

The Honest Truth About Causal Trees: Accuracy Limits for Heterogeneous Treatment Effect Estimation

Matias D. Cattaneo* Jason M. Klusowski* Ruiqi (Rae) Yu*

September 16, 2025

Abstract

Recursive decision trees have emerged as a leading methodology for heterogeneous causal treatment effect estimation and inference in experimental and observational settings. These procedures are fitted using the celebrated CART (Classification And Regression Tree) algorithm [Breiman et al., 1984], or custom variants thereof, and hence are believed to be “adaptive” to high-dimensional data, sparsity, or other specific features of the underlying data generating process. Athey and Imbens [2016] proposed several “honest” causal decision tree estimators, which have become the standard in both academia and industry. We study their estimators, and variants thereof, and establish lower bounds on their estimation error. We demonstrate that these popular heterogeneous treatment effect estimators cannot achieve a polynomial-in- n convergence rate under basic conditions, where n denotes the sample size. Contrary to common belief, honesty does not resolve these limitations and at best delivers negligible logarithmic improvements in sample size or dimension. As a result, these commonly used estimators can exhibit poor performance in practice, and even be inconsistent in some settings. Our theoretical insights are empirically validated through simulations.

Keywords: recursive partitioning, decision trees, causal inference, heterogeneous treatment effects

*Department of Operations Research and Financial Engineering, Princeton University.

1 Introduction

[Athey and Imbens \[2016\]](#) proposed to use recursive decision trees to estimate (and later conduct inference about) heterogeneous causal effects in experimental and observational settings. Their methodology is often called “honest” causal trees. Due in part to its simple, interpretable structure, their causal inference methodology has been widely adopted in academic and industry empirical research over the last decade. For example, to advocate for their proposal, the authors wrote that “[i]t enables researchers to let the data discover relevant subgroups while preserving the validity of confidence intervals constructed on treatment effects within subgroups” [[Athey and Imbens, 2016](#), page 7353].

Despite the widespread use of honest causal tree estimators, little is known about their theoretical properties for estimation and inference. Existing results typically require very strong assumptions on the tree-growing process [[Wager and Athey, 2018](#)], which we show are incompatible with canonical implementations of causal trees under basic conditions. Specifically, this paper establishes lower bounds on the estimation error of heterogeneous treatment effect estimators based on recursive adaptive partitioning. We demonstrate that such estimators cannot achieve a polynomial-in- n convergence rate under basic conditions, where n denotes the sample size. Instead, these popular estimators can exhibit arbitrarily slow convergence rates, if not become inconsistent in some cases. As a consequence, our theoretical insights demonstrate that honest causal tree estimators, and variant thereof, may be inaccurate for estimating heterogeneous causal effects, and invalid for constructing confidence intervals on treatment effects within subgroups.

Our work in the causal setting also complements the rich existing theoretical analyses of recursive adaptive partitioning estimators for regression [[Scornet et al., 2015](#), [Chi et al., 2022](#), [Klusowski and Tian, 2024](#), [Cattaneo et al., 2024](#), [Mazumder and Wang, 2024](#)] and contributes to the small but growing body of negative results. For example, [Ishwaran \[2015\]](#) showed that regression trees via CART methodology [[Breiman et al., 1984](#)] can create imbalanced cells containing a small number of samples. [Tan et al. \[2022\]](#) proved that regression trees are inefficient at estimating additive structure, regardless of the way in which they are optimized. [Tan et al. \[2024b\]](#) proved that mixing times for Bayesian Additive Regression Trees (BART) [[Chipman et al., 2010](#)] can increase with the training sample size. Finally, [Tan et al. \[2024a\]](#) established that adaptive regression trees with Boolean covariates can require exponentially many samples in the dimension and are high-dimensional inconsistent for learning ANOVA decompositions with certain interaction patterns.

The present paper supersedes the unpublished manuscript by [Cattaneo, Klusowski, and Tian \[2022\]](#), which showed that a one-dimensional regression stump (i.e., single-split regression trees with a single covariate) constructed via CART can suffer arbitrarily slow convergence rates, and furthermore conjectured (but did not prove) that causal trees might (i) exhibit the same pathology and (ii) fail to benefit from honesty. Our paper proves both conjectures, and goes further by establishing these results for arbitrary covariate dimension and for any causal tree structure with at least one split (i.e., allowing for an arbitrary number of splits or depth of the causal tree).

The supplemental appendix also reports analogous results for plain adaptive regression trees. As sketched in Section 4.1, with full details given in the supplemental appendix, our method of proof relies on new insights concerning non-asymptotic approximations for the suprema of partial sums and various Gaussian processes, which may be of independent theoretical interest. In particular, we correct an error in Eicker [1979].

2 Setup

The available data $\mathcal{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ is a random sample, where y_i is an outcome variable, $\mathbf{x}_i = (x_{1,1}, \dots, x_{1,p})^\top$ is a vector of (pre-treatment) covariates, and d_i is a binary treatment indicator. Employing standard potential outcomes notation [see, e.g., Hernán and Robins, 2020, for an introduction], we assume that

$$y_i = y_i(1)d_i + y_i(0)(1 - d_i),$$

where $y_i(1)$ is the potential outcome under treatment assignment ($d_i = 1$), and $y_i(0)$ is the potential outcome under control assignment ($d_i = 0$). In classical experimental settings, the treatment assignment mechanism is independent of both the potential outcomes and the covariates, that is, $(y_i(0), y_i(1), \mathbf{x}_i^\top) \perp\!\!\!\perp d_i$.

The parameter of interest is the conditional average treatment effect (CATE) function

$$\tau(\mathbf{x}) \equiv \mathbb{E}[y_i(1) - y_i(0) | \mathbf{x}_i = \mathbf{x}],$$

which captures average treatment effects for different values of observable (pre-treatment) covariates. In experimental settings, the CATE function is identifiable because

$$\tau(\mathbf{x}) = \mathbb{E}[y_i | d_i = 1, \mathbf{x}_i = \mathbf{x}] - \mathbb{E}[y_i | d_i = 0, \mathbf{x}_i = \mathbf{x}] \tag{1}$$

$$= \mathbb{E}\left[y_i \frac{d_i - \xi}{\xi(1 - \xi)} \middle| \mathbf{x}_i = \mathbf{x}\right], \tag{2}$$

where the probability of treatment assignment $\xi = \mathbb{P}(d_i = 1)$ is known by virtue of the known randomization mechanism. The first equality (1) represents $\tau(\mathbf{x})$ as the difference of two conditional expectation functions based on observed data, while the second equality (2) represents $\tau(\mathbf{x})$ as a single conditional expectation of the “transformed” outcome $y_i \frac{d_i - \xi}{\xi(1 - \xi)}$.

Traditional semiparametric methods would replace the unknown conditional expectations by estimators thereof to learn about heterogeneous treatment effects from experimental data. These methods do not cope well with high-dimensional data, sparsity, or other unknown specific features of the data generating process. Motivated by the recent success of modern (adaptive) machine learning methods, Athey and Imbens [2016] proposed to estimate $\tau(\mathbf{x})$ using recursive decision trees. While retaining the core ideas underlying the greedy recursive construction via standard

CART, their proposals customized the tree splitting criterion to the causal inference setting, and employed sample splitting (the so-called “honesty” property) to de-couple the tree construction from the estimation of $\tau(\mathbf{x})$ on the terminal nodes of the tree. This honesty modification has been viewed as a natural “fix,” since separating model selection from estimation is believed to reduce overfitting and improve the validity of inference. Despite this prevailing view, we show that honesty cannot overcome the fundamental limitations of recursive partitioning for heterogeneous causal effect estimation (or for plain adaptive regression trees), offering only at best negligible logarithmic improvements in sample size or dimension.

We perform a comprehensive study of the estimation accuracy of *nine* distinct causal tree methods, which differ on how their three key underlying parts are implemented: (i) *CATE estimator*, (ii) *tree construction*, and (iii) *sample splitting*.

2.1 CATE Estimator

Leveraging the identification results in (1)–(2), [Athey and Imbens \[2016\]](#) considered the following two CATE estimators based on a tree T and a dataset \mathcal{D}_τ . Sections 2.2 and 2.3 discuss specific choices of T and \mathcal{D}_τ , respectively. Let $\mathbf{1}(\cdot)$ be the indicator function.

Definition 1 (CATE Estimators). *Suppose T is the tree used, and $\mathcal{D}_\tau = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\tau\}$, with $n_\tau \leq n$, is the dataset used. Let \mathbf{t} be the unique terminal node in T containing $\mathbf{x} \in \mathcal{X}$.*

- *The Difference-in-Means (DIM) estimator is*

$$\hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau) = \frac{1}{n_1(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} d_i y_i - \frac{1}{n_0(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} (1 - d_i) y_i,$$

where $n_d(\mathbf{t}) = \sum_{i=1}^{n_\tau} \mathbf{1}(\mathbf{x}_i \in \mathbf{t}, d_i = d)$, for $d = 0, 1$, are the “local” sample sizes. We set $\hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau) = 0$ whenever $n_0(\mathbf{t}) = 0$ or $n_1(\mathbf{t}) = 0$.

- *The Inverse Probability Weighting (IPW) estimator is*

$$\hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau) = \frac{1}{n(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} \frac{d_i - \xi}{\xi(1 - \xi)} y_i,$$

where $n(\mathbf{t}) = n_0(\mathbf{t}) + n_1(\mathbf{t}) = \sum_{i=1}^{n_\tau} \mathbf{1}(\mathbf{x}_i \in \mathbf{t})$ is the “local” sample size. We set $\hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau) = 0$ whenever $n(\mathbf{t}) = 0$.

Both estimators, $\hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau)$ and $\hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau)$, rely on localization near \mathbf{x} via the tree construction: T forms a partition of the support of the covariates \mathcal{X} , and estimation of $\tau(\mathbf{x})$ uses only observations with covariates \mathbf{x}_i belonging to the cell in the partition covering $\mathbf{x} \in \mathcal{X}$. Therefore, given a tree (or partition), both estimators can be represented as nonparametric partitioning-based estimates of $\tau(\mathbf{x})$. See [Györfi et al. \[2002\]](#), [Cattaneo et al. \[2020\]](#), [Cattaneo et al. \[2025\]](#), and references therein.

Since the estimators $\hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}, \mathcal{D}_\tau)$ and $\hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbf{T}, \mathcal{D}_\tau)$ output a constant fit for all \mathbf{x} within each terminal node of \mathbf{T} (or cell in the partition), we define

$$\hat{\tau}_l(\mathbf{t}; \mathbf{T}, \mathcal{D}_\tau) = \hat{\tau}_l(\mathbf{x}; \mathbf{T}, \mathcal{D}_\tau), \quad l \in \{\text{DIM}, \text{IPW}\}, \quad \mathbf{x} \in \mathbf{t},$$

for all terminal nodes \mathbf{t} of \mathbf{T} .

2.2 Tree Construction

An axis-aligned recursive decision tree is a predictive model that makes decisions by repeatedly splitting the data into subsets based on both outcome and covariate values. At each node, the algorithm selects the feature and threshold that best separate the data according to some criterion (e.g., squared error, Gini impurity, or entropy), and this process continues recursively until a stopping condition is met (e.g., maximum depth or pure terminal nodes). See Berk [2020], Zhang and Singer [2010], and references therein.

The most popular implementation of recursive decision trees is via the CART algorithm, which proceeds in a top-down, greedy manner through recursive binary splitting. Given a dataset $\mathcal{D}_\mathbf{T} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\mathbf{T}\}$, with $n_\mathbf{T} \leq n$, a parent node \mathbf{t} in the tree (i.e., a region in \mathcal{X}) is divided into two child nodes, \mathbf{t}_L and \mathbf{t}_R , by minimizing the sum-of-squares error (SSE),

$$\min_{1 \leq j \leq p} \min_{\beta_\text{L}, \beta_\text{R}, \varsigma \in \mathbb{R}} \sum_{\mathbf{x}_i \in \mathbf{t}} (y_i - \beta_\text{L} \mathbf{1}(x_{ij} \leq \varsigma) - \beta_\text{R} \mathbf{1}(x_{ij} > \varsigma))^2, \quad (3)$$

where the solution yields estimates $(\hat{\beta}_\text{L}, \hat{\beta}_\text{R}, \hat{\varsigma}, \hat{j})$, being the two child nodes average output, split point and split direction, respectively. Because the splits occur along values of a single covariate, the induced partition of the input space \mathcal{X} is a collection of hyper-rectangles, and hence the resulting refinement of \mathbf{t} produces child nodes $\mathbf{t}_\text{L} = \{\mathbf{x} \in \mathbf{t} : \mathbf{e}_j^\top \mathbf{x} \leq \hat{\varsigma}\}$ and $\mathbf{t}_\text{R} = \{\mathbf{x} \in \mathbf{t} : \mathbf{e}_j^\top \mathbf{x} > \hat{\varsigma}\}$. More precisely, the normal equations imply that $\hat{\beta}_\text{L} = \frac{1}{n(\mathbf{t}_\text{L})} \sum_{\mathbf{x}_i \in \mathbf{t}_\text{L}} y_i$ and $\hat{\beta}_\text{R} = \frac{1}{n(\mathbf{t}_\text{R})} \sum_{\mathbf{x}_i \in \mathbf{t}_\text{R}} y_i$, the respective sample means after splitting the parent node at $\mathbf{e}_j^\top \mathbf{x} = \hat{\varsigma}$. These child nodes become new parent nodes at the next level of the tree construction, and can be further refined in the same manner, and so on and so forth, until a desired depth K is reached. While not every parent node needs to generate a new child node in a recursive tree construction, a maximal decision tree of depth K is a particular instance where the construction is iterated K times until (i) the node contains a single data point (y_i, \mathbf{x}_i^\top) or (ii) all input values \mathbf{x}_i and/or all response values y_i within the node are the same.

Building on the CART algorithm, Athey and Imbens [2016] proposed the following two custom criteria for constructing a tree \mathbf{T} to implement their causal tree estimators.

Definition 2 (Tree Construction). *Suppose $\mathcal{D}_\mathbf{T} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\mathbf{T}\}$, with $n_\mathbf{T} \leq n$, is the dataset used to construct the tree \mathbf{T} . There is a unique node $\mathbf{t}_0 = \mathcal{X}$ at initialization, and child nodes are generated by iterative axis-aligned splitting of the parent node based on either of the following two rules.*

- *Variance Maximization:* A parent node \mathbf{t} (i.e., a terminal node partitioning \mathcal{X}) in a previous tree \mathbf{T}' is divided into two child nodes, \mathbf{t}_L and \mathbf{t}_R , forming the new tree \mathbf{T} , by maximizing

$$\frac{n(\mathbf{t}_L)n(\mathbf{t}_R)}{n(\mathbf{t})} \left(\hat{\tau}_l(\mathbf{t}_L; \mathbf{T}, \mathcal{D}_\mathbf{T}) - \hat{\tau}_l(\mathbf{t}_R; \mathbf{T}, \mathcal{D}_\mathbf{T}) \right)^2, \quad l \in \{\text{DIM}, \text{IPW}\}. \quad (4)$$

Assuming at least one split, the two final causal trees are denoted by $\mathbf{T}^{\text{DIM}}(\mathcal{D}_\mathbf{T})$ and $\mathbf{T}^{\text{IPW}}(\mathcal{D}_\mathbf{T})$, respectively.

- *SSE Minimization:* A parent node \mathbf{t} (i.e., a terminal node partitioning \mathcal{X}) in the previous tree \mathbf{T}' is divided into two child nodes, \mathbf{t}_L and \mathbf{t}_R , forming the next tree \mathbf{T} , by solving

$$\min_{a_L, b_L, a_R, b_R \in \mathbb{R}} \sum_{\mathbf{x}_i \in \mathbf{t}_L} (y_i - a_L - b_L d_i)^2 + \sum_{\mathbf{x}_i \in \mathbf{t}_R} (y_i - a_R - b_R d_i)^2, \quad (5)$$

where only the data $\mathcal{D}_\mathbf{T}$ is used. Assuming at least one split, the final causal tree is denoted by $\mathbf{T}^{\text{SSE}}(\mathcal{D}_\mathbf{T})$.

The variance maximization splitting criterion is somewhat different than the original CART criteria (3), since it explicitly selects splits based on maximizing the squared difference of the child treatment effect estimates. For the IPW estimator, this rule is equivalent to applying the CART criterion in (3) to the transformed outcome $\tilde{y}_i = y_i \frac{d_i - \xi}{\xi(1 - \xi)}$. This transformation satisfies $\mathbb{E}[\tilde{y}_i \mid \mathbf{x}_i = \mathbf{x}] = \tau(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, and thus CART operates on an outcome whose conditional mean equals the CATE. The DIM estimator follows the same idea of predicting the within-node average treatment effect, but it constructs these predictions somewhat differently.

The SSE Minimization criterion resembles the original CART criteria (3), but its formulation still targets treatment effect heterogeneity as the splitting criteria: in Section SA-3.3 of the supplemental appendix we show that the objective function (5) can be recast as maximization of the sum of variances of treatment and control group outcomes given by

$$\begin{aligned} & \frac{n_1(\mathbf{t}_L)n_1(\mathbf{t}_R)}{n_1(\mathbf{t})} \left(\frac{1}{n_1(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} d_i y_i - \frac{1}{n_1(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} d_i y_i \right)^2 \\ & + \frac{n_0(\mathbf{t}_L)n_0(\mathbf{t}_R)}{n_0(\mathbf{t})} \left(\frac{1}{n_0(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} (1 - d_i) y_i - \frac{1}{n_0(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} (1 - d_i) y_i \right)^2. \end{aligned}$$

Each of the causal recursive tree constructions leads to a distinct data-driven partition of \mathcal{X} . A key observation in this paper is that they do not generate quasi-uniform partitions, and thus known results in the nonparametric partitioning-based estimation literature [Györfi et al., 2002, Cattaneo et al., 2020, 2025] are not applicable. The supplemental appendix considers other recursive partitioning constructions, including the standard CART algorithm and variants thereof.

2.3 Sample Splitting

The final ingredient of the causal tree estimators concerns the data used at each stage of their construction. It is believed that de-coupling the CATE estimation (Definition 1) from the tree implementation (Definition 2) can lead to better performance of the final estimator. In practice, this approach corresponds to sample splitting, and [Athey and Imbens \[2016\]](#) and others referred to it as “honesty.” To avoid confusion, we emphasize that procedures without sample splitting are not “dishonest” in any formal sense; they are simply harder to analyze formally.

To elucidate the relative merits of sample splitting, we consider two distinct scenarios: (i) no sample splitting, where the same data is used throughout (as the original CART procedure is often implemented); and (ii) honesty, where two independent datasets are used, one for tree construction and the other for CATE estimation (these are the procedures proposed by [Athey and Imbens \[2016\]](#) and many others). Formally, we consider the following data usages and resulting treatment effect estimators.

Definition 3 (Sample Splitting and Estimators). *Recall Definition 1 and Definition 2, and that $\mathcal{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ is the available random sample.*

- *No Sample Splitting (NSS): The dataset \mathcal{D} is used for both the tree construction and the treatment effect estimation, that is, $\mathcal{D}_\top = \mathcal{D}$ and $\mathcal{D}_\tau = \mathcal{D}$. The causal tree estimators are*

$$\begin{aligned}\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{DIM}}(\mathcal{D}), \mathcal{D}), \\ \hat{\tau}_{\text{IPW}}^{\text{NSS}}(\mathbf{x}) &= \hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbf{T}^{\text{IPW}}(\mathcal{D}), \mathcal{D}), \quad \text{and} \\ \hat{\tau}_{\text{SSE}}^{\text{NSS}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{SSE}}(\mathcal{D}), \mathcal{D}).\end{aligned}$$

- *Honesty (HON): The dataset \mathcal{D} is divided in two independent datasets \mathcal{D}_\top and \mathcal{D}_τ with sample sizes n_\top and n_τ , respectively, and satisfying $n \lesssim n_\top, n_\tau \lesssim n$. The causal tree estimators are*

$$\begin{aligned}\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{DIM}}(\mathcal{D}_\top), \mathcal{D}_\tau), \\ \hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}) &= \hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbf{T}^{\text{IPW}}(\mathcal{D}_\top), \mathcal{D}_\tau), \quad \text{and} \\ \hat{\tau}_{\text{SSE}}^{\text{HON}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{SSE}}(\mathcal{D}_\top), \mathcal{D}_\tau).\end{aligned}$$

The no-sample-splitting and honesty data usages are commonly encountered in the literature, and thus our results will speak directly to theoretical, methodological and empirical work relying on these sample splitting designs. While the estimators $\hat{\tau}_l^{\text{NSS}}(\mathbf{x})$ and $\hat{\tau}_l^{\text{HON}}(\mathbf{x})$, $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$, depend on the depth of the tree construction used, our notation does not make this dependence explicit because our results apply whenever at least one split takes place. See Section 5 for more discussion, and a setting where the number of splits is assumed to increase with the sample size.

3 Assumptions

We impose the following assumption throughout the paper.

Assumption 1 (Data Generating Process). $\mathcal{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : 1 \leq i \leq n\}$ is a random sample, where $y_i = d_i y_i(1) + (1 - d_i) y_i(0)$, $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^\top$, and the following conditions hold for all $d = 0, 1$ and $i = 1, 2, \dots, n$.

- (i) $(y_i(0), y_i(1), \mathbf{x}_i) \perp\!\!\!\perp d_i$, and $\xi = \mathbb{P}(d_i = 1) \in (0, 1)$.
- (ii) $y_i(d) = \mu_d(\mathbf{x}_i) + \varepsilon_i(d)$, with $\mathbb{E}[\varepsilon_i(d) | \mathbf{x}_i] = 0$ and $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i(d)$.
- (iii) $\mu_d(\mathbf{x}) = c_d$ for all $\mathbf{x} \in \mathcal{X}$, where c_d is some constant and \mathcal{X} is the support of \mathbf{x}_i .
- (iv) $x_{i,1}, \dots, x_{i,p}$ are independent and continuously distributed.
- (v) There exists $\alpha > 0$ such that $\mathbb{E}[\exp(\lambda \varepsilon_i(d))] < \infty$ for all $|\lambda| < 1/\alpha$ and $\mathbb{E}[\varepsilon_i^2(d)] > 0$.

Assumption 1(i) corresponds to simple randomized experiments. Assumption 1(ii) further assumes a canonical homoskedastic causal regression model, while Assumption 1(iii) implies that there is no heterogeneity in the causal treatment effect $\tau = c_1 - c_0$. Because trees are invariant with respect to monotone transformations of the coordinates of \mathbf{x}_i , without loss of generality, Assumption 1(iv) can be replaced by the assumption that covariates are uniformly distributed on $\mathcal{X} = [0, 1]^p$, i.e., $x_{i,j} \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}([0, 1])$ for $j = 1, 2, \dots, p$. Finally, Assumption 1(v) means that potential outcome errors are sub-exponential, or equivalently, they satisfy a Bernstein moment condition.

Since we are interested in establishing lower bounds on the estimation accuracy of the causal tree estimators in Definition 3, it is sufficient to consider the constant treatment effect model in Assumption 1 for several reasons. First, this statistical model is a canonical member of any interesting class of data generating processes because the constant function belongs to all classical smoothness function classes, as well as to the set of functions with bounded total variation. It follows that our results will shed light in settings where uniformity over any of the aforementioned classes of functions is of interest: our lower bounds can be applied directly in those cases because for any estimator $\hat{\tau}(\mathbf{x})$ of the parameter $\tau(\mathbf{x})$,

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| > \epsilon \right) \geq \mathbb{P}_1 \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| > \epsilon \right),$$

for all $\epsilon > 0$, and for any data generating class \mathcal{P} that includes the distribution \mathbb{P}_1 satisfying Assumption 1. In fact, the constant treatment effect model is a canonical case to consider in causal inference.

Second, Assumption 1 also removes issues related to smoothing (or misspecification) bias, heteroskedasticity, and heavy tail distributions. In particular, since the CATE function $\tau(\mathbf{x})$ is constant

for all $\mathbf{x} \in \mathcal{X}$, our results will not be driven by standard (boundary or other smoothing) bias in nonparametrics. For example, if the distributions of $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are symmetric about zero,

$$\mathbb{E}[\hat{\tau}_l^q(\mathbf{x})] = \tau, \quad q \in \{\text{NSS}\}, \quad \text{and} \quad \mathbb{E}[\hat{\tau}_l^{\text{HON}}(\mathbf{x})] = \tau - \tau \mathbb{P}(n(\mathbf{t}) = 0),$$

for $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ and $\mathbf{x} \in \mathbf{t}$ where \mathbf{t} is a terminal node in the tree. Unbiasedness of $\hat{\tau}_l^{\text{NSS}}(\mathbf{x})$ follows from the fact that the split points are symmetric functions of the residuals. In the case of $\hat{\tau}_l^{\text{HON}}(\mathbf{x})$, sample splitting can generate empty cells with positive probability, which is captured by the term $\tau \mathbb{P}(n(\mathbf{t}) = 0)$; see Lemma SA-37 in the supplemental appendix. It follows that, in particular, $\hat{\tau}_l^{\text{HON}}(\mathbf{x})$ is unbiased when $\tau = 0$ (or for any other known treatment effect value), as well as in tree constructions ensuring that $\mathbb{P}(n(\mathbf{t}) = 0) = 0$; otherwise, $\hat{\tau}_l^{\text{HON}}(\mathbf{x})$ is asymptotically unbiased whenever $\mathbb{P}(n(\mathbf{t}) = 0) \rightarrow 0$ as $n \rightarrow \infty$. Our results will be driven by the fact that canonical adaptive decision tree constructions can generate small cells containing only a handful of observations, thereby making the estimator highly inaccurate in some regions of \mathcal{X} , regardless of bias. In other words, inconsistency is due to a large variance problem, not a large bias problem.

Third, the local constant treatment effect model could also be interpreted as a first-order approximation of the smooth function $\tau(\mathbf{x})$. Because the recursive partitioning schemes lead to a partitioning-based estimator of the CATE function, it follows that $\tau(\mathbf{x})$ is approximated locally by a Haar basis (piecewise constant functions). In fact, our results can be extended to hold uniformly over appropriate shrinking neighborhoods of smooth functions local to the constant function, provided that the signal to noise ratio (bias-variance trade-off) is small.

4 Main Results

The following theorem summarizes our first main result. Let e denote Euler's constant.

Theorem 1 (Uniform Accuracy). *Suppose Assumption 1 holds, and the underlying causal tree has at least one split (i.e., at least two terminal nodes). Then, for $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ and all $b \in (0, 1)$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_l^{\text{NSS}}(\mathbf{x}) - \tau(\mathbf{x})| \geq C_1 n^{-b/2} \sqrt{\log \log n} \right) \geq b/e,$$

where the positive constant C_1 only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_l^{\text{HON}}(\mathbf{x}) - \tau(\mathbf{x})| \geq C_2 n^{-b/2} \right) \geq C_3 b,$$

where the positive constants C_2 and C_3 only depend on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and the sample splitting scheme via $\liminf_{n \rightarrow \infty} \frac{n_{\tau}}{n}$ and $\limsup_{n \rightarrow \infty} \frac{n_{\tau}}{n}$. The precise definitions of the constants are given in the supplemental appendix.

