{24}$$

where the remainder satisfies $\mathcal{R}_{RF} \lesssim \frac{2^l}{n(1+(n-1)2^{-l})^{1/2}} + \left(1-2^{-l}\right)^n$. In particular, when $B=1$, the global MSE of a single tree estimator is

$$MSE(\widehat{\mu}_{tree}) = \underbrace{\frac{1}{4}\mathbb{E}\left[\sum_{j=1}^{s}\beta_j^2(1-I_{lj})\right]}_{\text{single-tree squared bias}} + \underbrace{\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-I_{lj})\right)\frac{2^{\sum_{j=1}^{d}I_{lj}}}{n}\right]}_{\text{single-tree variance}} + \mathcal{R}_{tree}, \tag{25}$$

in which the remainder is bounded by $\mathcal{R}_{tree} \lesssim \frac{2^l}{n(1+(n-1)2^{-l})} + \left(1-2^{-l}\right)^n$. Furthermore, as $l = l_n \to \infty$ with $2^l/n \to 0$, we have the convergence rate result

$$\max\{MSE(\widehat{\mu}_{RF}), MSE(\widehat{\mu}_{tree})\} \lesssim s\left(1-\frac{3}{4}\gamma W_{\gamma,d}(s)\right)^{l+1} + \frac{2^l}{n}, \tag{26}$$

where the constant in $\lesssim$ depends only on β_j^2's and σ_0^2, and the positive function with $x \le d$

$$W_{\gamma,d}(x) := \left(1-\frac{x}{d}\right)\cdots\left(1-\frac{x}{d-\lceil\gamma d\rceil+1}\right) \tag{27}$$

depends only on feature subsampling rate $\gamma \in (0,1]$ and dimensionality d.

The results in Theorem 8 above are built upon the general Theorem 4, and its proof is provided in Section C.4. Some interesting insights can be obtained by examining the expressions in (24) and (25).

Remark 9 *For simplicity, we assume that B is large enough so that the terms involving $1/B$ are all negligible, and that β_j^2's all have similar magnitude $\beta_0^2 > 0$. Comparing (24) and (25), it is seen that the leading contributions to MSE for trees and forests are completely different. For $\widehat{\mu}_{tree}$, the variance and squared bias are driven by a single Markov process*

I_l. For $\widehat{\mu}_{RF}$, owing to ensemble, MSE is driven by the interaction effects between two independent processes I_l and I'_l, and the effects from individual tree estimates are negligible.

Additionally, ensemble reduces both the squared bias and variance. For the squared bias terms, it is seen from both (24) and (25) that they depend only on the s informative variables via the weighted sums of their β_j^2's. For a single tree, we have

$$\underbrace{\sum_{j=1}^{s} \beta_j^2 (1 - I_{lj})}_{\text{squared bias by a single tree}} \asymp \beta_0^2 \sum_{j=1}^{s} (1 - I_{lj}),$$

in which the right-hand side (RHS) is proportional to the number of unsplit informative variables by a single binary CART process I_l. For forest, we have

$$\underbrace{\sum_{j=1}^{s} \beta_j^2 (1 - \max\{I_{lj}, I'_{lj}\})}_{\text{squared bias by tree ensemble}} \asymp \beta_0^2 \sum_{j=1}^{s} (1 - \max\{I_{lj}, I'_{lj}\}),$$

where the RHS is proportional to the number of unsplit features shared by I_l and I'_l. By comparing the two results above, the squared bias for $\widehat{\mu}_{RF}$ is always smaller than that for $\widehat{\mu}_{tree}$, showing that ensemble reduces the squared bias.

For the variance terms, the order of the tree variance is determined by $2^{\sum_{j=1}^{d} I_{lj}}/n$, which is driven by the number of splits in I_l. The order of the forest variance is determined by $2^{\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}}/n$, which is driven by the number of shared splits between I_l and I'_l. Noting that $2^{\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}} \leq 2^{\sum_{j=1}^{d} I_{lj}}$, the forest estimator admits a smaller variance than the tree estimator. Intuitively, the exogenous randomness makes two individual trees likely split along distinct coordinators (that is, it is likely that $\min\{I_{lj}, I'_{lj}\} < I_{lj}$), and hence reduces the variance of forests.

Remark 10 *Theorem 8 provides a new perspective on how exogenous randomness helps reduce the pairwise correlation among trees. To understand this, note that using the definitions in (16) and (17), the expectation of kernel function K of two independent partitions P and P' generated by two separate CART trees can be written as $\mathbb{E}[K(P, P')] = \mathbb{E}\left[2^{\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}}\right]$, and the expectation of function Q for a single tree P is $\mathbb{E}[Q(P)] = \mathbb{E}\left[2^{\sum_{j=1}^{d} I_{lj}}\right]$; the derivations can be found at the end of Section C.4. When $l < \lceil \gamma d \rceil$, we have $Q(P) \equiv 2^l$. Thus, in this case, the expected cross-tree correlation function becomes*

$$\mathbb{E}[Corr(P, P')] = \mathbb{E}\left[2^{\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\} - l}\right]. \tag{28}$$

Recall that $\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}$ is the number of shared splits by level l and thus is always no larger than the total splits l. The correlation function takes values between 0 and 1, with a smaller number of shared splits corresponding to a smaller correlation. In this sense, the exogenous randomness helps increase the variability of splits across different trees and thus reduces $\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}$, which consequently reduces the variance of $\widehat{\mu}_{RF}$.

Theorem 8 provides a common upper bound (26) for the convergence rates for both $\widehat{\mu}_{tree}$ and $\widehat{\mu}_{RF}$; the bound is conservative for both estimators, more so for the latter. We derive the following corollary for the special case of $s = 1$ to demonstrate that random forests can achieve faster convergence rates than a single tree.

Corollary 11 *Assume that $l \leq d$ and $s = 1$ with $\beta_1 \neq 0$. Let I_l and I_l' be independent binary CART processes. Then it holds that*

$$\begin{aligned} MSE(\widehat{\mu}_{RF}) \lesssim & \underbrace{\frac{1}{4B}\beta_1^2(1-\gamma)^l + \frac{B-1}{4B}\beta_1^2(1-\gamma)^{2l}}_{\text{squared bias}} \\ & + \underbrace{\frac{1}{B}\sigma_0^2\frac{2^l}{n} + \frac{B-1}{B}\sigma_0^2\frac{\rho_{l,d}(\gamma)}{n}}_{\text{variance}} + \mathcal{R}'_{RF}, \end{aligned} \tag{29}$$

where $\mathcal{R}'_{RF} \lesssim (1-\gamma)^l\frac{2^l}{n} + (\frac{2^l}{n})^{3/2} + (1-2^{-l})^n$. Here, the rate $\rho_{l,d}(\gamma)$ is a quadratic form of $P_1 = 1-(1-\gamma)^l$

$$\rho_{l,d}(\gamma) = A_{l,d}P_1^2 + 2B_{l,d}P_1(1-P_1) + C_{l,d}(1-P_1)^2, \tag{30}$$

in which the coefficients

$$\begin{aligned} A_{l,d} &= 2\sum_{r=0}^{l-1}\frac{(l-1)_r^2}{(d-1)_r\, r!}, \qquad B_{l,d} = \sum_{r=0}^{l-1}\left(\frac{l}{l-r}\right)\frac{(l-1)_r^2}{(d-1)_r\, r!}, \\ C_{l,d} &= \sum_{r=0}^{l-1}\left(\frac{l}{l-r}\right)^2\frac{(l-1)_r^2}{(d-1)_r\, r!} \end{aligned} \tag{31}$$

depend only on d, l, not on γ, with the falling factorial $(x)_t = x(x-1)\cdots(x-t+1)$. Moreover, as $d, l \to \infty$ and $l/d \to 0$ (the high-dimensional shallow-tree regime), for any fixed $\gamma \in (0,1)$, we have $\log_2 \rho_{l,d}(\gamma) \lesssim \frac{l^2}{d} \ll l$. Thus, random forests achieves faster convergence rates both in squared bias and the variance than a single tree.

The detailed proof of Corollary 11 is given in Appendix C.5.

Remark 12 (Blessing of dimensionality) *Note that the squared bias rate and the single-tree variance are both independent of dimensionality d. As for the tree-ensemble variance, we recall that coefficients $A_{l,d}$, $B_{l,d}$, and $C_{l,d}$ in (31) are all strictly decreasing in d. Hence, the total variance part decreases in d and thus higher dimensionality improves the performance of random forests by reducing the variance.*

Remark 13 (Optimal $\gamma^\star \in (0,1)$) *Recall that the squared bias part decreases in γ. Meanwhile, we observe that*

$$\frac{d}{dP_1}\rho_{l,d}(\gamma) = 2(B_{l,d} - C_{l,d}) + 2(A_{l,d} - 2B_{l,d} + C_{l,d})P_1.$$

Note that $C_{l,d} > B_{l,d}$, $A_{l,d} > B_{l,d}$ and $A_{l,d} - 2B_{l,d} + C_{l,d} > 0$. Thus, the derivative above is strictly increasing in P_1 and changes sign exactly once at

$$P_1^\star = \frac{C_{l,d} - B_{l,d}}{A_{l,d} - 2B_{l,d} + C_{l,d}} \in (0,1).$$

Since $P_1 = 1-(1-\gamma)^l$ is strictly increasing in γ, it follows that $\rho_{l,d}(\gamma)$ is decreasing *in γ for $\gamma < \gamma^\star$ and* increasing *in γ for $\gamma > \gamma^\star$, where $\gamma^\star = 1-(1-P_1^\star)^{1/l} \in (0,1)$. This together with the fact that the squared bias decreases in γ shows that when the noise variance σ_0^2 is sufficiently large compared to the signal strength β_1^2, the total MSE exhibits a U-shaped curve in γ, implying that the random forests estimator with the optimal $\gamma^\star \in (0,1)$ outperforms a single tree estimator without feature subsampling (i.e., $\gamma = 1$).*

4.1.2 Model with Continuous Features

We now consider the case where $X_1, \cdots, X_d$ are i.i.d. uniform random variables $\mathcal{U}(0,1)$. Similar to the last subsection, we start by describing the population CART partitioning rule applied to this specific setting. For a cell $\mathbf{t} = \prod_{j=1}^d t_j$ with $t_j = (a_j, b_j) \subset [0,1]$, the conditional mean of y on $\mathbf{t}$ is given by $\mathbb{E}[y|\boldsymbol{X} \in \boldsymbol{t}] = \sum_{j=1}^s \beta_j \frac{b_j+a_j}{2}$. Then the impurity decrement in (3) by splitting along X_j is specified as

$$\Delta_j(\boldsymbol{t}) = \begin{cases} \beta_j^2 (b_j - a_j)^2/12 & \text{if } j \in \{1,\ldots,s\}; \\ 0 & \text{if } j \in \{s+1,\ldots,d\}. \end{cases} \tag{32}$$

We note that for any informative feature $j \in \{1,\ldots,s\}$, the optimal splitting position c_j^* along the jth coordinate is the midpoint $\frac{b_j+a_j}{2}$ of t_j. However, in the case of non-informative variables (for $j \in \{s+1,\ldots,d\}$), splitting at any position within t_j yields the same zero impurity decrement; we choose to split at the midpoint for simplicity. This simplifies the tie-breaking process by limiting randomness to split direction only.

When building trees, for each parent node $\mathbf{t}$ and a randomly sampled feature subset $\Xi \subset \{1,\ldots,d\}$ with size $\lceil \gamma d \rceil$, if $\Xi \cap \{1,\ldots,s\} \neq \varnothing$, we choose $j^* \in \Xi$ as the one with the maximal impurity decrement (ties are broken *randomly*); and if $\Xi \cap \{1,\ldots,s\} = \varnothing$, we randomly pick one from Ξ as j^*. We note that, *unlike* the binary case, an index $j \in \{1,\cdots,s\}$ in the uniform case can be split arbitrarily many times, with the impurity decrement always being positive. This difference also makes the continuous feature case more difficult to analyze.

We now define the *uniform CART process* and the algorithm for generating it.

Definition 14 (Uniform CART process) *For a terminal cell $\mathbf{t}$ in a depth l tree, the uniform CART process is a vector process $J_k = (J_{k1},\ldots,J_{kd})$ with J_{kj} being the number of splits along coordinate X_j by tree depth k, where $j \in \{1,\ldots,d\}$ and $k \in \{1,\cdots,l\}$.*

Similar to the binary case, the uniform CART process is also *Markovian* and parameterized by γ given model (23), since the conditional distribution of J_{k+1} given J_k is determined only by the exogenous randomness.

Each terminal cell $\mathbf{t}$ can be uniquely determined by its uniform CART process J_l. To understand this, note that given $x_0 \in (0,1)$, if interval $(0,1)$ is split equally into m parts, the

Algorithm 2 Uniform CART process

1: Initialize $J_0 = (0, \ldots, 0)$ of length d.
2: **for** $k = 1$ to l **do**
3: Randomly select a subset Ξ_k of size $\lceil \gamma d \rceil$ from $\{1, \ldots, d\}$.
4: Randomly select j^* from $\Xi_{1k} = \arg\max_{j \in \Xi_{0k}} \beta_j^2 2^{-2J_{k-1,j}}$.
5: Update J_k by setting $J_{k,j^*} = J_{k-1,j^*} + 1$ and $J_{k,j} = J_{k-1,j}$ for the remaining j's.
6: **end for**
7: **return** J_l.

unique interval containing x_0 is $t(x_0, m) = \left(\frac{K(x_0,m)-1}{2^m}, \frac{K(x_0,m)}{2^m} \right]$ with $K(x_0, m) = \lceil x_0 2^m \rceil$. Thus, for a given $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{0d})^\top$, the terminal cell $\boldsymbol{t}_{\boldsymbol{x}_0}$ can be expressed as

$$\boldsymbol{t}(\boldsymbol{x}_0, J_l) = \prod_{j=1}^{d} t(x_{0j}, J_{lj}), \tag{33}$$

in which J_l is independent of $\boldsymbol{x}_0$. Same as in the binary case, J_l and J_l' from two separate trees grown by the uniform CART partitioning rule are i.i.d. random vectors.

Theorem 15 *Let J_l and J_l' be two independent uniform CART processes. The global MSE of the related random forests estimator $\widehat{\mu}_{RF}$ is*

$$\begin{aligned} MSE(\widehat{\mu}_{RF}) = {} & \frac{B-1}{B} \underbrace{\mathbb{E}\left[\frac{1}{12} \sum_{i=1}^{s} \beta_i^2 2^{-2\max\{J_{li}, J_{li}'\}} \right]}_{\text{squared bias by tree ensemble}} + \frac{1}{B} \underbrace{\mathbb{E}\left[\sum_{i=1}^{s} \frac{1}{12} \beta_i^2 2^{-2J_{li}} \right]}_{\text{squared bias by a single tree}} \\ & + \frac{B-1}{B} \underbrace{\mathbb{E}\left[\left(\sigma_0^2 + \frac{1}{12} \sum_{i=1}^{s} \beta_i^2 2^{-2\max\{J_{li}, J_{li}'\}} \right) \frac{2^{\sum_{i=1}^{d} \min\{J_{li}, J_{li}'\}}}{n} \right]}_{\text{cross-tree covariance}} \\ & + \frac{1}{B} \underbrace{\mathbb{E}\left[\left(\sigma_0^2 + \frac{1}{12} \sum_{i=1}^{s} \beta_i^2 2^{-2J_{li}} \right) \frac{2^l}{n} \right]}_{\text{single-tree variance}} + \mathcal{R}_{RF}. \end{aligned} \tag{34}$$

In particular, when $B = 1$, the global MSE for a single tree estimator is

$$MSE(\widehat{\mu}_{tree}) = \underbrace{\mathbb{E}\left[\frac{1}{12} \sum_{i=1}^{s} \beta_i^2 2^{-2J_{li}} \right]}_{\text{squared bias by a single tree}} + \underbrace{\mathbb{E}\left[\left(\sigma_0^2 + \frac{1}{12} \sum_{i=1}^{s} \beta_i^2 2^{-2J_{li}} \right) \frac{2^l}{n} \right]}_{\text{single-tree variance}} + \mathcal{R}_{tree}. \tag{35}$$

Here, the remainders $\mathcal{R}_{RF}$ and $\mathcal{R}_{tree}$ are the same as in Theorem 8. Additionally, as $l = l_n \to \infty$ with $2^l/n \to 0$, $\max\{MSE(\widehat{\mu}_{RF}), MSE(\widehat{\mu}_{tree})\}$ has the same upper bound as in (26).

The proof of Theorem 15 above is also based on Theorem 4 and is detailed in Section C.6. It is seen that despite the distinct feature distributions in this example, the MSE

expansions exhibit surprisingly similar forms, and all insights presented in the last section remain true. We provide three remarks emphasizing the differences from the binary case.

Remark 16 *We make the same simplification assumption that β_j^2's all have similar magnitude β_0^2, and that B is large enough so that all terms involving $1/B$ are small enough. Theorem 15 also shows that ensemble reduces both squared bias and variance, and that the leading contributions to MSE for trees and forests are completely different.*

For the squared bias term, both $\widehat{\mu}_{tree}$ and $\widehat{\mu}_{RF}$ depend only on the informative features. However, the dependence mechanism here is different from the binary case. To make this clear, we define the diagonal signal length along the direction of informative features for a cell $\mathbf{t} = \prod_{i=1}^d t_i$ as $|\mathbf{t}|_{2,s} = \left(\sum_{i=1}^s |t_i|^2\right)^{1/2}$. For a target point $\boldsymbol{x}_0$, recall the terminal cell $\mathbf{t}(\boldsymbol{x}_0, J_l)$ in (33)*. Then we have*

$$\underbrace{\sum_{i=1}^s \beta_i^2 2^{-2J_{li}}}_{\text{squared bias by a single tree}} \asymp \beta_0^2 \sum_{i=1}^s 2^{-2J_{li}} = \beta_0^2 |\mathbf{t}(\boldsymbol{x}_0, J_l)|_{2,s}^2,$$

$$\underbrace{\sum_{i=1}^s \beta_i^2 2^{-2\max\{J_{li}, J'_{li}\}}}_{\text{squared bias by forest}} \asymp \beta_0^2 \sum_{i=1}^s 2^{-2\max\{J_{li}, J'_{li}\}} = \beta_0^2 |\mathbf{t}(\boldsymbol{x}_0, J_l) \cap \mathbf{t}(\boldsymbol{x}_0, J'_l)|_{2,s}^2.$$

Since the squared bias of $\widehat{\mu}_{RF}$ is associated with the intersection of two terminal nodes, it is naturally smaller than that of $\widehat{\mu}_{tree}$. The comparison and interpretation of variance terms are similar to the binary case and omitted here.

Remark 17 *Comparing the above upper bound to* (26)*, we see a surprising similarity between the results despite the different feature distributions. These upper bounds have an explicit dependence on γ. However, since these bounds can be conservative and the MSE expansions in* (24)*,* (25)*,* (34)*, and* (35) *are more accurate in characterizing the effect of γ, we will use simulation studies to simulate the Markov processes involved in these upper bounds to gain insights on how γ affects the MSE.*

Remark 18 *Using (16) and (17), the expectation of kernel function K of partitions P and P' generated by two independent uniform CART processes J_{lj} and J'_{lj} can be calculated as $\mathbb{E}[K(P, P')] = \mathbb{E}\left[2^{\sum_{j=1}^d \min\{J_{li}, J'_{li}\}}\right]$, and the expectation of function Q for one single tree can be calculated as $\mathbb{E}[Q(P)] = \mathbb{E}\left[2^{\sum_{i=1}^d J_{li}}\right] \equiv 2^l$; see the end of Section C.6 for detailed derivations. Thus, the expected* cross-tree correlation function *is*

$$\mathbb{E}[Corr(P, P')] = \mathbb{E}\left[2^{\sum_{i=1}^d \min\{J_{li}, J'_{li}\} - l}\right]. \tag{36}$$

Similar to the binary case, the correlation function is strictly less than 1 unless J_l and J'_l share all splits (that is, $\sum_{i=1}^d \min\{J_{li}, J'_{li}\} = l$). Thus, the exogenous randomness helps increase the variability of J_l and J'_l and consequently reduces the forest estimator variance.

Analogous to the binary-feature case, we next investigate the special case when $s = 1$, where the explicit convergence rates for a single tree and forest estimator can be derived.

Corollary 19 *Assume that $l \leq d$ and $s = 1$ with $\beta_1 \neq 0$ in model (23). Let J_l and J_l' be independent uniform CART processes. Then it holds that*

$$\begin{aligned} MSE(\widehat{\mu}_{RF}) \lesssim & \underbrace{\frac{B-1}{12B}\beta_1^2\left(1-\frac{\gamma}{2}\right)^{2l} F_l\left(\frac{\gamma}{2-\gamma}\right) + \frac{1}{12B}\beta_1^2\left(1-\frac{3\gamma}{4}\right)^{l}}_{squared\ bias} \\ & + \underbrace{\frac{B-1}{B}\sigma_0^2\frac{2^l}{n}G_{l,d}(\gamma) + \frac{1}{B}\sigma_0^2\frac{2^l}{n}}_{variance} + \mathcal{R}_{RF}, \end{aligned} \tag{37}$$

where $\mathcal{R}_{RF} \lesssim (1-3\gamma/4)^l\frac{2^l}{n} + (1-2^{-l})^n$ and functions F_l and $G_{l,d}$ are defined as follows. For $r \in (0,1)$, denote the Poisson kernel as

$$P_r(\theta) = \frac{1-r^2}{1-2r\cos\theta + r^2}, \quad \theta \in [-\pi, \pi],$$

and the related probability density function as $\mu_r(\theta) = P_r(\theta)/(2\pi)$. Let ψ be a random variable having a density $\mu_{1/2}(\theta')$, and $\varphi_1, \ldots, \varphi_d$ an i.i.d. sequence of random variables sharing a common density function $\mu_{1/\sqrt{2}}(\theta)$. Then we have that for any $q \in [0,1]$,

$$F_l(q) = \mathbb{E}_\psi\left[\left|(1-q) + qe^{\mathrm{i}\psi}\right|^{2l}\right], \tag{38}$$

$$G_{l,d}(q) = \mathbb{E}_{\varphi_1,\ldots,\varphi_d}\left[\left|qe^{\mathrm{i}\varphi_1} + \frac{1-q}{d-1}\sum_{j=2}^{d} e^{\mathrm{i}\varphi_j}\right|^{2l}\right], \tag{39}$$

where i *is the imaginary unit. Further, for any fixed d and $q \in (0,1)$, it holds that $F_l(q) = O(l^{-\frac{1}{2}})$ and $G_{l,d}(q) = O(l^{-\frac{d-1}{2}})$. Therefore, the tree ensemble indeed leads to faster convergence rates both in squared bias and the variance than a single tree.*

The proof of Corollary 19 is provided in Appendix C.7.

Remark 20 (Blessing of dimensionality) *In the MSE bound (37), both the squared bias part and the single-tree variance part are independent of dimensionality d. Meanwhile, we show that for any fixed l and $\gamma \in (0,1)$, the rate $G_{l,d}(\gamma)$ is a decreasing function of d; see Appendix C.7 for details. Hence, as dimensionality d grows, the variance part caused by the tree ensemble reduces and thus optimizes the performance of random forests.*

Remark 21 (Optimal $\gamma^\star \in (0,1)$) *We show in Appendix C.7 that for a sufficiently large depth l: 1) the squared bias part in (37) decreases in feature subsampling rate γ, and 2) $G_{l,d}(\gamma)$ first decreases over interval $\gamma \in [0, 1/d]$ and then increases over $\gamma \in [1/d, 1]$, forming a U-shaped curve in γ. With these observations, we see that when the noise variance σ_0^2 are comparable to the signal strength β_1^2, the MSE (37) will also exhibit a U-shaped curve as a function of γ, demonstrating that the random forests with the optimal ratio $\gamma^\star \in (0,1)$ outperforms the single tree without feature subsampling (i.e., $\gamma = 1$).*

4.2 Additional Insights via Theory-Guided Simulations

This section provides justifications on and some additional insights into the success of RF over individual trees via simulation studies. Our simulation examples are designed and guided by the theoretical results established in the last subsection. We numerically verify the phenomena described in Theorems 8 and 15, which state that RFs can reduce both squared bias and variance. In addition, the following key insights about the success of RF are discussed, accompanied by simulations: 1) Type II exogenous randomness from random tie-breaking helps RF outperform single trees even when $\gamma = 1$, and 2) exogenous randomness improves the performance of RF over tree both when $\gamma = 1$ and $\gamma < 1$, but the working mechanisms are different in these two cases.

We generate data from model (23) and set dimensionality $d = 100$, $s = 5$, sample size $n = 1000$, the number of trees $B = 100$, and noise variance $\sigma_0^2 = 1.69$. The true regression coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_5)^T$ has the following two different configurations:

(I) *Equal configuration:* $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.5$.

(II) *Unequal configuration:* $\beta_1 = 2$, $\beta_2 = 1.8$, $\beta_3 = 1.6$, $\beta_4 = 1.4$, $\beta_5 = 1.2$.

We vary feature subsampling rate γ from 0.1 to 1, and the tree depth from $l = 1$ to 9.

Motivated by our theoretical results, we record the values of the following performance measures across 1000 Monte Carlo simulations for comparing RF to single trees:

(a) *Squared bias:* Squared bias terms by tree ensemble and a single tree in (24) for the binary case, and those in (34) for the continuous case.

(b) *Number of unsplit signals in the binary case:* $s - \sum_{j=1}^{s} I_{lj}$ for a single tree, and $s - \sum_{j=1}^{s} \max\{I_{lj}, I'_{lj}\}$ for a random forest, as discussed in Remark 9.

Squared diagonal signal length in the continuous case: $\sum_{i=1}^{s} 2^{-2J_{li}}$ for a single tree, and $\sum_{i=1}^{s} 2^{-2\max\{J_{li}, J'_{li}\}}$ for a random forest, as discussed in Remark 16.

(c) *Variance:* Single-tree variance and cross-tree covariance in (24) for the binary case, and those in (34) for the continuous case.

(d) *Cross-tree correlation:* Expected cross-tree correlation (28) for the binary case, and (36) for the continuous case. The cross-tree correlation for tree is defined as 1.

(e) *Number of shared splits:* For a single tree, $\sum_{j=1}^{d} I_{lj}$ for the binary case and $\sum_{i=1}^{d} J_{li}$ for the continuous case; while for a forest, $\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}$ for the binary case and $\sum_{i=1}^{d} \min\{J_{li}, J'_{li}\}$ for the continuous case.

(f) *Global MSE bound:* The leading terms of the global MSE bounds in (24) and (34), respectively, for the binary and continuous cases, with the remainder terms ignored.

4.2.1 Random Forests can Reduce Both Squared Bias and Variance

This section aims at verifying Remarks 9 and 16 via simulation using Equal Coefficient Configuration (I). We present the results for performance measures (a)–(f) in Figures 1 and 2 for the binary and the continuous cases, respectively, with respect to varying γ.

From Figures 1(a) and (c), we observe that the squared bias and variance curves for RFs are consistently lower than those for tree, leading to a substantial improvement in MSE as shown in Figure 1(f). The reduction in squared bias is because the forest splits more signal features, as shown in Figure 1(b). The reduction in variance is due to fewer shared splits between I_l and I'_l, as illustrated in Figure 1(e). These results are consistent with

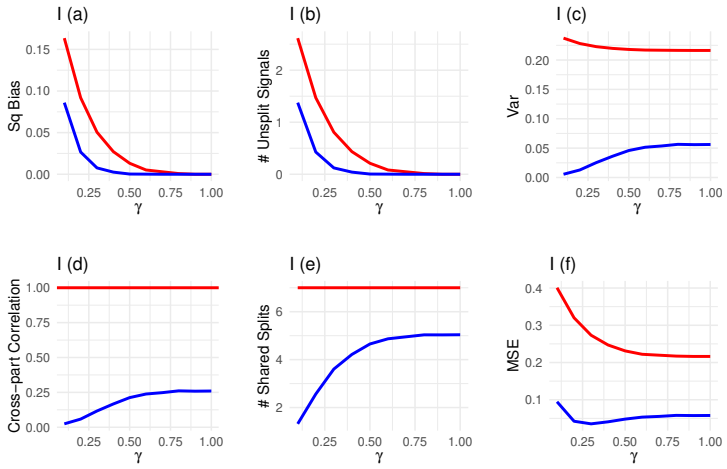

Figure 1: Performance measures (a)–(f) for tree and forests in the binary case under configuration (I) as γ varies. Tree depth is fixed at $l = 7$. Red: tree; blue: forest.

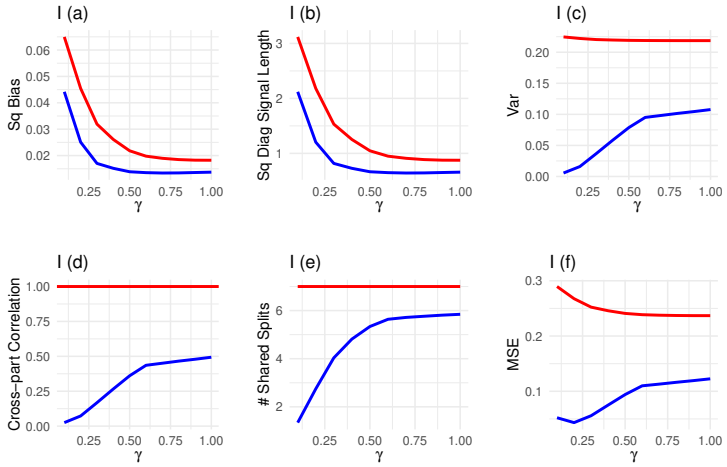

Figure 2: Performance measures (a)–(f) for tree and forest in the continuous case under configuration (I) as γ varies. Tree depth is fixed at $l = 7$. Red: tree; blue: forest.

our theoretical analysis in Remark 9. Additionally, the correlation between trees is low, as shown in Figure 1(d), also explaining the reduction in variance.

For the continuous features, the trends observed in Figure 2 for performance measures (a)–(f) exhibit similar patterns to those in the binary case. These results also support our conclusion in Remark 16, demonstrating that RF consistently achieves lower squared bias and variance compared to tree as γ varies. They also confirm the improved effectiveness of RF across different types of feature distribution.

Additionally, as seen in Figures 1(f) and 2(f), the MSE curves for RFs exhibit U-shaped patterns, with the optimal MSE differing in these two settings. These observations are consistent with Remarks 21 and 13, and indicate that the *optimal* γ can be sensitive to the model structure, calling for further investigation.

4.2.2 Type II exogenous randomness alone can enhance RFs' performance

Type I exogenous randomness from feature subsampling is widely recognized as a key factor in the success of random forests and has prompted extensive research. In contrast, Type II exogenous randomness has received little attention. Indeed, it is commonly believed that without Type I exogenous randomness (i.e., $\gamma = 1$), random forest is considered identical to a single decision tree. This subsection challenges such existing belief and demonstrates that Type II exogenous randomness can work *solely* to improve the performance of random forests compared to a single tree. Additionally, our simulation results support the "blessing of dimensionality" phenomenon: as the dimensionality grows, the presence of many non-informative features introduces tie-breaking randomness that leads to a clear reduction in variance. The detailed simulation results and interpretations are presented in Appendix Section A.1.

4.2.3 The role of Type I exogenous randomness in RFs' success

In this subsection, by comparing scenario when $\gamma < 1$ to that when $\gamma = 1$, we gain additional insights into the role of Type I randomness in the success of RF. We demonstrate that feature subsampling allows noise features to contribute earlier (to variance reduction), before tree depth reaches $l = s = 5$ ($l = 5$ means that all informative features have been split once), and thus further reduces cross-tree correlations across all tree depths (see Figure 3(d)). The performance measures (a)–(f) are plotted in Figures 3 and 4 below for Equal Configuration (I), and in Figures 5 and 6 for Unequal Configuration (II), respectively, as we vary l.

For configuration (I), feature subsampling adds Type I exogenous randomness into the tree-building process, and thus, non-informative features now have a chance to be split earlier than informative ones at the first s tree depths; see Figure 3(b) for the increasing number of unsplit informative features at each fixed tree depth when γ decreases. This early involvement of non-informative features has a strong effect on reducing cross-tree correlation in high-dimensional settings, leading to a significant decrease in variance and correlation as γ decreases at each fixed tree depth; see Figures 3(c) and (d). Indeed, Figure 3(d) shows that the cross-tree correlation stays below 1 for all $\gamma < 1$ even when tree depth $l = s = 5$, a distinction from the $\gamma = 1$ scenario where $l = s = 5$ has cross-tree correlation 1. The decreasing patterns of variance and cross-tree correlation are supported by the reduction in shared splits between I_l and I'_l as shown theoretically in (24) and empirically in Figure 3(e). However, such variance reduction comes at the cost of increased squared bias (especially for small l), as shown in Figure 3(a), because of the heavy involvement of non-informative features. This directly leads to a bias-variance tradeoff (see Figure 3(f)), and $\gamma = 1$ is not optimal as l increases.

For the continuous case, recall the discussion in Section A.1 that non-informative features never enter the tree-building process when $\gamma = 1$. Thanks to Type I exogenous randomness, for $\gamma < 1$, non-informative features have a chance to be split and thus reduce

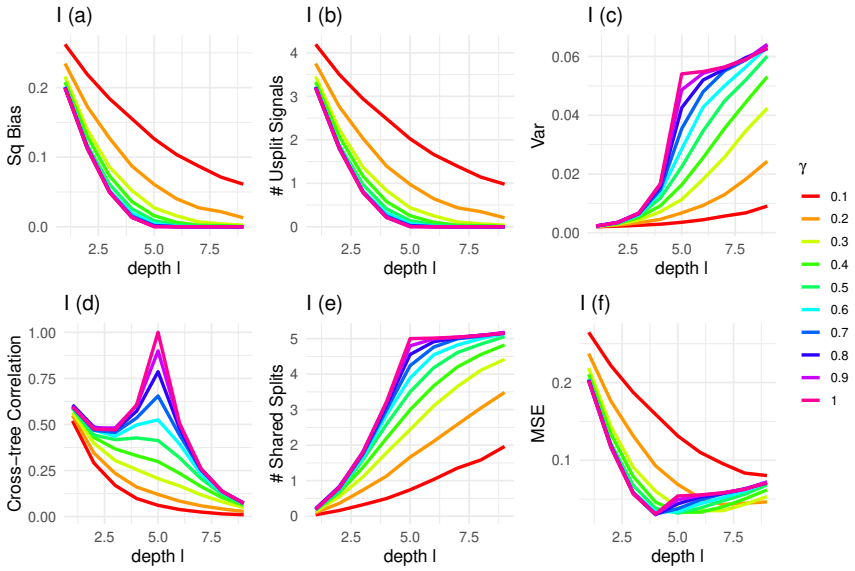

Figure 3: Performance measures (a)–(f) for forests in the binary case as γ and l vary under configuration (I).

