Journal of Machine Learning Research 25 (2024) 1-81 Submitted 5/23; Revised 10/24; Published 10/24

Optimal Weighted Random Forests

Xinyu Chen SA21204192@mail.ustc.edu.cn
International Institute of Finance
School of Management
University of Science and Technology of China
Hefei, 230026, Anhui, China

Dalei Yu* yudalei@126.com
School of Mathematics and Statistics
Xi'an Jiaotong University
Xi'an, 710049, Shaanxi, China

Xinyu Zhang xinyu@amss.ac.cn
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190, China
International Institute of Finance
School of Management
University of Science and Technology of China
Hefei, 230026, Anhui, China

Editor: Mladen Kolar

Abstract

The random forest (RF) algorithm has become a very popular prediction method for its great flexibility and promising accuracy. In RF, it is conventional to put equal weights on all the base learners (trees) to aggregate their predictions. However, the predictive performance of different trees within the forest can vary significantly due to the randomization of the embedded bootstrap sampling and feature selection. In this paper, we focus on RF for regression and propose two optimal weighting algorithms, namely the 1 Step Optimal Weighted RF (1step-WRF$_{\text{opt}}$) and 2 Steps Optimal Weighted RF (2steps-WRF$_{\text{opt}}$), that combine the base learners through the weights determined by weight choice criteria. Under some regularity conditions, we show that these algorithms are asymptotically optimal in the sense that the resulting squared loss and risk are asymptotically identical to those of the infeasible but best possible weighted RF. Numerical studies conducted on real-world data sets and semi-synthetic data sets indicate that these algorithms outperform the equal-weight forest and two other weighted RFs proposed in the existing literature in most cases.

Keywords: weighted random forest, model averaging, optimality, splitting criterion

*. Dalei Yu is the corresponding author.

1. Introduction

Random forest (RF) (Breiman, 2001), growing trees using Classification and Regression Trees (CART) algorithm, is one of the most successful machine learning algorithms that scale with the volume of information while maintaining sufficient statistical efficiency (Biau and Scornet, 2016). Due to its great flexibility and promising accuracy, RF has been widely used in diverse areas of data analysis, including policy-making (Yoon, 2021; Lin et al., 2021), business analysis (Pallathadka et al., 2023; Ghosh et al., 2022), chemoinformatics (Svetnik et al., 2003), real-time recognition of human pose (Shotton et al., 2011), and so on. RF ensembles multiple decision trees grown on bootstrap samples and yields highly accurate predictions. In the conventional implementation of RF, it is customary and convenient to allocate equal weight to each decision tree. Theoretically, the predictive performance varies from tree to tree due to the application of randomly selected sub-spaces of data and features. In other words, trees exhibit greater diversity due to the injected randomness. An immediate question then arises: Is it always optimal to employ equal weights? In fact, there is sufficient evidence to indicate that an averaging strategy with appropriately selected unequal weights may achieve better performance than simple averaging (that is, equally weighting) if individual learners exhibit non-identical strength (Zhou, 2012; Peng and Yang, 2022).

To solve the problem mentioned above, some efforts have been made in the literature regarding weighted RFs. Specifically, Trees Weighting Random Forest (TWRF) introduced by Li et al. (2010) employs the accuracy in the out-of-bag data as an index that measures the classification power of the tree and sets it as the weight. Winham et al. (2013) develop Weighted Random Forests (wRF), where the weights are determined based on tree-level prediction error. Based on wRF, Xuan et al. (2018) put forward Refined Weighted Random Forests (RWRF) using all training data, including in-bag and out-of-bag data. A novel weights formula is also developed in RWRF but cannot be manipulated into a regression pattern. Pham and Olafsson (2019) replace the regular average with a Cesáro average with theoretical analysis. However, these studies have predominantly focused on classification, and less attention has been paid to the regression pattern (that is, estimating the conditional expectation), although some mechanisms for classification can be transformed into corresponding regression patterns. In addition, none of the aforementioned studies have investigated the theoretical underpinnings regarding the optimality properties of their methods.

Recently, Qiu et al. (2020) propose a novel framework that averages the outputs of multiple machine learning algorithms by the weights determined by Mallows-type criteria. Motivated by their work, we extend this approach by developing an asymptotically optimal weighting strategy for RF. Specifically, we treat the individual trees within the RF as base learners and employ Mallows-type criterion to obtain their respective weights. It is worth noting that Qiu et al. (2020) implicitly assume the independence of the "hat matrix" from the response values (the term "hat matrix" generally refers to the matrix that maps the response vector to the corresponding vector of fitting values), which is deemed unrealistic in practical situations. In the current study, we remove this restriction and establish the asymptotic optimality under the standard setting where the "hat matrix" is determined by a response-based splitting criterion (Breiman, 2001; Chi et al., 2022). Moreover, to reduce

computational burden, we further propose an accelerated algorithm that requires only two quadratic optimization tasks. Asymptotic optimality is established for both the original and accelerated weighted RF estimators. Extensive analyses on real-world and semi-synthetic data sets demonstrate that the proposed methods show promising performance over existing RFs.

The remaining part of the paper proceeds as follows: Section 2 formulates the problem. Section 3 establishes our weighted RF algorithms and provides theoretical analysis. Section 4 shows their promising performance on 12 real-world data sets from UCI Machine Learning Repository and `Openml.org`, as well as on semi-synthetic data. Section 5 concludes. The data and codes are available publicly on https://github.com/XinyuChen-hey/Optimal-Weighted-Random-Forests.

2. Model and Problem Formulation

Let $\mathbf{x}_i^0 = (x_{i1}, x_{i2}, \ldots)^\top$ be a set of countably infinite predictors (or explanatory variables, attributes, features) and y_i be a univariate response variable for $i = 1, \ldots, n$. The data generating process is as follows

$$y_i = \mu_i + e_i,$$

where $\{e_i, \mathbf{x}_i^0\}_{i=1}^n$ are independent and identically distributed random variables with $\mathbb{E}(e_i|\mathbf{x}_i^0) = 0$ and $\mathbb{E}(e_i^2|\mathbf{x}_i^0) = \sigma_i^2$, and $\mu_i = \mathbb{E}(y_i|\mathbf{x}_i^0)$. So conditional heteroscedasticity is allowed here.

Consider an observable data set $\mathscr{D} = \{y_i, \mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ with p being the dimension of the feature. Given a predictor vector $\mathbf{x}_i$, the corresponding prediction for y_i by a tree (or base learner, BL) in the construction of RF can be written as follows

$$\widehat{y}_i = \mathbf{P}_{\mathrm{BL}}^\top(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})\mathbf{y},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is the vector of the response variable, $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ is an n-dimensional vector for tree configuration, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is the matrix of predictors. The variables $\mathscr{B}$ and $\boldsymbol{\Theta}$ play implicit roles in injecting randomness into RFs. First, each tree is fitted to an independent bootstrap sample from the original data, with $\mathscr{B}$ determining the randomization inherent in this bootstrap sampling process. Second, $\boldsymbol{\Theta}$ dictates the randomness in feature selection at each node within the trees. The specific nature and dimensions of $\mathscr{B}$ and $\boldsymbol{\Theta}$ are contingent upon the specific implementation of each tree. Note that $\mathscr{B}$ is irrelevant to $\mathbf{y}$, while $\boldsymbol{\Theta}$ relies on $\mathbf{y}$ for guiding splits.[1]

Let us assume that we have drawn M_n bootstrap data sets of size n and grown M_n trees on their bootstrap data, where M_n can grow with n or remain fixed. Take the m^{th} tree for example. Dropping an instance $(y_0, \mathbf{x}_0)$ down this base learner and end up with a specific tree leaf l with n_l observations $\mathscr{D}_l = \{(y_{i_1}, \mathbf{x}_{i_1}), \ldots, (y_{i_{n_l}}, \mathbf{x}_{i_{n_l}})\}$. Assume that the number of occurrences of instance $(y_i, \mathbf{x}_i)$ in this tree is h_i for all i because of the bootstrap sampling procedure. Then $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_0, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ for this tree is a sparse vector, with elements being h_i/n_l or zero, depending on the relationship between $\mathscr{D}$ and $\mathscr{D}_l$. More specifically, the i^{th} element of $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_0, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ equals 0 if $(y_i, \mathbf{x}_i) \notin \mathscr{D}_l$ and $(y_0, \mathbf{x}_0) \in \mathscr{D}_l$. Elements of $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_0, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ are weights put on elements of $\mathbf{y}$ to make a prediction for y_0.

1. In case where the tree structure is unaffected by $\mathbf{y}$, $\boldsymbol{\Theta}$ becomes independent of $\mathbf{y}$, therefore reducing $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ to $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_i, \mathbf{X}, \mathscr{B}, \boldsymbol{\Theta})$. Such a situation occurs with split-unsupervised trees as discussed in Section 3.2.

By randomly selecting sub-spaces of data and features, trees in RFs are given more randomness than trees without these randomization techniques. Specifically, bootstrap data are used to grow trees rather than the original training data. In addition, when searching for the best splitting variable at each node, we draw q ($q < p$) variables from the total pool of p variables, rather than using all p variables. If without the bootstrap procedure, we have $h_i \equiv 1$ for all $i \in \{1, \ldots, n\}$, and $\mathbf{P}_{\mathrm{BL}}(\mathbf{x}_0, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ will contain fewer zero elements.

The prediction for y_i by the m^{th} tree (or the m^{th} base learner) within the forest obeys the following relationship

$$\widehat{y}_i^{(m)} = \mathbf{P}_{\mathrm{BL}(m)}^{\top}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \boldsymbol{\Theta}_{(m)})\mathbf{y}, \tag{1}$$

where $\widehat{y}_i^{(m)}$ is the prediction for y_i by the m^{th} tree, and $\mathbf{P}_{\mathrm{BL}(m)}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \boldsymbol{\Theta}_{(m)})$ is the n-dimensional vector for configuring the m^{th} tree. The final output of the forest is integrated by

$$\widehat{y}_i(\mathbf{w}) = \sum_{m=1}^{M_n} w_{(m)} \widehat{y}_i^{(m)},$$

where $w_{(m)}$ is the weight put on the m^{th} tree. Our goal is to determine appropriate weights to improve prediction accuracy of RF, given a predictor vector $\mathbf{x}$. Clearly, the conventional RF has $w_{(m)} \equiv 1/M_n$ for $m = 1, \ldots, M_n$.

3. Mallows-type Weighted RFs

Let $\mathbf{P}_{\mathrm{BL}(m)}$ be an $n \times n$ "hat matrix", of which the i^{th} row is $\mathbf{P}_{\mathrm{BL}(m)}^{\top}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \boldsymbol{\Theta}_{(m)})$. Let

$$\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M_n} w_{(m)} \mathbf{P}_{\mathrm{BL}(m)},$$

and

$$\widehat{\mathbf{y}}(\mathbf{w}) = \sum_{m=1}^{M_n} w_{(m)} \widehat{\mathbf{y}}^{(m)},$$

with

$$\widehat{\mathbf{y}}^{(m)} = \left(\widehat{y}_1^{(m)}, \ldots, \widehat{y}_n^{(m)}\right)^{\top}.$$

Define the following averaged squared error function

$$L_n(\mathbf{w}) \equiv \|\widehat{\mathbf{y}}(\mathbf{w}) - \boldsymbol{\mu}\|^2, \tag{2}$$

which measures the sum of squared biases between the true $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\top}$ and its model averaging estimate $\widehat{\mathbf{y}}(\mathbf{w})$. Denote $R_n(\mathbf{w}) = \mathbb{E}\{L_n(\mathbf{w})|\mathscr{X}\}$, where $\mathscr{X}$ is the σ-algebra generated by $\{\mathbf{x}_1^0, \ldots, \mathbf{x}_n^0, \mathscr{B}_{(1)}, \ldots, \mathscr{B}_{(M_n)}, \boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(M_n)}\}$. We will propose some weight choice criteria to obtain weights based on $R_n(\mathbf{w})$.

3.1 Mallows-type Weight Choice Criteria

Considering the choice of weights, we use the solution obtained by minimizing the following Mallows-type criterion (3) with the restriction of $\mathbf{w} \in \mathscr{H} \equiv \left\{\mathbf{w} \in [0,1]^{M_n} : \sum_{m=1}^{M_n} w_{(m)} = 1\right\}$

$$C_n(\mathbf{w}) = \|\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\|^2 + 2\sum_{i=1}^{n} e_i^2 P_{ii}(\mathbf{w}), \tag{3}$$

where $P_{ii}(\mathbf{w})$ is the i^{th} diagonal term in $\mathbf{P}(\mathbf{w})$, and $\mathbf{e} = (e_1, \ldots, e_n)^\top$ is the true error term vector.

This criterion is originally proposed by Zhao et al. (2016) for considering linear models. In the context of linear models, $\mathbb{E}\{C_n(\mathbf{w}) \mid \mathcal{X}\}$ equals the conditional risk $R_n(\mathbf{w})$ up to a constant term that is irrelevant to $\mathbf{w}$. Zhao et al. (2016) further show that the criterion is asymptotically optimal in the context considered therein. However, e_i's are unobservable terms in practice. So they further consider the following feasible criterion, replacing the true error terms with averaged residuals

$$C'_n(\mathbf{w}) = \|\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\|^2 + 2\sum_{i=1}^{n} \hat{e}_i(\mathbf{w})^2 P_{ii}(\mathbf{w}), \tag{4}$$

where

$$\hat{\mathbf{e}}(\mathbf{w}) = \{\hat{e}_1(\mathbf{w}), \ldots, \hat{e}_n(\mathbf{w})\}^\top = \sum_{m=1}^{M_n} w_{(m)} \hat{\mathbf{e}}^{(m)} = \{\mathbf{I}_n - \mathbf{P}(\mathbf{w})\}\mathbf{y},$$

$\hat{\mathbf{e}}^{(m)}$ is the residual vector for the m^{th} candidate model, and $\mathbf{I}_n$ is an identity matrix with dimension n. This feasible criterion also accommodates conditional heteroscedasticity. Besides, it relies on all candidate models to estimate the true error vector, which avoids placing too much confidence on a single model. Similar criterion has also been considered in Qiu et al. (2020).

We apply criterion (4) to determine $\mathbf{w}$ in $\widehat{\mathbf{y}}(\mathbf{w})$. Criterion (4) comprises two terms. The first term measures the fitting error of the weighted RF in the training data, by computing the residual sum of squares. The second term penalizes the complexity of the trees in the forest. For each $m \in \{1, \ldots, M_n\}$, $P_{\text{BL}(m),ii}$ denotes the i^{th} diagonal term in $\mathbf{P}_{\text{BL}(m)}$. As explained in Section 2, $P_{\text{BL}(m),ii}$ is the proportion of the i^{th} observation to the total number of samples in the leaf that includes the i^{th} observation. Thus, for each $m \in \{1, \ldots, M_n\}$ and $i \in \{1, \ldots, n\}$, the larger the value of $P_{\text{BL}(m),ii}$, the smaller the gap between y_i and $\widehat{y}_i^{(m)}$. In the most extreme case where a tree is so deep that the leaf node containing the i^{th} observation is pure, $P_{\text{BL}(m),ii}$ equals 1, and $\widehat{y}_i^{(m)}$ equals y_i. Essentially, this tree has low prediction error within the training sample, but may exhibit poor generalization performance when applied to new data. To mitigate the contribution of overfitted trees in the ensemble output, this algorithm assigns lower weights to these trees, thereby decreasing the second term.

From another aspect, noting that $\sum_{i=1}^{n} \hat{e}_i(\mathbf{w})^2 P_{ii}(\mathbf{w}) \geqslant \min_{1 \leqslant i \leqslant n} \hat{e}_i(\mathbf{w})^2 \sum_{i=1}^{n} P_{ii}(\mathbf{w})$, the summation part $\sum_{i=1}^{n} P_{ii}(\mathbf{w}) = \sum_{m=1}^{M_n} w_{(m)} \sum_{i=1}^{n} P_{\text{BL}(m),ii} = \sum_{m=1}^{M_n} w_{(m)} \ell_{(m)}$ represents the weighted number of leaves of all trees, where $\ell_{(m)}$ is the number of leaves of the m^{th} tree, representing the complexity of the tree. The regularized objective for minimization in the

Extreme Gradient Boosting (XGBoost) algorithm, proposed by Chen and Guestrin (2016), also contains a penalty term that penalizes the number of leaves in the tree. In light of this, both the weighted RF with weights obtained by minimizing criterion (4) and XGBoost employ the number of leaves in a tree to measure its complexity. The regularization term helps to allocate the weights to avoid overfitting (Chen and Guestrin, 2016).

Inherent from the Mallows criterion (Hansen, 2007), the resulting weights of (4) exhibit sparsity. It is important to note that some trees within a RF might contribute to the deterioration of the ensemble's overall performance. Therefore, forming a more accurate committee via removal of trees with poor performance is a more rational strategy. It is called the theorem of MCBTA ("many could be better than all") introduced by Zhou et al. (2002), which indicates that for supervised learning, given a set of individual learners, it may be better to ensemble some instead of all of these individual learners. Employing sparse weights can be regarded as a form of adaptive tree selection, reducing the risk of integrating trees that could weaken the final outcome of the ensemble. Additionally, it offers advantages over model selection methods, which only choose a single tree and thereby ignore model uncertainty. In short, our approach with sparse weights provides a balanced and adaptive solution, selectively aggregating members while acknowledging the significance of trees diversity.

It is clear that criterion (4) is a cubic function of $\mathbf{w}$, whose optimization is substantially more time-consuming than that of quadratic programming. To expedite the process, we further suggest an accelerated algorithm that estimates $\mathbf{e}$ using a vector that is irrelevant to $\mathbf{w}$. The accelerated algorithm consists of two steps, where the first step involves calculating the estimated error terms, and the second step involves substituting the vector obtained in the first step for the true error terms in criterion (3). In specific, consider the following intermediate criterion,

$$C_n^{\circ}(\mathbf{w}) = \|\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\|^2 + 2\sum_{i=1}^{n} \hat{\sigma}^2 P_{ii}(\mathbf{w}), \tag{5}$$

where $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{P}(\mathbf{w}_0)\mathbf{y}\|^2/n$, and $\mathbf{w}_0$ is an n-dimensional vector with all elements being $1/M_n$. Solve this quadratic programming task over $\mathbf{w} \in \mathscr{H}$, and get a solution $\mathbf{w}^{\circ}$. Then, calculate the residual vector by

$$\tilde{\mathbf{e}} = (\tilde{e}_1, \ldots, \tilde{e}_n)^{\top} = \{\mathbf{I}_n - \mathbf{P}(\mathbf{w}^{\circ})\}\mathbf{y}.$$

Next, consider the following criterion

$$C_n''(\mathbf{w}) = \|\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\|^2 + 2\sum_{i=1}^{n} \tilde{e}_i^2 P_{ii}(\mathbf{w}). \tag{6}$$

Both (5) and (6) are quadratic functions of $\mathbf{w}$, while criterion (4) is a cubic function. Many contemporary software packages, such as `quadprog` in `R` or `MATLAB`, can effectively handle quadratic programming problems. In fact, from the real data analysis conducted in Section 4.1, it is observed that the time required to solve two quadratic programming problems is notably lower compared to that required to solve a more intricate nonlinear programming problem of higher order. Please see Table 5 for more details.

Algorithm 1: 1step-WRF$_{\text{opt}}$

Input: (1) The training data set $\mathscr{D} = \{y_i, \mathbf{x}_i\}_{i=1}^n$ (2) The number of trees in random forest M_n

Output: Weight vector $\widehat{\mathbf{w}} \in \mathscr{H}$

for *m = 1 to* M_n **do**
- Draw a bootstrap data set $\mathscr{D}_{(m)}$ of size n from the training data set $\mathscr{D}$;
- Grow a random-forest tree $\hat{f}_{(m)}$ to the bootstrap data $\mathscr{D}_{(m)}$, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size `nodesize` is reached ; `// nodesize,`q `are hyper parameters`
  - i. Select q variables at random from the p variables;
  - ii. Pick the best variable/ splitting point among the q;
  - iii. Split the node into two daughter nodes.
- **for** *i = 1 to* n **do**
  - Drop $\mathbf{x}_i$ down the the m^{th} tree and get $\mathbf{P}_{\text{BL}(m)}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \mathbf{\Theta}_{(m)})$.
- **end**
- $\mathbf{P}_{\text{BL}(m)} \leftarrow \{\mathbf{P}_{\text{BL}(m)}(\mathbf{x}_1, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \mathbf{\Theta}_{(m)}), \ldots, \mathbf{P}_{\text{BL}(m)}(\mathbf{x}_n, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \mathbf{\Theta}_{(m)})\}^\top$.

end

Solve the convex optimization problem:

$$\widehat{\mathbf{w}} = (\widehat{w}_{(1)}, \ldots, \widehat{w}_{(M_n)})^\top \leftarrow \underset{\mathbf{w} \in \mathscr{H}}{\arg\min}\ C'_n(\mathbf{w}).$$

We refer to the RF with tree-level weights derived from optimizing criterion (4) as 1 Step Optimal Weighted RF (1step-WRF$_{\text{opt}}$), and the RF with weights of trees obtained by optimizing criterion (6) as 2 Steps Optimal Weighted RF (2steps-WRF$_{\text{opt}}$). Their details are given in Algorithms 1 and B.1, respectively. For the sake of simplicity, we provide the complete description of Algorithm B.1 in Appendix B.1, since it shares a large part in common with Algorithm 1 except for the weight choice criteria. In the following, we use the WRF$_{\text{opt}}$ to refer to the algorithms including both the 1step-WRF$_{\text{opt}}$ and 2steps-WRF$_{\text{opt}}$.

3.2 Asymptotic Optimality

Denote the selected weight vectors from $C'_n(\mathbf{w})$ and $C''_n(\mathbf{w})$ by

$$\widehat{\mathbf{w}} = \underset{\mathbf{w} \in \mathscr{H}}{\arg\min} C'_n(\mathbf{w}) \text{ and } \widetilde{\mathbf{w}} = \underset{\mathbf{w} \in \mathscr{H}}{\arg\min} C''_n(\mathbf{w}),$$

respectively. In this section, we aim to analyze the loss and risk behavior associated with $\widehat{\mathbf{w}}$ and $\widetilde{\mathbf{w}}$. Before providing the theoretical support for the proposed algorithms, as an intermediate step, we first introduce a special tree-type algorithm which splits nodes without the guidance of response values in learning samples. Namely, the "hat matrix" is independent

	Tree-Based Candidate Models	BL-of-RF-Based Candidate Models	Dependence of $\mathbf{P}$ on $\mathbf{y}$
Least squares model averaging (Hansen, 2007)	✗	✗	✗
↓			
Mallows-type averaging (Qiu et al., 2020)	✓	✗	✗
↓			
WRF_{opt} with SUT	✓	✓	✗
↓			
WRF_{opt} with CART	✓	✓	✓

Table 1: Flowchart of Theoretical Analysis (Inside the dotted box are our works.)

of the output values of learning samples. In this context, the vector $\mathbf{P}_{\text{BL}}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}, \boldsymbol{\Theta})$ reduces to $\mathbf{P}_{\text{BL}}(\mathbf{x}_i, \mathbf{X}, \mathscr{B}, \boldsymbol{\Theta})$. This setup has also been imposed in the theoretical analysis of Geurts et al. (2006) and Biau (2012). Besides, this theoretical framework is referred to as "honesty" in the field of causal inference (Athey and Imbens, 2016). We term this methodology as Split-Unsupervised Tree (SUT) in contrast to CART whose splitting criterion relies on response information. Adopting the SUT simplifies theoretical analysis, aligning it with the framework of Mallows model averaging estimator (Hansen, 2007) in linear regression, where the "hat matrix" $\mathbf{P}$ is independent of response values. This approach is also similar to the estimators explored by Qiu et al. (2020), making it a useful intermediary for further analysis on the asymptotic optimality of our proposed weighted RFs. In the remaining part of this section, we first discuss the asymptotic optimality of the WRF_{opt} with SUT. Subsequently, we present the asymptotic optimality of the WRF_{opt} with CART, which is the most important conclusion in this paper. This is achieved by theoretically linking the large sample behavior of WRF_{opt} with CART to the behavior of WRF_{opt} with SUT. The corresponding theoretical analysis process is outlined in Table 1. Appendix A provides the details of the CART algorithm and an example of SUT algorithm (used in Section 4.1).

3.2.1 Asymptotic Optimality with SUT

In this section, we will establish the asymptotic optimality of the 1step-WRF_{opt} estimator and 2steps-WRF_{opt} estimator with SUT trees. To differentiate notations associated with the SUT methodology from those used in CART, we employ the script $\star$ on RF-related notations when referring to their SUT counterparts. This indicates that the notations pertain to the SUT methodology, while maintaining their fundamental meanings with the exception of the splitting criterion. For example, for $m = 1, \ldots, M_n$, $\mathbf{P}_{\star\text{BL}(m)}$ and $\mathbf{P}_{\text{BL}(m)}$ share the same fundamental meaning, with the former associated with SUT trees and the latter related to CART trees. Likewise, $\mathbf{P}_\star(\mathbf{w}) = \sum_{m=1}^{M_n} w_{(m)} \mathbf{P}_{\star\text{BL}(m)}$ contrasts with $\mathbf{P}(\mathbf{w})$ for CART trees. Let $\xi_n = \inf_{\mathbf{w}\in\mathscr{H}} R_n(\mathbf{w})$, $n_{(m),l}$ be the number of observations in the l^{th} leaf of the m^{th}

tree, and $\text{ℵ} = \min_{1\leqslant m\leqslant M_n} \min_{1\leqslant l\leqslant \ell_{(m)}} n_{(m),l}$ be the smallest sample size (controlled by the hyper parameter `nodesize` in `R` package `randomForest`) across all leaves in all trees within the CART trees. We employ $\xi_{\star n}$ and $\text{ℵ}_\star$ to denote the counterparts of ξ_n and ℵ pertaining to SUT trees, respectively. For brevity, we will not enumerate each SUT-corresponding notation individually. We list and discuss some technical conditions as follows.

Condition 1 *$\xi_{\star n}^{-1} M_n^2 = o(1)$ almost surely, and $\mathbb{E}\left(\xi_{\star n}^{-1} M_n^2\right)$ exists for all fixed $n \geqslant 1$.*

Condition 2 *There exists a positive constant v such that $\mathbb{E}\left(e_i^4 \mid \mathbf{x}_i^0\right) \leqslant v < \infty$ almost surely for $i = 1, \ldots, n$.*

Additionally, we define $h_{(m),i}$ as the number of occurrences of instance $(y_i, \mathbf{x}_i)$ in the m^{th} bootstrap sample. Following Chi et al. (2022), when the summation is over an empty set, we define its value as zero; also, we define $0/0 = 0$. Thus, without loss of generality, we assume that $h_{(m),i} > 0$ for all $m = 1, \ldots, M_n$ and $i = 1, \ldots, n$ in the remaining part of this study. Denote $\text{ℏ}_{\max} = \max_{1\leqslant m\leqslant M_n, 1\leqslant i\leqslant n} h_{(m),i}$.

Condition 3 *$\text{ℏ}_{\max} \text{ℵ}_\star^{-1} n^{1/2} = O(1)$ almost surely.*

Condition 4 *For each fixed $i, j \in \{1, \ldots, n\}$, there exists a positive constant c such that $c h_{(m),j} \leqslant h_{(m),i}$ each $m = 1, \ldots, M_n$, almost surely.*

Condition 5 *$\xi_{\star n}^{-1} M_n n^{1/2} = o(1)$ almost surely, and $\mathbb{E}\left(\xi_{\star n}^{-1} M_n n^{1/2}\right)$ exists for all fixed $n \geqslant 1$.*

Conditions 1 and 5 restrict the increasing rates of the number of trees M_n and regulate the behavior of the minimum averaging risk $\xi_{\star n}$. Similar conditions have been considered and discussed by Zhang et al. (2020), Zhang (2021), Zou et al. (2022), and others. One typical scenario that guarantees these two conditions occurs when all trees are misspecified, which precludes any trees within the RF yield perfect predictions and dominate others. The misspecification scenario is particularly common under high-dimensional data contexts, where important predictors might be omitted from tree construction due to the randomization process inherent in feature selection at each split. Besides, inspired by Chi et al. (2022), we first rigorously define the notion of "important predictors" in the context of RF under a simplified scenario in Appendix C.7. Then, under this situation, it is reasonable to expect that the key identity $\xi_{\star n}^{-1}$ converges to 0 at the rate $O(n^{-1})$. In such cases, the rate of convergence in Conditions 1 and 5 can both be further simplified to $M_n^2 n^{-1} = o(1)$. Moreover, even when all the important predictors are incorporated into the tree-building process, in a very simplified situation where a non-adaptive tree is considered, Lin and Jeon (2006) show that the mean square error of the limiting value of the RF estimator is bounded below by $C/\{\mathcal{N}^\star \log^{p-1}(n)\}$, where C is a positive constant and $\mathcal{N}^\star$ represents the maximal number of samples across all leaves and all trees. This indicates that $\xi_{\star n}^{-1} = O[1/\{\mathcal{N}^\star \log^{p-1}(n)\}]$. In this case, Conditions 1 and 5 can be further reduced to $M_n^2/\{\mathcal{N}^\star \log^{p-1}(n)\} = o(1)$ and $M_n n^{1/2}/\{\mathcal{N}^\star \log^{p-1}(n)\} = o(1)$, respectively. However, given the complexity of RF, achieving an explicit rate of convergence for $\xi_{\star n}^{-1}$ under general RF scenarios remains an open question, which we identify as a direction for future research.

Condition 2 imposes the boundedness of the conditional moments, which is a mild condition and can be found in much literature, including works by Hansen and Racine (2012), and Hansen (2007). Condition 3 is a high-level assumption that restricts the structure of the RF and its constituent trees, which is equivalent to Condition C.9 of Qiu et al. (2020) in the context therein. In fact, Condition 3 requires that the minimum sample size across all leaves and all trees grows with order no slower than $n^{1/2}$. Besides, it yields that $\text{trace}(\mathbf{P}_{\star \text{BL}(m)}) = O(n^{1/2})$, almost surely. This poses restrictions on the degrees-of-freedom or complexity of trees. In other words, trees should not be fully developed. Actually, as noted by Probst et al. (2019), increasing hyper parameter `nodesize`, leading to less tree leaves, can decreases the computation time approximately exponentially. Our practical experience, particularly with large data sets, suggests that setting this hyper parameter to a value higher than the default can significantly reduce runtime, often without markedly compromising prediction performance. Specifically, Segal (2004) show an example where increasing the number of noise variables results in a higher optimal `nodesize`. Therefore, it is advisable to construct trees in a RF that are moderately developed, balancing between being too shallow, which may result in underfitting (Hastie et al., 2009; James et al., 2013), and being excessively deep, which can impose a significant computational burden. Thus, Condition 3 is reasonable and easy to be satisfied in practice. Condition 4 excludes the unbalanced sampling in bootstrap sampling process, specifically ruling out the extreme case where certain samples disproportionately dominate the bootstrap samples.

The following theorems establish the asymptotic optimality of the 1step-WRF$_{\text{opt}}$ estimator and 2steps-WRF$_{\text{opt}}$ estimator, respectively.

Theorem 1 (Asymptotic Optimality for 1step-WRF$_{\text{opt}}$ with SUT) *Assume Conditions 1 - 4 hold. Then, as $n \to \infty$,*

$$\frac{L_n^\star(\widehat{\mathbf{w}}^\star)}{\inf_{\mathbf{w}\in\mathscr{H}} L_n^\star(\mathbf{w})} \xrightarrow{p} 1. \tag{7}$$

If, in addition, there exists an integrable random variable η such that

$$\left|\{L_n^\star(\widehat{\mathbf{w}}^\star) - \xi_{\star n}\}\,\xi_{\star n}^{-1}\right| \leqslant \eta,$$

then

$$\frac{R_n^\star(\widehat{\mathbf{w}}^\star)}{\inf_{\mathbf{w}\in\mathscr{H}} R_n^\star(\mathbf{w})} \xrightarrow{p} 1. \tag{8}$$

Theorem 2 (Asymptotic Optimality for 2steps-WRF$_{\text{opt}}$ with SUT) *Assume Conditions 1 - 5 hold. Then, as $n \to \infty$,*

$$\frac{L_n^\star(\widetilde{\mathbf{w}}^\star)}{\inf_{\mathbf{w}\in\mathscr{H}} L_n^\star(\mathbf{w})} \xrightarrow{p} 1. \tag{9}$$

If, in addition, there exists an integrable random variable η such that

$$\left|\{L_n^\star(\widetilde{\mathbf{w}}^\star) - \xi_{\star n}\}\,\xi_{\star n}^{-1}\right| \leqslant \eta,$$

then

$$\frac{R_n^\star(\widetilde{\mathbf{w}}^\star)}{\inf_{\mathbf{w}\in\mathscr{H}} R_n^\star(\mathbf{w})} \xrightarrow{p} 1. \tag{10}$$

Results obtained in (7) and (9) regard the asymptotic optimality in the sense of achieving the lowest possible squared loss, while (8) and (10) yield asymptotic optimality in the sense of achieving the lowest possible squared risk. Proofs of Theorems 1 and 2 are presented in Appendices C.2 and C.3, respectively.