Section 4.1 gives an overview of the proof strategy of Theorem 1, with all omitted technical details given in the supplemental appendix (see Section SA-1.2 for details). Our proof relies on

several non-asymptotic approximation steps for the suprema of partial sums and various Gaussian processes leveraging key technical results from Chernozhukov et al. [2017], Chernozhukov et al. [2022], Csörgö and Révész [1981], Csörgö and Horváth [1997], Eicker [1979], El-Yaniv and Pechyony [2009], Göing-Jaesche and Yor [2003], Horváth [1993], Latała and Matlak [2017], Petrov [2007], Shorack and Smythe [1976], and Skorski [2023]. As a technical by-product, we correct a mistake in Eicker [1979]: see Remark SA-1 in the supplemental appendix.

Theorem 1 presents precise lower bounds on the uniform convergence rate of the six causal tree estimators introduced in Section 2. Starting with procedures that do not employ sample splitting, Theorem 1 demonstrates that the three estimators $\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x})$, $\hat{\tau}_{\text{IPW}}^{\text{NSS}}(\mathbf{x})$ and $\hat{\tau}_{\text{SSE}}^{\text{NSS}}(\mathbf{x})$ cannot achieve a uniform convergence rate of $n^{-b/2}\sqrt{\log \log n}$, for any $b > 0$. That is, they must have a worse than polynomial-in- n uniform convergence rate, and thus suffer from low accuracy in estimating heterogeneous treatment effects in certain regions of the support \mathcal{X} .

Athey and Imbens [2016], and many others, argue that sample splitting (the so-called “honesty” property) can improve the performance of machine learning estimators, and in particular their proposed causal tree estimators, because such sample usage de-couples the causal tree construction and the CATE estimation steps. The second result in Theorem 1 considers exactly their honest causal tree estimators, $\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x})$, $\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x})$ and $\hat{\tau}_{\text{SSE}}^{\text{HON}}(\mathbf{x})$. It follows from the theorem that these estimators cannot achieve a uniform convergence rate that is polynomial-in- n either. Notably, our results show that sample splitting (or honesty) improves the best achievable uniform convergence rate of the estimators, but this improvement is quite modest: the penalty term $\sqrt{\log \log n}$ is removed, thereby improving the uniform convergence rate by a very slow factor.

The results in Theorem 1 offer a pessimistic outlook on the utility of adaptive decision tree methods in causal inference when the goal is to learn about heterogeneous treatment effects: the estimators cannot perform well pointwise (and hence uniformly) over the entire support of the covariates; see Section 4.1 for more formal details. As a point of contrast, the same procedures considered in Theorem 1 can achieve near-optimal convergence rates “on average” over \mathcal{X} , as the following theorem establishes. Here again, honesty delivers only negligible improvements of order $\log(p)$.

Theorem 2 (Mean Square Accuracy). *Suppose Assumption 1 holds and the underlying causal tree has depth at most $K \geq 1$, and let $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{x}_i \leq \mathbf{x})$. Then, for $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$,*

$$\mathbb{E} \left[\int_{\mathcal{X}} |\hat{\tau}_l^{\text{NSS}}(\mathbf{x}) - \tau(\mathbf{x})|^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C_1 \frac{2^K \log^4(n) \log(np)}{n},$$

where the constant C_1 only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and

$$\mathbb{E} \left[\int_{\mathcal{X}} |\hat{\tau}_l^{\text{HON}}(\mathbf{x}) - \tau(\mathbf{x})|^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C_2 \frac{2^K \log^5(n)}{n},$$

provided that $\rho \leq n_{\top}/n_{\tau} \leq 1 - \rho$ for some $\rho \in (0, 1)$, and the constant C_2 only depends on ρ and the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

The proof of this theorem is given in the supplemental appendix (see Section SA-1.2 for details). It leverages ideas and technical results from Györfi et al. [2002] and Klusowski and Tian [2024]. Crucially, the result applies only when Assumption 1 holds, that is, when $\tau(\mathbf{x})$ is constant. The main purpose of Theorem 2 is to demonstrate that in the same basic setting when uniform convergence fails, causal decision trees nonetheless achieve favorable performance on average in an integrated mean-squared sense. A natural way to interpret the juxtaposition between Theorem 1 and Theorem 2 is related to the often claimed tension between causal inference and prediction in machine learning settings: adaptive causal trees can perform poorly pointwise (hence uniformly), but excellently on average, over the feature space.

From a technical perspective, the results in Theorem 2 are new in the context of causal tree estimation and, notably, for the formal comparison between no-sample-splitting and honest implementations. Furthermore, our theoretical work in the supplemental appendix establishes the integrated mean-squared error bounds with high-probability, enabling a sharper comparison with Theorem 1. For example, for the case of no-sample-splitting, we show that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} |\hat{\tau}_l^{\text{NSS}}(\mathbf{x}) - \tau(\mathbf{x})|^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C_1 \frac{2^K \log^4(n) \log(np)}{n} \right) = 0,$$

where C_1 is the constant in Theorem 2.

4.1 Proof Strategy of Theorem 1

Underlying our theoretical insights are a collection of technical results concerning a decision stump, and hence a decision tree of depth one. For each tree splitting criteria and sample splitting design, we first study the probabilistic properties of the split location at the root node, and thus characterize the regions of the support \mathcal{X} where the first split index is most likely to realize. These theoretical results also characterize the effective sample size of the resulting child nodes. We establish that with non-vanishing probability, the first split will concentrate near a region of the boundary of the parent node (a cell in the partition of \mathcal{X}), from the beginning of any tree construction. More precisely, let $\hat{i} = n(\mathbf{t}_L)$ and \hat{j} be the CART split index and split variable at the root node, respectively, with $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$, noticing that the first split coincide for no-sample-splitting and honest constructions. For each $a, b \in (0, 1)$ with $a < b$ and $j \in \{1, 2, \dots, p\}$, and $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b, \hat{j} = j) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i} \leq n - n^a, \hat{j} = j) \geq \frac{b - a}{2pe}. \quad (6)$$

The slow uniform convergence rate of a decision stump estimator occurs because the optimal split point concentrates near the boundary of the support, causing the two nodes in the stump to be imbalanced, with one containing a much smaller number of samples, and therefore rendering a situation where local averaging is less accurate. This can be deduced from (6): for each coordinate $j = 1, 2, \dots, p$ and $b \in (0, 1)$, there is non-vanishing $b/(pe)$ probability that the child cells $\{\mathbf{x} \in \mathcal{X} :$

$x_j \leq \hat{\varsigma}\}$ or $\{\mathbf{x} \in \mathcal{X} : x_j > \hat{\varsigma}\}$ are highly anisotropic and will contain at most n^b samples. Thus, with non-vanishing probability, the causal tree procedures will exhibit arbitrarily slow convergence rate in a region of \mathcal{X} . These results are then carefully recycled to characterize the properties of the deeper trees: due to their recursive nature, and since $p > 1$, the problematic regions take the form of many hyper-rectangles, and will realize anywhere in \mathcal{X} , with non-vanishing probability.

The core of proof strategy is to study the tree construction as the maximizer of the split criterion from (4) and (5), as indexed by the optimal split location and covariate coordinate. We leverage non-asymptotic high-dimensional central limit theorems, Gaussian comparison inequalities, Gaussian process embeddings, the Darling-Erdős theorem, and empirical process techniques [El-Yaniv and Pechyony, 2009, Petrov, 2007, Shorack and Smythe, 1976, Skorski, 2023], as explained in the following four main steps.

Step 1: Split Criterion Approximation. Using empirical process theory techniques, we establish an asymptotic equivalence between the split criterion underlying each of the causal tree estimators and the split criterion of a standard (non-causal) decision regression tree employing CART. For $l = \text{DIM}$ and $l = \text{IPW}$, the latter can be viewed as a standard regression tree with transformed outcomes $y_i \frac{d_i - \xi}{\xi(1-\xi)}$. For $l = \text{SSE}$, approximating process is the sum of two independent split criterion processes, one with transformed outcome $\frac{d_i}{\xi} y_i$ for treated units, and the other with transformed outcome $\frac{1-d_i}{1-\xi} y_i$ for control units. We employ a careful truncation argument to remove extremely small or large split indices [Csörgő and Horváth, 1997, Theorem A.4.1], where empirical process techniques are hard to apply.

Step 2: Conditional Gaussian Approximation. We show that, conditional on the covariates ordering, the square root of the split criterion processes from step 1 can be approximated by Gaussian processes with the same conditional covariance structure. For $l = \text{DIM}$ and $l = \text{IPW}$, we view the split criterion process as a summation of i.i.d. high-dimensional random vectors, each entry corresponding to one pair of split index and coordinate. The high-dimensional central limit theorem of [Chernozhukov et al., 2017, Theorem 2.1] implies that the split criterion process in high-dimensional vector form is close to a high-dimensional Gaussian random vector with the same covariance matrix conditional the ordering, the latter can then be interpreted as a Gaussian process conditional on the ordering. Due to the structure of the splitting criteria, a high-dimensional CLT for hyper-rectangles is sufficient. For $l = \text{SSE}$, we stack the control and treatment groups process in a twice as long high-dimensional vector. However, due to the structure the splitting criteria in this case, we employ instead Chernozhukov et al. [2017, Proposition 3.1], which gives a high-dimensional CLT for convex sets.

Step 3: Unconditional Gaussian Approximation. For the special case of $p = 1$, this step is not necessary because there is only one ordering possible. However, for $p > 1$, recursive decision trees find the best split along each dimension of \mathbf{x}_i , which implies a different ordering of the vector. Nevertheless, we show that the conditional Gaussian process from step 2 is close to an unconditional Gaussian process with zero correlation for different split coordinate indexes. Zero correlation between splits of different coordinates implies that the (sub)-processes corresponding to

splitting different coordinates are asymptotically independent, reducing the problem to studying the arg max of the split criterion over one coordinate. The result is proven by applying a Gaussian-to-Gaussian comparison inequality [Chernozhuokov et al., 2022, Proposition 2.1], after establishing an upper bound on the matrix max norm of the difference between the conditional covariance matrix (which depends on the ordering) and the unconditional covariance matrix (which does not depend on the ordering). For $l = \text{DIM}$ and $l = \text{IPW}$, the results is immediate because the high-dimensional CLT was established over hyper-rectangles. For $l = \text{SSE}$, the additional error induced by considering a simple convex sets approximation is be controlled using Nazarov’s inequality [Nazarov, 2003].

Step 4: Lower bound on imbalanced split probability. The unconditional Gaussian approximation processes from Step 3 take the form of the square Euclidean norm of a univariate (for $l \in \{\text{DIM}, \text{IPW}\}$) or bivariate (for $l = \text{SSE}$) Ornstein-Uhlenbeck process, where the split and time of Ornstein-Uhlenbeck process satisfies a one-to-one transformation [Csörgö and Révész, 1981, Göing-Jaesche and Yor, 2003]. Since Darling-Erdös [Eicker, 1979, Horváth, 1993] allows for calculation of the maximum of norm of an O-U process within any time interval, we can find the lower bound on the probability of split occurs with a small or large index from (6) with the help of Gaussian correlation inequality [Latała and Matlak, 2017, Remark 3 (i)]. In turn, this characterizes precisely the effective sample sizes of each child node.

The remaining of our proofs leverage the technical insights above, applying then recursively to understand deeper tree constructions and the concentration in probability properties of the resulting CATE estimates.

5 X-Adaptivity and Inconsistency

The estimators considered in Theorem 1 either employ the full sample in their entire construction, or they rely on a two-sample independent split (honesty), where one subsample is use for training the tree, and the other is used for estimation of the conditional average treatment effects. As discussed in Devroye et al. [2013], and references therein, **X**-adaptivity offers a middle ground between the two sample usage designs considered in Definition 2: the tree construction and the final estimation step share the same covariates but each step employs different outcomes variables, that is, the two subsamples are independence conditional on the covariates.

We leverage the idea of **X**-Adaptivity, and study causal tree estimators where the outcome variable and treatment indicator are independent across all levels of the tree construction and the final CATE estimation step, but the same covariates are used throughout. This **X**-adaptive data design is of theoretical interest because it offers a bridge between no-sample-splitting and honesty. The following definition formalizes the construction of the **X**-adaptive causal tree estimators.

Definition 4 (**X**-Adaptive Estimation). *Recall Definition 1 and Definition 2, and that $\mathcal{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ is the available random sample.*

1. *The dataset \mathcal{D} is divided into $K + 1$ datasets $(\mathcal{D}_{T_1}, \dots, \mathcal{D}_{T_K}, \mathcal{D}_\tau)$, with sample sizes given by $(n_{T_1}, \dots, n_{T_K}, n_\tau)$, respectively, and satisfying $n_{T_1} = \dots = n_{T_K} = n_\tau$ (possibly after*

dropping $n \bmod K$ data points at random). For each of the datasets $\mathcal{D}_j = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, \dots, n_{T_j}\}$, $j = 1, \dots, K$, replace $\{(y_i, d_i) : i = 1, \dots, n_{T_j}\}$ with independent copies $\{(\tilde{y}_i, \tilde{d}_i) : i = 1, \dots, n_{T_j}\}$, while keeping the same $\{\mathbf{x}_i : i = 1, \dots, n_{T_j}\}$.

2. The maximal decision tree of depth K , $T_K^l(\mathcal{D}_{T_1}, \dots, \mathcal{D}_{T_K})$, is obtained by iterating K times the $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ splitting procedures in Definition 2, each time splitting all terminal nodes until (i) the node contains a single data point $(y_i, d_i, \mathbf{x}_i^\top)$, or (ii) the input values \mathbf{x}_i and/or all (d_i, y_i) within the node are the same.

3. The \mathbf{X} -adaptive estimators are

$$\begin{aligned}\hat{\tau}_{\text{DIM}}^{\mathbf{X}}(\mathbf{x}; K) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; T_K^{\text{DIM}}(\mathcal{D}_{T_1}, \dots, \mathcal{D}_{T_K}), \mathcal{D}_\tau), \\ \hat{\tau}_{\text{IPW}}^{\mathbf{X}}(\mathbf{x}; K) &= \hat{\tau}_{\text{IPW}}(\mathbf{x}; T_K^{\text{IPW}}(\mathcal{D}_{T_1}, \dots, \mathcal{D}_{T_K}), \mathcal{D}_\tau), \quad \text{and} \\ \hat{\tau}_{\text{SSE}}^{\mathbf{X}}(\mathbf{x}; K) &= \hat{\tau}_{\text{SSE}}(\mathbf{x}; T_K^{\text{SSE}}(\mathcal{D}_{T_1}, \dots, \mathcal{D}_{T_K}), \mathcal{D}_\tau).\end{aligned}$$

As in the previous cases, if the distributions of $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are symmetric about zero, then the \mathbf{X} -adaptive estimators are unbiased: $\mathbb{E}[\hat{\tau}_l^{\mathbf{X}}(\mathbf{x}; K)] = \tau$, for $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$.

Theorem 3 (Accuracy of \mathbf{X} -Adaptive Causal Tree Estimators). *Suppose Assumption 1 holds and additionally that $\mathbb{E}[\varepsilon_i^2(0)] = \mathbb{E}[\varepsilon_i^2(1)]$. Then, for $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P}\left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_l^{\mathbf{X}}(\mathbf{x}; K_n) - \tau(\mathbf{x})| \geq C_1\right) \geq C_2,$$

provided that $\liminf_{n \rightarrow \infty} \frac{K_n}{\log \log n} = \kappa > 0$, and where the positive constants C_1 and C_2 only depend on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$ and κ .

Furthermore, for $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ and any $K \geq 1$,

$$\mathbb{E}\left[\int_{\mathcal{X}} (\hat{\tau}_l^{\mathbf{X}}(\mathbf{x}, K) - \tau(\mathbf{x}))^2 dF_{\mathbf{X}}(\mathbf{x})\right] \leq C_3 \frac{K 2^K}{n},$$

where the positive constant C_3 only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

The theorem establishes uniform inconsistency of the \mathbf{X} -adaptive causal tree estimator so long as $K_n \gtrsim \log \log n$. To put this side rate restriction in perspective, if $n/K_n \approx 1$ billion then $\log \log(10^9) \approx 3$. Therefore, the inconsistency of the estimator will manifest as soon as $K_n \approx 3$, a shallow tree when compared to those commonly encountered in practice (even in settings with much more moderate sample sizes, that is, with n much smaller than K_n billions). This result also shows that the integrated mean square error (IMSE) of a uniformly inconsistent \mathcal{X} -adaptive causal tree estimator can nonetheless decay at the optimal \sqrt{n} rate, up to a poly-logarithmic- n factor. As demonstrated before, the performance of the causal tree estimators can vary widely depending on whether the input \mathbf{x} is average or worst case.

6 Discussion

6.1 Decision Stumps

The phenomenon of generating unbalanced cells in adaptive recursive partitioning schemes has been observed in various forms since the inception of CART. Historically, this phenomenon has been called the *end-cut preference*, where splits along noisy directions tend to concentrate along the boundary of the parent node. More specifically, considering the standard CART for regression estimation without sample splitting, Breiman et al. [1984, Theorem 11.1] and Ishwaran [2015, Theorem 4] showed that in one-dimension ($p = 1$), for each $\delta \in (0, 1)$, $\mathbb{P}(n(\mathbf{t}_L) \leq \delta n \text{ or } n(\mathbf{t}_R) \geq (1 - \delta)n) \rightarrow 1$ as $n \rightarrow \infty$. If applicable to the context of this paper, their result would only imply rates in uniform norm slower than any *constant multiple* of the already nearly optimal rate $\sqrt{n/\log \log(n)}$, i.e., for any $C > 0$,

$$\liminf_{n \rightarrow \infty} \mathbb{P}\left(\sup_{x \in \mathcal{X}} |\hat{\tau}_l^{\text{NSS}}(x) - \tau(x)| \geq C\sigma n^{-1/2} \sqrt{\log \log(n)}\right) = 1.$$

In contrast, our results hold for all $p \geq 1$ and characterize precisely the regions of the support \mathcal{X} where the pointwise rates of estimation are slower than any polynomial-in- n (see Corollary SA-7, Theorem SA-14, Corollary SA-21 in the supplemental appendix). Thus, past theoretical work is not strong enough to illustrate the weaknesses of causal trees for pointwise estimation (i.e., prior lower bounds in the literature would be too loose to be informative). Furthermore, our results also study settings where sample splitting (honesty) is used, and demonstrate that they cannot mitigate the low convergence rate of adaptive causal trees under Assumption 1. Last but not least, our results apply to the causal tree constructions which are different (and more complicated) than those in plain vanilla CART regression (Definition 2).

6.2 Deeper Trees, Multivariate Covariates, and the Location of Small Cells

Our theoretical results show that, under Assumption 1, the first split of any decision tree construction will generate a small child cell with non-vanishing probability. As a result, and due to their recursive nature, deeper tree constructions will have multiple regions with too small sample sizes (with non-vanishing probability). This problem is exacerbated in multiple dimensions ($p > 1$), which is exactly the setting where causal tree estimators would be potentially more useful to uncover treatment effect heterogeneity.

The small regions of the support \mathcal{X} , and hence the slower than any polynomial-in- n convergence rate (or inconsistency) of causal tree estimators, need not occur near a region of the boundary of \mathcal{X} . At each stage in the tree construction, a parent node \mathbf{t} will generate two child nodes, one small and the other large, but the splitting may realize anywhere on \mathbf{t} (parent cell) and along any individual covariate (in \mathbf{x}_i , or axis), thereby generating problematic hyper-rectangle cells all over the support \mathcal{X} with non-vanishing probability.

6.3 Regularization and Bias

It is tempting to try to regularize the decision tree estimator in order to eliminate the small cell problem, and thus improve its convergence rate. For instance, the tree construction algorithm may not split a parent node if the effective sample size is too small, or it may include a penalty term for overfitting. However, it is also important to note that adaptive decision tree constructions purposely select small cells for two opposing reasons: misspecification bias vs. low signal-to-noise ratio. More precisely, on the one hand, if the unknown conditional expectation function exhibits high curvature (bias) in a certain region of \mathcal{X} , then the tree construction will tend to generate a small child cell (node) in that region to reduce misspecification bias, which is precisely a celebrated feature of an “adaptive” procedure. On the other hand, as shown in this paper, small cells also emerge with non-vanishing probability when there is no misspecification bias in that region, that is, when the unknown conditional expectation function is locally constant. In practice, it is impossible to distinguish between the two equally possible scenarios.

Our theoretical results purposely remove misspecification bias by considering data generating processes with constant conditional expectation functions. In real application settings, however, the conditional expectation functions may exhibit heterogeneity (even if locally constant), in which case regularization to remove small cells may lead to large bias in the causal decision tree estimators, also affecting their convergence rate.

6.4 α -Regularity and Causal Random Forests

Under specific assumptions, [Wager and Athey \[2018\]](#) and others established polynomial-in- n convergence rates for honest causal trees and forests. The slow convergence rates established in [Theorem 1](#) do not contradict, but are rather precluded by existing polynomial-in- n convergence guarantees in the literature because they assume that each split generates two child nodes that contain a constant fraction of the number of observations in the parent node, i.e., $n(\mathbf{t}_L) \gtrsim n(\mathbf{t})$ and $n(\mathbf{t}_R) \gtrsim n(\mathbf{t})$. The key assumption is often called α -regularity, because it assumes that the tree construction generates an $\alpha > 0$ proportion of the data in each terminal node (cell).

Our theoretical results imply that assumptions such as α -regularity, or variants thereof, which require *balanced* cells almost surely, are incompatible with standard decision tree constructions employing causal trees [[Athey and Imbens, 2016](#)] or any other conventional CART methodology [e.g., [Behr et al., 2022](#), and references therein]. By implication, results for causal random forests relying on α -regularity, or variants thereof, do not apply to standard recursive partitioning using CART-type algorithms. Some form of (algorithmic and/or statistical) regularization is needed, thereby introducing a bias in the estimation as well as additional tuning parameters that would need to be chosen in practice.

6.5 Decision Tree Regression

The supplemental appendix also studies standard adaptive decision tree regression via CART for nonparametric estimation of the conditional expectation of an output given a collection of features. Section SA-2 in the supplemental appendix establishes an analogue of Theorem 1, demonstrating that adaptive decision tree regression exhibits slow convergence rate or inconsistency, as causal trees do, depending on the sample splitting design used.

Our results are connected to Bühlmann and Yu [2002] and Banerjee and McKeague [2007], and subsequent work in the statistical literature. They study large sample properties of the decision stump without sample splitting with a univariate covariate ($p = 1$ and $K = 1$), and show that the minimizers $(\hat{\beta}_L, \hat{\beta}_R, \hat{\varsigma})$ in (3) at the root node converge to well-defined population minimizers $(\beta_L^*, \beta_R^*, \varsigma^*)$ at a cube-root rate $n^{1/3}$ when the population minimizers are unique and the population conditional expectation function is continuously differentiable and has nonzero derivative at ς^* , among other technical conditions. Thus, our results show that the conclusion in Bühlmann and Yu [2002] and Banerjee and McKeague [2007] are not uniformly valid over the class of conditional expectation functions: the exclusion of the constant regression function from the allowed class of data generating processes is necessary for their results to hold for all values of the scalar covariate.

6.6 Invalidity of Inference Methods

Theorem 1 establishes lower bounds on the uniform convergence rate of causal decision tree estimators. The main technical observation is that these estimation procedures will generate a partition of \mathcal{X} with highly unbalanced cells, where potentially many cells will have a very small number of samples. These results are established under Assumption 1, which does not assume a parametric family of distributions on the data, but rather only independence and moment conditions.

From an inference perspective, our results also show that a valid (Gaussian or otherwise) distributional approximation for the causal decision tree estimators, after perhaps properly centering and scaling, does not hold in general. The main obstacle is that the effective sample size may not even increase for the approximation to apply in many regions of \mathcal{X} . In particular, standard inference methods, such as the usual confidence intervals of the form $\hat{\tau}_\ell^q(\mathbf{x}) \pm z_\alpha \cdot \text{Sd.Err.}(\hat{\tau}_\ell^q(\mathbf{x}))$ with z_α denoting the usual quantile of the standard Gaussian distribution, $\text{Sd.Err.}(\cdot)$ a standard error estimator, and $q \in \{\text{NSS}, \text{HON}, \text{X}\}$, will not deliver asymptotically valid inference for the parameter of interest $\tau(\mathbf{x})$.

7 Simulations

We illustrate the implications of Theorem 1 in the univariate case $p = 1$. Figure 1 reports the pointwise root mean squared error $\text{RMSE}(x) = \{\mathbb{E}[(\hat{\tau}_\ell^q(x) - \tau)^2]\}^{1/2}$, for $\ell \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ and $q \in \{\text{NSS}, \text{HON}, \text{X}\}$, estimated from 2,000 Monte Carlo replications under $\tau = \mu_0 = \mu_1 = 0$, $\varepsilon_i(0), \varepsilon_i(1) \stackrel{\text{i.i.d.}}{\sim} \text{N}(0, 1)$, $X_i \sim \text{Uniform}[0, 1]$, and $n = 1,000$. For each of the nine causal-tree estimators, we consider depths $K \in \{1, \dots, 5\}$, where curves are color-coded by K .

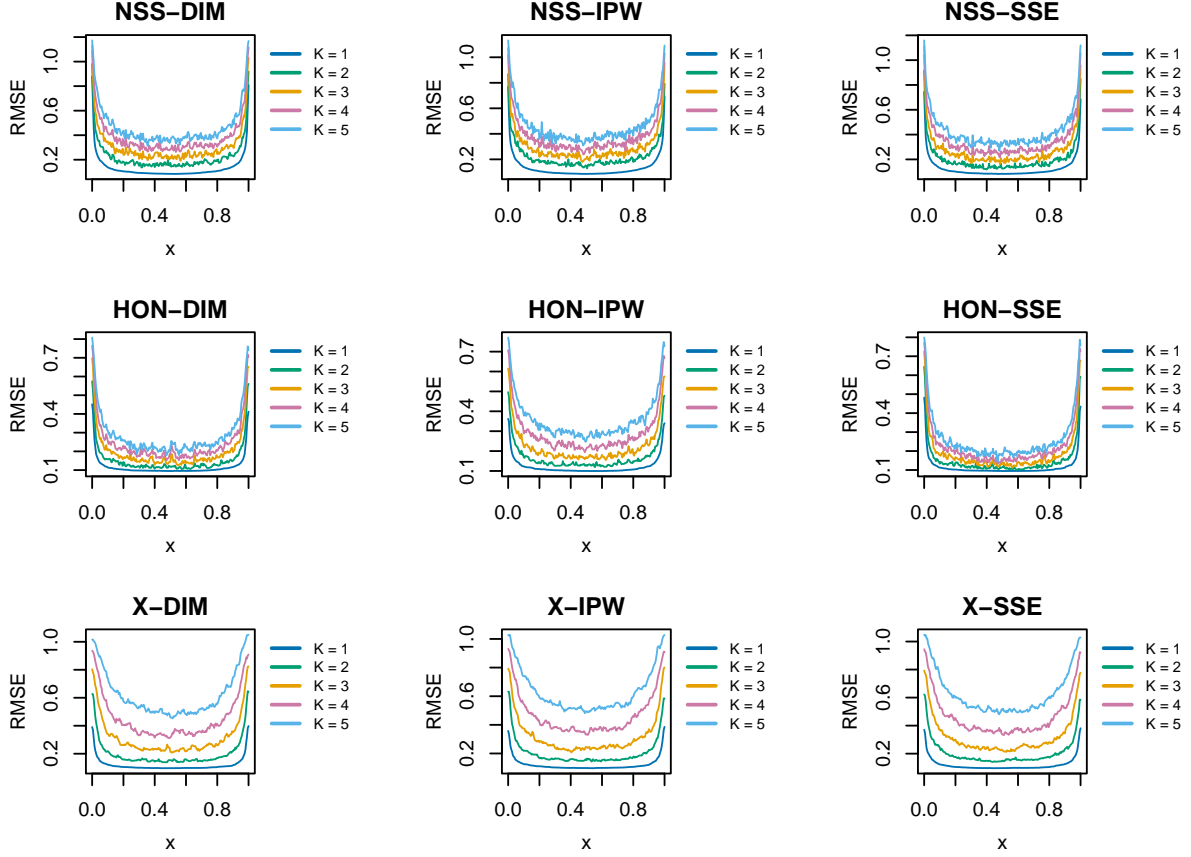


Figure 1: Plots of root mean-squared error (RMSE) of heterogeneous treatment effect estimation using nine distinct causal tree methods with depth $K = 1, 2, \dots, 5$. We chose $p = 1$, and the univariate covariate X is supported on $[0, 1]$. For all methods and depths, the causal tree has smallest pointwise RMSE near the center of the covariate space, but the performance degrades as the evaluation points move closer to the boundary. The experiment is conducted with 2,000 Monte-Carlo simulations.