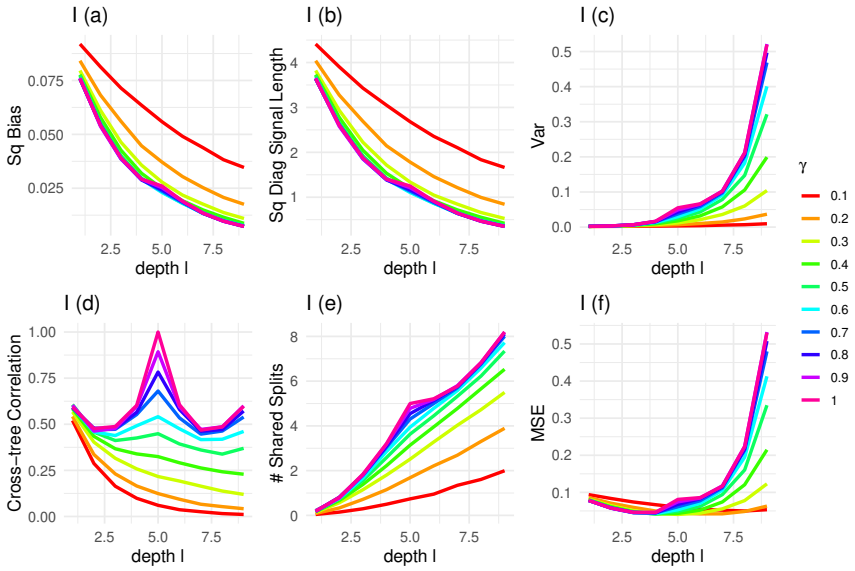

Figure 4: Performance measures (a)–(f) for forests in the continuous case as γ and l vary under configuration (I).

cross-tree correlation, leading to similar patterns in performance measures to those in the binary case; see Figure 4.

A surprising observation is that feature subsampling does not always increase the squared bias at the early stage of tree-building. To demonstrate it, consider Unequal Configuration (II) in Figures 5 and 6. Recall that as in Figures 9 and 7, the tree-building process is purely deterministic when $\gamma = 1$ and $l \leq s$: features are split deterministically in the decreasing order of impurity decrement. With $\gamma < 1$, however, since important features may not be selected into some feature subset, such deterministic splitting order is disrupted. This may seem harmful to bias reduction because less important variables can be split earlier on. However, feature subsampling creates a large number of trees, and ensemble allows (on average) many more informative features to be split earlier on. This results in a reduction of squared bias by encouraging fewer shared-unsplit-informative variables by two independent CART processes; see (24), (34), and Figures 5(a)–(b). Meanwhile, Type I exogenous randomness also introduces variability in splits along informative features, allows non-informative features to be involved earlier, and thus helps reduce the cross-tree correlation in Figure 5(d), leading to lower variance in Figure 5(c) and a reduction in MSE in Figure 5(e). The observed behaviors in Figure 6 resemble those in the binary setting, with similar interpretations. To summarize, in this example, we see that feature subsampling can introduce beneficial randomness to achieve *both* bias and variance reduction.

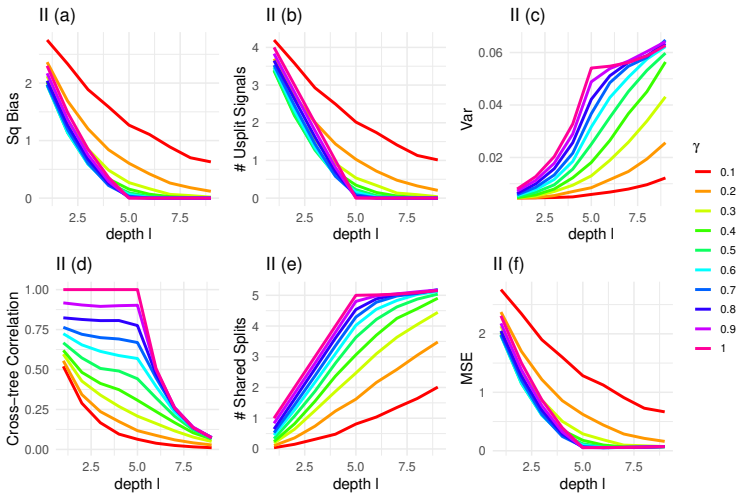

Figure 5: Performance measures (a)–(f) for forests in the binary case as γ and l vary under configuration (II).

Combining with the discussions in Section 4.2.1, our finding clearly shows that feature subsampling can reduce *both* bias and variance when γ is not too small, where the bias reduction phenomenon may be model-specific and the variance reduction could be a robust phenomenon across models. Our theory in Section 4.1 explains clearly how bias and variance reductions are achieved. Although (I) and (II) represent two specific configurations, they share the same mechanisms where feature subsampling improves RF's performance by *early* involvement of non-informative variables. These results underscore the *generality* of the feature subsampling effect in enhancing RF's performance across various coefficient configurations.

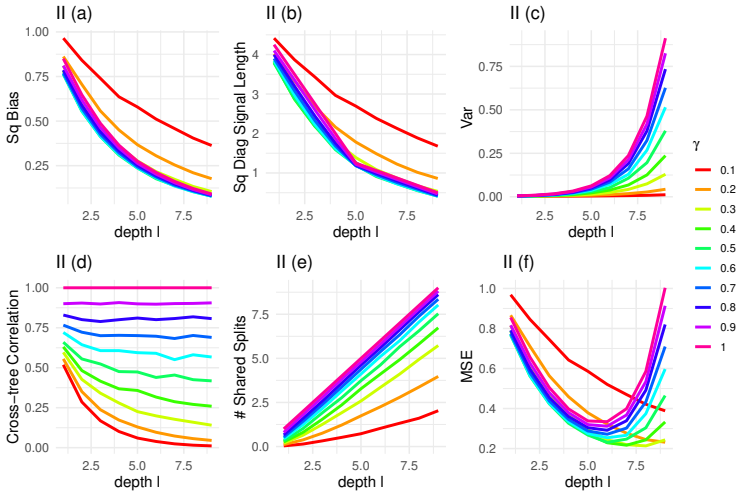

Figure 6: Performance measures (a)–(f) for forests in the continuous case as γ and l vary under configuration (II).

5. Discussions

In this paper, we have studied *both* theoretically and empirically the effects of exogenous randomness in the success of RFs with training-data independent partitioning rules. The exogenous randomness can be caused by random feature subsampling *or* random tie-breaking in forming an RF estimate. Our study has provided theoretical and empirical justifications on the advantage of exogenous randomness in the success of RFs. We have focused on training-data-independent partitioning rules, and an interesting direction for future research would be to explore data-adaptive partitioning rules, particularly the sample CART splitting criterion. It would also be interesting to extend the study to sequence data by exploiting the Transformer and BERT models.

Acknowledgments

This work was supported in part by NSF grants EF-2125142, DMS-2310981, and DMS-2324490.

Appendix A. Additional Simulations

We present additional simulation studies here to illustrate the effect of Type II exogenous randomness in RF's performance.

A.1 Type II exogenous randomness alone can enhance RFs' performance

We provide details for Section 4.2.2 of the main text. To illustrate this point, we examine the performance measures (a)–(f) and present the results in Figures 7–10 for forests and trees under various settings discussed at the beginning of this section while fixing $\gamma = 1$.

We note some key points before presenting our numerical results. First, with $\gamma = 1$, both binary and uniform CART processes prioritize informative features for splitting, as long as they are available. Second, along a tree branch connecting an end cell (at level l) with the root cell (at level 0), a continuous variable can be split as many as l times, while a binary variable can only be split at most once. These basic properties can assist us in understanding the numerical results.

Figures 7 and 8 present the results with continuous features, where each split can only be on informative features. Thus, under Unequal Configuration (II), no exogenous randomness is present and thus random forests collapses to a single tree; hence, all curves in panel (a)–(f) coincide as shown in Figure 7.

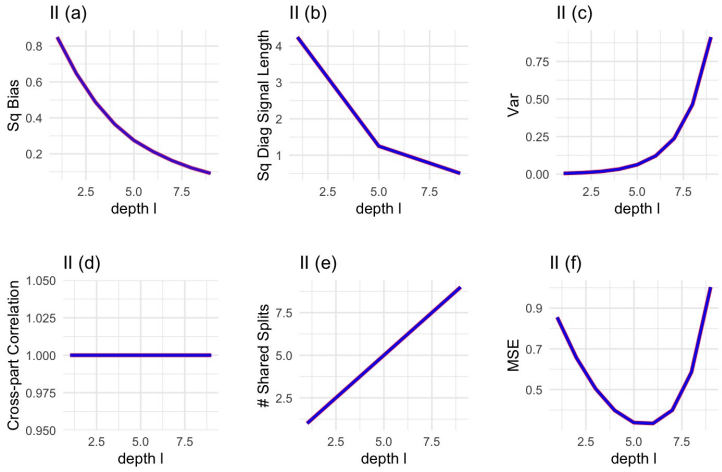

Figure 7: Performance measures (a)–(f) for tree and forest in the continuous case under configuration (II) as tree depth l varies when $\gamma = 1$. Red: tree; blue: forest. Without exogenous randomness, RF and single tree are identical.

Under Equal Coefficient Configuration (I) in Figure 8, on the other hand, random tie-breaking may happen when splitting informative features, causing Type II exogenous randomness and multiple trees in the forest. Figure 8(d) shows an intriguing pattern in the cross-tree correlation:

i) when $l < 5$, the cross-tree correlation is strictly less than 1;

ii) at $l = 5$, each informative feature has been split exactly once and thus there is no Type II exogenous randomness, resulting in identical tree and forest performance;

iii) for $l > 5$ but $l < 10$, each of the $s = 5$ informative features is split exactly one more time, leading to a second round of reduction in correlation caused by Type II exogenous randomness.

Additionally, we observe the interesting phenomenon of quasi-periodic fluctuations in Figures 8(d) and (e). Each time when trees grow from depth $ks + 1$ to depth $(k + 1)s$ for $k = 0, 1, ...,$ each informative feature is split exactly once. As k increases, the impurity decrement in each quasi-period gradually diminishes, resulting in slower and slower decreases each period in squared bias in Figure 8(a). However, the variance becomes increasingly larger as l increases, because the leading term $2^l/n$ in variance keeps growing as l increase. This explains the substantial rise in both variance and MSE when l is large, as shown in Figures 8(c) and (f).

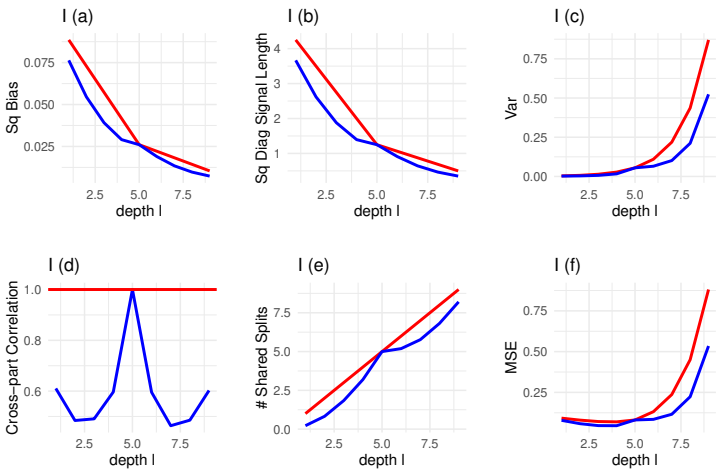

Figure 8: Performance measures (a)–(f) for tree and forest in the continuous case under configuration (I) as l varies. We fix $\gamma = 1$. Red: tree; blue: forest.

Next, we consider binary features, where each feature is split at most once along each tree branch and informative features are always split before uninformative ones when coexisting. The Unequal Configuration (II) results are presented in Figure 9, which can be explained theoretically using our findings in Section 4.1.1. In detail,

(i) When depth $l < s = 5$, since the coefficients are distinct (no Type II randomness), the performances of random forests and a single tree are identical. The bias quickly reaches 0 as tree depth grows to $s = 5$.

(ii) When depth $l \geq s + 1$, the binary CART process begins to split non-informative features, which causes Type II randomness and a significant variance reduction. Since $d - s$ is large, it is generally hard for I_l and I'_l to share any split on non-informative

features (see Figure 9(e)). This leads to a sharp decline in cross-tree correlation $2^{\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\} - l} \approx 2^{-(l-5)}$ for $l > 5$ as shown in Figure 9(d) and the significantly slow growth in variance and MSE in Figures 9(c) and (f).

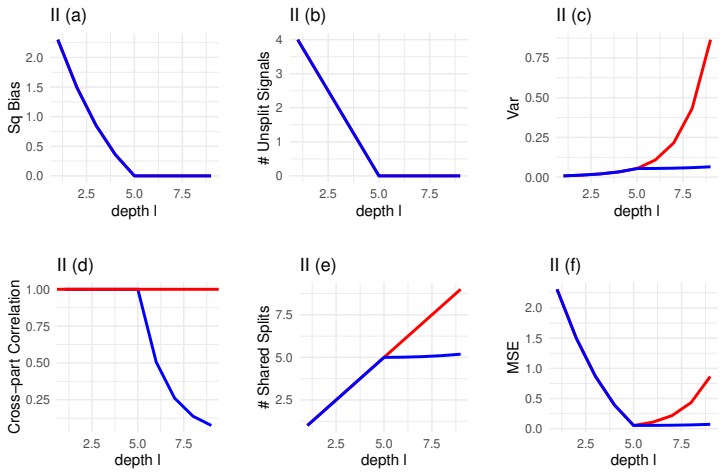

Figure 9: Performance measures (a)–(f) for tree and forest in the binary case under configuration (II) as tree depth l varies when $\gamma = 1$. Red: tree; blue: forest.

Figure 10 presents Equal Coefficient Configuration (I) with discrete features. The only difference from Figure 9 setting is that there exists Type II exogenous randomness among informative features, which explains the bias and variance reduction of RF compared to tree before depth reaches 5.

It is worth emphasizing that the binary feature example above reveals a notable phenomenon that the existence of many non-informative features can be a *blessing* to random forests' success, because of their contribution to reduced cross-tree correlations by limiting the number of shared splits. To further illustrate the "blessing of dimensionality", we fix the feature subsampling rate γ and gradually increase the number of noise features (i.e., the ambient dimensionality d). The simulations are conducted under two types of feature distributions and two configurations, configuration (I) with a fixed $\gamma = 0.33$ and configuration (II) with $\gamma = 0.6$, respectively. As shown in Figures 11–14, both the variance component and the overall MSE decrease steadily with d, whereas the squared bias fluctuates only slightly. This demonstrates that adding non-informative features can effectively optimize the performance of MSE via variance reduction—the essence of the "blessing of dimensionality."

We acknowledge that the blessing of high dimensionality is, to some extent, because of the population CART partitioning rule. For the sample CART partitioning rule, the effect of noise variables and high dimensionality can be more complicated because of the finite-sample estimation error involved. However, the phenomenon of non-informative features enhancing model performance has been observed in various statistical settings. Specifically, Mentch and Zhou (2022) identified this effect under the augmented bagging (AugBagg)

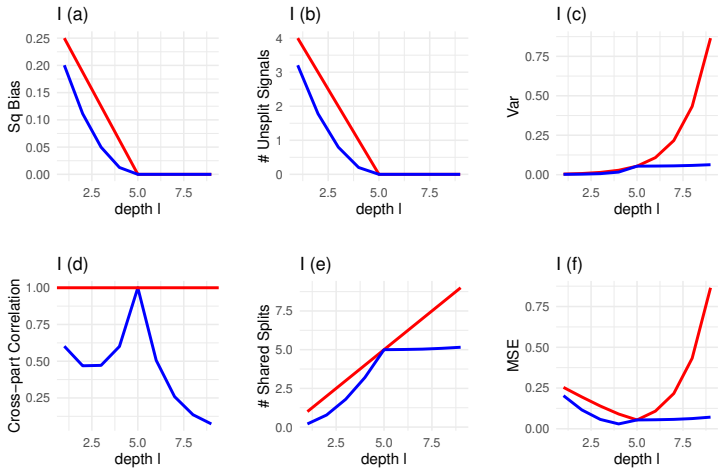

Figure 10: Performance measures (a)–(f) for tree and forest in the binary case under configuration (I) as l varies. We fix $\gamma = 1$. Red: tree; blue: forest.

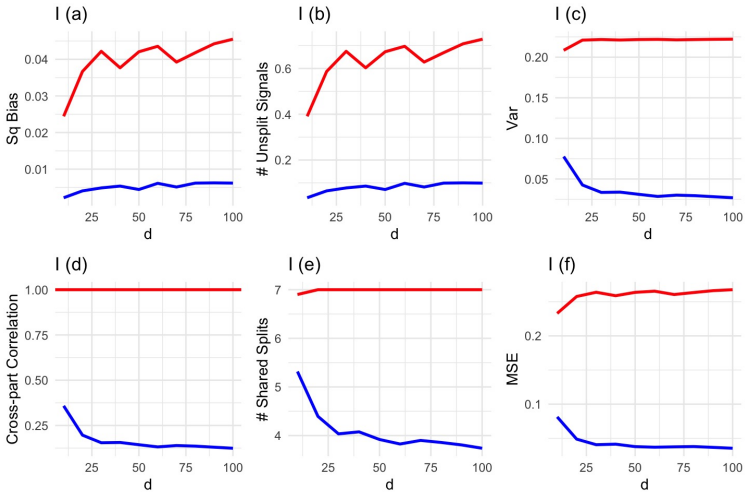

Figure 11: Performance measures (a)–(f) for tree and forest in the binary case under configuration (I) as d varies. We fix $l = 7$ and $\gamma = 0.33$. Red: tree; blue: forest.

learning framework, and Kobak et al. (2020) found similar benefits in the context of ridge regression. These findings highlight the potential generality of this phenomenon across different learning methods and underscore its value for further investigation, particularly for RF with sample CART partitioning rule.

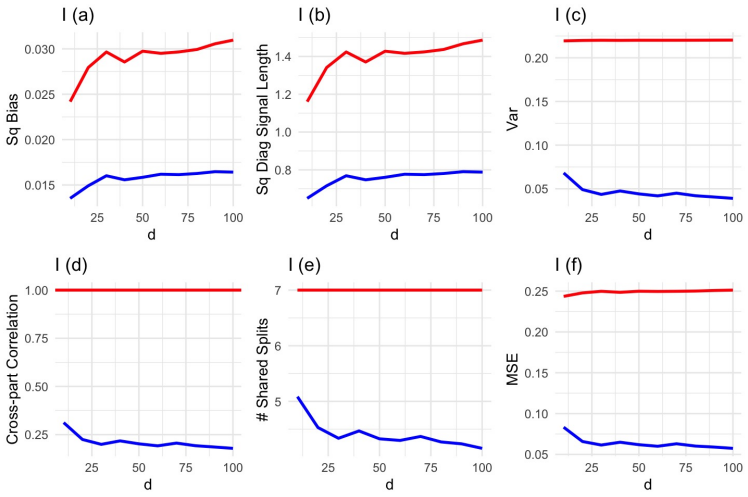

Figure 12: Performance measures (a)–(f) for tree and forest in the continuous case under configuration (I) as d varies. We fix $l = 7$ and $\gamma = 0.33$. Red: tree; blue: forest.

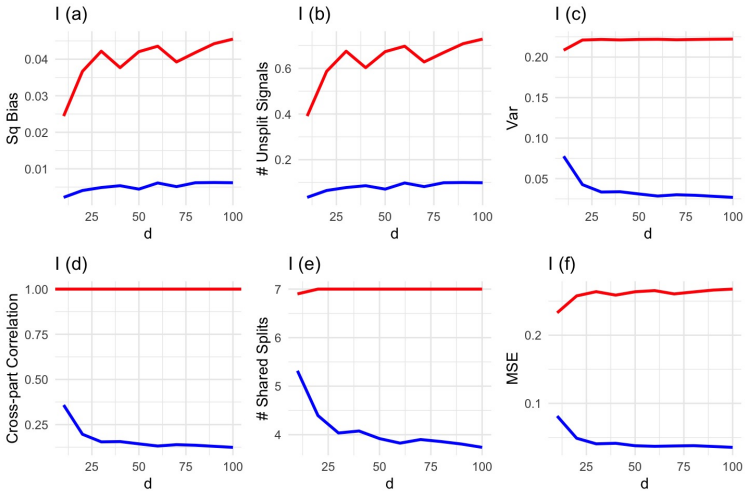

Figure 13: Performance measures (a)–(f) for tree and forest in the binary case under configuration (II) as d varies. We fix $l = 7$ and $\gamma = 0.6$. Red: tree; blue: forest.

Appendix B. Local MSE Bounds for Ensemble Estimators

Let $\widehat{\mu}(\boldsymbol{x}; \boldsymbol{Z}, \boldsymbol{P}^n(\Theta))$ be either $\widehat{\mu}_{\text{part}}$ or $\widehat{\mu}_{\text{ens}}$ as in (2) and (5) in Section 2. For each given $\boldsymbol{x}_0 \in \mathcal{X}^d$, the *local mean squared error (LMSE)* is defined as

$$\text{MSE}(\widehat{\mu}; \boldsymbol{x}_0) = \mathbb{E}_{\boldsymbol{Z},\Theta}\left[(\mu(\boldsymbol{x}_0) - \widehat{\mu}(\boldsymbol{x}_0; \boldsymbol{Z}, \boldsymbol{P}^n(\Theta)))^2\right]. \tag{B.1}$$

In this section, we aim to study the non-asymptotic expansion of the local MSE for ensemble estimators.

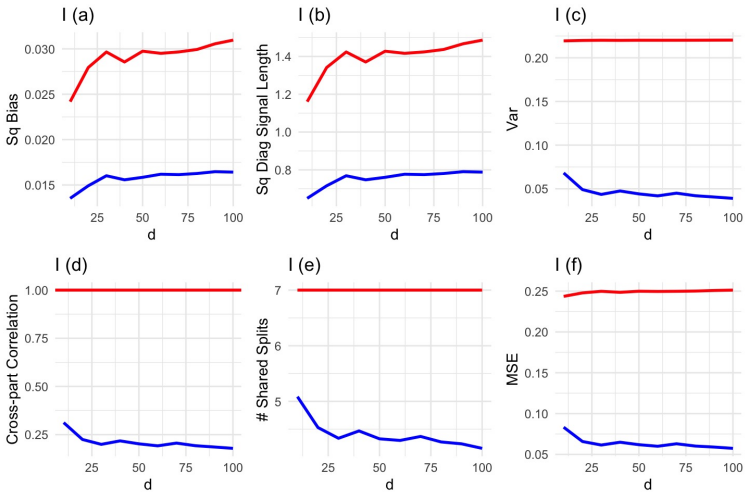

Figure 14: Performance measures (a)–(f) for tree and forest in the continuous case under configuration (II) as d varies. We fix $l = 7$ and $\gamma = 0.6$. Red: tree; blue: forest.

Let us define two kernel functions around the target point $\boldsymbol{x}_0$ as

$$\begin{aligned} K_\mu(P, P'; \boldsymbol{x}_0) &= \mathbb{E}_{\boldsymbol{X}}\left((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_{\boldsymbol{x}_0}))(\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P'_{\boldsymbol{x}_0}))|P_{\boldsymbol{x}_0} \cap P'_{\boldsymbol{x}_0}\right) \\ &\quad \times \frac{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0} \cap P'_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})}, \\ K_{\sigma^2}(P, P'; \boldsymbol{x}_0) &= \mathbb{E}_{\boldsymbol{X}}\left(\sigma^2|P_{\boldsymbol{x}_0} \cap P'_{\boldsymbol{x}_0}\right) \frac{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0} \cap P'_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})}, \end{aligned} \tag{B.2}$$

respectively, where P and P' are elements in $\mathcal{P}_0$. In particular, when $P = P'$, denote by

$$\begin{aligned} Q_\mu(P; \boldsymbol{x}_0) &= K_\mu(P, P; \boldsymbol{x}_0) = \frac{\mathrm{Var}_{\boldsymbol{X}}(\mu|P_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})}, \\ Q_{\sigma^2}(P; \boldsymbol{x}_0) &= K_{\sigma^2}(P, P; \boldsymbol{x}_0) = \frac{\mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})}. \end{aligned} \tag{B.3}$$

These functions capture the interaction pattern of partitioning estimators generated by P and P' locally around the target point $\boldsymbol{x}_0$, and they also exhibit a similar Cauchy–Schwarz property as the global cross-partition covariance functions.

Proposition 22 *Let P and P' be two components in $\mathcal{P}_0^n$ (defined in Assumption 4). Then it holds that*

$$\begin{aligned} K_\mu(P, P'; \boldsymbol{x}_0) &\le \sqrt{Q_\mu(P; \boldsymbol{x}_0) Q_\mu(P'; \boldsymbol{x}_0)}, \\ K_{\sigma^2}(P, P'; \boldsymbol{x}_0) &\le \sqrt{Q_{\sigma^2}(P; \boldsymbol{x}_0) Q_{\sigma^2}(P'; \boldsymbol{x}_0)}. \end{aligned} \tag{B.4}$$

For the first result, the equality holds if and only if $P_{\boldsymbol{x}_0}$ are $P'_{\boldsymbol{x}_0}$ are indistinguishable, or μ is constant on $P_{\boldsymbol{x}_0} \cup P'_{\boldsymbol{x}_0}$. For the second result, the equality holds if and only if $P_{\boldsymbol{x}_0}$ are

$P'_{\boldsymbol{x}_0}$ *are indistinguishable. Consequently, Assumption 4(3) ensures that the second result in* (B.4) *is a strict inequality, and that when* μ *is a non-constant function, the first result in* (B.4) *is a strict inequality.*

The proof of Proposition 22 above is provided later in Section B.3.

Theorem 23 (Local MSE bounds) *Assume that for any fixed target point* $\boldsymbol{x}_0 \in \mathcal{X}^d$*, the random cell* $\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)$ *containing the target point satisfies*

$$\sup_n \mathbb{E}_\Theta \left[\frac{1}{n\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta))} \right] < \infty. \tag{B.5}$$

Then under Assumptions 1–4, the local MSE of the ensemble estimator satisfies

$$\begin{aligned} MSE(\widehat{\mu}_{ens}; \boldsymbol{x}_0) = & \underbrace{\frac{B-1}{B} \left(\mu(\boldsymbol{x}_0) - \mathbb{E}_\Theta \left(\mathbb{E}_{\boldsymbol{X}}(\mu | \boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)) \right) \right)^2}_{\text{squared bias of ensemble}} \\ & + \underbrace{\frac{1}{B} \mathbb{E}_\Theta \left[(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu | \boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)))^2 \right]}_{\text{squared bias of a single partition}} \\ & + \underbrace{\frac{B-1}{Bn} \mathbb{E}_{\Theta,\Theta'} \left[K_\mu(\boldsymbol{P}^n(\Theta), \boldsymbol{P}^n(\Theta'); \boldsymbol{x}_0) + K_{\sigma^2}(\boldsymbol{P}^n(\Theta), \boldsymbol{P}^n(\Theta'); \boldsymbol{x}_0) \right]}_{\text{cross-partition covariance}} \\ & + \underbrace{\frac{1}{Bn} \mathbb{E}_\Theta \left[Q_\mu(\boldsymbol{P}^n(\Theta); \boldsymbol{x}_0) + Q_{\sigma^2}(\boldsymbol{P}^n(\Theta); \boldsymbol{x}_0) \right]}_{\text{single-partition variance}} + \mathcal{R}_{ens}(\mu; \boldsymbol{x}_0), \end{aligned} \tag{B.6}$$

in which Θ *and* Θ' *are mutually independent and the remainder* $\mathcal{R}_{ens}(\mu; \boldsymbol{x}_0)$ *satisfies*

$$\begin{aligned} \mathcal{R}_{ens}(\mu; \boldsymbol{x}_0) \lesssim & (\|\mu\|_\infty + \|\sigma^2\|_\infty)^2 \times \Bigg\{ \mathbb{E}_\Theta((1 - \mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)))^n) \\ & + \frac{B-1}{B} \mathbb{E}_{\Theta,\Theta'} \left(\frac{\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta) \cap \boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta'))}{n\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta))\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta'))(1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)))} \right) \\ & + \frac{B-1}{B} \mathbb{E}_\Theta \left(\frac{1}{(1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)))^{3/2}} \right) \\ & + \frac{1}{B} \mathbb{E}_\Theta \left(\frac{1}{n\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta))(1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}^n_{\boldsymbol{x}_0}(\Theta)))} \right) \Bigg\} \end{aligned}$$

with the constant in $\lesssim$ *independent of* $\boldsymbol{x}_0$*,* μ*,* σ^2*,* $\boldsymbol{P}^n(\Theta)$*, and* n*. In particular, the local MSE of a single partitioning estimator* μ_{part} *is the special case of* $B = 1$ *in (B.6).*

The proof of Theorem 23 above is provided in Section B.2 later.

B.1 Technical Preparation

This section is devoted to providing an expansion of the local MSE, based on which our main results are built. In what follows, denote by $\mathbb{E}_{\boldsymbol{X}|\Theta,\Theta'}(\mu) = \mathbb{E}(\mu(\boldsymbol{X};\Theta,\Theta')|\Theta,\Theta')$ the conditional expectation of function $\mu(\boldsymbol{X};\Theta,\Theta')$ given the exogenous factors (Θ,Θ'). In other words, given the remaining two Θ's, we integrate with respect to $\boldsymbol{X}$. Similarly, $\mathbb{E}_{\varepsilon|\boldsymbol{X},\Theta,\Theta'}(\cdot)$ is the conditional expectation with respect to ε given $(\boldsymbol{X},\Theta,\Theta')$. To simplify the notation, we omit the superscript n in $\boldsymbol{P}^n$ throughout the proof whenever there is no confusion.

We consider the local version decomposition of MSE as in (7)

$$\begin{aligned}\mathrm{MSE}(\widehat{\mu}_{ens};\boldsymbol{x}_0) &= \mathbb{E}_{\boldsymbol{Z},\Theta_{1:B}}[(\mu(\boldsymbol{x}_0)-\widehat{\mu}_{ens}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B}))^2] \\ &= \underbrace{\mathbb{E}_{\Theta_{1:B}}\left[(\mu(\boldsymbol{x}_0))-\mathbb{E}_{\boldsymbol{Z}|\Theta_{1:B}}[\widehat{\mu}_{ens}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B}))]\right)^2]}_{\mathrm{Bias}^2(\boldsymbol{x}_0)} \\ &\quad + \underbrace{\mathbb{E}_{\Theta_{1:B}}\left[\mathrm{Var}_{\boldsymbol{Z}|\Theta_{1:B}}[\widehat{\mu}_{ens}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B}))]\right]}_{\mathrm{Var}(\boldsymbol{x}_0)},\end{aligned} \tag{B.7}$$

where $\widetilde{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0;\boldsymbol{P}^n(\Theta)) = \mathbb{E}_{\boldsymbol{Z}}[\widehat{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}^n(\Theta))]$ with the expectation taken with respect to the training sample $\boldsymbol{Z}$.

For further investigation of the structure of the local MSE in (B.7), let us introduce the following notation

$$\widetilde{\mu}(\boldsymbol{x}_0;\Theta) := \mathbb{E}_{\boldsymbol{Z}|\Theta}(\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))), \tag{B.8}$$

$$V_1(\boldsymbol{x}_0;\Theta) := \mathrm{Var}_{\boldsymbol{Z}|\Theta}[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))], \tag{B.9}$$

$$V_2(\boldsymbol{x}_0;\Theta,\Theta') := \mathrm{Cov}_{\boldsymbol{Z}|\Theta,\Theta'}[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta)),\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta'))]. \tag{B.10}$$

The meanings of these quantities are given below. The notation $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$ stands for a function of the exogenous factor Θ by integrating over the training data of a single partitioning estimator. The function $V_1(\boldsymbol{x}_0;\Theta)$ refers to the conditional variance of the partitioning estimator given the partition $\boldsymbol{P}(\Theta)$ and thus is also a function of Θ. The bivariate function $V_2(\boldsymbol{x}_0;\Theta,\Theta')$ represents the conditional covariance of two partitioning estimators when the corresponding partitions $\boldsymbol{P}(\Theta)$ and $\boldsymbol{P}(\Theta')$ are fixed.