3.2.2 Asymptotic Optimality with CART

Theoretical analysis regarding CART is generally very challenging, as the splitting criterion relies on response information and the black-box nature of the procedure (Scornet et al., 2015). Nevertheless, there are some pioneering studies in the literature that employ the CART splitting criterion. For example, in the context of additive regression models, Scornet et al. (2015) study the consistency of Breiman's results (Breiman, 2001). Moreover, Klusowski (2021) establishes the universal consistency of CART in the context of high dimensional additive models. In addition, Syrgkanis and Zampetakis (2020) analyze the finite sample properties of CART with binary features, under a sparsity constraint. More recently, Chi et al. (2022) establish the consistency rates for the original CART. Now, we are equipped to establish the asymptotic optimality of WRF_{opt} using CART trees, which will be achieved by studying the difference between CART-based RF and its limiting version, based on the intermediate results established in Section 3.2.1. More specifically, following the recent work of Chi et al. (2022) and Scornet et al. (2015), we term the limiting version of CART-splitting criterion as the *theoretical* CART-splitting criterion. Unlike its practical counterpart, the theoretical CART-splitting criterion does not depend on response values and therefore can be considered as a special type of splitting criterion employed by SUT trees. Our analysis in this section focuses on the discrepancy between estimators using the CART-splitting criterion and those using the theoretical CART-splitting criterion (referred to as theoretical RF).

To facilitate the technical presentation, we first introduce additional notations for the structure of a tree and its cells. Following Chi et al. (2022), without loss of generality, we set $\mathbf{x}_i \in [0,1]^p$ and define a cell as a rectangle $\mathbf{t} = \times_{d=1}^{p} t_d$, representing the Cartesian product of the closed or half-closed interval t_d in $[0,1]$. Let $\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k,1}, \ldots, \boldsymbol{\theta}_{k,2^k}\}$ with elements $\boldsymbol{\theta}_{k,\cdot} \subset \{1, \ldots, p\}$ be the sets of available features for the 2^{k-1} splits at level $k-1$ that grow the 2^k cells (including empty cells). Given the CART-splitting criterion and a set of $\boldsymbol{\theta}_{1:k} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$, let $T(\boldsymbol{\theta}_{1:k}) = \{\mathbf{t}_{1:k,1}, \ldots, \mathbf{t}_{1:k,2^k}\}$ be the collection of all cell sequences connecting the root to the end cells at level k, which is determined by $\boldsymbol{\Theta}$. Each sequence $\mathbf{t}_{1:k,s}$ for $s = 1, \ldots, 2^k$ can be considered as a "tree branch", representing a partition of $[0,1]^p$. One can refer to Figure 1 of Chi et al. (2022) for a graphical illustration for this splitting scheme. It is natural to consider varying tree heights within a RF, implying that the maximum level k depends on m for each $m = 1, \ldots, M_n$. Therefore, we use $k_{(m)}$ to denote the maximum level of the m^{th} tree.

For a new instance $\mathbf{x}$, which is an independent copy of $\mathbf{x}_1$, the m^{th} tree estimator given $T(\boldsymbol{\theta}_{1:k_{(m)}})$ and $\{h_{(m),1}, \ldots, h_{(m),n}\}$ can be expressed as follows

$$\widehat{\mu}_{T(\boldsymbol{\theta}_{1:k_{(m)}})}(\mathbf{x}) \quad = \quad \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x} \in \mathbf{t}_{k_{(m)},s}\right) \frac{\sum_{i=1}^{n} h_{(m),i} y_i \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{i=1}^{n} h_{(m),i} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right)}$$

$$= \sum_{i=1}^{n} y_i \sum_{s=1}^{2^{k_{(m)}}} \frac{h_{(m),i}\mathbb{I}\left(\mathbf{x} \in \mathbf{t}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{i'=1}^{n} h_{(m),i'}\mathbb{I}\left(\mathbf{x}_{i'} \in \mathbf{t}_{k_{(m)},s}\right)}, \tag{11}$$

where $\mathbf{t}_{k_{(m)},s}$ is the end cell of the tree branch $\mathbf{t}_{1:k_{(m)},s}$ (also referred to leaves if it is not empty), and $\mathbb{I}(\cdot)$ denotes the indicator function. Now, by definition, the $(i,j)^{\text{th}}$ element of $\mathbf{P}_{\text{BL}(m)}$ can be expressed as

$$P_{\text{BL}(m),ij} = \sum_{s=1}^{2^{k_{(m)}}} \frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)}, \tag{12}$$

for all $m = 1, \ldots, M_n$. Consistent with Chi et al. (2022), given a cell $\mathbf{t}$, the CART-splitting criterion is defined as

$$\begin{aligned}\text{CART}_{\mathbf{t}}(d,c) &= \min_{c_1 \in \mathbb{R}^1} \frac{\sum_{i=1}^{n} \mathbb{I}(x_{id} < c, \mathbf{x}_i \in \mathbf{t})\,(y_i - c_1)^2}{n} \\ &\quad + \min_{c_2 \in \mathbb{R}^1} \frac{\sum_{i=1}^{n} \mathbb{I}(x_{id} \geqslant c, \mathbf{x}_i \in \mathbf{t})\,(y_i - c_2)^2}{n},\end{aligned}$$

with its *theoretical* version

$$\begin{aligned}\text{CART}^*_{\mathbf{t}}(d,c) &= \Pr\left(x_{1d} < c \mid \mathbf{x}_1 \in \mathbf{t}\right) \text{var}\left(y_1 \mid x_{1d} < c, \mathbf{x}_1 \in \mathbf{t}\right) \\ &\quad + \Pr\left(x_{1d} \geqslant c \mid \mathbf{x}_1 \in \mathbf{t}\right) \text{var}\left(y_1 \mid x_{1d} \geqslant c, \mathbf{x}_1 \in \mathbf{t}\right),\end{aligned}$$

where d is the label of splitting variable, c is a corresponding splitting point, and x_{id} is the d^{th} entry of $\mathbf{x}_i$. Analogous to (12), we denote the corresponding theoretical version of $\mathbf{P}_{\text{BL}(m)}$ pertaining to $\text{CART}^*_{\mathbf{t}}(d,c)$ for each $m \in \{1, \ldots, M_n\}$ as

$$\mathbf{P}_{*\text{BL}(m)} = \left(P^*_{\text{BL}(m),ij}\right)_{n\times n} = \left\{ \sum_{s=1}^{2^{k_{(m)}}} \frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right\}_{n\times n}, \tag{13}$$

where $\mathbf{t}^*_{k_{(m)},s}$ for each $m \in \{1, \ldots, M_n\}$ and $s \in \{1, \ldots, 2^{k_{(m)}}\}$ is the s^{th} end cell of the m^{th} base learner based on the theoretical CART-splitting criterion $\text{CART}^*_{\mathbf{t}}(d,c)$. Further, we continue employing the script $*$ on RF-related notations when referring to their theoretical counterparts. Detailed definitions corresponding to theoretical CART-splitting criterion will not be enumerated here, in the interest of brevity.

Note that the theoretical CART-splitting criterion is independent of response values and therefore can be viewed as a special type of splitting criterion used by SUT trees. Consequently, as an immediate application, Theorems 1 - 2 in Section 3.2.1 guarantee the asymptotic optimality of the WRF_{opt} with theoretical CART-splitting criterion. In what follows, we introduce additional notations crucial for quantifying the discrepancy between the RF and theoretical RF. Let

$$\Pr\left\{\mathbb{I}\left(x_i \in \mathbf{t}_{k_{(m)},s} \Delta \mathbf{t}^*_{k_{(m)},s}\right) = 1\right\} = p_{(m),is},$$

and

$$\delta_n = \max_{1\leqslant i\leqslant n, 1\leqslant m\leqslant M_n, 1\leqslant s\leqslant 2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s} \Delta \mathbf{t}^*_{k_{(m)},s}\right),$$

where $S_1 \Delta S_2 = (S_1 \cup S_2) - (S_1 \cap S_2)$ is the symmetric difference between two generic sets S_1, S_2. Let $\bar{K}_n = \max_{1\leqslant m\leqslant M_n} 2^{k_{(m)}}$ be the largest number of end cells (including empty end cells) among all trees, and $\mathscr{N} = \max_{1\leqslant m\leqslant M_n, 1\leqslant l\leqslant \ell_{(m)}} n_{(m),l}$ be the largest sample size across all leaves in all trees within the RF,[2] with $\mathscr{N}^*$ being the theoretical counterpart of $\mathscr{N}$. Now, we are equipped to explore the discrepancy between the RF and theoretical RF. The following result bounds the gap between two pivotal matrices $\mathbf{P}_{\mathrm{BL}(m)}$ and $\mathbf{P}_{*\mathrm{BL}(m)}$ for all $m = 1, \ldots, M_n$.

Lemma 1 *Under Condition 4, there exist two positive constants c_1, c_2 such that*

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|P_{\mathrm{BL}(m),ij} - P^*_{\mathrm{BL}(m),ij}\right| \leqslant c_1 \bar{K}_n(\mathscr{N} + \mathscr{N}^*)\delta_n, \tag{14}$$

and

$$\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}\left|P_{\mathrm{BL}(m),ij} - P^*_{\mathrm{BL}(m),ij}\right| \leqslant \left\{2 + c_2\left(1 + \frac{\mathscr{N}^*}{\mathscr{n}}\right)\right\}\delta_n, \tag{15}$$

almost surely, uniformly for all $m = 1, \ldots, M_n$.

Clearly, δ_n plays a pivotal role in bounding the proximity between the matrices $\mathbf{P}_{\mathrm{BL}(m)}$ and $\mathbf{P}_{*\mathrm{BL}(m)}$. The proof of Lemma 1 will be provided in Appendix C.4. Let ξ_{*n} and $\mathscr{n}_*$ be the counterparts of ξ_n and $\mathscr{n}$ under theoretical CART, we list and discuss some technical conditions as follows.

Condition 1′ $\xi_{*n}^{-1} M_n^2 = o(1)$ almost surely, and $\mathbb{E}\left(\xi_{*n}^{-1} M_n^2\right)$ exists for all fixed $n \geqslant 1$.

Condition 2′ There exist two positive constants v_1 and v_2 such that $\mathbb{E}\left(|e_i|^r \mid \mathbf{x}_i^0\right) \leqslant v_1^2 v_2^{r-2} r!/2$ almost surely for every $i = 1, \ldots, n$ and $r \geqslant 2$.

Condition 3′ $\mathscr{h}_{\max} \mathscr{n}_*^{-1} n^{1/2} = O(1)$ almost surely.

Condition 5′ $\xi_{*n}^{-1} M_n n^{1/2} = o(1)$ almost surely, and $\mathbb{E}\left(\xi_{*n}^{-1} M_n n^{1/2}\right)$ exists for all fixed $n \geqslant 1$.

Condition 6 *As $n \to \infty$, $\mathscr{N}^*/\mathscr{n}$, $\mathscr{N} + \mathscr{N}^*$ and $p_{(m),is}$ are bounded above by deterministic series, say, $\bar{\mathscr{r}}_n$, $\bar{\mathscr{N}}_n$ and $\bar{p}_n$, respectively.*

2. Note that for all $m = 1, \ldots, M_n$, $l \in \{1, \ldots, \ell_{(m)}\}$ represents the index of leaves (non-empty end cells) in the m^{th} tree, while $s \in \{1, \ldots, 2^{k_{(m)}}\}$ denotes the index of all end cells, inclusive of those that are empty.

Condition 7 *As $n \to \infty$, $\bar{p}_n < 1/2$,*

$$\bar{\epsilon}_n = \xi_n^{-1} \bar{K}_n \bar{\mathscr{N}}_n \log^2(n) \left\{ \sqrt{\frac{2\log(2nM_n\bar{K}_n)}{\log\left(\frac{1}{\bar{p}_n} - 1\right)}} + \bar{p}_n \right\}^{1/4} = o(1),$$

almost surely, and $\mathbb{E}(\bar{\epsilon}_n)$ exists all fixed $n \geqslant 1$.

Conditions $1'$ - $3'$ and $5'$ correspond to Conditions 1 - 3 and 5, respectively, with the focus shifted to theoretical CART-splitting. Moreover, Condition $2'$ guarantees that $\mathbb{E}|e_i|^r \leqslant v_1^2 v_2^{r-2} r!/2$, which is referred to as Bernstein's moment condition (Zhang and Chen, 2021). Particularly, if $\{e_i\}_{i=1}^n$ are generated independently by Normal$(0, \sigma^2)$, it follows that $v_1^2 = 2\sigma^2$ and $v_2 = \sigma^2$ (Zhang and Chen, 2021, Example 5.4). In addition, Condition 6 imposes restrictions on the hyper parameters of RF, requiring the behavior of these parameters does not become too erratic as the amount of data increases. Condition 7 requires that the CART criterion leads to reasonably accurate splits in the sense that the difference between $\mathbf{t}_{k_{(m)},s}$ and $\mathbf{t}^*_{k_{(m)},s}$ (measured by $\mathbb{E}(\delta_n)$, which is bounded above by $\sqrt{2\log(2nM_n\bar{K}_n)/\log\left(\bar{p}_n^{-1} - 1\right)} + \bar{p}_n$) is ignorable in comparison to ξ_n.

Theorem 3 (Asymptotic Optimality under CART and criterion 5) *Assume that Conditions $1'$ - $3'$, 4, and 6 - 7 hold. Then, as $n \to \infty$,*

$$\frac{L_n(\mathbf{w}^\circ)}{\inf_{\mathbf{w}\in\mathscr{H}} L_n(\mathbf{w})} \xrightarrow{p} 1,$$

where $\mathbf{w}^\circ$ is the solution of minimizing the criterion (5) over $\mathbf{w} \in \mathscr{H}$.

The proof of Theorem 3 is provided in Appendix C.5. Theorem 3 demonstrates the asymptotic optimality of the weighted RF with CART-splitting criterion, with weights obtained by criterion (5). This criterion is derived under a "working homoskedastic" framework and serves as a groundwork for further theoretical analysis. Based on Theorem 3, similar theoretical results regarding the heteroskedastic criteria (that is, $C'_n(\mathbf{w})$ and $C''_n(\mathbf{w})$) can be established.

Corollary 1 (Asymptotic Optimality for 1step-WRF$_{\text{opt}}$ with CART) *Assume Conditions $1'$ - $3'$, 4, and 6 - 7 hold. Then, as $n \to \infty$,*

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathscr{H}} L_n(\mathbf{w})} \xrightarrow{p} 1. \tag{16}$$

If, in addition, there exists an integrable random variable η such that

$$\left|\{L_n(\widehat{\mathbf{w}}) - \xi_n\}\, \xi_n^{-1}\right| \leqslant \eta,$$

then

$$\frac{R_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathscr{H}} R_n(\mathbf{w})} \xrightarrow{p} 1. \tag{17}$$

Corollary 2 (Asymptotic Optimality for 2steps-WRF$_{\text{opt}}$ with CART) *Assume Conditions 1′ - 3′, 4, 5′ and 6 - 7 hold. Then, as $n \to \infty$,*

$$\frac{L_n\left(\widetilde{\mathbf{w}}\right)}{\inf_{\mathbf{w}\in\mathscr{H}} L_n(\mathbf{w})} \xrightarrow{p} 1. \tag{18}$$

If, in addition, there exists an integrable random variable η such that

$$\left|\{L_n(\widetilde{\mathbf{w}}) - \xi_n\}\,\xi_n^{-1}\right| \leqslant \eta,$$

then

$$\frac{R_n(\widetilde{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathscr{H}} R_n(\mathbf{w})} \xrightarrow{p} 1. \tag{19}$$

With the result established in Theorem 3, Corollaries 1 and 2 can readily be verified within the same framework that proves Theorem 2 and thus is omitted. Theorem 3 and Corollaries 1 and 2 extend the results in Theorems 1 and 2 from SUT to CART. We are unaware of any similar results in this field.

4. Numerical Study

In this section, we will conduct experiments using real and semi-synthetic data, the latter derived from the former, to evaluate the performance of different weighted RFs.

4.1 Real Data Analysis

To assess the prediction performance of different weighted RFs in practical situations, we used 11 data sets from the UCI data repository for machine learning (Dua and Graff, 2017). Because most of these data sets are low-dimensional, one additional high-dimensional data set from `openml.org` (Vanschoren et al., 2013) was also included. The details of the 12 data sets are listed in Table 2. Appendix B features a demonstration of two competitors, namely wRF and CRF.

For the sake of brevity, in the following, we will refer to each data set by its abbreviation. We randomly partitioned each data set into training data, testing data and validation data, in the ratio of 0.5 : 0.3 : 0.2. The training data was used to construct trees and to calculate weights, and the test data was used to evaluate the predictive performance of different algorithms. The validation data was employed to select tuning parameters, such as the exponent in the expression for calculating weights in the wRF, and probability sequence in the SUT algorithm.

In this section, the number of trees M_n was set to 100. Before each split, the dimension of random feature sub-space q was set to $\lceil p/3 \rceil$, which is the default value in the regression mode of the `R` package `randomForest`. We set the minimum leaf size `nodesize` to $\lceil \sqrt{n} \rceil$ in CART trees and 5 in SUT trees, in order to control the depth of trees. We also tried other values of M_n and `nodesize`, and the patterns of the performance remain stable in general. Figures D.12 - D.23 in Appendix D will provide more information on the robustness of the proposed methods over different RF hyper parameters.[3]

3. In our robustness tests for M_n, we varied M_n across 100, 200, 400 and 800, while keeping `nodesize` fixed at $\lceil \sqrt{n} \rceil$. For `nodesize` robustness tests, `nodesize` was set to the quintiles within the range of $[5, \lceil \sqrt{n} \rceil]$, with M_n fixed at 100.

Data set	Abbreviation	Attributes	Samples
Boston Housing	BH	12	506
Servo	Servo	4	167
Concrete Compressive Strength	CCS	9	1030
Airfoil Self-Noise	ASN	5	1503
Combined Cycle Power Plant	CCPP	4	9568
Concrete Slump Test	CST	7	103
Energy Efficiency	EE	8	768
Parkinsons Telemonitoring	PT	20	5875
QSAR aquatic toxicity	QSAR	8	546
Synchronous Machine	SM	4	557
Yacht Hydrodynamics	YH	6	308
Tecator	Tecator	124	240

Table 2: Summary of 12 Data Sets

For each strategy, the number of replication was set to $D = 1000$ and the forecasting performance was accessed by the following two criteria:

$$\text{MSFE} = \frac{1}{D \times n_{\text{test}}} \sum_{d=1}^{D} \sum_{i=1}^{n_{\text{test}}} (y_i - \widehat{y}_{i,d})^2 \text{ and } \text{MAFE } = \frac{1}{D \times n_{\text{test}}} \sum_{d=1}^{D} \sum_{i=1}^{n_{\text{test}}} |y_i - \widehat{y}_{i,d}|,$$

where n_{test} is the size of testing data, and $\widehat{y}_{i,d}$ is the forecast for y_i in the d^{th} repetition. MSFE and MAFE are abbreviations of "Mean Squared Forecast Error" and "Mean Absolute Forecast Error", respectively.

As noted in Section 1, an averaging strategy with appropriately selected unequal weights may outperform simple averaging if individual learners display non-identical strength. To ascertain the relationship between the performance of the WRF_{opt} and the diversity level of base learners, we employ the following weighted correlation between the residuals $\mathbf{r}_{(m)} = \mathbf{y} - \widehat{\mathbf{y}}^{(m)}$ and $\mathbf{r}_{(m')} = \mathbf{y} - \widehat{\mathbf{y}}^{(m')}$ where $1 \leqslant m, m' \leqslant M_n$ and $m \neq m'$, as proposed by Breiman (2001),

$$\bar{\rho} = \left\{ \frac{2}{M_n(M_n-1)} \sum_{1 \leqslant m < m' \leqslant M_n} \rho(\mathbf{r}_{(m)}, \mathbf{r}_{(m')}) \, \text{sd}(\mathbf{r}_{(m)}) \, \text{sd}(\mathbf{r}_{(m')}) \right\} \Big/ \left\{ \frac{1}{M_n} \sum_{m=1}^{M_n} \text{sd}(\mathbf{r}_{(m)}) \right\}^2,$$

where $\rho(\mathbf{r}_{(m)}, \mathbf{r}_{(m')})$ is the correlation between $\mathbf{r}_{(m)}$ and $\mathbf{r}_{(m')}$, and $\text{sd}(\mathbf{r}_{(m)})$ is the standard deviation of $\mathbf{r}_{(m)}$. It is clear that a larger $\bar{\rho}$ signifies reduced diversity among base learners. In this scenario, it is expected that equal weights may also yield reasonable performance (Zhou, 2012). In fact, our numerical experiments below echo this speculation, showing that the WRF_{opt} strategies significantly outperform conventional RFs when $\bar{\rho}$ is relatively small, for instance, $\bar{\rho} < 0.5$. Consequently, with a relatively small $\bar{\rho}$, one can expect more pronounced improvements in predictive performance by adopting appropriate unequal weights. Next, we will provide the results of different weighting techniques on RFs with CART trees and RFs with SUT trees, respectively.

Data set	RF	2steps-WRF$_{\text{opt}}$	1step-WRF$_{\text{opt}}$	wRF	CRF
BH	$15.484^{(5)}$	$13.958^{(1)}$	$14.038^{(2)}$	$14.517^{(3)}$	$14.664^{(4)}$
Servo	$1.610^{(5)}$	$0.825^{(1)}$	$0.836^{(2)}$	$0.860^{(3)}$	$1.169^{(4)}$
CCS	$60.460^{(5)}$	$50.004^{(1)}$	$50.048^{(2)}$	$52.868^{(3)}$	$54.065^{(4)}$
ASN	$20.022^{(5)}$	$14.572^{(1)}$	$14.575^{(2)}$	$15.611^{(3)}$	$17.054^{(4)}$
CCPP	$18.016^{(5)}$	$16.065^{(2)}$	$16.062^{(1)}$	$16.237^{(3)}$	$16.556^{(4)}$
CST	$26.421^{(5)}$	$19.877^{(1)}$	$20.074^{(2)}$	$21.500^{(3)}$	$22.534^{(4)}$
EE	$4.332^{(5)}$	$3.643^{(2)}$	$3.642^{(1)}$	$3.964^{(3)}$	$4.087^{(4)}$
PT	$14.641^{(5)}$	$8.653^{(2)}$	$8.649^{(1)}$	$9.099^{(3)}$	$10.819^{(4)}$
QSAR	$1.436^{(5)}$	$1.423^{(3)}$	$1.434^{(4)}$	$1.417^{(1)}$	$1.420^{(2)}$
SM($\times 10^{-4}$)	$6.981^{(5)}$	$3.403^{(1)}$	$3.404^{(2)}$	$4.342^{(3)}$	$5.214^{(4)}$
YH	$35.442^{(5)}$	$3.727^{(1)}$	$3.735^{(2)}$	$5.603^{(3)}$	$13.422^{(4)}$
Tecator	$5.274^{(5)}$	$3.295^{(2)}$	$3.282^{(1)}$	$3.458^{(3)}$	$3.879^{(4)}$

Table 3: Test Error Comparisons by MSFE for Different Forests with CART Trees

4.1.1 RFs with CART Trees

Tables 3 and 4 exhibit the risks of RFs with CART trees in terms of MSFE and MAFE, respectively. Each row presents the results for a specific data set, comparing the risks associated with different RF algorithms across columns. Values in parentheses in the upper right corner indicate the risk ranking for each RF algorithm within the same data set, with a lower rank denoting a lower risk.

Regarding MSFE, the 1step-WRF$_{\text{opt}}$ or 2steps-WRF$_{\text{opt}}$ estimator manifests the best performance in 11 out of 12 data sets, whereas exhibits the best performance in 10 out of 12 data sets in terms of MAFE. It is observed that the wRF becomes the best method in some data sets. Of all cases considered, the CRF is found to never be the best method. It is also noticeable that the 2steps-WRF$_{\text{opt}}$ is superior to the 1step-WRF$_{\text{opt}}$ in most cases, albeit with minor differences.

Table 5 compares the time consumption of the 2steps-WRF$_{\text{opt}}$ and 1step-WRF$_{\text{opt}}$ algorithms for a single run, averaged over D repetitions, with the ratio of the latter to the former in the fourth column. Apparently, the 2steps-WRF$_{\text{opt}}$ can accelerate optimization by tens or hundreds of times when compared to the 1step-WRF$_{\text{opt}}$, given that solving quadratic optimization is considerably faster than solving a higher-order nonlinear optimization task.

To further assess the performance of the 2steps-WRF$_{\text{opt}}$ over other competing methods, we evaluated their relative risks with respect to the 2steps-WRF$_{\text{opt}}$. Specifically, we calculated the relative risks of the RF, 1step-WRFopt, wRF, and CRF by dividing their respective risks by that of the benchmark 2steps-WRF$_{\text{opt}}$. In the following, we assert that a relative risk is not essential if it falls in the interval of $(0.95, 1.05)$, while it is essential if it is lower than 0.95 or higher than 1.05. The relative MSFE and MAFE of each method on 12 data sets are reported in Figures 1 and 2, respectively. The results are depicted by blue, green, purple and red bars, respectively, for the RF, 1step-WRF$_{\text{opt}}$, wRF, and CRF. Furthermore, to illustrate the relationship between the performance of the WRF$_{\text{opt}}$ and the diversity level of base learners, we displayed averaged $\bar{\rho}$ over D replications below the names of the data sets, arranged in ascending order from the smallest to the largest.

Data set	RF	2steps-WRF$_{\text{opt}}$	1step-WRF$_{\text{opt}}$	wRF	CRF
BH	$2.608^{(5)}$	$2.536^{(1)}$	$2.549^{(3)}$	$2.539^{(2)}$	$2.562^{(4)}$
Servo	$0.900^{(5)}$	$0.550^{(2)}$	$0.550^{(3)}$	$0.535^{(1)}$	$0.754^{(4)}$
CCS	$6.092^{(5)}$	$5.500^{(1)}$	$5.503^{(2)}$	$5.668^{(3)}$	$5.751^{(4)}$
ASN	$3.607^{(5)}$	$3.013^{(1)}$	$3.013^{(2)}$	$3.121^{(3)}$	$3.295^{(4)}$
CCPP	$3.243^{(5)}$	$3.058^{(2)}$	$3.058^{(1)}$	$3.075^{(3)}$	$3.108^{(4)}$
CST	$4.023^{(5)}$	$3.425^{(1)}$	$3.445^{(2)}$	$3.568^{(3)}$	$3.676^{(4)}$
EE	$1.563^{(5)}$	$1.349^{(2)}$	$1.349^{(1)}$	$1.423^{(3)}$	$1.487^{(4)}$
PT	$2.933^{(5)}$	$2.201^{(2)}$	$2.201^{(1)}$	$2.241^{(3)}$	$2.493^{(4)}$
QSAR	$0.892^{(4)}$	$0.888^{(3)}$	$0.892^{(5)}$	$0.885^{(1)}$	$0.887^{(2)}$
SM($\times 10^{-2}$)	$2.044^{(5)}$	$1.374^{(1)}$	$1.375^{(2)}$	$1.551^{(3)}$	$1.746^{(4)}$
YH	$3.877^{(5)}$	$1.182^{(1)}$	$1.182^{(2)}$	$1.358^{(3)}$	$2.329^{(4)}$
Tecator	$1.637^{(5)}$	$1.319^{(2)}$	$1.318^{(1)}$	$1.362^{(3)}$	$1.435^{(4)}$

Table 4: Test Error Comparisons by MAFE for Different Forests with CART Trees

Data set	2steps-WRF$_{\text{opt}}$	1step-WRF$_{\text{opt}}$	Ratio
BH	0.065	3.898	60.371
Servo	0.072	1.347	18.778
CCS	0.081	6.822	83.982
ASN	0.094	6.468	68.840
CCPP	0.566	2.785	4.916
CST	0.075	2.061	27.630
EE	0.072	3.498	48.304
PT	0.291	40.445	138.967
QSAR	0.069	3.148	45.431
SM	0.060	1.409	23.327
YH	0.065	2.631	40.546
Tecator	0.041	2.710	66.098

Table 5: Time Consumption Comparisons (Unit: seconds)

Some findings are worth mentioning in Figure 1. First, the improvement of the WRF$_{\text{opt}}$ (including the 1step-WRF$_{\text{opt}}$ and 2steps-WRF$_{\text{opt}}$) over the conventional RF is essential in 11 out of 12 data sets. What stands out in the figure is that the relative MSFEs of others with respect to the benchmark are conspicuously large in the YH data set. This spells the great success of our WRF$_{\text{opt}}$ methods in practice. More importantly, the WRF$_{\text{opt}}$ methods outperform competitors essentially in 8 out of 12 data sets, while none of the competitors dominate the benchmark essentially in all cases, underscoring the robustness of the WRF$_{\text{opt}}$.

Figure 2 remains the similar qualitative results, albeit with less notable power of the WRF$_{\text{opt}}$ than Figure 1. Specifically, the WRF$_{\text{opt}}$ shows essential improvement over the conventional RF in 10 out of 12 data sets, and dominates all competitors essentially in 3 out of 12 data sets. These proportions are relatively lower than those in Figure 1. But none of the competitors surpass the benchmark essentially in all cases, which is consistent with Figure 1.

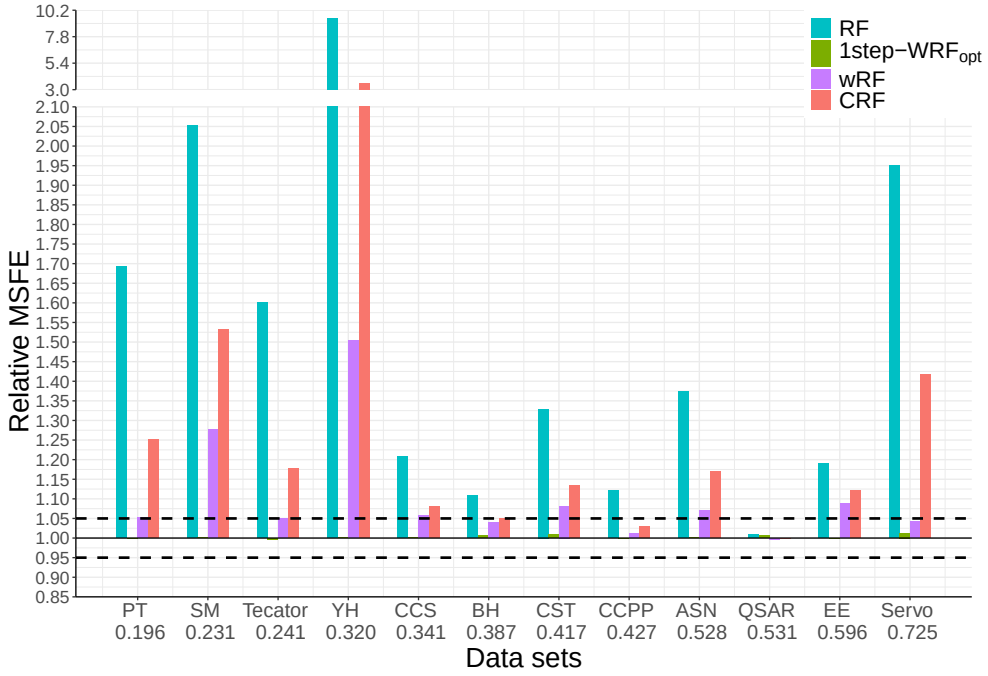

Figure 1: Relative MSFE for Different Forests with CART Trees (The horizontal axis shows the names of 12 data sets, arranged in ascending order of their corresponding averaged $\bar{\rho}$ displayed beneath each data set's abbreviation. The green bars are barely discernible, with a relative risk close to 1 due to their negligible predictive error difference compared to the 2steps-WRF$_{\text{opt}}$. This pattern consistent across subsequent Figures 2 - 4 and 7 - 8.)

Note that the wRF algorithm requires tuning a parameter outside of the training set, whereas the WRF$_{\text{opt}}$ and CRF do not. For the fairness of the comparison, all three weighted RFs should use identical tree models built in the same training data set. Were WRF$_{\text{opt}}$ and CRF not compared with wRF, they can employ more training samples, potentially leading to superior predictive performance than currently observed.