Two patterns emerge across all nine methods: (i) For any fixed K , the pointwise RMSE is smallest near the center of the covariate space and increases as x approaches the boundary; (ii) For any fixed $x \in [0, 1]$, the RMSE increases with tree depth K . The first pattern is due to the small cells near boundary predicted by (6), rendering a situation where local averaging is less accurate. The second is consistent with the X-results of Theorem 1 and, heuristically, extends to NSS and HON: at higher depths, a larger fraction of evaluation points lie near terminal node boundaries, where the same boundary effects that govern decision stumps degrade performance, leading to increased RMSE even for interior points.

Acknowledgments

The authors thank Benjamin Budway, Max Farrell, Boris Hanin, Felix Hoefer, Michael Jansson, Joowon Klusowski, Boris Shigida, Jantje Sönksen, Jennifer Sun, Rocío Titiunik, and Kevin Zhang

for comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through SES-1947805, SES-2019432, and SES-2241575. Klusowski gratefully acknowledges financial support from the National Science Foundation through CAREER DMS-2239448.

References

- Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- Moulinath Banerjee and Ian W. McKeague. Confidence sets for split points in decision trees. *Annals of Statistics*, 35(2):543 – 574, 2007.
- Merle Behr, Yu Wang, Xiao Li, and Bin Yu. Provable boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences*, 119(22):e2118636119, 2022.
- Richard A Berk. *Statistical learning from a regression perspective*. Springer Series in Statistics. Springer Nature, 2020.
- Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- Peter Bühlmann and Bin Yu. Analyzing bagging. *Annals of Statistics*, 30(4):927 – 961, 2002.
- Matias D. Cattaneo, Max H. Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators. *Annals of Statistics*, 48(3):1718–1741, 2020.
- Matias D Cattaneo, Jason M Klusowski, and Peter M Tian. On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *Technical report, arXiv preprint arXiv:2211.10805*, 2022.
- Matias D. Cattaneo, Rajita Chandak, and Jason M. Klusowski. Convergence rates of oblique regression trees for flexible function libraries. *Annals of Statistics*, 52(2):466 – 490, 2024.
- Matias D Cattaneo, Yingjie Feng, and Boris Shigida. Uniform estimation and inference for non-parametric partitioning-based m-estimators. *arXiv preprint arXiv:2409.05715*, 2025.
- Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309 – 2352, 2017.
- Victor Chernozhuokov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *Annals of Statistics*, 50(5):2562–2586, 2022.
- Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, December 2022.

- Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266 – 298, 2010.
- M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley, 1997.
- M. Csörgö and P. Révész. *Strong Approximations in Probability and Statistics*. Probability and Mathematical Statistics : a series of monographs and textbooks. Academic Press, 1981.
- Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.
- F. Eicker. The asymptotic distribution of the suprema of the standardized empirical processes. *Annals of Statistics*, 7(1):116 – 138, 1979.
- Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.
- Anja Göing-Jaeschke and Marc Yor. A survey and some generalizations of besell processes. *Bernoulli*, 9(2):313 – 349, 2003.
- László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, 2002.
- Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.
- Lajos Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of statistics*, 21(2):671–680, 1993.
- Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015.
- Jason M Klusowski and Peter M Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024.
- Rafał Latała and Dariusz Matlak. *Royen’s Proof of the Gaussian Correlation Inequality*, pages 265–275. Springer International Publishing, 2017.
- Rahul Mazumder and Haoyue Wang. On the convergence of CART under sufficient impurity decrease condition. *Advances in Neural Information Processing Systems*, 36, 2024.
- Fedor Nazarov. On the maximal perimeter of a convex set in \mathbb{R}^n with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis: Israel Seminar, 2001–2002*, pages 169–187. Springer, 2003.
- Valentin V. Petrov. On lower bounds for tail probabilities. *Journal of Statistical Planning and Inference*, 137(8):2703–2705, 2007.

- Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716 – 1741, 2015.
- Galen R Shorack and RT Smythe. Inequalities for $\max_{k \in \mathcal{N}} s_k/b_k$ where $k \in \mathcal{N}$. *Proceedings of the American Mathematical Society*, pages 331–336, 1976.
- Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.
- Yan Shuo Tan, Abhineet Agarwal, and Bin Yu. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 9663–9685. PMLR, 2022.
- Yan Shuo Tan, Jason M Klusowski, and Krishnakumar Balasubramanian. Statistical-computational trade-offs for recursive adaptive partitioning estimators. *arXiv preprint arXiv:2411.04394*, 2024a.
- Yan Shuo Tan, Omer Ronen, Theo Saarinen, and Bin Yu. The computational curse of big data for bayesian additive regression trees: A hitting time analysis. *arXiv preprint arXiv:2406.19958*, 2024b.
- Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- Heping Zhang and Burton H Singer. *Recursive Partitioning and Applications*. Springer, 2010.

The Honest Truth About Causal Trees: Accuracy Limits for Heterogeneous Treatment Effect Estimation Supplemental Appendix

Matias D. Cattaneo* Jason M. Klusowski* Ruiqi (Rae) Yu*

September 16, 2025

Abstract

This supplemental appendix presents more general theoretical results encompassing those discussed in the main paper, and their proofs.

Keywords: recursive partitioning, decision trees, causal inference, heterogeneous treatment effects

*Department of Operations Research and Financial Engineering, Princeton University.

Contents

SA-1 Overview	4
SA-1.1 Notations	4
SA-1.2 Proof of Main Paper Results	4
SA-2 Constant Regression Model	5
SA-2.1 No Sample Splitting	6
SA-2.2 Honest Sample Splitting	8
SA-2.3 X-adaptive Tree	9
SA-3 Heterogeneous Causal Effect Estimation	9
SA-3.1 IPW Estimator	11
SA-3.1.1 No Sample Splitting	12
SA-3.1.2 Honest Sample Splitting	13
SA-3.1.3 X-adaptive Tree	13
SA-3.2 DIM Estimator	13
SA-3.2.1 No Sample Splitting	14
SA-3.2.2 Honest Sample Splitting	16
SA-3.2.3 X-adaptive Tree	16
SA-3.3 SSE Estimator	17
SA-3.3.1 No Sample Splitting	17
SA-3.3.2 Honest Sample Splitting	19
SA-3.3.3 X-adaptive Tree	20
SA-3.4 Additional Results	20
SA-3.4.1 Squared T-statistic Estimators	20
SA-3.4.2 Unbiasedness under Symmetric Error	21
SA-4 Proofs	21
SA-4.1 Proof of Theorem SA-1	21
SA-4.1.1 Univariate Case	22
SA-4.1.2 Multivariate Case	24
SA-4.2 Proof of Remark SA-1	30
SA-4.3 Proof of Theorem SA-2	31
SA-4.4 Proof of Theorem SA-3	33
SA-4.5 Proof of Theorem SA-4	34
SA-4.6 Proof of Theorem SA-5	34
SA-4.7 Proof of Theorem SA-6	36
SA-4.8 Proof of Theorem SA-7	37
SA-4.9 Proof of Theorem SA-8	39
SA-4.10 Proof of Corollary SA-9	39
SA-4.11 Proof of Corollary SA-10	39
SA-4.12 Proof of Corollary SA-11	40
SA-4.13 Proof of Corollary SA-12	40
SA-4.14 Proof of Corollary SA-13	40

SA-4.15	Proof of Corollary SA-14	40
SA-4.16	Proof of Corollary SA-15	40
SA-4.17	Proof of Corollary SA-16	40
SA-4.18	Proof of Lemma SA-17	40
SA-4.19	Proof of Lemma SA-18	42
SA-4.20	Proof of Theorem SA-19	44
SA-4.21	Proof of Theorem SA-20	47
SA-4.22	Proof of Theorem SA-21	49
SA-4.23	Proof of Theorem SA-22	49
SA-4.24	Proof of Theorem SA-23	52
SA-4.25	Proof of Theorem SA-24	55
SA-4.26	Proof of Theorem SA-25	55
SA-4.27	Proof of Theorem SA-26	55
SA-4.28	Proof of Lemma SA-27	56
SA-4.29	Proof of Lemma SA-28	57
SA-4.30	Proof of Theorem SA-29	57
SA-4.31	Proof of Corollary SA-30	63
SA-4.32	Proof of Corollary SA-31	64
SA-4.33	Proof of Corollary SA-32	64
SA-4.34	Proof of Corollary SA-33	64
SA-4.35	Proof of Corollary SA-34	64
SA-4.36	Proof of Corollary SA-35	64
SA-4.37	Proof of Corollary SA-36	64
SA-4.38	Proof of Lemma SA-37	64

SA-1 Overview

This supplement presents proofs for the results in the main paper, and several additional theoretical results. We start with a homoskedastic constant regression model in Section SA-2, showing that the standard CART decision tree estimator of the (constant) conditional mean suffers from slow uniform convergence rates. In Section SA-3, we then study the more challenging heterogeneous causal effect estimators discussed in the main paper: inverse probability weighting (IPW) estimator, the difference in mean (DIM) estimator, and the sum-of-square-minimization (SSE) estimator are considered in Sections SA-3.1, SA-3.2 and SA-3.3, respectively. Section SA-1.2 links the results in this supplemental appendix to those presented in the main paper.

SA-1.1 Notations

Sets. \mathbb{R} is the set of real numbers and \mathbb{N} the positive integers. For $n \in \mathbb{N}$ we write $[n] = \{1, \dots, n\}$.

Vectors and matrices. Boldface lower-case letters (e.g. \mathbf{x}) denote column vectors, and boldface upper-case letters (e.g. \mathbf{A}) denote matrices. For a vector \mathbf{x} , its i -th component is x_i ; for a matrix \mathbf{A} , its (i, j) -th entry is A_{ij} . Denote by \mathbf{e}_j the j -th unit vector.

Norms. For $\mathbf{x} \in \mathbb{R}^d$, define $\|\mathbf{x}\| = (\sum_{i=1}^d x_i^2)^{1/2}$, and $\|\mathbf{x}\|_\infty = \max_{i \leq d} |x_i|$. For a matrix $A \in \mathbb{R}^{m \times n}$, the operator norm is $\|A\| = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$, and the max norm is $\|A\|_{\max} = \max_{1 \leq i \leq m, 1 \leq j \leq n} |A_{ij}|$. For a bounded measurable function g , $\|g\|_\infty = \sup_x |g(x)|$. For a random variable X with distribution P_X , denote the population L_2 norm by $\|X\| = (\int \|x\|^2 dP_X(x))^{1/2}$; and given a random sample $\mathcal{D} = \{X_1, \dots, X_n\}$, denote the empirical L_2 norm by $\|X\|_{\mathcal{D}} = (n^{-1} \sum_{i=1}^n \|X_i\|^2)^{1/2}$.

Asymptotics. For reals sequences $a_n \ll b_n$ (or $a_n = o(b_n)$) if $\limsup_{n \rightarrow \infty} \frac{|a_n|}{|b_n|} = 0$; $|a_n| \lesssim |b_n|$ (or $a_n = O(b_n)$) if there exists some constant C and $N > 0$ such that $n > N$ implies $|a_n| \leq C|b_n|$. For sequences of random variables $a_n = o_{\mathbb{P}}(b_n)$ if $\text{plim}_{n \rightarrow \infty} \frac{|a_n|}{|b_n|} = 0$, $|a_n| \lesssim_{\mathbb{P}} |b_n|$ if $\limsup_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P}[|\frac{a_n}{b_n}| \geq M] = 0$.

Other. $\mathbf{1}(\cdot)$ denotes the indicator function. For two random variables X and Y , $X \perp Y$ means X and Y are independent. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor and ceiling of x respectively. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\text{Beta}(\alpha, \beta)$ denotes the Beta distribution with parameter (α, β) . A stochastic process $\{B(t), 0 \leq t \leq 1\}$ is a Brownian bridge, if B is a continuous Gaussian process with $\mathbb{E}[B(t)] = 0$, and $\mathbb{E}[B(t)B(s)] = \min\{t, s\} - ts$.

SA-1.2 Proof of Main Paper Results

- **Proof of Theorem 1:** The conclusions follow from Corollary SA-11, Corollary SA-13, Theorem SA-21, Theorem SA-23, Corollary SA-31, and Corollary SA-33.
- **Proof of Theorem 2:** The conclusions follow from Corollary SA-12, Corollary SA-14, Theorem SA-22, Theorem SA-24, Corollary SA-32, and Corollary SA-34.
- **Proof of Theorem 3:** The conclusions follow from Corollary SA-15, Corollary SA-16, Theorem SA-25, Theorem SA-26, Corollary SA-35, and Corollary SA-36.

SA-2 Constant Regression Model

This section is self-contained, and substantially improves on the results reported in Cattaneo et al. [2022]. The results presented herein are of independent interest in regression estimation settings, and they also offer a gentle introduction to the more technically involved results discussed in Section SA-3.

Consider the canonical regression model where the observed data $\{(y_i, \mathbf{x}_i^T) : i = 1, 2, \dots, n\}$ is a random sample satisfying

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | \mathbf{x}_i] = 0, \quad \mathbb{E}[\varepsilon_i^2 | \mathbf{x}_i] = \sigma^2(\mathbf{x}_i), \quad (\text{SA-1})$$

with $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ a vector of p covariates taking values on some support set \mathcal{X} .

Assumption SA-1 (Location Regression Model). $\mathcal{D} = \{(y_i, \mathbf{x}_i^T) : 1 \leq i \leq n\}$ is a random sample such that the following conditions hold for all $i = 1, 2, \dots, n$, satisfying Equation (SA-1) and the following:

1. $y_i = \mu(\mathbf{x}_i) + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i | \mathbf{x}_i] = 0$ and $\mathbf{x}_i \perp \varepsilon_i$.
2. $\mu(\mathbf{x}) = c$ for all $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$, where c is some constant.
3. $x_{i,1}, \dots, x_{i,p}$ are independent and continuously distributed.
4. There exists $\alpha > 0$ such that $\mathbb{E}[\exp(\lambda \varepsilon_i)] < \infty$ for all $|\lambda| < 1/\alpha$ and $\sigma^2 = \mathbb{E}[\varepsilon_i^2] > 0$.

In what follows, we denote by P_X the marginal distribution of \mathbf{x}_i .

Now we illustrate the CART estimation strategy. Given any tree T , the CART estimator is given as follows:

Definition SA-1 (CART Estimate). Suppose T is the tree used, and $\mathcal{D}_\mu = \{(y_i, \mathbf{x}_i^T) : i = 1, 2, \dots, n_\mu\}$, with $n_\mu \leq n$, is the dataset used. Let \mathbf{t} be the unique terminal node in T containing $\mathbf{x} \in \mathcal{X}$. The CART estimator is

$$\hat{\mu}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\mu) = \frac{1}{n(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} y_i,$$

where $n(\mathbf{t}) = \sum_{i=1}^{n_\mu} \mathbf{1}(\mathbf{x}_i \in \mathbf{t})$ is the “local” sample sizes. In case $n(\mathbf{t}) = 0$, take $\hat{\mu}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\mu) = 0$.

Definition SA-2 (Tree Construction). Given a dataset $\mathcal{D}_\mathsf{T} = \{(y_i, \mathbf{x}_i^T) : i = 1, 2, \dots, n_\mathsf{T}\}$, with $n_\mathsf{T} \leq n$, a parent node \mathbf{t} in the tree (i.e., a region in \mathcal{X}) is divided into two child nodes, \mathbf{t}_L and \mathbf{t}_R , by minimizing the sum-of-squares error (SSE),

$$\min_{1 \leq j \leq p} \min_{\beta_\mathsf{L}, \beta_\mathsf{R}, \varsigma \in \mathbb{R}} \sum_{\mathbf{x}_i \in \mathbf{t}} (y_i - \beta_\mathsf{L} \mathbf{1}(x_{ij} \leq \varsigma) - \beta_\mathsf{R} \mathbf{1}(x_{ij} > \varsigma))^2, \quad (\text{SA-2})$$

where $(\beta_\mathsf{L}, \beta_\mathsf{R}, \varsigma, j)$ denote the two child nodes outputs, split point, and split direction, respectively. With at least one split, the final CART tree is denoted by $\mathsf{T}(\mathcal{D}_\mathsf{T})$.

Definition SA-3 (Sample Splitting). Recall Definition SA-1 and Definition SA-2, and that $\mathcal{D} = \{(y_i, \mathbf{x}_i^T) : i = 1, 2, \dots, n\}$ is the available random sample.

- **No Sample Splitting (NSS)**: The dataset \mathcal{D} is used for both the tree construction and the treatment effect estimation, that is, $\mathcal{D}_\mathsf{T} = \mathcal{D}$ and $\mathcal{D}_\mu = \mathcal{D}$. The CART tree estimator is

$$\hat{\mu}^{\text{NSS}}(\mathbf{x}) = \hat{\mu}(\mathbf{x}; \mathsf{T}(\mathcal{D}), \mathcal{D}).$$

- *Honesty (HON):* The dataset \mathcal{D} is divided in two independent datasets \mathcal{D}_\top and \mathcal{D}_μ with sample sizes n_\top and n_μ , respectively, and satisfying $n \lesssim n_\top, n_\mu \lesssim n$. The CART tree estimator is

$$\hat{\mu}^{\text{HON}}(\mathbf{x}) = \hat{\mu}(\mathbf{x}; \mathbb{T}(\mathcal{D}_\top), \mathcal{D}_\mu).$$

Definition SA-4 (X-Adaptive Estimation). Recall Definition SA-1 and Definition SA-2, and that $\mathcal{D} = \{(y_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n\}$ is the available random sample.

1. The dataset \mathcal{D} is divided into $K+1$ datasets $(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K}, \mathcal{D}_\mu)$, with sample sizes $(n_{\top_1}, \dots, n_{\top_K}, n_\mu)$, respectively, and satisfying $n_{\top_1} = \dots = n_{\top_K} = n_\mu$ (possibly after dropping $n \bmod K$ data points at random). For each of the datasets $\mathcal{D}_{\top_j} = \{(y_i, \mathbf{x}_i^\top) : i = 1, \dots, n_{\top_j}\}$, $j = 1, \dots, K$, replace $\{y_i : i = 1, \dots, n_{\top_j}\}$ with independent copies $\{\tilde{y}_i : i = 1, \dots, n_{\top_j}\}$, while keeping the same $\{\mathbf{x}_i : i = 1, \dots, n_{\top_j}\}$.
2. The maximal decision tree of depth K , $\mathbb{T}_K(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K})$, is obtained by iterating K times the $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ splitting procedures in Definition SA-2, each time splitting all terminal nodes until (i) the node contains a single data point (y_i, \mathbf{x}_i^\top) , or (ii) the input values \mathbf{x}_i and/or all y_i within the node are the same.
3. The **X**-adaptive estimator is

$$\hat{\mu}^{\text{X}}(\mathbf{x}; K) = \hat{\mu}(\mathbf{x}; \mathbb{T}_K(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K}), \mathcal{D}_\mu).$$

SA-2.1 No Sample Splitting

We start from the no sample splitting (NSS) case, and characterize the location of the first split.

Decision Stumps.

For each variable $j = 1, 2, \dots, p$, let π_j be the permutation such that $x_{\pi_j(i), j}$ is non-decreasing in the index $i = 1, 2, \dots, n$. Then, minimizing Equation (SA-2) can be equivalently recasted as maximizing the so-called *impurity gain*:

$$\begin{aligned} & \sum_{\mathbf{x}_i \in t} (y_i - \bar{y}_t)^2 - \sum_{\mathbf{x}_i \in t} (y_i - \bar{y}_{t_L} \mathbf{1}(\mathbf{x}_i \in t_L) - \bar{y}_{t_R} \mathbf{1}(\mathbf{x}_i \in t_R))^2 \\ &= \frac{\left(\frac{1}{\sqrt{n(t)}} \sum_{\mathbf{x}_i \in t_L} (y_i - \mu) - \frac{n(t_L)}{n(t)} \frac{1}{\sqrt{n(t)}} \sum_{\mathbf{x}_i \in t} (y_i - \mu) \right)^2}{(n(t_L)/n(t))(1 - n(t_L)/n(t))}, \end{aligned} \quad (\text{SA-3})$$

where $\bar{y}_t = n(t)^{-1} \sum_{\mathbf{x}_i \in t} y_i \mathbf{1}(\mathbf{x}_i \in t)$. We can show this is also equivalent to maximizing the *conditional variance given the split*:

$$\frac{n(t_L)n(t_R)}{n(t)} (\bar{y}_{t_L} - \bar{y}_{t_R})^2. \quad (\text{SA-4})$$

We start by considering the case when the tree is depth one ($K = 1$), i.e., a decision stump. Then optimization objectives are equivalent to choosing a splitting coordinate \hat{j} , and a splitting index \hat{i} such that

$$t_L = \{\mathbf{u} \in X : \mathbf{u}_j \leq x_{\pi_j(i), j}\}, \quad t_R = \{\mathbf{u} \in X : \mathbf{u}_j > x_{\pi_j(i), j}\}.$$

The tree output can then be written as

$$\hat{\mu}^{\text{NSS}}(\mathbf{x}) = \begin{cases} \bar{y}_{t_L}, & \mathbf{x} \in t_L, \\ \bar{y}_{t_R}, & \mathbf{x} \in t_R, \end{cases} \quad (\text{SA-5})$$

where x_j denotes the value of the j -th component of \mathbf{x} .

The following theorem formally (and very precisely) characterizes the regions of the support \mathcal{X} where the first CART split index \hat{i} , at the root node, has non-vanishing probability of realizing. As a consequence, the theorem also characterizes the effective sample size of the resulting cells (recall the data is ordered so that $\hat{\mu} = x_{i\hat{j}}$ and hence $\hat{i} = \#\{\mathbf{x}_i : x_{i\hat{j}} \leq \hat{\mu}\}$).

Theorem SA-1 (Imbalanced Splits). *Suppose Assumption SA-1 holds, and let (\hat{i}, \hat{j}) be the CART split index and split direction at the root node. For each $a, b \in (0, 1)$ with $a < b$, and $\ell \in [p]$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b, \hat{j} = \ell) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i} \leq n - n^a, \hat{j} = \ell) \geq \frac{b - a}{2pe}, \quad (\text{SA-6})$$

which implies

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i} \leq n - n^a) \geq \frac{b - a}{2e}.$$

As part of the technical proofs, we correct a statement in the limiting distribution of the maximum of an O-U process in Eicker [1979, Theorem 5] – the $2 \log(c)$ term appearing in the limiting probability should be $\log(c)$. A corrected version for a more general case (the maximum of the norm of possibly multivariate O-U process) is given in the following remark:

Remark SA-1 (A Markovian type result of Darling-Erdos Theorem for Vectors). *Let $\{V_1(t) : 0 \leq t < \infty\}$, \dots , $\{V_d(t) : 0 \leq t < \infty\}$ be independent identically distributed Ornstein-Uhlenbeck processes with $\mathbb{E}[V_i(t)] = 0$ and $\mathbb{E}[V_i(t)V_i(s)] = \exp(-|t - s|/2)$, $1 \leq i \leq d$. Define*

$$N(t) = \left(\sum_{1 \leq i \leq d} V_i^2(t) \right)^{1/2}.$$

For any $c > 0$, $z \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(a(\log(n)) \sup_{0 \leq t \leq c \log(n)} N(\log(n)) - b_d(\log(n)) \leq z \right) = \exp \left(-e^{-(z - \log(c))} \right),$$

where $a(t) = (2 \log(t))^{1/2}$ and $b_d(t) = 2 \log(t) + \frac{d}{2} \log \log(t) - \log \Gamma(d/2)$.

Theorem SA-2 (Convergence Rates for Decision Stumps). *Suppose Assumption SA-1 holds. Suppose the CART tree has depth $K = 1$. Then for any $a, b \in (0, 1)$ with $a < b$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}^{\text{NSS}}(\mathbf{x}) - \mu| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b}{e}, \quad (\text{SA-7})$$

and suppose w.l.o.g. that $\mathbf{x}_i \sim \text{Uniform}([0, 1]^p)$, then

$$\liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in \mathcal{X}_n} \mathbb{P} \left(|\hat{\mu}^{\text{NSS}}(\mathbf{x}) - \mu| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b - a}{2e}, \quad (\text{SA-8})$$

where $\mathcal{X}_n = \{\mathbf{x} \in [0, 1]^p : x_j = o(1)n^{a-1} \text{ or } 1 - x_j = o(1)n^{a-1} \text{ for some } j \in [p]\}$.

Deep Trees.

We will show that the imbalanced split issue is inherited from the decision stumps to trees of arbitrary depth.

Theorem SA-3 (Convergence Rates for Deep Trees). *Suppose Assumption SA-1 holds. Then for any $b \in (0, 1)$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}^{\text{NSS}}(\mathbf{x}) - \mu| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq b/e.$$

Therefore, decision trees grown with CART methodology cannot converge faster than any polynomial-in- n , when uniformity over the full support of the data \mathcal{X} , and over possible data generating processes, is of interest.

However, for the L_2 -risk we still have the following positive result. This is because the small cells that leads to issues in uniform consistency will have a small measure by $\mathbb{P}_{\mathbf{X}}$.

Theorem SA-4 (L_2 Consistency – NSS). *Suppose Assumption SA-1 holds. Then for the depth K (possibly non-maximal) tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\mu}^{\text{NSS}}(\mathbf{x}) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^4 \log(np)}{n},$$

where C is a positive constant that only depends on σ^2 . Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\mu}^{\text{NSS}}(\mathbf{x}) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^4 \log(np)}{n} \right) = 0,$$

where C' is a positive constant that only depends on the distribution of ε_i .

SA-2.2 Honest Sample Splitting

For honest sample splitting strategy, we also present a lower bound on uniform consistency and an upper bound on L_2 consistency.