On the one hand, for the conditional expectation part, it holds that

$$\mathbb{E}_{\boldsymbol{Z}|\Theta_{1:B}}(\widehat{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B}))) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_{\boldsymbol{Z}|\Theta_b}(\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_b))) = \frac{1}{B}\sum_{b=1}^{B}\widetilde{\mu}(\boldsymbol{x}_0;\Theta_b),$$

in which the right-hand side (RHS) is a sum of independent and identically distributed (i.i.d.) random variables. Hence, the conditional expectation part in (B.7) reduces to

$$\begin{aligned}&\mathbb{E}_{\Theta_{1:B}}\left[(\mu(\boldsymbol{x}_0)-\mathbb{E}_{\boldsymbol{Z}|\Theta_{1:B}}(\widehat{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B}))))^2\right] \\ =&\mathbb{E}_{\Theta_{1:B}}\left\{\left(\mu(\boldsymbol{x}_0)-\frac{1}{B}\sum_{b=1}^{B}\widetilde{\mu}(\boldsymbol{x}_0;\Theta_b)\right)^2\right\} \\ =&\frac{1}{B}\mathbb{E}_{\Theta}\left\{(\mu(\boldsymbol{x}_0)-\widetilde{\mu}(\boldsymbol{x}_0;\Theta))^2\right\}+\frac{B-1}{B}\mathbb{E}_{\Theta,\Theta'}\left\{(\mu(\boldsymbol{x}_0)-\widetilde{\mu}(\boldsymbol{x}_0;\Theta))(\mu(\boldsymbol{x}_0)-\widetilde{\mu}(\boldsymbol{x}_0;\Theta'))\right\} \\ =&\frac{1}{B}\mathbb{E}_{\Theta}\left\{(\mu(\boldsymbol{x}_0)-\widetilde{\mu}(\boldsymbol{x}_0;\Theta))^2\right\}+\frac{B-1}{B}(\mu(\boldsymbol{x}_0)-\mathbb{E}_{\Theta}[\widetilde{\mu}(\boldsymbol{x}_0;\Theta)])^2.\end{aligned}$$

Observe that $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$ and $\widetilde{\mu}(\boldsymbol{x}_0;\Theta')$ are independent and identically distributed. We immediately see that

$$\begin{aligned}\mathrm{Bias}^2(\boldsymbol{x}_0) =&\mathbb{E}_{\Theta_{1:B}}\left[(\mu(\boldsymbol{x}_0)-\mathbb{E}_{\boldsymbol{Z}|\Theta_{1:B}}(\widehat{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B}))))^2\right]\\ =&\frac{1}{B}\mathbb{E}_{\Theta}\left\{(\mu(\boldsymbol{x}_0)-\widetilde{\mu}(\boldsymbol{x}_0;\Theta))^2\right\}+\frac{B-1}{B}(\mu(\boldsymbol{x}_0)-\mathbb{E}_{\Theta}[\widetilde{\mu}(\boldsymbol{x}_0;\Theta)])^2\\ =&(\mu(\boldsymbol{x}_0)-\mathbb{E}_{\Theta}[\widetilde{\mu}(\boldsymbol{x}_0;\Theta)])^2+\frac{1}{B}\mathrm{Var}_{\Theta}(\widetilde{\mu}(\boldsymbol{x}_0;\Theta)).\end{aligned} \tag{B.11}$$

On the other hand, for the conditional variance part, we have

$$\begin{aligned}&\mathrm{Var}_{\boldsymbol{Z}|\Theta_{1:B}}(\widehat{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0);\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B})))=\frac{1}{B^2}\sum_{b=1}^{B}\mathrm{Var}_{\boldsymbol{Z}|\Theta_b}\left[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_b))\right]\\ &\quad+\frac{1}{B^2}\sum_{b\neq b'}\mathrm{Cov}_{\boldsymbol{Z}|\Theta_b,\Theta_{b'}}\left[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_b)),\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{b'}))\right]\\ &=\frac{1}{B^2}\sum_{b=1}^{B}V_1(\boldsymbol{x}_0;\Theta_b)+\frac{1}{B^2}\sum_{b\neq b'}V_2(\boldsymbol{x}_0;\Theta_b,\Theta_{b'}).\end{aligned}$$

Since $\Theta_1,\ldots,\Theta_B$ are i.i.d. copies of the exogenous decision factors, both $\{V_1(\boldsymbol{x}_0;\Theta_b)\}$ and $\{V_2(\boldsymbol{x}_0;\Theta_b,\Theta_{b'})\}$ are collections of identically distributed variables, and as such, the conditional variance term in (B.7) can be simplified as

$$\begin{aligned}\mathrm{Var}(\boldsymbol{x}_0) =&\mathbb{E}_{\Theta_{1:B}}[\mathrm{Var}_{\boldsymbol{Z}|\Theta_{1:B}}(\widehat{\mu}_{\mathrm{ens}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta_{1:B})))]\\ =&\frac{1}{B}\mathbb{E}_{\Theta}[V_1(\boldsymbol{x}_0;\Theta)]+\frac{B-1}{B}\mathbb{E}_{\Theta,\Theta'}[V_2(\boldsymbol{x}_0;\Theta,\Theta')].\end{aligned} \tag{B.12}$$

B.2 Proof of Theorem 23

According to (B.7), (B.11), and (B.12), it holds that

$$\begin{aligned}\mathrm{MSE}(\widehat{\mu}_{\mathrm{ens}};\boldsymbol{x}_0)=&\underbrace{(\mu(\boldsymbol{x}_0)-\mathbb{E}_{\Theta}(\widetilde{\mu}(\boldsymbol{x}_0;\Theta))^2+\frac{1}{B}\mathrm{Var}_{\Theta}(\widetilde{\mu}(\boldsymbol{x}_0;\Theta))}_{\mathrm{Bias}^2(\boldsymbol{x}_0)}\\ &+\underbrace{\frac{B-1}{B}\mathbb{E}_{\Theta,\Theta'}(V_2(\boldsymbol{x}_0;\Theta,\Theta'))+\frac{1}{B}\mathbb{E}_{\Theta}(V_1(\boldsymbol{x}_0;\Theta))}_{\mathrm{Var}(\boldsymbol{x}_0)},\end{aligned} \tag{B.13}$$

in which $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$, $V_1(\boldsymbol{x}_0;\Theta)$, and $V_2(\boldsymbol{x}_0;\Theta,\Theta')$ are defined as in (B.8), (B.9), and (B.10), respectively. The proof of Theorem 23 primarily involves deriving the asymptotic expansions for these three quantities. We omit the superscript n from $\boldsymbol{P}^n$ whenever there is no ambiguity.

B.2.1 Asymptotic Expansion of $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$ in (B.8)

This section is devoted to showing the connection of $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$ in (B.14) and $\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]$, the conditional expectation of the ground truth μ over the target cell $\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)$.

In view of the definitions in (2) and (B.8), we have

$$\begin{aligned}\widetilde{\mu}(\boldsymbol{x}_0;\Theta) &= \mathbb{E}_{\boldsymbol{Z}|\Theta}[\widehat{\mu}_{\text{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))] \\ &= \mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)}\sum_{i=1}^{n} Y_i I\{\boldsymbol{X}_i \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ &= \mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)}\sum_{i=1}^{n} (\mu(\boldsymbol{X}_i)+\varepsilon_i) I\{\boldsymbol{X}_i \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ &= \mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)}\sum_{i=1}^{n} \mu(\boldsymbol{X}_i) I\{\boldsymbol{X}_i \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right],\end{aligned}$$

since $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = 0$. Recall that the cell $\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)$ is independent of observations $\{\boldsymbol{X}_j\}$, and hence, $N_{\boldsymbol{x}_0}(\Theta) = \sum_{i=1}^{n} I\{\boldsymbol{X}_i \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}$ follows a binomial distribution $\mathcal{B}(n, p(\boldsymbol{x}_0;\Theta))$ with $p(\boldsymbol{x}_0;\Theta) = \mathbb{P}_{\boldsymbol{X}|\Theta}(\boldsymbol{X} \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))$ when the partition is given. Thus, it holds that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)}\sum_{i=1}^{n} \mu(\boldsymbol{X}_i) I\{\boldsymbol{X}_i \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\sum_{k=1}^{n}\left(\frac{1}{k}\sum_{i=1}^{n} \mu(\boldsymbol{X}_i) I\{\boldsymbol{X}_i \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right) I\{N_{\boldsymbol{x}_0}(\Theta) = k\}\right] \\ =&\sum_{k=1}^{n}\binom{n}{k}\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\left(\frac{1}{k}\sum_{i=1}^{k} \mu(\boldsymbol{X}_i)\right) I\{\boldsymbol{X}_{1:k} \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta); \boldsymbol{X}_{(k+1):n} \notin \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ =&\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]\sum_{k=1}^{n}\binom{n}{k} p(\boldsymbol{x}_0;\Theta)^k (1-p(\boldsymbol{x}_0;\Theta))^{n-k} \\ =&\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]\left(1-(1-p(\boldsymbol{x}_0;\Theta))^n\right).\end{aligned}$$

In conclusion, we now find that

$$\widetilde{\mu}(\boldsymbol{x}_0;\Theta) = \mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))\left(1-(1-p(\boldsymbol{x}_0;\Theta))^n\right) \tag{B.14}$$

with $p(\boldsymbol{x}_0;\Theta) = \mathbb{P}_{\boldsymbol{X}|\Theta}(\boldsymbol{X} \in \boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))$.

Back to (B.13), it follows immediately from (B.14) that

$$\begin{aligned}(\mu(\boldsymbol{x}_0) - \mathbb{E}_\Theta(\widetilde{\mu}(\boldsymbol{x}_0;\Theta)))^2 &= (\mu(\boldsymbol{x}_0) - \mathbb{E}_\Theta(\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))))^2 \\ &\quad + \|\mu\|_\infty^2 O(\mathbb{E}_\Theta((1-p(\boldsymbol{x}_0;\Theta))^n)), \\ \text{Var}_\Theta(\widetilde{\mu}(\boldsymbol{x}_0;\Theta)) &= \text{Var}_\Theta(\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)))) \\ &\quad + \|\mu\|_\infty^2 O(\mathbb{E}_\Theta((1-p(\boldsymbol{x}_0;\Theta))^n)),\end{aligned} \tag{B.15}$$

which gives the asymptotic expansion for terms in the squared bias part.

B.2.2 Asymptotic Expansion of Variance $V_1(\boldsymbol{x}_0;\Theta)$ in (B.9)

In this section, we focus on the nonasymptotic expansion of $V_1(\boldsymbol{x}_0;\Theta)$ by deriving its leading term and the order of remainders in (B.18) and (B.19).

Recall that

$$\begin{aligned} V_1(\boldsymbol{x}_0;\Theta) &= \mathrm{Var}_{\boldsymbol{Z}|\Theta}(\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))) \\ &= \mathbb{E}_{\boldsymbol{Z}|\Theta}[(\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))^2] - (\widetilde{\mu}(\boldsymbol{x}_0;\Theta))^2, \end{aligned} \tag{B.16}$$

in which the expansion of $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$ has been derived in the last section. Hence, it suffices to focus on the first squared term.

Since $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$ and noises $\varepsilon_1,\ldots,\varepsilon_n$ are mutually independent, we can deduce that

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z}|\Theta}[(\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))^2] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\left(\frac{1}{N_{\boldsymbol{x}_0}(\Theta)}\sum_{i=1}^n(\mu(\boldsymbol{X}_i)+\varepsilon_i)I\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right)^2\right] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{i,j=1}^n(\mu(\boldsymbol{X}_i)+\varepsilon_i)(\mu(\boldsymbol{X}_j)+\varepsilon_j)I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{i=1}^n(\mu(\boldsymbol{X}_i)^2+\varepsilon_i^2)I\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ &+\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{1\le i\ne j\le n}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{i=1}^n(\mu(\boldsymbol{X}_i)^2+\sigma^2(\boldsymbol{X}_i))I\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ &+\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{1\le i\ne j\le n}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ =:&M_1+M_2. \end{aligned}$$

Applying a similar technique as in the proof of (B.14), we immediately find that the first term satisfies

$$\begin{aligned} M_1 &= \mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{i=1}^n(\mu(\boldsymbol{X}_i)^2+\sigma^2(\boldsymbol{X}_i))I\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ &=\sum_{k=1}^n\binom{n}{k}\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\left(\frac{1}{k^2}\sum_{i=1}^k(\mu(\boldsymbol{X}_i)^2+\sigma^2(\boldsymbol{X}_i)\right)I\{\boldsymbol{X}_{1:k}\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta);\boldsymbol{X}_{(k+1):n}\notin\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right] \\ &=\mathbb{E}_{\boldsymbol{X}}[\mu^2+\sigma^2|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]\sum_{k=1}^n\frac{1}{k}\binom{n}{k}p(\boldsymbol{x}_0;\Theta)^k(1-p(\boldsymbol{x}_0;\Theta))^{n-k} \\ &=\mathbb{E}_{\boldsymbol{X}}[\mu^2+\sigma^2|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{I\{N_{\boldsymbol{x}_0}(\Theta)\ge 1\}}{N_{\boldsymbol{x}_0}(\Theta)}\right], \end{aligned}$$

where $p(\boldsymbol{x}_0;\Theta) = \mathbb{P}_{\boldsymbol{X}|\Theta}(\boldsymbol{X}\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))$ and $N_{\boldsymbol{x}_0}(\Theta)\sim\mathcal{B}(n,p(\boldsymbol{x}_0;\Theta))$ given Θ.

Meanwhile, for the second term, we can also show that

$$\begin{aligned}
M_2 =& \mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)^2}\sum_{1\le i\neq j\le n}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right]\\
=&\sum_{k=1}^{n}\binom{n}{k}\frac{k(k-1)}{k^2}\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\mu(\boldsymbol{X}_1)\mu(\boldsymbol{X}_2)I\{\boldsymbol{X}_{1:k}\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta);\boldsymbol{X}_{(k+1):n}\notin\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\}\right]\\
=&(\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)])^2\sum_{k=1}^{n}\frac{k-1}{k}\binom{n}{k}p(\boldsymbol{x}_0;\Theta)^k(1-p(\boldsymbol{x}_0;\Theta))^{n-k}\\
=&(\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)])^2\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{(N_{\boldsymbol{x}_0}(\Theta)-1)I\{N_{\boldsymbol{x}_0}(\Theta)\ge 1\}}{N_{\boldsymbol{x}_0}(\Theta)}\right]\\
=&(\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)])^2(1-(1-p(\boldsymbol{x}_0;\Theta))^n)\\
&-(\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)])^2\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{I\{N_{\boldsymbol{x}_0}(\Theta)\ge 1\}}{N_{\boldsymbol{x}_0}(\Theta)}\right].
\end{aligned}$$

Substituting the results above into (B.16), we can obtain that

$$\begin{aligned}
V_1(\boldsymbol{x}_0;\Theta) =& \left\{\mathrm{Var}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]+\mathbb{E}_{\boldsymbol{X}}[\sigma^2|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]\right\}\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{I\{N_{\boldsymbol{x}_0}(\Theta)\ge 1\}}{N_{\boldsymbol{x}_0}(\Theta)}\right]\\
&+(\mathbb{E}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)])^2(1-(1-p(\boldsymbol{x}_0;\Theta))^n)(1-p(\boldsymbol{x}_0;\Theta))^n.
\end{aligned}\tag{B.17}$$

To further expand $V_1(\boldsymbol{x}_0;\Theta)$, we need to deal with term

$$\mathbb{E}_{\boldsymbol{Z}|\Theta}\left[\frac{I\{N_{\boldsymbol{x}_0}(\Theta)\ge 1\}}{N_{\boldsymbol{x}_0}(\Theta)}\right].$$

The lemma below provides a *sharp* approximation of this inverse moment of a binomial variable, and its proof is contained in Section D.1.

Lemma 24 *Assume that $N\sim\mathcal{B}(n,p)$ with $p>0$. Then we have*

$$\left|\mathbb{E}\left\{\frac{I\{N\ge 1\}}{N}\right\}-\frac{1}{np}\right|\le C\left(1+\frac{1}{np}\right)\frac{1}{(1+(n-1)p)^2},$$

in which C is a positive constant independent of n and p.

Combining the discussion above and the definition in (B.3), we conclude that

$$\begin{aligned}
V_1(\boldsymbol{x}_0;\Theta) &= \frac{\mathrm{Var}_{\boldsymbol{X}}[\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]+\mathbb{E}_{\boldsymbol{X}}[\sigma^2|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)]}{np(\boldsymbol{x}_0;\Theta)}+\mathcal{R}_{V1}(\boldsymbol{x}_0;\Theta)\\
&= \frac{1}{n}Q_{\mu}(\boldsymbol{P}(\Theta);\boldsymbol{x}_0)+\frac{1}{n}Q_{\sigma^2}(\boldsymbol{P}(\Theta);\boldsymbol{x}_0)+\mathcal{R}_{V_1}(\boldsymbol{x}_0;\Theta),
\end{aligned}\tag{B.18}$$

where the remainder satisfies

$$\begin{aligned}
\mathcal{R}_{V1}(\boldsymbol{x}_0;\Theta)\lesssim&(\|\mu\|_\infty+\|\sigma^2\|_\infty)\\
&\times\left(\frac{1}{(1+(n-1)p(\boldsymbol{x}_0;\Theta))^2}+(1-p(\boldsymbol{x}_0;\Theta))^n\right)\left(1+\frac{1}{np(\boldsymbol{x}_0;\Theta)}\right)
\end{aligned}\tag{B.19}$$

with the positive constant in $\lesssim$ independent of μ, σ^2, $\boldsymbol{P}$, Θ, and n.

Back to (B.13), we can immediately deduce that

$$\begin{aligned}\mathbb{E}_\Theta(V_1(\boldsymbol{x}_0;\Theta)) = \frac{1}{n}\mathbb{E}_\Theta\left\{Q_\mu(\boldsymbol{P}(\Theta);\boldsymbol{x}_0) + Q_{\sigma^2}(\boldsymbol{P}(\Theta);\boldsymbol{x}_0)\right\} \\ + (\|\mu\|_\infty + \|\sigma^2\|_\infty)^2 \times O\Bigg(\mathbb{E}_\Theta\left(\frac{1}{np(\boldsymbol{x}_0;\Theta)(1+(n-1)p(\boldsymbol{x}_0;\Theta))}\right) \\ + \mathbb{E}_\Theta((1-p(\boldsymbol{x}_0;\Theta))^n)\Bigg).\end{aligned} \tag{B.20}$$

B.2.3 Asymptotic Expansion of Covariance $V_2(\boldsymbol{x}_0;\Theta,\Theta')$ in (B.10)

Recall that $V_2(\boldsymbol{x};\Theta,\Theta')$ is a covariance function and thus possesses the decomposition

$$\begin{aligned}&V_2(\boldsymbol{x}_0;\Theta,\Theta') = \mathrm{Cov}_{\boldsymbol{Z}|\Theta,\Theta'}[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta)),\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta'))] \\ &=\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta'))] - \widetilde{\mu}(\boldsymbol{x}_0;\Theta)\widetilde{\mu}(\boldsymbol{x}_0;\Theta').\end{aligned} \tag{B.21}$$

Since we have already derived in (B.14) the nonasymptotic expansion of $\widetilde{\mu}(\boldsymbol{x}_0;\Theta)$ in the second term above for the product of expectations, it suffices to focus on the first term for the expectation of products. To this end, observe that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}[\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta'))] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)N_{\boldsymbol{x}_0}(\Theta')}\sum_{i,j=1}^n Y_iY_jI\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta),\boldsymbol{X}_j\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta')\}\right] \\ =&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)N_{\boldsymbol{x}_0}(\Theta')}\sum_{i=1}^n Y_i^2I\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\cap\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta')\}\right] \\ +&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{N_{\boldsymbol{x}_0}(\Theta)N_{\boldsymbol{x}_0}(\Theta')}\sum_{1\le i\ne j\le n} Y_iY_jI\{\boldsymbol{X}_i\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta),\boldsymbol{X}_j\in\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta')\}\right].\end{aligned}$$

To further elucidate the structure of the equations, we introduce some additional notation for simplicity. Without ambiguity, let us denote $\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)$ and $N_{\boldsymbol{x}_0}(\Theta)$ by P and N, respectively, and likewise, $\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta')$ and $N_{\boldsymbol{x}_0}(\Theta')$ by P' and N', respectively. Next, define $P_0 = P\cap P'$, $P_1 = P\setminus P'$, and $P_2 = P'\setminus P$. For $j=0,1,2$, let $N_j = \sum_{i=1}^n I\{\boldsymbol{X}_i\in P_j\}$ be the number of observations in cell P_j. Clearly, given Θ and Θ', it holds that $N_j\sim\mathcal{B}(n,p_j)$, where $p_j = \mathbb{P}_{\boldsymbol{X}|\Theta,\Theta'}(\boldsymbol{X}\in P_j)$ for $j=0,1,2$. Moreover, we notice that $N = N_0+N_1$ and $N' = N_0+N_2$.

The first term in Y_i^2's can be further simplified as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i=1}^{n}Y_i^2 I\{\boldsymbol{X}_i\in P_0\}\right]\\
=&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i)^2+\varepsilon_i^2) I\{\boldsymbol{X}_i\in P_0\}\right]\\
=&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i)^2+\sigma^2(\boldsymbol{X}_i)) I\{\boldsymbol{X}_i\in P_0\}\right]\\
=&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{(N_0+N_1)(N_0+N_2)}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i)^2+\sigma^2(\boldsymbol{X}_i)) I\{\boldsymbol{X}_i\in P_0\}\right]\\
=&\mathbb{E}_{\boldsymbol{X}}[\mu^2+\sigma^2|P_0]\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_0}{(N_0+N_1)(N_0+N_2)}\right].
\end{aligned}\tag{B.22}$$

The second term in the cross-product Y_iY_j's can be further decomposed according to the different configurations of the two distinct observations $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ across the disjoint parts P_0, P_1, and P_2

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}Y_iY_j I\{\boldsymbol{X}_i\in P,\boldsymbol{X}_j\in P'\}\right]\\
=&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j) I\{\boldsymbol{X}_i\in P,\boldsymbol{X}_j\in P'\}\right]\\
=&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j) I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in P_0\}\right]\\
&+\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j) I\{\boldsymbol{X}_i\in P_0,\boldsymbol{X}_j\in P_2\}\right]\\
&+\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j) I\{\boldsymbol{X}_i\in P_1,\boldsymbol{X}_j\in P_0\}\right]\\
&+\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j) I\{\boldsymbol{X}_i\in P_1,\boldsymbol{X}_j\in P_2\}\right]\\
=:&I_1+I_2+I_3+I_4,
\end{aligned}$$

in which

$$
\begin{aligned}
I_1 &= \mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in P_0\}\right] \\
&= (\mathbb{E}_{\boldsymbol{X}}[\mu|P_0])^2\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_0(N_0-1)}{(N_0+N_1)(N_0+N_2)}\right], \\
I_2 &= \mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i\in P_0,\boldsymbol{X}_j\in P_2\}\right] \\
&= (\mathbb{E}_{\boldsymbol{X}}[\mu|P_0])(\mathbb{E}_{\boldsymbol{X}}[\mu|P_2])\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_0N_2}{(N_0+N_1)(N_0+N_2)}\right], \\
I_3 &= \mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i\in P_1,\boldsymbol{X}_j\in P_0\}\right] \\
&= (\mathbb{E}_{\boldsymbol{X}}[\mu|P_1])(\mathbb{E}_{\boldsymbol{X}}[\mu|P_0])\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_1N_0}{(N_0+N_1)(N_0+N_2)}\right], \\
I_4 &= \mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{1}{NN'}\sum_{i\neq j}\mu(\boldsymbol{X}_i)\mu(\boldsymbol{X}_j)I\{\boldsymbol{X}_i\in P_1,\boldsymbol{X}_j\in P_2\}\right] \\
&= (\mathbb{E}_{\boldsymbol{X}}[\mu|P_1])(\mathbb{E}_{\boldsymbol{X}}[\mu|P_2])\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_1N_2}{(N_0+N_1)(N_0+N_2)}\right].
\end{aligned}
$$

Combining the decompositions above, we can deduce that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}[\widehat{\mu}_{\text{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta))\widehat{\mu}_{\text{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta'))] \\
=&(\mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_0)+\text{Var}_{\boldsymbol{X}}(\mu|P_0))\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_0}{(N_0+N_1)(N_0+N_2)}\right] \\
+&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\left(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)\frac{N_0}{N_0+N_1}+\mathbb{E}_{\boldsymbol{X}}(\mu|P_1)\frac{N_1}{N_0+N_1}\right)\right. \\
&\left.\times\left(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)\frac{N_0}{N_0+N_2}+\mathbb{E}_{\boldsymbol{X}}(\mu|P_2)\frac{N_2}{N_0+N_2}\right)\right].
\end{aligned}
\tag{B.23}
$$

Consequently, an application of (B.14), (B.21), and (B.23) leads to the covariance formula

$$
\begin{aligned}
V_2(\boldsymbol{x}_0;\Theta,\Theta') =&\text{Cov}_{\boldsymbol{Z}|\Theta,\Theta'}[\widehat{\mu}_{\text{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta)),\widehat{\mu}_{\text{part}}(\boldsymbol{x}_0;\boldsymbol{Z},\boldsymbol{P}(\Theta'))] \\
=&(\mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_0)+\text{Var}_{\boldsymbol{X}}(\mu|P_0))\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\frac{N_0}{(N_0+N_1)(N_0+N_2)}\right] \\
+&\mathbb{E}_{\boldsymbol{Z}|\Theta,\Theta'}\left[\left(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)\frac{N_0}{N_0+N_1}+\mathbb{E}_{\boldsymbol{X}}(\mu|P_1)\frac{N_1}{N_0+N_1}\right)\right. \\
&\left.\times\left(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)\frac{N_0}{N_0+N_2}+\mathbb{E}_{\boldsymbol{X}}(\mu|P_2)\frac{N_2}{N_0+N_2}\right)\right] \\
-&(\mathbb{E}_{\boldsymbol{X}}(\mu|P))(\mathbb{E}_{\boldsymbol{X}}(\mu|P'))\left(1-(1-p)^n\right)\left(1-(1-p')^n\right),
\end{aligned}
\tag{B.24}
$$

where $p = \mathbb{P}_{\boldsymbol{X}|\Theta}(\boldsymbol{X} \in P)$ and $p' = \mathbb{P}_{\boldsymbol{X}|\Theta'}(\boldsymbol{X} \in P')$.

Lemma 25 *Assume that we have an i.i.d. sample $\boldsymbol{Z} = \{\boldsymbol{X}_i\}_{1\le i\le n}$ from a population $\boldsymbol{X}$ in $\mathcal{X}^d$. Let P and P' be two non-empty cells in $\mathcal{X}^d$ with non-empty intersection $P_0 = P \cap P' \neq \varnothing$. Denote by $N = \sum_{i=1}^n I\{\boldsymbol{X}_i \in P\}$ with $p = \mathbb{P}(\boldsymbol{X} \in P)$, and similarly, N' and p' for P'. In addition, define $P_1 = P \setminus P'$ and $P_2 = P' \setminus P$. Let $N_j = \sum_{i=1}^n I\{\boldsymbol{X}_i \in P_j\}$ and $p_j = \mathbb{P}(\boldsymbol{X} \in P_j)$. Then when $p_0 > 0$, it holds that*

(1)

$$\mathbb{E}\left[\frac{N_0}{NN'}\right] = \frac{p_0}{npp'} + R_{21}, \tag{B.25}$$

where the remainder satisfies

$$R_{21} \lesssim \frac{p_0}{npp'}\left(\frac{1}{1+(n-1)p} + \frac{1}{1+(n-1)p'}\right)$$

with the constant C in $\lesssim$ independent of cells P, P', and the population distribution of $\boldsymbol{X}$.

(2) For any real numbers α, β, and γ, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\alpha\frac{N_0}{N} + \beta\frac{N_1}{N}\right)\left(\alpha\frac{N_0}{N'} + \gamma\frac{N_1}{N'}\right)\right] \\
&= \left(\alpha\frac{p_0}{p} + \beta\frac{p_1}{p}\right)\left(\alpha\frac{p_0}{p'} + \gamma\frac{p_2}{p'}\right) \\
&+(\alpha-\beta)(\alpha-\gamma)\left(\frac{p_1p_2}{pp'}\right)\left(\frac{p_0}{npp'}\right) + R_{22},
\end{aligned} \tag{B.26}$$

where the remainder R_{22} satisfies

$$\begin{aligned}
R_{22} \lesssim \max\{|\alpha|,|\beta|,|\gamma|\}^2 \Bigg\{ &\frac{p_0}{npp'}\left\{\frac{1}{1+(n-1)p} + \frac{1}{1+(n-1)p'}\right\} \\
&+ \frac{1}{(1+(n-1)p)^{3/2}} + \frac{1}{(1+(n-1)p')^{3/2}} \\
&+ (1-p)^n + (1-p')^n \Bigg\}.
\end{aligned}$$

The proof of Lemma 25 above is provided in Section D.2. Substituting (B.25) and (B.26) into (B.24), by setting $\alpha = \mathbb{E}_{\boldsymbol{X}}(\mu|P_0)$, $\beta = \mathbb{E}_{\boldsymbol{X}}(\mu|P_1)$, and $\gamma = \mathbb{E}_{\boldsymbol{X}}(\mu|P_3)$ we can deduce

that

$$\begin{aligned}V_2(\boldsymbol{x}_0;\Theta,\Theta') =& (\mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_0)+\mathrm{Var}_{\boldsymbol{X}}(\mu|P_0))\frac{p_0}{npp'}\\ &+(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)-\mathbb{E}_{\boldsymbol{X}}(\mu|P_1))(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)-\mathbb{E}_{\boldsymbol{X}}(\mu|P_2))\frac{p_1p_2}{pp'}\frac{p_0}{npp'}\\ &+(\|\mu\|_\infty+\|\sigma^2\|_\infty)^2\times O\bigg(\frac{p_0}{npp'}\bigg\{\frac{1}{1+(n-1)p}+\frac{1}{1+(n-1)p'}\bigg\}\\ &\quad+\frac{1}{(1+(n-1)p)^{3/2}}+\frac{1}{(1+(n-1)p')^{3/2}}\\ &\quad+(1-p)^n-(1-p')^n\bigg).\end{aligned}$$

Note that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'))|P_0)\\ =&\mathrm{Var}_{\boldsymbol{X}}(\mu|P_0)+(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)-\mathbb{E}_{\boldsymbol{X}}(\mu|P))(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)-\mathbb{E}_{\boldsymbol{X}}(\mu|P'))\end{aligned}\tag{B.27}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}}(\mu|P)&=\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)\frac{p_0}{p}+\mathbb{E}_{\boldsymbol{X}}(\mu|P_1)\frac{p_1}{p},\\ \mathbb{E}_{\boldsymbol{X}}(\mu|P')&=\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)\frac{p_0}{p'}+\mathbb{E}_{\boldsymbol{X}}(\mu|P_2)\frac{p_2}{p'}.\end{aligned}$$

We can then obtain the decomposition formula

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'))|P_0)\\ &=\mathrm{Var}_{\boldsymbol{X}}(\mu|P_0)\\ &+(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)-\mathbb{E}_{\boldsymbol{X}}(\mu|P_1))(\mathbb{E}_{\boldsymbol{X}}(\mu|P_0)-\mathbb{E}_{\boldsymbol{X}}(\mu|P_2))\frac{p_1p_2}{pp'},\end{aligned}$$

which along with (B.2) entails that

$$\begin{aligned}V_2(\boldsymbol{x}_0;\Theta,\Theta') =& \mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_0)\frac{p_0}{npp'}+\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'))|P_0)\frac{p_0}{npp'}\\ &+(\|\mu\|_\infty+\|\sigma^2\|_\infty)^2O\bigg(\frac{p_0}{npp'}\bigg\{\frac{1}{1+(n-1)p}+\frac{1}{1+(n-1)p'}\bigg\}\\ &\quad+\frac{1}{(1+(n-1)p)^{3/2}}+\frac{1}{(1+(n-1)p')^{3/2}}+(1-p)^n+(1-p')^n\bigg)\\ =&\frac{1}{n}\left\{K_\mu(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta');\boldsymbol{x}_0)+K_{\sigma^2}(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta');\boldsymbol{x}_0)\right\}\\ &+(\|\mu\|_\infty+\|\sigma^2\|_\infty)^2\times O\bigg(\frac{p_0}{npp'}\bigg\{\frac{1}{1+(n-1)p}+\frac{1}{1+(n-1)p'}\bigg\}\\ &\quad+\frac{1}{(1+(n-1)p)^{3/2}}+\frac{1}{(1+(n-1)p')^{3/2}}+(1-p)^n+(1-p')^n\bigg).\end{aligned}$$

Finally, by integrating with respect to the exogenous factors Θ and Θ', we can conclude that

$$\begin{aligned}\mathbb{E}_{\Theta,\Theta'}(V_2(\boldsymbol{x}_0;\Theta,\Theta')) = \frac{1}{n}\mathbb{E}_{\Theta,\Theta'}\Big\{ & K_\mu(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta');\boldsymbol{x}_0) \\ & + K_{\sigma^2}(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta');\boldsymbol{x}_0)\Big\} \\ & + (\|\mu\|_\infty + \|\sigma^2\|_\infty)^2 \\ & \times O\Bigg(\mathbb{E}_{\Theta,\Theta'}\left(\frac{p(\boldsymbol{x}_0;\Theta,\Theta')}{np(\boldsymbol{x}_0;\Theta)p(\boldsymbol{x}_0;\Theta')(1+(n-1)p(\boldsymbol{x}_0;\Theta))}\right) \\ & + \mathbb{E}_\Theta\left(\frac{1}{(1+(n-1)p(\boldsymbol{x}_0;\Theta))^{3/2}}\right) \\ & + \mathbb{E}_\Theta((1-p(\boldsymbol{x}_0;\Theta))^n)\Bigg)\end{aligned} \tag{B.28}$$

with $p(\boldsymbol{x}_0;\Theta,\Theta') = \mathbb{P}_{\boldsymbol{X}}(\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)\cap\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta'))$.

The bound in (B.6) can be derived directly by a combination of (B.13), (B.15), (B.20), and (B.28). This completes the proof of Theorem 23.