Combining all the findings together, we can conclude that the proposed WRF$_{\text{opt}}$ methods yield more accurate predictions compared to the conventional RF and other existing weighted RFs in most cases. Additionally, it is clear from Figures 1 and 2 that in the first four data sets, which have relatively low averaged $\bar{\rho}$, WRF$_{\text{opt}}$ methods tend to yield more pronounced performance over their competitors, with the exception of the Servo data set. In other words, these first four data sets grow trees with greater diversity, suggesting a stronger preference for unequal weights over equal weights, thus yielding better performance of the WRF$_{\text{opt}}$. This observation aligns with our anticipation, and exactly the motivation for adopting unequal weights, as discussed in Section 1. The Servo data set stands out as an exception, possessing the highest $\bar{\rho}$ value among the 12 data sets yet exhibiting relatively pronounced predictive

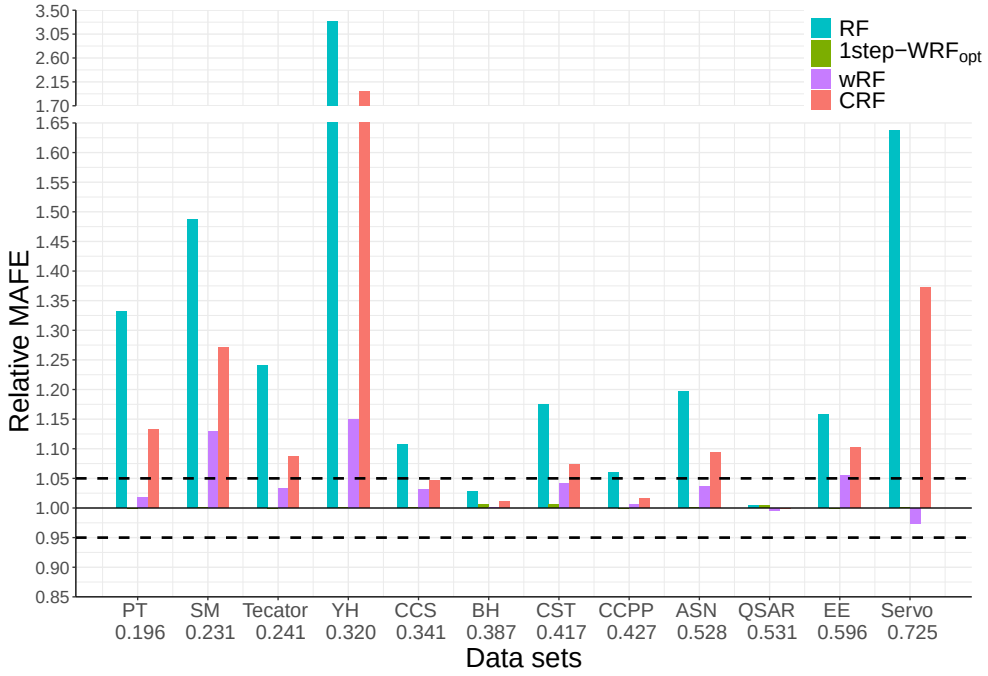

Figure 2: Relative MAFE for Different Forests with CART Trees (The horizontal axis shows the names of 12 data sets, arranged in ascending order of their corresponding averaged $\bar{\rho}$ displayed beneath each data set's abbreviation.)

performance. This anomaly is likely due to its small number of samples and attributes ($n = 167$ and $p = 4$). Similar phenomenon has also been observed in the literature, as seen in Hansen (2007), Zhang et al. (2013) and Zhang et al. (2016).

4.1.2 RFs with SUT Trees

For comparison purpose, we also study the performance of our proposed methods based on SUT trees, reporting results in Tables 6 and 7. The 1step-WRF$_{\text{opt}}$ or 2steps-WRF$_{\text{opt}}$ estimator consistently outperforms the conventional RFs and the two competitors in terms of MSFE, while performing best in 11 out of 12 data sets in terms of MAFE. Additionally, the gaps between the 1step-WRF$_{\text{opt}}$ and 2steps-WRF$_{\text{opt}}$ are relatively small, akin to RFs with CART trees.

The relative MSFE and MAFE are depicted in Figures 3 and 4, respectively. With SUT trees rather than CART trees, the WRF$_{\text{opt}}$ methods perform better at upgrading equal-weight forests. Concerning the MSFE and MAFE, the number of supporting data sets remains 11 and 10 out of 12 data sets, respectively. Additionally, the proportion of outperforming rivals climbs to 10 out of 12 data sets in terms of MSFE and 8 out of 12 data sets in terms of MAFE. Notably, the advantages in the SM, YH, and EE data sets are

Data set	RF	2steps-WRF$_{\text{opt}}$	1step-WRF$_{\text{opt}}$	wRF	CRF
BH	$38.213^{(5)}$	$24.516^{(1)}$	$24.522^{(2)}$	$28.797^{(3)}$	$29.974^{(4)}$
Servo	$1.604^{(5)}$	$0.964^{(1)}$	$0.968^{(2)}$	$0.998^{(3)}$	$1.232^{(4)}$
CCS	$149.276^{(5)}$	$119.471^{(1)}$	$119.505^{(2)}$	$136.243^{(4)}$	$132.435^{(3)}$
ASN	$36.391^{(5)}$	$33.465^{(1)}$	$33.472^{(2)}$	$35.675^{(4)}$	$35.419^{(3)}$
CCPP	$50.329^{(5)}$	$36.619^{(2)}$	$36.613^{(1)}$	$39.145^{(4)}$	$38.600^{(3)}$
CST	$42.899^{(5)}$	$25.933^{(2)}$	$25.928^{(1)}$	$32.806^{(3)}$	$36.783^{(4)}$
EE	$17.768^{(5)}$	$5.140^{(2)}$	$5.140^{(1)}$	$6.193^{(3)}$	$8.026^{(4)}$
PT	$98.864^{(5)}$	$89.299^{(1)}$	$89.325^{(2)}$	$97.924^{(4)}$	$95.321^{(3)}$
QSAR	$1.737^{(5)}$	$1.657^{(1)}$	$1.661^{(2)}$	$1.680^{(4)}$	$1.668^{(3)}$
SM($\times 10^{-4}$)	$20.963^{(5)}$	$0.212^{(1)}$	$0.212^{(2)}$	$0.251^{(3)}$	$4.460^{(4)}$
YH	$33.241^{(5)}$	$2.433^{(2)}$	$2.431^{(1)}$	$3.171^{(3)}$	$8.152^{(4)}$
Tecator	$143.584^{(5)}$	$101.741^{(1)}$	$101.940^{(2)}$	$126.409^{(3)}$	$128.264^{(4)}$

Table 6: Test Error Comparisons by MSFE for Different Forests with SUT Trees

Data set	RF	2steps-WRF$_{\text{opt}}$	1step-WRF$_{\text{opt}}$	wRF	CRF
BH	$3.759^{(5)}$	$3.128^{(1)}$	$3.131^{(2)}$	$3.314^{(3)}$	$3.349^{(4)}$
Servo	$0.847^{(5)}$	$0.578^{(1)}$	$0.578^{(2)}$	$0.606^{(3)}$	$0.701^{(4)}$
CCS	$9.914^{(5)}$	$8.764^{(1)}$	$8.764^{(2)}$	$9.421^{(4)}$	$9.273^{(3)}$
ASN	$4.903^{(5)}$	$4.689^{(2)}$	$4.685^{(1)}$	$4.813^{(4)}$	$4.765^{(3)}$
CCPP	$5.805^{(5)}$	$4.860^{(2)}$	$4.858^{(1)}$	$5.019^{(4)}$	$4.999^{(3)}$
CST	$5.209^{(5)}$	$3.927^{(2)}$	$3.926^{(1)}$	$4.455^{(3)}$	$4.773^{(4)}$
EE	$3.356^{(5)}$	$1.611^{(2)}$	$1.609^{(1)}$	$1.799^{(3)}$	$2.094^{(4)}$
PT	$7.995^{(5)}$	$7.553^{(2)}$	$7.544^{(1)}$	$7.948^{(4)}$	$7.815^{(3)}$
QSAR	$1.001^{(5)}$	$0.979^{(3)}$	$0.980^{(4)}$	$0.977^{(2)}$	$0.972^{(1)}$
SM($\times 10^{-2}$)	$3.674^{(5)}$	$0.290^{(1)}$	$0.291^{(2)}$	$0.304^{(3)}$	$1.628^{(4)}$
YH	$3.508^{(5)}$	$0.856^{(1)}$	$0.857^{(2)}$	$0.902^{(3)}$	$1.483^{(4)}$
Tecator	$9.624^{(5)}$	$7.787^{(1)}$	$7.805^{(2)}$	$8.886^{(3)}$	$8.993^{(4)}$

Table 7: Test Error Comparisons by MAFE for Different Forests with SUT Trees

particularly substantial. Besides, the relationship between the performance of the WRF$_{\text{opt}}$ and the diversity level of SUT trees follows a similar pattern to that observed in Figures 1 and 2.

Without response data for guiding splits, the WRF$_{\text{opt}}$ methods with SUT trees yield worse predictive performance than their counterparts with CART trees. However, it is worthwhile noting that the improvement of the WRF$_{\text{opt}}$ methods over RFs employing equal weights become more significant, demonstrating the potential advantage of our WRF$_{\text{opt}}$ with weaker base learners.

4.2 Semi-Synthetic Experiments

To further investigate the data set characteristics that affect the effectiveness of the WRF$_{\text{opt}}$, we conducted semi-synthetic experiments using the real data sets in Section 4.1. Clearly,

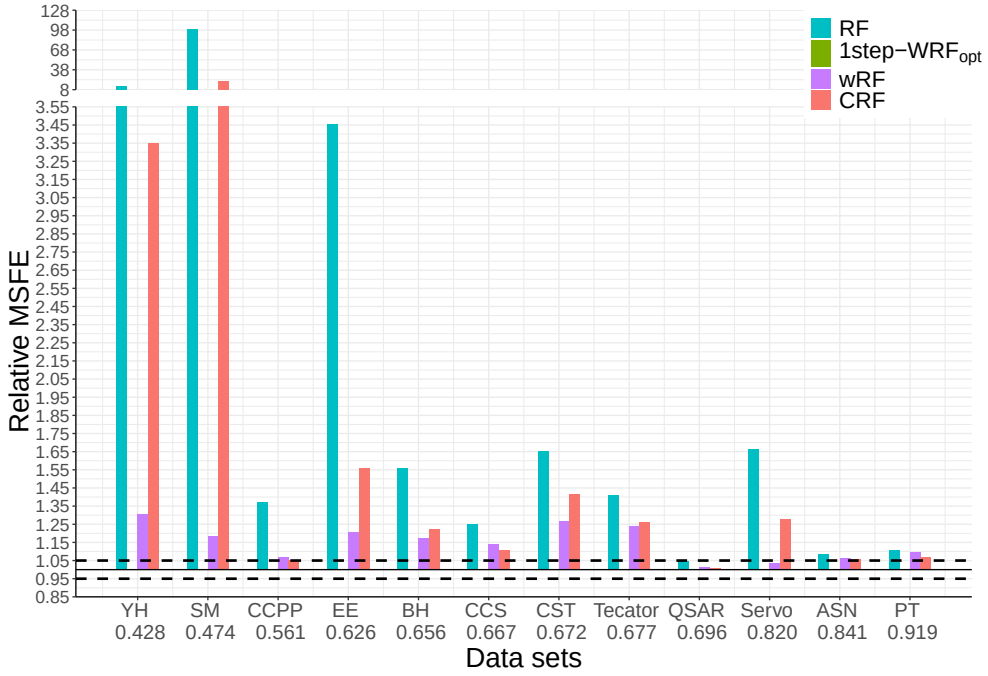

Figure 3: Relative MSFE for Different Forests with SUT Trees (The horizontal axis shows the names of 12 data sets, arranged in ascending order of their corresponding averaged $\bar{\rho}$ displayed beneath each data set's abbreviation.)

both the signal-to-noise ratio (SNR) and dimension are critical characteristics of data sets. Therefore, we will first assess the performance of WRF_{opt} on semi-synthetic data sets under various noise scenarios. Moreover, given our prior analysis on low-dimensional data sets, we now turn our attention to the performance of different algorithms on high-dimensional data, which has been augmented via feature engineering.

4.2.1 Improvement Ratio vs SNR

In order to examine the effectiveness of the WRF_{opt} on noisy data sets, we consider three different noise injection schemes, similar to Reis et al. (2018):[4]

(a) Noise in the predictive variables only (that is, $\mathbf{X}$),

(b) Noise in the response variable only (that is, $\mathbf{y}$),

(c) Noise both in the predictive and response variables (that is, $\mathbf{X}$ and $\mathbf{y}$).

We follow Reis et al. (2018) to configure noise injection:

4. They focus on classification scenarios.

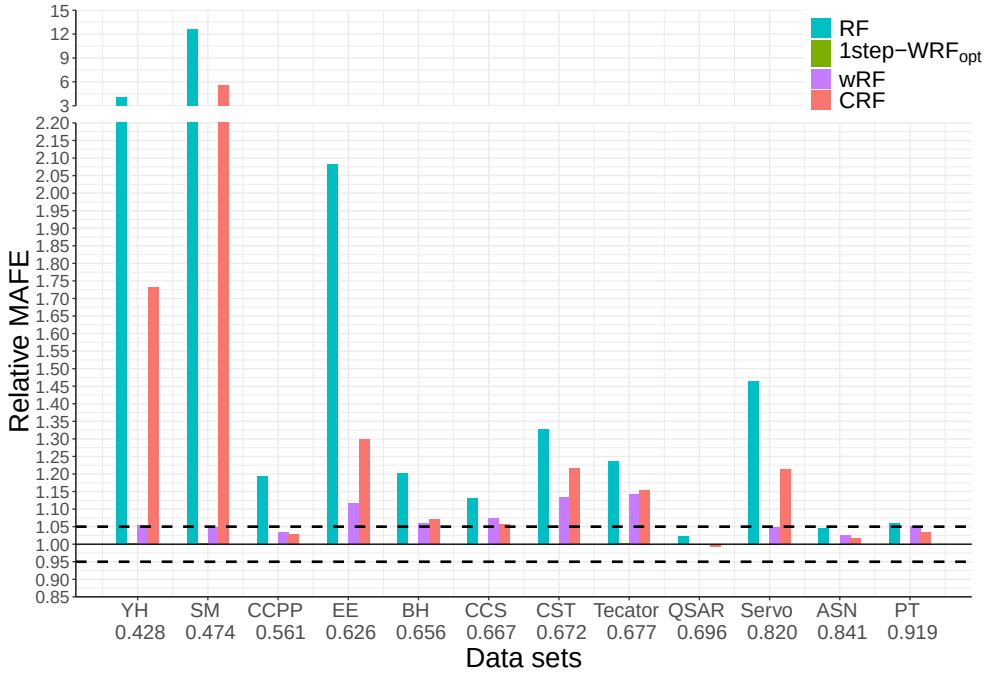

Figure 4: Relative MAFE for Different Forests with SUT Trees (The horizontal axis shows the names of 12 data sets, arranged in ascending order of their corresponding averaged $\bar{\rho}$ displayed beneath each data set's abbreviation.)

- *For response variable or continuous predictive variables*—The noise comes from a Gaussian distribution, with its magnitude randomly set for each object and feature in the data set. Specifically, a per-object noise coefficient, N_o for each $o \in \{1, \ldots, n\}$, is randomly drawn from a uniform distribution between 0 and 1. Likewise, a per-feature noise coefficient, N_f for each $f = 1, \ldots, p$, follows the same uniform distribution. Then, the noise coefficient for a specific measurement, corresponding to a particular object-feature pair, is defined as $N_{o,f} = N_o \times N_f \times N_s$. Here, N_s is the overall noise coefficient for the data set, which will be varied from 0 to 1 throughout the experiment. The synthetic noise for the f^{th} feature is defined by $\sigma_{o,f} = N_{o,f} \times \sigma_f$, with σ_f being the standard deviation of the given feature across all objects. This multiplication ensures that the noisy data retains the same physical units as the original. Finally, the noisy measurement for each object-feature pair is drawn from the normal distribution Normal $\left(x_{o,f}, \sigma_{o,f}^2\right)$, where $x_{o,f}$ denotes the original measurement. To evaluate performances of different strategies as a function of the noise level in the data set, $\sigma_{o,f}/\sigma_f$, averaged across different features and objects, can be used to represent the average scatter due to the noise with respect to the intrinsic scatter of the original data set (Reis et al., 2018). Thus, it can be considered as the average inverse of the relative

SNR of the semi-synthetic data set, with respect to the real data set. An N_s value of 0 implies no noise injection, while a value of 1 yields an relative SNR close to the scale of 4. By combining the inherent noise from the original data with the injected noise, the real SNR of the noisy data (that is, semi-synthetic data) is sufficiently low to mimic the actual noisy scenarios encountered in real world.

- *For categorical predictive variables*—the probability of a class switch is determined by $p_{o,f} = p_o \times p_f$. In this context, p_o and p_f are independently drawn from a uniform distribution ranging between 0 and 0.5. Consequently, the class of the object-feature pair is randomly switched to another class with a probability of $p_{o,f}/(C-1)$, where C represents the total number of classes for the given feature. That is, the class remains unchanged with a probability of $1-p_{o,f}$. Note that the number of categorical features in the data set is considerably less than that of continuous features, resulting in a relatively minor impact on the overall noise level of the data. For simplicity, we opt not to introduce an additional parameter for varying noise levels specifically for categorical predictive variables.

The 12 noisy data sets serve as inputs to the RF algorithms, and we focus on the relative performance of four weighted RFs over conventional RFs, rather than the predictive performance itself. Hence, the improvement ratio (IR) is used to evaluate different algorithms:

$$\mathrm{IR} = \frac{1}{D}\sum_{d=1}^{D}\left\{\frac{\sum_{i=1}^{n_{\text{test}}}\left(y_i - \widehat{y}_{i,d}^{\mathrm{RF}}\right)^2}{\sum_{i=1}^{n_{\text{test}}}\left(y_i - \widehat{y}_{i,d}^{\mathrm{XRF}}\right)^2} - 1\right\},$$

where $\widehat{y}_{i,d}^{\mathrm{RF}}$ and $\widehat{y}_{i,d}^{\mathrm{XRF}}$ represent the forecasts for y_i by conventional RF and any weighted RF in the d^{th} repetition, respectively. If not specified, other settings remain the same as those in Section 4.1.1.

The IR values of all weighted strategies under various noise levels were calculated for all 12 data sets. For the sake of simplicity, we display only two representative figures here, while the remaining figures can be found in Appendix D. Figures 5 and 6 illustrate the IR of the 1step-WRF$_{\text{opt}}$, 2steps-WRF$_{\text{opt}}$, wRF and CRF on the ASN and CCPP data sets, represented by blue, green, purple, and red lines, respectively. Additionally, the averaged $\bar{\rho}$ under different noise scenarios are shown using yellow bars.

It is clear from Figures 5 and 6 that the WRF$_{\text{opt}}$ algorithms consistently outperform the conventional RF and two competitors across all noise scenarios. As expected, the IR values of all weighted RFs decrease as the noise level increases. This phenomenon can be attributed to data sets with higher levels of noise yielding trees that are less informative and possess reduced predictive power, thereby diminishing the benefits from post-processing (that is, weighting strategy). What stands out in the figures is the inverse relationship between $\bar{\rho}$ and IR, highlighting the practical significance of $\bar{\rho}$ as an indicator for the effectiveness of WRF$_{\text{opt}}$. Regarding Figures D.1 - D.10 in Appendix D, the trends are mostly align to those observed in the ASN and CCPP data sets. However, note that the WRF$_{\text{opt}}$ methods continue to underperform compared to the conventional RF or the two competitors in varied noise scenarios on the data sets where they failed in Section 4.1. To summarize, SNR plays an unneglectable role in affecting the improvability of the WRF$_{\text{opt}}$ over equal weight strategy.

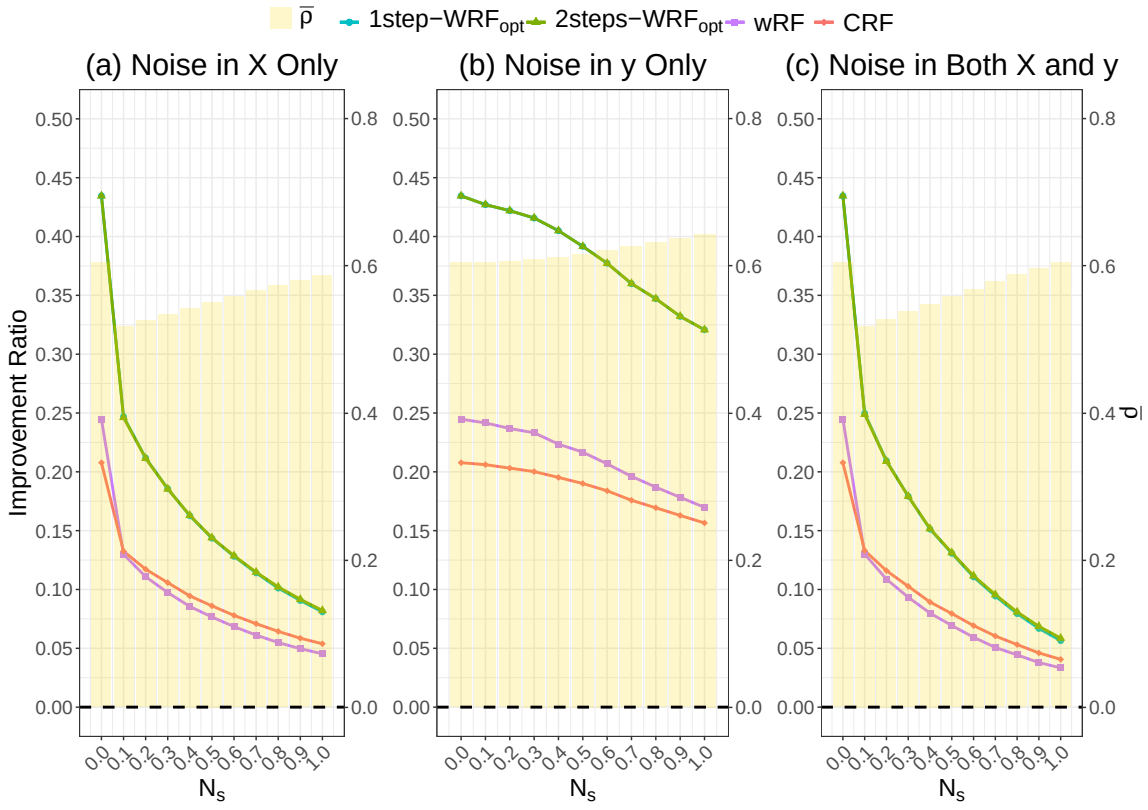

Figure 5: Improvement Ratio vs Noise on ASN Data Set (The green and blue lines are nearly indistinguishable from each other, owing to the similar performance between the 1step-WRF$_{\text{opt}}$ and 2steps-WRF$_{\text{opt}}$. This pattern may consistent across subsequent Figures 6, D.1 - D.10 and D.12 - D.23).

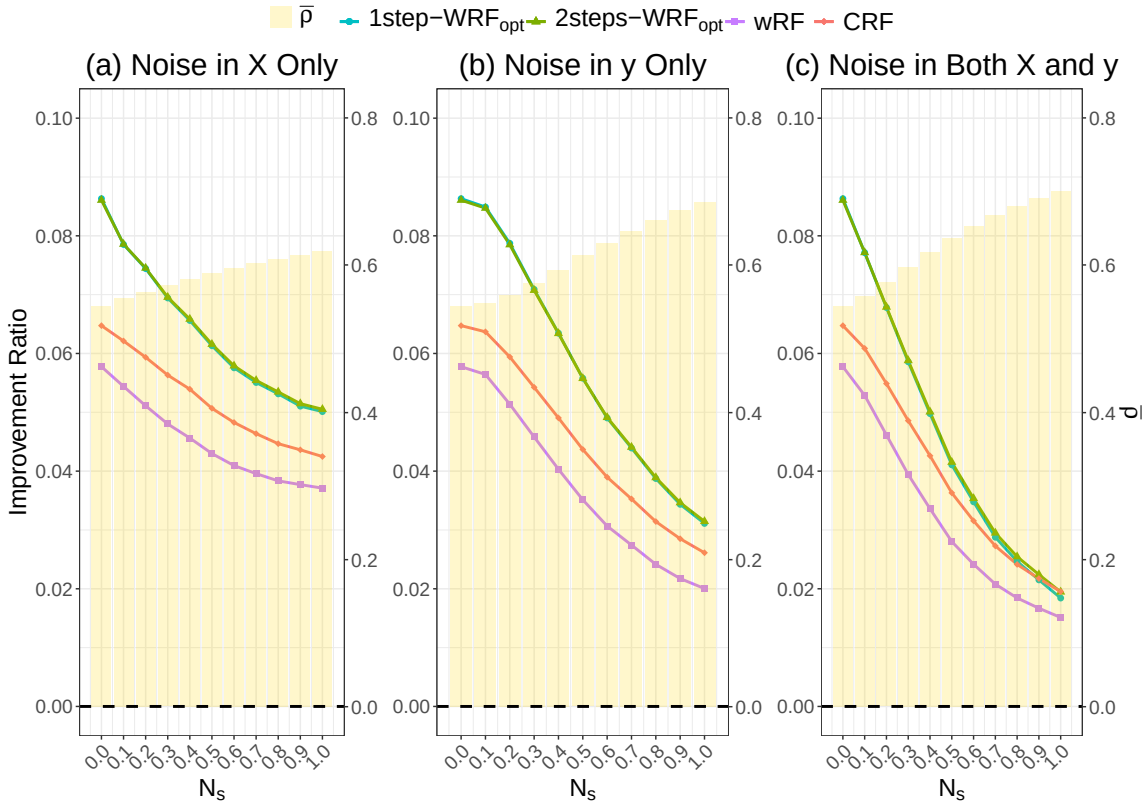

Figure 6: Improvement Ratio vs Noise on CCPP Data Set

More specifically, it is observed from numerical study that when the data are deemed to have relatively low noise levels, WRF$_{\text{opt}}$ strategies typically provide more obvious improvements.

4.2.2 Performance under High-Dimensional Setting

Having analyzed performance on low-dimensional data sets earlier (that is, where the sample size is much larger than the number of attributes), we now shift our focus to high-dimensional data sets. For this purpose, we semi-synthesize high-dimensional data using all the data sets from Section 4.1, generating additional attributes by `sklearn.preprocessing.PolynomialFeatures` (Pedregosa et al., 2011). More specifically, we create new attributes comprising all polynomial combinations of the original attributes up to a specified degree. For instance, in a two-dimensional feature set represented as $\{a, b\}$, the degree-2 polynomial attributes would be $\{1, a, b, a^2, ab, b^2\}$. The details of these 12 semi-synthetic data sets are provided in Table 8, with the last column indicating the `degree` parameter used in `PolynomialFeatures`.

In alignment with the analytical methods in Section 4.1, we present the risks of different RFs regarding MSFE and MAFE in Tables D.1 and D.2, respectively. For brevity, these tables are included in Appendix D, as their patterns are similar to those in Tables 3 and 4

Data set	Abbreviation	Attributes	Samples	Degree
Boston Housing	BH	558	506	3
Servo	Servo	363	167	3
Concrete Compressive Strength	CCS	1286	1030	5
Airfoil Self-Noise	ASN	1286	1503	8
Combined Cycle Power Plant	CCPP	10625	9568	20
Concrete Slump Test	CST	119	103	3
Energy Efficiency	EE	1286	768	5
Parkinsons Telemonitoring	PT	10625	5875	4
QSAR aquatic toxicity	QSAR	494	546	4
Synchronous Machine	SM	494	557	8
Yacht Hydrodynamics	YH	461	308	5
Tecator	Tecator	7874	240	1

Table 8: Summary of 12 High-Dimensional Data Sets

but with notably smaller values.[5] Additionally, the relative MSFE and MAFE, based on the 2steps-WRF_{opt}, are depicted in Figures 7 and 8, respectively. While the scale of improvement for conventional RFs is less pronounced in these figures compared to their original counterparts, which is expected given the significant risk reduction achieved through feature engineering. Generally, it is challenging to achieve comparable levels of improvement in a model that has already undergone substantial optimization. Nonetheless, the WRF_{opt} methods continue to exhibit commendable performance on high-dimensional data. Specifically, they significantly improve conventional RFs in 5 out of 12 data sets in terms of MSFE, and in 4 out of 12 data sets in terms of MAFE. Moreover, none of the competitors consistently outperform the WRF_{opt} across all scenarios, highlighting the robustness of WRF_{opt} in high-dimensional situations. Once again, it is clear that the WRF_{opt} methods show more pronounced improvement when associated with relatively low $\bar{\rho}$, similar to the low-dimensional scenario. As mentioned before, this phenomenon is owing to the fact that the data set with large $\bar{\rho}$ yields similar base learners and thus the conventional RF with equal weights can also provide promising prediction in this situation.

5. Conclusion

This study proposes an optimal weighted RF algorithm for regression and the corresponding accelerated variant is also studied. These methods are proven to be asymptotically optimal. We also employ a cost-efficient indicator, $\bar{\rho}$, to guide RF users in deciding whether to consider the WRF_{opt} methods as a post-processing technique to enhance predictive performance. Empirical evidence demonstrates that the proposed methods yield lower risks compared to RFs with equal weights and other existing unequally weighted forests, on both low-

5. Conventional RFs exhibit enhanced predictive power after feature engineering, potentially reducing risks to as little as one percent of the original. For a comprehensive comparison of relative risks between conventional RFs using augmented (that is, high-dimensional) data sets and original data sets, refer to Figure D.11 in Appendix D. It is a common practice in machine learning to employ automated feature engineering tools to boost the predictive capabilities of models.

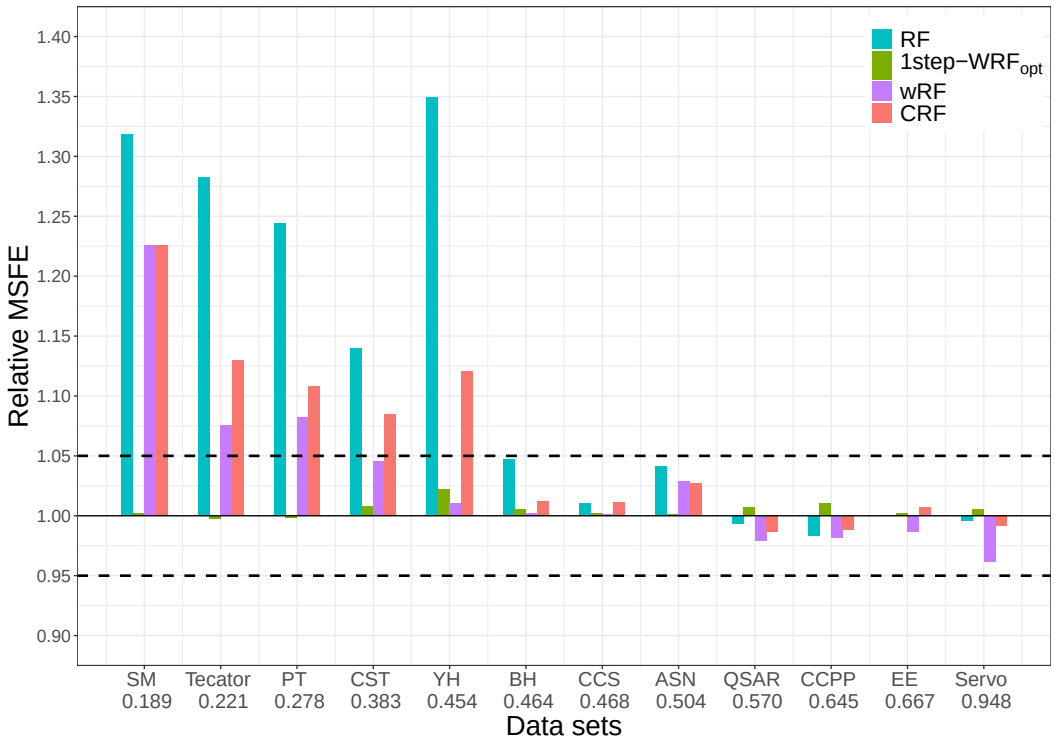

Figure 7: Relative MSFE for Different Forests on High-Dimensional Data (The horizontal axis shows the names of 12 data sets, arranged in ascending order of their corresponding averaged $\bar{\rho}$ displayed beneath each data set's abbreviation.)

dimensional and high-dimensional data. Additionally, our weighting methods show good robustness across various configurations of key hyper parameters within the RF algorithm, as verified in numerical experiments. In light of the results obtained, we provide the following suggestions for RF users seeking to improve the predictive capabilities of their finely-tuned models:

1. We suggest RF users to consider the WRF$_{\text{opt}}$ under the following scenarios:

   (a) Pre-RF building: The data are considered to exhibit a relatively high signal-to-noise ratio or to be well cleansed.

   (b) Post-RF building: The $\bar{\rho}$ of the RF model is relatively small, for instance, $\bar{\rho} < 0.5$.

2. We recommend using the 2steps-WRF$_{\text{opt}}$ rather than the 1step-WRF$_{\text{opt}}$ as the former offers comparable performance but is less computationally burdensome.

While the current study focuses on regression, it is also important to study the optimal weighted RF for classification with different loss functions. We identify this as a promising future research direction.