Theorem SA-5. *Suppose Assumption SA-1 holds. Then for any $b \in (0, 1)$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}^{\text{HON}}(\mathbf{x}) - \mu| \geq \frac{C \mathbb{E}[|y_i - \mu|]}{n^{b/2}} \right) \geq C \frac{\mathbb{E}[|y_i - \mu|^2]}{\mathbb{V}[y_i]} b,$$

where C is some constant only depending on $\liminf_{n \rightarrow \infty} \frac{n_{\top}}{n_{\mu}}$ and $\limsup_{n \rightarrow \infty} \frac{n_{\top}}{n_{\mu}}$.

Theorem SA-6 (L_2 Consistency – HON). *Suppose Assumption SA-1 holds. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\mu}^{\text{HON}}(\mathbf{x}) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^5}{n},$$

provided $\rho^{-1} \leq \frac{n_{\top}}{n_{\mu}} \leq \rho$ for some $\rho \in (0, 1)$, and C is a positive constant that only depends on σ^2 and ρ .

Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\mu}^{\text{HON}}(\mathbf{x}) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^5}{n} \right) = 0,$$

where C' is some constant only depending on ρ and the distribution of ε_i .

Compared to Theorem SA-3, the lower bound on the LHS of Theorem SA-5 that we characterize has one less $\sqrt{(2 + o(1)) \log \log(n)}$. Compared to Theorem SA-4, the upper bound on the RHS of Theorem SA-6 has $\log(np)$ replaced by $\log(n)$. These changes are due to the honest sample splitting strategy.

SA-2.3 X-adaptive Tree

For X-adaptive trees, we leverage the decision stump result from Theorem SA-1 using an iterative argument to infer inconsistency of trees of depth $K_n \gtrsim \log \log(n)$.

Theorem SA-7 (Pointwise Inconsistency). *Suppose Assumption SA-1 holds. If $\liminf_{n \rightarrow \infty} \frac{K_n}{\log \log(n)} > 0$, then there exists a positive constant C not depending on n such that*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}^{\mathbf{x}}(\mathbf{x}; K_n) - \mu| > C \right) > 0.$$

Since we keep the \mathbf{x}_i 's and refresh the (d_i, y_i) 's, the tree estimator has a simple form condition on \mathbf{x}_i 's. Hence a direct variance calculation gives us the following L_2 -consistency result.

Theorem SA-8 (L2 Consistency – X). *Suppose Assumption SA-1 holds. Then*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\mu}^{\mathbf{x}}(\mathbf{x}; K) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq \frac{2^{K+1}(K+1)\sigma^2}{n+1}.$$

Using the same argument as Theorem SA-6, we can show

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\mu}^{\mathbf{x}}(\mathbf{x}; K) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{K 2^K \log(n)^5}{n},$$

where C is a positive constant that only depends on σ^2 . The direct variance calculation allows us to remove extra poly-log terms.

SA-3 Heterogeneous Causal Effect Estimation

In this section, we consider the heterogeneous causal effect estimation problem from the main paper. The assumptions on the data generating process and the definitions of causal trees are the same as in the main paper. For completeness, we include them here:

Assumption SA-2 (Data Generating Process). $\mathcal{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : 1 \leq i \leq n\}$ is a random sample, where $y_i = d_i y_i(1) + (1 - d_i) y_i(0)$, $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^\top$, and the following conditions hold for all $d = 0, 1$ and $i = 1, 2, \dots, n$.

1. $(y_i(0), y_i(1), \mathbf{x}_i) \perp\!\!\!\perp d_i$, and $\xi = \mathbb{P}[d_i = 1] \in (0, 1)$.
2. $y_i(d) = \mu_d(\mathbf{x}_i) + \varepsilon_i(d)$, with $\mathbb{E}[\varepsilon_i(d)|\mathbf{x}_i] = 0$ and $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i(d)$.
3. $\mu_d(\mathbf{x}) = c_d$ for all $\mathbf{x} \in \mathcal{X}$, where c_d is some constant, and \mathcal{X} is the support of \mathbf{x}_i .

4. $x_{i,1}, \dots, x_{i,p}$ are independent and continuously distributed.

5. There exists $\alpha > 0$ such that $\mathbb{E}[\exp(\lambda \varepsilon_i(d))] < \infty$ for all $|\lambda| < 1/\alpha$ and $\mathbb{E}[\varepsilon_i^2(d)] > 0$.

And the causal trees are constructed based on the following rules:

Definition SA-5 (CATE Estimators). Suppose \mathbb{T} is the tree used, and $\mathcal{D}_\tau = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\tau\}$, with $n_\tau \leq n$, is the dataset used. Let \mathbf{t} be the unique terminal node in \mathbb{T} containing $\mathbf{x} \in \mathcal{X}$.

- The *Difference-in-Means (DIM)* estimator is

$$\hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbb{T}, \mathcal{D}_\tau) = \frac{1}{n_1(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} d_i y_i - \frac{1}{n_0(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} (1 - d_i) y_i,$$

where $n_d(\mathbf{t}) = \sum_{i=1}^{n_\tau} \mathbf{1}(\mathbf{x}_i \in \mathbf{t}, d_i = d)$, for $d = 0, 1$, are the “local” sample sizes. In case $n_0(\mathbf{t}) = 0$ or $n_1(\mathbf{t}) = 0$, take $\hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbb{T}, \mathcal{D}_\tau) = 0$.

- The *Inverse Probability Weighting (IPW)* estimator is

$$\hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbb{T}, \mathcal{D}_\tau) = \frac{1}{n(\mathbf{t})} \sum_{i: \mathbf{x}_i \in \mathbf{t}} \frac{d_i - \xi}{\xi(1 - \xi)} y_i,$$

where $n(\mathbf{t}) = n_0(\mathbf{t}) + n_1(\mathbf{t}) = \sum_{i=1}^{n_\tau} \mathbf{1}(\mathbf{x}_i \in \mathbf{t})$ is the “local” sample size. In case $n(\mathbf{t}) = 0$, take $\hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbb{T}, \mathcal{D}_\tau) = 0$.

Definition SA-6 (Tree Construction). Suppose $\mathcal{D}_\mathbb{T} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\mathbb{T}\}$, with $n_\mathbb{T} \leq n$, is the dataset used to construct the tree \mathbb{T} .

- *Variance Maximization*: A parent node \mathbf{t} (i.e., a terminal node partitioning \mathcal{X}) in a previous tree \mathbb{T}' is divided into two child nodes, \mathbf{t}_L and \mathbf{t}_R , forming the new tree \mathbb{T} , by maximizing

$$\frac{n(\mathbf{t}_L)n(\mathbf{t}_R)}{n(\mathbf{t})} \left(\hat{\tau}_l(\mathbf{t}_L; \mathbb{T}, \mathcal{D}_\mathbb{T}) - \hat{\tau}_l(\mathbf{t}_R; \mathbb{T}, \mathcal{D}_\mathbb{T}) \right)^2, \quad l \in \{\text{DIM}, \text{IPW}\}. \quad (\text{SA-9})$$

With at least one split, the two final causal trees are denoted by $\mathbb{T}^{\text{DIM}}(\mathcal{D}_\mathbb{T})$ and $\mathbb{T}^{\text{IPW}}(\mathcal{D}_\mathbb{T})$, respectively, for $l \in \{\text{DIM}, \text{IPW}\}$.

- *SSE Minimization*: A parent node \mathbf{t} (i.e., a terminal node partitioning \mathcal{X}) in the previous tree \mathbb{T}' is divided into two child nodes, \mathbf{t}_L and \mathbf{t}_R , forming the next tree \mathbb{T} , by solving

$$\min_{a_L, b_L, a_R, b_R \in \mathbb{R}} \sum_{\mathbf{x}_i \in \mathbf{t}_L} (y_i - a_L - b_L d_i)^2 + \sum_{\mathbf{x}_i \in \mathbf{t}_R} (y_i - a_R - b_R d_i)^2, \quad (\text{SA-10})$$

where only the data $\mathcal{D}_\mathbb{T}$ is used. With at least one split, the final causal tree is denoted by $\mathbb{T}^{\text{SSE}}(\mathcal{D}_\mathbb{T})$.

Definition SA-7 (Sample Splitting and Estimators). Recall Definition SA-5 and Definition SA-6, and that $\mathcal{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ is the available random sample.

- *No Sample Splitting (NSS)*: The dataset \mathcal{D} is used for both the tree construction and the treatment

effect estimation, that is, $\mathcal{D}_\top = \mathcal{D}$ and $\mathcal{D}_\tau = \mathcal{D}$. The causal tree estimators are

$$\begin{aligned}\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{DIM}}(\mathcal{D}), \mathcal{D}), \\ \hat{\tau}_{\text{IPW}}^{\text{NSS}}(\mathbf{x}) &= \hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbf{T}^{\text{IPW}}(\mathcal{D}), \mathcal{D}), \quad \text{and} \\ \hat{\tau}_{\text{SSE}}^{\text{NSS}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{SSE}}(\mathcal{D}), \mathcal{D}),\end{aligned}$$

- *Honesty (HON)*: The dataset \mathcal{D} is divided in two independent datasets \mathcal{D}_\top and \mathcal{D}_τ with sample sizes n_\top and n_τ , respectively, and satisfying $n \lesssim n_\top, n_\tau \lesssim n$. The causal tree estimators are

$$\begin{aligned}\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{DIM}}(\mathcal{D}_\top), \mathcal{D}_\tau), \\ \hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}) &= \hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbf{T}^{\text{IPW}}(\mathcal{D}_\top), \mathcal{D}_\tau), \quad \text{and} \\ \hat{\tau}_{\text{SSE}}^{\text{HON}}(\mathbf{x}) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}^{\text{SSE}}(\mathcal{D}_\top), \mathcal{D}_\tau).\end{aligned}$$

While the estimators $\hat{\tau}_l^{\text{NSS}}(\mathbf{x})$ and $\hat{\tau}_l^{\text{HON}}(\mathbf{x})$, $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ depend on the depth of the tree construction used, our notation does not make this dependence explicit because our results only require (at least) one single split.

X-Adaptive Trees.

Definition SA-8 (X-Adaptive Estimation). Recall Definition SA-5 and Definition SA-6, and that $\mathcal{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ is the available random sample.

1. The dataset \mathcal{D} is divided into $K+1$ datasets $(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K}, \mathcal{D}_\tau)$, with sample sizes $(n_{\top_1}, \dots, n_{\top_K}, n_\tau)$, respectively, and satisfying $n_{\top_1} = \dots = n_{\top_K} = n_\tau$ (possibly after dropping $n \bmod K$ data points at random). For each of the datasets $\mathcal{D}_j = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, \dots, n_{\top_j}\}$, $j = 1, \dots, K$, replace $\{(y_i, d_i) : i = 1, \dots, n_{\top_j}\}$ with independent copies $\{(\tilde{y}_i, \tilde{d}_i) : i = 1, \dots, n_{\top_j}\}$, while keeping the same $\{\mathbf{x}_i : i = 1, \dots, n_{\top_j}\}$.
2. The maximal decision tree of depth K , $\mathbf{T}_K^l(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K})$, is obtained by iterating K times the $l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}$ splitting procedures in Definition SA-6, each time splitting all terminal nodes until (i) the node contains a single data point $(y_i, d_i, \mathbf{x}_i^\top)$, or (ii) the input values \mathbf{x}_i and/or all (d_i, y_i) within the node are the same.
3. The **X**-adaptive estimators are

$$\begin{aligned}\hat{\tau}_{\text{DIM}}^{\mathbf{X}}(\mathbf{x}; K) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}_K^{\text{DIM}}(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K}), \mathcal{D}_\tau), \\ \hat{\tau}_{\text{IPW}}^{\mathbf{X}}(\mathbf{x}; K) &= \hat{\tau}_{\text{IPW}}(\mathbf{x}; \mathbf{T}_K^{\text{IPW}}(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K}), \mathcal{D}_\tau), \quad \text{and} \\ \hat{\tau}_{\text{SSE}}^{\mathbf{X}}(\mathbf{x}; K) &= \hat{\tau}_{\text{DIM}}(\mathbf{x}; \mathbf{T}_K^{\text{SSE}}(\mathcal{D}_{\top_1}, \dots, \mathcal{D}_{\top_K}), \mathcal{D}_\tau).\end{aligned}$$

SA-3.1 IPW Estimator

The transformed outcomes $y_i \frac{d_i - \xi}{\xi(1 - \xi)}$, $1 \leq i \leq n$, are i.i.d, with

$$\mathbb{E} \left[y_i \frac{d_i - \xi}{\xi(1 - \xi)} \middle| \mathbf{x}_i \right] = \mathbb{E}[y_i(1) - y_i(0) | \mathbf{x}_i] = c_1 - c_0,$$

and

$$\tilde{\varepsilon}_i = y_i \frac{d_i - \xi}{\xi(1 - \xi)} - (c_1 - c_0) = (c_1 + \varepsilon_i(1)) \frac{d_i}{\xi} - (c_0 + \varepsilon_i(0)) \frac{1 - d_i}{1 - \xi} - (c_1 - c_0) \perp \mathbf{x}_i.$$

Assumption SA-2 implies $\mathbb{E}[\exp(\lambda \tilde{\varepsilon}_i)] < \infty$ for all $|\lambda| \leq 1/\beta$ with β only depending on ξ and α , and $\mathbb{E}[\tilde{\varepsilon}_i^2] > 0$. Hence the following results are immediate corollaries from the results in Section SA-2.

SA-3.1.1 No Sample Splitting

Corollary SA-9 (Imbalanced Split). *Suppose Assumption SA-2 holds. Then for each $a, b \in (0, 1)$ with $a < b$, for every $\ell \in [p]$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b, \hat{j} = \ell) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i} \leq n - n^a, \hat{j} = \ell) \geq \frac{b - a}{2pe}.$$

Corollary SA-10 (Stump). *Suppose Assumption SA-2 holds, and the tree has depth $K = 1$. Then for any $a, b \in (0, 1)$ with $a < b$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P}\left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)}\right) \geq \frac{b}{e},$$

where $\sigma^2 = \mathbb{V}\left[\frac{d_i y_i(1)}{\xi} + \frac{(1 - d_i) y_i(0)}{1 - \xi}\right]$. Moreover, if \mathbf{x}_i has a density that is continuous and positive on $[0, 1]^p$, then

$$\liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in \mathcal{X}_n} \mathbb{P}\left(|\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)}\right) \geq \frac{b - a}{2e},$$

where $\mathcal{X}_n = \{\mathbf{x} \in [0, 1]^p : x_j = o(1)n^{a-1} \text{ or } 1 - x_j = o(1)n^{a-1} \text{ for some } j \in [p]\}$.

Corollary SA-11 (Rates). *Suppose Assumption SA-2 holds. Then for any $b \in (0, 1)$ and arbitrary depth tree, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P}\left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)}\right) \geq \frac{b}{e}.$$

Corollary SA-12 (L_2 Consistency – NSS). *Suppose Assumption SA-2 holds. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E}\left[\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x})\right] \leq C \frac{2^K \log(n)^4 \log(np)}{n},$$

where C is a positive constant that only depends on the distribution of $\tilde{\varepsilon}_i = y_i \frac{d_i - \xi}{\xi(1 - \xi)} - \tau$. Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P}\left(\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^4 \log(np)}{n}\right) = 0,$$

where C' is a positive constant that only depends on the distribution of $\tilde{\varepsilon}_i$.

SA-3.1.2 Honest Sample Splitting

Corollary SA-13 (Honest Causal Output). *Suppose Assumption SA-2 holds. Then for any $b \in (0, 1)$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}) - \tau| \geq \frac{C \mathbb{E}[|\tilde{\varepsilon}_i|]}{8n^{b/2}} \right) \geq C \frac{\mathbb{E}[|\tilde{\varepsilon}_i|^2]}{\mathbb{V}[\tilde{\varepsilon}_i]} b,$$

where C is some constant only depending on the distribution of $\tilde{\varepsilon}_i = y_i \frac{d_i - \xi}{\xi(1 - \xi)} - \tau$, $\liminf_{n \rightarrow \infty} \frac{n\tau}{n_\tau}$ and $\limsup_{n \rightarrow \infty} \frac{n\tau}{n_\tau}$.

Corollary SA-14 (L_2 Consistency – HON). *Suppose Assumption SA-2 holds. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^5}{n},$$

provided $\rho^{-1} \leq \frac{n\tau}{n_\tau} \leq \rho$ for some $\rho \in (0, 1)$, and C is some constant only depending on the distribution of $\tilde{\varepsilon}_i = y_i \frac{d_i - \xi}{\xi(1 - \xi)} - \tau$ and ρ . Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^5}{n} \right) = 0,$$

where C' is some constant only depending on the distribution of $\tilde{\varepsilon}_i$ and ρ .

SA-3.1.3 X-adaptive Tree

Corollary SA-15 (Honest CART+). *Suppose Assumption SA-2 holds. Suppose $\liminf_{n \rightarrow \infty} \frac{K_n}{\log \log(n)} > 0$. Then, there exists a positive constant C not depending on n such that*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{IPW}}^{\text{X}}(\mathbf{x}; K_n) - \tau| > C \right) > 0.$$

Corollary SA-16 (L_2 Consistency – X). *Suppose Assumption SA-2 holds. Then*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{IPW}}^{\text{X}}(\mathbf{x}; K) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K K \sigma^2}{n},$$

where C is some constant only depending on the distribution of $\tilde{\varepsilon}_i = y_i \frac{d_i - \xi}{\xi(1 - \xi)} - \tau$.

SA-3.2 DIM Estimator

The DIM estimator can not be directly written as a regression tree with transformed outcome. However, we show that it can be approximated by an IPW-tree. More specifically, we view the split criterion with different splitting index and coordinate as an empirical process, and show that the split criterion for DIM and IPW approximate each other.

SA-3.2.1 No Sample Splitting

Approximation Results on Decision Stumps.

Denote by π_ℓ permutation of index $[n]$ such that $x_{\pi_\ell(1),\ell} \leq x_{\pi_\ell(2),\ell} \leq \dots \leq x_{\pi_\ell(n),\ell}$, $1 \leq \ell \leq p$. Consider the split criterion for the regression and ipw trees when splitting at the root note when $\#\{\mathbf{x}_{\pi_\ell(i)} \in t_L\} = k$: For $1 \leq \ell \leq p$, $1 \leq k \leq n$, consider

$$\begin{aligned}\mathcal{J}^{\text{DIM}}(k, \ell) &= \frac{k(n-k)}{n} \left(\hat{\tau}_L^{\text{DIM}}(k, \ell) - \hat{\tau}_R^{\text{DIM}}(k, \ell) \right)^2, \\ \bar{\mathcal{J}}^{\text{IPW}}(k, \ell) &= \frac{k(n-k)}{n} \left(\bar{\tau}_L^{\text{IPW}}(k, \ell) - \bar{\tau}_R^{\text{IPW}}(k, \ell) \right)^2,\end{aligned}$$

where

$$\begin{aligned}\hat{\tau}_L^{\text{DIM}}(k, \ell) &= \frac{\sum_{i=1}^k d_{\pi_\ell(i)} y_{\pi_\ell(i)}}{\sum_{i=1}^k d_{\pi_\ell(i)}} - \frac{\sum_{i=1}^k (1 - d_{\pi_\ell(i)}) y_{\pi_\ell(i)}}{\sum_{i=1}^k (1 - d_{\pi_\ell(i)})}, \\ \hat{\tau}_R^{\text{DIM}}(k, \ell) &= \frac{\sum_{i=k+1}^n d_{\pi_\ell(i)} y_{\pi_\ell(i)}}{\sum_{i=k+1}^n d_{\pi_\ell(i)}} - \frac{\sum_{i=k+1}^n (1 - d_{\pi_\ell(i)}) y_{\pi_\ell(i)}}{\sum_{i=k+1}^n (1 - d_{\pi_\ell(i)})}, \\ \bar{\tau}_L^{\text{IPW}}(k, \ell) &= \frac{1}{k} \sum_{i=1}^k \frac{d_{\pi_\ell(i)}}{\xi} \varepsilon_{\pi_\ell(i)}(1) - \frac{1}{k} \sum_{i=1}^k \frac{1 - d_{\pi_\ell(i)}}{1 - \xi} \varepsilon_{\pi_\ell(i)}(0), \\ \bar{\tau}_R^{\text{IPW}}(k, \ell) &= \frac{1}{n-k} \sum_{i=k+1}^n \frac{d_{\pi_\ell(i)}}{\xi} \varepsilon_{\pi_\ell(i)}(1) - \frac{1}{n-k} \sum_{i=k+1}^n \frac{1 - d_{\pi_\ell(i)}}{1 - \xi} \varepsilon_{\pi_\ell(i)}(0).\end{aligned}$$

Notice that if we replace $\varepsilon_{\pi_\ell(i)}$ by $y_{\pi_\ell(i)}$, we would get $\hat{\tau}_L^{\text{IPW}}$ (or $\hat{\tau}_R^{\text{IPW}}$) instead of $\bar{\tau}_L^{\text{IPW}}$ (or $\bar{\tau}_R^{\text{IPW}}$). But putting $\varepsilon_{\pi_\ell(i)}$ here allows us to approximate the $\mathcal{J}^{\text{DIM}}(\cdot, \ell)$ processes.

The optimization objective based on Definition SA-6 for the regression based estimator with variance maximization is equivalent to choosing a splitting coordinate \hat{j}_{DIM} , and a splitting index \hat{i}_{DIM} such that

$$t_L = \{\mathbf{u} \in \mathcal{X} : \mathbf{u}_{\hat{j}_{\text{DIM}}} \leq x_{\pi_{\hat{j}_{\text{DIM}}}(\hat{i}_{\text{DIM}}), \hat{j}_{\text{DIM}}}\}, \quad t_R = \{\mathbf{u} \in \mathcal{X} : \mathbf{u}_{\hat{j}_{\text{DIM}}} > x_{\pi_{\hat{j}_{\text{DIM}}}(\hat{i}_{\text{DIM}}), \hat{j}_{\text{DIM}}}\},$$

that maximizes

$$\frac{n(t_L)n(t_R)}{n(t)} \left(\hat{\tau}_{\text{DIM}}(t_L) - \hat{\tau}_{\text{DIM}}(t_R) \right)^2,$$

that is,

$$(\hat{i}_{\text{DIM}}, \hat{j}_{\text{DIM}}) = \arg \max_{k, \ell} \mathcal{J}^{\text{DIM}}(k, \ell).$$

A technical aspect is to control for fluctuations of objects of the form $\frac{\sum_{i=1}^k d_{\pi_\ell(i)} y_{\pi_\ell(i)}}{\sum_{i=1}^k d_{\pi_\ell(i)}}$, for which we will use a truncation argument that requires $\sum_{i=1}^k d_{\pi_\ell(i)} \geq r_n$ with $r_n \rightarrow \infty$. This gives the following lemma:

Lemma SA-17 (Approximation Error). *Suppose Assumption SA-2 holds. Let $(r_n)_{n \in \mathbb{N}}$ be a sequence of real numbers such that $r_n \rightarrow \infty$. Then*

$$\max_{1 \leq \ell \leq p} \max_{r_n \leq k \leq n - r_n} \left| \mathcal{J}^{\text{DIM}}(k, \ell) - \bar{\mathcal{J}}^{\text{IPW}}(k, \ell) \right| = O_{\mathbb{P}} \left(\frac{\log \log(n)}{\sqrt{r_n}} \right).$$

We also control for the truncation error:

Lemma SA-18 (Truncation Error). *Suppose Assumption SA-2 holds. Let ρ_n be a sequence taking values in $(0, 1)$ such that $\limsup_{n \rightarrow \infty} \rho_n \log \log(n) = 0$, and take $s_n = \exp((\log n)^{\rho_n})$. Then*

$$\max_{1 \leq \ell \leq p} \max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \left| \mathcal{J}^{\text{DIM}}(k, \ell) - \mathcal{J}^{\text{IPW}}(k, \ell) \right| = O_{\mathbb{P}} \left(\rho_n \log \log(n) + \frac{s_n}{n - s_n} \log \log(n) \right).$$

Rates for Decision Stumps.

The previous two lemmas imply that we can study $\arg \max$ of \mathcal{J}^{DIM} in terms of $\arg \max$ of \mathcal{J}^{IPW} . The latter is the split criterion based on CART with *transformed outcome* $\frac{d_i}{\xi} \varepsilon_i(1) - \frac{1-d_i}{1-\xi} \varepsilon_i(0)$, and results from Section SA-2 can be applied.

Theorem SA-19 (Imbalanced Split). *Suppose Assumption SA-2 holds. Then for each $a, b \in (0, 1)$ with $a < b$, for every $\ell \in [p]$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i}_{\text{DIM}} \leq n^b, \hat{j}_{\text{DIM}} = \ell) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i}_{\text{DIM}} \leq n - n^a, \hat{j}_{\text{DIM}} = \ell) \geq \frac{b-a}{2pe}.$$

The issue of imbalanced cells gives rise to the slow uniform convergence rate.

Theorem SA-20 (Rates for Stump). *Suppose Assumption SA-2 holds, and the tree has depth $K = 1$. Then for any $a, b \in (0, 1)$ with $a < b$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b}{e},$$

where $\sigma^2 = \mathbb{V}[\tilde{\varepsilon}_i]$, with $\tilde{\varepsilon}_i = \frac{d_i}{\xi} \varepsilon_i(1) - \frac{1-d_i}{1-\xi} \varepsilon_i(0)$. Suppose w.l.o.g. that $\mathbf{x}_i \sim \text{Uniform}([0, 1]^p)$, then

$$\liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in \mathcal{X}_n} \mathbb{P} \left(|\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b-a}{2e},$$

where $\mathcal{X}_n = \{\mathbf{x} \in [0, 1]^p : x_j = o(1)n^{a-1} \text{ or } 1 - x_j = o(1)n^{a-1} \text{ for some } j \in [p]\}$.

Deeper Trees.

We generalize the above results on decision stumps to trees of arbitrary depths.

Theorem SA-21 (Deeper Trees). *Suppose Assumption SA-2 holds. Then for any $b \in (0, 1)$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq b/e.$$

In comparison to the uniform convergence rate, for L_2 convergence rate we can give an upper bound as follows.

Theorem SA-22 (L_2 Consistency – NSS). *Suppose Assumption SA-2 holds. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^4 \log(np)}{n},$$

where C is a positive constant that only depends on the distribution of $(d_i, \varepsilon_i(0), \varepsilon_i(1))$. Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^4 \log(np)}{n} \right) = 0,$$

where C' is a positive constant that only depends on the distribution of $(d_i, \varepsilon_i(0), \varepsilon_i(1))$.

SA-3.2.2 Honest Sample Splitting

With the honest sample splitting strategy, we also give a lower bound on uniform convergence rate and an upper bound on L_2 convergence rate. The difference in rates from the rates in the previous section is due to the different sample splitting strategies.

Theorem SA-23 (Honest Causal Output). *Suppose Assumption SA-2 holds. Then for any $b \in (0, 1)$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x}) - \tau| \geq C n^{-b/2} \right) \geq C \xi (1 - \xi) b.$$

where C is some positive constant only depending on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, $\liminf_{n \rightarrow \infty} \frac{n\tau}{n_\tau}$ and $\limsup_{n \rightarrow \infty} \frac{n\tau}{n_\tau}$.