B.3 Proof of Proposition 22

Observe that for each $P_i \in P$, it holds that $\mu_P I_{P_i} = \mathbb{E}_{\boldsymbol{X}}(\mu|P_i)I_{P_i}$, and similarly, for each $P' \in P'$, it holds that $\mu_{P'} I_{P'_j} = \mathbb{E}_{\boldsymbol{X}}(\mu|P'_j)I_{P'_j}$. Thus it follows that

$$(\mu-\mu_P)I_{P_{\boldsymbol{x}_0}} = (\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_{\boldsymbol{x}_0}))I_{P_{\boldsymbol{x}_0}}, \quad (\mu-\mu_{P'})I_{P'_{\boldsymbol{x}_0}} = (\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P'_{\boldsymbol{x}_0}))I_{P'_{\boldsymbol{x}_0}}.$$

Meanwhile, by the Cauchy–Schwarz inequality, we have

$$\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_P)(\mu-\mu_{P'})I_{P_{\boldsymbol{x}_0}}I_{P'_{\boldsymbol{x}_0}}) \le \sqrt{\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_P)^2 I_{P_{\boldsymbol{x}_0}})\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_{P'})^2 I_{P'_{\boldsymbol{x}_0}})}.$$

By the definition of the signal-induced local cross-partition covariance function in (B.2), it holds that

$$\begin{aligned}K_\mu(P,P';\boldsymbol{x}_0) &= \frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_P)(\mu-\mu_{P'})I_{P_{\boldsymbol{x}_0}\cap P'_{\boldsymbol{x}_0}})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})} = \frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_P)(\mu-\mu_{P'})I_{P_{\boldsymbol{x}_0}}I_{P'_{\boldsymbol{x}_0}})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})} \\ &\le \sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_P)^2|P_{\boldsymbol{x}_0})\mathbb{E}_{\boldsymbol{X}}((\mu-\mu_{P'})^2|P_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})}} \\ &= \sqrt{Q_\mu(P;\boldsymbol{x}_0)Q_\mu(P';\boldsymbol{x}_0)}.\end{aligned}$$

The equality above holds if and only if $(\mu-\mu_P)I_{P_{\boldsymbol{x}_0}} = (\mu-\mu_{P'})I_{P'_{\boldsymbol{x}_0}}$ almost surely under $\mathbb{P}_{\boldsymbol{X}}$. We next discuss the sufficient and necessary condition for obtaining such equality.

On the one hand, since $\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0}\cap P'_{\boldsymbol{x}_0}) > 0$, by Assumption 4 we immediately see that $\mathbb{E}_{\boldsymbol{X}}(\mu|P_{\boldsymbol{x}_0}) = \mathbb{E}_{\boldsymbol{X}}(\mu|P'_{\boldsymbol{x}_0})$ because $\mu-\mu_P$ and $\mu-\mu_{P'}$ are identical on $P_{\boldsymbol{x}_0}\cap P'_{\boldsymbol{x}_0}$. On the other hand, the equation also implies that $\mu - \mathbb{E}(\mu|P_{\boldsymbol{x}_0}) = 0$ on $P_{\boldsymbol{x}_0}\backslash P'_{\boldsymbol{x}_0}$ and $P'_{\boldsymbol{x}_0}\backslash P_{\boldsymbol{x}_0}$. When $P_{\boldsymbol{x}_0}$ and $P'_{\boldsymbol{x}_0}$ are indistinguishable, μ must be a constant on $P_{\boldsymbol{x}_0}\cup P_{\boldsymbol{x}_0}$.

We now move on to the second part of the proof. Note that by the Cauchy–Schwarz inequality,

$$\mathbb{E}_{\boldsymbol{X}}(\sigma^2 I_{P_{\boldsymbol{x}_0}} I_{P'_{\boldsymbol{x}_0}}) \le \sqrt{\mathbb{E}_{\boldsymbol{X}}(\sigma^2 I_{P_{\boldsymbol{x}_0}})\mathbb{E}_{\boldsymbol{X}}(\sigma^2 I_{P'_{\boldsymbol{x}_0}})}.$$

Then it follows from the definition of the model-error-induced local cross-partition covariance function in (B.2) that

$$\begin{aligned} K_{\sigma^2}(P, P'; \boldsymbol{x}_0) &= \frac{\mathbb{E}_{\boldsymbol{X}}(\sigma^2 I_{P_{\boldsymbol{x}_0} \cap P'_{\boldsymbol{x}_0}})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})} = \frac{\mathbb{E}_{\boldsymbol{X}}(\sigma^2 I_{P_{\boldsymbol{x}_0}} I_{P'_{\boldsymbol{x}_0}})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})} \\ &\le \sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}(\sigma^2 | P_{\boldsymbol{x}_0})\mathbb{E}_{\boldsymbol{X}}(\sigma^2 | P_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})}} \\ &= \sqrt{Q_{\sigma^2}(P; \boldsymbol{x}_0) Q_{\sigma^2}(P'; \boldsymbol{x}_0)}. \end{aligned}$$

The equality above holds if and only if $\sigma^2 I_{P_{\boldsymbol{x}_0}} = \sigma^2 I_{P'_{\boldsymbol{x}_0}}$ almost surely under $\mathbb{P}_{\boldsymbol{X}}$, which is equivalent to the indistinguishability of $P_{\boldsymbol{x}_0}$ and $P'_{\boldsymbol{x}_0}$. This concludes the proof of Proposition 22.

Appendix C. Proofs of Proposition 3, Theorems 4, 5, 8, and 15, and Corollaries 11 and 19

This section contains the detailed proofs of theorems in the main body. The proofs are presented in the following order: Proposition 3 first and then Theorems 4 and 5, next Theorem 8 with Corollary 11, and finally Theorem 15 with Corollary 19.

C.1 Proof of Proposition 3

According to the definitions of Φ_μ and Φ_{σ^2}, it is straightforward to verify that K_μ and K_{σ^2} are generated by these two feature mappings, respectively. In what follows, we provide an alternative way to prove the triangle inequality for K_μ and K_{σ^2}, by which the condition for equality can be derived easily.

Let us recall that

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{X}}((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))(\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))|P_i \cap P'_j) \frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j)^2}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)} \\ =&\mathbb{E}_{\boldsymbol{X}}((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))(\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))I_{P_i \cap P'_j}) \frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)} \\ \le&\sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i)\mathbb{E}_{\boldsymbol{X}}((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))^2|P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}} \mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j). \end{aligned}$$

Meanwhile, it holds that

$$\begin{aligned} Q_\mu(P) &= \sum_{P_i \in P} \mathrm{Var}(\mu|P_i) = \sum_{P_i \in P} \mathbb{E}((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i) \\ &= \mathbb{E}_{\boldsymbol{X}'} \left\{ \sum_{P_i \in P} \sqrt{\frac{\mathbb{E}((\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i)}{\mathbb{P}_{\boldsymbol{X}}(P_i)}} I\{\boldsymbol{X}' \in P_i\} \right\}^2, \end{aligned}$$

where $\boldsymbol{X'}$ is an independent copy of $\boldsymbol{X}$.

Then it follows that

$$\begin{aligned}
K_\mu(P,P') &\le \sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i)\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))^2|P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}}\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j).\\
&= \mathbb{E}_{\boldsymbol{X'}}\Bigg\{\left(\sum_{P_i\in P}\sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i)}{\mathbb{P}_{\boldsymbol{X}}(P_i)}}I\{\boldsymbol{X'}\in P_i\}\right)\\
&\qquad\times\left(\sum_{P'_j\in P'}\sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))^2|P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P'_j)}}I\{\boldsymbol{X'}\in P'_j\}\right)\Bigg\}\\
&\le \sqrt{\mathbb{E}_{\boldsymbol{X'}}\left\{\left(\sum_{P_i\in P}\sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i)}{\mathbb{P}_{\boldsymbol{X}}(P_i)}}I\{\boldsymbol{X'}\in P_i\}\right)^2\right\}}\\
&\times\sqrt{\mathbb{E}_{\boldsymbol{X'}}\left\{\left(\sum_{P'_j\in P'}\sqrt{\frac{\mathbb{E}_{\boldsymbol{X}}((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))^2|Q')}{\mathbb{P}_{\boldsymbol{X}}(P'_j)}}I\{\boldsymbol{X'}\in P'_j\}\right)^2\right\}}\\
&\le \sqrt{Q_\mu(P)Q_\mu(P')},
\end{aligned}$$

which proves the triangle inequality for the kernel function K_μ. The equality above holds if and only if for any $P_i\in P$, $P'_j\in P'$ with $P_i\cap P'_j\neq\varnothing$, $(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))I_{P_i}=(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))I_{P'_j}$. According to the discussion in Section B.3, we know that this holds if and only if μ is constant on $P_i\cup P'_j$ or P_i and P'_j are indistinguishable.

We next show the triangle inequality for the kernel function K_{σ^2}. Since $\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)\le\min\{\mathbb{P}_{\boldsymbol{X}}(P_i),\mathbb{P}_{\boldsymbol{X}}(P'_j)\}$, it holds that

$$\begin{aligned}
K_{\sigma^2}(P,P') &= \sum_{P_i\in P}\frac{1}{\mathbb{P}_{\boldsymbol{X}}(P_i)}\left(\sum_{P'_j\in P'}\mathbb{E}_{\boldsymbol{X}}[\sigma^2|P_i\cap P'_j]\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)^2}{\mathbb{P}_{\boldsymbol{X}}(P'_j)}\right)\\
&\le \sum_{P_i\in P}\frac{1}{\mathbb{P}_{\boldsymbol{X}}(P_i)}\left(\sum_{P'_j\in P'}\mathbb{E}_{\boldsymbol{X}}[\sigma^2 I_{P_i\cap P'_j}]\right)\\
&= \sum_{P_i\in P}\frac{\mathbb{E}_{\boldsymbol{X}}(\sigma^2 I_{P_i})}{\mathbb{P}_{\boldsymbol{X}}(P_i)} = \sum_{P_i\in P}\mathbb{E}_{\boldsymbol{X}}[\sigma^2|P_i] = Q_{\sigma^2}(P).
\end{aligned}$$

Similarly, we also have $K_{\sigma^2}(P,P')\le Q_{\sigma^2}(P')$ and thus $K_{\sigma^2}(P,P')\le\sqrt{Q_{\sigma^2}(P)Q_{\sigma^2}(P')}$. The equality above holds if and only if

$$\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P'_j)}=\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)}=1$$

for any P_i and P'_j, which entails that P and P' are indistinguishable. This completes the proof of Proposition 3.

C.2 Proof of Theorem 4

The global MSE can be derived by integrating the local MSE with respect to the test data $\boldsymbol{X'}$ independent of the training data $\boldsymbol{Z}$, that is,

$$\mathrm{MSE}(\widehat{\mu}_{\mathrm{ens}}) = \mathbb{E}_{\boldsymbol{X'}}[\mathrm{MSE}(\widehat{\mu}_{\mathrm{ens}}; \boldsymbol{X'})].$$

Still, we omit the superscript n from $\boldsymbol{P}^n$ with no ambiguity. Recall that given a partition P, the corresponding partitioning estimator $\widehat{\mu}_{\mathrm{part}}(\cdot; \boldsymbol{Z}, P)$ is defined by

$$\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0; \boldsymbol{Z}, P) = \frac{1}{N(P_{\boldsymbol{x}_0})} \sum_{i=1}^{n} Y_i I\{\boldsymbol{X}_i \in P_{\boldsymbol{x}_0}\}.$$

However, this representation is inconvenient in analyzing the global MSE since the target point and the cell in partition are mixed up in $\boldsymbol{P}_{\boldsymbol{x}_0}$ so it is not easy to separate the test data and the exogenous factors to exchange the order of integration. Instead, we prefer to work with an *alternative* representation

$$\widehat{\mu}_{\mathrm{part}}(\boldsymbol{x}_0; \boldsymbol{Z}, P) = \sum_{P_i \in P} \left(\frac{1}{N(P_i)} \sum_{i=1}^{n} Y_i I\{\boldsymbol{X}_i \in P_i\} \right) I\{\boldsymbol{x}_0 \in P_i\}. \tag{C.29}$$

Clearly, the form in (C.29) above is more convenient in exchanging the order of integration since the information of the target point $\boldsymbol{x}_0$ is separated by an indicator function $I\{\boldsymbol{x}_0 \in P_0\}$.

C.2.1 Terms in The Squared Bias Part

Let us first deal with the two terms in the squared bias part. Observe that

$$\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))) = \sum_{P_i \in \boldsymbol{P}(\Theta)} \mathbb{E}_{\boldsymbol{X}}(\mu|P_i) I\{\boldsymbol{x}_0 \in P_i\} = \mu_{\boldsymbol{P}(\Theta)}(\boldsymbol{x}_0).$$

Hence, we can show that

$$\mathbb{E}_{\boldsymbol{X'},\Theta}[(\mu(\boldsymbol{X'}) - \mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{X'}}(\Theta)))^2] = \mathbb{E}_{\boldsymbol{X},\Theta}[(\mu - \mu_{\boldsymbol{P}(\Theta)})^2]$$

and

$$\mathbb{E}_{\boldsymbol{X'}}\left[\mathrm{Var}_{\Theta}(\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{X'}}(\Theta)))\right] = \mathbb{E}_{\boldsymbol{X'}}\left[\mathrm{Var}_{\Theta}(\mu_{\boldsymbol{P}(\Theta)}(\boldsymbol{X'}))\right].$$

C.2.2 Terms in The Variance Part

Next, we focus on the two terms in the variance part. Since the local cross-partition covariance function is related to cells $P_{\boldsymbol{x}_0}$ and $P'_{\boldsymbol{x}_0}$, we need to change to another form of

the representation of these functions. The *new* representation is given by

$$K_\mu(P,P';\boldsymbol{x}_0)=\sum_{\substack{P_i\in P,\\ P'_j\in P'}}\Big\{\mathbb{E}_{\boldsymbol{X}}\left((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))|P_i\cap P'_j\right)\\ \times\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}I\{\boldsymbol{x}_0\in P_i\cap P'_j\}\Big\},$$

$$K_{\sigma^2}(P,P';\boldsymbol{x}_0)=\sum_{\substack{P_i\in P,\\ P'_j\in P'}}\mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_i\cap P'_j)\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}I\{\boldsymbol{x}_0\in P_i\cap P'_j\}.$$

In particular, when $P=P'$, the representations for the Q-functions are given by

$$Q_\mu(P;\boldsymbol{x}_0)=\sum_{P_i\in P}\mathbb{E}_{\boldsymbol{X}}\left((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))^2|P_i\right)\frac{1}{\mathbb{P}_{\boldsymbol{X}}(P_i)}I\{\boldsymbol{x}_0\in P_i\},$$

$$Q_{\sigma^2}(P;\boldsymbol{x}_0)=\sum_{P_i\in P}\mathbb{E}_{\boldsymbol{X}}(\sigma^2|P_i)\frac{1}{\mathbb{P}_{\boldsymbol{X}}(P_i)}I\{\boldsymbol{x}_0\in P_i\cap P'_j\}.$$

Substituting $\boldsymbol{X'}$ into the expression above and taking the expectation, we can deduce that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X'},\Theta,\Theta'}\left\{K_\mu(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta');\boldsymbol{X'})\right\}\\
=&\mathbb{E}_{\boldsymbol{X'},\Theta,\Theta'}\left\{\sum_{P_i\in P,P'_j\in P'}\left\{\mathbb{E}_{\boldsymbol{X}}\left((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))|P_i\cap P'_j\right)\right.\right.\\
&\left.\left.\times\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}I\{\boldsymbol{X'}\in P_i\cap P'_j\}\right\}\right\}\\
=&\mathbb{E}_{\Theta,\Theta'}\left\{\sum_{P_i\in P,P'_j\in P'}\left\{\mathbb{E}_{\boldsymbol{X}}\left((\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P_i))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|P'_j))|P_i\cap P'_j\right)\right.\right.\\
&\left.\left.\times\frac{\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}\mathbb{P}_{\boldsymbol{X}}(P_i\cap P'_j)\right\}\right\}\\
=&\mathbb{E}_{\Theta,\Theta'}\left\{K_\mu(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta'))\right\}.
\end{aligned}$$

Applying a similar technique, we can also obtain that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X'},\Theta,\Theta'}\left\{K_{\sigma^2}(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta');\boldsymbol{X'})\right\}&=\mathbb{E}_{\Theta,\Theta'}\left\{K_{\sigma^2}(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta'))\right\},\\
\mathbb{E}_{\boldsymbol{X'},\Theta}\left\{Q_\mu(\boldsymbol{P}(\Theta);\boldsymbol{X'})\right\}&=\mathbb{E}_{\Theta,\Theta'}\left\{Q_\mu(\boldsymbol{P}(\Theta))\right\},\\
\mathbb{E}_{\boldsymbol{X'},\Theta}\left\{Q_{\sigma^2}(\boldsymbol{P}(\Theta);\boldsymbol{X'})\right\}&=\mathbb{E}_{\Theta,\Theta'}\left\{Q_{\sigma^2}(\boldsymbol{P}(\Theta))\right\}.
\end{aligned}$$

Finally, the remainders can be derived similarly by combining an alternative representation separating $\boldsymbol{X'}$ from the cell of partitions and the exchange of integrations. We omit

the details to save space here and provide the final results

$$\mathcal{R}_{\text{ens}}(\mu) \lesssim (\|\mu\|_\infty + \|\sigma^2\|_\infty)^2 \times \Bigg\{ \mathbb{E}_\Theta \left(\sum_{P_i \in \boldsymbol{P}(\Theta)} (1 - \mathbb{P}_{\boldsymbol{X}}(P_i))^n \mathbb{P}_{\boldsymbol{X}}(P_i) \right) + \frac{B-1}{B} \mathbb{E}_\Theta \left(\frac{1}{n} \sum_{P_i \in \boldsymbol{P}(\Theta)} \frac{1}{\sqrt{1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)}} \right) + \mathbb{E}_\Theta \left(\frac{1}{n} \sum_{P_i \in \boldsymbol{P}(\Theta)} \frac{1}{1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)} \right) \Bigg\}.$$

Recall that $(1-x)^n \le e^{-x}$ and $e^{-x}x \le \frac{1}{1+x}$ hold for all $x \in [0, \infty)$. The first term on the right-hand side above satisfies

$$\begin{aligned} \mathbb{E}_\Theta \left(\sum_{P_i \in \boldsymbol{P}(\Theta)} (1 - \mathbb{P}_{\boldsymbol{X}}(P_i))^n \mathbb{P}_{\boldsymbol{X}}(P_i) \right) &\lesssim \mathbb{E}_\Theta \left(\frac{1}{n} \sum_{P_i \in \boldsymbol{P}(\Theta)} e^{-n\mathbb{P}_{\boldsymbol{X}}(P_i)} n \mathbb{P}_{\boldsymbol{X}}(P_i) \right) \\ &\lesssim \mathbb{E}_\Theta \left(\frac{1}{n} \sum_{P_i \in \boldsymbol{P}(\Theta)} \frac{1}{1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)} \right). \end{aligned}$$

In view of this observation, we see that when $B \ge 2$, the second term above leads the remainder and it follows that

$$\mathcal{R}_{\text{ens}}(\mu) \lesssim (\|\mu\|_\infty + \|\sigma^2\|_\infty + 1)^2 \times \mathbb{E}_\Theta \left(\frac{1}{n} \sum_{P_i \in \boldsymbol{P}(\Theta)} \frac{1}{\sqrt{1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)}} \right).$$

In particular, when $B = 1$, the second term in $\mathcal{R}_{\text{ens}}$ vanishes and thus we can obtain that

$$\mathcal{R}_{\text{part}} \lesssim (\|\mu\|_\infty + \|\sigma^2\|_\infty + 1)^2 \times \mathbb{E}_\Theta \left(\frac{1}{n} \sum_{P_i \in \boldsymbol{P}(\Theta)} \frac{1}{1 + (n-1)\mathbb{P}_{\boldsymbol{X}}(P_i)} \right).$$

This concludes the proof of Theorem 4.

C.3 Proof of Theorem 5

According to the dominated convergence theorem, conditions (14) and $n\mathbb{P}_{\boldsymbol{X}}(P_i^n) \to +\infty$ imply that all the remainder terms in (15) tend to zero as $n \to \infty$. Hence, the weak consistency holds *if and only if* the leading terms tend to zero.

It is clear that condition $\mathbb{E}_{\boldsymbol{X},\Theta}[(\mu - \mu_{\boldsymbol{P}^n(\Theta)})^2] \to 0$ in (21) entails the convergence of the second term of the squared bias parts in the first line of (15) with $B = 1$. Similarly, condition $\mathbb{E}_{\boldsymbol{X}}[(\mu - \mathbb{E}_\Theta(\mu_{\boldsymbol{P}^n(\Theta)}))^2] \to 0$ in (22) together with $B \to \infty$ entails the convergence of the first term of the squared bias parts in the first line of (15). Thus, it remains to consider the convergence for the variance parts.

Observe that μ and σ^2 are upper bounded so that we have

$$K_\mu(P, P') \le \|\mu\|_\infty^2 K(P, P'), \quad K_{\sigma^2}(P, P') \le \|\sigma^2\|_\infty K(P, P').$$

Meanwhile, since σ^2 is lower bounded by σ_0^2, it holds that

$$K_{\sigma^2}(P, P') \ge \sigma_0^2 K(P, P').$$

It then follows that

$$\begin{aligned}\sigma_0^2 \mathbb{E}_\Theta[Q(\boldsymbol{P}(\Theta))] &\le \mathbb{E}_\Theta[Q_\mu(\boldsymbol{P}(\Theta)) + Q_{\sigma^2}(\boldsymbol{P}(\Theta))] \\ &\le (\|\mu\|_\infty^2 + \|\sigma^2\|_\infty)\mathbb{E}_\Theta[Q(\boldsymbol{P}(\Theta))],\end{aligned}$$

since both Q_μ and Q_{σ^2} are nonnegative. In light of $Q(\boldsymbol{P}(\Theta)) = |\boldsymbol{P}(\Theta)|$, we can immediately conclude that the variance part in the third line of (15) converges to zero if and only if condition $n^{-1}\mathbb{E}_\Theta[|\boldsymbol{P}^n(\Theta)|] \to 0$ in (21) holds.

If in addition $\mathbb{E}_{\Theta,\Theta'}[K_\mu(\boldsymbol{P}(\Theta), \boldsymbol{P}(\Theta'))] \ge 0$, we also have

$$\begin{aligned}\sigma_0^2 \mathbb{E}_{\Theta,\Theta'}[K(\boldsymbol{P}(\Theta), \boldsymbol{P}(\Theta'))] &\le \mathbb{E}_{\Theta,\Theta'}[K_\mu(\boldsymbol{P}(\Theta), \boldsymbol{P}(\Theta')) + K_{\sigma^2}(\boldsymbol{P}(\Theta), \boldsymbol{P}(\Theta'))] \\ &\le (\|\mu\|_\infty^2 + \|\sigma^2\|_\infty)\mathbb{E}_{\Theta,\Theta'}[K(\boldsymbol{P}(\Theta), \boldsymbol{P}(\Theta'))].\end{aligned}$$

The equivalence of the convergence of the variance part for the ensemble estimator in (15) and condition $n^{-1}\mathbb{E}_{\Theta,\Theta'}[\mathrm{Cov}(\boldsymbol{P}^n(\Theta), \boldsymbol{P}^n(\Theta'))] \to 0$ (22) immediately follows. Hence, it suffices to show that $\mathbb{E}_{\Theta,\Theta'}[K_\mu(\boldsymbol{P}(\Theta), \boldsymbol{P}(\Theta'))] \ge 0$.

For any $\theta \in \mathcal{D}^n$, denote by $\boldsymbol{P}(\theta) = \{P_r^\theta : r = 1, \ldots, R_\theta\}$. From the definition of the global cross-partition covariance function in (8), it holds that

$$K_\mu(\boldsymbol{P}(\theta), \boldsymbol{P}(\theta')) = \sum_{r=1}^{R_\theta}\sum_{s=1}^{R_{\theta'}} \frac{\mathbb{E}_{\boldsymbol{X}}\left((\mu - \mu_{\boldsymbol{P}(\theta)})(\mu - \mu_{\boldsymbol{P}(\theta')})I_{P_r^\theta} I_{P_s^{\theta'}}\right)\mathbb{P}_{\boldsymbol{X}}(P_r^\theta \cap P_s^{\theta'})}{\mathbb{P}_{\boldsymbol{X}}(P_r^\theta)\mathbb{P}_{\boldsymbol{X}}(P_s^{\theta'})}.$$

Let us define random vectors

$$\boldsymbol{\alpha}(\boldsymbol{X}; \theta) = \left(\frac{(\mu(\boldsymbol{X}) - \mu_{\boldsymbol{P}(\theta)}(\boldsymbol{X}))I_{P_r^\theta}(\boldsymbol{X})}{\sqrt{\mathbb{P}_{\boldsymbol{X}}(P_r^\theta)}} : r = 1, \ldots, R_\theta\right)^\top,$$

$$\boldsymbol{\beta}(\boldsymbol{X}; \theta) = \left(\frac{I_{P_r^\theta}(\boldsymbol{X})}{\sqrt{\mathbb{P}_{\boldsymbol{X}}(P_r^\theta)}} : r = 1, \ldots, R_\theta\right)^\top.$$

Further, we introduce the matrix-valued bivariate functions

$$\begin{aligned}\boldsymbol{A}(\theta, \theta') &= \mathbb{E}_{\boldsymbol{X}}\left(\boldsymbol{\alpha}(\boldsymbol{X}; \theta)\boldsymbol{\alpha}(\boldsymbol{X}; \theta')^\top\right) \\ &= \left(\frac{\mathbb{E}_{\boldsymbol{X}}\left((\mu - \mu_{\boldsymbol{P}(\theta)})(\mu - \mu_{\boldsymbol{P}(\theta')})I_{P_r^\theta \cap P_s^{\theta'}}\right)}{\sqrt{\mathbb{P}_{\boldsymbol{X}}(P_r^\theta)\mathbb{P}_{\boldsymbol{X}}(P_s^{\theta'})}} : r = 1, \ldots, R_\theta; s = 1, \ldots, R_{\theta'}\right), \\ \boldsymbol{B}(\theta, \theta') &= \mathbb{E}_{\boldsymbol{X}}\left(\boldsymbol{\beta}(\boldsymbol{X}; \theta)\boldsymbol{\beta}(\boldsymbol{X}; \theta')^\top\right) \\ &= \left(\frac{\mathbb{P}_{\boldsymbol{X}}(P_r^\theta \cap P_s^{\theta'})}{\sqrt{\mathbb{P}_{\boldsymbol{X}}(P_r^\theta)\mathbb{P}_{\boldsymbol{X}}(P_s^{\theta'})}} : r = 1, \ldots, R_\theta; s = 1, \ldots, R_{\theta'}\right),\end{aligned}$$

where $\boldsymbol{A}(\theta,\theta') = \boldsymbol{A}(\theta',\theta)^\top$ and $\boldsymbol{B}(\theta,\theta') = \boldsymbol{B}(\theta',\theta)^\top$. Then we see that

$$K_\mu(\boldsymbol{P}(\theta),\boldsymbol{P}(\theta')) = \mathrm{tr}(\boldsymbol{A}(\theta,\theta')\boldsymbol{B}(\theta',\theta)),$$

Let $\boldsymbol{Y}$ be an independent copy of $\boldsymbol{X}$. By defining $h(\boldsymbol{X},\boldsymbol{Y};\theta) = \boldsymbol{\alpha}(\boldsymbol{X};\theta)^\top\boldsymbol{\beta}(\boldsymbol{Y};\theta)$, we can deduce that

$$\begin{aligned}
K_\mu(\boldsymbol{P}(\theta),\boldsymbol{P}(\theta')) &= \mathrm{tr}(\boldsymbol{A}(\theta,\theta')\boldsymbol{B}(\theta',\theta)) \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left\{\mathrm{tr}\left(\boldsymbol{\alpha}(\boldsymbol{X};\theta)\boldsymbol{\alpha}(\boldsymbol{X};\theta')^\top\boldsymbol{\beta}(\boldsymbol{Y};\theta')\boldsymbol{\beta}(\boldsymbol{Y};\theta)^\top\right)\right\} \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left\{(\boldsymbol{\alpha}(\boldsymbol{X};\theta)^\top\boldsymbol{\beta}(\boldsymbol{Y};\theta)(\boldsymbol{\alpha}(\boldsymbol{X};\theta')^\top\boldsymbol{\beta}(\boldsymbol{Y};\theta')\right\} \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left\{h(\boldsymbol{X},\boldsymbol{Y};\theta)h(\boldsymbol{X},\boldsymbol{Y};\theta')\right\}.
\end{aligned}$$

By Assumption 4, we see that $h(\boldsymbol{X},\boldsymbol{Y};\theta)$ is a ternary measurable function. Denote by $H(\boldsymbol{X},\boldsymbol{Y}) = \mathbb{E}_\Theta[h(\boldsymbol{X},\boldsymbol{Y};\Theta)]$. An application of Fubini's theorem yields that

$$\begin{aligned}
\mathbb{E}_{\Theta,\Theta'}[K_\mu(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta'))] &= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left\{\mathbb{E}_\Theta(h(\boldsymbol{X},\boldsymbol{Y};\Theta))\mathbb{E}_{\Theta'}(h(\boldsymbol{X},\boldsymbol{Y};\Theta'))\right\} \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left\{H(\boldsymbol{X},\boldsymbol{Y})^2\right\} \geq 0,
\end{aligned}$$

which verifies the non-negativeness of $\mathbb{E}_{\Theta,\Theta'}[K_\mu(\boldsymbol{P}(\Theta),\boldsymbol{P}(\Theta'))]$. This completes the proof of Theorem 5.

C.4 Proof of Theorem 8

For a given target point $\boldsymbol{x}_0$, the terminal node $\mathbf{t}_{\boldsymbol{x}_0} = \mathbf{t}(\boldsymbol{x}_0; I_l)$ containing $\boldsymbol{x}_0$ can be uniquely determined by $\boldsymbol{x}_0$ and a binary CART process, where the process is completely independent of $\boldsymbol{x}_0$. The independence of the binary CART process and $\boldsymbol{x}_0$ enables us to exchange the order of integration with respect to the binary CART process and the test data $\boldsymbol{X}'$ replacing $\boldsymbol{x}_0$.