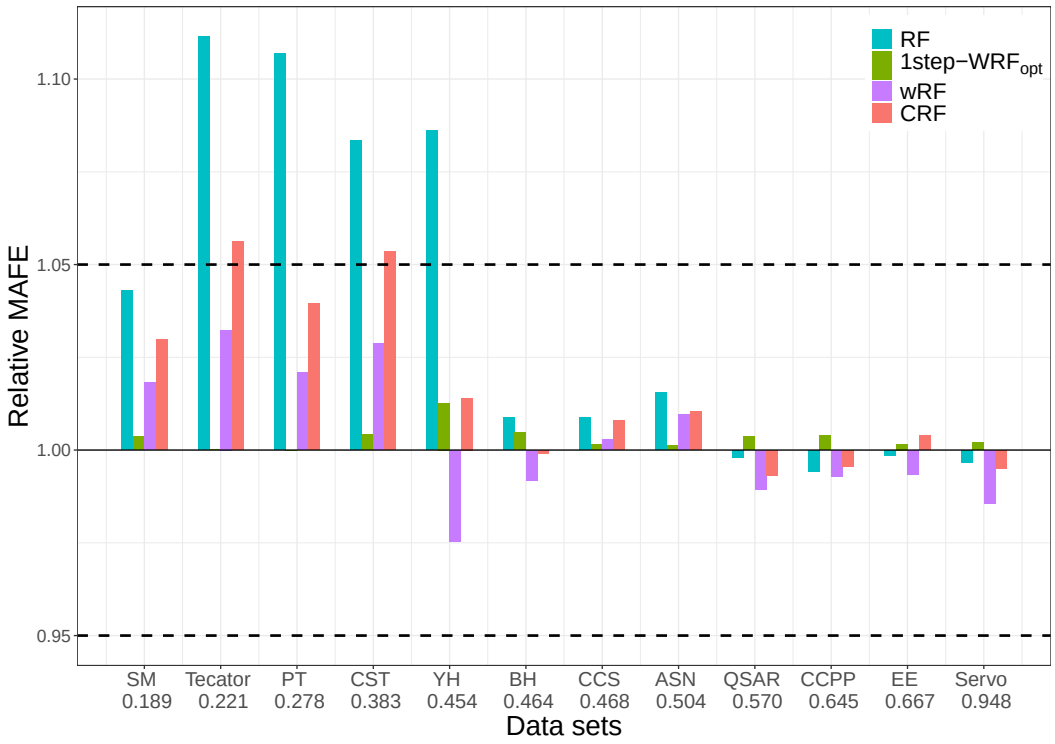

Figure 8: Relative MAFE for Different Forests on High-Dimensional Data (The horizontal axis shows the names of 12 data sets, arranged in ascending order of their corresponding averaged $\bar{\rho}$ displayed beneath each data set's abbreviation.)

Acknowledgments

The authors would like to thank the Action Editor Professor Mladen Kolar and two anonymous reviewers for their constructive suggestions and comments that have substantially improved earlier version of this article. Dalei Yu was supported by National Natural Science Foundation of China (NNSFC, grant no. 12071414). Xinyu Zhang was supported by National Natural Science Foundation of China (NNSFC, grant nos. 71925007, 72091212 and 12288201), Beijing Natural Science Foundation (Z240004) and the CAS Project for Young Scientists in Basic Research (YSBR-008). All authors equally contributed to this work and are listed in the alphabetical order.

Appendix Appendix A. Tree-Building Algorithms

We will elucidate the differences between the two practical splitting criteria in this appendix. When constructing RFs using CART trees, we employ Algorithm A.1, and when building them with SUT trees, we adopt Algorithm A.2. The structures of SUT trees are developed in an unsupervised manner, eliminating the reliance on response values during splits, whereas CART trees use the information of $\mathbf{y}$ to obtain the best splitting variables and cut points. They are the same in other procedures, such as growing on the bootstrap data. Note that there are many ways to grow SUT trees, provided that their splitting processes are not dependent on response values. Algorithm A.2 represents just one of these methods.

When selecting the probability sequence $\mathscr{P}$ in Algorithm A.2, we built conventional RFs with CART trees in the validation data to compute variables importance. The variable importance is the total decrease in node impurities from splitting on the variable, averaged over all trees. For regression, the node impurity is measured by residual sum of squares. After that, the probability sequence $\mathscr{P}$ was determined by the normalized variables importance.

Appendix Appendix B. Detailed Demonstration of the 2steps-WRF$_{\text{opt}}$ and Competitors

In this appendix, we provide details of the 2steps-WRF$_{\text{opt}}$ algorithm, and present an exposition of two weighted RF algorithms introduced in Section 1. Since these two competitors are proposed for classification trees, we further describe a methodology to transform classification patterns into regression patterns to address regression tasks.

B.1 2steps-WRF$_{\text{opt}}$

The following Algorithm B.1 presents the complete 2steps-WRF$_{\text{opt}}$ algorithm described in Section 3.1.

B.2 Weighted RF (wRF)

Much of the current literature on binary classification pay particular attention to out-of-bag data. Namely, Li et al. (2010) use the accuracy in the out-of-bag data as an index of the classification ability of a given tree. This metric is subsequently employed to assign weights to the individual trees. Winham et al. (2013) provide a family of weights choice based on the prediction error in the out-of-bag data of each tree. The reason why using out-of-bag individuals instead of another shared data set is that it gives internal estimates that are helpful in understanding the predictive performance and how to improve it without testing data set aside (Breiman, 2001).

Specifically, Winham et al. (2013) define the tree-level prediction error (tPE), measuring the predictive ability of the m^{th} tree as follows

$$t\text{PE}_m = \frac{1}{\sum_{i=1}^{n} \text{OOB}_{im}} \sum_{i=1}^{n} |v_{im} - y_i| \cdot \text{OOB}_{im}, \tag{B.1}$$

where v_{im} is the vote for the i^{th} subject in the m^{th} tree, and OOB_{im} is the indicator for the out-of-bag status of the i^{th} subject in the m^{th} tree. By drawing on the concept of tPE, they

Algorithm A.1: CART

Split_a_node(S)

Input: The local learning subset S corresponding to the node we want to split

Output: A split $[a < c]$ or nothing

-If **Stop_split**(S) is TRUE then return nothing.

-Otherwise select q attributes $A_q = \left\{a_{j_1}, \ldots, a_{j_q}\right\}$ randomly among all non constant (in S) candidate attributes;

-Return the best split s_*, where $s_* =$ **Find_the_best_split**(S, A_q).

Find_the_best_split(S, A_q)

Input: The subset S and the selected attribute list A_q

Output: The best split

- Seek the splitting variable a_j and cut point c that solve $\min_{d\in\{j_1,\ldots,j_q\},c} \mathrm{CART}_{\mathbf{t}}(d, c)$;
- Return the split $[a_j < c]$.

Stop_split(S)

Input: A subset S

Output: A boolean

- If $|S| <$ `nodesize`, then return TRUE;
- If all attributes are constant in S, then return TRUE;
- If the output is constant in S, then return TRUE;
- Otherwise, return FALSE.

have been able to show that weights inversely related to tPE are appropriate. Such as

$$w_{(m)} = 1 - t\mathrm{PE}_m, \tag{B.2}$$

$$w_{(m)} = \exp\left(\frac{1}{t\mathrm{PE}_m}\right), \tag{B.3}$$

and

$$w_{(m)} = \left(\frac{1}{t\mathrm{PE}_m}\right)^{\lambda} \text{for some } \lambda. \tag{B.4}$$

In their proposed wRF algorithm, they normalized weights of the form

$$w_{(m)} = \frac{w_{(m)}}{\sum_{m=1}^{M_n} w_{(m)}}.$$

The classification model can be easily turned into a regression model by simply changing (B.1) to the following definition

$$t\mathrm{PE}'_m = \frac{1}{\sum_{i=1}^n \mathrm{OOB}_{im}} \sum_{i=1}^n \left| \hat{f}_{(m)}(\mathbf{x}_i) - y_i \right| \cdot \mathrm{OOB}_{im}, \tag{B.5}$$

where $\hat{f}_{(m)}(\mathbf{x}_i)$ is the prediction for y_i by the m^{th} tree. The details of the wRF in regression pattern is in Algorithm B.2, which selects (B.4) for example. For simplicity, we only present the best result of the wRF family as a representative in Section 4.

B.3 Cesáro RF (CRF)

Another unequally weighted RF mentioned earlier is the CRF proposed by Pham and Olafsson (2019), which replaces the regular average with the Cesáro average. Their method is based on a renowned theory that if a sequence converges to a number c, then the Cesáro sequence also converges to c. To implement the CRF, a strategy for sequencing M_n trees from best to worst must be established. This can be done by ranking trees based on their out-of-bag error rates or accuracy on a separate training set. Next, a weight sequence $\{w_{(m)}\}_{m=1}^{M_n}$ is obtained by arranging weights in descending order, where $w_{(m)} = \sum_{\nu=m}^{M_n} \nu^{-1}$, with normalizer being $\sum_{m=1}^{M_n} \sum_{\nu=m}^{M_n} \nu^{-1}$.

This classification model can be easily converted into a regression model as well through a simple modification in the sequencing methods. We can draw $t\mathrm{PE}'$ defined by the wRF algorithm, and subsequently rank trees using out-of-bag data. The details of the CRF in regression pattern are in Algorithm B.3.

Appendix Appendix C. Proofs and Derivations

In the current appendix, we provide proofs of the lemmas and theorems in Section 3.2, along with derivations and further discussions to support our findings.

C.1 Preliminary Results

The following preliminary results will be used in the proofs. Note that this appendix and the following ones, Appendices C.2 - C.3, focus on the SUT methodology and are based on the work by Qiu et al. (2020). Their theoretical framework presumes tree structures are independent of response values (that is, "hat matrix" is independent of $\mathbf{y}$). Consequently, notations throughout these appendices are distinguished with the $\star$ script for clarity and consistency.

Lemma C.1 *For each $m \in \{1, \ldots, M_n\}$, let $\mathbf{P}_{\star \mathrm{X}(m)}$ denote the "hat matrix" corresponding to any algorithm for ensemble. In addition to the independence of $\mathbf{P}_{\star \mathrm{X}(m)}$ from $\mathbf{y}$, assume the following Conditions C.4 - C.9 hold as well.*[6]

C.4 There exists a positive constant c_0 such that for all $m, r \in \{1, \ldots, M_n\}$,

$$\operatorname{trace}\left(\mathbf{P}_{\star \mathrm{X}(m)} \mathbf{P}_{\star \mathrm{X}(m)}^\top\right) \geqslant c_0 > 0 \quad \text{and} \quad \operatorname{trace}\left(\mathbf{P}_{\star \mathrm{X}(m)} \mathbf{P}_{\star \mathrm{X}(r)}^\top\right) \geqslant 0,$$

6. The condition labels here are directly adopted from Qiu et al. (2020) for ease of reference.

almost surely.

C.5 There exists a positive constant c_1 *such that for all* $m \in \{1, \ldots, M_n\}$,

$$\zeta_{\max}\left(\mathbf{P}_{\star \mathrm{X}(m)}\mathbf{P}_{\star \mathrm{X}(m)}^{\top}\right) \leqslant c_1,$$

almost surely, where $\zeta_{\max}(\mathbf{B})$ *denotes the largest singular value of a generic matrix* $\mathbf{B}$.

C.6 There exists a positive constant c_2 *such that for all* $m, r \in \{1, \ldots, M_n\}$,

$$\operatorname{trace}\left(\mathbf{P}_{\star \mathrm{X}(m)}^{2}\right) \leqslant c_2 \operatorname{trace}\left(\mathbf{P}_{\star \mathrm{X}(m)}^{\top}\mathbf{P}_{\star \mathrm{X}(m)}\right),$$

and

$$\operatorname{trace}\left(\mathbf{P}_{\star \mathrm{X}(r)}^{\top}\mathbf{P}_{\star \mathrm{X}(m)}\mathbf{P}_{\star \mathrm{X}(r)}^{\top}\mathbf{P}_{\star \mathrm{X}(m)}\right) \leqslant c_2 \operatorname{trace}\left(\mathbf{P}_{\star \mathrm{X}(m)}^{\top}\mathbf{P}_{\star \mathrm{X}(m)}\right),$$

almost surely.

C.7 $\xi_{\star n}^{-1} M_n^2 \to 0$ *almost surely, as* $n \to \infty$, *and* $\mathbb{E}(\xi_{\star n}^{-1} M_n^2)$ *exists for any fixed* $n \geqslant 1$.

C.8 There exists a positive constant v *such that* $\mathbb{E}\left(e_i^4 \mid \mathbf{x}_i^0\right) \leqslant v < \infty$ *almost surely for* $i = 1, \ldots, n$.

C.9 $\bar{\iota}_n = \max_{1 \leqslant m \leqslant M_n} \max_{1 \leqslant i \leqslant n} \iota_{ii}^{(m)\star} = O(n^{-1/2})$ *almost surely, as* $n \to \infty$, *where* $\iota_{ii}^{(m)\star}$ *is the* i^{th} *diagonal element of* $\mathbf{P}_{\star \mathrm{X}(m)}$.

Then, as $n \to \infty$, *we have (7).*

The original version of Lemma C.1 is established in the proof of Theorem 1 in Qiu et al. (2020) for *non-stochastic* $\mathbf{X}$. However, as demonstrated in Appendix C.6, by substituting expectations with conditional expectations, and applying the Law of Iterated (or Total) Expectation, Pull-out rule and Lebesgue's Dominated Convergence Theorem, the proof by Qiu et al. (2020) can be readily extended to accommodate scenarios with *stochastic* $\mathbf{X}$. Thus we omit the step-by-step proof in the current study.

Next, we introduce four other lemmas for proving Theorems 1 - 3.

Lemma C.2 (Gao et al., 2019) *Let*

$$\widetilde{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathscr{H}} \left\{ L_n(\mathbf{w}) + a_n(\mathbf{w}) + b_n \right\},$$

where $a_n(\mathbf{w})$ *is a term related to* $\mathbf{w}$ *and* b_n *is a term unrelated to* $\mathbf{w}$. *If*

$$\sup_{\mathbf{w} \in \mathscr{H}} \left| a_n(\mathbf{w}) \right| / R_n(\mathbf{w}) = o_p(1), \sup_{\mathbf{w} \in \mathscr{H}} \left| R_n(\mathbf{w}) - L_n(\mathbf{w}) \right| / R_n(\mathbf{w}) = o_p(1),$$

and there exists a constant c *and a positive integer* n_0 *so that when* $n \geqslant n_0$ *and* $\inf_{\mathbf{w} \in \mathscr{H}} R_n(\mathbf{w}) \geqslant c > 0$ *almost surely, then* $L_n(\widetilde{\mathbf{w}}) / \inf_{\mathbf{w} \in \mathscr{H}} L_n(\mathbf{w}) \to 1$ *in probability.*

Lemma C.3 (Saniuk and Rhodes, 1987) *For any $n \times n$ matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ with both $\mathbf{G}_1, \mathbf{G}_2 \geqslant 0$,*

$$\operatorname{trace}(\mathbf{G}_1\mathbf{G}_2) \leqslant \|\mathbf{G}_1\|_2 \operatorname{trace}(\mathbf{G}_2),$$

where $\|\cdot\|_2$ denotes the spectral norm or largest singular value. Besides, for any $n \times n$ symmetric matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ with $\mathbf{G}_2 \geqslant 0$,

$$\operatorname{trace}(\mathbf{G}_1\mathbf{G}_2) \leqslant \lambda_{\max}(\mathbf{G}_1) \operatorname{trace}(\mathbf{G}_2),$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue.

Lemma C.4 (Buldygin and Moskvichova, 2013) *Let $\beta(p)$ denote a Bernoulli random variable with probabilities $p \in [0,1]$ and $q = 1-p$. Let $\beta^{(0)}(p)$ be the centered Bernoulli random variable with parameter p, that is*

$$\beta^{(0)}(p) = \beta(p) - \mathbb{E}\{\beta(p)\} = \beta(p) - p.$$

Then, when $p \in (0, 1/2)$, the variance proxy (or the square of sub-Gaussian norm) of $\beta^{(0)}(p)$ is

$$\tau^2(p) = \frac{\frac{1}{2} - p}{\log\left(\frac{1}{p} - 1\right)},$$

which is a monotonically increasing function of p.

Lemma C.5 (Zhang and Chen, 2021) *Let $\{Z_i\}_{i=1}^n$ be sub-Gaussian random variables (without independence assumption) with mean zero and variance proxy τ^2. Then, we have*

$$\mathbb{E}\left(\max_{1 \leqslant i \leqslant n} |Z_i|\right) \leqslant \tau\sqrt{2\log(2n)}.$$

C.2 Proof of Theorem 1

To verify (7), it remains to verify that Conditions 1 - 4 in the main paper can guarantee Conditions C.4 - C.9 in Lemma C.1. Having clearly configured the base learners within RFs, we now verify that Conditions C.4 - C.6 are satisfied when $\mathbf{P}_{\star \mathrm{BL}(m)}$ is used in place of $\mathbf{P}_{\star \mathrm{X}(m)}$.

For each $m = 1, \ldots, M_n$, $i = 1, \ldots, n$ and $j = 1, \ldots, n$, recall that

$$P^{\star}_{\mathrm{BL}(m),ij} = \sum_{s=1}^{2^{k(m)}} \frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^{\star}_{k(m),s}\right)\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^{\star}_{k(m),s}\right)}{\sum_{j=1}^{n} h_{(m),j}\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^{\star}_{k(m),s}\right)},$$

which is the counterpart of $P_{\mathrm{BL}(m)i,j}$ (the precise definition of $P_{\mathrm{BL}(m)i,j}$ is given in Equation 12),[7] pertaining to SUT trees. Then, there exists a positive constant c such that

$$\operatorname{trace}\left(\mathbf{P}^{\top}_{\star \mathrm{BL}(m)}\mathbf{P}_{\star \mathrm{BL}(m)}\right)$$

7. This notation is introduced in Section 3.2.2, under the context of CART trees, but is also applicable to SUT trees.

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{2^{k_{(m)}}}\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}\right\}^2\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{2^{k_{(m)}}}\sum_{k=1}^{2^{k_{(m)}}}\frac{h^2_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},k}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},k}\right)}{\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},k}\right)}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{2^{k_{(m)}}}\frac{h^2_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\left\{\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\right\}^2}\\
&\geqslant c\sum_{s=1}^{2^{k_{(m)}}}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{h_{(m),i}h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\left\{\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\right\}^2}\\
&= c\sum_{s=1}^{2^{k_{(m)}}}\frac{\sum_{i=1}^{n}h_{(m),i}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\left\{\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\right\}^2}\\
&= c\,\ell_{(m)}\\
&> 0, \qquad\qquad (C.1)
\end{aligned}
$$

almost surely, where the fourth step is from Condition 4. Clearly, for all $m=1,\ldots,M_n$, the elements of $\mathbf{P}_{\star\mathrm{BL}(m)}$ are non-negative. Therefore, for all $m,r\in\{1,\ldots,M_n\}$, it is clear that

$$\operatorname{trace}\left(\mathbf{P}_{\star\mathrm{BL}(m)}\mathbf{P}^{\top}_{\star\mathrm{BL}(r)}\right)\geqslant 0.$$

Thus, Condition C.4 in Lemma C.1 is satisfied. Besides, we have

$$
\begin{aligned}
\left\|\mathbf{P}_{\star\mathrm{BL}(m)}\right\|_{\infty} &= \max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}\sum_{s=1}^{2^{k_{(m)}}}\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\sum_{l=1}^{n}h_{(m),l}\mathbb{I}\left(\mathbf{x}_l\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}\\
&= \max_{1\leqslant i\leqslant n}\sum_{s=1}^{2^{k_{(m)}}}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\frac{\sum_{j=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\sum_{l=1}^{n}h_{(m),l}\mathbb{I}\left(\mathbf{x}_l\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}\\
&= 1, \qquad\qquad (C.2)
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|\mathbf{P}_{\star\mathrm{BL}(m)}\right\|_{1} &= \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}\sum_{s=1}^{2^{k_{(m)}}}\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\sum_{l=1}^{n}h_{(m),l}\mathbb{I}\left(\mathbf{x}_l\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}\\
&= \max_{1\leqslant j\leqslant n}\sum_{s=1}^{2^{k_{(m)}}}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\frac{\sum_{i=1}^{n}h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\sum_{l=1}^{n}h_{(m),l}\mathbb{I}\left(\mathbf{x}_l\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}\\
&\leqslant c\max_{1\leqslant j\leqslant n}\sum_{s=1}^{2^{k_{(m)}}}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^{\star}_{k_{(m)},s}\right)\frac{\sum_{i=1}^{n}h_{(m),i}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}{\sum_{l=1}^{n}h_{(m),l}\mathbb{I}\left(\mathbf{x}_l\in\mathbf{t}^{\star}_{k_{(m)},s}\right)}
\end{aligned}
$$

$$
\begin{aligned}
=& c \max_{1 \leqslant j \leqslant n} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^{\star}_{k_{(m)},s}\right) \\
=& c, \qquad \text{(C.3)}
\end{aligned}
$$

almost surely. Here, the last step in (C.2) is from the fact that $\sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^{\star}_{k_{(m)},s}\right) \equiv 1$, and the inequality in (C.3) comes from Condition 4. Combining (C.2) and (C.3), we have

$$
\begin{aligned}
\zeta^2_{\max}\left(\mathbf{P}_{\star \mathrm{BL}(m)} \mathbf{P}^{\top}_{\star \mathrm{BL}(m)}\right) =& \zeta^2_{\max}\left(\mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(m)}\right) \\
=& \lambda_{\max}\left(\mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(m)} \mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(m)}\right) \\
=& \left\| \mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(m)} \right\|^2 \\
\leqslant& \left\| \mathbf{P}_{\star \mathrm{BL}(m)} \right\|_{\infty} \left\| \mathbf{P}_{\star \mathrm{BL}(m)} \right\|_1 \\
\leqslant& c,
\end{aligned}
$$

almost surely. Thus, Condition C.5 in Lemma C.1 is satisfied. In addition, from Condition 4 and (C.1), we have

$$
\begin{aligned}
\mathrm{trace}\left(\mathbf{P}^2_{\star \mathrm{BL}(m)}\right) &= \sum_{i=1}^{n} \sum_{l=1}^{n} \sum_{s=1}^{2^{k_{(m)}}} \frac{h_{(m),l} h_{(m),i} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^{\star}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_l \in \mathbf{t}^{\star}_{k_{(m)},s}\right)}{\left\{\sum_{j=1}^{n} h_{(m),j} \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^{\star}_{k_{(m)},s}\right)\right\}^2} \\
&\leqslant c \sum_{i=1}^{n} \sum_{l=1}^{n} \sum_{s=1}^{2^{k_{(m)}}} \frac{h^2_{(m),l} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^{\star}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_l \in \mathbf{t}^{\star}_{k_{(m)},s}\right)}{\left\{\sum_{j=1}^{n} h_{(m),j} \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^{\star}_{k_{(m)},s}\right)\right\}^2} \\
&= c\, \mathrm{trace}\left(\mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(m)}\right),
\end{aligned}
$$

almost surely. Additionally, by Lemma C.3, for each $m, r \in \{1, \ldots, M_n\}$, we have

$$
\begin{aligned}
& \mathrm{trace}\left(\mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(r)} \mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(r)}\right) \\
\leqslant& \mathrm{trace}\left(\mathbf{P}_{\star \mathrm{BL}(r)} \mathbf{P}^{\top}_{\star \mathrm{BL}(r)} \mathbf{P}_{\star \mathrm{BL}(m)} \mathbf{P}^{\top}_{\star \mathrm{BL}(m)}\right) \\
\leqslant& \lambda_{\max}\left(\mathbf{P}_{\star \mathrm{BL}(r)} \mathbf{P}^{\top}_{\star \mathrm{BL}(r)}\right) \mathrm{trace}\left(\mathbf{P}_{\star \mathrm{BL}(m)} \mathbf{P}^{\top}_{\star \mathrm{BL}(m)}\right) \\
\leqslant& c\, \mathrm{trace}\left(\mathbf{P}^{\top}_{\star \mathrm{BL}(m)} \mathbf{P}_{\star \mathrm{BL}(m)}\right),
\end{aligned}
$$

almost surely, where the first inequality comes from the fact that $\mathrm{trace}\left\{\left(\mathbf{A}^{\top}\mathbf{B}\right)^2\right\} \leqslant \mathrm{trace}\left(\mathbf{A}\mathbf{A}^{\top}\mathbf{B}\mathbf{B}^{\top}\right)$ for any generic matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, and the second step stems from Lemma C.3. Thus, Condition C.6 in Lemma C.1 is satisfied.

Further, Conditions 1 and 2 in the main text are equivalent to Conditions C.7 and C.8 in Lemma C.1, respectively. Besides, Condition 3 guarantees Conditions C.9. Based on these observations, it is readily seen that Lemma C.1 holds under Conditions 1 - 4, and this proves (7).

By (7) in Section 3, and (A34) in Proof of Theorem 2 by Qiu et al. (2020) , we further have

$$L_n^\star(\widehat{\mathbf{w}}^\star)\xi_{\star n}^{-1} = 1 + o_p(1). \tag{C.4}$$

This is owing to the fact that

$$\begin{aligned} L_n^\star(\widehat{\mathbf{w}}^\star)\xi_{\star n}^{-1} &\leqslant \sup_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\widehat{\mathbf{w}}^\star) L_n^{\star\,-1}(\mathbf{w}) \right\} \sup_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\mathbf{w}) R_n^{\star\,-1}(\mathbf{w}) \right\} \\ &\leqslant \sup_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\widehat{\mathbf{w}}^\star) L_n^{\star\,-1}(\mathbf{w}) \right\} \times \left[1 + \sup_{\mathbf{w}\in\mathscr{H}} \left\{ \left| L_n^\star(\mathbf{w}) - R_n^\star(\mathbf{w}) \right| R_n^{\star\,-1}(\mathbf{w}) \right\} \right] \\ &= 1 + o_p(1), \end{aligned} \tag{C.5}$$

and

$$\begin{aligned} L_n^\star(\widehat{\mathbf{w}}^\star)\xi_{\star n}^{-1} &\geqslant \sup_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\widehat{\mathbf{w}}^\star) L_n^{\star\,-1}(\mathbf{w}) \right\} \inf_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\mathbf{w}) R_n^{\star\,-1}(\mathbf{w}) \right\} \\ &= \sup_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\widehat{\mathbf{w}}^\star) L_n^{\star\,-1}(\mathbf{w}) \right\} \times \left[1 + \inf_{\mathbf{w}\in\mathscr{H}} \left\{ \left(L_n^\star(\mathbf{w}) - R_n^\star(\mathbf{w}) \right) R_n^{\star\,-1}(\mathbf{w}) \right\} \right] \\ &\geqslant \sup_{\mathbf{w}\in\mathscr{H}} \left\{ L_n^\star(\widehat{\mathbf{w}}^\star) L_n^{\star\,-1}(\mathbf{w}) \right\} \times \left[1 - \sup_{\mathbf{w}\in\mathscr{H}} \left\{ \left| L_n^\star(\mathbf{w}) - R_n^\star(\mathbf{w}) \right| R_n^{\star\,-1}(\mathbf{w}) \right\} \right] \\ &= 1 + o_p(1). \end{aligned} \tag{C.6}$$

Then, under the same framework of Appendix A.4 in Zhang et al. (2020), it is readily seen from Lebesgue's Dominated Convergence Theorem that $\mathbb{E}\left| \{ L_n^\star(\widehat{\mathbf{w}}^\star) - \xi_{\star n} \} \xi_{\star n}^{-1} \right| \to 0$, and this proves (8).

C.3 Proof of Theorem 2

Based on Lemma C.2, we now present the proof of Theorem 2. It is seen that

$$C_n^{\star\prime\prime}(\mathbf{w}) = C_n^\star(\mathbf{w}) + 2\sum_{i=1}^n \left(\tilde{e}_i^{\star 2} - e_i^2 \right) P_{ii}^\star(\mathbf{w}),$$

where $C_n^\star(\mathbf{w})$ and $C_n^{\star\prime\prime}(\mathbf{w})$, related to SUT trees, are the counterparts of $C_n(\mathbf{w})$ and $C_n''(\mathbf{w})$, respectively. Hence, from Lemma C.2 above and the derivation of (24) in Qiu et al. (2020), in order to prove (9), we need only to verify that

$$\sup_{\mathbf{w}\in\mathscr{H}} \left\{ \left| \sum_{i=1}^n \left(\tilde{e}_i^{\star 2} - e_i^2 \right) P_{ii}^\star(\mathbf{w}) \right| / R_n^\star(\mathbf{w}) \right\} = o_p(1). \tag{C.7}$$

For each $m = 1, \ldots, M_n$, let $\iota_{ii}^{(m)\star}$ be the i^{th} diagonal element of $\mathbf{P}_{\star \mathrm{BL}(m)}$,

$$\mathbf{Q}_{(m)}^\star = \operatorname{diag}\left(\iota_{11}^{(m)\star}, \ldots, \iota_{nn}^{(m)\star} \right),$$

$$\mathbf{Q}^\star(\mathbf{w}) = \sum_{m=1}^{M_n} w_{(m)} \mathbf{Q}_{(m)}^\star,$$

and

$$\mathbf{K}_n^\star = \operatorname{diag}\left(\tilde{e}_1^{\star 2} - e_1^2, \ldots, \tilde{e}_n^{\star 2} - e_n^2\right).$$

Then, we have

$$\sup_{\mathbf{w}\in\mathscr{K}} \left\{ \left| \sum_{i=1}^n \left(\tilde{e}_i^{\star 2} - e_i^2\right) P_{ii}^\star(\mathbf{w}) \right| / R_n^\star(\mathbf{w}) \right\} = \sup_{\mathbf{w}\in\mathscr{K}} \frac{\left|\operatorname{trace}\left\{\mathbf{Q}^\star(\mathbf{w})\mathbf{K}_n^\star\right\}\right|}{R_n^\star(\mathbf{w})}.$$

We observe that for any $\delta > 0$, under Conditions 2 and 3,

$$\begin{aligned}
\Pr\left[\sup_{\mathbf{w}\in\mathscr{K}} \left| \frac{\operatorname{trace}\left\{\mathbf{Q}^\star(\mathbf{w})\mathbf{K}_n^\star\right\}}{R_n^\star(\mathbf{w})} \right| > \delta\right] &\leqslant \sum_{m=1}^{M_n} \Pr\left\{ \left| \frac{\operatorname{trace}\left(\mathbf{Q}_{(m)}^\star \mathbf{K}_n^\star\right)}{R_n^\star(\mathbf{w})} \right| > \delta \right\} \\
&\leqslant \delta^{-1} \sum_{m=1}^{M_n} \mathbb{E}\left\{ \left| \frac{\operatorname{trace}\left(\mathbf{Q}_{(m)}^\star \mathbf{K}_n^\star\right)}{R_n^\star(\mathbf{w})} \right| \right\} \\
&\leqslant \delta^{-1} \sum_{m=1}^{M_n} \mathbb{E}\left\{ \left| \operatorname{trace}\left(\mathbf{Q}_{(m)}^\star \mathbf{K}_n^\star\right) \right| \xi_{\star n}^{-1} \right\} \\
&\leqslant \delta^{-1} \sum_{m=1}^{M_n} \mathbb{E}\left\{ \operatorname{trace}\left(\mathbf{Q}_{(m)}^\star\right) \left\|\mathbf{K}_n^\star\right\| \xi_{\star n}^{-1} \right\} \\
&= \delta^{-1} \sum_{m=1}^{M_n} \mathbb{E}\left(\ell_{(m)}^\star \left\|\mathbf{K}_n^\star\right\| \xi_{\star n}^{-1} \right) \\
&\leqslant c_1 \delta^{-1} M_n \mathbb{E}\left(n^{1/2} \left\|\mathbf{K}_n^\star\right\| \xi_{\star n}^{-1} \right) \\
&\leqslant c_1 \delta^{-1} M_n \mathbb{E}^{1/2}\left(n^{1/2} \xi_{\star n}^{-1} \right)^2 \mathbb{E}^{1/2} \left\|\mathbf{K}_n^\star\right\|^2 \\
&\leqslant c_2 \delta^{-1} \mathbb{E}^{1/2}\left(M_n n^{1/2} \xi_{\star n}^{-1} \right)^2,
\end{aligned}$$

where c_1 and c_2 are positive constants, the second inequality follows from the Markov's Inequality, the fourth inequality is obtained by Lemma C.3, the fifth inequality comes from Condition 3, the last second inequality is from the Cauchy-Schwarz Inequality, and the last step stems from the boundedness of $\mathbb{E}^{1/2}\left\|\mathbf{K}_n^\star\right\|^2$ by combining equality (C.2) and Conditions 2 and 3. Thus, (9) is proved by Condition 5 and the Lebesgue's Dominated Convergence Theorem. Similar to the proof techniques of Theorem 1, we have $L_n^\star(\widetilde{\mathbf{w}}^\star)\xi_{\star n}^{-1} = 1 + o_p(1)$, which yields (10).