Theorem SA-24 (L_2 Consistency – HON). *Suppose Assumption SA-2 holds. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^5}{n},$$

provided $\rho^{-1} \leq \frac{n\tau}{n_\tau} \leq \rho$ for some $\rho \in (0, 1)$, and C is a positive constant that only depends on ρ and the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$. Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^5}{n} \right) = 0,$$

where C' is a positive constant that only depends on ρ and the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

SA-3.2.3 X-adaptive Tree

We leverage Theorem SA-19 with an iterative argument to get

Theorem SA-25 (CART+). *Suppose Assumption SA-2 holds. Suppose $\liminf_{n \rightarrow \infty} \frac{K_n}{\log \log(K_n)} > 0$. Then*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\mathcal{X}}(\mathbf{x}; K_n) - \tau| > C \right) > 0,$$

where C is some positive constant not depending on n .

A direct variance calculation gives

Theorem SA-26 (L_2 Consistency). *Suppose Assumption SA-2 holds. Then*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\mathcal{X}}(\mathbf{x}; K) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{K 2^K}{n},$$

where C is some positive constant that only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

Using the same argument as Theorem SA-24, we can show

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{DIM}}^{\mathcal{X}}(\mathbf{x}; K) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{K 2^K \log(n)^5}{n},$$

where C is a positive constant that only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$. The direct variance calculation allows us to remove extra poly-log terms.

SA-3.3 SSE Estimator

While the CATE estimators given the tree of the SSE strategy coincides with the DIM strategy, the tree construction methods differ. Similar to DIM, for SSE we also characterize the distribution of split index via a Gaussian approximation. Here we show the split criterion with SSE strategy can be approximated by the split criterion from two transformed outcome regressions, one for treatment and one for control. A careful high dimensional Gaussian approximation with respect to the geometry of simple convex sets then enables us to characterize the limiting distribution of splitting indices.

SA-3.3.1 No Sample Splitting

Decision Stump.

For each variable $j = 1, 2, \dots, p$, the data $\{x_{ij} : \mathbf{x}_i \in \mathbf{t}\}$ is relabeled so that x_{ij} is increasing in the index $i = 1, 2, \dots, n(\mathbf{t})$, where $n(\mathbf{t}) = \#\{\mathbf{x}_i \in \mathbf{t}\}$. The fit-based objective is to minimize

$$\min_{a_L, b_L, a_R, b_R \in \mathbb{R}} \sum_{\mathbf{x}_i \in t_L} (y_i - a_{t_L} - b_{t_L} d_i)^2 + \sum_{\mathbf{x}_i \in t_R} (y_i - a_{t_R} - b_{t_R} d_i)^2 \quad (\text{SA-11})$$

with respect to the index i and variable j . Again, the maximizers are denoted by $(\hat{i}_{\text{SSE}}, \hat{j}_{\text{SSE}})$, and the optimal split point $\hat{\tau}$ that maximizes (SA-11) can be expressed as $x_{\hat{i}_{\text{SSE}}, \hat{j}_{\text{SSE}}}$.

To break down the criterion (SA-11), denote

$$\begin{aligned} \hat{\mu}_{L,0}(k, \ell) &= \frac{\sum_{i=1}^k (1 - d_{\pi_\ell(i)}) y_{\pi_\ell(i)}}{\sum_{i=1}^k (1 - d_{\pi_\ell(i)})}, & \hat{\mu}_{L,1}(k, \ell) &= \frac{\sum_{i=1}^k d_{\pi_\ell(i)} y_{\pi_\ell(i)}}{\sum_{i=1}^k d_{\pi_\ell(i)}}, \\ \hat{\mu}_{R,0}(k, \ell) &= \frac{\sum_{i=k+1}^n (1 - d_{\pi_\ell(i)}) y_{\pi_\ell(i)}}{\sum_{i=k+1}^n (1 - d_{\pi_\ell(i)})}, & \hat{\mu}_{R,1}(k, \ell) &= \frac{\sum_{i=k+1}^n d_{\pi_\ell(i)} y_{\pi_\ell(i)}}{\sum_{i=k+1}^n d_{\pi_\ell(i)}}. \end{aligned}$$

Also to denote the counts compactly, $n_0 = \sum_{i=1}^n (1 - d_i)$, $n_{L,0}(k) = \sum_{i=1}^k (1 - d_{\pi_\ell(i)})$, $n_{R,0}(k) = \sum_{i=k+1}^n (1 - d_{\pi_\ell(i)})$, and $n_1 = \sum_{i=1}^n d_i$, $n_{L,1}(k) = \sum_{i=1}^k d_{\pi_\ell(i)}$, $n_{R,1}(k) = \sum_{i=k+1}^n d_{\pi_\ell(i)}$. Then we can show that maximizing Equation (SA-11) is equivalent to maximizing

$$\mathcal{J}^{\text{SSE}}(k, \ell) = \frac{n_{L,0} n_{R,0}}{n_0} (\hat{\mu}_{L,0}(k, \ell) - \hat{\mu}_{R,0}(k, \ell))^2 + \frac{n_{L,1} n_{R,1}}{n_1} (\hat{\mu}_{L,1}(k, \ell) - \hat{\mu}_{R,1}(k, \ell))^2.$$

We want to show the above empirical process can be approximated by

$$\mathcal{J}^{\text{prox}}(k, \ell) = (1 - \xi) \frac{k(n-k)}{n} (\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell))^2 + \xi \frac{k(n-k)}{n} (\bar{\mu}_{L,1}(k, \ell) - \bar{\mu}_{R,1}(k, \ell))^2,$$

with

$$\begin{aligned}\bar{\mu}_{L,0}(k, \ell) &= \frac{1}{k} \sum_{i \leq k} \frac{1 - d_{\pi_\ell(i)}}{1 - \xi} Y_{\pi_\ell(i)}, & \bar{\mu}_{L,1}(k, \ell) &= \frac{1}{k} \sum_{i \leq k} \frac{d_{\pi_\ell(i)}}{\xi} Y_{\pi_\ell(i)}, \\ \bar{\mu}_{R,0}(k, \ell) &= \frac{1}{n - k} \sum_{i > k} \frac{1 - d_{\pi_\ell(i)}}{1 - \xi} Y_{\pi_\ell(i)}, & \bar{\mu}_{R,1}(k, \ell) &= \frac{1}{n - k} \sum_{i > k} \frac{d_{\pi_\ell(i)}}{\xi} Y_{\pi_\ell(i)}.\end{aligned}$$

The latter can be approximated by the summation of two independent time-transformed O-U process (which is again a time-transformed O-U process), for fixed coordinate $\ell \in [p]$. More precisely, we present the approximation lemmas:

Lemma SA-27 (Approximation Error). *Suppose Assumption SA-2 holds. Let $(r_n)_{n \in \mathbb{N}}$ be a sequence of real numbers such that $r_n \rightarrow \infty$. Then*

$$\max_{1 \leq \ell \leq p} \max_{r_n \leq k < n - r_n} \left| \mathcal{J}^{SSE}(k, \ell) - \mathcal{J}^{prox}(k, \ell) \right| = O_{\mathbb{P}} \left(\frac{\log \log(n)^{3/2}}{\sqrt{r_n}} \right).$$

Lemma SA-28 (Truncation Error). *Suppose Assumption SA-2 holds. Let ρ_n be a sequence taking values in $(0, 1)$ such that $\limsup_{n \rightarrow \infty} \rho_n \log \log(n) = \infty$, and take $s_n = \exp((\log n)^{\rho_n})$. Then*

$$\max_{1 \leq \ell \leq p} \max_{1 \leq k \leq s_n, n - s_n \leq k \leq n} \left| \mathcal{J}^{SSE}(k, \ell) - \mathcal{J}^{prox}(k, \ell) \right| = O_{\mathbb{P}} \left(\rho_n \log \log(n) + \frac{s_n}{n - s_n} \log \log(n) \right).$$

Theorem SA-29. *Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Then for each $a, b \in (0, 1)$ with $a < b$, for every $\ell \in [p]$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i}_{SSE} \leq n^b, \hat{j}_{SSE} = \ell) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i}_{SSE} \leq n - n^a, \hat{j}_{SSE} = \ell) \geq \frac{b - a}{2pe}.$$

Remark SA-2. *We add the condition that $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$ so that a two-dimensional Darling-Erdos theorem [Horváth, 1993, Lemma 2.1] can be applied. We conjecture that without $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$, the conclusion still holds with a Darling-Erdos theorem for i.n.i.d O-U process, but this is out of the scope of this paper.*

Notice that although the splitting criteria is different from the regression tree, once cells are given the estimator given by the fit-based tree is exactly the same as the regression tree (see Section SA-3.2). Hence the following results can be proved based on Theorem SA-29 and the same logic as Theorem SA-20 to Theorem SA-25.

Corollary SA-30 (Rates for Stump). *Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. For any $a, b \in (0, 1)$ with $a < b$, we have*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{SSE}^{NSS}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b}{e},$$

and suppose w.l.o.g. that $\mathbf{x}_i \sim \text{Uniform}([0, 1]^p)$, then

$$\liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in \mathcal{X}_n} \mathbb{P} \left(|\hat{\tau}_{SSE}^{NSS}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b - a}{2e},$$

where $\mathcal{X}_n = \{\mathbf{x} \in [0, 1]^p : x_j = o(1)n^{a-1} \text{ or } 1 - x_j = o(1)n^{a-1} \text{ for some } j \in [p]\}$, and $\sigma^2 = \mathbb{V}[\frac{d_i y_i(1)}{\xi}] +$

$$\frac{(1-d_i)y_i(0)}{1-\xi}].$$

Deeper Trees.

Corollary SA-31 (Deeper Trees). *Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Then for any $b \in (0, 1)$, for any sequence K_n taking values in \mathbb{N} ,*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{SSE}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq b/e.$$

Corollary SA-32 (L_2 Consistency – NSS). *Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{SSE}}^{\text{NSS}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^4 \log(np)}{n},$$

where C is a positive constant that only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$. Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\tau}_{\text{SSE}}^{\text{NSS}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^4 \log(np)}{n} \right) = 0,$$

where C' is a positive constant that only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

SA-3.3.2 Honest Sample Splitting

Corollary SA-33 (Honest Causal Output). *Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Then for any $b \in (0, 1)$, for any sequence K_n taking values in \mathbb{N} ,*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{SSE}}^{\text{HON}}(\mathbf{x}) - \tau| \geq C n^{-b/2} \right) \geq C\xi(1 - \xi)b.$$

where C is some constant only depending on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and $\liminf_{n \rightarrow \infty} \frac{n\tau}{n_\tau}$ and $\limsup_{n \rightarrow \infty} \frac{n\tau}{n_\tau}$.

Corollary SA-34 (L_2 Consistency – HON). *Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Then for the depth K (possibly non-maximal) causal tree,*

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{SSE}}^{\text{HON}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{2^K \log(n)^5}{n},$$

provided $\rho^{-1} \leq \frac{n\tau}{n_\tau} \leq \rho$ for some $\rho \in (0, 1)$, and C is a positive constant that only depends on ρ and the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$. Moreover,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\int_{\mathcal{X}} (\hat{\tau}_{\text{SSE}}^{\text{HON}}(\mathbf{x}) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C' \frac{2^K \log(n)^5}{n} \right) = 0,$$

where C' is a positive constant that only depends on ρ and the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

SA-3.3.3 X-adaptive Tree

Corollary SA-35. Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Suppose $\liminf_{n \rightarrow \infty} \frac{K_n}{\log \log(n)} > 0$. Then

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{SSE}}^{\mathbf{x}}(\mathbf{x}; K_n) - \tau| > C \right) > 0.$$

Corollary SA-36 (L2 Consistency). Suppose Assumption SA-2 holds with $\mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$. Then

$$\mathbb{E} \left[\int_{\mathcal{X}} (\hat{\tau}_{\text{SSE}}^{\mathbf{x}}(\mathbf{x}; K) - \tau)^2 dF_{\mathbf{X}}(\mathbf{x}) \right] \leq C \frac{K 2^K}{n},$$

where C is some constant not depending on n .

SA-3.4 Additional Results

SA-3.4.1 Squared T-statistic Estimators

The fourth method proposed by [Athey and Imbens \[2016\]](#) is the squared T-statistic trees, where at the root node the index and coordinate to split (\hat{i}, \hat{j}) are chosen so that the *squared T-statistics metric* is maximized, that is,

$$\hat{i}, \hat{j} = \arg \max_{k \in [n], \ell \in [p]} n \frac{(\hat{\tau}_L(k, \ell) - \hat{\tau}_R(k, \ell))^2}{S(k, \ell)^2/k + S(k, \ell)^2/(n-k)},$$

where $\hat{\tau}_L(k, \ell)$ and $\hat{\tau}_R(k, \ell)$ are the causal tree estimators for the left and right nodes respectively based on split coordinate ℓ and index k , and $S(k, \ell)^2$ is the conditional sample variance given the split, that is,

$$\begin{aligned} S(k, \ell)^2 &= \frac{1}{n-2} \sum_{i \leq k} (\tau_i - \hat{\tau}_L(k, \ell))^2 + \frac{1}{n-2} \sum_{i > k} (\tau_i - \hat{\tau}_R(k, \ell))^2 \\ &= \frac{1}{n-2} \left[\sum_{i=1}^n (\tau_i - n^{-1} \sum_{j=1}^n \tau_j)^2 - \frac{k(n-k)}{n} (\hat{\tau}_L(k, \ell) - \hat{\tau}_R(k, \ell))^2 \right]. \end{aligned}$$

Putting together, we see the *squared T-statistics metric* is a monotone transformation of the split criterion of previously studied estimators,

$$\begin{aligned} & n \frac{(\hat{\tau}_L(k, \ell) - \hat{\tau}_R(k, \ell))^2}{S(k, \ell)^2/k + S(k, \ell)^2/(n-k)} \\ &= n \frac{k(n-k)}{n} (\hat{\tau}_L(k, \ell) - \hat{\tau}_R(k, \ell))^2 \left(\frac{1}{n-2} \sum_{i=1}^n (\tau_i - n^{-1} \sum_{j=1}^n \tau_j)^2 - \frac{1}{n-2} \frac{k(n-k)}{n} (\hat{\tau}_L(k, \ell) - \hat{\tau}_R(k, \ell))^2 \right)^{-1}. \end{aligned}$$

Hence the split is always the same as the split by the split criterion studied in Section SA-3.1 and Section SA-3.2.

SA-3.4.2 Unbiasedness under Symmetric Error

Lemma SA-37 (Unbiasedness). *Suppose Assumption SA-2 holds, and $\varepsilon_i(0)$, $\varepsilon_i(1)$ are symmetrically distributed around zero. Then*

$$\mathbb{E}[\hat{\tau}_l^q(\mathbf{x}; K)] = \tau, \quad l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}, \quad q \in \{\text{NSS}, \text{X}\}, \quad K \geq 1,$$

and suppose \mathbf{t} is the node containing \mathbf{x} , then

$$\begin{aligned} \mathbb{E}[\hat{\tau}_l^{\text{HON}}(\mathbf{x}; K)] &= \tau - \tau \mathbb{P}(n(\mathbf{t}) = 0), \quad l \in \{\text{IPW}\}, \\ \mathbb{E}[\hat{\tau}_l^{\text{HON}}(\mathbf{x}; K)] &= \tau - \tau \mathbb{P}(n_0(\mathbf{t}) = 0 \text{ or } n_1(\mathbf{t}) = 0), \quad l \in \{\text{DIM}, \text{SSE}\}. \end{aligned}$$

SA-4 Proofs

SA-4.1 Proof of Theorem SA-1

First, we introduce some notations. Recall for $\ell \in [p]$, π_ℓ denotes the permutation such that $(x_{\pi_\ell(i)} : 1 \leq i \leq n)$ is non-decreasing. Define sample mean at the left and right leave at index $k \in [n]$ based on coordinate $\ell \in [p]$ by

$$\hat{\mu}_L(k, \ell) = \frac{1}{k} \sum_{i=1}^k y_{\pi_\ell(i)}, \quad \hat{\mu}_R(k, \ell) = \frac{1}{n-k} \sum_{i=k+1}^n y_{\pi_\ell(i)}, \quad k \in [n], \quad \ell \in [p].$$

We can check that minimizing the *sum of squares* criterion Equation (SA-2) is equivalent to maximizing the split criterion

$$(\hat{i}, \hat{j}) = \arg \max_{(i,j) \in [n] \times [p]} \mathcal{J}(i, j).$$

where

$$\mathcal{J}(k, \ell) = \frac{k(n-k)}{n} \left(\hat{\mu}_L(k, \ell) - \hat{\mu}_R(k, \ell) \right)^2, \quad k \in [n], \quad \ell \in [p].$$

Moreover, under the constant conditional mean assumption, Assumption SA-1 (1), we have that $\hat{\mu}_L(k, \ell) - \hat{\mu}_R(k, \ell) = \frac{1}{k} \sum_{i=1}^k \varepsilon_{\pi_\ell(i)} - \frac{1}{n-k} \sum_{i=k+1}^n \varepsilon_{\pi_\ell(i)}$. Hence we can w.l.o.g. replace y_i by ε_i in the definition of $\hat{\mu}_L$ and $\hat{\mu}_R$, that is,

$$\hat{\mu}_L(k, \ell) = \frac{1}{k} \sum_{i=1}^k \varepsilon_{\pi_\ell(i)}, \quad \hat{\mu}_R(k, \ell) = \frac{1}{n-k} \sum_{i=k+1}^n \varepsilon_{\pi_\ell(i)}, \quad k \in [n], \quad \ell \in [p].$$

The rest of the proof is organized as follows. In Section SA-4.1.1, we prove the results under $p = 1$, showing a strong approximation of the split criterion $(\mathcal{J}(k, 1) : k \in [n])$ by the square of a time-transformed Ornstein-Uhlenbeck (O-U) process, and studying the argmax of the split criterion through the argmax of the O-U process. In Section SA-4.1.2, we generalize to allow for $p \geq 1$. We show that the split criterion over different coordinates, that is, $(\mathcal{J}(k, \ell) : k \in [n])$ for different ℓ 's, are asymptotically independent. This reduces our problem to one-dimensional calculations, and the same technique of approximation by O-U process from

Section SA-4.1.1 can be used.

SA-4.1.1 Univariate Case

This the case when $p = 1$. For notational simplicity, define partial sums by

$$S_k = \sum_{i=1}^k \varepsilon_{\pi_1(i)}, \quad k \in [n].$$

By Csörgő and Horváth [1997, Equation A.4.37], we can define a sequence of Brownian bridges $\{B_n(t) : 0 \leq t \leq 1\}$ on a suitable probability space such that

$$\left| \max_{1 \leq k < n} \sqrt{\mathcal{J}(k, 1)} - \sup_{1/n \leq t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} \right| = \left| \max_{1 \leq k < n} \frac{\left| \frac{1}{\sqrt{n}} S_k - \frac{k}{n} \frac{1}{\sqrt{n}} S_n \right|}{\sqrt{(k/n)(1-k/n)}} - \sup_{1/n \leq t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} \right| = \epsilon_n, \quad (\text{SA-12})$$

where $\epsilon_n = o_{\mathbb{P}}((\log \log(n))^{-1/2})$. We note that while Csörgő and Horváth [1997, Equation A.4.37] bounds the approximation error of the maximum over the full range $1 \leq k < n$ as in (SA-12), its proof, which relies on invariance principles for partial sums of i.i.d. random variables, can be generalized to bound the approximation error over $1 \leq k < n^a$, $n^b < k < n$. Thus,

$$\left| \max_{1 \leq k < n^a, n^b < k < n} \frac{\left| \frac{1}{\sqrt{n}} S_k - \frac{k}{n} \frac{1}{\sqrt{n}} S_n \right|}{\sqrt{(k/n)(1-k/n)}} - \sup_{1/n \leq t < n^{a-1}, n^{b-1} < t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} \right| = \epsilon_n. \quad (\text{SA-13})$$

We note that the standardized Brownian bridge $\{B_n(t)/\sqrt{t(1-t)} : 0 < t < 1\}$ is distributionally equivalent to a time-transformed Ornstein-Uhlenbeck (O-U) process $\{U(\log(t/(1-t))) : 0 < t < 1\}$, where $\{U(t) : t \in \mathbb{R}\}$ is an O-U process with mean $\mathbb{E}[U(t)] = 0$ and covariance $\mathbb{E}[U(s)U(t)] = e^{-|s-t|/2}$ [Csörgő and Révész, 1981, Section 1.9], and thus

$$\begin{aligned} & \mathbb{P} \left(\sup_{1/n \leq t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} > \sup_{1/n \leq t < n^{a-1}, n^{b-1} < t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} + 2\epsilon_n \right) \\ &= \mathbb{P} \left(\sup_{-\log(n-1) \leq t \leq \log(n-1)} |U(t)| > \sup_{-\log(n-1) \leq t < \log(\frac{n^{a-1}-1}{1-n^{a-1}}), \log(\frac{n^{b-1}-1}{1-n^{b-1}}) < t \leq \log(n-1)} |U(t)| + 2\epsilon_n \right) \\ &= \mathbb{P} \left(\sup_{0 \leq t \leq 2 \log(n-1)} |U(t)| > \sup_{0 \leq t < \log(\frac{n^{a-1}-1}{1-n^{a-1}}), \log(\frac{n^{b-1}-1}{1-n^{b-1}}) < t \leq 2 \log(n-1)} |U(t)| + 2\epsilon_n \right), \quad (\text{SA-14}) \end{aligned}$$

where the last equality follows from stationarity of the process $|U(t)|$, the square of which is a Cox-Ingersoll-Ross (CIR) process [Göing-Jaesche and Yor, 2003]. Continuing from (SA-14), for any sequence u_n , we

have

$$\begin{aligned}
& \mathbb{P} \left(\sup_{0 \leq t \leq 2 \log(n-1)} |U(t)| > \sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}}), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |U(t)| + 2\epsilon_n \right) \\
& \geq \mathbb{P} \left(\sup_{0 \leq t \leq 2 \log(n-1)} |U(t)| \geq u_n, \sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}}), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |U(t)| < u_n - 2\epsilon_n \right) \\
& \geq \mathbb{P} \left(\sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}}), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |U(t)| < u_n - 2\epsilon_n \right) \\
& \quad - \mathbb{P} \left(\sup_{0 \leq t \leq 2 \log(n-1)} |U(t)| < u_n \right). \tag{SA-15}
\end{aligned}$$

Now, since $U(t)$ is a continuous, mean-zero Gaussian process, it induces a centered Gaussian measure on the space of all continuous functions on $[0, 2 \log(n-1)]$ equipped with the supremum norm (a separable Banach space). Thus, by the Gaussian correlation inequality [Latała and Matlak, 2017, Remark 3 (i)], we have that

$$\begin{aligned}
& \mathbb{P} \left(\sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}}), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |U(t)| < u_n - 2\epsilon_n \right) \\
& \geq \mathbb{P} \left(\sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}})} |U(t)| < u_n - 2\epsilon_n \right) \cdot \mathbb{P} \left(\sup_{\log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |U(t)| < u_n - 2\epsilon_n \right) \\
& = \mathbb{P} \left(\sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}})} |U(t)| < u_n - 2\epsilon_n \right) \cdot \mathbb{P} \left(\sup_{0 < t \leq \log(\frac{n^{1-b}(n-1)}{1-n^{b-1}})} |U(t)| < u_n - 2\epsilon_n \right), \tag{SA-16}
\end{aligned}$$

where the last equality follows from stationarity.

Remark SA-3. *The next step of our proof relies on a precise characterization of weak convergence for the suprema of a standardized empirical process, as studied in [Eicker, 1979]. However, Eicker [1979, Theorem 5] is incorrectly stated: the $2 \log(c)$ term appearing in the limiting probability should be $\log(c)$. This correction has important implications in our proof.*

By the Darling-Erdős Limit Theorem for the O-U process [Csörgő and Révész, 1981, Theorem 1.9.1] and [Eicker, 1979, Theorem 2.2 and the correct version of Theorem 5], for all $c > 0$ and $z \in \mathbb{R}$, we have

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \mathbb{P} \left(\sup_{0 \leq t \leq (c+o(1)) \log(n)} |U(t)| < \frac{2 \log \log(n) + (1/2) \log \log \log(n) + z - (1/2) \log(\pi)}{\sqrt{2 \log \log(n)}} \right) \\
& = \exp \left(-e^{-(z - \log(c))} \right). \tag{SA-17}
\end{aligned}$$

For a detailed proof of a generalized result on multidimensional O-U process, see Remark SA-1.

Let z^* maximize $z \mapsto \exp(-2e^{-(z - \log(2 - (b-a)))}) - \exp(-2e^{-(z - \log(2))})$, and set

$$u_n = \frac{2 \log \log(n) + (1/2) \log \log \log(n) + z^* - (1/2) \log(\pi)}{\sqrt{2 \log \log(n)}}.$$

We combine (SA-14), (SA-15), and (SA-16), and employ (SA-17) three times with $c = 2$ and $c = 2 - b$, and

$c = a$, together with the fact that $\epsilon_n = o_{\mathbb{P}}((\log \log(n))^{-1/2})$. We have that

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{1/n \leq t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} > \sup_{1/n \leq t < n^{a-1}, n^{b-1} < t \leq 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} + 2\epsilon_n \right) \\
& \geq \exp \left(-2e^{-(z^* - \log(a))} \right) \cdot \exp \left(-2e^{-(z^* - \log(2-b))} \right) - \exp \left(-2e^{-(z^* - \log(2))} \right) \\
& = \exp \left(-2e^{-(z^* - \log(2-(b-a)))} \right) - \exp \left(-2e^{-(z^* - \log(2))} \right) \\
& = \frac{b-a}{2} \left(1 - \frac{b-a}{2} \right)^{\frac{2}{b-a}-1} \\
& \geq \frac{b-a}{2e}.
\end{aligned} \tag{SA-18}$$

Remark SA-4. Alternatively, for any $0 < A < B < C$, we have

$$\mathbb{P} \left(\sup_{0 \leq t \leq C} |U(t)| > \sup_{0 \leq t \leq A, B \leq t \leq C} |U(t)| \right) = \frac{B-A}{C}. \tag{SA-19}$$

This can readily be shown using the fact that the absolute value of a zero-mean O-U process is stationary, Markov, and has continuous paths. Consequently, ignoring the stochastic error ϵ_n from approximating the split criterion (SA-3) by the square of a standardized Brownian bridge (not yet justified), using (SA-19), we can approximate the probability $\mathbb{P}(\max_{1 \leq k \leq n} \mathcal{J}(k, 1) > \max_{1 \leq k < n^a, n^b < k < n} \mathcal{J}(k, 1))$ by

$$\begin{aligned}
& \mathbb{P} \left(\sup_{0 \leq t \leq 2 \log(n-1)} |U(t)| > \sup_{0 \leq t < \log(\frac{n^a-1}{1-n^{a-1}}), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |U(t)| \right) \\
& = \frac{\log \frac{n^{b-1}(n-1)}{1-n^{b-1}} - \log(\frac{n^a-1}{1-n^{a-1}})}{2 \log(n-1)} \rightarrow \frac{b-a}{2}, \quad n \rightarrow \infty.
\end{aligned} \tag{SA-20}$$

SA-4.1.2 Multivariate Case

Now we prove for the general case of $p \geq 1$. As a sketch of the proof, we show that the split criterion over different coordinates, that is, $(\mathcal{J}(k, \ell) : k \in [n])$ for different ℓ 's, are asymptotically independent, which will imply

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b, \hat{j} = \ell) \\
& = \liminf_n \mathbb{P} \left(\max_k \mathcal{J}(k, 1) > \max_{k, j \neq 1} \mathcal{J}(k, j), \max_k \mathcal{J}(k, 1) > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1) \right) \\
& \geq \liminf_n \mathbb{P} \left(\max_k \mathcal{J}(k, 1) > z_n > \max_{k, j \neq 1} \mathcal{J}(k, j), \max_k \mathcal{J}(k, 1) > z_n > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1) \right) \\
& \stackrel{(*)}{=} \left(\liminf_n \mathbb{P} \left(\max_k \mathcal{J}(k, 1) < z_n \right) \right)^{p-1} \liminf_n \mathbb{P} \left(\max_k \mathcal{J}(k, 1) > z_n > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1) \right).
\end{aligned}$$

where in equality (*) we use asymptotic independence between $\mathcal{J}(\cdot, \ell)$ for different ℓ 's, and the last line are one-dimensional probabilities that can be handled by O-U process approximation like in Section SA-4.1.1.