C.4.1 Derivation of The Local MSE Formula

We first determine the local MSE for the related random forests estimator $\widehat{\mu}_{\mathrm{RF}}$ according to our general Theorem 23. We begin with the squared bias terms in (B.6). Observe that in (B.6), we have

$$\begin{aligned}
&(\mu(\boldsymbol{x}_0) - \mathbb{E}_\Theta\left(\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))\right))^2 \\
=&(\mu(\boldsymbol{x}_0) - \mathbb{E}_\Theta\left(\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta))\right)\left(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\Theta'}\left(\mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta'))\right)\right) \\
=&\mathbb{E}_{\Theta,\Theta'}\left[(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta)))\left(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\boldsymbol{P}_{\boldsymbol{x}_0}(\Theta'))\right)\right],
\end{aligned} \tag{C.30}$$

where the exogenous factors Θ and Θ' are mutually independent. Replacing $\boldsymbol{P}(\Theta)$ and $\boldsymbol{P}(\Theta')$ by two independent bi-partitions with the population CART driven by two independent binary CART processes I_l and I_l', we see that the squared bias part in the local MSE formula is given by

$$\begin{aligned}
&\text{The squared bias of tree ensemble} \\
&= \mathbb{E}_{I_l,I_l'}\left[(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)))(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(x_0;I_l')))\right].
\end{aligned}$$

Recall that for points in $\mathbf{t}(\boldsymbol{x}_0; I_l)$, coordinates x_j with $I_{lj} = 1$ are identical to x_{0j}, and the rest can take values 0 or 1. Hence, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; I_l)) &= \sum_{j:1\le j\le s, I_{lj}=1} \mu(x_{0j}) + \sum_{j:1\le j\le s, I_{lj}=0} \beta_j \mathbb{E}(X_j) \\ &= \sum_{j:1\le j\le s, I_{lj}=1} \mu(x_{0j}) + \sum_{j:1\le j\le s, I_{lj}=0} \frac{\beta_j}{2}\end{aligned}$$

and consequently,

$$\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(x_0; I_l)) = \sum_{j:1\le j\le s, I_{lj}=0} \beta_j \left(x_{0j} - \frac{1}{2}\right).$$

It then follows that

$$\begin{aligned}&\text{The squared bias of tree ensemble} \\ &= \mathbb{E}_{I_l, I_l'}\left[\left(\sum_{j:1\le j\le s, I_{lj}=0} \beta_j \left(x_{0j} - \frac{1}{2}\right)\right)\left(\sum_{j':1\le j'\le s, I'_{lj'}=0} \beta_{j'} \left(x_{0j'} - \frac{1}{2}\right)\right)\right].\end{aligned}$$

Similarly, for one single tree, we can show that

$$\begin{aligned}\text{The squared bias of a single tree} =& \mathbb{E}_{I_l}\left[(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu; \mathbf{t}(\boldsymbol{x}_0; I_l)))^2\right] \\ =& \mathbb{E}_{I_l}\left[\left(\sum_{j:1\le j\le s, I_{lj}=0} \beta_j^2 \left(x_{0j} - \frac{1}{2}\right)\right)^2\right].\end{aligned}$$

We next deal with the cross-partition covariance and the single-partition variance terms in (B.2). Note that a coordinate in the intersection of two terminal nodes is split as long as it is split by either I_l or I_l'. Given I_l and I_l', we have

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l) \cap \mathbf{t}(\boldsymbol{x}_0; I_l')) = 2^{-\sum_{j=1}^d \max\{I_{lj}, I'_{lj}\}}, \\ &\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l)) = 2^{-\sum_{j=1}^d I_{lj}},\ \mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l')) = 2^{-\sum_{j=1}^d I'_{lj}},\end{aligned}$$

and thus,

$$\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l) \cap \mathbf{t}(\boldsymbol{x}_0; I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l'))} = 2^{\sum_{j=1}^d I_{lj} + I'_{lj} - \max\{I_{lj}, I'_{lj}\}} = 2^{\sum_{j=1}^d \min\{I_{lj}, I'_{lj}\}}.$$

Since the model error has constant variance, in light of (B.2), for two independent bi-partitions P and P' driven by two independent binary CART processes I_l and I_l', the kernel function K_{σ^2} is given by

$$K_{\sigma^2}(P, P'; \boldsymbol{x}_0) = \sigma_0^2 \frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l) \cap \mathbf{t}(\boldsymbol{x}_0; I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l'))} = \sigma_0^2 2^{\sum_{j=1}^d \min\{I_{lj}, I'_{lj}\}}.$$

Thus, it holds that

$$Q_{\sigma^2}(P;\boldsymbol{x}_0)=\sigma_0^2\frac{1}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0,I_l))}=\sigma_0^2 2^{\sum_{j=1}^d I_{lj}}.$$

Let us proceed with analyzing the kernel function K_μ. Recall the decomposition formula (B.27) that the coefficient in (B.2) satisfies

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}}\left[(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l')))\big|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l')\right]\\&=\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))+\Big(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))\\&\quad-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l))\Big)\Big(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l'))\Big).\end{aligned}$$

Observe that

$$\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))=\frac{1}{4}\sum_{j:1\le j\le s,I_{lj}=I_{lj}'=0}\beta_j^2=\frac{1}{4}\sum_{j=1}^s\beta_j^2 I\{\max\{I_{lj},I_{lj}'\}=0\}$$

and

$$\begin{aligned}&(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)))\\&\times(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l')))\\&=\left(\sum_{j:1\le j\le s,I_{lj}'-I_{lj}=1}\beta_j\left(x_{0j}-\frac{1}{2}\right)\right)\left(\sum_{j':1\le j'\le s,I_{lj'}-I_{lj'}'=1}\beta_{j'}\left(x_{0j'}-\frac{1}{2}\right)\right).\end{aligned}$$

Hence, for two independent bi-partitions P and P' driven by I_l and I_l', we can deduce that

$$\begin{aligned}K_\mu(P,P';\boldsymbol{x}_0)&=\mathbb{E}_{\boldsymbol{X}}\left[(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l')))\big|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l')\right]\\&\quad\times\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l'))}\\&=\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l'))}\\&\quad+(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)))\\&\quad\times(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l')))\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l)\cap\mathbf{t}(\boldsymbol{x}_0;I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l'))}\\&=\left(\frac{1}{4}\sum_{j=1}^s\beta_j^2 I\{\max\{I_{lj},I_{lj}'\}=0\}\right)2^{\sum_{j=1}^d\min\{I_{lj},I_{lj}'\}}\\&\quad+\left(\sum_{j:1\le j\le s,I_{lj}'-I_{lj}=1}\beta_j\left(x_{0j}-\frac{1}{2}\right)\right)\left(\sum_{j':1\le j'\le s,I_{lj'}-I_{lj'}'=1}\beta_{j'}\left(x_{0j'}-\frac{1}{2}\right)\right)\\&\quad\times 2^{\sum_{j=1}^d\min\{I_{lj},I_{lj}'\}}.\end{aligned}$$

Meanwhile, we also have

$$\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l))=\frac{1}{4}\sum_{j:1\le j\le s,I_{lj}=0}\beta_j^2=\frac{1}{4}\sum_{j=1}^{s}\beta_j^2 I\{I_{lj}=0\}.$$

Thus, from (B.3) it holds that

$$Q_\mu(P;\boldsymbol{x}_0)=\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;I_l))/\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l))=\left(\frac{1}{4}\sum_{j=1}^{s}\beta_j^2 I\{I_{lj}=0\}\right)2^{\sum_{j=1}^{d}I_{lj}}.$$

Substituting the equations derived above into (B.2) and (B.6), we can finally obtain that

$$\begin{aligned}
&\mathrm{MSE}(\widehat{\mu}_{RF};\boldsymbol{x}_0)\\
&=\frac{B-1}{B}\mathbb{E}_{I_l,I_l'}\left[\left(\sum_{j:1\le j\le s,I_{lj}=0}\beta_j\left(\frac{1}{2}-x_{0j}\right)\right)\left(\sum_{j':1\le j'\le s,I'_{lj'}=0}\beta_{j'}\left(\frac{1}{2}-x_{0j'}\right)\right)\right]\\
&+\frac{1}{B}\mathbb{E}_{I_l}\left[\left(\sum_{j:1\le j\le s,I_{lj}=0}\beta_j\left(\frac{1}{2}-x_{0j}\right)\right)^2\right]\\
&+\frac{B-1}{Bn}\mathbb{E}_{I_l,I_l'}\left[\left(\sigma^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2 I\{\max\{I_{lj},I'_{lj}\}=0\}\right)2^{\sum_{j=1}^{d}\min\{I_{lj},I'_{lj}\}}\right]\\
&+\frac{B-1}{Bn}\mathbb{E}_{I_l,I_l'}\left[\left(\sum_{j:1\le j\le s,I'_{lj}-I_{lj}=1}\beta_j\left(x_{0j}-\frac{1}{2}\right)\right)\right.\\
&\left.\times\left(\sum_{j':1\le j'\le s,I_{lj'}-I'_{lj'}=1}\beta_{j'}\left(x_{0j'}-\frac{1}{2}\right)\right)2^{\sum_{j=1}^{d}\min\{I_{lj},I'_{lj}\}}\right]\\
&+\frac{1}{Bn}\mathbb{E}_{I_l}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2 I\{I_{lj}=0\}\right)2^{\sum_{j=1}^{d}I_{lj}}\right]+\mathcal{R}_{RF}(\boldsymbol{x}_0),
\end{aligned}$$

in which the remainder satisfies

$$\mathcal{R}_{RF}(\boldsymbol{x}_0)\lesssim(1-2^{-l})^n+\frac{2^l}{n(1+(n-1)2^{-l})^{1/2}}.$$

C.4.2 Derivation of (24) and (25)

To further derive the global MSE, we replace $\boldsymbol{x}_0$ in $\mathrm{MSE}(\widehat{\mu}_{RF};\boldsymbol{x}_0)$ with a random vector $\boldsymbol{X}'$ independent of I_l and I_l'. Notice that

$$\{j:1\le j\le s,I_{lj}-I_{l'j}=1\}\cap\{j':1\le j'\le s,I_{l'j'}-I_{lj'}=1\}=\varnothing.$$

Due to the component independence of $\boldsymbol{X}'$, it holds that

$$\mathbb{E}_{\boldsymbol{X}'|I_l,I_l}\left[\left(\sum_{j:1\le j\le s, I'_{lj}-I_{lj}=1}\beta_j\left(X'_{0j}-\frac{1}{2}\right)\right)\left(\sum_{j':1\le j'\le s, I_{lj'}-I'_{lj'}=1}\beta_{j'}\left(X'_{0j'}-\frac{1}{2}\right)\right)\right]=0.$$

Hence, the fifth term in the local MSE formula will vanish after replacing $\boldsymbol{x}_0$ with $\boldsymbol{X}'$ and integrating out. Further, we have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}'|I_l,I'_l}\left[\left(\sum_{j:1\le j\le s, I_{lj}=0}\beta_j\left(\frac{1}{2}-X'_{0j}\right)\right)\left(\sum_{j':1\le j'\le s, I'_{lj'}=0}\beta_{j'}\left(\frac{1}{2}-X'_{0j'}\right)\right)\right]\\ =&\sum_{j:1\le j\le s, I_{lj}=I'_{lj}=0}\beta_j^2\mathrm{Var}(X'_j)=\frac{1}{4}\sum_{j=1}^{s}I\{\max\{I_{lj},I'_{lj}\}=0\}.\end{aligned}$$

Consequently, by exchanging the order of integration with respect to $\boldsymbol{X}'$ and binary CART processes in the local MSE formula, we can deduce that

$$\begin{aligned}\mathrm{MSE}(\widehat{\mu}_{\mathrm{RF}})=&\frac{B-1}{4B}\mathbb{E}\left[\sum_{j=1}^{s}\beta_j^2(1-\max\{I_{lj},I'_{lj}\})\right]+\frac{1}{4B}\mathbb{E}\left[\sum_{j=1}^{s}\beta_j^2(1-I_{lj})\right]\\&+\frac{B-1}{B}\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-\max\{I_{lj},I'_{lj}\})\right)\frac{2^{\sum_{j=1}^{d}\min\{I_{lj},I'_{lj}\}}}{n}\right]\\&+\frac{1}{B}\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2(1-I_{lj})\right)\frac{2^{\sum_{j=1}^{d}I_{lj}}}{n}\right]+\mathcal{R}_{\mathrm{RF}},\end{aligned}$$

where the bound of the remainder after integration remains the same, that is,

$$\mathcal{R}_{\mathrm{RF}}\lesssim(1-2^{-l})^n+\frac{2^l}{n(1+(n-1)2^{-l})^{1/2}}.$$

C.4.3 Derivation of Convergence Rate in (26)

In what follows, we will study the converge rate for the squared bias part in (25). Let us first observe that the variance part in (24) satisfies

$$\begin{aligned}&\frac{B-1}{B}\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2I\{\max\{I_{lj},I'_{lj}\}=0\}\right)\frac{2^{\sum_{j=1}^{d}\min\{I_{lj},I'_{lj}\}}}{n}\right]\\&+\frac{1}{B}\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2I\{I_{lj}=0\}\right)\frac{2^{\sum_{j=1}^{d}I_{lj}}}{n}\right]\\&\le\left(\sigma_0^2+\frac{1}{4}\sum_{j=1}^{s}\beta_j^2\right)\frac{2^l}{n}.\end{aligned}\tag{C.31}$$

Meanwhile, denote by $N_l = \sum_{j=1}^{s} I\{I_{lj} = 0\}$ the number of informative variables not being split by time l. It is easy to see that the squared bias part in (24) satisfies

$$\frac{B-1}{4B}\sum_{j=1}^{s}\beta_j^2\mathbb{E}[I\{\max\{I_{lj}, I'_{lj}\} = 0\}] + \frac{1}{4B}\sum_{j=1}^{s}\beta_j^2\mathbb{E}[I\{I_{lj} = 0\}] \le \left(\max_{1\le j\le s}\frac{\beta_j^2}{4}\right)\mathbb{E}[N_l].$$

Hence, it suffices to analyze the convergence rate for $\mathbb{E}[N_l]$.

Note that the process $\{N_k\}_{0\le k\le l}$ is a *homogenous absorbing Markov chain* on the state space $E = \{0, 1, \ldots, s\}$ with transition probability matrix

$$Q = (q_{ij}) = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ s-1 \\ s \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & \ldots & s-1 & s \end{array} \\ \begin{pmatrix} 1 & & & & \\ 1-q_1 & q_1 & & & \\ & \ddots & \ddots & & \\ & & & q_{s-1} & \\ & & & 1-q_s & q_s \end{pmatrix} \end{array},$$

in which

$$q_i = \frac{\binom{d-i}{\lceil \gamma d\rceil}}{\binom{d}{\lceil \gamma d\rceil}} = \left(1-\frac{i}{d}\right)\cdots\left(1-\frac{i}{d-\lceil \gamma d\rceil+1}\right)$$

represents the probability that the next split is on non-informative variables, given that there are i informative variables that have not been split yet. It then follows that

$$\mathbb{E}(N_{l+1}|N_l = i) = iq_i + (i-1)(1-q_i) = i - (1-q_i).$$

Recall that function $W_{\gamma,d}(x)$ in (27) is decreasing as x increases and its derivative satisfies

$$W'_{\gamma,d}(x) = -W_{\gamma,d}(x)\left(\sum_{k=0}^{\lceil \gamma d\rceil-1}\frac{1}{d-j-x}\right). \tag{C.32}$$

According to Taylor's theorem, for each $0 \le i \le s$ there exists some $\theta \in (0,1)$ such that

$$\begin{aligned} 1 - q_i = W_{\gamma,d}(0) - W_{\gamma,d}(i) &= -W'_{\gamma,d}(\theta i)i \\ &= iW(\theta i)\left(\sum_{k=0}^{\lceil \gamma d\rceil-1}\frac{1}{d-j}\right) \\ &\ge iW_{\gamma,d}(s)\frac{\lceil \gamma d\rceil}{d} = iW_{\gamma,d}(s)\gamma\frac{3}{4}. \end{aligned}$$

From such observation, it follows that

$$\mathbb{E}(N_{l+1}|N_l = i) \le i(1-\frac{3}{4}\gamma W_{\gamma,d}(s)),$$

and as such, since $N_0 \equiv s$ we have

$$\mathbb{E}(N_l) \leq \left(1 - \frac{3}{4}\gamma W_{\gamma,d}(s)\right)^{l+1} s.$$

Combining the results above, we can finally obtain that

$$\begin{aligned}
&\max\{\mathrm{MSE}(\widehat{\mu}_{RF}), \mathrm{MSE}(\widehat{\mu}_{\mathrm{tree}})\} \\
&\lesssim \left(\max_{1\leq j\leq s} \frac{\beta_j^2}{4}\right)\left(1 - \frac{3}{4}\gamma W(s)\right)^{l+1} s + \left(\sigma_0^2 + \frac{1}{4}\sum_{j=1}^{s}\beta_j^2\right)\frac{2^l}{n},
\end{aligned}$$

which completes the proof.

C.4.4 Derivation of The Cross-Tree Covariance Formula in Remark 10

In view of the definition in (17), the variance function of the partition generated by a single binary CART process is the total number of terminal nodes. Notice that at depth l, we have a total of $\sum_{j=1}^{d} I_{lj}$ splits, which means that the total number of nodes is $2^{\sum_{j=1}^{d} I_{lj}}$. Hence, it holds that

$$\mathbb{E}[Q(P)] = \mathbb{E}[2^{\sum_{j=1}^{d} I_{lj}}].$$

To derive the formula for function $K(P, P')$, we first make some general claims. For any partitions P and P', given a target point $\boldsymbol{x}_0$ let us define

$$K(P, P'; \boldsymbol{x}_0) = \frac{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0} \cap P'_{\boldsymbol{x}_0})}{\mathbb{P}_{\boldsymbol{X}}(P_{\boldsymbol{x}_0})\mathbb{P}_{\boldsymbol{X}}(P'_{\boldsymbol{x}_0})}, \tag{C.33}$$

which is identical to the local function (B.2) with $\sigma \equiv 1$. According to the definition in (16), it is clear that

$$K(P, P') = \mathbb{E}_{\boldsymbol{X}'}\left[K(P, P'; \boldsymbol{X}')\right], \tag{C.34}$$

where $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$. Indeed, from (C.33) it follows that

$$K(P, P'; \boldsymbol{x}_0) = \sum_{P_i \in P}\sum_{P'_j \in P'} I\{\boldsymbol{x}_0 \in P_i \cap P'_j\}\frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}.$$

Replacing $\boldsymbol{x}_0$ with $\boldsymbol{X}'$ and then integrating with respect to it, we can deduce that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}'}\left[K(P, P'; \boldsymbol{X}')\right] &= \mathbb{E}_{\boldsymbol{X}'}\left[\sum_{P_i \in P}\sum_{P'_j \in P'} I\{\boldsymbol{X}' \in P_i \cap P'_j\}\frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}\right] \\
&= \sum_{P_i \in P}\sum_{P'_j \in P'} \frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j)}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)}\mathbb{P}_{\boldsymbol{X}'}(\boldsymbol{X}' \in P_i \cap P_j) \\
&= \sum_{P_i \in P}\sum_{P'_j \in P'} \frac{\mathbb{P}_{\boldsymbol{X}}(P_i \cap P'_j)^2}{\mathbb{P}_{\boldsymbol{X}}(P_i)\mathbb{P}_{\boldsymbol{X}}(P'_j)} = K(P, P').
\end{aligned}$$

For any nodes, say, P_i and P_j' from P and P' with non-empty intersection, there exist certain $\boldsymbol{x}_0$ and two independent binary CART processes I_l and I_l' such that $P_i = \mathbf{t}(\boldsymbol{x}_0; I_l)$ and $P_j' = \mathbf{t}(x_0; I_l')$. Note that a coordinate in the intersection of two nodes is split as long as it is split by either I_l or I_l'. Given I_l and I_l', it holds that

$$\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l) \cap \mathbf{t}(\boldsymbol{x}_0; I_l')) = 2^{-\sum_{j=1}^d \max\{I_{lj}, I_{lj}'\}}$$

and

$$\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l)) = 2^{-\sum_{j=1}^d I_{lj}}, \quad \mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l')) = 2^{-\sum_{j=1}^d I_{lj}'}.$$

Consequently, for the ratio we have

$$\begin{aligned} K(P, P'; \boldsymbol{x}_0) = \frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l) \cap \mathbf{t}(\boldsymbol{x}_0; I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l'))} &= 2^{\sum_{j=1}^d I_{lj} + I_{lj}' - \max\{I_{lj}, I_{lj}'\}} \\ &= 2^{\sum_{j=1}^d \min\{I_{lj}, I_{lj}'\}}, \end{aligned}$$

which is constant for any $\boldsymbol{x}_0$. Hence, by (C.34) we can obtain that

$$K(P, P') = \mathbb{E}_{\boldsymbol{X'}}\left[K(P, P'; \boldsymbol{X'})\right] = 2^{\sum_{j=1}^d \min\{I_{lj}, I_{lj}'\}}.$$

Therefore, the final conclusion follows by integrating with respect to I_l and I_l'. This completes the proof of Theorem 8.

C.5 Proof of Corollary 11

In light of the definition of the binary CART process $(I_t)_{t\geq 0}$, it holds that

$$I_t = I_{t-1} + A_t, \quad I_0 \equiv 0,$$

where $A_t \in \{e_1, \ldots, e_d\}$ selects *one* previously unsplit coordinate e_j per step

$$\mathbb{P}(A_t = e_j \mid I_{t-1}) \simeq \begin{cases} 0 & \text{if } I_{t-1,j} = 1, \\ \gamma & \text{if } I_{t-1,1} = 0 \text{ and } j = 1, \\ \dfrac{1-\gamma}{\#\{k \neq 1 : \ I_{t-1,k} = 0\}} & \text{if } I_{t-1,1} = 0 \text{ and } j \neq 1, \\ \dfrac{1}{\#\{k : \ I_{t-1,k} = 0\}} & \text{if } I_{t-1,1} = 1 \text{ and } I_{t-1,j} = 0. \end{cases}$$

After l steps, we have $I_{l,j} \in \{0, 1\}$ and $\sum_{j=1}^d I_{l,j} = l$ (assuming $1 \leq l \leq d$). Let us write

$$S := \{j : I_{l,j} = 1\}, \quad S' := \{j : I_{l,j}' = 1\}$$

for two independent copies (two trees). Denote by $K = |S \cap S'|$.

In (24), we can find that the leading terms for the squared bias part are

$$\begin{aligned} \mathbb{E}[1 - I_{l1}] &= \mathbb{P}(I_{l1} = 0) = (1-\gamma)^l, \\ \mathbb{E}[1 - \max\{I_{l1}, I_{l1}'\}] &= \mathbb{P}(I_{lj} = I_{lj}' = 0) \\ &= \mathbb{P}(I_{lj} = 0)^2 = (1-\gamma)^{2l}, \end{aligned}$$

and the leading terms for the variance part are

$$\begin{aligned}\mathbb{E}[2^{\sum_{j=1}^{d} I_{lj}}] &= 2^l,\\ \mathbb{E}[2^{\sum_{j=1}^{d} \min\{I_{lj}, I'_{lj}\}}] &= \mathbb{E}[2^{\#\{j: I_{lj}=I'_{lj}=1\}}]\\ &= \mathbb{E}[2^{|S\cap S'|}] = \mathbb{E}[2^K],\end{aligned}$$

in which the quantity $\rho_{l,d}(\gamma)$ in (30) is equal to $\mathbb{E}[2^K]$. The interaction terms in the variance part will be handled later after we derive the explicit expression of $\rho_{l,d}(\gamma)$.

To analyze term $\mathbb{E}[2^K]$, it follows from the identity

$$2^K = \prod_{j\in[d]} (1 + I\{j \in S \cap S'\}) = \sum_{T\subseteq[d]} I\{T \subseteq S \cap S'\}$$

and the independence of S and S' that

$$\mathbb{E}[2^K] = \sum_{T\subseteq[d]} \big(\mathbb{P}(T \subseteq S)\big)^2.$$

Note that only subsets T with $|T| \le l$ will be shown in the summation since $|S| = |S'| = l$. Next, we group by $t = |T|$ and whether $1 \in T$ or not to derive the expansion of $\rho_{l,d}(\gamma)$. In fact, if $1 \in T$ and $1 \le |T| = t \le l$, it holds that

$$\mathbb{P}(T \subseteq S) = P_1 \cdot \frac{\binom{d-t}{l-t}}{\binom{d-1}{l-1}}$$

by conditioning on including 1 and then choosing the remaining $l-1$ uniformly from $d-1$ coordinates. If $1 \notin T$ and $0 \le |T| = t \le l$, we split by whether $1 \in S$

$$\mathbb{P}(T \subseteq S) = P_1 \frac{\binom{d-1-t}{l-1-t}}{\binom{d-1}{l-1}} + (1-P_1)\frac{\binom{d-1-t}{l-t}}{\binom{d-1}{l}}.$$

Counting the number of T with/without 1 yields that

$$\begin{aligned}\rho_{l,d}(\gamma) = \mathbb{E}[2^K] &= \sum_{t=1}^{l} \binom{d-1}{t-1} \left(P_1 \frac{\binom{d-t}{l-t}}{\binom{d-1}{l-1}}\right)^2\\ &+ \sum_{t=0}^{l} \binom{d-1}{t} \left(P_1 \frac{\binom{d-1-t}{l-1-t}}{\binom{d-1}{l-1}} + (1-P_1)\frac{\binom{d-1-t}{l-t}}{\binom{d-1}{l}}\right)^2\\ &= A_{l,d}P_1^2 + 2B_{l,d}P_1(1-P_1) + C_{l,d}(1-P_1)^2,\end{aligned}$$

where

$$\begin{aligned}A_{l,d} &= 2\sum_{t=1}^{l} \binom{d-1}{t-1} \left(\frac{\binom{d-t}{l-t}}{\binom{d-1}{l-1}}\right)^2,\\ B_{l,d} &= \sum_{t=0}^{l} \binom{d-1}{t} \frac{\binom{d-1-t}{l-1-t}}{\binom{d-1}{l-1}} \frac{\binom{d-1-t}{l-t}}{\binom{d-1}{l}},\\ C_{l,d} &= \sum_{t=0}^{l} \binom{d-1}{t} \left(\frac{\binom{d-1-t}{l-t}}{\binom{d-1}{l}}\right)^2.\end{aligned}$$

Fix integers $d \geq 2$ and $l \geq 1$. Recall the following binomial identities

$$\binom{d-1}{r} = \frac{(d-1)_r}{r!}, \qquad \frac{\binom{d-1-r}{l-1-r}}{\binom{d-1}{l-1}} = \frac{(l-1)_r}{(d-1)_r}, \qquad \frac{\binom{d-1-r}{l-r}}{\binom{d-1}{l}} = \frac{(l)_r}{(d-1)_r}. \tag{C.35}$$

Denote by $A_r := \frac{(l-1)_r^2}{(d-1)_r\, r!}$. Then we can easily deduce that

$$A_{l,d} = 2\sum_{r=0}^{l-1} A_r, \qquad B_{l,d} = \sum_{r=0}^{l-1} \left(\frac{l}{l-r}\right) A_r, \qquad C_{l,d} = \sum_{r=0}^{l-1} \left(\frac{l}{l-r}\right)^2 A_r. \tag{C.36}$$

Hence, combining (C.35) and (C.36), we can obtain equation (30).

Back to the MSE formula (24), we notice that the interaction terms in the variance part

$$\begin{aligned}
\mathbb{E}\left[(1-I_{l1})2^{\sum_{j=1}^d I_{lj}}\right] &= (1-\gamma)^l 2^l, \\
\mathbb{E}\left[\left(1-\max\{I_{l1}, I'_{l1}\}\right) 2^{\sum_{j=1}^d \min\{I_{lj}, I'_{lj}\}}\right] &= \mathbb{E}[I\{1 \notin S \cap S'\}2^K] \\
&= \mathbb{E}[2^K]\frac{(1-P_1)^2 C_{l,d}}{\mathbb{E}[2^K]} \\
&\lesssim \mathbb{E}[2^K](1-\gamma)^{2l}
\end{aligned}$$

are negligible compared to the corresponding leading terms, where the last inequality above holds since $C_{l,d} \leq 2A_{l,d}$. Consequently, by substituting all the calculations above into (24), we immediately find

$$\begin{aligned}
\mathrm{MSE}(\widehat{\mu}_{\mathrm{RF}}) &\lesssim \frac{B-1}{4B}\beta_1^2 \mathbb{E}[1-\max\{I_{l1}, I'_{l1}\}] + \frac{1}{4B}\beta_1^2 \mathbb{E}[1-I_{l1}] \\
&\quad + \frac{B-1}{B}\sigma_0^2 \frac{\mathbb{E}[2^{\sum_{j=1}^d \min\{I_{lj}, I'_{lj}\}}]}{n} + \frac{1}{B}\sigma_0^2 \frac{2^l}{n} + \mathcal{R}'_{\mathrm{RF}} \\
&= \frac{B-1}{4B}\beta_1^2 (1-\gamma)^{2l} + \frac{1}{4B}\beta_1^2 (1-\gamma)^l \\
&\quad + \frac{B-1}{B}\sigma_0^2 \frac{\mathbb{E}[2^K]}{n} + \frac{1}{B}\sigma_0^2 \frac{2^l}{n} + \mathcal{R}'_{\mathrm{RF}},
\end{aligned}$$

where

$$\mathcal{R}'_{\mathrm{RF}} \lesssim (1-\gamma)^l \frac{2^l}{n} + \left(\frac{2^l}{n}\right)^{3/2} + (1-2^{-l})^n.$$

This completes the proof of (29).

It remains to study the convergence rate of $\rho_{l,d}(\gamma)$ and make comparison with 2^l as $l, d \to \infty$ with $l/d \to 0$. We recall that for any fixed $t \leq l$, it holds that

$$\frac{\binom{d-1-t}{l-1-t}}{\binom{d-1}{l-1}} = \frac{(l-1)_t}{(d-1)_t}, \qquad \frac{\binom{d-1-t}{l-t}}{\binom{d-1}{l}} = \frac{(l)_t}{(d-1)_t}, \qquad \binom{d-1}{t}(d-1)_t^{-2} = \frac{1}{t!}(d-1)_t^{-1}.$$

Let us write $a = (l-1)/(d-1)$ and $P_1 = 1-(1-\gamma)^l$ with fixed $\gamma \in (0,1)$. Since $1-P_1 = (1-\gamma)^l \to 0$ exponentially in l, by some standard calculations using the three

identities above, we can find the formula in (30)

$$\rho_{l,d}(\gamma) = 2P_1^2 \sum_{t=0}^{l-1} \binom{d-1}{t} \left(\frac{(l-1)_t}{(d-1)_t} \right)^2 + o((1+a^2)^{d-1}).$$

For the main sum on the right-hand side above, it follows from the uniform ratio approximation that

$$\frac{(l-1)_t}{(d-1)_t} = a^t\big(1+\varepsilon_t\big), \qquad |\varepsilon_t| \le C\,\frac{t^2}{d} \quad (0 \le t \le l-1),$$

with C an uniform constant independent of d, l and t, and thus

$$\begin{aligned}\sum_{t=0}^{l-1} \binom{d-1}{t} \left(\frac{(l-1)_t}{(d-1)_t} \right)^2 &= (1+o(1)) \sum_{t=0}^{l-1} \binom{d-1}{t} a^{2t} \\ &= (1+a^2)^{d-1}\big(1+o(1)\big).\end{aligned}$$

Combining $P_1 \to 1$ with the result above, we can finally conclude that

$$\rho_{l,d}(\gamma) = 2\,(1+a^2)^{d-1}\,(1+o(1))$$

as $l/d \to 0$. Therefore, since $\log \rho_{l,d}(\gamma) \sim (d-1)\log(1+a^2) \sim l^2/d$, we have that

$$\frac{\rho_{l,d}}{2^l} \simeq \frac{\exp\big(l^2/d\big)}{2^l} \to 0,$$

which proves the final conclusion of Corollary 11.

C.6 Proof of Theorem 15

C.6.1 Derivation of The Local MSE Formula

We first calculate the local MSE for the related random forests estimator $\widehat{\mu}_{\mathrm{RF}}$ according to our general Theorem 23. Let $\mathbf{x}_0 = (x_{01}, \ldots, x_{0d})^\top \in [0,1]^d$ be the target point. The proof here follows similar lines as in Section C.4, where we first address the squared bias part and then the variance part.

On the one hand, let us recall the representation (C.30) of the squared bias part for the tree ensemble. Then for two independent bipartitions driven by two independent uniform CART processes J_l and J_l', the squared bias part in the local MSE formula can be written as

$$\begin{aligned}&\text{The squared bias of tree ensemble} \\ =&\mathbb{E}_{J_l, J_l'}\left[(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(x_0; J_l)))(\mu(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(x_0; J_l'))) \right].\end{aligned}$$

Given a realization of a uniform CART process J_l, in light of the discussion related to (33), the terminal node containing $\boldsymbol{x}_0$ is given by

$$\mathbf{t}(\boldsymbol{x}_0, J_l) = \prod_{i=1}^{d} \left(\frac{K(x_{0i}, J_{li}) - 1}{2^{J_{li}}}, \frac{K(x_{0i}, J_{li})}{2^{J_{li}}} \right],$$

where $K(x_0, m) = \lceil x_0 2^m \rceil$. In addition, it holds that

$$\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0, J_l)) = \sum_{i=1}^{s} \beta_i M(x_{0i}, J_{li}),$$

where $M(x_{0i}, J_{li}) = (2K(x_{0i}, J_{li}) - 1)/2^{J_{li}+1}$. Then we can show that

$$\begin{aligned}&\text{The squared bias of tree ensemble}\\ =&\mathbb{E}_{J_l, J_l'}\left[\left(\sum_{i=1}^{s}\beta_i\left(x_{0i} - M(x_{0i}, J_{li})\right)\right)\left(\sum_{i'=1}^{s}\beta_{i'}\left(x_{0i'} - M(x_{0i'}, J_{li'}')\right)\right)\right].\end{aligned}$$

Similarly, it also holds that

$$\text{The squared bias of one single tree} = \mathbb{E}_{J_l}\left[\left(\sum_{i=1}^{s}\beta_i\left(x_{0i} - M(x_{0i}, J_{li})\right)\right)^2\right].$$

On the other hand, observe that for two terminal nodes $\mathbf{t}(\boldsymbol{x}_0; J_l)$ and $\mathbf{t}(\boldsymbol{x}_0; J_l')$ in P and P', the edge length of their intersection is $2^{-\max\{J_{li}, J_{li}'\}}$ for any $i = 1, \ldots, p$. It follows that

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l')) = 2^{-\sum_{i=1}^{d}\max\{J_{li}, J_{li}'\}},\\ &\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l)) = \mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l')) = 2^{-l},\end{aligned}$$

and thus,

$$\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l) \cap \mathbf{t}(\boldsymbol{x}_0; I_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; I_l'))} = 2^{\sum_{i=1}^{d}\min\{J_{li}, J_{li}'\}}.$$

For the kernel function $K_{\sigma^2}(; \boldsymbol{x}_0)$ in (B.2), since the model error has a constant variance, we have that

$$K_{\sigma^2}(P, P'; \boldsymbol{x}_0) = \sigma_0^2 \frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l'))} = \sigma_0^2 2^{\sum_{i=1}^{d}\min\{J_{li}, J_{li}'\}},$$

where P and P' are two independent bi-partitions driven by two independent binary CART processes J_l and J_l'. Thus, when $P = P'$, we have

$$Q_{\sigma^2}(P; \boldsymbol{x}_0) = \sigma_0^2 \frac{1}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0, J_l))} = \sigma_0^2 2^{\sum_{i=1}^{d} J_{li}} = \sigma_0^2 2^l.$$

As for function $K_\mu(; \boldsymbol{x}_0)$, we recall the decomposition formula (B.27) and can see that the coefficient in (B.2) satisfies

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}}\left[(\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l)))(\mu - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l')))\big|\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l')\right]\\ =&\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l'))\\ +&(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l')) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l)))\\ \times&(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l')) - \mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0; J_l'))).\end{aligned}$$

Notice that

$$\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))=\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2\max\{J_{li},J_{li}'\}}$$

and

$$\begin{aligned}
&\big(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l))\big)\\
&\quad\times\big(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l'))\big)\\
&=\left(\sum_{i:1\le i\le s,J_{li}'>J_{li}}\beta_i\left(M(x_{0i},J_{li}')-M(x_{0i},J_{li})\right)\right)\\
&\quad\times\left(\sum_{i':1\le i'\le s,J_{li'}>J_{li'}'}\beta_{i'}\left(M(x_{0i'},J_{li'})-M(x_{0i'},J_{li'}')\right)\right).
\end{aligned}$$

Hence, for two independent bi-partitions P and P' driven by J_l and J_l', we can deduce that

$$\begin{aligned}
K_\mu(P,P';\boldsymbol{x}_0)&=\mathbb{E}_{\boldsymbol{X}}\left[(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)))(\mu-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l')))\big|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l')\right]\\
&\quad\times\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l'))}\\
&=\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l'))}\\
&\quad+(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)))\\
&\quad\times(\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))-\mathbb{E}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l')))\times\frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l)\cap\mathbf{t}(\boldsymbol{x}_0;J_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;I_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l'))}\\
&=\left(\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2\max\{J_{li},J_{li}'\}}\right)2^{\sum_{i=1}^{d}\min\{J_{li},J_{li}'\}}\\
&\quad+\left(\sum_{i:1\le i\le s,J_{li}'>J_{li}}\beta_i\left(M(x_{0i},J_{li}')-M(x_{0i},J_{li})\right)\right)\\
&\quad\times\left(\sum_{i':1\le i'\le s,J_{li'}>J_{li'}'}\beta_{i'}\left(M(x_{0i'},J_{li'})-M(x_{0i'},J_{li'}')\right)\right)\times 2^{\sum_{i=1}^{d}\min\{J_{li},J_{li}'\}}.
\end{aligned}$$