C.4 Proof of Lemma 1

It follows from the definitions of $P_{\mathrm{BL}(m),ij}$ and $P_{\mathrm{BL}(m),ij}^*$ that

$$\sum_{i=1}^n \sum_{j=1}^n \left| P_{\mathrm{BL}(m),ij} - P_{\mathrm{BL}(m),ij}^* \right|$$

$$
\begin{aligned}
&\leqslant \sum_{s=1}^{2^{k_{(m)}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \right. \\
&\qquad \left. - \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right| \\
&+ \sum_{s=1}^{2^{k_{(m)}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right. \\
&\qquad \left. - \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right| \\
&\equiv \quad \Xi_{C41} + \Xi_{C42}.
\end{aligned} \tag{C.8}
$$

In addition, note that

$$
\left| \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right) \right| = \left| \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right) \right|^2.
$$

From Condition 4, one has

$$
\begin{aligned}
\Xi_{C41} \leqslant & \sum_{s=1}^{2^{k_{(m)}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \right. \\
& \qquad \left. - \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \right| \\
& + \sum_{s=1}^{2^{k_{(m)}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \right. \\
& \qquad \left. - \frac{h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right| \\
\leqslant & c \sum_{s=1}^{2^{k_{(m)}}} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\left| \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right) \right| h_{(m),i} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \\
& + c \sum_{s=1}^{2^{k_{(m)}}} \left[\frac{\sum_{j'=1}^{n} \left| \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right) \right|}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right) \sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right. \\
& \qquad \left. \times \left\{ \sum_{j=1}^{n} \sum_{i=1}^{n} h_{(m),i} h_{(m),j} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right) \right\} \right]
\end{aligned}
$$

$$
\begin{aligned}
=&2c\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\left|\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}_{k_{(m)},s}\right)-\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)\right|\\
=&2c\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\left|\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}_{k_{(m)},s}\right)-\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)\right|^2\\
\leqslant&2c\delta_n\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\left\{\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}_{k_{(m)},s}\right)+\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)\right\}\\
\leqslant&2c\bar{K}_n(\mathscr{N}+\mathscr{N}^*)\delta_n,
\end{aligned}
\tag{C.9}
$$

almost surely, and

$$
\begin{aligned}
\Xi_{C42}\leqslant&\sum_{s=1}^{2^{k_{(m)}}}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{h_{(m),j}\left|\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}_{k_{(m)},s}\right)-\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\right|\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}^*_{k_{(m)},s}\right)}\\
\leqslant&\bar{K}_n(\mathscr{N}+\mathscr{N}^*)\delta_n.
\end{aligned}
\tag{C.10}
$$

Thus, (14) is verified by combining (C.9) - (C.10) with (C.8).

Besides, for each $i=1,\ldots,n$, we have

$$
\begin{aligned}
&\sum_{j=1}^{n}\left|P_{\mathrm{BL}(m),ij}-P^*_{\mathrm{BL}(m),ij}\right|\\
\leqslant&\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\left|\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}_{k_{(m)},s}\right)}-\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}_{k_{(m)},s}\right)}\right|\\
+&\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\left|\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}_{k_{(m)},s}\right)}-\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}^*_{k_{(m)},s}\right)}\right|\\
\leqslant&\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\left|\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}_{k_{(m)},s}\right)-\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)\right|}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}_{k_{(m)},s}\right)}\\
+&\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}_{k_{(m)},s}\right)\left|\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}_{k_{(m)},s}\right)-\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\right|}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}_{k_{(m)},s}\right)}\\
+&\sum_{s=1}^{2^{k_{(m)}}}\sum_{j=1}^{n}\left|\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}_{k_{(m)},s}\right)}-\frac{h_{(m),j}\mathbb{I}\left(\mathbf{x}_i\in\mathbf{t}^*_{k_{(m)},s}\right)\mathbb{I}\left(\mathbf{x}_j\in\mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n}h_{(m),j'}\mathbb{I}\left(\mathbf{x}_{j'}\in\mathbf{t}^*_{k_{(m)},s}\right)}\right|\\
\equiv&\quad\Xi_{C43}+\Xi_{C44}+\Xi_{C45}.
\end{aligned}
\tag{C.11}
$$

Note that $\sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \equiv 1$ and $\sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \equiv 1$ for all $i = 1, \ldots, n$ and $m = 1, \ldots, M_n$. Similar to the proofs of (C.9) and (C.10), it is seen that under Condition 4,

$$\begin{aligned}
\Xi_{C43} =& \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \sum_{j=1}^{n} \frac{h_{(m),i} \left|\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)\right|}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \\
\leqslant & c\delta_n \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \frac{\sum_{j=1}^{n} h_{(m),j} \left|\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right) + \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)\right|}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \\
\leqslant & c\delta_n \left(1 + \frac{\mathscr{N}^*}{\mathscr{n}}\right), \qquad \text{(C.12)}
\end{aligned}$$

almost surely,

$$\begin{aligned}
\Xi_{C44} =& \sum_{s=1}^{2^{k_{(m)}}} \left|\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right)\right| \\
\leqslant & \delta_n \sum_{s=1}^{2^{k_{(m)}}} \left\{\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) + \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right)\right\} \\
= & 2\delta_n, \qquad \text{(C.13)}
\end{aligned}$$

and

$$\begin{aligned}
\Xi_{C45} \leqslant & \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \left\{ \frac{\sum_{j=1}^{n} h_{(m),j} \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}^*_{k_{(m)},s}\right)}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right)} \right. \\
& \left. \times \frac{\sum_{j'=1}^{n} h_{(m),j'} \left|\mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right) - \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)\right|}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \right\} \\
\leqslant & \delta_n \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \frac{\sum_{j'=1}^{n} h_{(m),j'} \left\{\mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}^*_{k_{(m)},s}\right) + \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)\right\}}{\sum_{j'=1}^{n} h_{(m),j'} \mathbb{I}\left(\mathbf{x}_{j'} \in \mathbf{t}_{k_{(m)},s}\right)} \\
\leqslant & \delta_n \left(1 + \frac{\mathscr{N}^*}{\mathscr{n}}\right) \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}^*_{k_{(m)},s}\right) \\
= & \delta_n \left(1 + \frac{\mathscr{N}^*}{\mathscr{n}}\right). \qquad \text{(C.14)}
\end{aligned}$$

Thus, (15) can be verified by combining (C.11) - (C.14) together and this concludes the proof.

C.5 Proof of Theorem 3

In accordance with Section 3.2.2, notations without the $*$ script correspond to the CART-splitting criterion, while those with the $*$ script are related to the theoretical CART-splitting

criterion. It is evident that δ_n is a key factor bounding the difference between matrices $\mathbf{P}_{\mathrm{BL}(m)}$ and $\mathbf{P}_{*\mathrm{BL}(m)}$. It is clear that

$$\Pr\left\{\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\Delta\mathbf{t}^*_{k_{(m)},s}\right)=0\right\}=1-p_{(m),is},$$

from the definition of $p_{(m),is}$. Therefore, when $0 < p_{(m),is} < 1/2$, from Lemma C.4, $\mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\Delta\mathbf{t}^*_{k_{(m)},s}\right)-p_{(m),is}$ is a centered Bernoulli random variable with variance proxy

$$\tau^2\left(p_{(m),is}\right)=\frac{\frac{1}{2}-p_{(m),is}}{\log\left(\frac{1}{p_{(m),is}}-1\right)},$$

where $\tau^2\left(p_{(m),is}\right)$ is a strictly increasing function of $p_{(m),is}$. Then, for $p_{(m),is} \in (0,1/2)$, we have

$$\max_{1\leqslant i\leqslant n,1\leqslant m\leqslant M_n,1\leqslant s\leqslant 2^{k_{(m)}}}\tau^2\left(p_{(m),is}\right)\leqslant\frac{\frac{1}{2}-\bar{p}_n}{\log\left(\frac{1}{\bar{p}_n}-1\right)}\leqslant\frac{\frac{1}{2}}{\log\left(\frac{1}{\bar{p}_n}-1\right)}.$$

By Lemma C.5 (which does not require the assumption of independence), for each $r \geqslant 1$, we have

$$\begin{aligned}\mathbb{E}|\delta_n|^r=\mathbb{E}|\delta_n| &\leqslant \mathbb{E}\left\{\max_{1\leqslant i\leqslant n,1\leqslant m\leqslant M_n,1\leqslant s\leqslant 2^{k_{(m)}}}\left|\mathbb{I}\left(x_i\in\mathbf{t}_{k_{(m)},s}\Delta\mathbf{t}^*_{k_{(m)},s}\right)-p_{(m),is}\right|\right\}+\bar{p}_n\\ &\leqslant\sqrt{\frac{2\log(2nM_n\bar{K}_n)}{\log\left(\frac{1}{\bar{p}_n}-1\right)}}+\bar{p}_n,\end{aligned}\tag{C.15}$$

and this provides the upper bound of $\mathbb{E}|\delta_n|$.

With the above point addressed, we now resume the primary trajectory of the proof. It is sufficient to verify that

$$\sup_{\mathbf{w}\in\mathscr{H}}\frac{|R_n(\mathbf{w})-L_n(\mathbf{w})|}{R_n(\mathbf{w})}=o_p(1),\tag{C.16}$$

and

$$\sup_{\mathbf{w}\in\mathscr{H}}\frac{|C^\circ_n(\mathbf{w})-L_n(\mathbf{w})|}{R_n(\mathbf{w})}=o_p(1).\tag{C.17}$$

We will verify them successively. Note that

$$\begin{aligned}\sup_{\mathbf{w}\in\mathscr{H}}\frac{|R_n(\mathbf{w})-L_n(\mathbf{w})|}{R_n(\mathbf{w})} &\leqslant \sup_{\mathbf{w}\in\mathscr{H}}\frac{|R_n(\mathbf{w})-R^*_n(\mathbf{w})|}{R_n(\mathbf{w})}+\sup_{\mathbf{w}\in\mathscr{H}}\frac{|L^*_n(\mathbf{w})-L_n(\mathbf{w})|}{R_n(\mathbf{w})}\\ &\quad+\sup_{\mathbf{w}\in\mathscr{H}}\frac{|R^*_n(\mathbf{w})-L^*_n(\mathbf{w})|}{R_n(\mathbf{w})},\end{aligned}\tag{C.18}$$

and

$$\sup_{\mathbf{w}\in\mathscr{H}}\frac{|C^\circ_n(\mathbf{w})-L_n(\mathbf{w})|}{R_n(\mathbf{w})} \leqslant \sup_{\mathbf{w}\in\mathscr{H}}\frac{|\{C^\circ_n(\mathbf{w})-L_n(\mathbf{w})\}-\{C^{\circ*}_n(\mathbf{w})-L^*_n(\mathbf{w})\}|}{R_n(\mathbf{w})}$$

$$+\sup_{\mathbf{w}\in\mathscr{H}}\frac{|C_n^{\circ *}(\mathbf{w})-L_n^*(\mathbf{w})|}{R_n(\mathbf{w})}. \tag{C.19}$$

Moreover, we have

$$\begin{aligned}
&|R_n(\mathbf{w})-R_n^*(\mathbf{w})|\\
&\leqslant\Big|\mathbb{E}\left\{\boldsymbol{\mu}^\top\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\mu}+2\boldsymbol{\mu}^\top\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}+\mathbf{e}^\top\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\mid\mathscr{X}\right\}\\
&\quad-\mathbb{E}\left\{\boldsymbol{\mu}^\top\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\boldsymbol{\mu}+2\boldsymbol{\mu}^\top\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\mathbf{e}+\mathbf{e}^\top\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\mathbf{e}\mid\mathscr{X}\right\}\Big|\\
&\quad+2\left|\mathbb{E}\left\{\boldsymbol{\mu}^\top\mathbf{P}(\mathbf{w})\boldsymbol{\mu}+\boldsymbol{\mu}^\top\mathbf{P}(\mathbf{w})\mathbf{e}\mid\mathscr{X}\right\}-\mathbb{E}\left\{\boldsymbol{\mu}^\top\mathbf{P}_*(\mathbf{w})\boldsymbol{\mu}+\boldsymbol{\mu}^\top\mathbf{P}_*(\mathbf{w})\mathbf{e}\mid\mathscr{X}\right\}\right|\\
&\leqslant\left|\mathbb{E}\left[\mathbf{e}^\top\left\{\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})-\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\right\}\mathbf{e}\mid\mathscr{X}\right]\right|\\
&\quad+2\left|\mathbb{E}\left[\boldsymbol{\mu}^\top\left\{\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})-\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\right\}\mathbf{e}\mid\mathscr{X}\right]\right|\\
&\quad+\left|\boldsymbol{\mu}^\top\left\{\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})-\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\right\}\boldsymbol{\mu}\right|\\
&\quad+2\left|\mathbb{E}\left[\boldsymbol{\mu}^\top\left\{\mathbf{P}(\mathbf{w})-\mathbf{P}_*(\mathbf{w})\right\}\mathbf{e}\mid\mathscr{X}\right]\right|+2\left|\boldsymbol{\mu}^\top\left\{\mathbf{P}(\mathbf{w})-\mathbf{P}_*(\mathbf{w})\right\}\boldsymbol{\mu}\right|,
\end{aligned}\tag{C.20}$$

and

$$\begin{aligned}
&|L_n(\mathbf{w})-L_n^*(\mathbf{w})|\\
&\leqslant\Big|\boldsymbol{\mu}^\top\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\mu}+2\boldsymbol{\mu}^\top\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}+\mathbf{e}^\top\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\\
&\quad-\left\{\boldsymbol{\mu}^\top\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\boldsymbol{\mu}+2\boldsymbol{\mu}^\top\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\mathbf{e}+\mathbf{e}^\top\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\mathbf{e}\right\}\Big|\\
&\quad+2\left|\left\{\boldsymbol{\mu}^\top\mathbf{P}(\mathbf{w})\boldsymbol{\mu}+\boldsymbol{\mu}^\top\mathbf{P}(\mathbf{w})\mathbf{e}\right\}-\left\{\boldsymbol{\mu}^\top\mathbf{P}_*(\mathbf{w})\boldsymbol{\mu}+\boldsymbol{\mu}^\top\mathbf{P}_*(\mathbf{w})\mathbf{e}\right\}\right|.
\end{aligned}\tag{C.21}$$

Now, by (14) of Lemma 1, there exists a positive constant c_1 such that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathscr{H}}\left|\mathbf{e}^\top\left\{\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})-\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\right\}\mathbf{e}\right|\\
&\leqslant\sup_{\mathbf{w}\in\mathscr{H}}\sum_{m=1}^{M_n}\sum_{r=1}^{M_n}w_{(m)}w_{(r)}\left|\mathbf{e}^\top\left(\mathbf{P}_{\mathrm{BL}(m)}^\top\mathbf{P}_{\mathrm{BL}(r)}-\mathbf{P}_{*\mathrm{BL}(m)}^\top\mathbf{P}_{*\mathrm{BL}(r)}\right)\mathbf{e}\right|\\
&\leqslant\max_{1\leqslant m,r\leqslant M_n}\left|\mathbf{e}^\top\left(\mathbf{P}_{\mathrm{BL}(m)}^\top\mathbf{P}_{\mathrm{BL}(r)}-\mathbf{P}_{*\mathrm{BL}(m)}^\top\mathbf{P}_{*\mathrm{BL}(r)}\right)\mathbf{e}\right|\\
&\leqslant\max_{1\leqslant m,r\leqslant M_n}\|\mathbf{e}\|_\infty^2\sum_{i=1}^n\sum_{j=1}^n\sum_{t=1}^n\Bigg(\left|P_{\mathrm{BL}(m),ti}P_{\mathrm{BL}(r),tj}-P_{\mathrm{BL}(m),ti}P^*_{\mathrm{BL}(r),tj}\right|\\
&\quad+\left|P_{\mathrm{BL}(m),ti}P^*_{\mathrm{BL}(r),tj}-P^*_{\mathrm{BL}(m),ti}P^*_{\mathrm{BL}(r),tj}\right|\Bigg)\\
&\leqslant\max_{1\leqslant m,r\leqslant M_n}\|\mathbf{e}\|_\infty^2\Bigg(\sum_{j=1}^n\sum_{t=1}^n\left|P_{\mathrm{BL}(r),tj}-P^*_{\mathrm{BL}(r),tj}\right|\sum_{i=1}^n P_{\mathrm{BL}(m),ti}\\
&\quad+\sum_{i=1}^n\sum_{t=1}^n\left|P_{\mathrm{BL}(m),ti}-P^*_{\mathrm{BL}(m),ti}\right|\sum_{j=1}^n P^*_{\mathrm{BL}(r),tj}\Bigg)
\end{aligned}$$

$$\leqslant c_1\|\mathbf{e}\|_\infty^2\bar{K}_n(\mathscr{N}+\mathscr{N}^*)\delta_n$$
$$\leqslant c_1\|\mathbf{e}\|_\infty^2\bar{K}_n\bar{\mathscr{N}}_n\delta_n, \tag{C.22}$$

almost surely, where the last second step is from the fact that

$$\sum_{i=1}^n P_{\mathrm{BL}(m),ti}=\sum_{i=1}^n P^*_{\mathrm{BL}(m),ti}\equiv 1,$$

for all $m=1,\ldots,M_n$ and $t=1,\ldots,n$, and the last step comes from Condition 6. Similarly, one has

$$\sup_{\mathbf{w}\in\mathscr{H}}\left|\boldsymbol{\mu}^\top\left\{\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})-\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\right\}\mathbf{e}\right|\leqslant c_1\|\boldsymbol{\mu}\|_\infty\|\mathbf{e}\|_\infty\bar{K}_n\bar{\mathscr{N}}_n\delta_n, \tag{C.23}$$

almost surely, and

$$\sup_{\mathbf{w}\in\mathscr{H}}\left|\boldsymbol{\mu}^\top\left\{\mathbf{P}^\top(\mathbf{w})\mathbf{P}(\mathbf{w})-\mathbf{P}_*^\top(\mathbf{w})\mathbf{P}_*(\mathbf{w})\right\}\boldsymbol{\mu}\right|\leqslant c_1\|\boldsymbol{\mu}\|_\infty^2\bar{K}_n\bar{\mathscr{N}}_n\delta_n, \tag{C.24}$$

almost surely. Besides, we have

$$\begin{aligned}\sup_{\mathbf{w}\in\mathscr{H}}\left|\boldsymbol{\mu}^\top\left\{\mathbf{P}(\mathbf{w})-\mathbf{P}_*(\mathbf{w})\right\}\mathbf{e}\right|&\leqslant\sup_{\mathbf{w}\in\mathscr{H}}\sum_{m=1}^{M_n}w_{(m)}\left|\boldsymbol{\mu}^\top\left(\mathbf{P}_{\mathrm{BL}(m)}-\mathbf{P}_{*\mathrm{BL}(m)}\right)\mathbf{e}\right|\\&\leqslant\max_{1\leqslant m\leqslant M_n}\left|\boldsymbol{\mu}^\top\left(\mathbf{P}_{\mathrm{BL}(m)}-\mathbf{P}_{*\mathrm{BL}(m)}\right)\mathbf{e}\right|\\&\leqslant\max_{1\leqslant m\leqslant M_n}\|\boldsymbol{\mu}\|_\infty\|\mathbf{e}\|_\infty\sum_{i=1}^n\sum_{j=1}^n\left|P_{\mathrm{BL}(m),ij}-P^*_{\mathrm{BL}(m),ij}\right|\\&\leqslant c_1\|\boldsymbol{\mu}\|_\infty\|\mathbf{e}\|_\infty\bar{K}_n\bar{\mathscr{N}}_n\delta_n,\end{aligned} \tag{C.25}$$

almost surely, and

$$\sup_{\mathbf{w}\in\mathscr{H}}\left|\boldsymbol{\mu}^\top\left\{\mathbf{P}(\mathbf{w})-\mathbf{P}_*(\mathbf{w})\right\}\boldsymbol{\mu}\right|\leqslant c_1\|\boldsymbol{\mu}\|_\infty^2\bar{K}_n\bar{\mathscr{N}}_n\delta_n, \tag{C.26}$$

almost surely. Then, under Conditions 2′, 4, 6, and 7, by combining (C.20), (C.22) and (C.23) - (C.26) together, it is readily seen that there exists a positive constant c_3 such that

$$\begin{aligned}&\mathbb{E}\left\{\frac{\sup_{\mathbf{w}\in\mathscr{H}}|R_n(\mathbf{w})-R_n^*(\mathbf{w})|}{\xi_n}\right\}\\&\leqslant c_3\bar{K}_n\bar{\mathscr{N}}_n\mathbb{E}\left\{\frac{\delta_n(1+\|\mathbf{e}\|_\infty+\|\mathbf{e}\|_\infty^2)}{\xi_n}\right\}\\&\leqslant c_3\bar{K}_n\bar{\mathscr{N}}_n\left\{\mathbb{E}\left(\frac{\delta_n}{\xi_n}\right)+\mathbb{E}^{1/2}\|\mathbf{e}\|_\infty^2\mathbb{E}^{1/2}\left(\frac{\delta_n}{\xi_n^2}\right)+\mathbb{E}^{1/2}\|\mathbf{e}\|_\infty^4\mathbb{E}^{1/2}\left(\frac{\delta_n}{\xi_n^2}\right)\right\}\\&\leqslant 3c_3\mathbb{E}^{1/2}\left\{\frac{\bar{K}_n^2\bar{\mathscr{N}}_n^2\log^4(n)\delta_n}{\xi_n^2}\right\}\\&\leqslant 3c_3\mathbb{E}^{1/4}(\delta_n)\mathbb{E}^{1/4}\left\{\frac{\bar{K}_n^4\bar{\mathscr{N}}_n^4\log^8(n)}{\xi_n^4}\right\}\end{aligned}$$

$$\leqslant 3c_3 \left\{ \sqrt{\frac{2\log(2nM_n\bar{K}_n)}{\log\left(\frac{1}{\bar{p}_n}-1\right)}} + \bar{p}_n \right\}^{1/4} \mathbb{E}^{1/4}\left\{ \frac{\bar{K}_n^4 \bar{\mathscr{N}}_n^4 \log^8(n)}{\xi_n^4} \right\}$$
$$= o(1), \tag{C.27}$$

almost surely. Here, several facts contribute to obtaining (C.27). The second inequality arises from the Cauchy-Schwartz Inequality. The third is owing to the Maximal Inequality with Bernstein's moment conditions (Zhang and Chen, 2021, Proposition 7.1), along with the fact that $\mathbb{E}^{1/r}|Z|^r$ is a non-decreasing function of r for $r > 0$ and any generic random variable Z. The penultimate step results from (C.15). Finally, the last equality is from Condition 7 combined with the Lebesgue's Dominated Convergence Theorem. By (C.21), under the same conditions and framework that derives (C.27), we have

$$\mathbb{E}\left\{ \frac{\sup_{\mathbf{w}\in\mathscr{H}} |L_n(\mathbf{w}) - L_n^*(\mathbf{w})|}{\xi_n} \right\} = o(1), \tag{C.28}$$

almost surely. Besides, since the theoretical CART-splitting criterion is independent of response values, it can be considered as a special type of splitting criterion used by SUT trees. Therefore, under Conditions 1′ - 3′ and 4, by combining Theorem 1, (C.27) and the proof of Theorem 1 by Qiu et al. (2020), we have

$$\sup_{\mathbf{w}\in\mathscr{H}} \frac{|R_n^*(\mathbf{w}) - L_n^*(\mathbf{w})|}{R_n(\mathbf{w})} \leqslant \sup_{\mathbf{w}\in\mathscr{H}} \frac{|R_n^*(\mathbf{w}) - L_n^*(\mathbf{w})|}{R_n^*(\mathbf{w})} \left\{ 1 + \sup_{\mathbf{w}\in\mathscr{H}} \frac{|R_n(\mathbf{w}) - R_n^*(\mathbf{w})|}{R_n(\mathbf{w})} \right\}$$
$$= o_p(1). \tag{C.29}$$

Then, (C.16) can be verified by combining this with (C.18), (C.20), (C.21), (C.27) and (C.28) together.

Additionally,

$$\begin{aligned}
&|\{C_n^\circ(\mathbf{w}) - L_n(\mathbf{w})\} - \{C_n^{\circ *}(\mathbf{w}) - L_n^*(\mathbf{w})\}| \\
&= \Big| 2\mathbf{e}^\top \{\mathbf{P}_*(\mathbf{w}) - \mathbf{P}(\mathbf{w})\}\boldsymbol{\mu} - 2\mathbf{e}^\top \{\mathbf{P}_*(\mathbf{w}) - \mathbf{P}(\mathbf{w})\}\mathbf{e} \\
&\quad + 2\widehat{\sigma}^2 \operatorname{trace}\{\mathbf{P}(\mathbf{w})\} - 2\widehat{\sigma}_*^2 \operatorname{trace}\{\mathbf{P}_*(\mathbf{w})\} \Big| \\
&\leqslant 2\left| \mathbf{e}^\top \{\mathbf{P}_*(\mathbf{w}) - \mathbf{P}(\mathbf{w})\}\boldsymbol{\mu} \right| + 2\left| \mathbf{e}^\top \{\mathbf{P}_*(\mathbf{w}) - \mathbf{P}(\mathbf{w})\}\mathbf{e} \right| \\
&\quad + 2\widehat{\sigma}^2 \Big| \operatorname{trace}\{\mathbf{P}(\mathbf{w}) - \mathbf{P}_*(\mathbf{w})\} \Big| + 2\left|\widehat{\sigma}^2 - \widehat{\sigma}_*^2\right| \cdot \Big| \operatorname{trace}\{\mathbf{P}_*(\mathbf{w})\} \Big|.
\end{aligned} \tag{C.30}$$

Under the same framework that establishes (C.25), one has

$$\sup_{\mathbf{w}\in\mathscr{H}} \left| \mathbf{e}^\top \{\mathbf{P}(\mathbf{w}) - \mathbf{P}_*(\mathbf{w})\} \mathbf{e} \right| \leqslant c_1 \|\mathbf{e}\|_\infty^2 \bar{K}_n \bar{\mathscr{N}}_n \delta_n. \tag{C.31}$$

In the light of (15) in Lemma 1 and Condition 6, there exists a positive constant c_2 such that

$$\sup_{\mathbf{w}\in\mathscr{H}} |\operatorname{trace}\{\mathbf{P}(\mathbf{w}) - \mathbf{P}_*(\mathbf{w})\}| \leqslant \max_{1\leqslant m\leqslant M_n} \sum_{i=1}^{n} \left| P_{\mathrm{BL}(m),ii} - P^*_{\mathrm{BL}(m),ii} \right|$$

$$\leqslant \delta_n\{2 + c_2(1 + \bar{r}_n)\}, \tag{C.32}$$

almost surely. As analogues of (C.2) and (C.3) in the context of CART trees, under Condition 4, there exists a positive constant c such that

$$\left\|\mathbf{P}_{\mathrm{BL}(m)}\right\|_\infty = 1,$$

and

$$\left\|\mathbf{P}_{\mathrm{BL}(m)}\right\|_1 \leqslant c,$$

almost surely, for all $m = 1, \ldots, M_n$. Then, we have

$$\|\mathbf{P}(\mathbf{w}_0)\| = \left\|\sum_{m=1}^{M_n} \frac{1}{M_n}\mathbf{P}_{\mathrm{BL}(m)}\right\| \leqslant \max_{1\leqslant m\leqslant M_n} \left\|\mathbf{P}_{\mathrm{BL}(m)}\right\| \leqslant \max_{1\leqslant m\leqslant M_n} \left\|\mathbf{P}_{\mathrm{BL}(m)}\right\|_1 \left\|\mathbf{P}_{\mathrm{BL}(m)}\right\|_\infty \leqslant c, \tag{C.33}$$

almost surely. Therefore, it is seen that

$$\begin{aligned}
\mathbb{E}^{1/2}\left|\widehat{\sigma}^2\right|^2 &= \mathbb{E}^{1/2}\left\{\frac{\|\mathbf{y} - \mathbf{P}(\mathbf{w}_0)\mathbf{y}\|^2}{n}\right\}^2 \\
&\leqslant 2\mathbb{E}^{1/2}\left\{\frac{\|\mathbf{y}\|^2 + \|\mathbf{P}(\mathbf{w}_0)\|^2\|\mathbf{y}\|^2}{n}\right\}^2 \\
&\leqslant 2(1 + c^2)\mathbb{E}^{1/2}\left(\frac{\|\mathbf{y}\|^2}{n}\right)^2 \\
&= O(1),
\end{aligned} \tag{C.34}$$

almost surely, where the third step comes from (C.33), and the last step is from Condition $2'$. Likewise, we have

$$\begin{aligned}
\mathbb{E}^{1/2}\left|\widehat{\sigma}^2 - \widehat{\sigma}_*^2\right|^2 &\leqslant \mathbb{E}^{1/2}\left\{\frac{\|\mathbf{y} - \mathbf{P}(\mathbf{w}_0)\mathbf{y}\|^2}{n}\right\}^2 + \mathbb{E}^{1/2}\left\{\frac{\|\mathbf{y} - \mathbf{P}_*(\mathbf{w}_0)\mathbf{y}\|^2}{n}\right\}^2 \\
&= O(1),
\end{aligned} \tag{C.35}$$

almost surely. By combining this with (C.25), (C.31), (C.32), (C.34) and (C.35) with (C.30), we have

$$\mathbb{E}\left[\frac{\sup_{\mathbf{w}\in\mathscr{K}} |\{C_n^\circ(\mathbf{w}) - L_n(\mathbf{w})\} - \{C_n^{\circ *}(\mathbf{w}) - L_n^*(\mathbf{w})\}|}{\xi_n}\right] = o(1). \tag{C.36}$$

Similar to (C.29), under Conditions $1'$ - $3'$ and 4, by combining Theorem 1, (C.27) and the proof of Theorem 1 in Qiu et al. (2020), we have

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathscr{K}} \frac{|C_n^{\circ *}(\mathbf{w}) - L_n^*(\mathbf{w})|}{R_n(\mathbf{w})} &\leqslant \sup_{\mathbf{w}\in\mathscr{K}} \frac{|C_n^{\circ *}(\mathbf{w}) - L_n^*(\mathbf{w})|}{R_n^*(\mathbf{w})}\left\{1 + \sup_{\mathbf{w}\in\mathscr{K}} \frac{|R_n(\mathbf{w}) - R_n^*(\mathbf{w})|}{R_n(\mathbf{w})}\right\} \\
&= o_p(1).
\end{aligned} \tag{C.37}$$

By combining (C.36), (C.37) with (C.19), we obtain (C.17) and this concludes the proof.

C.6 Verifying Results of Qiu et al. (2020) for Stochastic X

For the sake of convenience, Qiu et al. (2020) assume in all proofs that $\mathbf{X}$ is non-stochastic rather than stochastic. However, the framework developed in Qiu et al. (2020) can readily be extended for stochastic $\mathbf{X}$. We take

$$\sup_{\mathbf{w}\in\mathscr{H}} \left|\mathbf{e}^\top \mathbf{A}_\star(\mathbf{w})\boldsymbol{\mu}\right| / R_n(\mathbf{w}) = o_p(1)$$

for example, where $\mathbf{A}_\star(\mathbf{w}) = \mathbf{I}_n - \mathbf{P}_\star(\mathbf{w})$. In Appendix A.2 of Qiu et al. (2020), this result is labeled as (A6). Note that the framework proposed by Qiu et al. (2020) necessitates the independence of the "hat matrix" from response values, warranting notations in this section to carry the script $\star$.

Proof of (A6) in Qiu et al. (2020) when X is Stochastic. Let $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, $\mathbf{A}_{\star(m)} = \mathbf{I}_n - \mathbf{P}_{\star\mathrm{BL}(m)}$ for all $m = 1, \ldots, M_n$. Let $\boldsymbol{\Phi} = \left(\boldsymbol{\mu}^\top \mathbf{A}_{\star(m)}^\top \mathbf{A}_{\star(s)}\boldsymbol{\mu}\right)_{M_n\times M_n}$, that is, the $(m,s)^{\text{th}}$ component of $\boldsymbol{\Phi}$ is $\boldsymbol{\mu}^\top \mathbf{A}_{\star(m)}^\top \mathbf{A}_{\star(s)}\boldsymbol{\mu}$, $\mathbf{G}_{n\times M_n} = \left(\mathbf{A}_{\star(1)}\boldsymbol{\mu}, \ldots, \mathbf{A}_{\star(M_n)}\boldsymbol{\mu}\right)$, $\boldsymbol{\Psi} = \left\{\mathrm{trace}\left(\mathbf{P}_{\star\mathrm{BL}(m)}\mathbf{P}_{\star\mathrm{BL}(s)}^\top\boldsymbol{\Omega}\right)\right\}_{M_n\times M_n}$, and

$$\boldsymbol{\Psi}_0 = \mathrm{diag}\left\{\mathrm{trace}\left(\mathbf{P}_{\star\mathrm{BL}(1)}\mathbf{P}_{\star\mathrm{BL}(1)}^\top\boldsymbol{\Omega}\right), \ldots, \mathrm{trace}\left(\mathbf{P}_{\star\mathrm{BL}(M_n)}\mathbf{P}_{\star\mathrm{BL}(M_n)}^\top\boldsymbol{\Omega}\right)\right\}.$$

So $\boldsymbol{\Phi} = \mathbf{G}^\top\mathbf{G}$. For any $\mathbf{w} \in \mathscr{H}$,

$$\mathbf{w}^\top\boldsymbol{\Psi}_0\mathbf{w} \leqslant \mathbf{w}^\top\boldsymbol{\Psi}\mathbf{w}, \tag{C.38}$$

because for any $m, s \in \{1, \ldots, M_n\}, w_{(m)} \geqslant 0, w_{(s)} \geqslant 0$ and $\mathrm{trace}\left(\mathbf{P}_{\star\mathrm{BL}(m)}\mathbf{P}_{\star\mathrm{BL}(s)}^\top\boldsymbol{\Omega}\right) \geqslant 0$ by Condition C.4. In addition, it is clear that

$$\begin{aligned} R_n^\star(\mathbf{w}) &= \mathbb{E}\left\{\|\mathbf{P}_\star(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu} + \mathbf{P}_\star(\mathbf{w})\mathbf{e}\|^2 \middle| \mathscr{X}\right\} \\ &= \|\mathbf{A}_\star(\mathbf{w})\boldsymbol{\mu}\|^2 + \mathrm{trace}\left\{\mathbf{P}_\star(\mathbf{w})\mathbf{P}_\star^\top(\mathbf{w})\boldsymbol{\Omega}\right\} \\ &= \mathbf{w}^\top(\boldsymbol{\Phi} + \boldsymbol{\Psi})\mathbf{w} \\ &\geqslant \mathbf{w}^\top\left(\boldsymbol{\Phi} + \boldsymbol{\Psi}_0\right)\mathbf{w}, \end{aligned} \tag{C.39}$$

where the last step is from (C.38). We also have

$$\boldsymbol{\Phi} + \boldsymbol{\Psi}_0 > 0, \tag{C.40}$$

because $\boldsymbol{\Phi} = \mathbf{G}^\top\mathbf{G}$ and $\boldsymbol{\Psi}_0 > 0$ by definition.