To show the split criteria over different coordinates are asymptotically independent, we break down into two steps: In the first step, we show the partial sum process n indices and p coordinates can be approximated by another partial sum process with Gaussian increments (hence a Gaussian process), with the same covariance structure. In the second step, we show the covariance between the split criteria over

any two different coordinates and any indices are vanishing. Together with Gaussianity, this implies that the split criteria over different coordinates are asymptotically independent.

Step 1: Non-Gaussian to Gaussian Coupling.

For $1 \leq \ell \leq p$, denote by $H_n^\ell(\frac{k}{n})$ the scaled partial sum for the ℓ -th coordinate evaluated at time $\frac{k}{n}$, that is,

$$\begin{aligned} H_n^\ell\left(\frac{k}{n}\right) &= \sqrt{\frac{n}{k(n-k)}} \left\{ \sum_{i=1}^k \varepsilon_{\pi^\ell(i)} - \frac{k}{n} \sum_{i=1}^n \varepsilon_{\pi^\ell(i)} \right\} \\ &= \sqrt{\frac{n}{k(n-k)}} \sum_{i=1}^n \left(\mathbf{1}(\# \pi^\ell(i) \leq k) - \frac{k}{n} \right) \varepsilon_i, \end{aligned}$$

where $\# \pi^\ell : [n] \rightarrow [n]$ is the inverse mapping of π^ℓ .

We use a truncation argument for the proof. Fix $\varepsilon \in (0, 1)$. Take $r_n = \exp((\log n)^\varepsilon)$. And consider

$$\mathbf{C}_i = \sqrt{n} \left(\left(\sqrt{\frac{n}{k(n-k)}} \left(\mathbf{1}(\# \pi^\ell(i) \leq k) - \frac{k}{n} \right) : r_n \leq k \leq n - r_n \right)^\top : 1 \leq \ell \leq p \right)^\top \varepsilon_i,$$

where $\# \pi^\ell$ denotes the inverse mapping of π^ℓ . Notice that we add the \sqrt{n} factor for standardization. Then we can check that condition on \mathcal{B} , the σ -algebra generated by the p permutations π^1, \dots, π^p , \mathbf{C}_i 's are independent, and for all $1 \leq j \leq p(n - 2r_n)$, $1 \leq \ell \leq p$, we have

$$n^{-1} \sum_{i=1}^n \mathbb{E}[C_{ij}^2 | \mathcal{B}] = \frac{n}{k(n-k)} \left[k \left(\frac{n-k}{n} \right)^2 + (n-k) \left(\frac{k}{n} \right)^2 \right] = 1,$$

where we assume row j in \mathbf{C}_i corresponds to $\sqrt{n} \sqrt{\frac{n}{k(n-k)}} \left(\mathbf{1}(\# \pi^\ell(i) \leq k) - \frac{k}{n} \right)$. To use the coupling result [Chernozhukov et al., 2017, Theorem 2.1], we bound a few quantities: Suppose K_1 and K_2 are the universal constants given in the cited theorem,

$$\begin{aligned} L_n &= \max_{1 \leq j \leq p(n-2r_n)} \sum_{i=1}^n \mathbb{E}[|C_{ij}|^3 | \mathcal{B}] / n \\ &= \max_{1 \leq \ell \leq p} \max_{r_n \leq k \leq n-r_n} n^{3/2} \left(\frac{n}{k(n-k)} \right)^{3/2} \left[k(1 - k/n)^3 + (n-k)(-k/n)^3 \right] \mathbb{E}[|\varepsilon_i|^3] / n \\ &\lesssim \max_{1 \leq \ell \leq p} \max_{r_n \leq k \leq n-r_n} \frac{(n-2k)\sqrt{n}}{\sqrt{(n-k)nk}} \\ &\lesssim \sqrt{n/r_n}. \end{aligned} \tag{SA-21}$$

For notational simplicity, denote $\mathbf{P} = p(n - 2r_n)$. Take $\bar{L}_n = L_n$, then

$$\phi_n = K_2 \left(\frac{\bar{L}_n^2 \log^4(\mathbf{P})}{n} \right)^{-1/6} = K_2 \left(\frac{r_n}{\log^4(\mathbf{P})} \right)^{1/6}.$$

The definition of \mathbf{C}_i implies C_{ij} is $\sqrt{n/r_n}$ -exponential. Hence

$$\begin{aligned}
M_{n,X}(\phi_n) &= n^{-1} \sum_{i=1}^n \mathbb{E} \left[\max_{1 \leq j \leq \mathbf{P}} |C_{ij}|^3 \mathbf{1} \left(\max_{1 \leq j \leq \mathbf{P}} |C_{ij}| > \sqrt{n}/(4\phi_n \log(\mathbf{P})) \right) \middle| \mathcal{B} \right] \\
&\leq n^{-1} \sum_{i=1}^n \mathbb{E} \left[\max_{1 \leq j \leq \mathbf{P}} |C_{ij}|^6 \middle| \mathcal{B} \right]^{1/2} \mathbb{P} \left[\max_{1 \leq j \leq \mathbf{P}} |C_{ij}| > \sqrt{n}/(4\phi_n \log(\mathbf{P})) \middle| \mathcal{B} \right]^{1/2} \\
&\leq n^{-1} \sum_{i=1}^n \left[\sum_{1 \leq j \leq \mathbf{P}} \mathbb{E}[C_{ij}^6 | \mathcal{B}] \right]^{1/2} \left[\sum_{1 \leq j \leq \mathbf{P}} \mathbb{P}(|C_{ij}| > \sqrt{n}/(4\phi_n \log(\mathbf{P})) | \mathcal{B}) \right]^{1/2} \\
&\lesssim n^{-1} \sum_{i=1}^n (\mathbf{P}(n/r_n)^3)^{1/2} \left[\mathbf{P} \exp \left(- \frac{\sqrt{n}/(4\phi_n \log(\mathbf{P}))}{\sqrt{n/r_n}} \right) \right]^{1/2} \\
&\lesssim \mathbf{P}(n/r_n^3)^{1/2} \exp \left(- \frac{1}{4} \left(\frac{r_n}{\log \mathbf{P}} \right)^{1/3} \right) \\
&\lesssim n^{-2},
\end{aligned}$$

since $r_n = \exp((\log n)^\varepsilon)$ and ε, p are fixed. Now condition on \mathcal{B} , let $\mathbf{D}_i, 1 \leq i \leq n$ to be independent mean-zero Gaussian random vectors such that

$$\mathbf{D}_i \sim N(\mathbf{0}, \mathbb{E}[\mathbf{C}_i \mathbf{C}_i^\top | \mathcal{B}]), \quad \text{condition on } \mathcal{B}.$$

Then for each $1 \leq j \leq \mathbf{P}$, $1 \leq i \leq n$, we have D_{ij} is r_n^{-1} -subGaussian. Hence the same argument implies

$$M_{n,Y}(\phi_n) \lesssim n^{-2}.$$

[Chernozhukov et al., 2017, Theorem 2.1] then implies

$$\begin{aligned}
\sup_{A \in \mathcal{A}^{\text{re}}} \left| \mathbb{P} \left(\sum_{i=1}^n \mathbf{C}_i \in A \middle| \mathcal{B} \right) - \mathbb{P} \left(\sum_{i=1}^n \mathbf{D}_i \in A \middle| \mathcal{B} \right) \right| &\leq K_1 \left[\left(\frac{\bar{L}_n^2 \log^7(\mathbf{P})}{n} \right)^{1/6} + \frac{M_{n,X}(\phi_n) + M_{n,Y}(\phi_n)}{\bar{L}_n} \right] \\
&\lesssim \left(\frac{\log^7(\mathbf{P})}{r_n} \right)^{1/6} + \sqrt{\frac{r_n}{n}} \frac{1}{n^2} \\
&\lesssim \left(\frac{\log^7(n)}{r_n} \right)^{1/6}, \tag{SA-22}
\end{aligned}$$

where \mathcal{A}^{re} is the class of all rectangles A of the form

$$A = \{\mathbf{u} \in \mathbb{R}^{\mathbf{P}} : a_j \leq u_j \leq b_j, \forall j = 1, 2, \dots, \mathbf{P}\},$$

for some $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, 2, \dots, \mathbf{P}$. In particular, suppose $u_i, 1 \leq i \leq n$ are i.i.d $N(0, \mathbb{E}[\varepsilon_i^2])$ random variables, then \mathbf{D}_i can be taken such that

$$\mathbf{D}_i = \sqrt{n} \left(\left(\sqrt{\frac{n}{k(n-k)}} (\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) : r_n \leq k \leq n - r_n \right)^\top : 1 \leq \ell \leq p \right)^\top u_i.$$

The above result shows if we define

$$G_n^\ell\left(\frac{k}{n}\right) = \sqrt{\frac{n}{k(n-k)}} \left\{ \sum_{i=1}^k u_{\pi^\ell(i)} - \frac{k}{n} \sum_{i=1}^n u_{\pi^\ell(i)} \right\},$$

then Equation (SA-22) and unconditioning on \mathcal{B} , we get

$$\sup_{t_1, \dots, t_p \in \mathbb{R}} \left| \mathbb{P}\left(\max_{r_n \leq k \leq n-r_n} |H_n^\ell(k/n)| \leq t_\ell, 1 \leq \ell \leq p\right) - \mathbb{P}\left(\max_{r_n \leq k \leq n-r_n} |G_n^\ell(k/n)| \leq t_\ell, 1 \leq \ell \leq p\right) \right| \lesssim \left(\frac{\log^7(n)}{r_n}\right)^{1/6}.$$

Step 2: Gaussian to Gaussian Coupling.

For $1 \leq \ell \leq p$, denote by $G_n^\ell(\frac{k}{n})$ the partial sum for the ℓ -th coordinate evaluated at time $\frac{k}{n}$, that is,

$$G_n^\ell\left(\frac{k}{n}\right) = \sqrt{\frac{n}{k(n-k)}} \left\{ \sum_{i=1}^k u_{\pi^\ell(i)} - \frac{k}{n} \sum_{i=1}^n u_{\pi^\ell(i)} \right\}.$$

Then $\mathbf{G}_n = ((G_n^1(1/n), G_n^1(2/n), \dots, G_n^1(n/n))^\top, \dots, (G_n^p(1/n), G_n^p(2/n), \dots, G_n^p(n/n))^\top)^\top$. Then \mathbf{G}_n is a np -dimensional Gaussian random vector, and denote by Σ_n its covariance matrix. We want to show that Σ_n is close to one with covariance between different coordinates zero.

Consider two different coordinates, $\ell_1, \ell_2 \in [p]$. W.l.o.g, we can assume $\ell_1 = 1$ and $\ell_2 = 2$. Let $k, j \in [n]$. Denote by σ the sigma-algebra generated by π_1, \dots, π_p . Then

$$\begin{aligned} & \text{Cov}\left[G_n^1\left(\frac{k}{n}\right), G_n^2\left(\frac{j}{n}\right) \middle| \sigma\right] \\ &= \sqrt{\frac{n}{k(n-k)} \frac{n}{j(n-j)}} \left\{ \sum_{i=1}^k \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma] - \frac{j}{n} \sum_{i=1}^k \sum_{i'=1}^n \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma] \right. \\ & \quad \left. - \frac{k}{n} \sum_{i=1}^n \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma] + \frac{kj}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma] \right\} \\ &= \sqrt{\frac{n}{k(n-k)} \frac{n}{j(n-j)}} \frac{jk}{n} \left\{ \frac{n}{jk} \sum_{i=1}^k \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma] - 1 \right\}. \end{aligned}$$

To calculate $\sum_{i=1}^k \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma]$, we can first condition on π_1 , and let $\mathcal{J} = \{\pi_1(i) : 1 \leq i \leq k\}$. Observe that $\sum_{i=1}^k \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \sigma] = |\{i' \in [j] : \pi_1(i') \in \mathcal{J}\}|$. Now consider

$$f(\pi) = \frac{n}{jk} |\{i \in [j] : \pi(i) \in \mathcal{J}\}|,$$

π is a random permutation of $[n]$. Changing the order of the first j values of π does not change the value of $f(\pi)$, and $|f(\pi) - f(\pi^{s,t})| \leq \frac{n}{jk}$ for all π , $s \in \{1, \dots, j\}$, $t \in \{j+1, \dots, n\}$, where the permutation $\pi^{s,t}$ is obtained from π by transposition of its s th and t th coordinates. We will show later that w.l.o.g. we can

assume $j, k \leq \lceil n/2 \rceil$. Then by Lemma 2 from [El-Yaniv and Pechyony \[2009\]](#), for any $t \geq 0$,

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{n}{jk} \sum_{i=1}^k \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \boldsymbol{\sigma}] - 1\right| \geq t \mid \pi_1\right) \\ &= \mathbb{P}(|f(\pi_2) - \mathbb{E}[f(\pi_2)]| \geq t \mid \pi_1) \\ &\leq 2 \exp\left(-\frac{2t^2}{j(\frac{n}{jk})^2} \frac{n-1/2}{n-j} \left(1 - \frac{1}{2\max(j, n-j)}\right)\right). \end{aligned}$$

Since $\frac{n-1/2}{n-j} \left(1 - \frac{1}{2\max(j, n-j)}\right) \geq 1 - \frac{1}{n}$, we can marginalize over π_1 and uncondition on $\boldsymbol{\sigma}$ to get there exists a positive constant C such that for n large enough, for all $j, k \in [n]$,

$$\left| \frac{n}{jk} \sum_{i=1}^k \sum_{i'=1}^j \mathbb{E}[u_{\pi_1(i)} u_{\pi_2(i')} | \boldsymbol{\sigma}] - 1 \right| \leq C \frac{n}{\sqrt{jk}}.$$

which implies

$$|\text{Cov}[G_n^1(\frac{k}{n}), G_n^2(\frac{j}{n})]| \leq C \sqrt{\frac{jk}{(n-k)(n-j)}} \frac{n}{\sqrt{jk}} \frac{1}{\sqrt{k}} \leq C k^{-1/2}. \quad (\text{SA-23})$$

The reduction to $j, k \leq \lceil n/2 \rceil$ is because

$$\begin{aligned} G_n^\ell\left(\frac{k}{n}\right) &= \sqrt{\frac{n}{k(n-k)}} \left\{ \sum_{i=1}^k u_{\pi^\ell(i)} - \frac{k}{n} \sum_{i=1}^n u_{\pi^\ell(i)} \right\} \\ &= -\sqrt{\frac{n}{k(n-k)}} \left\{ \sum_{i=k+1}^n u_{\pi^\ell(i)} - \frac{n-k}{n} \sum_{i=1}^n u_{\pi^\ell(i)} \right\}. \end{aligned}$$

Now consider a np -dimensional mean-zero Gaussian random vector

$$\mathbf{Z}_n = ((Z_n^1(1/n), Z_n^1(2/n), \dots, Z_n^1(n/n))^\top, \dots, (Z_n^p(1/n), Z_n^p(2/n), \dots, Z_n^p(n/n))^\top)^\top,$$

where for each $1 \leq \ell \leq p$, $(Z_n^\ell(1/n), Z_n^\ell(2/n), \dots, Z_n^\ell(n/n))^\top$ has the same joint distribution as the partial sum random vector $(G_n^\ell(1/n), G_n^\ell(2/n), \dots, G_n^\ell(n/n))^\top$, and for any $\ell \neq \ell'$ and any $j, k \in [n]$,

$$\text{Cov}[Z_n^\ell(j/n), Z_n^{\ell'}(k/n)] = 0.$$

Denote by $\boldsymbol{\Gamma}_n$ the covariance matrix of \mathbf{Z}_n . We want to show $\boldsymbol{\Gamma}_n$ is close to $\boldsymbol{\Sigma}_n$. For a tight control on the rate of convergence, consider the truncated random vector,

$$\begin{aligned} T_{r_n}(\mathbf{G}_n) &= ((G_n^\ell(k/n) : r_n \leq k \leq n - r_n)^\top : 1 \leq \ell \leq p)^\top, \\ T_{r_n}(\mathbf{Z}_n) &= ((Z_n^\ell(k/n) : r_n \leq k \leq n - r_n)^\top : 1 \leq \ell \leq p)^\top. \end{aligned}$$

Also by an abuse of notations, denote by $T_{r_n}(\boldsymbol{\Sigma}_n)$ and $T_{r_n}(\boldsymbol{\Gamma}_n)$ the covariance matrix of $T_{r_n}(\mathbf{G}_n)$ and $T_{r_n}(\mathbf{Z}_n)$, respectively. Then Equation (SA-23) implies

$$\|T_{r_n}(\boldsymbol{\Sigma}_n) - T_{r_n}(\boldsymbol{\Gamma}_n)\|_{\max} = O(r_n^{-1/2}). \quad (\text{SA-24})$$

Additionally, we can lower bound the variance of each item of $T_{r_n}(\mathbf{Z}_n)$ by the following conditioning argument: Condition on the permutations π_ℓ , $1 \leq \ell \leq p$, then

$$\begin{aligned}
\mathbb{V}[\mathbf{Z}_n^\ell(k/n)|\pi_\ell, 1 \leq \ell \leq p] &= \mathbb{V}[\mathbf{G}_n^\ell(k/n)|\pi_\ell, 1 \leq \ell \leq p] \\
&= \mathbb{V}\left[\sqrt{\frac{n}{k(n-k)}}\left(\sum_{i=1}^k u_{\pi_\ell(i)} - \frac{k}{n} \sum_{i=1}^n u_{\pi_\ell(i)}\right) \middle| \pi_\ell, 1 \leq \ell \leq p\right] \\
&= \mathbb{V}\left[\sqrt{\frac{n}{k(n-k)}}\left(\sum_{i=1}^k u_i - \frac{k}{n} \sum_{i=1}^n u_i\right)\right] \\
&= 1, \quad 1 \leq k < n, 1 \leq \ell \leq p,
\end{aligned}$$

where in the third line, we have used the fact that condition on π_ℓ , $1 \leq \ell \leq p$, $(u_{\pi_\ell(i)})_{i \in [n]}$'s are i.i.d. $\mathcal{N}(0, 1)$. By the Gaussian-to-Gaussian Comparison result [Chernozhuov et al., 2022, Proposition 2.1],

$$\sup_{\mathbf{y} \in \mathbb{R}^{pT(n)}} |\mathbb{P}(T_{r_n}(\mathbf{G}_n) \leq \mathbf{y}) - \mathbb{P}(T_{r_n}(\mathbf{Z}_n) \leq \mathbf{y})| \leq C \log(n) \|T_{r_n}(\mathbf{\Sigma}_n) - T_{r_n}(\mathbf{\Gamma}_n)\|_{\max},$$

where C is an absolute constant, and $T(n) = \lceil n - r_n \rceil - \lfloor r_n \rfloor$. Combining with Equation (SA-24) and taking $\mathbf{y} = (t_1 \mathbf{1}^\top, \dots, t_p \mathbf{1}^\top)$, $\mathbf{y} = -(t_1 \mathbf{1}^\top, \dots, t_p \mathbf{1}^\top)$ separately with $\mathbf{1}$ a vector of $T(n)$ 1's, we get

$$\begin{aligned}
&\sup_{t_1, \dots, t_p \in \mathbb{R}} \left| \mathbb{P}\left(\max_{r_n \leq k \leq n-r_n} |G_n^\ell(k/n)| \leq t_\ell, 1 \leq \ell \leq p\right) - \mathbb{P}\left(\max_{r_n \leq k \leq n-r_n} |Z_n^\ell(k/n)| \leq t_\ell, 1 \leq \ell \leq p\right) \right| \\
&= O(\log(n) r_n^{-1/2}).
\end{aligned} \tag{SA-25}$$

Step 3: Reduction of calculations of one-dimensional O-U process

As in the previous two sections, fix $\varepsilon > 0$, and take $r_n = \exp((\log n)^\varepsilon)$. Let $\mathcal{E} = \{\exists \ell \in [p] : \arg \max_k \mathcal{J}(k, \ell) < r_n \text{ or } \arg \max_k \mathcal{J}(k, \ell) > n - r_n\}$. Then by [Csörgő and Horváth, 1997, proof of Theorem A.4.2], and a union bound argument, we have

$$\mathbb{P}(\mathcal{E}) \leq \sum_{\ell=1}^p \mathbb{P}(\arg \max_k \mathcal{J}(k, \ell) < r_n \text{ or } \arg \max_k \mathcal{J}(k, \ell) > n - r_n) = o(1).$$

Hence we can effectively restrict the candidates of $\arg \max$ to $[r_n, n - r_n]$. W.l.o.g., we consider coordinate $\ell = 1$, and

$$\begin{aligned}
&\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b, \hat{j} = \ell) \\
&= \mathbb{P}\left(\max_{k \in [n]} \mathcal{J}(k, 1) > \max_{k, j \neq 1} \mathcal{J}(k, j), \max_{k \in [n]} \mathcal{J}(k, 1) > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1)\right) \\
&\geq \mathbb{P}\left(\max_{k \in [n]} \mathcal{J}(k, 1) > \max_{k, j \neq 1} \mathcal{J}(k, j), \max_{k \in [n]} \mathcal{J}(k, 1) > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1), \mathcal{E}^c\right) - \mathbb{P}(\mathcal{E}) \\
&\geq \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} \mathcal{J}(k, 1) > \max_{k, j \neq 1} \mathcal{J}(k, j), \max_{k \in [r_n, n-r_n]} \mathcal{J}(k, 1) > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1)\right) - 2\mathbb{P}(\mathcal{E}) \\
&\geq \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} \mathcal{J}(k, 1) > \max_{k, j \neq 1} \mathcal{J}(k, j), \max_{k \in [r_n, n-r_n]} \mathcal{J}(k, 1) > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1)\right) + o(1).
\end{aligned}$$

Now we can using the *coupling result* developed previously. Using our notation, we have $\mathcal{J}(k, \ell) =$

$(H_n^\ell(k/n))^2$. Hence

$$\begin{aligned}
& \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} \mathcal{J}(k, 1) > \max_{j \neq 1, k \in [r_n, n-r_n]} \mathcal{J}(k, j), \max_{k \in [r_n, n-r_n]} \mathcal{J}(k, 1) > \max_{k \notin [n^a, n^b]} \mathcal{J}(k, 1)\right) \\
&= \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} |H_n^1\left(\frac{k}{n}\right)| > \max_{\ell \neq 1, k \in [r_n, n-r_n]} |G_n^\ell\left(\frac{k}{n}\right)|, \max_{k \in [r_n, n-r_n]} |H_n^1\left(\frac{k}{n}\right)| > \max_{k \notin [n^a, n^b]} |H_n^1\left(\frac{k}{n}\right)|\right) \\
&\geq \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} |H_n^1\left(\frac{k}{n}\right)| > z > \max_{\ell \neq 1, k \in [r_n, n-r_n]} |H_n^\ell\left(\frac{k}{n}\right)|, \max_{k \in [r_n, n-r_n]} |H_n^1\left(\frac{k}{n}\right)| > z > \max_{k \notin [n^a, n^b]} |H_n^1\left(\frac{k}{n}\right)|\right) \\
&\geq \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} |Z_n^1\left(\frac{k}{n}\right)| > z > \max_{\ell \neq 1, k \in [r_n, n-r_n]} |Z_n^\ell\left(\frac{k}{n}\right)|, \max_{k \in [r_n, n-r_n]} |Z_n^1\left(\frac{k}{n}\right)| > z > \max_{k \notin [n^a, n^b]} |Z_n^1\left(\frac{k}{n}\right)|\right) \\
&\quad + O(\log(n)^{7/6} r_n^{-1/6}),
\end{aligned}$$

where we have used Lemma SA-17 and Lemma SA-18. Since we choose $r_n = \exp((\log n)^\varepsilon)$, we have $\log(n)^{7/6} r_n^{-1/6} = o(1)$. It then follows from independence and symmetry between Z_n^ℓ 's across different ℓ 's that

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} |Z_n^1\left(\frac{k}{n}\right)| > z > \max_{\ell \neq 1, k \in [r_n, n-r_n]} |Z_n^\ell\left(\frac{k}{n}\right)|, \max_{k \in [r_n, n-r_n]} |Z_n^1\left(\frac{k}{n}\right)| > z > \max_{k \notin [n^a, n^b]} |Z_n^1\left(\frac{k}{n}\right)|\right) \\
&\geq \liminf_{n \rightarrow \infty} \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} |Z_n^1\left(\frac{k}{n}\right)| < z\right)^{p-1} \mathbb{P}\left(\max_{k \in [r_n, n-r_n]} |Z_n^1\left(\frac{k}{n}\right)| > z > \max_{k \in [r_n, n-r_n], k \notin [n^a, n^b]} |Z_n^1\left(\frac{k}{n}\right)|\right) \\
&\geq \sup_z \exp\left(-2(p-1)e^{-(z-\log(2))}\right) \left(\exp\left(-2e^{-(z-\log(2-(b-a)))}\right) - \exp\left(-2e^{-(z-\log(2))}\right)\right) \\
&= \frac{b-a}{2p} \left(1 - \frac{b-a}{2p}\right)^{\frac{2p}{b-a}-1} \\
&\geq \frac{b-a}{2pe},
\end{aligned}$$

where the third line is by similar calculation as in Section SA-4.1.1. Putting together, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i} \leq n^b, \hat{j} = \ell) \geq \frac{b-a}{2pe},$$

and by symmetry, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n - n^a \leq \hat{i} \leq n - n^b, \hat{j} = \ell) \geq \frac{b-a}{2pe}.$$

SA-4.2 Proof of Remark SA-1

Taking $T = c \log(n)$ in [Horváth, 1993, Lemma 2.1], we have

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\sup_{0 \leq t \leq c \log(n)} N(\log(n)) \leq \frac{z + b_d(c \log(n))}{a(c \log(n))}\right) = \exp(-e^{-z}).$$

Now we expand the term $\frac{z + b_d(c \log(n))}{a(c \log(n))}$. For notational simplicity, denote

$$L = \log \log n, \quad A = \log c, \quad L \rightarrow \infty \quad (n \rightarrow \infty).$$

First, we present some elementary expansions,

$$\begin{aligned}\sqrt{2(A+L)} &= \sqrt{2L} \sqrt{1 + \frac{A}{L}} = \sqrt{2L} \left(1 + \frac{A}{2L} - \frac{A^2}{8L^2} + O(L^{-3})\right), \\ \frac{1}{\sqrt{2(A+L)}} &= \frac{1}{\sqrt{2L}} \left(1 - \frac{A}{2L} + \frac{3A^2}{8L^2} + O(L^{-3})\right), \\ \log(L+A) &= \log L + \frac{A}{L} - \frac{A^2}{2L^2} + O(L^{-3}).\end{aligned}$$

Now we expand the terms for the numerator $b_d(c \log(n))$,

$$\begin{aligned}N_1 &= z + 2A + 2L + \frac{d}{2} \log(\log(c \log n)) - \log \Gamma(d/2), \\ N_2 &= z + 2A + 2L + \frac{d}{2} \log L - \log \Gamma(d/2), \\ N_3 &= z + A + 2L + \frac{d}{2} \log L - \log \Gamma(d/2).\end{aligned}$$

Then

$$\begin{aligned}& \frac{z + b_d(c \log n)}{a(c \log(n))} - \frac{z + \log(c) + b_d(\log n)}{a(\log(n))} \\ &= \frac{N_1}{\sqrt{2(A+L)}} - \frac{N_3}{\sqrt{2L}} \\ &= N_1 \left(\frac{1}{\sqrt{2(A+L)}} - \frac{1}{\sqrt{2L}} \right) + \frac{1}{\sqrt{2L}} (N_1 - N_3) \\ &= N_1 \frac{1}{\sqrt{2L}} \left(-\frac{A}{2L} + \frac{3A^2}{8L^2} + O(L^{-3}) \right) + \frac{1}{\sqrt{2L}} \left(\frac{d}{2} \left(\frac{A}{L} - \frac{A^2}{2L^2} + O(L^{-3}) \right) + A \right).\end{aligned}$$

Since $N_1 = 2L + O(\log \log \log(n))$, we have

$$\frac{z + b_d(c \log n)}{a(c \log(n))} - \frac{z + \log(c) + b_d(\log n)}{a(\log(n))} = \frac{3A^2}{4\sqrt{2}L^{3/2}} + \frac{dA}{2\sqrt{2}L^{3/2}} + o(L^{-3/2}) = o(L^{-1/2}).$$

Since $a(\log(n)) = \Theta(L^{1/2})$, we have

$$\begin{aligned}& \mathbb{P} \left(\sup_{0 \leq t \leq c \log(n)} N(\log(n)) \leq \frac{z + c \log(n) + b_d(\log(n))}{a(\log(n))} \right) \\ &= \mathbb{P} \left(\sup_{0 \leq t \leq c \log(n)} N(\log(n)) \leq \frac{z + o(1) + b_d(c \log(n))}{a(c \log(n))} \right) \\ &= \mathbb{P} \left(a(c \log(n)) \sup_{0 \leq t \leq c \log(n)} N(\log(n)) - b_d(c \log(n)) \leq z + o(1) \right) \rightarrow \exp(-e^{-z}) \text{ as } n \rightarrow \infty,\end{aligned}$$

where the last line follows from convergence in distribution of $a(c \log(n)) \sup_{0 \leq t \leq c \log(n)} N(\log(n)) - b_d(c \log(n))$ to a continuous distribution and Slutsky's Theorem.