Meanwhile, since

$$\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0,J_l))=\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2J_{li}},$$

it follows from (B.3) that

$$Q_\mu(P;\boldsymbol{x}_0)=\mathrm{Var}_{\boldsymbol{X}}(\mu|\mathbf{t}(\boldsymbol{x}_0;J_l))/\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0;J_l))=\left(\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2J_{li}}\right)2^{l}.$$

Substituting the equations derived above into (B.2) and (B.6), we can finally obtain that

$$\begin{aligned}
\mathrm{MSE}(\widehat{\mu}_{\mathrm{RF}};\boldsymbol{x}_0) &= \frac{B-1}{B}\mathbb{E}_{J_l,J_l'}\left[\left(\sum_{i=1}^{s}\beta_i\left(x_{0i}-M(x_{0i},J_{li})\right)\right)\left(\sum_{i'=1}^{s}\beta_{i'}\left(x_{0i'}-M(x_{0i'},J_{li'}')\right)\right)\right] \\
&\quad + \frac{1}{B}\mathbb{E}_{J_l}\left[\left(\sum_{i=1}^{s}\beta_i\left(x_{0i}-M(x_{0i},J_{li})\right)\right)^2\right] \\
&\quad + \frac{B-1}{Bn}\mathbb{E}_{J_l,J_l'}\left[\left(\sigma_0^2+\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2\max\{J_{li},J_{li}'\}}\right)2^{\sum_{i=1}^{d}\min\{J_{li},J_{li}'\}}\right] \\
&\quad + \frac{B-1}{Bn}\mathbb{E}_{J_l,J_l'}\Bigg[\left(\sum_{i:1\le i\le s,J_{li}'>J_{li}}\beta_i\left(M(x_{0i},J_{li}')-M(x_{0i},J_{li})\right)\right) \\
&\quad \times\left(\sum_{i':1\le i'\le s,J_{li'}>J_{li'}'}\beta_{i'}\left(M(x_{0i'},J_{li'})-M(x_{0i'},J_{li'}')\right)\right)2^{\sum_{i=1}^{d}\min\{J_{li},J_{li}'\}}\Bigg] \\
&\quad + \frac{1}{Bn}\mathbb{E}_{J_l}\left[\left(\sigma_0^2+\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2J_{li}}\right)2^{l}\right]+\mathcal{R}_{\mathrm{RF}}(\boldsymbol{x}_0),
\end{aligned}$$

in which the remainder satisfies

$$\mathcal{R}_{\mathrm{RF}}(\boldsymbol{x}_0)\lesssim(1-2^{-l})^n+\frac{2^l}{n(1+(n-1)2^{-l})^{1/2}}.$$

C.6.2 Derivation of (34) and (35)

We now proceed with examining the global MSE formula. Observe that the uniform CART process is independent of $\boldsymbol{x}_0$. Then by replacing x_{0i} with a uniform random variable X_i, it holds that

$$\begin{aligned}
\mathbb{E}_{X_i|J_{li}}(M(X_i,J_{li})-X_i) &= \mathbb{E}_{X_i|J_{li}}\left[\sum_{k=1}^{2^{J_{li}}}\left(\frac{2k-1}{2^{J_{li}+1}}-X_j\right)I\left\{\frac{k-1}{2^{J_{li}}}<X_j\le\frac{k}{2^{J_{li}}}\right\}\right] \\
&= \sum_{k=1}^{2^{J_{li}}}2^{-J_{li}}\left(\frac{2k-1}{2^{J_{li}+1}}-\frac{2k-1}{2^{J_{li}+1}}\right)=0, \\
\mathbb{E}_{X_i|J_{li}}[(M(X_i,J_{li})-X_i)^2] &= \mathbb{E}_{X_i|J_{li}}\left[\sum_{k=1}^{2^{J_{li}}}\left(\frac{2k-1}{2^{J_{li}+1}}-X_j\right)^2 I\left\{\frac{k-1}{2^{J_{li}}}<X_j\le\frac{k}{2^{J_{li}}}\right\}\right] \\
&= \sum_{k=1}^{2^{J_{li}}}\frac{1}{2^{J_{li}}}\frac{1}{3}\left(\frac{1}{2^{J_{li}+1}}\right)^2=\frac{1}{12}2^{-2J_{jl}}.
\end{aligned}$$

In addition, we have that

$$\{i:\ 1\le i\le s, J_{li}<J_{li}\}\cap\{i':\ 1\le i'\le s, J_{li'}>J_{li'}\}=\varnothing.$$

Thanks to the component independence of $\boldsymbol{X}'$, we can show that

$$\mathbb{E}_{\boldsymbol{X}'|J_l,J'_l}\Bigg[\left(\sum_{i:1\le i\le s,J'_{li}>J_{li}}\beta_i\left(M(X'_i,J'_{li})-M(X'_i,J_{li})\right)\right)\times\left(\sum_{i':1\le i'\le s,J_{li'}>J'_{li'}}\beta_{i'}\left(M(X'_{i'},J_{li'})-M(X'_{i'},J'_{li'})\right)\right)\Bigg]=0.$$

Meanwhile, it follows that

$$\mathbb{E}_{\boldsymbol{X}'|J_l,J'_l}\left[\left(\sum_{i=1}^{s}\beta_i\left(X_i-M(X_i,J_{li})\right)\right)\left(\sum_{i'=1}^{s}\beta_{i'}\left(X_{i'}-M(X_{i'},J'_{li'})\right)\right)\right]=\frac{1}{12}\sum_{i=1}^{s}2^{-2\max\{J_{li},J'_{li}\}}.$$

Summarizing the results derived above, we can deduce that

$$\begin{aligned}\mathrm{MSE}(\widehat{\mu}_{\mathrm{RF}})=&\frac{1}{12B}\mathbb{E}\left[\sum_{j=1}^{s}\beta_i^2 2^{-2J_{li}}\right]+\frac{B-1}{12B}\mathbb{E}\left[\sum_{i=1}^{s}\beta_i^2 2^{-2\max\{J_{li},J'_{li}\}}\right]\\&+\frac{1}{B}\mathbb{E}\left[\left(\sigma_0^2+\frac{1}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2J_{li}}\right)\frac{2^l}{n}\right]\\&+\frac{B-1}{B}\mathbb{E}\left[\left(\sigma_0^2+\frac{a^2}{12}\sum_{i=1}^{s}\beta_i^2 2^{-2\max\{J_{li},J'_{li}\}}\right)\frac{2^{\sum_{i=1}^{d}\min\{J_{li},J'_{li}\}}}{n}\right]\\&+\mathcal{R}_{\mathrm{RF}},\end{aligned}$$

where the remainder is bounded by the same one as in the local version formula

$$\mathcal{R}_{\mathrm{RF}}\lesssim(1-2^{-l})^n+\frac{2^l}{n(1+(n-1)2^{-l})^{1/2}}.$$

C.6.3 Derivation of The Convergence Rate

To investigate the convergence rate, we first notice that the convergence rates of the variance parts in (34) and (35) are bounded in order $2^l/n$. Hence, it suffices to focus on the convergence rate for the squared bias part, which is bounded by

$$\left(\max_{1\le i\le s}\frac{\beta_i^2}{12}\right)\mathbb{E}\left[\sum_{i=1}^{s}2^{-2J_{li}}\right].$$

Denote by $H_l=\sum_{i=1}^{s}2^{-2J_{li}}$. We see that H_l is non-increasing and satisfies

$$\begin{aligned}H_l-H_{l+1}&=\sum_{i=1}^{s}\left(2^{-2J_{li}}-2^{-2J_{l+1,i}}\right)I\{\text{the }(l+1)\text{th split is on }i\}\\&=\frac{3}{4}\sum_{i=1}^{s}2^{-2J_{l,i}}I\{\text{the }(l+1)\text{th split is on }i\}.\end{aligned}$$

Let us define $P_{J_l}(i) = \mathbb{P}(\text{the } (l+1)\text{th split is on } i | J_l)$ for $1 \le i \le s$. It holds that

$$\mathbb{E}(H_{l+1}|J_l) = \sum_{i=1}^{s} 2^{-2J_{li}} \left(1 - \frac{3}{4} P_{J_l}(i)\right).$$

We next aim to show that

$$\min_{1\le i\le s} P_{J_l}(i) \ge \gamma \left(1 - \frac{s}{d}\right) \cdots \left(1 - \frac{s}{d - \lceil \gamma d \rceil + 1}\right) = \gamma W_{\gamma,d}(s), \tag{C.37}$$

where the RHS above is independent of J_l, and $W(s)$ is the same as defined in the last subsection. To this end, denote by $n_{l,i} = \#\{1 \le j \le s : \beta_j^2 2^{-2J_{l,j}} = \beta_i^2 2^{-2J_{l,i}}\}$ and $m_{l,i} = \#\{1 \le j \le s : \beta_j^2 2^{-2J_{l,j}} < \beta_i^2 2^{-2J_{l,i}}\}$. It follows from the definition that

$$\begin{aligned} P_{J_l}(i) &= \frac{1}{n_{l,i}} \frac{\binom{d-s+m_{l,i}+n_{l,i}}{\lceil \gamma d \rceil} - \binom{d-s+n_{l,i}}{\lceil \gamma d \rceil}}{\binom{d}{\lceil \gamma d \rceil}} \\ &= \frac{1}{n_{l,i}} \Bigg\{ \left(1 - \frac{s - m_{l,i} - n_{l,i}}{d}\right) \cdots \left(1 - \frac{s - m_{l,i} - n_{l,i}}{d - \lceil \gamma d \rceil + 1}\right) \\ &\qquad - \left(1 - \frac{s - m_{l,i}}{d}\right) \cdots \left(1 - \frac{s - m_{l,i}}{d - \lceil \gamma d \rceil + 1}\right) \Bigg\}. \end{aligned}$$

With an application of (C.32) and Taylor's theorem, it holds that for $0 \le x \le y \le s$, there exists a $\theta \in (0,1)$ such that

$$\begin{aligned} &W_{\gamma,d}(x) - W_{\gamma,d}(y) = W'_{\gamma,d}(x + \theta y)(x - y) \\ &= (y - x) W_{\gamma,d}(x + \theta y) \left(\sum_{k=0}^{\lceil \gamma d \rceil - 1} \frac{1}{d - j - x - \theta y} \right) \ge (y - x) W_{\gamma,d}(s) \frac{\lceil \gamma d \rceil}{d}. \end{aligned}$$

Substituting $x = s - m_{l,i} - n_{l,i}$ and $y = s - m_{l,i}$ into the inequality above, we have

$$\begin{aligned} P_{J_l}(i) &= \frac{1}{n_{l,i}} (W_{\gamma,d}(s - m_{l,i} - n_{l,i}) - W_{\gamma,d}(s - n_{l,i})) \\ &\ge W_{\gamma,d}(s) \frac{\lceil \gamma d \rceil}{d} \ge \gamma W_{\gamma,d}(s). \end{aligned}$$

As such, it follows that

$$\mathbb{E}(H_{l+1}) \le \left(1 - \frac{3}{4}\gamma W_{\gamma,d}(s)\right) \mathbb{E}(H_l) \le \cdots \le \left(1 - \frac{3}{4}\gamma W_{\gamma,d}(s)\right)^{l+2} s.$$

Combining the results above, we can finally obtain that

$$\begin{aligned} &\max\{\mathrm{MSE}(\widehat{\mu}_{\mathrm{RF}}), \mathrm{MSE}(\widehat{\mu}_{\mathrm{tree}})\} \\ &\lesssim \left(\max_{1\le j\le s} \frac{\beta_j^2}{12} \right) (1 - \gamma W(s))^{l+1} s + \left(\sigma_0^2 + \frac{1}{12} \sum_{j=1}^{s} \beta_j^2 \right) \frac{2^l}{n}, \end{aligned}$$

which completes the proof.

C.6.4 Derivation of The Cross-Tree Covariance Formula in Remark 18

The derivation of the formula in Remark 18 is generally similar to that in Remark 10. Thus, we only emphasize the differences and omit repeated details here. In view of (17), since there are $\sum_{j=1}^{d} I_{lj}$ splits at depth l, the size of terminal nodes is $2^{\sum_{j=1}^{d} I_{lj}}$ and hence $Q(P) = 2^{\sum_{j=1}^{d} I_{lj}}$.

Meanwhile, for two terminal nodes $\mathbf{t}(\boldsymbol{x}_0; J_l)$ and $\mathbf{t}(x_0; J_l')$ in P and P', the edge length of their intersection is $2^{-\max\{J_{li}, J_{li}'\}}$ for each $i = 1, \ldots, p$. Then we can deduce that

$$K(P, P'; \boldsymbol{x}_0) = \frac{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l) \cap \mathbf{t}(\boldsymbol{x}_0; J_l'))}{\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l))\mathbb{P}_{\boldsymbol{X}}(\mathbf{t}(\boldsymbol{x}_0; J_l'))} = 2^{\sum_{i=1}^{d} J_{li} + J_{li}' - \max\{J_{li}, J_{li}'\}} = 2^{\sum_{i=1}^{d} \min\{J_{li}, J_{li}'\}},$$

which remains constant for any $\boldsymbol{x}_0$. Therefore, it follows from (C.34) that

$$K(P, P') = \mathbb{E}_{\boldsymbol{X}'}\left[K(P, P'; \boldsymbol{X}')\right] = 2^{\sum_{i=1}^{d} \min\{J_{li}, J_{li}'\}},$$

which yields the final conclusion. This concludes the proof of Theorem 15.

C.7 Proof of Corollary 19

From the definition of the uniform CART process, we see that for the special setting of $s = 1$, the evolution pattern of J_t is given by

$$J_t = J_{t-1} + A_t = \sum_{s=1}^{t} A_s, \quad J_0 = 0 \in \mathbb{R}^d,$$

where $(A_s)_{s \geq 1}$ is a sequence of i.i.d. random vectors taking values over the set of unit coordinates $\{e_j\}$ with

$$\mathbb{P}(A_s = e_j) \simeq \begin{cases} \gamma & \text{for } j = 1, \\ \frac{1-\gamma}{d-1} & \text{otherwise.} \end{cases}$$

Hence, we have that i) the coordinate process $J_{l1} \sim \mathcal{B}(l, \gamma)$, a binomial random variable; and ii) the whole process J_l is multinomial distributed, say $\mathcal{M}(n, p(\gamma))$, where

$$p(\gamma) = \big(\gamma, \underbrace{\frac{1-\gamma}{d-1}, \ldots, \frac{1-\gamma}{d-1}}_{d-1}\big).$$

Let us go back to the MSE bound in (34). For terms related to a single tree, the squared bias and the leading term of the variance part become

$$\mathbb{E}\Big[2^{-2J_{l1}}\Big] = \left(1 - \frac{3}{4}\gamma\right)^l, \quad \mathbb{E}\Big[2^{\sum_{j=1}^{d} J_{lj}}\Big] = 2^l, \tag{C.38}$$

respectively. Meanwhile, for the tree ensemble, we will show that the squared bias and the leading term of the variance part satisfy

$$\mathbb{E}\Big[2^{-2\max\{J_{l1},J'_{l1}\}}\Big] = \Big(1-\frac{\gamma}{2}\Big)^{2l} F_l\left(\frac{\gamma}{2-\gamma}\right), \tag{C.39}$$

$$\mathbb{E}\Big[2^{\sum_{j=1}^{d}\min\{J_{lj},J'_{lj}\}}\Big] = 2^l G_{l,d}(\gamma), \tag{C.40}$$

in which F_l and $G_{l,d}$ are as given in the corollary. Moreover, using a similar argument as in the discussion on the leading terms, accompanied by the Cauchy–Schwartz inequality, we can show that the interaction terms in the variance part are minor and have an order of $(1-\frac{3\gamma}{4})^l\frac{2^l}{n}+(1-2^{-l})^n$. Consequently, combining (C.38)–(C.40) with (34), we can derive the target MSE bound (37).

The rest of the proof is outlined as follows. We first prove equations (C.39) and (C.40), and provide upper bounds for the rate functions $F_l(q)$ and $G_{l,d}(\gamma)$. Next, we investigate the monotonicity of product $(1-\gamma/2)^{2l}F_l(\gamma/(2-\gamma))$ in γ, which reflects how the squared bias for the tree ensemble decreases as γ varies. Finally, we focus on $G_{l,d}(\gamma)$ and study its monotonicity in dimensionality d and γ, respectively.

Proof of equation (C.39). Recall that J_{l1} and J'_{l1} are independent binomial random variables. Using the identity $\max\{x,y\}=\frac{x+y}{2}+\frac{|x-y|}{2}$, we can deduce that

$$\mathbb{E}\Big[2^{-2\max\{J_{l1},J'_{l1}\}}\Big] = \Big(1-\frac{\gamma}{2}\Big)^{2l}\mathbb{E}_{\mathbb{Q}}\Big[2^{-|J_{l1}-J'_{l1}|}\Big], \tag{C.41}$$

where probability measure $\mathbb{Q}$ is defined via a transform

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{2^{-(J_{l1}+J'_{l1})}}{\mathbb{E}\big[2^{-(J_{l1}+J'_{l1})}\big]} \quad \text{with } \mathbb{E}\big[2^{-(J_{l1}+J'_{l1})}\big] = \Big(1-\frac{\gamma}{2}\Big)^{2l}.$$

In addition, we surprisingly find that under probability measure $\mathbb{Q}$, i) J_{l1} and J'_{l1} are still independent of each other, and ii) they follow the binomial distribution $\mathcal{B}(l,q(\gamma))$ with

$$q(\gamma)=\frac{\gamma}{2-\gamma}. \tag{C.42}$$

In fact, the first conclusion follows from

$$\begin{aligned}\mathbb{Q}(J_{l1}=m,J'_{l1}=m') &= \frac{\mathbb{E}[2^{-(J_{l1}+J'_{l1})}I\{J_{l1}=m,J'_{l1}=m'\}]}{\mathbb{E}[2^{-J_{l1}}]\mathbb{E}[2^{-J'_{l1}}]}\\ &= \frac{\mathbb{E}[2^{-J_{l1}}I\{J_{l1}=m\}]}{\mathbb{E}[2^{-J_{l1}}]}\frac{\mathbb{E}[2^{-J'_{l1}}I\{J'_{l1}=m'\}]}{\mathbb{E}[2^{-J'_{l1}}]}\\ &= \mathbb{Q}(J_{l1}=m)\mathbb{Q}(J'_{l1}=m'),\end{aligned}$$

while the second equality above holds since

$$\begin{aligned}\mathbb{Q}(J_{l1}=m) &= \frac{\mathbb{E}[2^{-J_{l1}}I\{J_{l1}=m\}]}{\mathbb{E}[2^{-J_{l1}}]} = \frac{2^{-m}}{(1-\frac{\gamma}{2})^l}\mathbb{P}(J_{l1}=m)\\ &= \binom{l}{m}\Big(\frac{\gamma}{2-\gamma}\Big)^m\Big(1-\frac{\gamma}{2-\gamma}\Big)^{l-m}.\end{aligned}$$

Recall that the Fourier coefficient of a Poisson kernel satisfies

$$\int_{-\pi}^{\pi} e^{\mathrm{i}k\theta}\mu_r(\theta)d\theta = r^{|k|}, \quad k \in \mathbb{Z}.$$

It then follows by taking $r = 1/2$ that

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\Big[2^{-|J_{l1}-J'_{l1}|}\Big] &= \int_{-\pi}^{\pi} \mathbb{E}_{\mathbb{Q}}\Big[e^{\mathrm{i}\theta(J_{l1}-J'_{l1})}\Big]\mu_{1/2}(\theta)d\theta \\
&= \int_{-\pi}^{\pi} \Big|\mathbb{E}_{\mathbb{Q}}\Big[e^{\mathrm{i}\theta J_{l1}}\Big]\Big|^2\mu_{1/2}(\theta)d\theta \\
&= \int_{-\pi}^{\pi} \Big|(1-q(\gamma)) + q(\gamma)e^{\mathrm{i}\theta}\Big|^{2l}\mu_{1/2}(\theta)d\theta,
\end{aligned}$$

where the last equality above holds due to the fact that the characteristic function of a binomial random variable $J_{l1} \sim_{\mathbb{Q}} \mathcal{B}(l, q(\gamma))$ is given by

$$\mathbb{E}_{\mathbb{Q}}\Big[e^{\mathrm{i}\theta J_{l1}}\Big] = \Big((1-q(\gamma)) + q(\gamma)e^{\mathrm{i}\theta}\Big)^l.$$

Combining the results above, we can complete the proofs of equation (38) and the fact that

$$\begin{aligned}
\mathbb{E}\Big[2^{-2\max\{J_{l1},J'_{l1}\}}\Big] &= \Big(1-\frac{\gamma}{2}\Big)^{2l}\mathbb{E}_{\psi\sim\mu_{1/2}}\left[\Big|(1-q(\gamma)) + q(\gamma)e^{\mathrm{i}\psi}\Big|^{2l}\right] \\
&= \Big(1-\frac{\gamma}{2}\Big)^{2l} F_l(q(\gamma)).
\end{aligned}$$

Upper bound for $F_l(q)$. In what follows, we will provide an upper bound for term $F_l(q)$ for each $q \in (0,1)$. Observe that

$$\Big|(1-q) + qe^{\mathrm{i}\theta}\Big|^2 = 1 - 2q(1-q)\big(1-\cos\theta\big) =: 1 - c(1-\cos\theta),$$

where $c := 2q(1-q) \in (0, \frac{1}{2}]$ when $q \in (0,1)$. Using the elementary inequalities

$$1 - x \le e^{-x}, \quad 1-\cos\theta \;\ge\; \frac{2}{\pi^2}\,\theta^2$$

that are valid for all $x \ge 0$ and all $\theta \in [-\pi, \pi]$, we can show that

$$\big(1 - c(1-\cos\theta)\big)^l \le \exp\big(-\ell\, c(1-\cos\theta)\big) \le \exp\Big(-\frac{2c}{\pi^2}\,\ell\,\theta^2\Big).$$

Hence, with the aid of the fact that $P_{1/2}$ has a positive upper bound $\|P_{1/2}\|_\infty$ over $[-\pi, \pi]$, we have that

$$F_l(q) \le \frac{\|P_{1/2}\|_\infty}{2\pi}\int_{-\pi}^{\pi} \exp\Big(-\frac{2c}{\pi^2}l\theta^2\Big)d\theta \le \frac{\|P_{1/2}\|_\infty}{2}\sqrt{\frac{\pi}{2cl}},$$

which verifies the conclusion that $F_l(q) = O(l^{-\frac{1}{2}})$ for any $q \in (0,1)$.

Proof of equation (C.40). Recall that $\sum_{j=1}^d J_{lj} = \sum_{j=1}^d J'_{lj} = l$ and $\min\{x, y\} = \frac{x+y}{2} - \frac{|x-y|}{2}$. Then it holds that

$$\begin{aligned}\mathbb{E}\left[2^{\sum_{j=1}^d \min\{J_{lj}, J'_{lj}\}}\right] &= \mathbb{E}\left[2^{\frac{1}{2}\sum_{j=1}^d J_{lj} + \frac{1}{2}\sum_{j=1}^d J'_{lj} - \frac{1}{2}\sum_{j=1}^d |J_{lj} - J'_{lj}|}\right] \\ &= 2^l \mathbb{E}\Big[2^{-\frac{1}{2}\|J_l - J'_l\|_1}\Big],\end{aligned} \tag{C.43}$$

where $\|\cdot\|_1$ denotes the 1-norm of vectors in $\mathbb{R}^d$. Next, using the identity of the Poisson kernel and the characteristic function for a multinomial distribution $\mathcal{M}(l, p)$

$$\mathbb{E}\left[e^{\mathrm{i}\langle\theta, J_l\rangle}\right] = \left(\sum_{j=1}^d p_j e^{\mathrm{i}\theta_j}\right)^l,$$

we can deduce that

$$\begin{aligned}\mathbb{E}\Big[2^{-\frac{1}{2}\|J_l - J'_l\|_1}\Big] &= \mathbb{E}\left[\int_{[-\pi,\pi]^d} e^{\mathrm{i}\langle\theta, J_l - J'_l\rangle} \mu_{1/\sqrt{2}}(\theta_1)\cdots\mu_{1/\sqrt{2}}(\theta_d) d\theta_1\cdots d\theta_d\right] \\ &= \int_{[-\pi,\pi]^d} \left|\mathbb{E}\left[e^{\mathrm{i}\langle\theta, J_l\rangle}\right]\right|^2 \mu_{1/\sqrt{2}}(\theta_1)\cdots\mu_{1/\sqrt{2}}(\theta_d) d\theta_1\cdots d\theta_d \\ &= \int_{[-\pi,\pi]^d} \left|\sum_{j=1}^d p_j e^{\theta_j}\right|^{2l} \mu_{1/\sqrt{2}}(\theta_1)\cdots\mu_{1/\sqrt{2}}(\theta_d) d\theta_1\cdots d\theta_d \\ &= \mathbb{E}_{\varphi_1,\ldots,\varphi_d}\Big[\Big|\sum_{j=1}^d p_j e^{\varphi_j}\Big|^{2l}\Big] = \widetilde{G}_{l,d}(p).\end{aligned}$$

Thus, substituting $p_1 = \gamma$ and $p_2 = \cdots = p_d = (1-\gamma)/(d-1)$ into $\widetilde{G}_{l,d}(p)$, we can derive the representation of $G_{l,d}(\gamma)$in (39).

Upper bounds for $G_{l,d}(\gamma)$. Let us define

$$S := \frac{1}{d-1}\sum_{j=2}^d e^{i\theta_j}.$$

Then it follows that

$$\begin{aligned}A_d(\gamma, \theta) &:= \Big|\ \gamma e^{\mathrm{i}\theta_1} + (1-\gamma)\frac{1}{d-1}\sum_{j=2}^d e^{\mathrm{i}\theta_j}\ \Big|^2 = \Big|\gamma e^{i\theta_1} + (1-\gamma)S\Big|^2 \\ &= \gamma^2 + (1-\gamma)^2|S|^2 + \frac{2\gamma(1-\gamma)}{d-1}\sum_{j=2}^d \cos(\theta_j - \theta_1).\end{aligned}$$

Using $|S| \le 1$ and $\cos(\cdot) \le 1$, we can show that

$$A_d(\gamma, \theta) \le 1 - c_\gamma \frac{1}{d-1}\sum_{j=2}^d (1 - \cos(\theta_j - \theta_1)),$$

where $c_\gamma := 2\gamma(1-\gamma) \in (0, 1/2]$ for $\gamma \in (0,1)$. Since $1 - x \le e^{-x}$, it holds that

$$A_d(\gamma,\theta)^l \le \exp\left\{ - l c_\gamma \Big(\frac{1}{d-1} \sum_{j=2}^{d} (1 - \cos(\theta_j - \theta_1)) \Big) \right\}. \tag{C.44}$$

For all $|x| \le 2\pi$, we have that

$$1 - \cos x \;\ge\; \frac{2}{\pi^2}\, \rho(x)^2, \quad \rho(x) := \min\{\, |x|,\, 2\pi - |x| \,\} \in [0, \pi].$$

Let $\varphi_j := \Theta_j - \Theta_1 \in [-2\pi, 2\pi]$ for $j = 2, \ldots, d$. Then from (C.44), it follows that

$$A_d(\gamma,\theta)^l \le \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{1}{d-1} \sum_{j=2}^{d} \rho(\varphi_j)^2 \right\}.$$

Note that $P_{1/\sqrt{2}}(\theta) \le \|P_{1/\sqrt{2}}\|_\infty$. We can deduce that

$$\begin{aligned}
&G_{l,d}(\gamma) = \mathbb{E}[A_d(\gamma, \mathbf{\Theta})^l] \\
&\le \Big(\frac{\|P_{1/\sqrt{2}}\|_\infty}{2\pi} \Big)^d \int_{[-\pi,\pi]^d} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \sum_{j=2}^{d} \rho(\theta_j - \theta_1)^2 \right\} d\theta_1 \cdots d\theta_d \\
&\le \Big(\frac{\|P_{1/\sqrt{2}}\|_\infty}{2\pi} \Big)^d \int_{[-\pi,\pi]} \left(\prod_{j=2}^{d} \int_{[-\pi,\pi]} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \rho(\theta_j - \theta_1)^2 \right\} d\theta_j \right) d\theta_1.
\end{aligned}$$

For each fixed θ_1, let us set $\phi_j = \theta_j - \theta_1 \in [-2\pi, 2\pi]$. Then it holds that

$$\begin{aligned}
&\int_{[-\pi,\pi]} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \rho(\theta_j - \theta_1)^2 \right\} d\theta_j \\
&= \int_{[-\pi-\theta_1, \pi-\theta_1]} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \rho(\phi_j)^2 \right\} d\phi_j \\
&\le \int_{[-2\pi, 2\pi]} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \rho(\phi_j)^2 \right\} d\phi_j.
\end{aligned}$$

Since each $u \in [-\pi, \pi]$ has at most two preimages under this folding and $\int_{-2\pi}^{2\pi} e^{-\kappa\rho(\phi)^2}\, d\phi \le 2 \int_{-\pi}^{\pi} e^{-\kappa u^2}\, du$ for any $\kappa > 0$, an application of Fubini's theorem leads to

$$\prod_{j=2}^{d} \int_{[-\pi,\pi]} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \rho(\theta_j - \theta_1)^2 \right\} d\theta_j \;\le\; 2^{\,d-1} \int_{[-\pi,\pi]^{d-1}} \exp\Big\{ -\kappa \sum_{j=2}^{d} u_j^2 \Big\}\, du_2 \cdots du_d,$$

where $\kappa = \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1}$. Hence, we have that

$$G_{l,d}(\gamma) \;\le\; C_0(d,r) \int_{[-\pi,\pi]^{d-1}} \exp\left\{ - \frac{2c_\gamma}{\pi^2} \frac{\ell}{d-1} \sum_{j=2}^{d} \varphi_j^2 \right\} d\varphi_2 \cdots d\varphi_d,$$

where $C_0(d,r) = (\|P_{1/\sqrt{2}}\|_\infty/(2\pi))^d \cdot (2\pi) \cdot 2^{d-1}$.

Further, bounding the cube by $\mathbb{R}^{d-1}$ gives that

$$\int_{[-\pi,\pi]^{d-1}} e^{-\alpha\sum_{j=2}^d \varphi_j^2}\, d\varphi \;\le\; \int_{\mathbb{R}^{d-1}} e^{-\alpha\|v\|^2}\, dv = \Big(\frac{\pi}{\alpha}\Big)^{\frac{d-1}{2}}, \quad \alpha = \frac{2c_\gamma}{\pi^2}\,\frac{\ell}{d-1}.$$

Therefore, we can obtain that

$$G_{l,d}(\gamma) \le C_0(d,r)\,\Big(\frac{\pi(d-1)}{\alpha}\Big)^{\frac{d-1}{2}} = C_0(d,r)\,\Big(\frac{\pi^3(d-1)}{2c_\gamma}\Big)^{\frac{d-1}{2}}\,\ell^{-\frac{d-1}{2}}.$$

In short, for each fixed $d \ge 2$ and $\gamma \in (0,1)$, we have $G_{l,d}(\gamma) = O\big(\ell^{-(d-1)/2}\big)$ as $\ell \to \infty$.

Monotonicity of $\mathcal{M}_l(\gamma) = (1-\gamma/2)^{2l}F_l(\gamma)$ in γ. We now aim to show that $M_l(\gamma)$ locally decreases in γ for sufficiently large l. In addition, we have

$$\frac{2\sqrt{\pi}}{\sqrt{l}}\big(1-\frac{\gamma}{2}\big)^{2l} \le \mathcal{M}_l(\gamma) \le \big(1-\frac{\gamma}{2}\big)^{2l}.$$

To see this, let us write

$$K(q,\theta) = \big|(1-q(\gamma)) + q(\gamma)e^{i\theta}\big|^2 = 1 - 2q(1-q)(1-\cos\theta) \ge 0.$$

Then it holds that $F_l(q) = \mathbb{E}_\psi[K(q,\psi)^l]$ and $F_l'(q) = l\mathbb{E}_\psi[K(q,\psi)^{l-1}\partial_q K(q,\psi)]$, where

$$\partial_q K(q,\theta) = -2(1-2q)(1-\cos\theta).$$

It is clear that $F_l'(q) = 0$ has a unique zero at $q_0 = 1/2$. Moreover, if $0 \le q \le 1/2$, we have $F_l'(q) < 0$ and thus $F_l(q)$ decreases; if $1/2 < q \le 1$, we have $F_l'(q) > 0$ and thus $F_l(q)$ increases. Hence, it follows that

$$F_l(1/2) \le F_l(\gamma) \le F_l(1) = F_l(0) = 1.$$

By the Beta integral formula and Stirling's formula, we can deduce that

$$\begin{aligned} F_l(1/2) &= \mathbb{E}_\psi\Big[\Big(\frac{1+\cos\psi}{2}\Big)^l\Big] = \mathbb{E}_\psi\Big[\cos^{2l}\Big(\frac{\psi}{2}\Big)\Big] \\ &= 4\left(\int_0^{\pi/2}\cos^{2l}(u)du\right)(1+o(1)) \\ &= 2\sqrt{\pi}\frac{\Gamma(l+1/2)}{\Gamma(l+1)} \sim 2\sqrt{\pi}l^{-1/2}. \end{aligned}$$

Hence, for any $0 \le \gamma_1 < \gamma_2 \le 1$, we have that

$$\begin{aligned} \frac{\mathcal{M}_l(\gamma_2)}{\mathcal{M}_l(\gamma_1)} &= \Big(\frac{1-\gamma_2/2}{1-\gamma_1/2}\Big)^{2l}\frac{F_l(\gamma_2)}{F_l(\gamma_1)} \le \Big(\frac{1-\gamma_2/2}{1-\gamma_1/2}\Big)^{2l}\frac{1}{F_l(1/2)} \\ &= \frac{\sqrt{l}}{2\sqrt{\pi}}\Big(\frac{1-\gamma_2/2}{1-\gamma_1/2}\Big)^{2l} < 1 \end{aligned}$$

for sufficiently large l, which proves the assertion.