Let $\boldsymbol{\rho} = \left(\mathbf{e}^\top\mathbf{A}_{\star(1)}\boldsymbol{\mu} \ldots, \mathbf{e}^\top\mathbf{A}_{\star(M_n)}\boldsymbol{\mu}\right)^\top$. It is straightforward to show that

$$\mathbb{E}(\boldsymbol{\rho}|\mathscr{X}) = 0. \tag{C.41}$$

Besides, under Condition 2, there exists a positive constant v such that

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\rho}|\mathscr{X}) &= \mathbb{E}\left(\boldsymbol{\rho}\boldsymbol{\rho}^\top \middle| \mathscr{X}\right) \\ &= \mathbb{E}\left\{\left(\mathbf{e}^\top\mathbf{A}_{\star(m)}\boldsymbol{\mu}\boldsymbol{\mu}^\top\mathbf{A}_{\star(s)}^\top\mathbf{e}\right)_{M_n\times M_n} \middle| \mathscr{X}\right\} \end{aligned}$$

$$
\begin{aligned}
&= \left(\boldsymbol{\mu}^\top \mathbf{A}_{\star(s)}^\top \boldsymbol{\Omega} \mathbf{A}_{\star(m)} \boldsymbol{\mu}\right)_{M_n \times M_n} \\
&\leqslant v\boldsymbol{\Phi}, \qquad \text{(C.42)}
\end{aligned}
$$

almost surely. It is seen that

$$
\begin{aligned}
\sup_{\mathbf{w}\in\mathscr{H}} \frac{\left\{\mathbf{e}^\top \mathbf{A}_\star(\mathbf{w})\boldsymbol{\mu}\right\}^2}{R_n^{\star 2}(\mathbf{w})} &= \sup_{\mathbf{w}\in\mathscr{H}} \frac{\left(\sum_{m=1}^{M_n} \mathbf{w}_{(m)} \mathbf{e}^\top \mathbf{A}_{\star(m)} \boldsymbol{\mu}\right)^2}{R_n^{\star 2}(\mathbf{w})} \\
&= \sup_{\mathbf{w}\in\mathscr{H}} \frac{\left(\mathbf{w}^\top \boldsymbol{\rho}\right)^2}{R_n^{\star 2}(\mathbf{w})} \\
&\leqslant \sup_{\mathbf{w}\in\mathscr{H}} \frac{\left(\mathbf{w}^\top \boldsymbol{\rho}\right)^2}{\mathbf{w}^\top \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)\mathbf{w}} \sup_{\mathbf{w}\in\mathscr{H}} \frac{1}{R_n^\star(\mathbf{w})} \\
&\leqslant \xi_{\star n}^{-1} \boldsymbol{\rho}^\top \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\rho}, \qquad \text{(C.43)}
\end{aligned}
$$

where the third step is from (C.39), and the last step is from (C.40) and Lemma 1 in Qiu et al. (2020). By Markov Inequality, we have that for any $\delta > 0$,

$$
\begin{aligned}
&\Pr\left\{\xi_{\star n}^{-1} \boldsymbol{\rho}^\top \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\rho} > \delta\right\} \\
&\leqslant \delta^{-1} \mathbb{E}\left\{\xi_{\star n}^{-1} \boldsymbol{\rho}^\top \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\rho}\right\} \\
&= \delta^{-1} \mathbb{E}\left[\mathbb{E}\left\{\xi_{\star n}^{-1} \boldsymbol{\rho}^\top \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\rho} \middle| \mathscr{X}\right\}\right] \\
&= \delta^{-1} \mathbb{E}\left[\xi_{\star n}^{-1} \mathbb{E}\left\{\boldsymbol{\rho}^\top \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\rho} \middle| \mathscr{X}\right\}\right] \\
&= v\delta^{-1} \mathbb{E}\left[\xi_{\star n}^{-1} \operatorname{trace}\left\{\left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\Phi}\right\}\right] \\
&\leqslant v\delta^{-1} \mathbb{E}\left[\xi_{\star n}^{-1} \operatorname{trace}\left\{\left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\Phi} + \boldsymbol{\Psi}_0^{1/2} \left(\boldsymbol{\Phi}+\boldsymbol{\Psi}_0\right)^{-1} \boldsymbol{\Psi}_0^{1/2}\right\}\right] \\
&= v\delta^{-1} \mathbb{E}\left(\xi_{\star n}^{-1} M_n\right), \qquad \text{(C.44)}
\end{aligned}
$$

where the second step follows from the Law of Iterated (or Total) Expectation, the third step is obtained by the Pull-out rule, and the fourth step is guaranteed by (C.41) and (C.42). Combining (C.43), (C.44) and Condition 1, we obtain similar result in (A6) of Qiu et al. (2020) by the Lebesgue's Dominated Convergence Theorem. ■

This demonstration bears a striking resemblance to Proof of (A6) in Appendix A.2 of Qiu et al. (2020). The only modification lies in the substitution of expectations with conditional expectations in (C.41) and (C.42), as well as the applications of the Law of Iterated Expectation, Pull-out rule, and Lebesgue Dominated Convergence Theorem in (C.44). Similarly, (A7) - (A9) and (A34) - (A35) for proving Theorems 1 and 2 in Qiu et al. (2020) can also be extrapolated using the same techniques.

C.7 Some Additional Discussions on the Behavior of Risk Function When Relevant/Important Features Are Not Involved in the Model

In many practical situations, investigators often cannot include all the relevant features in the model (Flynn et al., 2013), which leads to unignorable misspecification. In the current

section, we will adopt a simplified setting to demonstrate the impact of ignoring important or relevant features in weighted random forest.

For simplicity, we use the full sample to grow random forest and only theoretical CART is considered (similar setting has also been considered in Chi et al. (2022)). In this case, the cells are only constructed based on $\mathbf{\Theta} = \{\mathbf{\Theta}_{(1)}, \ldots, \mathbf{\Theta}_{(M_n)}\}$, and $\mathscr{X}$ reduces to the σ-algebra generated by $\{\mathbf{x}_1^0, \ldots, \mathbf{x}_n^0, \mathbf{\Theta}\}$. Now we formally introduce the notion of "ignoring relevant/important features" in our context. Inspired by the Definition 1 in Chi et al. (2022), we say our variable pool misses relevant/important features when

$$\mathbb{E}\left|\mu(\mathbf{x}_1^0) - \mathbb{E}\left\{\mu(\mathbf{x}_1^0) \mid \mathbf{x}_1, \mathbf{\Theta}\right\}\right|^2 \geqslant c_0 > 0 \tag{C.45}$$

for some positive constant c_0, where $\mu(\mathbf{x}_1^0) = \mu_1$. In this scenario, we will demonstrate that the conditional risk is asymptotically bounded below by a positive constant. To see this, assume that $\min_{1\leqslant m\leqslant M_n} \mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\} \geqslant c_0$. Then, for a given tree, we have

$$\begin{aligned}
&\mu(\mathbf{x}_i^0) - \sum_{m=1}^{M_n} w_{(m)}\widehat{\mu}_m(\mathbf{x}_i) \\
&= \sum_{m=1}^{M_n} w_{(m)}\{\mu(\mathbf{x}_i^0) - \widehat{\mu}_m(\mathbf{x}_i)\} \\
&= \sum_{m=1}^{M_n} w_{(m)} \left\{\mu(\mathbf{x}_i^0) - \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \frac{\sum_{j=1}^n y_j \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}\right\} \\
&= \sum_{m=1}^{M_n} w_{(m)} \left[\mu(\mathbf{x}_i^0) - \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \frac{\sum_{j=1}^n \left\{\mu(\mathbf{x}_j^0) + \varepsilon_j\right\} \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}\right] \\
&= \sum_{m=1}^{M_n} w_{(m)} \left[\mu(\mathbf{x}_i^0) - \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0) \mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}\right] \\
&+ \sum_{m=1}^{M_n} w_{(m)} \left(\sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \right. \\
&\quad \left. \times \left[\frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0) \mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}} - \frac{\frac{\sum_{j=1}^n \mu(\mathbf{x}_j^0) \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}}{\frac{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}}\right]\right) \\
&- \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \frac{\sum_{j=1}^n \varepsilon_j \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)} \\
&\equiv \sum_{m=1}^{M_n} w_{(m)} \left[\mu(\mathbf{x}_i^0) - \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0) \mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}\right] + r_{1,i} + r_{2,i}.
\end{aligned} \tag{C.46}$$

Now we aim to bound $r_{1,i}$ and $r_{2,i}$. In specific, if we assume that

$$\boldsymbol{n} \geqslant \bar{\boldsymbol{n}}_n,$$

and

$$\max_{1 \leqslant m \leqslant M_n} 2^{k_{(m)}} \leqslant \bar{K}_n,$$

where $\bar{\boldsymbol{n}}_n$ and $\bar{K}_n$ are deterministic positive series that grow to infinity as $n \to \infty$, and c_0 is a positive constant. Then, by Condition 2 and the fact that $\mathbf{x}_j^0$'s are independent conditionally on $\boldsymbol{\Theta}$, uniformly for every $i = 1, \ldots, n$, we have

$$\begin{aligned}
&\mathbb{E}^{1/2}|r_{1,i}|^2 \\
&\leqslant \sum_{m=1}^{M_n} w_{(m)} \\
&\quad \times \sum_{s=1}^{2^{k_{(m)}}} \mathbb{E}^{1/2} \left| \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \left[\frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}}{\mathbb{E}\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\}} - \frac{\frac{\sum_{j=1}^n \mu(\mathbf{x}_j^0)\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}}{\frac{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}} \right] \right|^2 \\
&\leqslant \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{E}^{1/2} \left| \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}} - \frac{\frac{\sum_{j=1}^n \mu(\mathbf{x}_j^0)\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}}{\frac{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}} \right|^2 \\
&\leqslant \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{E}^{1/2} \left| \frac{\frac{\sum_{j=1}^n \mu(\mathbf{x}_j^0)\mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n} - \mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}}{\frac{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}} \right|^2 \\
&\quad + \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{E}^{1/2} \left| \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}}{\frac{\sum_{j=1}^n \mathbb{I}\left(\mathbf{x}_j \in \mathbf{t}_{k_{(m)},s}\right)}{n}} - \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}}{\mathbb{E}\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\}} \right|^2 \\
&= O\left(\frac{n^{1/2}\bar{K}_n}{\bar{\boldsymbol{n}}_n}\right),
\end{aligned} \tag{C.47}$$

where the last inequality is based on Chebyshev's inequality. Similarly, under Condition 2, uniformly for every $i = 1, \ldots, n$, we also have

$$\mathbb{E}^{1/2}|r_{2,i}|^2 = O\left(\frac{n^{1/2}\bar{K}_n}{\bar{\boldsymbol{n}}_n}\right). \tag{C.48}$$

Combine (C.47) and (C.48) with (C.46), it is readily seen that

$$\begin{aligned}
&\frac{\sum_{i=1}^n \mathbb{E}\left\{|\mu(\mathbf{x}_i^0) - \sum_{m=1}^{M_n} w_{(m)}\widehat{\mu}_m(\mathbf{x}_i)|^2 \mid \mathscr{X}\right\}}{n} \\
&= \frac{\sum_{i=1}^n \left| \mu(\mathbf{x}_i^0) - \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_i \in \mathbf{t}_{k_{(m)},s}\right) \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \boldsymbol{\Theta}\right\}} \right|^2}{n}
\end{aligned}$$

$$
\begin{aligned}
&+ O_p\left(\frac{n^{1/2}\bar{K}_n}{\bar{\pi}_n}\right) \\
&= \mathbb{E}_{\mathbf{x}_{1,c}|\mathbf{x}_1,\mathbf{\Theta}}\left|\mu(\mathbf{x}_1^0) - \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}\right|^2 \\
&\quad + O_p\left(\frac{1}{n^{1/2}} + \frac{n^{1/2}\bar{K}_n}{\bar{\pi}_n}\right), \qquad \text{(C.49)}
\end{aligned}
$$

where $\mathbf{x}_{1,c}$ represents the vector containing the features that are not included in the model, and the $\mathbb{E}_{\mathbf{x}_{1,c}|\mathbf{x}_1,\mathbf{\Theta}}(\cdot)$ is taken with respect to those missing features, conditional on $\mathbf{x}_1$ and $\mathbf{\Theta}$. Then, by the Projection Theorem, it is readily seen that the leading term on the right-hand-side of the last line of (C.49) satisfies that

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}_{1,c}|\mathbf{x}_1,\mathbf{\Theta}}\left|\mu(\mathbf{x}_1^0) - \sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}\right|^2 \\
&\geqslant \mathbb{E}_{\mathbf{x}_{1,c}|\mathbf{x}_1,\mathbf{\Theta}}\left|\mu(\mathbf{x}_1^0) - \mathbb{E}\left\{\mu(\mathbf{x}_1^0) \mid \mathbf{x}_1, \mathbf{\Theta}\right\}\right|^2 \\
&\geqslant c_0 > 0,
\end{aligned}
$$

where we have used the fact that

$$
\sum_{m=1}^{M_n} w_{(m)} \sum_{s=1}^{2^{k_{(m)}}} \mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \frac{\mathbb{E}\left\{\mu(\mathbf{x}_1^0)\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}{\mathbb{E}\left\{\mathbb{I}\left(\mathbf{x}_1 \in \mathbf{t}_{k_{(m)},s}\right) \mid \mathbf{\Theta}\right\}}
$$

is $\sigma(\mathbf{x}_1, \mathbf{\Theta})$-measurable with $\sigma(\mathbf{x}_1, \mathbf{\Theta})$ being the σ-algebra generated by $\{\mathbf{x}_1, \mathbf{\Theta}\}$. This yields that if we further assume that $n^{1/2}\bar{K}_n/\bar{\pi}_n = o(1)$, one has

$$
\inf_{\mathbf{w}\in\mathscr{H}} \frac{1}{\sum_{i=1}^n \mathbb{E}\left\{|\mu(\mathbf{x}_i^0) - \sum_{m=1}^{M_n} w_{(m)}\widehat{\mu}_m(\mathbf{x}_i)|^2 \mid \mathscr{X}\right\}} = O_p\left(\frac{1}{n}\right).
$$

These discussions indicate that if some relevant/important features are missing in the model, it is expected that ξ_n^{-1} achieves a satisfactory rate of convergence.

Appendix Appendix D. Additional Tables and Figures

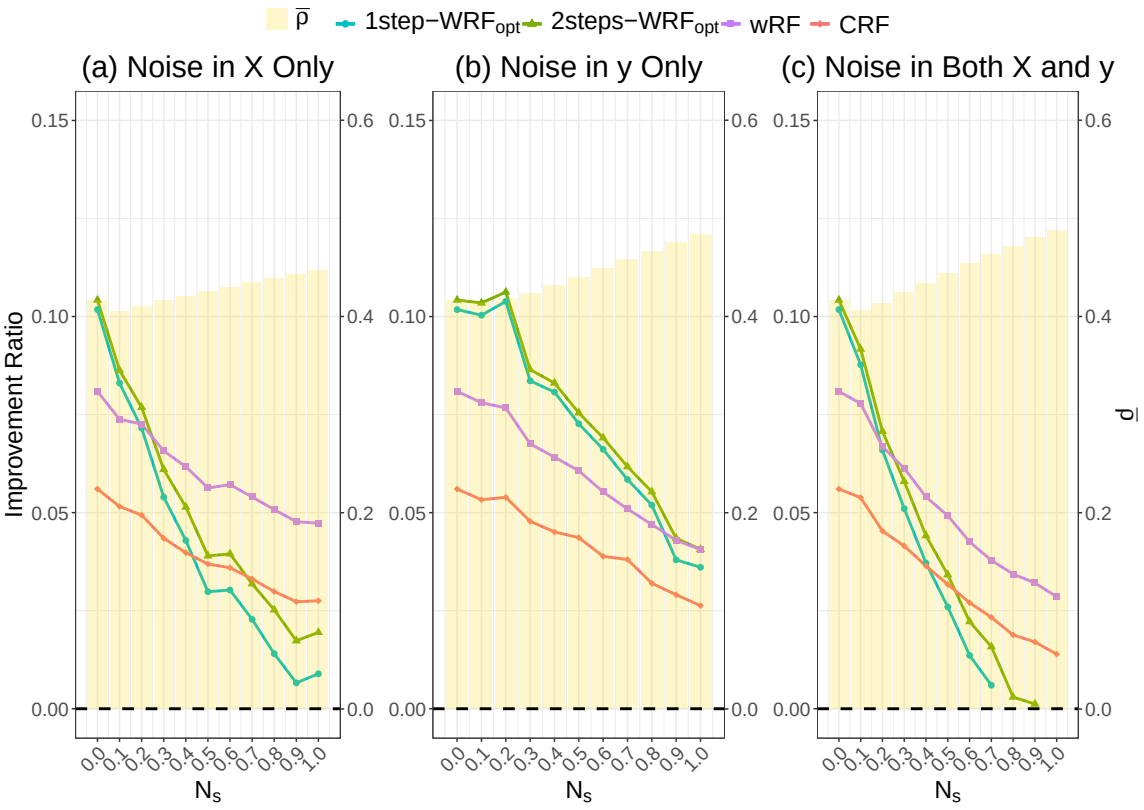

Figure D.1: Improvement Ratio vs Noise on BH Data Set

Algorithm A.2: SUT

Split_a_node(S)

Input: The local learning subset S corresponding to the node we want to split

Output: A split $[a < c]$ or nothing

-If **Stop_split**(S) is TRUE then return nothing.

-Otherwise select q attributes $\{a_1, \ldots, a_q\}$ by probability sequence $\mathscr{P}$ among all non constant (in S) candidate attributes ; `// Hyper parameter: probability sequence` $\mathscr{P} = \{P_1, \ldots, P_p\}$`, where` $P_j \in [0,1], \forall j = 1, \ldots, p$ `and` $\sum_{j=1}^{p} P_j = 1$

-Draw q splits $\{s_1, \ldots, s_q\}$, where $s_i = \textbf{Pick\_a\_split}(S, a_i), \forall i = 1, \ldots, q$;

-Return a split s_* such that $\textbf{Score}(s_*, S) = \max_{i=1,\ldots,q} \textbf{Score}(s_i, S)$.

Pick_a_split(S, a)

Input: A subset S and an attribute a

Output: A split

- Let $a^S_{\max}$ and $a^S_{\min}$ be the maximal and minimal value of a in S;

- Calculate the cut-point $c \leftarrow (a^S_{\min}, a^S_{\max})/2$;

- Return the split $[a < c]$.

Stop_split(S)

Input: A subset S

Output: A boolean

- If $|S| <$ `nodesize`, then return TRUE.

- If all attributes are constant in S, then return TRUE.

- If the output is constant in S, then return TRUE.

- Otherwise, return FALSE.

Score(s, S)

Input: A split s and a subset S

Output: The score of this split method

-Let $\mathbf{X}_P, \mathbf{X}_L, \mathbf{X}_R$ be the attribute matrix of this local parent node, left daughter, right daughter, respectively;

-Let n_P, n_L, n_R be the number of samples contained in the local parent node, left daughter, right daughter, respectively;

-Obtain $\tilde{\mathbf{X}}_P, \tilde{\mathbf{X}}_L, \tilde{\mathbf{X}}_R$ by centering and scaling of each column of the matrices $\mathbf{X}_P, \mathbf{X}_L, \mathbf{X}_R$, respectively;

-score $\leftarrow \frac{\|\tilde{\mathbf{X}}_P\| - \frac{n_L}{n_P}\|\tilde{\mathbf{X}}_L\| - \frac{n_R}{n_P}\|\tilde{\mathbf{X}}_R\|}{\|\tilde{\mathbf{X}}_P\|}$;

-Return score.

Algorithm B.1: 2steps-WRF$_{\text{opt}}$

Input: (1) The training data set $\mathscr{D} = \{y_i, \mathbf{x}_i\}_{i=1}^n$ (2) The number of trees in random forest M_n

Output: Weight vector $\widetilde{\mathbf{w}} \in \mathscr{H}$

for *m = 1 to* M_n **do**

  Draw a bootstrap data set $\mathscr{D}_{(m)}$ of size n from the training data set $\mathscr{D}$;

  Grow a random-forest tree $\hat{f}_{(m)}$ to the bootstrap data $\mathscr{D}_{(m)}$, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size `nodesize` is reached ; `// nodesize,` q `are hyper parameters`

    i. Select q variables at random from the p variables;

    ii. Pick the best variable/ splitting point among the q;

    iii. Split the node into two daughter nodes.

  for *i = 1 to* n **do**

    Drop $\mathbf{x}_i$ down the the m^{th} tree and get $\mathbf{P}_{\text{BL}(m)}(\mathbf{x}_i, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \mathbf{\Theta}_{(m)})$.

  end

  $\mathbf{P}_{\text{BL}(m)} \leftarrow \{\mathbf{P}_{\text{BL}(m)}(\mathbf{x}_1, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \mathbf{\Theta}_{(m)}), \ldots, \mathbf{P}_{\text{BL}(m)}(\mathbf{x}_n, \mathbf{X}, \mathbf{y}, \mathscr{B}_{(m)}, \mathbf{\Theta}_{(m)})\}^\top$.

end

Solve the quadratic programming problem:

$$\mathbf{w}^\circ = \left(w^\circ_{(1)}, \ldots, w^\circ_{(M_n)}\right)^\top \leftarrow \underset{\mathbf{w} \in \mathscr{H}}{\arg\min}\ C_n^\circ(\mathbf{w});$$

$\tilde{\mathbf{e}} \leftarrow \{\mathbf{I}_n - \mathbf{P}(\mathbf{w}^\circ)\}\mathbf{y}$ with $\mathbf{P}(\mathbf{w}^\circ) = \sum_{m=1}^{M_n} w^\circ_{(m)} \mathbf{P}_{\text{BL}(m)}$;

Solve the quadratic programming problem:

$$\widetilde{\mathbf{w}} = \left(\widetilde{w}_{(1)}, \ldots, \widetilde{w}_{(M_n)}\right)^\top \leftarrow \underset{\mathbf{w} \in \mathscr{H}}{\arg\min}\ C_n''(\mathbf{w}).$$

Algorithm B.2: wRF

Input: (1) The training data set $\mathscr{D} = \{y_i, \mathbf{x}_i\}_{i=1}^n$ (2) The number of trees in RF M_n (3) Parameter λ

Output: Weight vector $\widehat{\mathbf{w}} \in \mathscr{H}$

for *m = 1 to* M_n **do**

  Draw a bootstrap data set $\mathscr{D}_{(m)}$ of size n from the training data set $\mathscr{D}$;

  Grow a random-forest tree $\hat{f}_{(m)}$ to the bootstrap data $\mathscr{D}_{(m)}$, by recursively repeating the following steps for each node of the tree, until the minimum node size `nodesize` is reached ; `// nodesize,` q `are hyper parameters`

    i. Select q variables at random from the p variables;

    ii. Pick the best variable/ split-point among the q;

    iii. Split the node into two daughter nodes.

  $t\mathrm{PE}'_m \leftarrow \frac{1}{\sum_{i=1}^n \mathrm{OOB}_{im}} \sum_{i=1}^n \left| \hat{f}_{(m)}(\mathbf{x}_i) - y_i \right| \cdot \mathrm{OOB}_{im}$;

  $\widehat{w}_{(m)} \leftarrow (\frac{1}{t\mathrm{PE}'_m})^{\lambda}$;

end

$\widehat{\mathbf{w}} \leftarrow \left(\widehat{w}_{(1)}, \ldots, \widehat{w}_{(M_n)}\right)^\top$;

$\widehat{\mathbf{w}} \leftarrow \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|_1}$.

Algorithm B.3: CRF

Input: (1) The training data set $\mathscr{D} = \{y_i, \mathbf{x}_i\}_{i=1}^n$ (2) The number of trees in RF M_n

Output: Weight vector $\widehat{\mathbf{w}} \in \mathscr{H}$

for *m = 1 to* M_n **do**

  Draw a bootstrap data set $\mathscr{D}_{(m)}$ of size n from the training data set $\mathscr{D}$;

  Grow a random-forest tree $\hat{f}_{(m)}$ to the bootstrap data $\mathscr{D}_{(m)}$, by recursively repeating the following steps for each node of the tree, until the minimum node size `nodesize` is reached ; `// nodesize,` q `are hyper parameters`

    i. Select q variables at random from the p variables;

    ii. Pick the best variable/ split-point among the q;

    iii. Split the node into two daughter nodes.

  $t\mathrm{PE}'_m \leftarrow \frac{1}{\sum_{i=1}^n \mathrm{OOB}_{im}} \sum_{i=1}^n \left| \hat{f}_{(m)}(\mathbf{x}_i) - y_i \right| \cdot \mathrm{OOB}_{im}$;

end

Sequence $\{t\mathrm{PE}'_1, \ldots, t\mathrm{PE}'_{M_n}\}$ from smallest to largest;

for *m = 1 to* M_n **do**

  $r_m \leftarrow$ the order of the m^{th} tree in sorted sequence;

  $\widehat{w}_{(m)} \leftarrow \sum_{\nu=r_m}^{M_n} \frac{1}{\nu}$;

end

$\widehat{\mathbf{w}} \leftarrow \left(\widehat{w}_{(1)}, \ldots, \widehat{w}_{(M_n)}\right)^\top$;

$\widehat{\mathbf{w}} \leftarrow \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|_1}$.

Data set	RF	2steps-WRF$_{\text{opt}}$	1step-WRF$_{\text{opt}}$	wRF	CRF
BH	$11.913^{(5)}$	$11.380^{(1)}$	$11.445^{(3)}$	$11.401^{(2)}$	$11.521^{(4)}$
Servo	$1.053^{(3)}$	$1.057^{(4)}$	$1.063^{(5)}$	$1.017^{(1)}$	$1.049^{(2)}$
CCS	$27.944^{(4)}$	$27.658^{(1)}$	$27.721^{(3)}$	$27.703^{(2)}$	$27.962^{(5)}$
ASN	$5.308^{(5)}$	$5.095^{(1)}$	$5.103^{(2)}$	$5.241^{(4)}$	$5.232^{(3)}$
CCPP	$15.818^{(2)}$	$16.084^{(4)}$	$16.256^{(5)}$	$15.790^{(1)}$	$15.901^{(3)}$
CST	$10.618^{(5)}$	$9.319^{(1)}$	$9.395^{(2)}$	$9.740^{(3)}$	$10.109^{(4)}$
EE	$3.533^{(2)}$	$3.534^{(3)}$	$3.541^{(4)}$	$3.485^{(1)}$	$3.559^{(5)}$
PT	$1.806^{(5)}$	$1.452^{(2)}$	$1.449^{(1)}$	$1.571^{(3)}$	$1.609^{(4)}$
QSAR	$1.378^{(3)}$	$1.387^{(4)}$	$1.397^{(5)}$	$1.358^{(1)}$	$1.368^{(2)}$
SM($\times 10^{-5}$)	$2.629^{(5)}$	$1.994^{(1)}$	$1.999^{(2)}$	$2.444^{(3)}$	$2.444^{(4)}$
YH	$2.495^{(5)}$	$1.850^{(1)}$	$1.891^{(3)}$	$1.869^{(2)}$	$2.072^{(4)}$
Tecator	$3.790^{(5)}$	$2.955^{(2)}$	$2.948^{(1)}$	$3.178^{(3)}$	$3.338^{(4)}$

Table D.1: Test Error Comparisons by MSFE for Different Forests on High-Dimensional Data

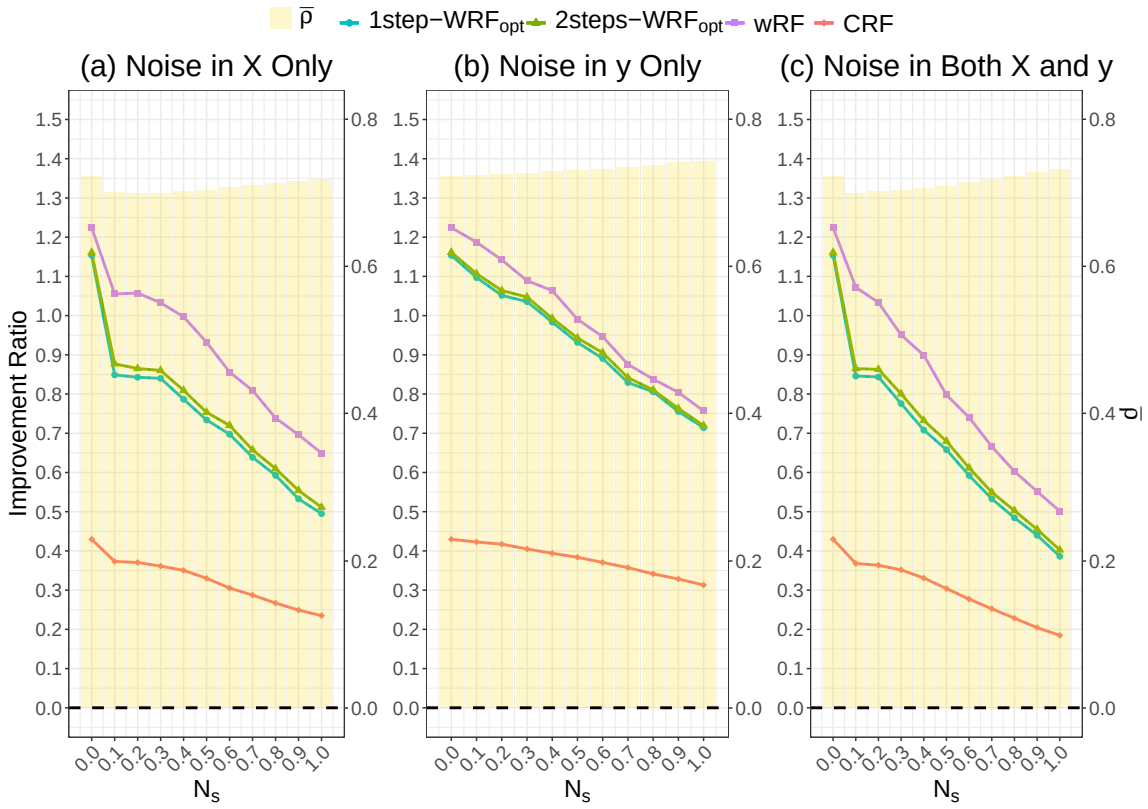

Figure D.2: Improvement Ratio vs Noise on Servo Data Set

Data set	RF	2steps-WRF$_{opt}$	1step-WRF$_{opt}$	wRF	CRF
BH	$2.293^{(5)}$	$2.273^{(3)}$	$2.284^{(4)}$	$2.254^{(1)}$	$2.271^{(2)}$
Servo	$0.610^{(3)}$	$0.613^{(4)}$	$0.614^{(5)}$	$0.604^{(1)}$	$0.609^{(2)}$
CCS	$3.870^{(5)}$	$3.836^{(1)}$	$3.841^{(2)}$	$3.847^{(3)}$	$3.866^{(4)}$
ASN	$1.672^{(5)}$	$1.646^{(1)}$	$1.648^{(2)}$	$1.662^{(3)}$	$1.663^{(4)}$
CCPP	$3.030^{(2)}$	$3.048^{(4)}$	$3.060^{(5)}$	$3.026^{(1)}$	$3.034^{(3)}$
CST	$2.504^{(5)}$	$2.311^{(1)}$	$2.320^{(2)}$	$2.377^{(3)}$	$2.434^{(4)}$
EE	$1.191^{(2)}$	$1.192^{(3)}$	$1.194^{(4)}$	$1.184^{(1)}$	$1.197^{(5)}$
PT	$0.860^{(5)}$	$0.777^{(1)}$	$0.777^{(1)}$	$0.794^{(3)}$	$0.808^{(4)}$
QSAR	$0.871^{(3)}$	$0.873^{(4)}$	$0.876^{(5)}$	$0.864^{(1)}$	$0.867^{(2)}$
SM($\times 10^{-3}$)	$2.882^{(5)}$	$2.763^{(1)}$	$2.773^{(2)}$	$2.813^{(3)}$	$2.846^{(4)}$
YH	$0.724^{(5)}$	$0.667^{(2)}$	$0.675^{(3)}$	$0.651^{(1)}$	$0.676^{(4)}$
Tecator	$1.342^{(5)}$	$1.207^{(1)}$	$1.207^{(1)}$	$1.246^{(3)}$	$1.275^{(4)}$