SA-4.3 Proof of Theorem SA-2

For simplicity, we denote $\hat{\mu}^{\text{NSS}}(\mathbf{x})$ by $\hat{\mu}(\mathbf{x})$. We divide the proofs into two parts, one for uniform estimation and one for pointwise results near the boundary.

Part 1: Inconsistency for Uniform Estimation Rates

For notational simplicity, introduce the *partial sum based on ordering for the ℓ 's coordinate*,

$$S(k, \ell) = \sum_{i=1}^k \varepsilon_{\pi_\ell(i)}, \quad k \in [n], \quad \ell \in [p],$$

and define the optimal index for splitting based on the ℓ 's coordinate by

$$\iota_\ell = \arg \max_{k \in [n]} \mathcal{J}(k, \ell), \quad \ell \in [p].$$

Consider the event

$$\begin{aligned} \mathbf{Imbalance}_\ell &= \{\hat{j} = \ell, \hat{i} < n^b \text{ or } \hat{i} > n - n^b\} \\ &= \{\max_k \mathcal{J}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}(k, j), \max_k \mathcal{J}(k, \ell) > \max_{k \in [n^b, n-n^b]} \mathcal{J}(k, \ell)\}, \quad \ell \in [p]. \end{aligned}$$

Consider the case $\hat{i} < n^b$ on $\mathbf{Imbalance}_\ell$. The other case where $\hat{i} > n - n^b$ can be dealt with by symmetry. Then

$$\begin{aligned} &\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}(\mathbf{x}) - \mu|^2 \\ &\geq \frac{S(\iota_\ell, \ell)^2}{\iota_\ell^2} \\ &\geq \frac{1}{\iota_\ell} \left[\frac{S(\iota_\ell, \ell)^2}{\iota_\ell} + \frac{(S(n, \ell) - S(\iota_\ell, \ell))^2}{n - \iota_\ell} - \frac{(S(n, \ell) - S(\iota_\ell, \ell))^2}{n - \iota_\ell} \right] \\ &\geq \frac{1}{\min\{\iota_\ell, n - \iota_\ell\}} \left(\max_{k \in [n]} \left(\frac{S(k, \ell)^2}{k} + \frac{(n - S(k, \ell))^2}{n - k} \right) - \max_{\lfloor n/2 \rfloor \leq k \leq n} \frac{S(k, \ell)^2}{k} - \max_{1 \leq k \leq \lfloor n/2 \rfloor} \frac{(n - S(k, \ell))^2}{n - k} \right). \end{aligned}$$

where the last line is because ι_ℓ is the index that maximize the split criterion based on the ℓ 's coordinate, i.e.,

$$\begin{aligned} \iota_\ell &= \arg \max_{k \in [n]} \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^k (y_{\pi_\ell(i)} - S(k, \ell)/k)^2 - \sum_{i=k+1}^n (y_i - (S(n, \ell) - S(k, \ell))/(n - k))^2 \\ &= \arg \max_{k \in [n]} \frac{S(k, \ell)^2}{k} + \frac{(S(n, \ell) - S(k, \ell))^2}{n - k}. \end{aligned}$$

Fix $\epsilon > 0$. Consider the events

$$\begin{aligned} A_\ell^\epsilon &= \left\{ \max_{k \in [n]} \frac{S(k, \ell)^2}{k} + \frac{(n - S(k, \ell))^2}{n - k} \geq (2 - \epsilon) \log \log(n) \right\}, \\ B_\ell^\epsilon &= \left\{ \max_{\lfloor n/2 \rfloor \leq k \leq n} \frac{S(k, \ell)^2}{k} + \max_{1 \leq k \leq \lfloor n/2 \rfloor} \frac{(n - S(k, \ell))^2}{n - k} \leq 2\epsilon \log \log(n) \right\}. \end{aligned}$$

By [Csörgö and Horváth, 1997, Theorem A.4.1] $\limsup_{n \rightarrow \infty} \mathbb{P}(A_\ell^\epsilon) = \limsup_{n \rightarrow \infty} \mathbb{P}(B_\ell^\epsilon) = 1$. Hence for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}(\mathbf{x}) - \mu|^2 \geq \frac{(2 - 3\epsilon) \log \log(n)}{n^b}\right) \geq \sum_{\ell=1}^p \mathbb{P}(\text{Imbalance}_\ell \cap A_\ell^\epsilon \cap B_\ell^\epsilon) \geq \frac{b}{e} + o(1),$$

where we have used the fact that Imbalance_ℓ 's are disjoint for different ℓ 's and Theorem SA-1. Equation (SA-7) then follows.

Part 2: Inconsistency for Points Near the Boundary

Consider the event

$$\begin{aligned} \text{Off}_\ell &= \{\hat{j} = \ell, \hat{i} \in [n^a, n^b]\} \\ &= \{\max_k \mathcal{J}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}(k, j), \max_k \mathcal{J}(k, \ell) > \max_{k \notin [n^a, n-n^b]} \mathcal{J}(k, \ell)\}, \quad \ell \in [p]. \end{aligned}$$

Since π_ℓ is the uniform permutation, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}(x_{\ell, \iota_\ell} \geq n^{a-1}) \geq \liminf_{n \rightarrow \infty} \mathbb{P}(x_{\ell, \pi_\ell(n^a)} \geq n^{a-1}) = 1.$$

Together with Theorem SA-1,

$$\mathbb{P}(\text{Off}_\ell, x_{\ell, \iota_\ell} \geq n^{a-1}) \geq \frac{b-a}{2pe} + o(1).$$

Then on the event Off_ℓ and $x_{\ell, \iota_\ell} \geq n^{a-1}$, for any $\mathbf{x} \in [0, 1]^p$ such that $x_\ell \leq n^{a-1}$, we have $x_\ell \leq x_{\ell, \iota_\ell}$, and

$$\begin{aligned} |\hat{\mu}(\mathbf{x}) - \mu|^2 &= \frac{S(\iota_\ell, \ell)^2}{\iota_\ell^2} \\ &= \frac{1}{\iota_\ell} \left(\frac{S(\iota_\ell, \ell)^2}{\iota_\ell} + \frac{(S(n, \ell) - S(\iota_\ell, \ell))^2}{n - \iota_\ell} - \frac{(S(n, \ell) - S(\iota_\ell, \ell))^2}{n - \iota_\ell} \right) \\ &\geq \frac{1}{\iota_\ell} \left(\max_{1 \leq k \leq n} \frac{S(k, \ell)^2}{k} + \frac{(S(n, \ell) - S(k, \ell))^2}{n - k} - \max_{1 \leq k \leq n^b} \frac{(S(n, \ell) - S(k, \ell))^2}{n - k} \right). \end{aligned}$$

By similar arguments as Part 1, we can show

$$\liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in \mathcal{X}_n} \mathbb{P}\left(|\hat{\mu}(\mathbf{x}) - \mu|^2 \geq \frac{(2 + o(1)) \log \log(n)}{n^b}\right) \geq \frac{b-a}{2pe},$$

which is Equation (SA-8).

SA-4.4 Proof of Theorem SA-3

Due to the recursive splitting and Theorem SA-1, the optimal split index \hat{i} at the K_n -th split ($K_n \geq 1$) also satisfies

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\hat{i} \leq n^b) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i}) \geq \frac{b}{2e}.$$

Hence the same argument as Part 1 in the proof of Theorem SA-2 leads to the result.

SA-4.5 Proof of Theorem SA-4

This follows directly from [Klusowski and Tian \[2024, Theorem 4.3\]](#), choosing $g^* \equiv \mu$ and $g \equiv \mu$, and changing the sub-Gaussian rate to the sub-exponential rate by choosing $U \asymp \log(n)$ instead of $U \asymp \sqrt{\log(n)}$ in the truncation argument step. The last statement follows from the proof of [Klusowski and Tian \[2024, Theorem 4.3\]](#).

SA-4.6 Proof of Theorem SA-5

Throughout the proof, we abbreviate the honest tree $\hat{\mu}^{\text{HON}}(\mathbf{x})$ by $\check{\mu}(\mathbf{x})$. Recall (\hat{i}, \hat{j}) denotes the optimal splitting index and coordinate for the decision stump. We use $(y_i, \mathbf{x}_i^\top)_{i=1}^M$ to denote $\mathcal{D}_{\text{HON},1}$, which we used to construct the causal tree. Denote by (\hat{i}, \hat{j}) the splitting index and coordinate at the K_n -th step.

Use $(\tilde{y}_i, \tilde{\mathbf{x}}_i^\top)_{i=1}^N$ to denote $\mathcal{D}_{\text{HON},2}$. By Definition SA-2, $n \lesssim M, N \lesssim n$. Then

$$\begin{aligned} \sup_{\mathbf{x} \in \mathcal{X}} |\check{\mu}(\mathbf{x}) - \mu| &\geq |\check{\mu}(\mathbf{0}) - \mu| \\ &= \left| \frac{\sum_{i=1}^N (\tilde{y}_i - \mu) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}), \hat{j}})}{\sum_{i=1}^N \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}), \hat{j}})} \right|. \end{aligned}$$

Since $\tilde{y}_i \perp \tilde{\mathbf{x}}_i$, condition on \hat{i}, \hat{j} and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)$, we have

$$\frac{\sum_{i=1}^N (\tilde{y}_i - \mu) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}), \hat{j}})}{\sum_{i=1}^N \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}), \hat{j}})} \stackrel{d}{=} \frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu),$$

where

$$\tilde{i} = \sum_{i=1}^N \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}), \hat{j}}).$$

By Marcinkiewicz–Zygmund inequality, for some positive absolute constant C , we have

$$\begin{aligned} \mathbb{E} \left[\left| \frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu) \right| \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}' \right] &\geq C \mathbb{E} \left[\left| \frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} \frac{(y_i - \mu)^2}{\tilde{i}} \right|^{1/2} \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}' \right] \\ &\geq C \mathbb{E} \left[\frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} \left| \frac{(y_i - \mu)^2}{\tilde{i}} \right|^{1/2} \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}' \right] \\ &\geq \frac{C \mathbb{E}[|y_i - \mu|]}{\tilde{i}^{1/2}}, \end{aligned}$$

where in the second to last line, we have used Jensen's inequality, and in the last line we have used \tilde{i} is measurable with respect to the σ -algebra generated by $\hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}'$. Then by Paley–Zygmund inequality, for

any $\theta \in (0, 1)$,

$$\begin{aligned}
& \mathbb{P}\left(\left|\frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu)\right| \geq \theta \frac{C\mathbb{E}[|y_i - \mu|]}{\tilde{i}^{1/2}} \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}'\right) \\
& \geq \mathbb{P}\left(\left|\frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu)\right| \geq \theta \mathbb{E}\left[\left|\frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu)\right| \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}'\right] \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}'\right) \\
& \geq (1 - \theta)^2 \frac{\mathbb{E}\left[\left|\frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu)\right| \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}'\right]^2}{\mathbb{E}\left[\left(\frac{1}{\tilde{i}} \sum_{i=1}^{\tilde{i}} (y_i - \mu)\right)^2 \middle| \hat{i}, \hat{j}, \mathbf{X}, \mathbf{X}'\right]} \\
& \geq C(1 - \theta)^2 \frac{\mathbb{E}[|y_i - \mu|^2]}{\mathbb{V}[y_i]}.
\end{aligned}$$

Now we want to obtain a high probability upper bound on \tilde{i} given ι . Let F be the cumulative distribution function of \mathbf{x}_i . Suppose $1 \leq k \leq N/2$. Then $F(\mathbf{x}_{(k)}) \sim \text{Beta}(k, M - k + 1)$. By a Bernstein bound for Beta variables [Skorski, 2023, Theorem 1], we have for all $\epsilon > 0$,

$$\mathbb{P}(F(\mathbf{x}_{(k)}) > k/M + \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2v + \frac{c\epsilon}{3}}\right),$$

where for large enough n ,

$$\begin{aligned}
v &= \frac{k(M - k + 1)}{(M + 1)^2(M + 2)} \leq 2\frac{k}{M^2}, \\
c &= \frac{2(M - 2k + 1)}{M(M + 2)} \leq \frac{2}{M}.
\end{aligned}$$

Hence with probability at least $1 - M^{-1}$,

$$F(\mathbf{x}_{(k)}) \leq k/M + 2\frac{\sqrt{\log(M)k}}{M} + 3\frac{\log(M)}{M}.$$

Condition on \mathbf{X} , $\mathbf{1}(\tilde{\mathbf{x}}_i \leq \mathbf{x}_{(k)})$'s are i.i.d Bernoulli($F(\mathbf{x}_{(k)})$). Hence condition on \mathbf{X} and \hat{i} , with probability at least $1 - N^{-1}$,

$$\tilde{i}/N = n^{-1} \sum_{i=1}^N \mathbf{1}(\tilde{\mathbf{x}}_i \leq \mathbf{x}_{(i)}) \leq F(\mathbf{x}_{(i)}) + 2\sqrt{\frac{\log(N)F(\mathbf{x}_{(i)})}{N}}.$$

Hence condition on the event $\hat{i} \leq M^b$, we have with probability at least $1 - 2N^{-1}$,

$$\tilde{i}/n \leq 4M^{b-1} \leq Cn^{b-1},$$

where C is some constant only depending on $\liminf_{n \rightarrow \infty} |D_{\text{HON},1}|/|\mathcal{D}_{\text{HON},2}|$ and $\limsup_{n \rightarrow \infty} |D_{\text{HON},1}|/|\mathcal{D}_{\text{HON},2}|$.

Due to the iterative partitioning, the conclusion for Theorem SA-1 holds not only for decision stump, but also for the splitting index at arbitrary depth K_n , that is, for any $b \in (0, 1)$, we have

$$\liminf_{M \rightarrow \infty} \mathbb{P}(\hat{i} \leq M^b) \geq \frac{b}{2e}.$$

Hence we have

$$\begin{aligned}
\mathbb{P}\left(|\hat{\mu}(\mathbf{0}) - \mu| \geq \theta \frac{C\mathbb{E}[|y_i - \mu|]}{Cn^{b/2}}\right) &\geq \sum_{k \leq M^b} \mathbb{P}\left(|\hat{\mu}(\mathbf{0}) - \mu| \geq \theta \frac{C\mathbb{E}[|y_i - \mu|]}{Cn^{b/2}} \middle| \hat{i} = k\right) \mathbb{P}(\hat{i} = k) \\
&\geq \sum_{k \leq M^b} \mathbb{P}\left(|\hat{\mu}(\mathbf{0}) - \mu| \geq \theta \frac{C\mathbb{E}[|y_i - \mu|]}{\hat{i}^{1/2}} \middle| \hat{i} = k\right) \mathbb{P}(\hat{i} = k) - 2n^{-1} \\
&\geq C(1 - \theta)^2 \frac{\mathbb{E}[|y_i - \mu|^2]}{\mathbb{V}[y_i]} \frac{b}{2e} - 2n^{-1}.
\end{aligned}$$

This proves the conclusion.

SA-4.7 Proof of Theorem SA-6

For notational simplicity, we use \mathbf{T} to denote the data-driven decision tree. We will follow the proof strategy from [Klusowski and Tian \[2024, Theorem 4.3\]](#) condition on $\mathcal{D}_{\mathbf{T}}$. Denote by \mathcal{G}_0 the class of constant functions. Decompose $\|\hat{\mu}(\mathbf{T}) - \mu\|^2 = E_1 + E_2$, where

$$E_1 = \|\hat{\mu}(\mathbf{T}) - \mu\|^2 - 2(\|y - \hat{\mu}(\mathbf{T})\|_{\mathcal{D}_{\mu}}^2 - \|y - \mu\|_{\mathcal{D}_{\mu}}^2) - \alpha - \beta,$$

and

$$E_2 = 2(\|y - \hat{\mu}(\mathbf{T})\|_{\mathcal{D}_{\mu}}^2 - \|y - \mu\|_{\mathcal{D}_{\mu}}^2) + \alpha + \beta.$$

Denote the partition for \mathbf{T} by \mathcal{P} . Since \mathcal{P} is independent to \mathcal{D}_{μ} , the bound (E.27) from [Klusowski and Tian \[2024\]](#) does not apply automatically. Instead, we consider \mathcal{G}_0 as the reference class. Given the partitions of \mathbf{T} , the values of leaf nodes are obtained by least-square projection using \mathcal{D}_{μ} . This immediately implies

$$\|y - \hat{\mu}(\mathbf{T})\|_{\mathcal{D}_{\mu}}^2 \leq \|y - \bar{y}\|_{\mathcal{D}_{\mu}}^2 \leq \|y - g\|_{\mathcal{D}_{\mu}}^2,$$

for any constant function $g \in \mathcal{G}_0$. Hence for all $g \in \mathcal{G}_0$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\mu}}[E_2 | \mathcal{D}_{\mathbf{T}}] &\leq 2\mathbb{E}_{\mathcal{D}_{\mu}}[\|y - g\|_{\mathcal{D}_{\mu}}^2 - \|y - \mu\|_{\mathcal{D}_{\mu}}^2 | \mathcal{D}_{\mathbf{T}}] + \alpha + \beta \\
&= 2\|g - \mu\|^2 + \alpha + \beta.
\end{aligned}$$

For the term E_1 , we first assume $|y_i| \leq U$. Observe that condition on $\mathcal{D}_{\mathbf{T}}$, $\hat{\mu}(\mathbf{T})$ is still a member of the class $\mathcal{G}_{n_{\mathbf{T}}}[\mathcal{P}]$, which is the collection of all piecewise constant functions (bounded by U) on the partition \mathcal{P} . Since for any $\varepsilon \in (0, 1)$,

$$N(\varepsilon U, \mathcal{G}_{n_{\mathbf{T}}}[\mathcal{P}], \|\cdot\|_{P_{X^{n_{\mu}}}}) \leq N(\varepsilon U, \mathcal{G}_{n_{\mathbf{T}}}[\mathcal{P}], \|\cdot\|_{\infty}) \leq \left(\frac{2}{\varepsilon}\right)^{2^{\kappa}},$$

we can still use [Györfi et al. \[2002, Theorem 11.4\]](#) and the same argument from Equation (B.30) to (B.33) in [Klusowski and Tian \[2024\]](#) to get

$$\mathbb{P}_{\mathcal{D}_{\mu}}(E_1 \geq 0 | \mathcal{D}_{\mathbf{T}}) \leq 14 \left(\frac{2U^2}{\beta}\right)^{2^{\kappa}} \exp\left(-\frac{\alpha n_{\mathbf{T}}}{2568U^4}\right).$$

The result then follows choosing $\alpha \asymp \frac{U^4 2^K \log(n)}{n}$ and $\beta \asymp \frac{U^2}{n}$, and truncation argument over the sub-exponential ε_i 's.

SA-4.8 Proof of Theorem SA-7

In this section, we prove Theorem SA-7. First, we define some notation related to the tree construction which will be used in the proofs. Let \tilde{n}_k be the number of observations in the node containing $x = 0$ at depth k , \tilde{i}_k be the CART split index of this node, and \tilde{j}_k be the CART split coordinate of this node, with $\tilde{n}_0 = n$ and $\tilde{i}_0 = \hat{i}$ (recall that \hat{i} is the split index for the decision stump (SA-5)). Then, the left-most cell at the k -th level can be expressed as $\mathbf{t} \cap [0, x_{\pi_{\tilde{j}, \tilde{j}}(\tilde{i}_{k-1})}]$ and $\tilde{n}_k = \tilde{i}_{k-1}$.

Lemma SA-38. *There exist $\delta \in (0, 1)$, $c > 1$, and a positive integer M such that for any depth $k \geq 1$ and $m \geq M$, we have $\mathbb{P}(\tilde{n}_k \leq m) \geq (1 - \delta) \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m) + \delta \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m^c)$.*

Proof. Observe that if m is a positive integer, then $\tilde{i}_{k-1} \mid \tilde{n}_{k-1} = m$ has the same distribution as $\tilde{i}_0 \mid \tilde{n}_0 = m$, because of the honest tree construction and Assumption SA-1. Therefore, we can apply (SA-6) to obtain

$$\mathbb{P}(m^a \leq \tilde{i}_{k-1} \leq m^b \mid \tilde{n}_{k-1} = m) \geq \delta > 0, \quad (\text{SA-26})$$

for some $\delta > 0$ and sufficiently large m . Hence, by (SA-26), we have for m sufficiently large,

$$\begin{aligned} & \mathbb{P}(\tilde{n}_k \leq m \mid m < \tilde{n}_{k-1} \leq m^{1/b}) \\ & \geq \min_{m < i \leq m^{1/b}} \mathbb{P}(i^a \leq \tilde{i}_{k-1} \leq i^b \mid \tilde{n}_{k-1} = i) \mathbb{P}(\tilde{n}_k \leq m \mid i^a \leq \tilde{i}_{k-1} \leq i^b) \\ & \geq \delta \min_{m < i \leq m^{1/b}} \mathbb{P}(\tilde{n}_k \leq m \mid i^a \leq \tilde{i}_{k-1} \leq i^b) \\ & \geq \delta \min_{m^a < i \leq m} \mathbb{P}(\tilde{n}_k \leq \tilde{i}_{k-1} \mid \tilde{i}_{k-1} = i) \\ & = \delta. \end{aligned} \quad (\text{SA-27})$$

Now, taking $c = 1/b$, note that (SA-27) implies Lemma SA-38 since, for m sufficiently large, we have

$$\mathbb{P}(\tilde{n}_k \leq m) = \mathbb{P}(\tilde{n}_k \leq m, \tilde{n}_{k-1} > m^c) + \mathbb{P}(\tilde{n}_k \leq m, \tilde{n}_{k-1} \leq m^c) \quad (\text{SA-28})$$

$$\geq \mathbb{P}(\tilde{n}_k \leq m, \tilde{n}_{k-1} \leq m^c) \quad (\text{SA-29})$$

$$= \mathbb{P}(\tilde{n}_k \leq m, \tilde{n}_{k-1} \leq m) + \mathbb{P}(\tilde{n}_k \leq m, m < \tilde{n}_{k-1} \leq m^c) \quad (\text{SA-30})$$

$$\geq \mathbb{P}(\tilde{n}_{k-1} \leq m) + \delta \cdot \mathbb{P}(m < \tilde{n}_{k-1} \leq m^c) \quad (\text{SA-31})$$

$$= (1 - \delta) \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m) + \delta \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m^c). \quad \square$$

Next, we use Lemma SA-38 to finish the proof of Theorem SA-7. The main idea is to establish that the terminal nodes in a shallow tree will be small with constant probability.

Proof of Theorem SA-7. For notational simplicity, we denote $\hat{\mu}^{\mathbf{x}}(\mathbf{x}; K)$ by $\tilde{\mu}(T_K)(\mathbf{x})$.

Define $n_\ell = N^{(1/c)^\ell}$, where $N = n/K_n$. We will show by induction that for any $k \geq 0$ and $\ell \geq 1$ such that $n_\ell \geq M$,

$$\mathbb{P}(\tilde{n}_k \leq n_\ell) \geq \sum_{k'=\ell}^k \binom{k'-1}{\ell-1} (1 - \delta)^{k'-\ell} \delta^\ell. \quad (\text{SA-32})$$

The base case of $k = 0$ is trivial since $\tilde{n}_0 = N$. Now, assume that for some fixed $k \geq 1$ and any $\ell' \geq 1$ such that $n_{\ell'} \geq M$, we have

$$\mathbb{P}(\tilde{n}_{k-1} \leq n_{\ell'}) \geq \sum_{k'=\ell'}^{k-1} \binom{k'-1}{\ell'-1} (1-\delta)^{k'-\ell'} \delta^{\ell'}. \quad (\text{SA-33})$$

If $\ell \geq 2$, then substituting our induction hypothesis (SA-33) with $\ell' = \ell$ and $\ell' = \ell - 1$ into Lemma SA-38, we get that

$$\mathbb{P}(\tilde{n}_k \leq n_\ell) \geq (1-\delta) \sum_{k'=\ell}^{k-1} \binom{k'-1}{\ell-1} (1-\delta)^{k'-\ell} \delta^\ell + \delta \sum_{k'=\ell-1}^{k-1} \binom{k'-1}{\ell-2} (1-\delta)^{k'-\ell+1} \delta^{\ell-1} \quad (\text{SA-34})$$

$$= \sum_{k'=\ell}^k \binom{k'-1}{\ell-1} (1-\delta)^{k'-\ell} \delta^\ell, \quad (\text{SA-35})$$

where we used Pascal's identity. This completes the inductive proof of (SA-32).