Monotonicity of $G_{l,d}(\gamma)$ in dimensionality d. We now prove that $G_{l,d+1}(\gamma) \le G_{l,d}(\gamma)$ for each fixed $\gamma \in (0,1)$ and fixed $l \in \mathbb{N}$. Recall that

$$G_{l,d}(\gamma) = \mathbb{E}_{\varphi_1,\dots,\varphi_d}\Big| \gamma e^{i\varphi_1} + (1-\gamma)\frac{1}{d-1}\sum_{j=2}^{d} e^{i\varphi_j} \Big|^{2l},$$

where $\varphi_1, \dots, \varphi_d$ are i.i.d. random variables with distribution $\mu_{1/\sqrt{2}}$ (the Poisson kernel measure).

Lemma 26 (Averaging decreases convex expectations) *Let $Y_1, Y_2, \dots$ be i.i.d. random vectors and $\overline{Y}_m = \frac{1}{m}\sum_{k=1}^{m} Y_k$. Then for any convex $g : \mathbb{C} \to \mathbb{R}$, we have*

$$\mathbb{E}[g(\overline{Y}_{m+1})] \le \mathbb{E}[g(\overline{Y}_m)], \quad m \ge 1.$$

Proof. For each $i = 1, \dots, m+1$, let us define the leave-one-out mean

$$\overline{Y}^{(-i)} := \frac{1}{m}\sum_{j \ne i} Y_j.$$

Then it holds that

$$\overline{Y}_{m+1} = \frac{1}{m+1}\sum_{i=1}^{m+1} \overline{Y}^{(-i)}.$$

By the convexity of function g, we have

$$g(\overline{Y}_{m+1}) = g\left(\frac{1}{m+1}\sum_{i=1}^{m+1} \overline{Y}^{(-i)}\right) \le \frac{1}{m+1}\sum_{i=1}^{m+1} g(\overline{Y}^{(-i)}).$$

Taking expectations and using the fact that all $\overline{Y}^{(-i)}$ share the same distribution as $\overline{Y}_m$, we can obtain that

$$\mathbb{E}[g(\overline{Y}_{m+1})] \le \frac{1}{m+1}\sum_{i=1}^{m+1} \mathbb{E}[g(\overline{Y}^{(-i)})] = \mathbb{E}[g(\overline{Y}_m)].$$

In addition, if g is strictly convex and the distribution of Y_k's are non-degenerate, the inequality above becomes a strict one. This completes the proof of Lemma 26.

We now apply Lemma 26 above to $Y_k := e^{\mathrm{i}\varphi_{k+1}}$ and convex function

$$g_a(w) := |\, a + (1-\gamma)w \,|^{2l} \quad \text{with } a = \gamma e^{\mathrm{i}\varphi_1}.$$

Since $S_m = \overline{Y}_m$, we can obtain that

$$\mathbb{E}\big[g_a(S_{m+1}) \mid \varphi_1\big] \le \mathbb{E}\big[g_a(S_m) \mid \varphi_1\big].$$

Taking expectation with respect to φ_1 yields the monotonicity of $G_{l,d}$ in dimensionality d. Since $\mu_{1/\sqrt{2}}$ is non-degenerate (e.g., the Poisson kernel) and $l \ge 1$, the inequality above is in fact a strict one.

Monotonicity of $G_{l,d}(\gamma)$ in γ. We finally aim to show that $G_{l,d}(\gamma)$ exhibits a U-shaped curve and attains its unique minimum at $\gamma^* = 1/d$. Let

$$\Delta^{d-1} = \{p \in \mathbb{R}^d : p_j \geq 0 \ \text{ and } \ \sum_j p_j = 1\}$$

be the collection of probability measures on $\{1, \ldots, d\}$. Recall that function $\widetilde{G}_{l,d} : \Delta^{d-1} \to \mathbb{R}$ is defined as

$$\widetilde{G}_{l,d}(p) = \mathbb{E}_{\varphi_1,\ldots,\varphi_d}\Big[\Big|\sum_{j=1}^{d} p_j e^{\varphi_j}\Big|^{2l}\Big].$$

We assert that $\widetilde{G}_{l,d}$ is in fact *symmetric* and *convex* on the probability simplex. Indeed, for each fixed $\theta = (\theta_1, \ldots, \theta_d)$, the map

$$p \ \longmapsto \ \Big|\sum_{j=1}^{d} p_j e^{i\theta_j}\Big|^{2l}$$

is the composition of a linear map in p and the convex function $z \mapsto |z|^{2l}$ on $\mathbb{C} \simeq \mathbb{R}^2$ (with $2l \geq 2$). Hence, such map is convex in p. The integration is taken with respect to a nonnegative product measure that is symmetric in the coordinates, and thus we establish the assertion.

Next, we exploit the concept of the Schur convexity in Marshall et al. (2011)[Chp 3] to show that $\widetilde{G}_{l,d}(p)$ attains its minimum at $p = (1/d, \cdots, 1/d)$. Recall that $\widetilde{G}_{l,d}$ is symmetric and convex, by Proposition C. 2 of Marshall et al. (2011)[Chp 3], $\widetilde{G}_{l,d}(p)$ is Schur-convex; that is, if $p \succ q$ (majorization and see Marshall et al. (2011)[Chp 1. A.1] for its definition), we have $\widetilde{G}_{l,d}(p) \geq \widetilde{G}_{l,d}(q)$. Since the uniform distribution $u = (1/d, \ldots, 1/d)$ is majorized by each $p \in \Delta^{d-1}$, the Schur-convexity yields that

$$\widetilde{G}_{l,d}(p) \geq \widetilde{G}_{l,d}(u).$$

Consequently, the unique global minimizer of $\widetilde{G}_{l,d}(p)$ on the simplex is the uniform distribution.

Observe that $G_{l,d}(\gamma) = \widetilde{G}_{l,d}(p(\gamma))$ with the affine transform

$$p(\gamma) = \Big(\gamma, \underbrace{\tfrac{1-\gamma}{d-1}, \ldots, \tfrac{1-\gamma}{d-1}}_{d-1}\Big), \quad \gamma \in [0,1].$$

Because $\widetilde{G}_{l,d}$ is convex on the simplex, the restriction $\gamma \mapsto G_{l,d}(\gamma)$ is also convex on $[0,1]$. By symmetry, the unique minimizer along this path is where all coordinates are equal, i.e., $\gamma^* = 1/d$. Therefore, we see that $G_{l,d}(\gamma)$ is a U-shaped curve: it decreases on $[0, 1/d]$ and increases on $[1/d, 1]$, and has a unique minimum at $\gamma^* = 1/d$. This completes the proof of Corollary 19.

Appendix D. Some Key Lemmas and Their Proofs

In this section, we provide a list of lemmas on the asymptotic inverse moments involving sample counts over subsets, which are used in the proof of our main theorems. The proofs of these lemmas are detailed in the later subsections.

In the first lemma, we derive the upper bound for inverse moments involving a multinomial distribution.

Lemma 27 *Let $(N_1, \ldots, N_k) \sim \mathcal{M}(n, (p_1, \ldots, p_K))$. Then for any $a_{ij} \geq 1$, $i = 1, \ldots, K$, $j = 1, \ldots, m_i$, it holds that*

$$\mathbb{E}\left[\prod_{i=1}^{K}\prod_{j=1}^{m_i}\left(\frac{1}{N_i + a_{ij}}\right)\right] \lesssim \prod_{i=1}^{K}\prod_{j=1}^{m_i}\left(\frac{1}{np_i + a_{ij}}\right).$$

The next lemma provides the higher-order expansion of the expected inverse power of a binomial random variable.

Lemma 28 *Let $N \sim \mathcal{B}(n, p)$ and r be a positive integer. Then for any constant $a \geq 1$, it holds that*

$$0 \leq \mathbb{E}\left[\frac{1}{(a+N)^r}\right] - \frac{1}{(a+np)^r} \lesssim \frac{1}{(a+np)^{r+1}}.$$

Further, we have

$$\mathbb{E}\left[\frac{1}{(a+N)^r}\right] = \frac{1}{(a+np)^r} + \frac{r(r+1)np(1-p)}{2(a+np)^{r+2}} + R_4,$$

in which the remainder satisfies

$$R_4 \lesssim \frac{1}{(a+np)^{r+3/2}}.$$

Further, we consider the higher-order expansion of expected products of inverse powers for a binomial random variable.

Lemma 29 *Let $N \sim \mathcal{B}(n, p)$, and r and s be two nonnegative integers with $r, s \geq 1$. Then for any constants $a, b \geq 1$, it holds that*

$$\begin{aligned} 0 &\leq \mathbb{E}\left[\frac{1}{(a+N)^r(b+N)^s}\right] - \frac{1}{(a+np)^r(b+np)^s} \\ &\lesssim \frac{1}{(a+np)^{r+1}(b+np)^s} + \frac{1}{(a+np)^r(b+np)^{s+1}}. \end{aligned}$$

Further, we have

$$\begin{aligned} \mathbb{E}\left[\frac{1}{(a+N)^r(b+N)^s}\right] &= \frac{1}{(a+np)^r(b+np)^s} + \frac{r(r+1)np(1-p)}{2(a+np)^{r+2}(b+np)^s} \\ &\quad + \frac{s(s+1)np(1-p)}{2(a+np)^r(b+np)^{s+2}} + \frac{rsnp(1-p)}{(a+np)^{r+1}(b+np)^{s+1}} + R_5, \end{aligned}$$

in which the remainder satisfies

$$R_5 \lesssim \frac{1}{(a+np)^{r+3/2}(b+np)^{s}} + \frac{1}{(a+np)^{r}(b+np)^{s+3/2}} + \frac{1}{(a+np)^{r+1}(b+np)^{s+1/2}} + \frac{1}{(a+np)^{r+1/2}(b+np)^{s+1}}.$$

The next lemma gives the expansion of inverse moments of the product of sample counts without overlapping.

Lemma 30 *Let A_1 and A_2 be two disjoint, non-empty subset with positive probability $p_1 = \mathbb{P}_{\boldsymbol{X}}(A_1)$ and $p_2 = \mathbb{P}_{\boldsymbol{X}}(A_2)$. Denote by $N_i = \sum_{i=1}^{n} I\{\boldsymbol{X}_i \in P_i\}$ for $i = 1, 2$. Then for any constants $a, b \geq 1$, it holds that*

$$\mathbb{E}\left[\frac{1}{(a+N_1)(b+N_2)}\right] = \frac{1}{(a+np_1)(b+np_2)} + \frac{np_1(1-p_1)}{(a+np_1)^3(b+np_2)} + \frac{np_2(1-p_2)}{(a+np_1)(b+np_2)^3} - \frac{np_1p_2}{(a+np_1)^2(b+np_2)^2} + R_6,$$

in which the remainder satisfies

$$R_6 \lesssim \frac{1}{(a+np_1)^{5/2}(b+np_2)} + \frac{1}{(a+np_1)(b+np_2)^{5/2}} + \frac{1}{(a+np_1)^{2}(b+np_2)^{3/2}} + \frac{1}{(a+np_1)^{3/2}(b+np_2)^{2}}.$$

In particular, we have

$$0 \leq \mathbb{E}\left[\frac{1}{(a+N_1)(b+N_2)}\right] - \frac{1}{(a+np_1)(b+np_2)} \lesssim \frac{1}{(a+np_1)(b+np_2)}\left(\frac{1}{a+np_1} + \frac{1}{b+np_2}\right).$$

Combining Lemmas 28–30, we finally derive the expansion of inverse moments of the product of sample counts with overlapping.

Lemma 31 *Let P and P' be two subsets in $\mathcal{X}^d$ with a non-empty intersection $P_0 = P \cap P' \neq \varnothing$ having positive probability $p_0 = \mathbb{P}(\boldsymbol{X} \in P_0)$. Define $p = \mathbb{P}(\boldsymbol{X} \in A)$ and $p' = \mathbb{P}(\boldsymbol{X} \in A)$. For a given sample $\{\boldsymbol{X}_i\}_{1 \leq i \leq n}$, denote by $N = \sum_{i=1}^{n} I\{\boldsymbol{X}_i \in P\}$ and $N' \sum_{i=1}^{n} I\{\boldsymbol{X}_i \in P'\}$ the sample counts of P and P', respectively. Then for any constant $a \geq 1$, it holds that*

$$\mathbb{E}\left[\frac{1}{(a+N)(a+N')}\right] = \frac{1}{(a+np)(a+np')}\left(1 + \frac{1-p}{a+np}\right)\left(1 + \frac{1-p'}{a+np'}\right) - \frac{n(p_0 - pp')}{(a+np)^2(a+np')^2} + R_7,$$

where the remainder satisfies

$$R_7 \lesssim \frac{1}{(a+np)^{5/2}(a+np')} + \frac{1}{(a+np)(a+np')^{5/2}} + \frac{1}{(a+np)^{2}(a+np')^{3/2}} + \frac{1}{(a+np)^{3/2}(a+np')^{2}}.$$

In particular, we have

$$0 \leq \mathbb{E}\left[\frac{1}{(a+N)(a+N')}\right] - \frac{1}{(a+np)(a+np')} \lesssim \frac{1}{(a+np)(a+np')}\left(\frac{1}{a+np} + \frac{1}{b+np'}\right).$$

The results displayed in Lemmas 27–31 above are inspired by those in Lemmas 3 and 7 of Klusowski and Tian (2024), and *extend* the scope of these findings by providing a more refined and comprehensive nonasymptotic framework. Specifically, we address the *higher-order* inverse moments for binomial distributions and also cover the cases related to the multinomial distributions.

D.1 Proof of Lemma 24

Assume that $N = \sum_{i=1}^n I_i$, where $\{I_j\}$ is a sequence of i.i.d. Bernoulli random variables with a success probability p. To examine the asymptotic expansion of the inverse moment of N, we adopt the *leave-one-out technique* with the following observation

$$\begin{aligned}\mathbb{E}\left\{\frac{I\{N \geq 1\}}{N}\right\} &= \mathbb{E}\left\{\frac{N}{N^2} I\{N \geq 1\}\right\} = \sum_{i=1}^n \mathbb{E}\left\{\frac{I_i}{N^2} I\{N \geq 1\}\right\} \\ &= \sum_{i=1}^n \mathbb{E}\left\{\frac{I_i}{N^2}\right\} = n\mathbb{E}\left\{\frac{I_1}{(I_1 + N_{-1})^2}\right\},\end{aligned}$$

in which $N_{-1} = \sum_{j=2}^n I_j \sim \mathcal{B}((n-1), p)$ and is independent of I_1. Hence, it holds that

$$\mathbb{E}\left\{\frac{I\{N \geq 1\}}{N}\right\} = n\mathbb{E}\left\{\frac{I_1}{(I_1 + N_{-1})^2}\right\} = np\mathbb{E}\left\{\frac{1}{(1 + N_{-1})^2}\right\}.$$

With an application of Lemma 28 with $r = 2$, we can obtain that

$$0 \leq \mathbb{E}\left\{\frac{1}{(1+N_{-1})^2}\right\} - \frac{1}{(1+(n-1)p)^2} \lesssim \frac{1}{(1+(n-1)p)^3}.$$

Consequently, we have that

$$0 \leq \mathbb{E}\left\{\frac{I\{N \geq 1\}}{N}\right\} - \frac{np}{(1+(n-1)p)^2} \lesssim \frac{np}{(1+(n-1)p)^3}.$$

Then it immediately follows that

$$\begin{aligned}\left|\mathbb{E}\left\{\frac{I\{N \geq 1\}}{N}\right\} - \frac{1}{np}\right| &\lesssim \frac{np}{(1+(n-1)p)^3} + \left|\frac{1}{np} - \frac{np}{(1+(n-1)p)^2}\right| \\ &= \frac{np}{(1+(n-1)p)^3} + \frac{2(1-p)}{(1+(n-1)p)^2} + \frac{(1-p)^2}{np(1+(n-1)p)^2} \\ &\lesssim \left(1 + \frac{1}{np}\right)\frac{1}{(1+(n-1)p)^2},\end{aligned}$$

which concludes the proof of Lemma 24.

D.2 Proof of Lemma 25

Proof of conclusion (1)

Let us first focus on justifying bound (B.25) of Lemma 25. Denote by

$$N_{-1} = \sum_{i=2}^{n} I\{\boldsymbol{X}_i \in P\} \quad \text{and} \quad N'_{-1} = \sum_{i=2}^{n} I\{\boldsymbol{X}_i \in P'\}.$$

Then it holds that

$$\frac{I\{\boldsymbol{X}_1 \in P \cap P'\}}{NN'} = \frac{I\{\boldsymbol{X}_1 \in P \cap P'\}}{(1+N_{-1})(1+N'_{-1})}.$$

Recall that $N_0 = \sum_{i=1}^{n} I\{\boldsymbol{X}_i \in P \cap P'\}$. We exploit the leave-one-out technique to deduce that

$$\begin{aligned}\mathbb{E}\left[\frac{N_0}{NN'}\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{I\{\boldsymbol{X}_i \in P \cap P'\}}{NN'}\right] = n\mathbb{E}\left[\frac{I\{X_1 \in P \cap P'\}}{(1+N_{-1})(1+N'_{-1})}\right] \\ &= np_0\mathbb{E}\left[\frac{1}{(1+N_{-1})(1+N'_{-1})}\right],\end{aligned}$$

where the last step above is because $\boldsymbol{X}_1$ is independent of N_{-1} and N'_{-1}.

An application of Lemma 31 with $a = 1$ leads to

$$\begin{aligned}\mathbb{E}\left[\frac{1}{(a+N_{-1})(a+N'_{-1})}\right] = &\left\{\frac{1}{a+(n-1)p}\left(1+\frac{1-p}{a+(n-1)p}\right)\right\} \\ &\times\left\{\frac{1}{a+(n-1)p'}\left(1+\frac{1-p'}{a+(n-1)p'}\right)\right\} + R_{11},\end{aligned} \tag{D.45}$$

in which the remainder satisfies

$$R_{11} \lesssim \frac{1}{(1+(n-1)p)^2(1+(n-1)p')} + \frac{1}{(1+(n-1)p)(1+(n-1)p')^2}.$$

Observe that

$$\frac{1}{1+(n-1)p}\left\{1+\frac{1}{1+(n-1)p}\right\} = \frac{1}{np} - \frac{(1-p)^2}{np(1+(n-1)p)^2}. \tag{D.46}$$

Then the leading term in (D.45) becomes

$$\begin{aligned}&\left\{\frac{1}{a+(n-1)p}\left(1+\frac{1-p}{a+(n-1)p}\right)\right\}\left\{\frac{1}{a+(n-1)p'}\left(1+\frac{1-p'}{a+(n-1)p'}\right)\right\} \\ =&\left\{\frac{1}{np} - \frac{(1-p)^2}{np(1+(n-1)p)^2}\right\}\left\{\frac{1}{np'} - \frac{(1-p')^2}{np(1+(n-1)p')^2}\right\} \\ =&\frac{1}{n^2pp'} + R_{12},\end{aligned}$$

in which the remainder R_{12} satisfies

$$R_{12} \lesssim \frac{1}{n^2pp'}\left(\frac{1}{(1+(n-1)p)^2} + \frac{1}{(1+(n-1)p')^2}\right).$$

Combining the expressions above, we can obtain that

$$\begin{aligned}\mathbb{E}\left[\frac{N_0}{NN'}\right] &= np_0\mathbb{E}\left[\frac{1}{(1+N_{-1})(1+N'_{-1})}\right] \\ &= \frac{p_0}{npp'} + np_0(R_{11}+R_{12}),\end{aligned}$$

where the remainder is controlled by

$$R_{21} = np_0(R_{11}+R_{12}) \lesssim \frac{p_0}{npp'}\left(\frac{1}{1+(n-1)p}+\frac{1}{1+(n-1)p'}\right).$$

Proof of conclusion (2)

It is easy to see that

$$\frac{N_1}{N} = \left(1-\frac{N_0}{N}\right) I\{N\geq 1\} \quad \text{and} \quad \frac{N_1}{N'} = \left(1-\frac{N_0}{N'}\right) I\{N'\geq 1\}.$$

The left-hand side (LHS) in (B.26) can be decomposed as

$$\begin{aligned}&\mathbb{E}\left[\left(\alpha\frac{N_0}{N}+\beta\frac{N_1}{N}\right)\left(\alpha\frac{N_0}{N'}+\gamma\frac{N_1}{N'}\right)\right] \\ =&\mathbb{E}\left[\left(\alpha\frac{N_0}{N}+\beta\left(1-\frac{N_0}{N}\right)\right)\left(\alpha\frac{N_0}{N'}+\gamma\left(1-\frac{N_0}{N'}\right)\right) I\{N\geq 1, N'\geq 1\}\right] \\ =&(\alpha-\beta)(\alpha-\gamma)\mathbb{E}\left[\frac{N_0^2}{NN'}\right] \\ &+\gamma(\alpha-\beta)\mathbb{E}\left[\frac{N_0}{N}\right]+\beta(\alpha-\gamma)\mathbb{E}\left[\frac{N_0}{N'}\right] \\ &+\beta\gamma\mathbb{P}(N\geq 1, N'\geq 1) \\ =:&J_1+J_2+J_3+J_4.\end{aligned}$$

Observe that

$$\mathbb{P}(N\geq 1, N'\geq 1) = 1-\mathbb{P}(\min\{N,N'\}=0) \geq 1-(1-p)^n-(1-p')^n.$$

We immediately see that

$$J_4 := \beta\gamma\mathbb{P}(N\geq 1, N'\geq 1) = \beta\gamma + R_{23} \tag{D.47}$$

with $R_{23}\leq |\beta||\gamma|((1-p)^n+(1-p')^n)$.

Next, for term J_2 we can apply the leave-one-out technique to show that

$$\mathbb{E}\left[\frac{N_0}{N}\right] = \sum_{i=1}^{n}\mathbb{E}\left[\frac{I\{X_i\in P\cap P'\}}{N}\right] = np_0\mathbb{E}\left[\frac{1}{1+N_{-1}}\right].$$

An application of Lemma 28 with $r=1$ and $a=1$ results in

$$\mathbb{E}\left\{\frac{1}{1+N_{-1}}\right\} = \frac{1}{1+(n-1)p}\left(1+\frac{1-p}{1+(n-1)p}\right)+R_{24},$$

where the remainder satisfies $R_{24} \lesssim 1/(1+(n-1)p)^{5/2}$. Notice that

$$\frac{1}{1+(n-1)p}\left(1+\frac{1-p}{1+(n-1)p}\right)=\frac{1}{np}-\frac{(1-p)^2}{np(1+(n-1)p)^2}.$$

It holds that

$$\mathbb{E}\left[\frac{N_0}{N}\right]=\frac{p_0}{p}+R_{25} \text{ with } R_{25}\lesssim\frac{p_0}{p(1+(n-1)p)^{3/2}}.$$

Similarly, by replacing N and p with N' and p', respectively, the same assertion still holds. Consequently, we have that

$$\begin{aligned} J_2+J_3&=\gamma(\alpha-\beta)\mathbb{E}\left[\frac{N_0}{N}\right]+\beta(\alpha-\gamma)\mathbb{E}\left[\frac{N_0}{N'}\right]\\ &=\gamma(\alpha-\beta)\frac{p_0}{p}+\beta(\alpha-\gamma)\frac{p_0}{p'}+R_{26}, \end{aligned} \tag{D.48}$$

where the remainder term satisfies

$$R_{26}\lesssim\max\{|\alpha|,|\beta|,|\gamma|\}^2\left(\frac{1}{(1+(n-1)p)^{3/2}}+\frac{1}{(1+(n-1)p')^{3/2}}\right).$$

For term J_1, we still adopt the leave-one-out technique and deduce that

$$\begin{aligned} \mathbb{E}\left[\frac{N_0^2}{NN'}\right]&=\sum_{i,j=1}^{n}\mathbb{E}\left[\frac{I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in P\cap P'\}}{NN'}\right]\\ &=\sum_{i=1}^{n}\mathbb{E}\left[\frac{I\{\boldsymbol{X}_i\in P\cap P'\}}{NN'}\right]+\sum_{1\le i\ne j\le n}\mathbb{E}\left[\frac{I\{\boldsymbol{X}_i,\boldsymbol{X}_j\in P\cap P'\}}{NN'}\right]\\ &=np_0\mathbb{E}\left[\frac{1}{(1+N_{-1})(1+N'_{-1})}\right]+n(n-1)p_0^2\mathbb{E}\left[\frac{1}{(2+N_{-2})(2+N_{-2})}\right]\\ &=:J_{11}+J_{12}, \end{aligned}$$

in which $N_{-2}=\sum_{i=3}^{n}I\{\boldsymbol{X}_i\in P\}$ and $N'_{-2}=\sum_{i=3}^{n}I\{\boldsymbol{X}_i\in P'\}$.

According to the discussion in conclusion (1), it holds that

$$J_{11}=\frac{p_0}{npp'}+R_{21} \quad \text{with } R_{21}\lesssim\frac{p_0}{npp'}\left(\frac{1}{1+(n-1)p}+\frac{1}{1+(n-1)p'}\right). \tag{D.49}$$

Meanwhile, note that

$$\frac{1}{2+(n-2)p}\left\{1+\frac{1-p}{2+(n-2)p}\right\}=\frac{1}{1+(n-1)p}-\frac{(1-p)^2}{(1+(n-1)p)(2+(n-2)p)^2}.$$

Then we can show that

$$\begin{aligned}
&\left\{\frac{1}{2+(n-2)p}\left\{1+\frac{1-p}{2+(n-2)p}\right\}\right\}\times\left\{\frac{1}{2+(n-2)p'}\left\{1+\frac{1-p'}{2+(n-2)p'}\right\}\right\}\\
=&\left\{\frac{1}{1+(n-1)p}-\frac{(1-p)^2}{(1+(n-1)p)(2+(n-2)p)^2}\right\}\\
\times&\left\{\frac{1}{1+(n-1)p'}-\frac{(1-p')^2}{(1+(n-1)p')(2+(n-2)p')^2}\right\}\\
=&\frac{1}{(1+(n-1)p)(1+(n-1)p')}+R_{27},
\end{aligned}$$

where the remainder R_{27} satisfies

$$R_{27}\lesssim\frac{1}{(1+(n-1)p)(1+(n-1)p')}\left(\frac{1}{(1+(n-1)p)^2}+\frac{1}{(1+(n-1)p')^2}\right).$$

Applying Lemma 31, we can obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{(2+N_{-2})(2+N_{-2})}\right]=&\frac{1}{(1+(n-1)p)(1+(n-1)p')}\\
&+\frac{(n-1)(p_0-pp')}{(2+(n-2)p)^2(2+(n-2)p')^2}+R_{27}.
\end{aligned}$$

To further simplify the leading terms, let us observe that

$$\begin{aligned}
\frac{1}{(1+(n-1)p)(1+(n-1)p')}=&\frac{1}{n(n-1)pp'}-\frac{n(p'+p-pp')}{n(n-1)pp'(1+(n-1)p)(1+(n-1)p')}\\
&-\frac{(1-p)(1-p')}{n(n-1)pp'(1+(n-1)p)(1+(n-1)p')},\\
\frac{1}{(1+(n-1)p)(1+(n-1)p')}=&\frac{1}{n^2pp'}-\frac{1-p'}{n^2pp'(1+(n-1)p')}-\frac{1-p}{n^2pp'(1+(n-1)p)}\\
&+\frac{(1-p)(1-p')}{n^2pp'(1+(n-1)p)(1+(n-1)p')}.
\end{aligned}$$

Combining the above results together, it follows that

$$\frac{1}{(1+(n-1)p)(1+(n-1)p')}=\frac{1}{n(n-1)pp'}-\frac{p'+p-pp'}{n^2(n-1)(pp')^2}+R_{28},$$

where the remainder satisfies

$$\begin{aligned}
R_{28}\lesssim&\frac{p+p'-pp'}{n^2(n-1)(pp')^2}\left\{\frac{1}{1+(n-1)p}+\frac{1}{1+(n-1)p'}\right\}\\
&+\frac{1}{n(n-1)pp'(1+(n-1)p)(1+(n-1)p')}.
\end{aligned}$$

Moreover, it holds that

$$\begin{aligned}\frac{1}{(2+(n-2)p)(2+(n-2)p')} =& \frac{1}{(n-1)^2pp'} - \frac{2-p'}{(n-1)^2pp'(2+(n-2)p')} \\ & - \frac{2-p}{(n-1)^2pp'(2+(n-2)p)} \\ & + \frac{(2-p)(2-p')}{(n-1)^2pp'(2+(n-2)p)(2+(n-2)p')}, \\ \frac{1}{(2+(n-2)p)(2+(n-2)p')} =& \frac{1}{n^2pp'} - \frac{2(1-p')}{n^2pp'(2+(n-2)p')} \\ & - \frac{2(1-p)}{n^2pp'(2+(n-2)p)} \\ & + \frac{4(1-p)(1-p')}{n^2pp'(2+(n-2)p)(2+(n-2)p')}.\end{aligned}$$

A combination of the results above gives

$$\frac{(n-1)(p_0-pp')}{(2+(n-2)p)^2(2+(n-2)p')^2} = \frac{p_0-pp'}{n^2(n-1)(pp')^2} + R_{29},$$

where the remainder satisfies

$$R_{29} \lesssim \frac{p_0-pp'}{n^2(n-1)(pp')^2}\left(\frac{1}{2+(n-2)p} + \frac{1}{2+(n-2)p'}\right).$$

In summary, we now have

$$\mathbb{E}\left[\frac{1}{(2+N_{-2})(2+N'_{-2})}\right] = \frac{1}{n(n-1)pp'} + \frac{p_0-p-p'}{n^2(n-1)(pp')^2} + R_{27} + R_{28} + R_{29}$$

and thus,

$$J_{12} = n(n-1)p_0^2\mathbb{E}\left[\frac{1}{(2+N_{-2})(2+N_{-2})}\right] = \frac{p_0^2}{pp'} + \frac{p_0^2(p_0-p-p')}{n(pp')^2} + R_{210}, \tag{D.50}$$

where the remainder satisfies

$$\begin{aligned}R_{210} &= n(n-1)p_0^2(R_{27}+R_{28}+R_{29}) \\ &\lesssim \frac{p_0^2(p+p'-p_0)}{n(pp')^2}\left\{\frac{1}{1+(n-1)p} + \frac{1}{1+(n-1)p'}\right\} \\ &\quad + \frac{p_0^2}{pp'(1+(n-1)p)(1+(n-1)p')} \\ &\lesssim \frac{p_0}{npp'}\left\{\frac{1}{1+(n-1)p} + \frac{1}{1+(n-1)p'}\right\}.\end{aligned}$$

Then substituting (D.49) and (D.50) into the expression, we can conclude that

$$\begin{aligned} J_1 = \mathbb{E}\left[\frac{N_0^2}{NN'}\right] &= \frac{p_0^2}{pp'} + \frac{p_0^2(p_0-p-p')}{n(pp')^2} + \frac{p_0}{npp'} + R_{21} + R_{210} \\ &= \frac{p_0^2}{pp'} + \frac{p_0}{npp'}\left(1 - \frac{p_0(p+p'-p_0)}{pp'}\right) + R_{21} + R_{210} \\ &= \frac{p_0^2}{pp'} + \frac{p_0}{npp'}\left(\frac{p_1p_2}{pp'}\right) + R_{21} + R_{210}, \end{aligned} \tag{D.51}$$

where the remainder satisfies

$$R_{21} + R_{210} \lesssim \frac{p_0}{npp'}\left\{\frac{1}{1+(n-1)p} + \frac{1}{1+(n-1)p'}\right\}.$$

Finally, combining bounds (D.47), (D.48), and (D.51), we can deduce that

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z}}\left[\left(\alpha\frac{N_0}{N} + \beta\frac{N_1}{N}\right)\left(\alpha\frac{N_0}{N'} + \gamma\frac{N_1}{N'}\right)\right] \\ =&(\alpha-\beta)(\alpha-\gamma)\left\{\frac{p_0^2}{pp'} + \frac{p_0}{npp'}\left(\frac{p_1p_2}{pp'}\right)\right\} + \gamma(\alpha-\beta)\frac{p_0}{p} + \beta(\alpha-\gamma)\frac{p_0}{p'} + \beta\gamma + R_{22} \\ =&\left(\alpha\frac{p_0}{p} + \beta\frac{p_1}{p}\right)\left(\alpha\frac{p_0}{p'} + \gamma\frac{p_2}{p'}\right) + (\alpha-\beta)(\alpha-\gamma)\left(\frac{p_1p_2}{pp'}\right)\left(\frac{p_0}{npp'}\right) + R_{22}, \end{aligned}$$

where the remainder satisfies

$$\begin{aligned} R_{22} \lesssim \max\{|\alpha|,|\beta|,|\gamma|\}^2 \Bigg\{ &\frac{p_0}{npp'}\left\{\frac{1}{1+(n-1)p} + \frac{1}{1+(n-1)p'}\right\} \\ &+ \frac{1}{(1+(n-1)p)^{3/2}} + \frac{1}{(1+(n-1)p')^{3/2}} \\ &+ (1-p)^n + (1-p')^n \Bigg\}. \end{aligned}$$

This completes the proof of Lemma 25.

D.3 Proof of Lemma 27

For each $i = 1, \ldots, K$, since $N_i \sim \mathcal{B}(n, p_i)$, by Lemma 3 in Cattaneo et al. (2024) we have

$$\mathbb{E}\left[\left(\prod_{j=1}^{m_i}\frac{1}{N_i + a_{ij}}\right)^K\right] \lesssim \left(\prod_{j=1}^{m_i}\frac{1}{np_i + a_{ij}}\right)^K.$$

It then follows from the Hölder inequality that

$$\mathbb{E}\left[\prod_{i=1}^{K}\prod_{j=1}^{m_i}\left(\frac{1}{N_i + a_{ij}}\right)\right] \le \prod_{i=1}^{K}\left\{\mathbb{E}\left(\prod_{j=1}^{m_i}\frac{1}{N_i + a_{ij}}\right)^K\right\}^{1/K} \lesssim \prod_{i=1}^{K}\prod_{j=1}^{m_i}\frac{1}{np_i + a_{ij}},$$

which proves the conclusion. This concludes the proof of Lemma 27.