Table D.2: Test Error Comparisons by MAFE for Different Forests on High-Dimensional Data

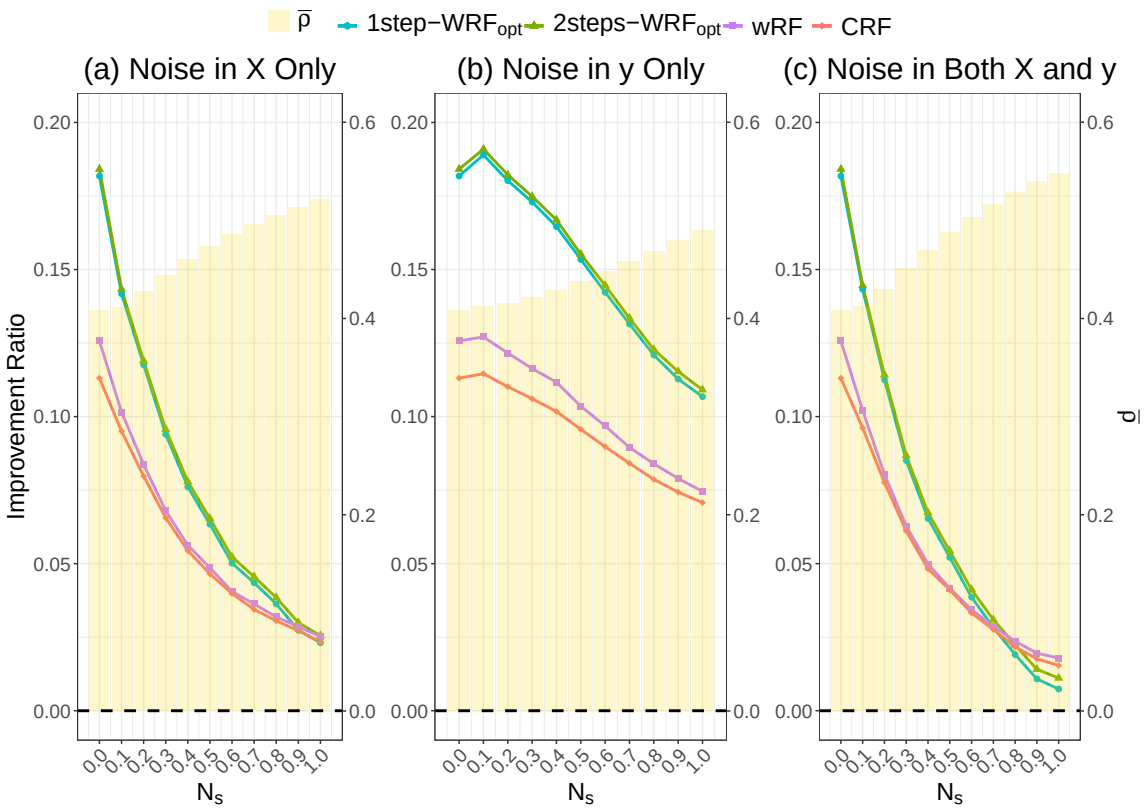

Figure D.3: Improvement Ratio vs Noise on CCS Data Set

Figure D.4: Improvement Ratio vs Noise on CST Data Set

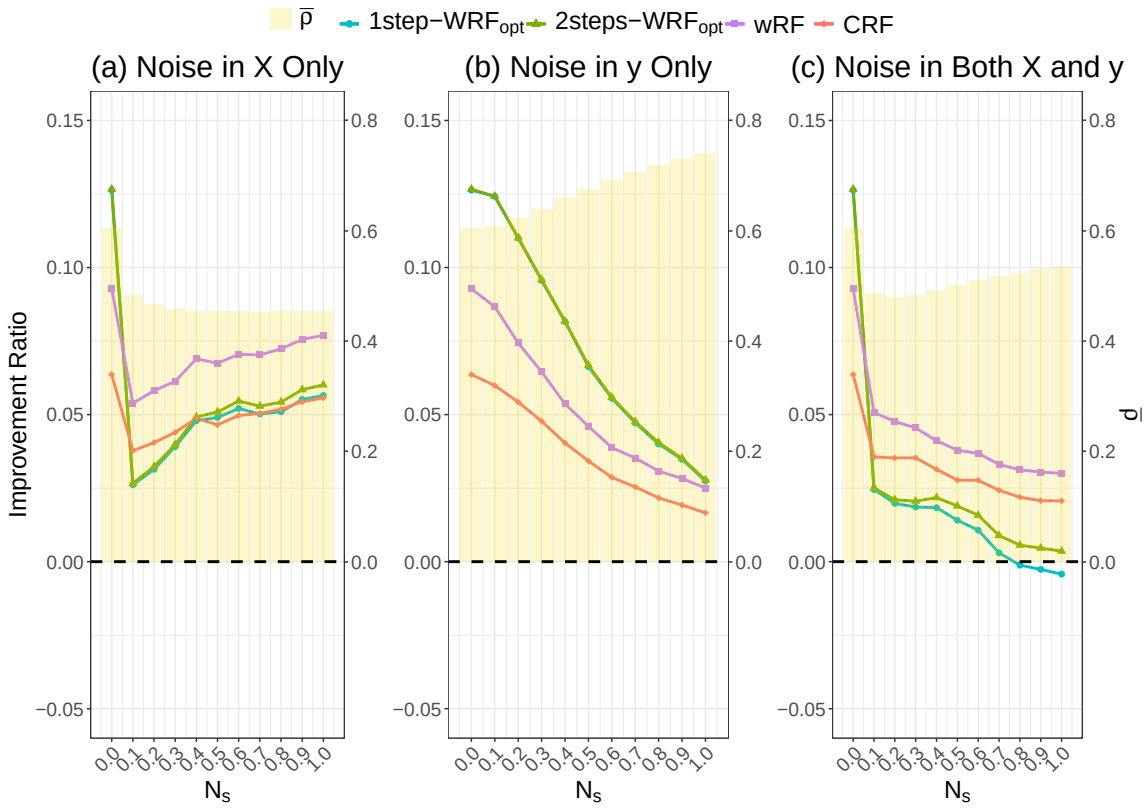

Figure D.5: Improvement Ratio vs Noise on EE Data Set

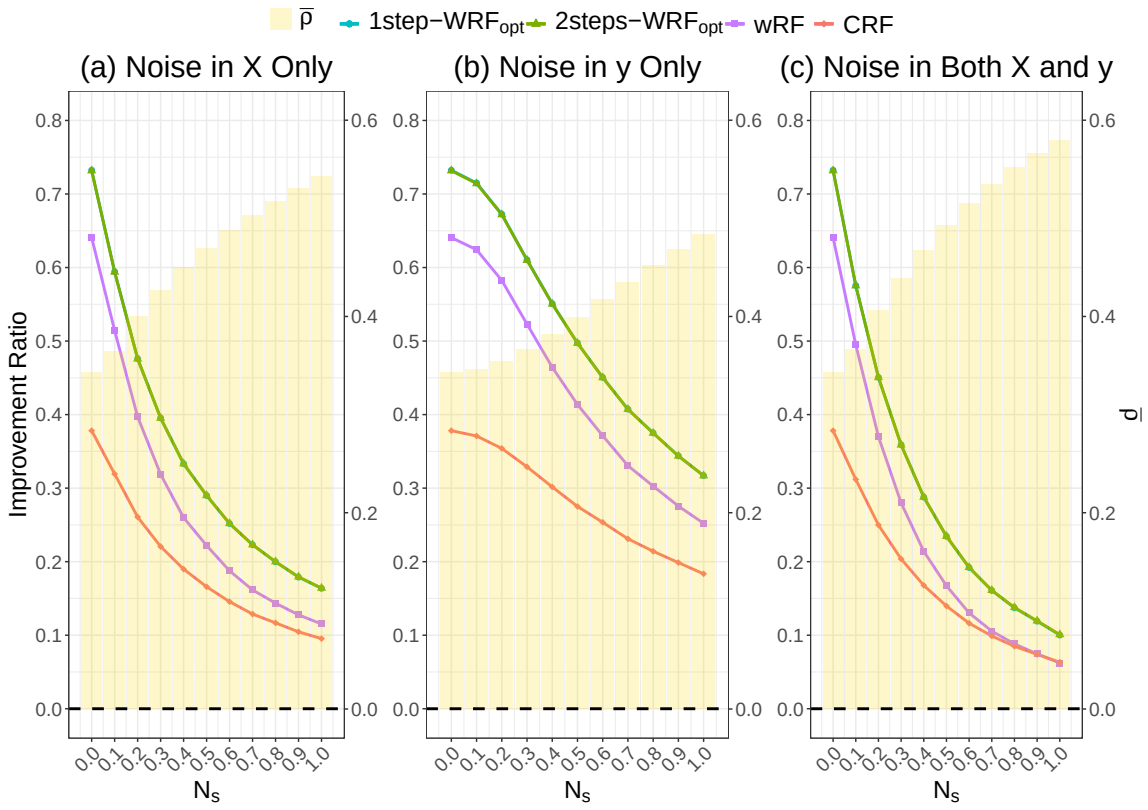

Figure D.6: Improvement Ratio vs Noise on PT Data Set

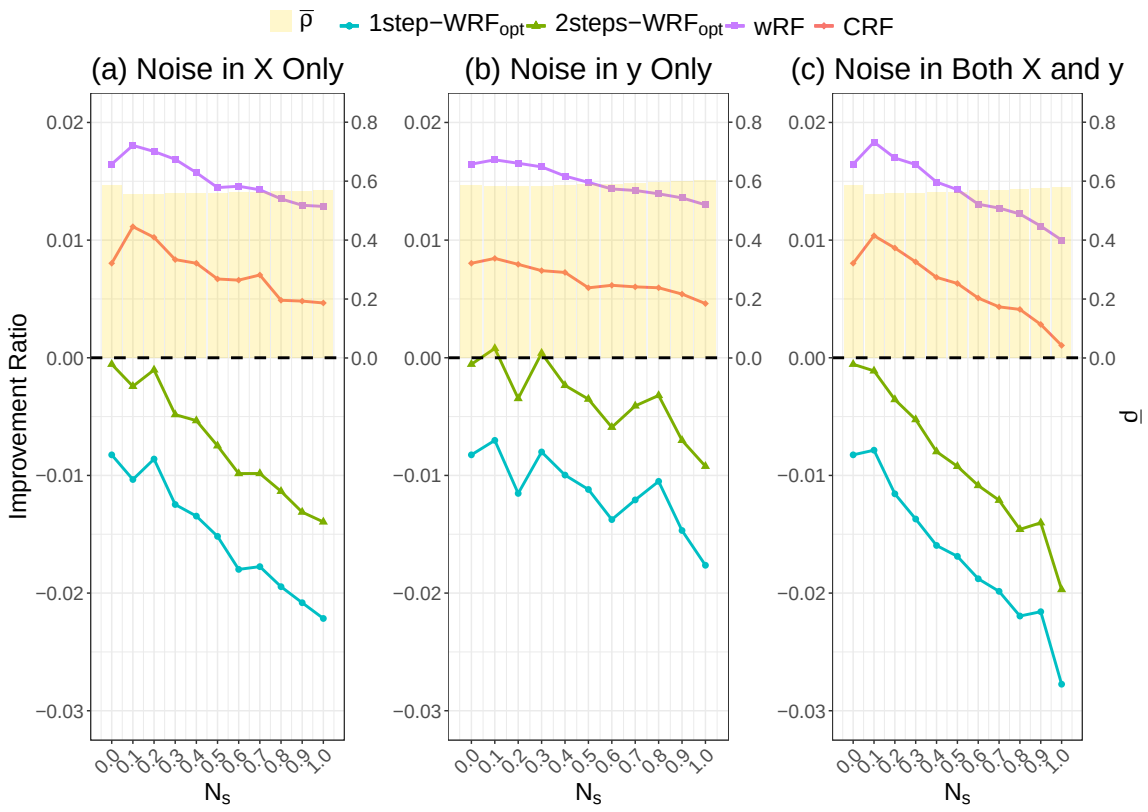

Figure D.7: Improvement Ratio vs Noise on QSAR Data Set

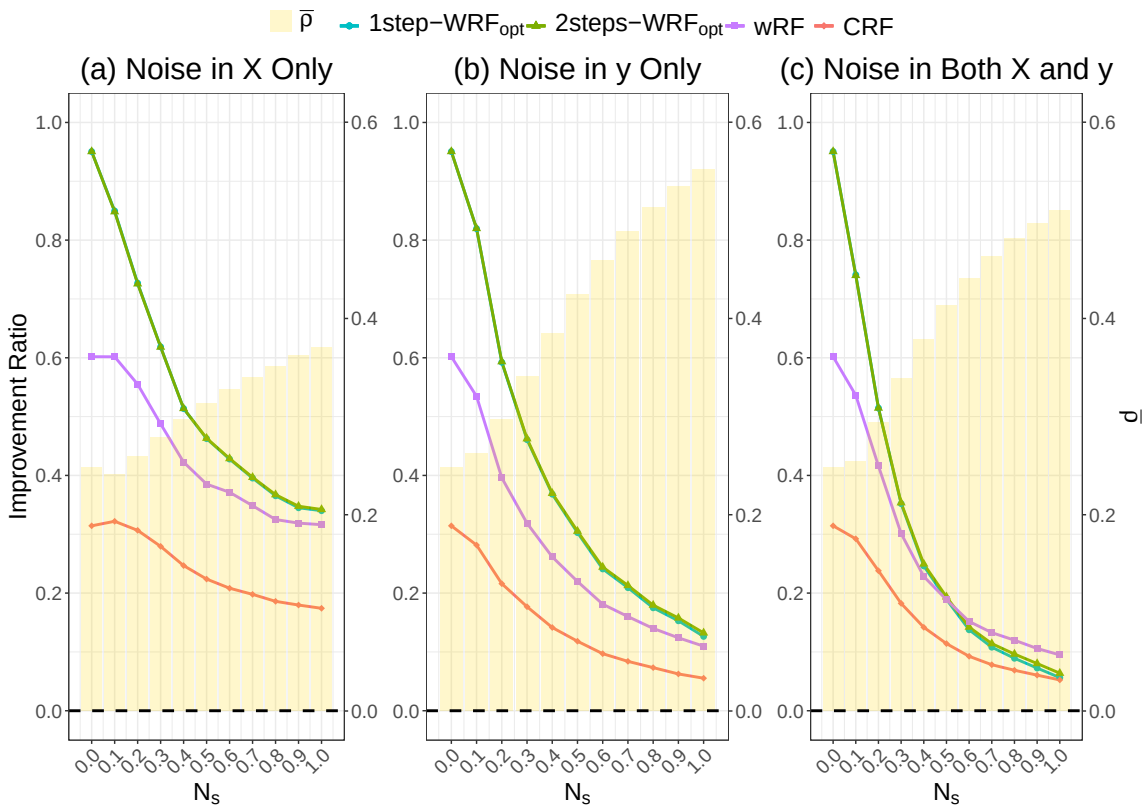

Figure D.8: Improvement Ratio vs Noise on SM Data Set

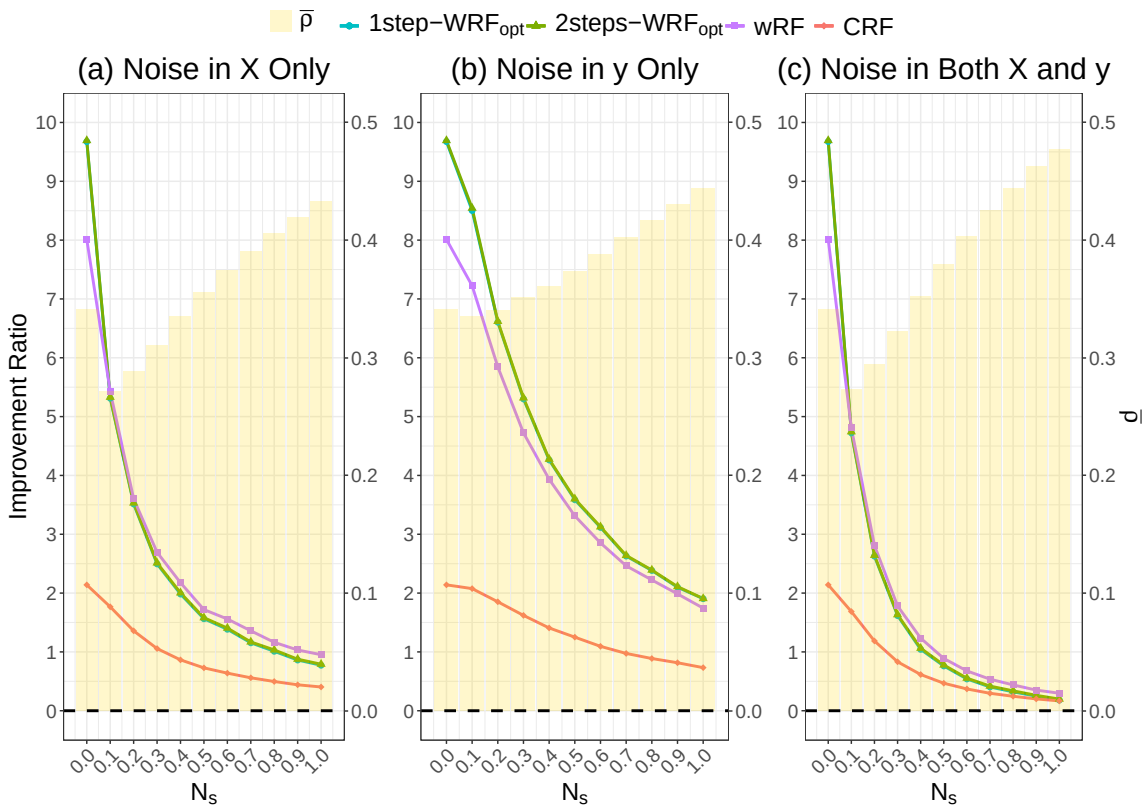

Figure D.9: Improvement Ratio vs Noise on YH Data Set

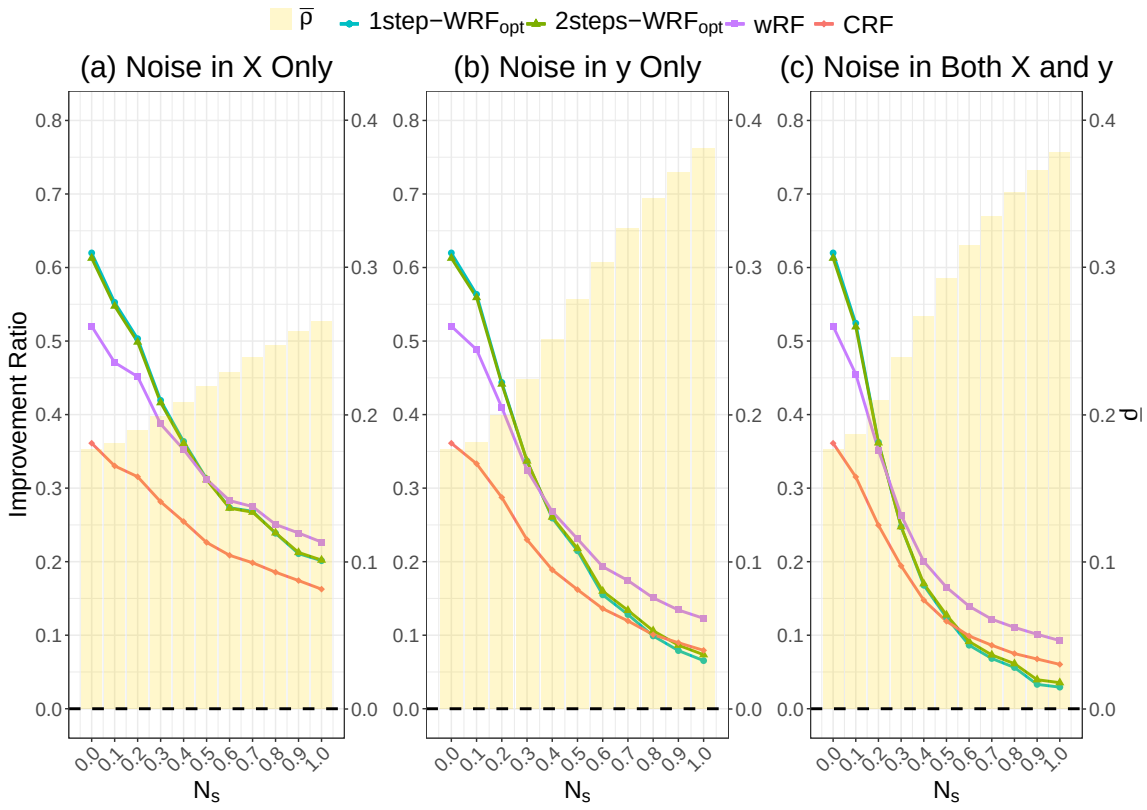

Figure D.10: Improvement Ratio vs Noise on Tecator Data Set

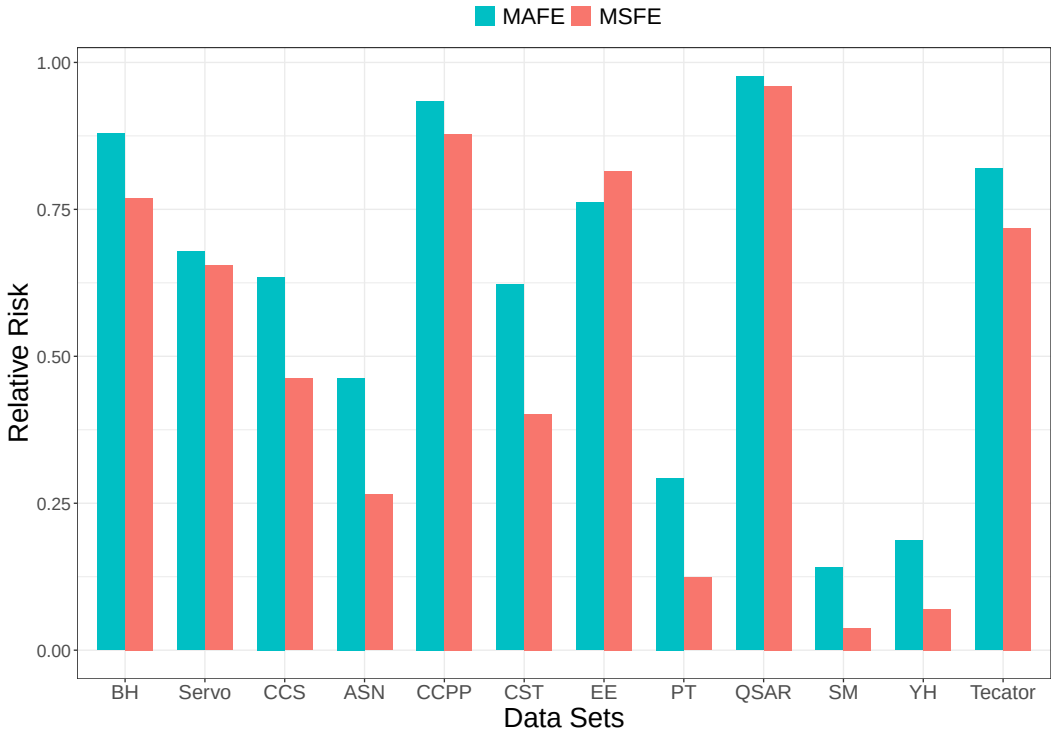

Figure D.11: Comparative Analysis of Conventional RF Predictive Performance Between Original and Augmented Data Sets (Relative risks are calculated as the ratio of conventional RF risks on augmented data sets to those on original data sets.)

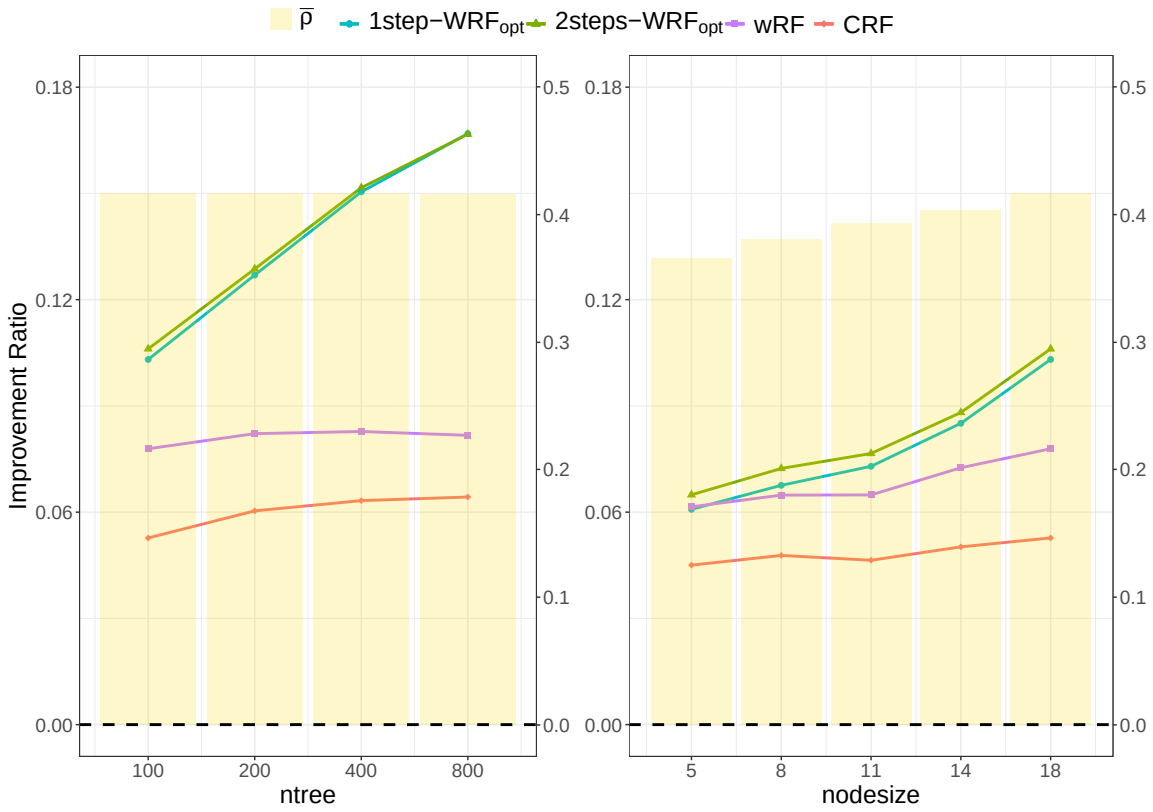

Figure D.12: Improvement Ratio vs Hyper Parameters of RF on BH Data Set

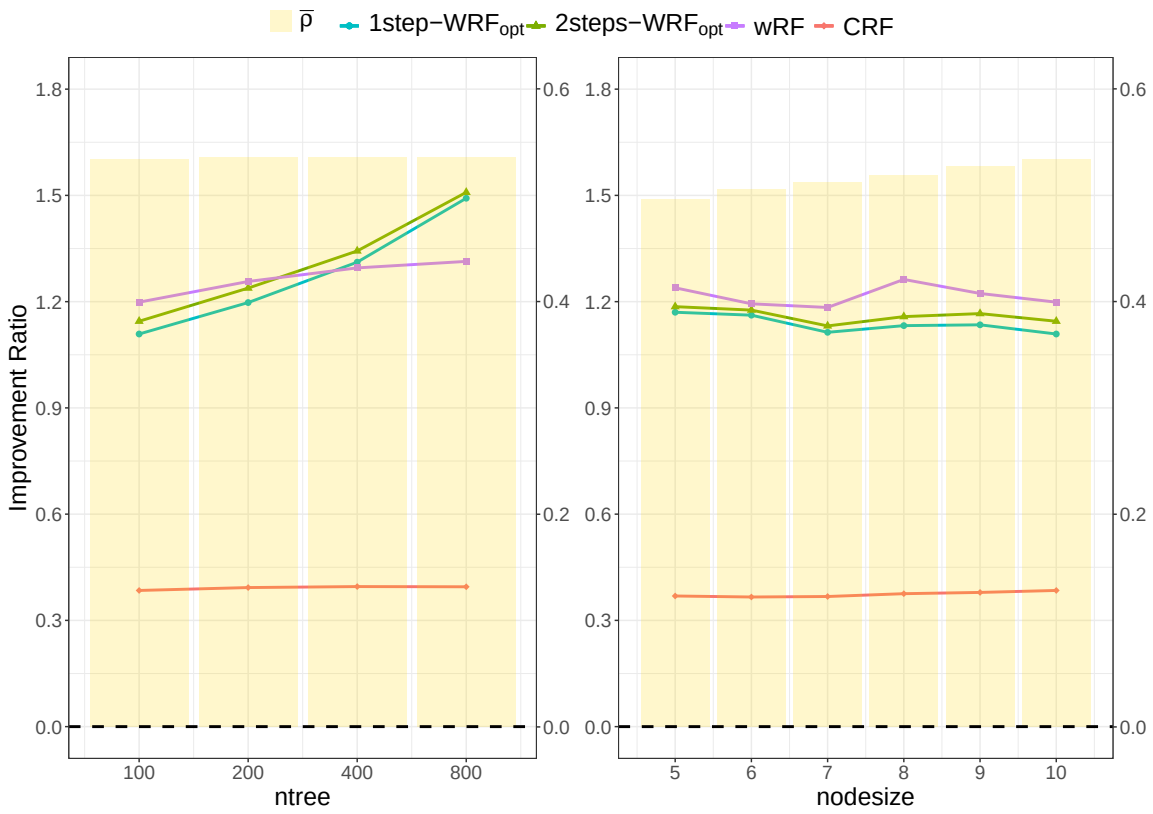

Figure D.13: Improvement Ratio vs Hyper Parameters of RF on Servo Data Set

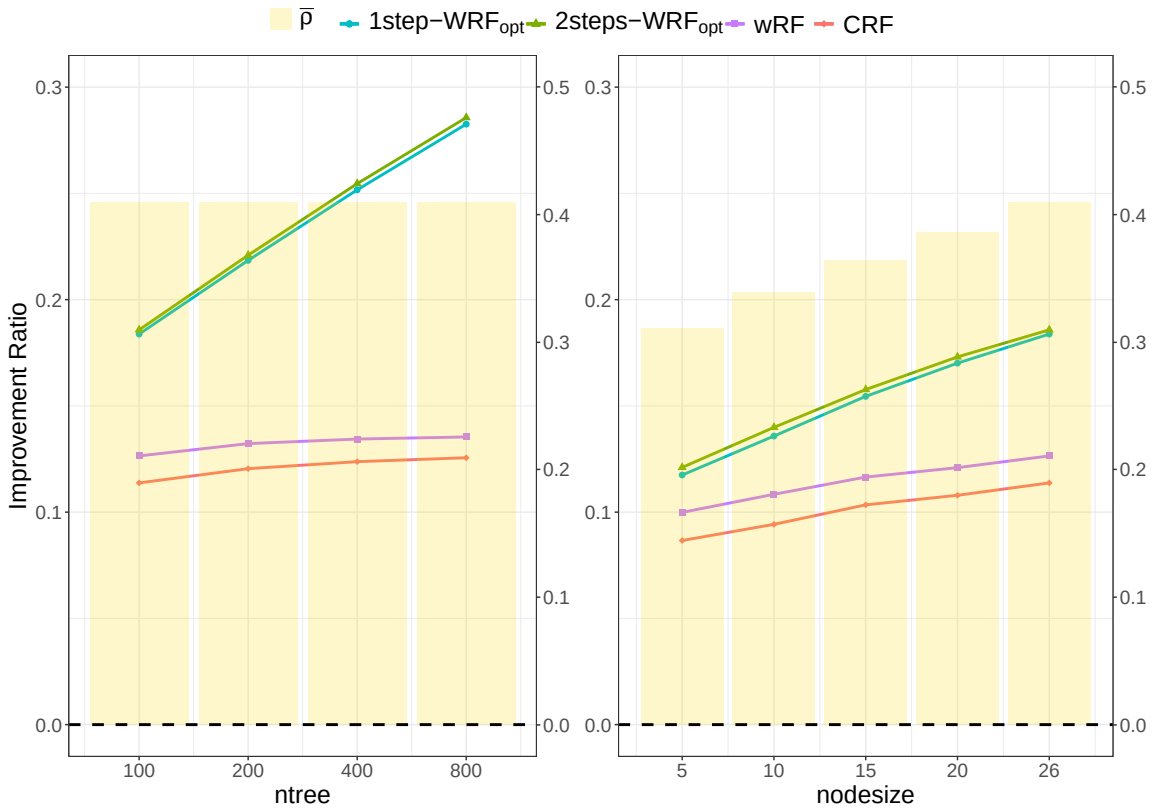

Figure D.14: Improvement Ratio vs Hyper Parameters of RF on CCS Data Set

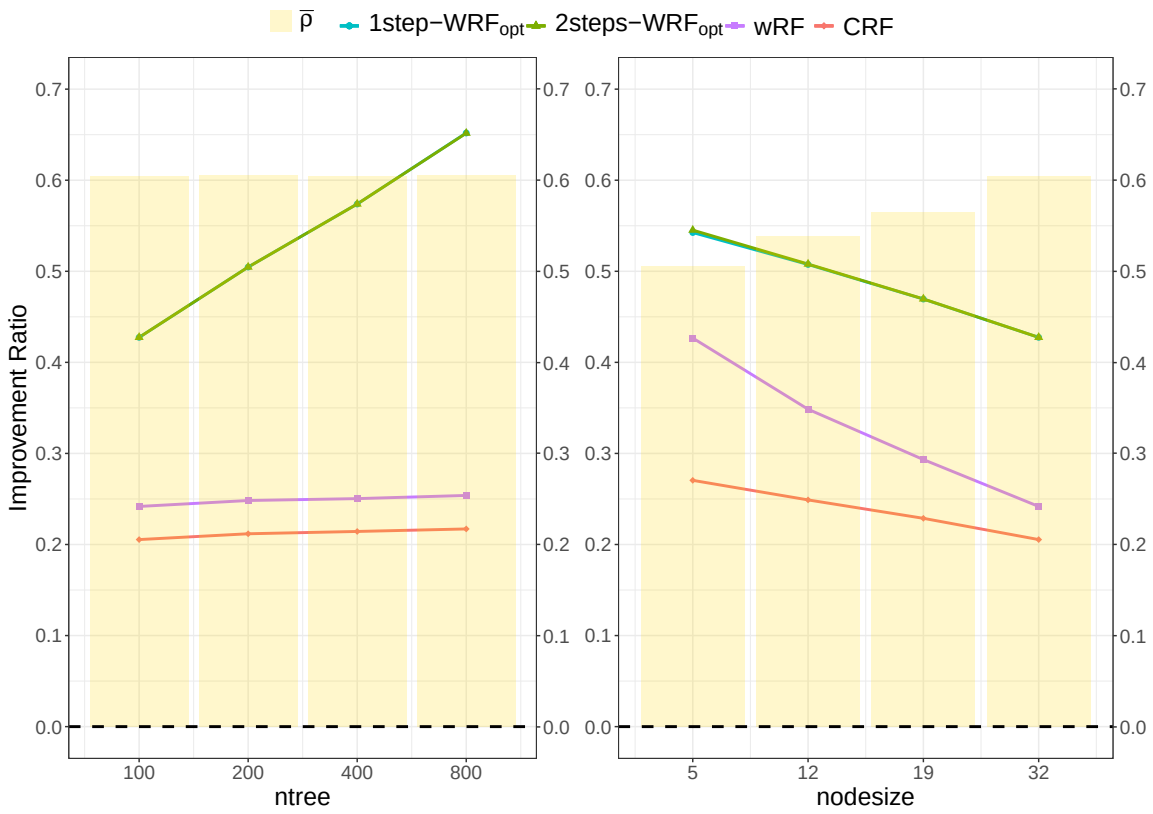

Figure D.15: Improvement Ratio vs Hyper Parameters of RF on ASN Data Set

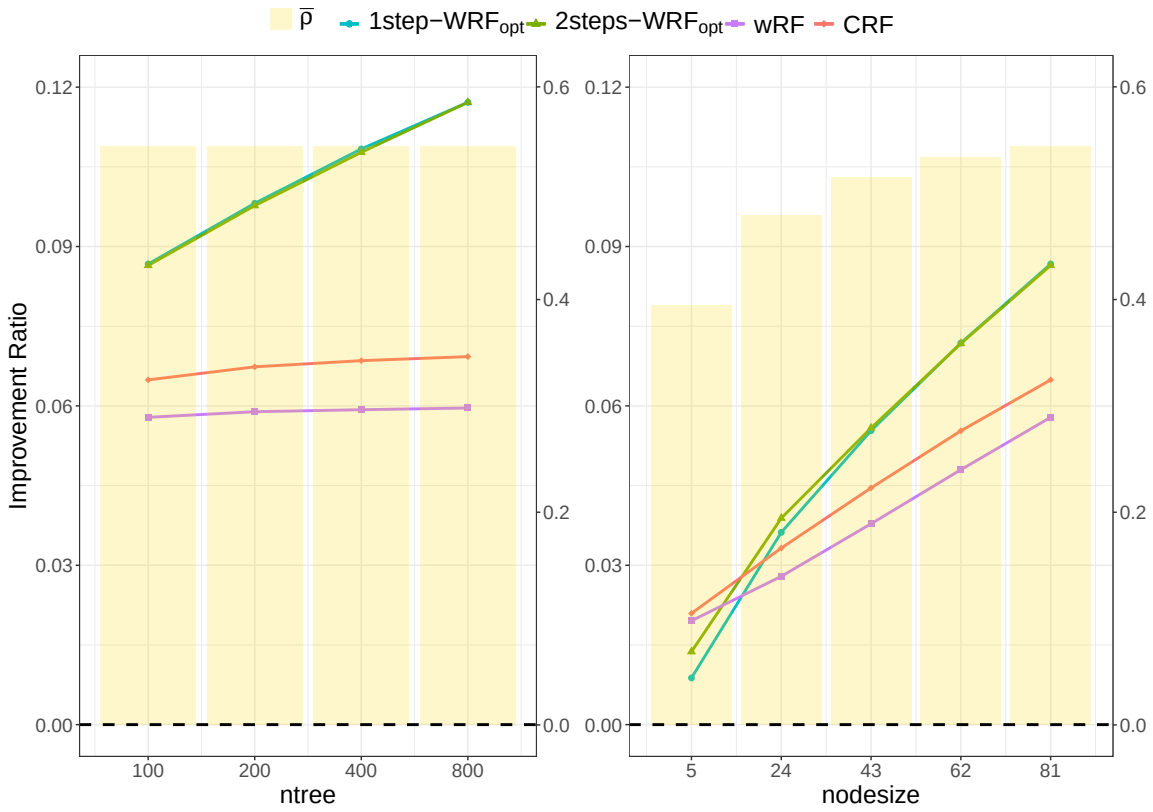

Figure D.16: Improvement Ratio vs Hyper Parameters of RF on CCPP Data Set

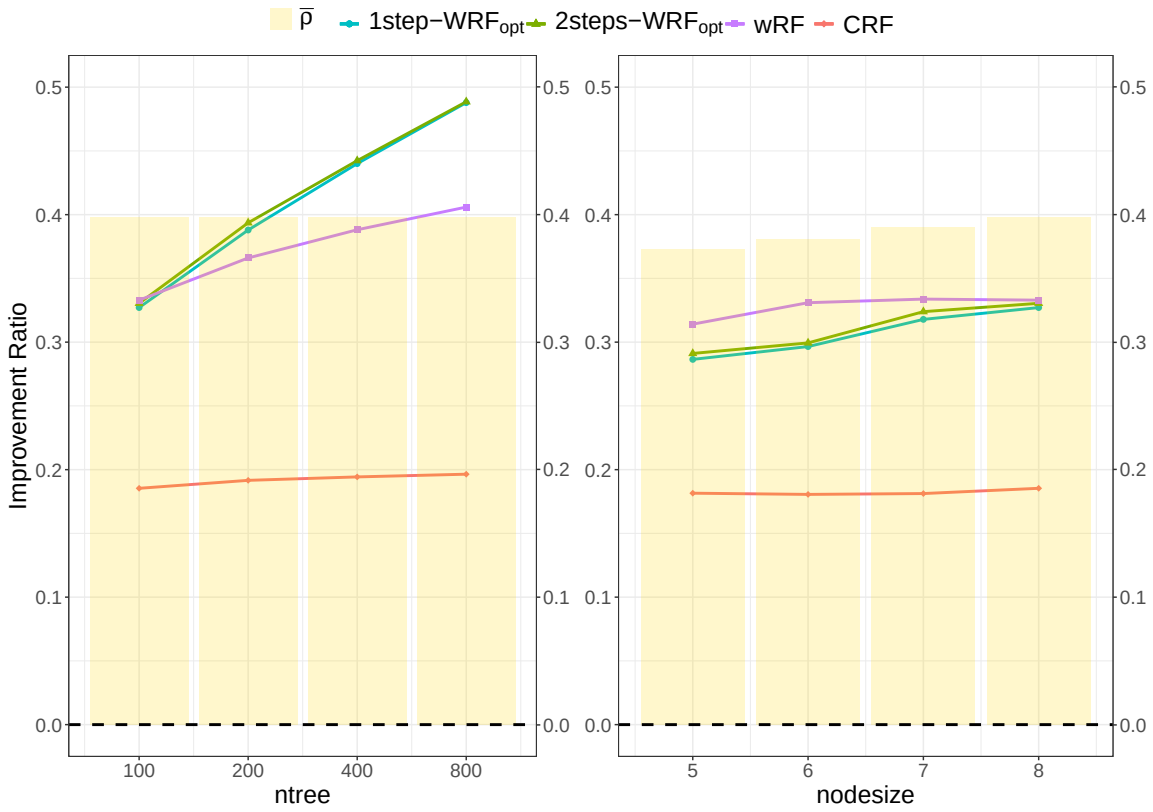

Figure D.17: Improvement Ratio vs Hyper Parameters of RF on CST Data Set

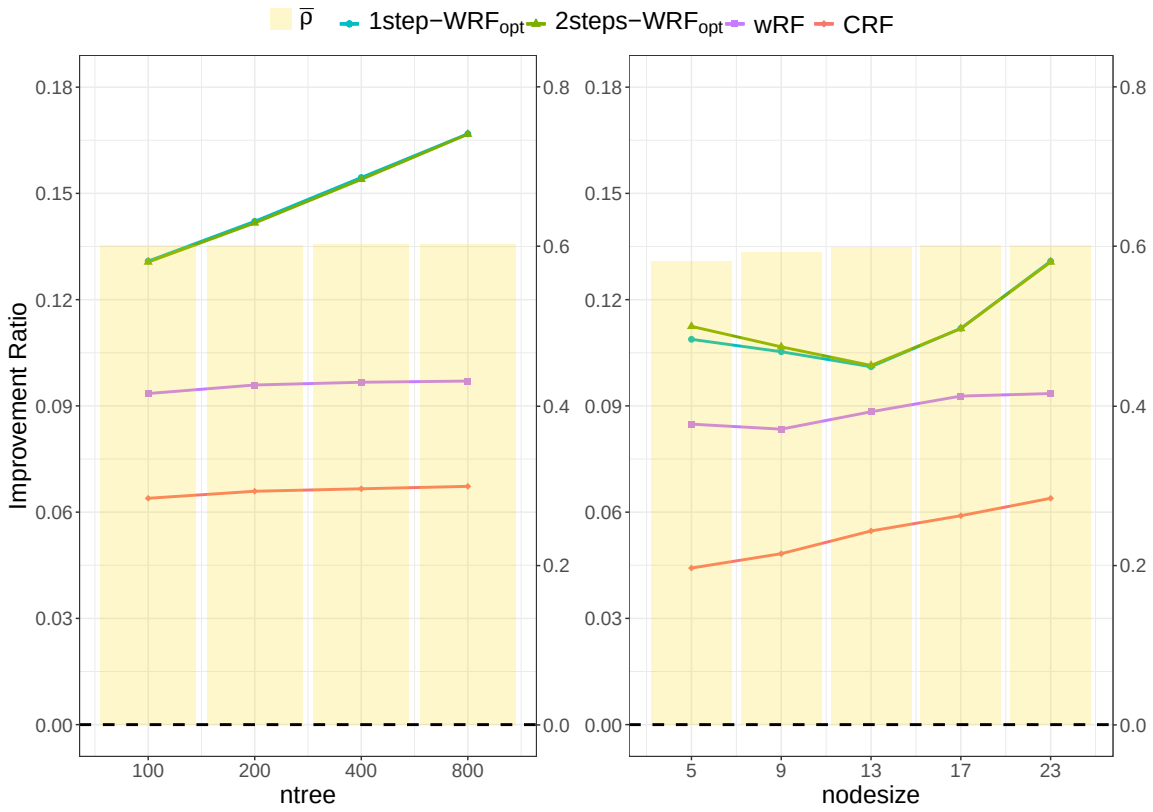

Figure D.18: Improvement Ratio vs Hyper Parameters of RF on EE Data Set

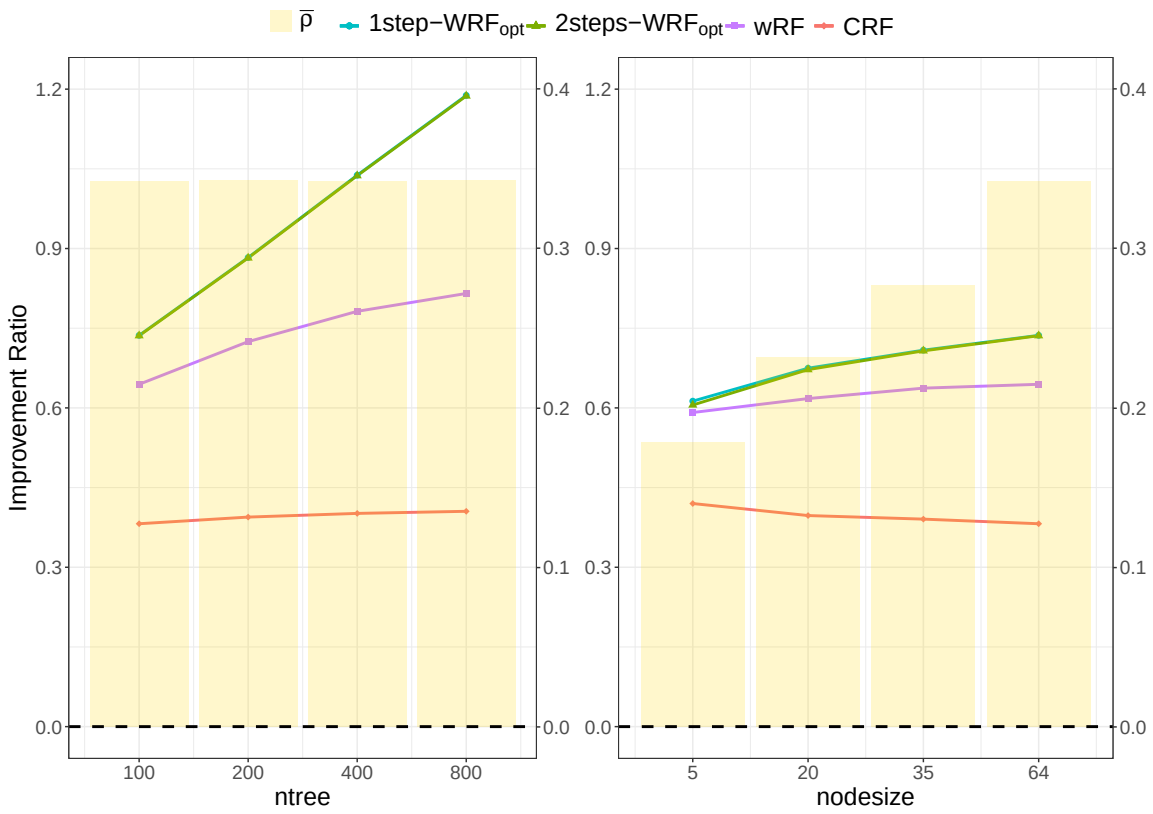

Figure D.19: Improvement Ratio vs Hyper Parameters of RF on PT Data Set

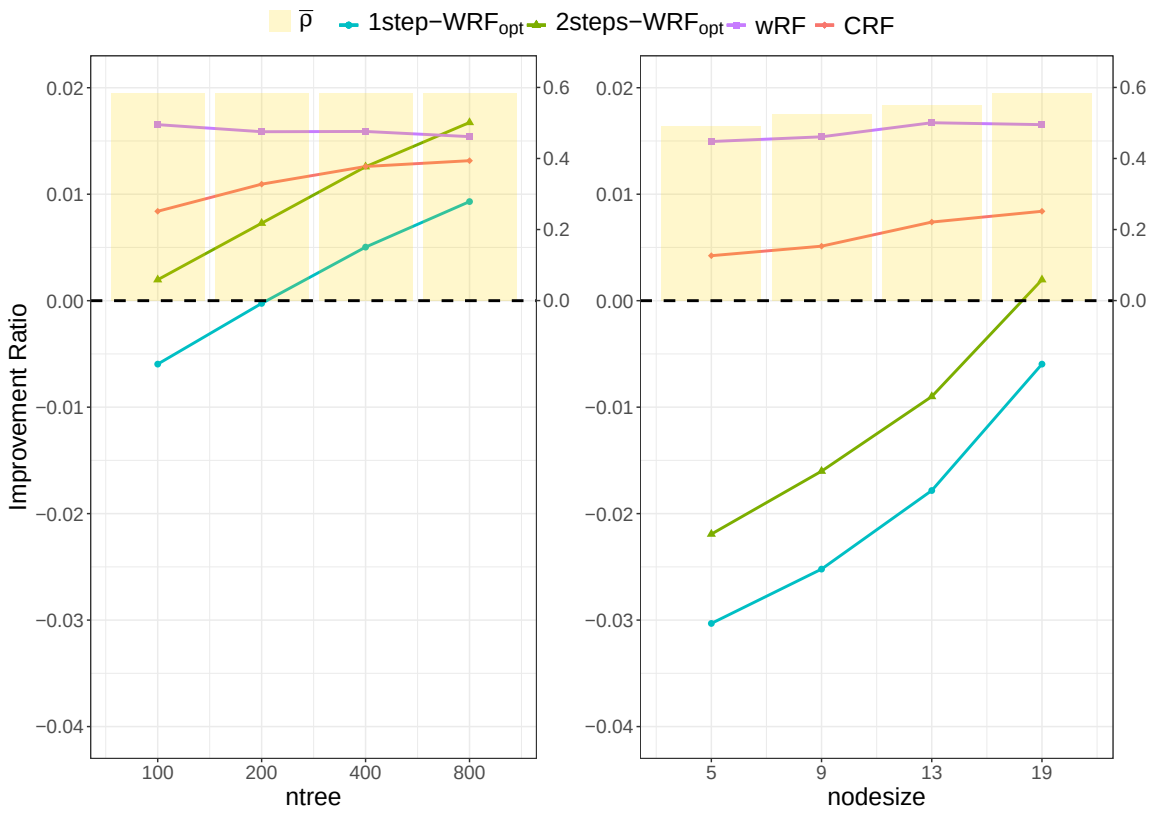

Figure D.20: Improvement Ratio vs Hyper Parameters of RF on QSAR Data Set

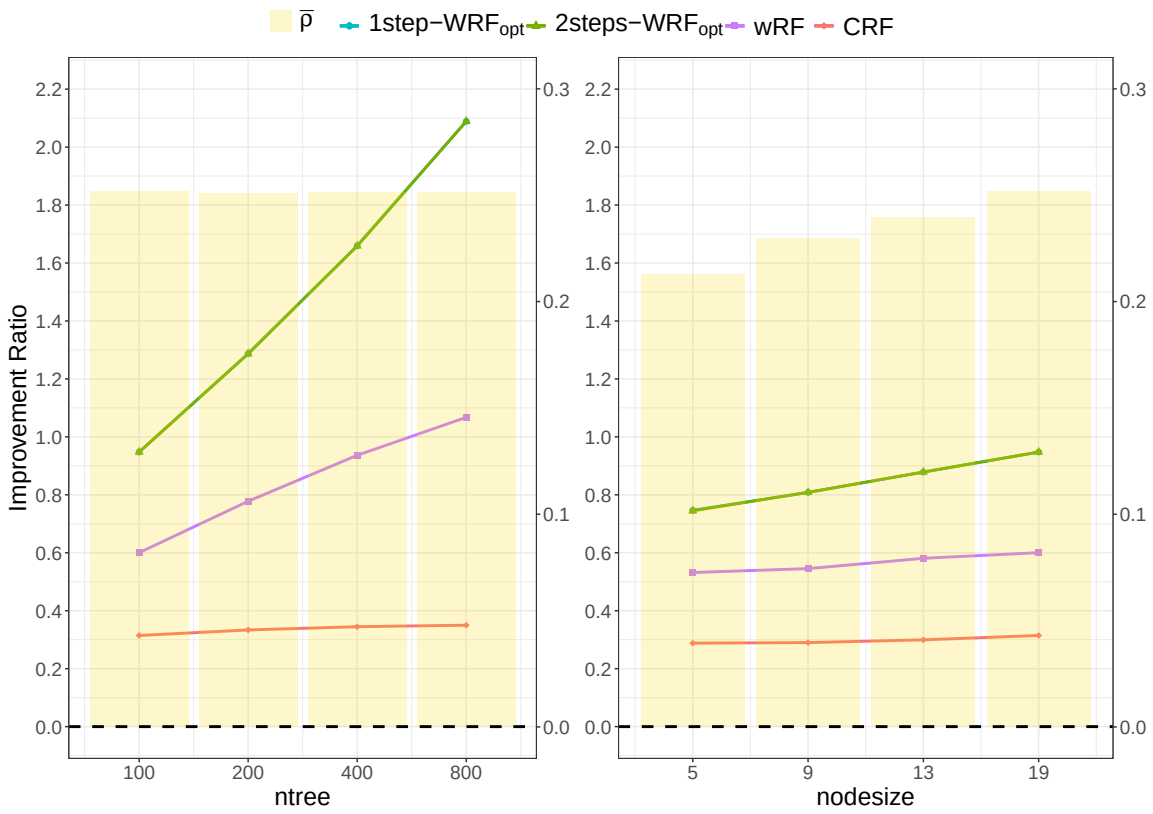

Figure D.21: Improvement Ratio vs Hyper Parameters of RF on SM Data Set

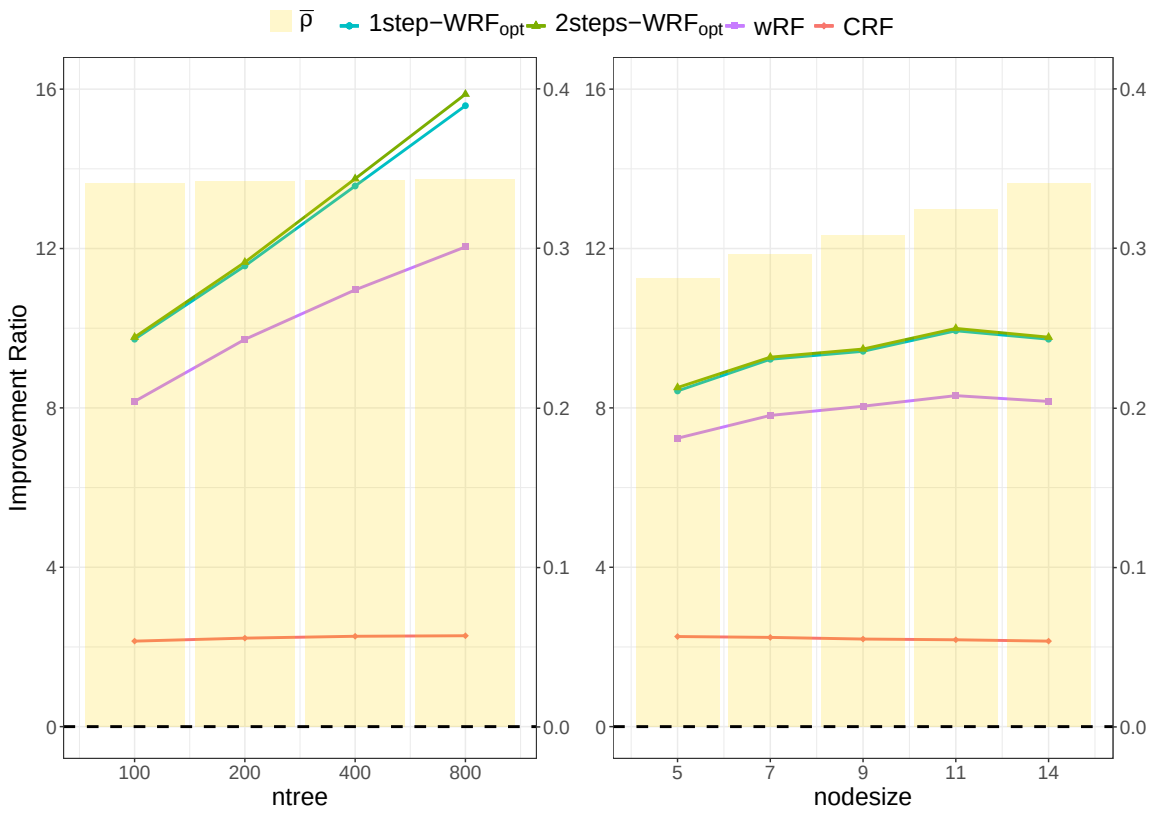

Figure D.22: Improvement Ratio vs Hyper Parameters of RF on YH Data Set

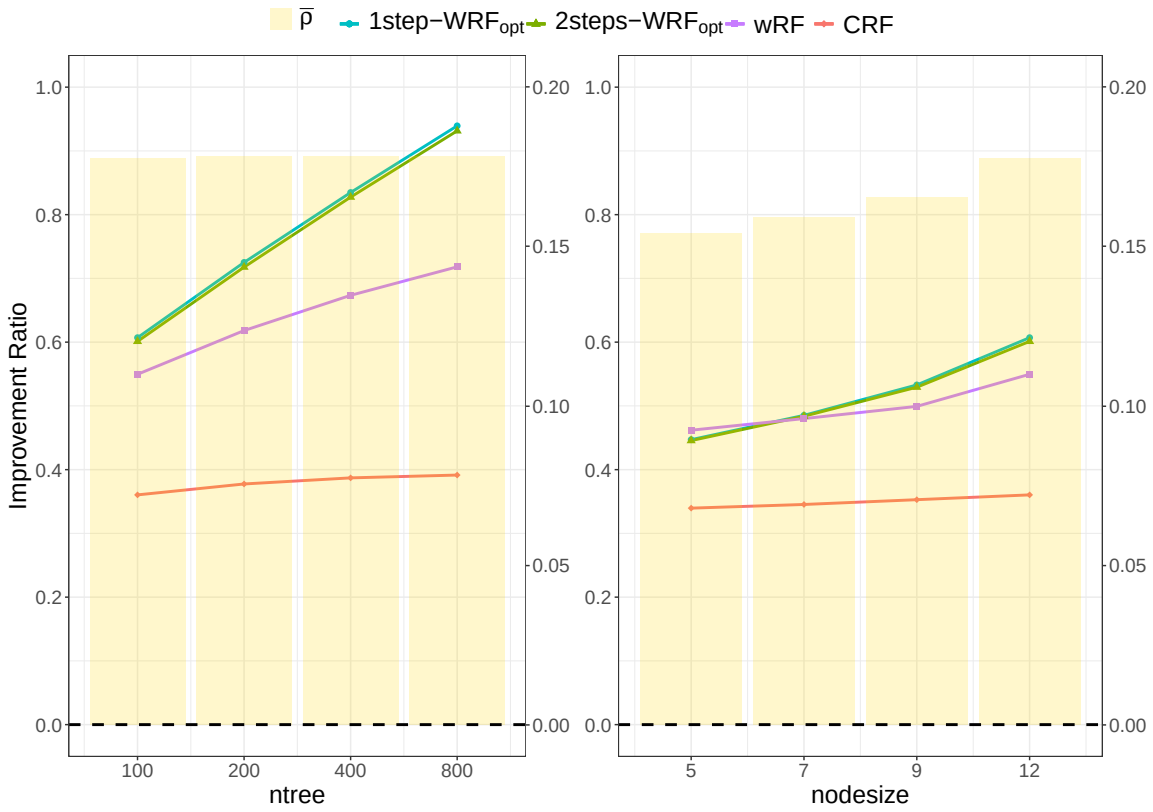

Figure D.23: Improvement Ratio vs Hyper Parameters of RF on Tecator Data Set

References

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

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

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

V. V. Buldygin and K. K. Moskvichova. The sub-Gaussian norm of a binary random variable. *Theory of Probability and Mathematical Statistics*, 86:33–49, 2013.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

C. M. Chi, P. Vossler, Y. Fan, and J. Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.

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

C. J. Flynn, C. M. Hurvich, and J. S. Simonoff. Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association*, 108:1031–1043, 2013.

Y. Gao, X. Zhang, S. Wang, T. T. Chong, and G. Zou. Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics*, 71(2):275–306, 2019.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63 (1):3–42, 2006.

P. Ghosh, A. Neufeld, and J. K. Sahoo. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. *Finance Research Letters*, 46(Part A):102280, 2022.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007.

B. E. Hansen and J. S. Racine. Jackknife model averaging. *Journal of Econometrics*, 167 (1):38–46, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, chapter 9, pages 307–308. Springer, 2009.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, chapter 8, pages 311–314. Springer, 2013.

J. M. Klusowski. Universal consistency of decision trees in high dimensions. *arXiv preprint arXiv:2104.13881*, 2021.

H. Li, W. Wang, H. Ding, and J. Dong. Trees weighting random forest method for classifying high-dimensional noisy data. In *2010 IEEE 7th International Conference on E-Business Engineering*, pages 160–163, 2010.

J. Lin, S. Lu, X. He, and F. Wang. Analyzing the impact of three-dimensional building structure on CO_2 emissions based on random forest regression. *Energy*, 236(1):121502, 2021.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

H. Pallathadka, E. H. Ramirez-Asis, T. P. Loli-Poma, K. Kaliyaperumal, R. J. M. Ventayen, and M. Naved. Applications of artificial intelligence in business management, e-commerce and finance. *Materials Today: Proceedings*, 80:2610–2613, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Peng and Y. Yang. On improvability of model selection by model averaging. *Journal of Econometrics*, 229(2):246–262, 2022.

H. Pham and S. Olafsson. On Cesáro averages for weighted trees in the random forest. *Journal of Classification*, 37(1):1–14, 2019.

P. Probst, M. N. Wright, and A. L. Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

Y. Qiu, T. Xie, J. Yu, and X. Zhang. Mallows-type averaging machine learning techniques. Working paper, 2020.

I. Reis, D. Baron, and S. Shahaf. Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, 157(1):16, 2018.

J. Saniuk and I. Rhodes. A matrix inequality associated with bounds on solutions of algebraic riccati and lyapunov equations. *IEEE Transactions on Automatic Control*, 32(8):739–740, 1987.

E. Scornet, G. Biau, and J. P. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

M. R. Segal. Machine learning benchmarks and random forest regression. *Center for Bioinformatics and Molecular Biostatistics*, 2004.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003.

V. Syrgkanis and M. Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Proceedings of Thirty Third Conference on Learning Theory: PMLR*, volume 125, pages 3453–3454, 2020.

J. Vanschoren, J. N. Van Rijn, B. Bischl, and L. Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL `http://doi.acm.org/10.1145/2641190.2641198`.

S. J. Winham, R. R. Freimuth, and J. M. Biernacka. A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):496–505, 2013.

S. Xuan, G. Liu, and Z. Li. Refined weighted random forest and its application to credit card fraud detection. In *Computational Data and Social Networks*, pages 343–355, 2018.

J. Yoon. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, 57(1):247–265, 2021.

H. Zhang and S. Chen. Concentration inequalities for statistical inference. *Communications in Mathematical Research*, 37(1):1–85, 2021.

X. Zhang. A new study on asymptotic optimality of least squares model averaging. *Econometric Theory*, 37(2):388–407, 2021.

X. Zhang, A. T. K. Wan, and G. Zou. Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, 174(2):82–94, 2013.

X. Zhang, D. Yu, G. Zou, and H. Liang. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111(516):1775–1790, 2016.

X. Zhang, G. Zou, H. Liang, and R. J. Carroll. Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115 (530):972–984, 2020.

S. Zhao, X. Zhang, and Y. Gao. Model averaging with averaging covariance matrix. *Economics Letters*, 145:214–217, 2016.

Z. Zhou. *Ensemble Methods: Foundations and Algorithms*, chapter 4, pages 68–71. CRC Press, 2012.

Z. Zhou, J. Wu, and W. Tang. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002.

J. Zou, W. Wang, X. Zhang, and G. Zou. Optimal model averaging for divergent-dimensional Poisson regressions. *Econometric Reviews*, 41(7):775–805, 2022.