D.4 Proof of Lemma 28

Let us define an auxiliary function $J(x) = 1/(a+x)^r$. The derivatives of function J are given by

$$J'(x) = -\frac{r}{(a+x)^{r+1}}, \quad J''(x) = \frac{r(r+1)}{(a+x)^{r+2}}, \quad J'''(x) = -\frac{r(r+1)(r+2)}{(a+x)^{r+3}}.$$

By resorting to Taylor's theorem, we have that

$$\begin{aligned} J(N) = J(np) + (N-np)J'(np) + \frac{1}{2}(N-np)^2 J''(np) \\ + \frac{1}{6}(N-np)^3 J'''(\theta N + (1-\theta)np), \end{aligned}$$

where θ is an unknown random parameter in $(0,1)$. Observe that $-J'''$ is a positive convex function. Hence, it holds that

$$0 \le R_4 := -J'''(\theta N + (1-\theta)np) \le \frac{r(r+1)(r+2)}{(a+N)^{r+3}} + \frac{r(r+1)(r+2)}{(a+np)^{r+3}}.$$

Further, we can show that

$$\begin{aligned} R_4 &:= \frac{1}{6}\mathbb{E}\left|(N-np)^3 J'''(\theta N + (1-\theta)np)\right| \\ &\lesssim \sqrt{\mathbb{E}[(N-np)^6]} \times \sqrt{\mathbb{E}\left(\frac{r(r+1)(r+2)}{(a+N)^{r+3}} + \frac{r(r+1)(r+2)}{(a+np)^{r+3}}\right)^2} \\ &\lesssim \sqrt{(1+np)^3} \times \frac{1}{(a+np)^{r+3}} \le \frac{1}{(a+np)^{r+3/2}}. \end{aligned}$$

The final assertion follows by taking the expectation with respect to N on both sides of the Taylor expansion, which completes the proof of Lemma 28.

D.5 Proof of Lemma 29

We define an auxiliary function

$$F(x) = \frac{1}{(a+x)^r (b+x)^s}.$$

Note that the derivatives of function $F(x)$ are given by

$$\begin{aligned} F'(x) &= -\frac{r}{(a+x)^{r+1}(b+x)^s} - \frac{s}{(a+x)^r (b+x)^{s+1}}, \\ F''(x) &= \frac{r(r+1)}{(a+x)^{r+2}(b+x)^s} + \frac{s(s+1)}{(a+x)^r (b+x)^{s+2}} \\ &\quad + \frac{2rs}{(a+x)^{r+1}(b+x)^{s+1}}, \\ F'''(x) &= -\frac{r(r+1)(r+2)}{(a+x)^{r+3}(b+x)^s} - \frac{3sr(r+1)}{(a+x)^{r+2}(b+x)^{s+1}} \\ &\quad - \frac{s(s+1)(s+2)}{(a+x)^r (b+x)^{s+3}} - \frac{3rs(s+1)}{(a+x)^{r+1}(b+x)^{s+2}} \\ &=: -(F_1'''(x) + F_2'''(x) + F_3'''(x) + F_4'''(x)). \end{aligned}$$

Exploiting Talyor's theorem, we can expand function $F(N)$ around np and obtain that

$$F(N) = F(np) + (N-np)F'(np) + \frac{1}{2}(N-np)^2F''(np) + R_5,$$

in which the remainder R_5 depends on an unknown random parameter $\theta = \theta(N, np) \in (0,1)$ and satisfies

$$\begin{aligned} R_5 &= \frac{1}{6}(N-np)^3F'''(\theta N + (1-\theta)np) \\ &= -\frac{1}{6}(N-np)^3(F_1''' + F_2''' + F_3''' + F_4''')(\theta N + (1-\theta)np) \\ &=: R_{51} + R_{52} + R_{53} + R_{54}. \end{aligned}$$

Hence, it suffices to establish θ-independent bounds for these four terms above.

Notice that $F_1'''(x)$ can be written as a product of two positive and convex functions $r(r+1)(r+2)/(a+x)^{r+3}$ and $1/(b+x)^s$. Due to the convexity and positivity, it holds that

$$\begin{aligned} \frac{1}{(a+\theta N + (1-\theta)np)^{r+3}} &\le \frac{1}{(a+N)^{r+3}} + \frac{1}{(a+np)^{r+3}}, \\ \frac{1}{(b+\theta N + (1-\theta)np)^{s}} &\le \frac{1}{(b+N)^{s}} + \frac{1}{(b+np)^{s}} \end{aligned}$$

and thus,

$$\begin{aligned} &F_1'''(\theta N + (1-\theta)np) \\ &\le r(r+1)(r+2)\left(\frac{1}{(a+N)^{r+3}} + \frac{1}{(a+np)^{r+3}}\right)\left(\frac{1}{(b+N)^{s}} + \frac{1}{(b+np)^{s}}\right), \end{aligned}$$

the RHS of which is independent of θ.

It then follows from the Cauchy–Schwartz inequality that

$$\begin{aligned} \mathbb{E}|R_{51}| &\lesssim \mathbb{E}\left|(N-np)^3\left(\frac{1}{(a+N)^{r+3}} + \frac{1}{(a+np)^{r+3}}\right)\left(\frac{1}{(b+N)^{s}} + \frac{1}{(b+np)^{s}}\right)\right| \\ &\lesssim \sqrt{\mathbb{E}[(N-np)^6]} \times \sqrt{\mathbb{E}\left[\left(\frac{1}{(a+N)^{r+3}} + \frac{1}{(a+np)^{r+3}}\right)^2\left(\frac{1}{(b+N)^{s}} + \frac{1}{(b+np)^{s}}\right)^2\right]} \\ &\lesssim \sqrt{(1+np)^3} \times \frac{1}{(a+np)^{r+3}(b+np)^s} = \frac{1}{(a+np)^{r+3/2}(b+np)^s}. \end{aligned}$$

Thanks to the similarity in structures, we can obtain that

$$\mathbb{E}|R_{53}| = \mathbb{E}\left|\frac{1}{6}(N-np)^3F_3'''(\theta N + (1-\theta)np)\right| \lesssim \frac{1}{(a+np)^r(b+np)^{s+3/2}}.$$

We next focus on term in F_2'''. Observe that F_2''' can also be decomposed into a product of convex and positive functions $r(r+1)/(a+x)^{r+2}$ and $3s/(b+x)^s$. Then we have that

$$\begin{aligned} &F_2'''(\theta N + (1-\theta)np) \\ &\le 3sr(r+1)\left(\frac{1}{(a+N)^{r+2}} + \frac{1}{(a+np)^{r+2}}\right)\left(\frac{1}{(b+N)^{s+1}} + \frac{1}{(b+np)^{s+1}}\right), \end{aligned}$$

the RHS of which is independent of θ. An application of the Cauchy–Schwartz inequality yields that

$$\begin{aligned}
\mathbb{E}|R_{52}| &\lesssim \mathbb{E}\left|(N-np)^3\left(\frac{1}{(a+N)^{r+2}}+\frac{1}{(a+np)^{r+2}}\right)\left(\frac{1}{(b+N)^{s+1}}+\frac{1}{(b+np)^{s+1}}\right)\right| \\
&\lesssim \sqrt{\mathbb{E}[(N-np)^6]}\times\sqrt{\mathbb{E}\left[\left(\frac{1}{(a+N)^{r+2}}+\frac{1}{(a+np)^{r+2}}\right)^2\left(\frac{1}{(b+N)^{s+1}}+\frac{1}{(b+np)^{s+1}}\right)^2\right]} \\
&\lesssim \sqrt{(1+np)^3}\times\frac{1}{(a+np)^{r+2}(b+np)^{s+1}} \\
&= \frac{1}{(a+np)^{r+1}(b+np)^{s+1/2}}.
\end{aligned}$$

It follows from the similarity in structures that

$$\mathbb{E}|R_{54}| = \mathbb{E}\left|\frac{1}{6}(N-np)^3F_4'''(\theta N+(1-\theta)np)\right| \lesssim \frac{1}{(a+np)^{r+1/2}(b+np)^{s+1}}.$$

Therefore, combining the results above, we can obtain that

$$\begin{aligned}
\mathbb{E}|R_5| \lesssim{}& \frac{1}{(a+np)^{r+3/2}(b+np)^s}+\frac{1}{(a+np)^r(b+np)^{s+3/2}} \\
&+\frac{1}{(a+np)^{r+1}(b+np)^{s+1/2}}+\frac{1}{(a+np)^{r+1/2}(b+np)^{s+1}}.
\end{aligned}$$

The final assertion follows by integrating out N in the Taylor expansion, which concludes the proof of Lemma 29.

D.6 Proof of Lemma 30

Let us define the auxiliary function as

$$H(x,y)=\frac{1}{(a+x)(b+y)}.$$

It is easy to verify that the derivatives of H with respect to x and y are given by

$$\begin{aligned}
H_x(x,y) &= -\frac{1}{(a+x)^2(b+y)}, \qquad H_y(x,y) = -\frac{1}{(a+x)(b+y)^2}, \\
H_{xx}(x,y) &= \frac{2}{(a+x)^3(b+y)}, \qquad H_{yy}(x,y) = \frac{2}{(a+x)(b+y)^3}, \\
H_{xy}(x,y) &= H_{yx}(x,y) = \frac{1}{(a+x)^2(b+y)^2}, \\
H_{xxx}(x,y) &= -\frac{6}{(a+x)^4(b+y)}, \qquad H_{yyy}(x,y) = -\frac{6}{(a+x)(b+y)^4}, \\
H_{xxy}(x,y) &= H_{xyx}(x,y) = H_{yxx}(x,y) = -\frac{2}{(a+x)^3(b+y)^2}, \\
H_{yyx}(x,y) &= H_{yxy}(x,y) = H_{xyy}(x,y) = -\frac{2}{(a+x)^2(b+y)^3}.
\end{aligned}$$

Applying Taylor's theorem, we can expand function $H(N_1, N_2)$ around (nq_1, nq_2) as

$$\begin{aligned} H(N_1, N_2) = & H(nq_1, nq_2) \\ & + (N_1 - nq_1)H_x(nq_1, nq_2) + (N_2 - nq_2)H_y(nq_1, nq_2) \\ & + \frac{1}{2}(N_1 - nq_1)^2 H_{xx}(nq_1, nq_2) + \frac{1}{2}(N_2 - nq_2)^2 H_{yy}(nq_1, nq_2) \\ & + (N_1 - nq_1)(N_2 - nq_2)H_{xy}(nq_1, nq_2) + R_6, \end{aligned}$$

where the remainder R_6 is expressed as

$$\begin{aligned} R_6 &:= R_{61} + R_{62} + R_{63} + R_{64} \\ &= \frac{1}{6}(N_1 - nq_1)^3 H_{xxx}(\theta N_1 + (1-\theta)nq_1, \theta N_2 + (1-\theta)nq_2) \\ &+ \frac{1}{6}(N_2 - nq_2)^3 H_{yyy}(\theta N_1 + (1-\theta)nq_1, \theta N_2 + (1-\theta)nq_2) \\ &+ \frac{1}{2}(N_1 - nq_1)^2(N_2 - nq_2) H_{xxy}(\theta N_1 + (1-\theta)nq_1, \theta N_2 + (1-\theta)nq_2) \\ &+ \frac{1}{2}(N_1 - nq_1)(N_2 - nq_2)^2 H_{yyx}(\theta N_1 + (1-\theta)nq_1, \theta N_2 + (1-\theta)nq_2) \end{aligned}$$

with an unknown random parameter $\theta = \theta(N_1, N_2, nq_1, nq_2) \in (0, 1)$. In what follows, we will establish θ-independent bounds for the four remainder terms above.

To bound the first term R_{61} involving H_{xxx}, a useful observation is that $H_{xxx} = -F(x)G(y)$ is separable, where

$$F(x) = \frac{1}{(a+x)^4} \quad \text{and} \quad G(y) = \frac{1}{b+y}$$

are univariate positive convex functions. The convexity and positivity of the functions entail that for any $\theta \in (0, 1)$,

$$F(\theta x_1 + (1-\theta)x_2)G(\theta y_1 + (1-\theta y_2)) \le (F(x_1) + F(x_2))\,(G(y_1) + G(y_2))\,,$$

in which the RHS is independent of θ. As a result, it holds that

$$\begin{aligned} \mathbb{E}|R_{61}| &\lesssim \mathbb{E}\left|(N_1 - nq_1)^3 \left(\frac{1}{(a+N_1)^4} + \frac{1}{(a+nq_1)^4}\right)\left(\frac{1}{b+N_2} + \frac{1}{b+nq_2}\right)\right| \\ &\lesssim \sqrt{\mathbb{E}(N_1 - nq_1)^6}\sqrt{\mathbb{E}\left\{\left(\frac{1}{(a+N_1)^4} + \frac{1}{(a+nq_1)^4}\right)^2\left(\frac{1}{b+N_2} + \frac{1}{b+nq_2}\right)^2\right\}} \\ &\lesssim \sqrt{\mathbb{E}(N_1 - nq_1)^6}\sqrt{\mathbb{E}\left\{\left(\frac{1}{(a+N_1)^8} + \frac{1}{(a+nq_1)^8}\right)\left(\frac{1}{(b+N_2)^2} + \frac{1}{(b+nq_2)^2}\right)\right\}}. \end{aligned}$$

On the one hand, we have that

$$\mathbb{E}(N_1 - nq_1)^6 \lesssim (nq_1 + 1)^3.$$

On the other hand, it follows from Lemma 27 that

$$\mathbb{E}\left[\frac{1}{(a+N_1)^8(b+N_2)^2}\right] \le \sqrt{\mathbb{E}\left(\frac{1}{(a+N_1)^{16}}\right)\mathbb{E}\left(\frac{1}{(b+N_2)^4}\right)} \lesssim \frac{1}{(a+nq_1)^8(b+nq_2)^2}$$

and thus

$$\mathbb{E}\left\{\left(\frac{1}{(a+N_1)^8}+\frac{1}{(a+nq_1)^8}\right)\left(\frac{1}{(b+N_2)^2}+\frac{1}{(b+nq_2)^2}\right)\right\} \lesssim \frac{1}{(a+nq_1)^8(b+nq_2)^2}.$$

Combining the results above leads to

$$\mathbb{E}|R_{61}| \lesssim \frac{(nq_1+1)^{3/2}}{(a+nq_1)^4(b+nq_2)} \lesssim \frac{1}{(a+nq_1)^{5/2}(b+nq_2)}.$$

Similarly, for term R_{62} we can obtain that

$$\mathbb{E}|R_{62}| \le C'\frac{(nq_2+1)^{3/2}}{(a+nq_1)(b+nq_2)^4} \le C'\frac{1}{(a+nq_1)(b+nq_2)^{5/2}}.$$

Next, we proceed with bounding term R_{63}. Similar to the arguments for term R_{61}, notice that $-H_{xxy}(x,y) = F'(x)G'(y)$ can be separated into the product of two univariate positive convex functions, where

$$F'(x) = \frac{1}{(a+x)^3} \quad \text{and} \quad G'(y) = \frac{1}{(b+y)^2}.$$

Then for any $\theta \in (0,1)$, we can establish a θ-independent bound for

$$F'(\theta x_1 + (1-\theta)x_2)G'(\theta y_1 + (1-\theta y_2)) \le \left(F'(x_1)+F'(x_2)\right)\left(G'(y_1)+G'(y_2)\right).$$

As a result, applying Hölder's inequality repeatedly, we can deduce that

$$\begin{aligned}
\mathbb{E}|R_{63}| &\lesssim \mathbb{E}\left|(N_1-nq_1)^2(N_2-nq_2)\left(\frac{1}{(a+N_1)^3}+\frac{1}{(a+nq_1)^3}\right)\left(\frac{1}{(b+N_2)^2}+\frac{1}{(b+nq_2)^2}\right)\right| \\
&\lesssim \sqrt{\mathbb{E}(N_1-nq_1)^4(N_2-nq_2)^2} \\
&\quad \times \sqrt{\mathbb{E}\left\{\left(\frac{1}{(a+N_1)^3}+\frac{1}{(a+nq_1)^3}\right)^2\left(\frac{1}{(b+N_2)^2}+\frac{1}{(b+nq_2)^2}\right)^2\right\}} \\
&\lesssim \sqrt{\left(\mathbb{E}(N_1-nq_1)^6\right)^{2/3}\left(\mathbb{E}(N_2-nq_2)^6\right)^{1/3}} \\
&\quad \times \sqrt{\mathbb{E}\left\{\left(\frac{1}{(a+N_1)^3}+\frac{1}{(a+nq_1)^3}\right)^2\left(\frac{1}{(b+N_2)^2}+\frac{1}{(b+nq_2)^2}\right)^2\right\}} \\
&\lesssim \frac{(nq_1+1)(nq_2+1)^{1/2}}{(a+nq_1)^3(b+nq_2)^2} \lesssim \frac{1}{(a+nq_1)^2(b+nq_2)^{3/2}}.
\end{aligned}$$

Similarly, we can also obtain that

$$\mathbb{E}|R_{64}| \lesssim \frac{1}{(a+nq_1)^{3/2}(b+nq_2)^2}.$$

Consequently, we see that remainder R_6 satisfies that

$$\begin{aligned}\mathbb{E}|R_6| \lesssim & \frac{1}{(a+nq_1)^{5/2}(b+nq_2)} + \frac{1}{(a+nq_1)(b+nq_2)^{5/2}} \\ & + \frac{1}{(a+nq_1)^2(b+nq_2)^{3/2}} + \frac{1}{(a+nq_1)^{3/2}(b+nq_2)^2}.\end{aligned}$$

Therefore, we can conclude by integrating out N in the leading terms of the Taylor expansion, which completes the proof of Lemma 30.

D.7 Proof of Lemma 31

Denote by $P_1 = P\backslash P_0$ and $P_2 = P'\backslash P_0$. The sample counts for subsets P_0, P_1, and P_2 are denoted as N_0, N_1, and N_2, respectively. It is clear that $N_i \sim \mathcal{B}(n, p_i)$, where the success probability $p_i = \mathbb{P}(\boldsymbol{X} \in P_i)$ for $i = 0, 1, 2$. In addition, it holds that $(N_0, N_1, N_2) \sim \mathcal{M}(n, (p_0, p_1, p_2))$.

A simple observation is that $N = N_0 + N_1$ and $N' = N_0 + N_2$. Let us first consider the conditional expectation of the target inverse moment given $N_0 = n_0$. Recall that $(N_1, N_2)|N_0 = n_0 \sim \mathcal{M}(n', (p_1', p_2'))$, in which

$$n' = n - n_0, \quad p_1' = \frac{p_1}{1-p_0}, \quad p_2' = \frac{p_2}{1-p_0}. \tag{D.52}$$

With an application of Lemma 30, we can deduce that

$$\begin{aligned}& \mathbb{E}\left[\frac{1}{(a+n_0+N_1)(a+n_0+N_2)}|N_0 = n_0\right] \\ =& \frac{1}{(a+n_0+n'p_1')(a+n_0+n'p_2')} \\ +& \frac{n'p_1'(1-p_1')}{(a+n_0+n'p_1')^3(a+n_0+n'p_2')} \\ +& \frac{n'p_2'(1-p_2')}{(a+n_0+n'p_1')(a+n_0+n'p_2')^3} \\ -& \frac{n'p_1'p_2'}{(a+n_0+n'p_1')^2(a+n_0+n'p_2)^2} + R_4(n_0) \\ =:& L_{71}(n_0) + L_{72}(n_0) + L_{73}(n_0) + L_{74}(n_0) + R_7(n_0),\end{aligned} \tag{D.53}$$

in which remainder R_7 satisfies

$$\begin{aligned}R_7(n_0) \lesssim & \frac{1}{(a+n_0+n'p_1')^{5/2}(a+n_0+n'p_2')} + \frac{1}{(a+n_0+n'p_1')(a+n_0+n'p_2')^{5/2}} \\ & + \frac{1}{(a+n_0+n'p_1')^2(a+n_0+n'p_2')^{3/2}} + \frac{1}{(a+n_0+n'p_1')^{3/2}(a+n_0+n'p_2')^2}.\end{aligned}$$

To further integrate with respect to N_0, we need to separate n_0 from n' in the present expressions. Notice that $n_0 + n'p_1' = np_1' + n_0(1-p_1')$, $n_0 + n'p_2' = np_2' + n_0(1-p_2')$. In addition, we have that

$$a + np_1' + np_0(1-p_1') = a + \frac{n}{1-p_0}(p_1 + p_0(1-p_0-p_1)) = a + n(p_1+p_0) = a + np$$

and $a + np_2' + np_0(1-p_2') = a + np'$. Since $N_0 \sim \mathcal{B}(n, p_0)$, an application of Lemma 29 with $r = s = 1$ yields that

$$\begin{aligned}\mathbb{E}[L_{71}(N_0)] = &\frac{1}{(a+np)(a+np')} + \frac{np_0(1-p_0)(1-p_1')^2}{(a+np)^3(a+np')} \\ &+ \frac{np_0(1-p_0)(1-p_2')^2}{(a+np))(a+np')^3} + \frac{np_0(1-p_0)(1-p_1')(1-p_2')}{(a+np)^2(a+np')^2} + W_{71},\end{aligned} \tag{D.54}$$

where remainder W_{41} satisfies

$$\begin{aligned}W_{71} \lesssim &\frac{1}{(a+np)^{5/2}(a+np')} + \frac{1}{(a+np)(a+np')^{5/2}} \\ &+ \frac{1}{(a+np)^2(a+np')^{3/2}} + \frac{1}{(a+np)^{3/2}(a+np')^2}.\end{aligned}$$

Meanwhile, by invoking Lemma 27, we see that remainder R_7 satisfies

$$\begin{aligned}\mathbb{E}[R_7(N_0)] \lesssim &\frac{1}{(a+np)^{5/2}(a+np')} + \frac{1}{(a+np)(a+np')^{5/2}} \\ &+ \frac{1}{(a+np)^2(a+np')^{3/2}} + \frac{1}{(a+np)^{3/2}(a+np')^2}.\end{aligned}$$

Observe that

$$\begin{aligned}\frac{n'p_1'}{a+n_0+n'p_1'} &= \frac{np_1' - n_0p_1'}{a+np_1'+n_0(1-p_1')} = \frac{np_1' - n_0p_1'}{a+np_1'+n_0(1-p_1')} \\ &= \frac{np_1'}{a+np_1'+n_0(1-p_1')} - \frac{p_1'}{1-p_1'}\left(1 - \frac{a+np_1'}{a+np_1'+n_0(1-p_1')}\right) \\ &= \frac{1}{1-p_1'}\left(\frac{np_1'+ap_1'}{a+np_1'+n_0(1-p_1')} - p_1'\right)\end{aligned}$$

and similarly,

$$\frac{n'p_2'}{a+n_0+n'p_2'} = \frac{1}{1-p_2'}\left(\frac{np_2'+ap_2'}{a+np_2'+n_0(1-p_2')} - p_2'\right).$$

Then it follows that

$$\begin{aligned}L_{72}(n_0) &= \frac{n'p_1'(1-p_1')}{(a+np_1'+n_0(1-p_1'))^3(a+np_2'+n_0(1-p_2'))} \\ &= \left(\frac{np_1'+ap_1'}{a+np_1'+n_0(1-p_1')} - p_1'\right)(a+np_1'+n_0(1-p_1'))^2(a+np_2'+n_0(1-p_2')) \\ &= -\frac{p_1'}{(a+np_1'+n_0(1-p_1'))^2(a+np_2'+n_0(1-p_2'))} \\ &+ \frac{np_1'+ap_1'}{(a+np_1'+n_0(1-p_1'))^3(a+np_2'+n_0(1-p_2'))}.\end{aligned}$$

Applying Lemma 29 with $r = 2, s = 1$ and $r = 3, s = 1$, respectively, we can deduce that

$$\begin{aligned}\mathbb{E}[L_{72}(N_0)] &= -\frac{p_1'}{(a+np_1'+n_0(1-p_1'))^2(a+np_2'+n_0(1-p_2'))} \\ &\quad + \frac{np_1'+ap_1'}{(a+np_1'+n_0(1-p_1'))^3(a+np_2'+n_0(1-p_2'))} \\ &= -\frac{p_1'}{(a+np)^2(a+np')} + \frac{np_1'+ap_1'}{(a+np)^3(a+np')} + W_{72} \\ &= \frac{np_1'(1-p)}{(a+np)^3(a+np')} + W_{72},\end{aligned} \tag{D.55}$$

where remainder W_{72} satisfies

$$W_{72} \lesssim \frac{1}{(a+np)^3(a+np')} + \frac{1}{(a+np)^2(a+np')^2}.$$

Similarly, we can also establish that

$$\begin{aligned}\mathbb{E}[L_{73}(N_0)] &= -\frac{p_2'}{(a+np_1'+n_0(1-p_1'))(a+np_2'+n_0(1-p_2'))^2} \\ &\quad + \frac{np_2'+ap_2'}{(a+np_1'+n_0(1-p_1'))(a+np_2'+n_0(1-p_2'))^3} \\ &= -\frac{p_2'}{(a+np)(a+np')^2} + \frac{np_2'+ap_2'}{(a+np)(a+np')^3} + W_{73} \\ &= \frac{np_2'(1-p)}{(a+np)(a+np')^3} + W_{73},\end{aligned} \tag{D.56}$$

where remainder W_{73} satisfies

$$W_{73} \lesssim \frac{1}{(a+np)(a+np')^3} + \frac{1}{(a+np)^2(a+np')^2}.$$

Moreover, it holds that

$$\begin{aligned}L_{74}(n_0) &= -\frac{n'p_1'p_2'}{(a+np_1'+n_0(1-p_1'))^2(a+np_2'+n_0(1-p_2'))^2} \\ &= -\frac{p_2'}{1-p_1'}\left(\frac{np_1'+ap_1'}{a+np_1'+n_0(1-p_1')} - p_1'\right) \\ &\quad \times \frac{1}{(a+np_1'+n_0(1-p_1'))(a+np_2'+n_0(1-p_2'))^2}.\end{aligned}$$

Hence, we have that

$$\begin{aligned}\mathbb{E}[L_{74}(N_0)] &= -\frac{p_2'}{1-p_1'}\left(\frac{np_1'+ap_1'}{a+np} - p_1'\right)\frac{1}{(a+np)(a+np')^2} + W_{74} \\ &= -\frac{np_1p_2}{(1-p_0)(a+np)^2(a+np')^2} + W_{74},\end{aligned} \tag{D.57}$$

where remainder W_{74} satisfies

$$W_{74} \lesssim \frac{1}{(a+np)^2(a+np')^2} + \frac{1}{(a+np)(a+np')^3}.$$

Substituting (D.54)–(D.57) into (D.53), we can obtain that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{(a+N)(a+N')}\right] &= \frac{1}{(a+np)(a+np')} + \frac{np_0(1-p_0)(1-p_1')^2}{(a+np)^3(a+np')} \\ &\quad + \frac{np_1'(1-p)}{(a+np)^3(a+np')} + \frac{np_0(1-p_0)(1-p_2')^2}{(a+np))(a+np')^3} + \frac{np_2'(1-p)}{(a+np)(a+np')^2} \\ &\quad + \frac{np_0(1-p_0)(1-p_1')(1-p_2')}{(a+np)^2(a+np')^2} - \frac{np_1'p_2'(1-p_0)}{(a+np)^2(a+np')^2} + \widetilde{R}_7,\end{aligned}$$

where remainder $\widetilde{R}_7$ satisfies

$$\begin{aligned}\widetilde{R}_7 &\lesssim \frac{1}{(a+np)^{5/2}(a+np')} + \frac{1}{(a+np)(a+np')^{5/2}} \\ &\quad + \frac{1}{(a+np)^2(a+np')^{3/2}} + \frac{1}{(a+np)^{3/2}(a+np')^2}.\end{aligned}$$

Some simple calculations lead to

$$\begin{aligned}&\frac{np_0(1-p_0)(1-p_1')^2}{(a+np)^3(a+np')} + \frac{np_1'(1-p)}{(a+np)^3(a+np')} \\ &\quad = \frac{n(1-p)}{(1-p_0)(a+np)^3(a+np')}\left(p_0(1-p)+p_1\right) \\ &\quad = \frac{np(1-p)}{(a+np)^3(a+np')} \\ &\quad = \frac{1-p}{(a+np)^2(a+np')} - \frac{a(1-p)}{(a+np)^3(a+np')}, \\ &\frac{np_0(1-p_0)(1-p_2')^2}{(a+np))(a+np')^3} + \frac{np_2'(1-p)}{(a+np)(a+np')^2} = \frac{np'(1-p')}{(a+np)(a+np')^3} \\ &\quad = \frac{1-p'}{(a+np)(a+np')^2} - \frac{a(1-p')}{(a+np)(a+np')^3}, \\ &\frac{np_0(1-p_0)(1-p_1')(1-p_2')}{(a+np)^2(a+np')^2} - \frac{np_1'p_2'(1-p_0)}{(a+np)^2(a+np')^2} \\ &\quad = \frac{n(1-p_0)}{(a+np)^2(a+np')^2}(p_0(1-p_1')(1-p_2') - p_1'p_2') \\ &\quad = \frac{n}{(1-p_0)(a+np)^2(a+np')^2}(p_0(1-p)(1-p') - p_1p_2) \\ &\quad = \frac{n(p_0 - p_0(p+p') + p_0pp' - (p-p_0)(p'-p_0))}{(1-p_0)(a+np)^2(a+np')^2} \\ &\quad = \frac{n(p_0-pp')}{(a+np)^2(a+np')^2}.\end{aligned}$$

Therefore, substituting the simplified expressions above into the asymptotic expansion of the inverse moment, we can obtain that

$$\mathbb{E}\left[\frac{1}{(a+N)(a+N')}\right] = \frac{1}{(a+np)(a+np')} + \frac{1-p}{(a+np)^2(a+np')} + \frac{1-p'}{(a+np)(a+np')^2} + \frac{n(p_0-pp')}{(a+np)^2(a+np')^2} + \widetilde{R}_7',$$

where remainder $\widetilde{R}_7'$ has the same bound as $\widetilde{R}_7$. The final conclusion follows from a further step of simplification, which concludes the proof of Lemma 31.

References

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

Simon Bernard, Laurent Heutte, and Sébastien Adam. Influence of hyperparameters on random forest accuracy. In *Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings 8*, pages 171–180. Springer, 2009.

Gérard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(38):1063–1095, 2012. URL http://jmlr.org/papers/v13/biau12a.html.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002. doi: 10.1214/aos/1031689014. URL https://doi.org/10.1214/aos/1031689014.

Matias D. Cattaneo, Jason M. Klusowski, and William G. Underwood. Inference with mondrian random forests. *Manuscript*, 2024. URL https://arxiv.org/abs/2310.09702. arXiv:2310.09702.

Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022. doi: 10.1214/22-AOS2234. URL https://doi.org/10.1214/22-AOS2234.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. Why do random forests work? understanding tree ensembles as self-regularizing adaptive smoothers. *Manuscript*, 2024. URL https://arxiv.org/abs/2402.01502. arXiv: 2402.01502.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014. URL http://jmlr.org/papers/v15/delgado14a.html.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Jason Klusowski. Sharp analysis of a simple model for random forests. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 757–765, 2021. URL https://proceedings.mlr.press/v130/klusowski21b.html.

Jason M. Klusowski. Analyzing CART. *Manuscript*, 2020. URL https://arxiv.org/abs/1906.10086. arXiv:1906.10086.

Jason M. Klusowski and Peter M. Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024. doi: 10.1080/01621459.2022.2126782. URL https://doi.org/10.1080/01621459.2022.2126782.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020. URL http://jmlr.org/papers/v21/19-844.html.

Hong-Lam Le, Thanh-Tuoi Le, Doan-Hieu Tran, Dinh Van Chau, et al. A survey on the impact of hyperparameters on random forest performance using multiple accelerometer datasets. *International Journal for Computers & Their Applications*, 30(4):351–361, 2023.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006. doi: 10.1198/016214505000001230. URL https://doi.org/10.1198/016214505000001230.

Brian Liu and Rahul Mazumder. Randomization can reduce both bias and variance: A case study in random forests. *Manuscript*, 2024. URL https://arxiv.org/abs/2402.12668. arXiv:2402.12668.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*, volume 143. Springer, second edition, 2011. doi: 10.1007/978-0-387-68276-1.

Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020. URL http://jmlr.org/papers/v21/19-905.html.

Lucas Mentch and Siyu Zhou. Getting better from worse: Augmented bagging and a cautionary tale of variable importance. *Journal of Machine Learning Research*, 23(224):1–32, 2022. URL http://jmlr.org/papers/v23/20-1264.html.

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for Mondrian trees and forests. *The Annals of Statistics*, 48(4):2253–2276, 2020. doi: 10.1214/19-AOS1886. URL https://doi.org/10.1214/19-AOS1886.

Philipp Probst and Anne-Laure Boulesteix. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181):1–18, 2018. URL http://jmlr.org/papers/v18/17-269.html.

Erwan Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83, 2016. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2015.06.009. URL https://www.sciencedirect.com/science/article/pii/S0047259X15001542.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015. doi: 10.1214/15-AOS1321. URL https://doi.org/10.1214/15-AOS1321.

Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3453–3454, 2020. URL https://proceedings.mlr.press/v125/syrgkanis20a.html.

Yan Shuo Tan, Jason M. Klusowski, and Krishnakumar Balasubramanian. Statistical-computational trade-offs for greedy recursive partitioning estimators. *Manuscript*, 2024. URL https://arxiv.org/abs/2411.04394. arXiv:2411.04394.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

Siyu Zhou. *Random Forests and Regularization*. PhD thesis, University of Pittsburgh, 2022.

Siyu Zhou and Lucas Mentch. Trees, forests, chickens, and eggs: when and why to prune trees in a random forest. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 16(1):45–64, 2023.