Let $X \sim \text{NB}(L, \delta)$, i.e., the number of independent trials, each occurring with probability δ , until L successes. Choose

$$L = \lceil \log_c \log_c(N) - \log_c \log_c(M) - 1 \rceil \asymp \log \log(N), \quad n_L = N^{(1/c)^L} \in [M, M^c].$$

By (SA-32) and Markov's inequality applied to the tail probability of X , we have that

$$\begin{aligned} \mathbb{P}(\tilde{n}_K \leq n_L) &\geq \sum_{k'=L}^K \binom{k'-1}{L-1} (1-\delta)^{k'-L} \delta^L \\ &= 1 - \mathbb{P}(X \geq K+1) \\ &\geq 1 - \frac{\mathbb{E}[X]}{K+1} \\ &= 1 - \frac{L}{\delta(K+1)} \\ &\geq \frac{1}{2}, \end{aligned} \quad (\text{SA-36})$$

as long as $K \geq 2L/\delta \gtrsim \log \log(N)$. By the Paley-Zygmund inequality [Petrov, 2007] and the fact that $\text{Var}(\tilde{\mu}(T_K)(0)) = \mathbb{E}[1/\tilde{n}_K] \leq 1$, we have

$$\mathbb{P}\left(|\tilde{\mu}(T_K)(0)| > \frac{\mathbb{E}[|\tilde{\mu}(T_K)(0)|]}{2}\right) \geq \frac{(\mathbb{E}[|\tilde{\mu}(T_K)(0)|])^2}{4\text{Var}(\tilde{\mu}(T_K)(0))} \geq \frac{(\mathbb{E}[|\tilde{\mu}(T_K)(0)|])^2}{4}. \quad (\text{SA-37})$$

By the honest construction of the tree and (SA-36), we have the lower bound

$$\begin{aligned} \mathbb{E}[|\tilde{\mu}(T_K)(0)|] &= \sum_{k=1}^n \mathbb{E}\left[\left|\frac{1}{k} \sum_{i=1}^k \tilde{y}_i\right|\right] \mathbb{P}(\tilde{n}_K = k) \\ &\geq \min_{k=1,2,\dots,\lceil n_L \rceil} \mathbb{E}\left[\left|\frac{1}{k} \sum_{i=1}^k \tilde{y}_i\right|\right] \mathbb{P}(\tilde{n}_K \leq \lceil n_L \rceil) \\ &\geq \frac{1}{2} \min_{k=1,2,\dots,\lceil n_L \rceil} \mathbb{E}\left[\left|\frac{1}{k} \sum_{i=1}^k \tilde{y}_i\right|\right]. \end{aligned} \quad (\text{SA-38})$$

Since a sum of independent random variables is almost surely constant if and only if the individual random variables are almost surely constant, it follows that the last expression in (SA-38) is bounded away from zero. Returning to (SA-37) completes the proof. \square

SA-4.9 Proof of Theorem SA-8

For simplicity, denote $\hat{\mu}^{\mathbf{x}}(\mathbf{x})$ by $\tilde{\mu}(T_K)(\mathbf{x})$, and $N = n/(K+1)$ denotes the sample size for each folds in the \mathbf{x} sample splitting scheme.

Let $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{2^K}$ denote the 2^K leaf nodes in the decision tree, (if a node cannot be further refined, we duplicate the split indices and values at the next level). And let N_1, N_2, \dots, N_{2^K} and m_1, m_2, \dots, m_{2^K} denote the number of observations and the Lebesgue measure of the 2^K leaf nodes, respectively. Note that $\vec{N} = (N_1, \dots, N_{2^K})$ are independent of the \tilde{y}_i data by the honest condition and the x_i data per Assumption SA-1.

Claim: Condition on \vec{N} , $m_k \sim \text{Beta}(N_k, N - N_k + 1)$

Thus, the IMSE can be bounded as follows:

$$\begin{aligned} \mathbb{E} \left[\int_{\mathcal{X}} (\tilde{\mu}(T_K)(x))^2 \mathbb{P}_x(dx) \right] &= \sum_{k=1}^{2^K} \mathbb{E} \left[m_k \left(\frac{\mathbf{1}(N_k > 0)}{N_k} \sum_{i=1}^n \tilde{y}_i \mathbf{1}(\mathbf{x}_i \in \mathbf{t}_k) \right)^2 \right] \\ &= \sum_{k=1}^{2^K} \mathbb{E} \left[\frac{m_k}{N_k} \mathbf{1}(N_k > 0) \right] \sigma^2 \\ &\leq \sum_{k=1}^{2^K} \mathbb{E} \left[\frac{1}{N+1} \right] \sigma^2 \\ &\leq \frac{2^{K+1}}{N+1} \sigma^2. \end{aligned}$$

Proof of Claim: We show by induction. Base Case: $K=1$. For decision stumps, for some coordinate $j \in [p]$, we have $m_1 = x_{(N_1)}$, and $m_2 = x_{(N_1+N_2)} - x_{(N_1)} = 1 - x_{(N_1)}$. By Assumption SA-1, the order statistics $x_{j,(i)}$ is independent to \vec{N} . Hence $m_k \sim \text{Beta}(N_k, N - N_k + 1)$, $k=1, 2$.

Induction Step: $K \geq 2$. Let $\mathbf{t}_l^{\text{prev}}$ be a $(K-1)$ -th level node, we annotate all relevant depth $K-1$ information with superscript prev. We already know condition on $N_1^{\text{prev}}, \dots, N_{2^{K-1}}^{\text{prev}}$, $m_l^{\text{prev}} \sim \text{Beta}(N_l^{\text{prev}}, N - N_l^{\text{prev}} + 1)$. Suppose $\mathbf{t}_l^{\text{prev}}$ is divided into $\mathbf{t}_{2l}, \mathbf{t}_{2l+1}$ with Lebesgue measure and number of observations given by m_{2l}, m_{2l+1} and N_{2l}, N_{2l+1} , respectively, and the split is based on coordinate $j \in [p]$. By Assumption SA-1, condition on $\mathbf{x}_i \in \mathbf{t}_l^{\text{prev}}$, $\mathbf{x}_i \sim \text{Uniform}(\mathbf{t}_l^{\text{prev}})$. Hence condition on $N_l^{\text{prev}}, N_{2l}$ and m_l^{prev} , we have $m_{2l}/m_l^{\text{prev}} \sim \text{Beta}(N_l^{\text{prev}}, N - N_l^{\text{prev}} + 1)$. Hence condition on $\vec{N} = (N_1, \dots, N_{2^K})$, we have $m_k \sim \text{Beta}(N_k, N - N_k + 1)$, $1 \leq k \leq 2^K$. Induction then concludes the proof.

SA-4.10 Proof of Corollary SA-9

This is an immediate corollary from Theorem SA-1.

SA-4.11 Proof of Corollary SA-10

This is an immediate corollary from Theorem SA-2.

SA-4.12 Proof of Corollary SA-11

This is an immediate corollary from Theorem SA-3.

SA-4.13 Proof of Corollary SA-12

This is an immediate corollary from Theorem SA-4.

SA-4.14 Proof of Corollary SA-13

This is an immediate corollary from Theorem SA-5.

SA-4.15 Proof of Corollary SA-14

This is an immediate corollary from Theorem SA-6.

SA-4.16 Proof of Corollary SA-15

This is an immediate corollary from Theorem SA-7.

SA-4.17 Proof of Corollary SA-16

This is an immediate corollary from Theorem SA-8.

SA-4.18 Proof of Lemma SA-17

Since the number of coordinate p is fixed, we can use a union bound over the approximation error for the p coordinates. Hence w.l.o.g. we can assume $p = 1$ and drop the second index on the coordinate ℓ from $\mathcal{J}^{\text{DIM}}(k, \ell)$ and $\mathcal{J}^{\text{IPW}}(k, \ell)$ everywhere. And throughout, we assume the data is already sorted so that

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$

Expand the square, we have for any $k = 1, 2, \dots, n$,

$$\mathcal{J}^{\text{DIM}}(k) - \mathcal{J}^{\text{IPW}}(k) = \frac{k(n-k)}{n} \underbrace{\left(\hat{\tau}_{t_L}^{\text{DIM}}(k) - \hat{\tau}_{t_R}^{\text{DIM}}(k) + \bar{\tau}_{t_L}^{\text{IPW}}(k) - \bar{\tau}_{t_R}^{\text{IPW}}(k) \right)}_{=: R_1(k)} \underbrace{\left(\hat{\tau}_{t_L}^{\text{DIM}}(k) - \hat{\tau}_{t_R}^{\text{DIM}}(k) - \bar{\tau}_{t_L}^{\text{IPW}}(k) + \bar{\tau}_{t_R}^{\text{IPW}}(k) \right)}_{=: R_2(k)}. \quad (\text{SA-39})$$

We focus on the case where $1 \leq k \leq \frac{n}{2}$, the other case where $\frac{n}{2} < k \leq n$ follow from symmetry. Consider the term $R_2(k)$. First, consider the term corresponding to i from 1 to k . The other term corresponding to i from $k+1$ to n can be handled similarly. Breaking down $y_i(1) = \mu_1(x_i) + \varepsilon_i(1)$ and $y_i(0) = \mu_0(x_i) + \varepsilon_i(0)$,

we have

$$\begin{aligned}
|R_2(k)| &= \left| \frac{\sum_{i=1}^k d_i y_i(1)}{\sum_{i=1}^k d_i} - \frac{1}{k} \sum_{i=1}^k \frac{d_i}{\xi} \varepsilon_i(1) - \frac{\sum_{i=1}^k (1-d_i) y_i(0)}{\sum_{i=1}^k (1-d_i)} + \frac{1}{k} \sum_{i=1}^k \frac{1-d_i}{1-\xi} \varepsilon_i(0) + \text{counterpart for } t_R \right| \\
&\leq \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| \cdot \left| \frac{1}{k} \sum_{i=1}^k \left(\frac{d_i}{\xi} - 1 \right) \right| + \left| \frac{\sum_{i=1}^k (1-d_i) \varepsilon_i(0)}{\sum_{i=1}^k (1-d_i)} \right| \cdot \left| \frac{1}{k} \sum_{i=1}^k \left(\frac{1-d_i}{1-\xi} - 1 \right) \right| \\
&\quad + \left| \frac{\sum_{i=k+1}^n d_i \varepsilon_i(1)}{\sum_{i=k+1}^n d_i} \right| \cdot \left| \frac{1}{n-k} \sum_{i=k+1}^n \left(\frac{d_i}{\xi} - 1 \right) \right| + \left| \frac{\sum_{i=k+1}^n (1-d_i) \varepsilon_i(0)}{\sum_{i=k+1}^n (1-d_i)} \right| \cdot \left| \frac{1}{n-k} \sum_{i=k+1}^n \left(\frac{1-d_i}{1-\xi} - 1 \right) \right| \\
&\quad + \left| \frac{\sum_{i=1}^k d_i \mu_1(x_i)}{\sum_{i=1}^k d_i} - \frac{\sum_{i=1}^k (1-d_i) \mu_0(x_i)}{\sum_{i=1}^k (1-d_i)} - \frac{\sum_{i=k+1}^n d_i \mu_1(x_i)}{\sum_{i=k+1}^n d_i} + \frac{\sum_{i=k+1}^n (1-d_i) \mu_0(x_i)}{\sum_{i=k+1}^n (1-d_i)} \right|. \quad (\text{SA-40})
\end{aligned}$$

Notice that Assumption SA-2 (ii) implies that the last term is zero. Since $x_i \perp d_i$, even though the data is ordered according to x_i , $\{d_i/\xi - 1 : 1 \leq i \leq n\}$ are i.i.d mean-zero with bounded second moment. By Theorem A.4.1 in Csörgő and Horváth [1997],

$$\max_{r_n \leq k < n-r_n} \sqrt{k} \cdot \left| \frac{1}{k} \sum_{i=1}^k \left(\frac{d_i}{\xi} - 1 \right) \right| = O_{\mathbb{P}}(\sqrt{\log \log(n)}).$$

Take $b_i = \sum_{1 \leq \ell \leq i} d_\ell$. By Equation (8) from Shorack and Smythe [1976], for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P} \left(\max_{r_n \leq k \leq n-r_n} \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| \geq \lambda \mid (d_i)_{1 \leq i \leq n} \right) &\leq 16 \sum_{r_n \leq i \leq n-r_n} \frac{d_i \mathbb{V}[\varepsilon_i(1)]}{b_i^2} \lambda^{-2} \\
&\leq 16 \sum_{i \geq b_{r_n}} \frac{1}{i^2} \lambda^{-2} \mathbb{V}[\varepsilon_i(1)] \\
&\leq \frac{8}{3} \pi^2 \lambda^{-2} \mathbb{V}[\varepsilon_i(1)] \frac{1}{b_{r_n}},
\end{aligned}$$

The assumption that $\liminf_{n \rightarrow \infty} \rho_n \log \log(n) = \infty$ implies $\liminf_{n \rightarrow \infty} r_n = \infty$. Hence

$$(b_{r_n})^{-1} = r_n^{-1} \left(\xi + \frac{1}{r_n} \sum_{i=1}^{r_n} (d_i - \xi) \right)^{-1} = O_{\mathbb{P}}(r_n^{-1}).$$

Hence uncondition on $(d_i)_{1 \leq i \leq n}$, and we have

$$\max_{r_n \leq k \leq n-r_n} \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| = O_{\mathbb{P}}(r_n^{-1/2}). \quad (\text{SA-41})$$

Hence

$$\max_{r_n \leq k < n-r_n} \sqrt{k} \cdot \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| \cdot \left| \frac{1}{k} \sum_{i=1}^k \left(\frac{d_i}{\xi} - 1 \right) \right| = O_{\mathbb{P}} \left(\sqrt{\frac{\log \log(n)}{r_n}} \right).$$

By similar arguments, we can show the same bound holds for other terms in the first two lines of Equa-

tion (SA-40). Hence

$$\max_{r_n \leq k < n-r_n} \sqrt{k} |R_2(k)| = O_{\mathbb{P}} \left(\sqrt{\frac{\log \log(n)}{r_n}} \right).$$

Under the assumption that $\mu_0 \equiv c_0$ and $\mu_1 \equiv c_1$, we have

$$\begin{aligned} R_1(k) &= \left| \frac{\sum_{i=1}^k d_i y_i}{\sum_{i=1}^k d_i} + \frac{1}{k} \sum_{i=1}^k \frac{d_i}{\xi} \varepsilon_i(1) - \frac{\sum_{i=1}^k (1-d_i) y_i}{\sum_{i=1}^k (1-d_i)} - \frac{1}{k} \sum_{i=1}^k \frac{1-d_i}{1-\xi} \varepsilon_i(0) + \text{counterpart for } t_R \right| \\ &= \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} + \frac{1}{k} \sum_{i=1}^k \frac{d_i}{\xi} \varepsilon_i(1) - \frac{\sum_{i=1}^k (1-d_i) \varepsilon_i(0)}{\sum_{i=1}^k (1-d_i)} - \frac{1}{k} \sum_{i=1}^k \frac{1-d_i}{1-\xi} \varepsilon_i(0) + \text{counterpart for } t_R \right|. \end{aligned}$$

By Equation (SA-41) and Theorem A.4.1 in Csörgö and Horváth [1997] for the terms $k^{-1} \sum_{i=1}^k \xi^{-1} d_i \varepsilon_i(1)$, $k^{-1} \sum_{i=1}^k (1-\xi)^{-1} (1-d_i) \varepsilon_i(0)$ and the counterparts for t_R , we have

$$\max_{r_n \leq k < n-r_n} \sqrt{k} |R_1(k)| = O_{\mathbb{P}} \left(\sqrt{\log \log(n)} \right).$$

Putting together the parts for R_1 and R_2 , we have

$$\max_{r_n \leq k < n-r_n} |\mathcal{J}^{\text{DIM}}(k) - \mathcal{J}^{\text{IPW}}(k)| = O_{\mathbb{P}} \left(\frac{\log \log(n)}{r_n^{1/2}} \right).$$

SA-4.19 Proof of Lemma SA-18

Since the number of coordinates p is fixed, we can use a union bound over the approximation error for the p coordinates. Hence w.l.o.g. we can assume $p = 1$ and drop the second index on the coordinate ℓ from $\mathcal{J}^{\text{DIM}}(k, \ell)$ and $\mathcal{J}^{\text{IPW}}(k, \ell)$ everywhere. And throughout, we assume the data is already sorted so that

$$x_1 \leq x_2 \leq \dots \leq x_n.$$

For $1 \leq k \leq s_n$ and $n-s_n \leq k \leq n$, Equations (SA-39) and (SA-40) still hold. W.l.o.g assume $1 \leq k \leq s_n$. First, we upper bound the IPW terms. Definition of s_n and Equation (A.4.3) in Csörgö and Horváth [1997] imply

$$\max_{1 \leq k \leq s_n} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \frac{d_i}{\xi} \varepsilon_i(1) \right| + \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \frac{1-d_i}{1-\xi} \varepsilon_i(0) \right| = O_{\mathbb{P}}(u_n), \quad (\text{SA-42})$$

with $u_n = \sqrt{\rho_n \log \log(n)}$. Also Equation (A.4.2) in Csörgö and Horváth [1997] imply

$$\max_{1 \leq k \leq s_n} \sqrt{k} \cdot \left| \frac{1}{n-k} \sum_{i=k+1}^n \frac{d_i}{\xi} \varepsilon_i(1) \right| + \sqrt{k} \cdot \left| \frac{1}{n-k} \sum_{i=k+1}^n \frac{1-d_i}{1-\xi} \varepsilon_i(0) \right| = O_{\mathbb{P}}(v_n), \quad (\text{SA-43})$$

where $v_n = \sqrt{\frac{s_n}{n-s_n} \log \log(n)}$. Again Equation (A.4.3) from Csörgő and Horváth [1997] imply that

$$\max_{1 \leq k \leq s_n} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \left(\frac{d_i}{\xi} - 1 \right) \right| = O_{\mathbb{P}}(u_n).$$

Take $b_i = \sum_{1 \leq \ell \leq i} d_\ell$. By Equation (8) from Shorack and Smythe [1976], for any $\lambda > 0$,

$$\begin{aligned} \mathbb{P} \left(\max_{1 \leq k \leq s_n} \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| \geq \lambda \mid (d_i)_{1 \leq i \leq n} \right) &\leq 16 \sum_{1 \leq i \leq s_n} \frac{d_i \mathbb{V}[\varepsilon_i(1)]}{b_i^2} \lambda^{-2} \\ &\leq 16 \sum_{1 \leq i \leq s_n} \frac{1}{i^2} \lambda^{-2} \mathbb{V}[\varepsilon_i(1)] \\ &\leq \frac{8}{3} \pi^2 \lambda^{-2} \mathbb{V}[\varepsilon_i(1)], \end{aligned}$$

Hence uncondition on $(d_i)_{1 \leq i \leq n}$, and we have

$$\max_{1 \leq k \leq s_n} \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| = O_{\mathbb{P}}(1).$$

It follows that

$$\max_{1 \leq k \leq s_n} \sqrt{k} \cdot \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} - \frac{1}{k} \sum_{i=1}^k \frac{d_i}{\xi} \varepsilon_i(1) \right| = \max_{1 \leq k \leq s_n} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \left(\frac{d_i}{\xi} - 1 \right) \cdot \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| = O_{\mathbb{P}}(u_n). \quad (\text{SA-44})$$

Putting together the above equation with Equation (SA-42) and using a similar argument for the control group,

$$\max_{1 \leq k \leq s_n} \sqrt{k} \cdot \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| + \sqrt{k} \cdot \left| \frac{\sum_{i=1}^k (1-d_i) \varepsilon_i(0)}{\sum_{i=1}^k (1-d_i)} \right| = O_{\mathbb{P}}(u_n). \quad (\text{SA-45})$$

Apply Equation (A.4.2) in Csörgő and Horváth [1997] for the partial sum with at least $n - s_n$ terms and using $\max_{1 \leq k \leq s_n} \left| \frac{1}{n-k} \sum_{i=k+1}^n (d_i - \xi) \right| = o_{\mathbb{P}}(1)$, we have

$$\begin{aligned} \max_{1 \leq k \leq s_n} \sqrt{k} \cdot \left| \frac{\sum_{i=k+1}^n d_i \varepsilon_i(1)}{\sum_{i=k+1}^n d_i} \right| &= \max_{1 \leq k \leq s_n} \sqrt{k} \cdot \left| \frac{n-k}{\sum_{i=k+1}^n d_i} \right| \cdot \left| \frac{1}{n-k} \sum_{i=k+1}^n d_i \varepsilon_i(1) \right| \\ &\leq \sqrt{\frac{s_n}{n-s_n}} \left(\xi + \min_{1 \leq k \leq s_n} \frac{1}{n-k} \sum_{i=k+1}^n (d_i - \xi) \right)^{-1} \cdot \max_{1 \leq k \leq s_n} \left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^n d_i \varepsilon_i(1) \right| \\ &= O_{\mathbb{P}}(v_n). \end{aligned} \quad (\text{SA-46})$$

The same bound hold for $\max_{1 \leq k \leq s_n} \sqrt{k} \cdot \left| \frac{\sum_{i=k+1}^n (1-d_i) \varepsilon_i(1)}{\sum_{i=k+1}^n (1-d_i)} \right|$ by a similar argument. Putting together Equations (SA-42), (SA-43), (SA-45), (SA-46), we have

$$\max_{\ell=1,2} \max_{1 \leq k \leq s_n} \sqrt{k} |R_\ell(k)| = O_{\mathbb{P}}(u_n + v_n).$$

From the decomposition in Equation (SA-40) and the symmetry for $k \in [1, s_n]$ and $k \in [n - s_n, n]$, the conclusion follows.

SA-4.20 Proof of Theorem SA-19

We break down the proofs into two steps.

Step 1: Approximation of reg-score by ipw-score

Let $0 < a < b < 1$. Let ρ_n be a sequence of real numbers taking values in $(0, 1)$ to be determined, and take $s_n = \exp((\log n)^{\rho_n})$. Then for large enough n , we have $s_n \leq n^a \leq n^b \leq n - s_n$. Consider the event $A_n := \{\exists \ell \in [p] : \max_{k \in [n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k \notin [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell)\}$. By Equation (A.4.18) from Csörgő and Horváth [1997],

$$\max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} = O_{\mathbb{P}}(\sqrt{\rho_n \log \log(n)}).$$

Then controlling the difference between $\mathcal{J}^{\text{IPW}}(k, \ell)$ and $\mathcal{J}^{\text{DIM}}(k, \ell)$ by Lemma SA-18,

$$\max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \mathcal{J}^{\text{DIM}}(k, \ell) = O_{\mathbb{P}}\left(\rho_n \log \log(n) + \frac{s_n}{n-s_n} \log \log(n)\right) \quad (\text{SA-47})$$

By Lemma SA-17 with the choice $r_n = s_n$,

$$\begin{aligned} \max_{s_n < k < n-s_n} \sqrt{\mathcal{J}^{\text{DIM}}(k, \ell)} &= \max_{s_n < k < n-s_n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} + O_{\mathbb{P}}\left(\frac{\log \log(n)^{1/2}}{s_n^{1/4}}\right) \\ &\geq \max_{1 \leq k \leq n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} - \max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} + O_{\mathbb{P}}\left(\frac{\log \log(n)^{1/2}}{s_n^{1/4}}\right). \end{aligned}$$

Equation (A.4.20) in Csörgő and Horváth [1997] imply that $(2 \log \log(n))^{-1/2} \max_{1 \leq k \leq n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} = 1 + o_{\mathbb{P}}(1)$ and $(2 \log \log(n))^{-1/2} \max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} = \rho_n(1 + o_{\mathbb{P}}(1))$. Hence

$$\max_{1 \leq k \leq n} \sqrt{\mathcal{J}^{\text{IPW}}(k, \ell)} \geq \sqrt{2 \log \log(n)} + O_{\mathbb{P}}(\sqrt{\rho_n \log \log(n)}) + O_{\mathbb{P}}\left(\frac{\log \log(n)^{1/2}}{s_n^{1/4}}\right) \quad (\text{SA-48})$$

Choose $\log \log \log(n) / \log \log(n) \ll \rho_n \ll 1$, then by Equation (SA-47) and (SA-48),

$$\max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \mathcal{J}^{\text{DIM}}(k, \ell) = o_{\mathbb{P}}(\log \log(n)), \text{ and } \max_{s_n \leq k \leq n-s_n} \mathcal{J}^{\text{DIM}}(k, \ell) = \sqrt{2 \log \log(n)}(1 + o_{\mathbb{P}}(1)).$$

Hence

$$\max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \mathcal{J}^{\text{DIM}}(k, \ell) = o_{\mathbb{P}}\left(\max_{s_n \leq k \leq n-s_n} \mathcal{J}^{\text{DIM}}(k, \ell)\right), \quad \ell \in [p],$$

which by a union bound implies

$$\limsup_{n \rightarrow \infty} \mathbb{P}(A_n) = 0.$$

Observe that on the event A_n^c , the argmax for \mathcal{J}^{DIM} should be inside $[s_n, n - s_n]$. Hence

$$\begin{aligned}
& \mathbb{P}\left(\exists \ell \in [p] : \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k \notin [n^a, n^b]} \mathcal{J}^{\text{DIM}}(k, \ell)\right) \\
& \geq \mathbb{P}\left(\exists \ell \in [p] : \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k \notin [n^a, n^b]} \mathcal{J}^{\text{DIM}}(k, \ell) \text{ and } A_n^c\right) - \mathbb{P}(A_n) \\
& \geq \mathbb{P}\left(\exists \ell \in [p] : \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \right. \\
& \quad \left. \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k \notin [n^a, n^b], k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell)\right) - 2\mathbb{P}(A_n).
\end{aligned}$$

Now we focus on the first term. By symmetry in the p coordinates,

$$\begin{aligned}
& \mathbb{P}\left(\exists \ell \in [p] : \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, \ell)\right) \\
& = p\mathbb{P}\left(\max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, 1) > \max_{j \neq 1} \mathcal{J}^{\text{DIM}}(k, j), \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, 1) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, 1)\right) \\
& \geq p \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, j) < z, \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, 1) > z > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, 1)\right) \\
& \geq p \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, j) < z, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, 1) < z\right) \\
& \quad - p\mathbb{P}\left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, j) < z, \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, 1) > z\right).
\end{aligned}$$

Then using the fact that $\bar{\mathcal{J}}^{\text{IPW}}(k, \ell)$ approximates $\mathcal{J}^{\text{DIM}}(k, \ell)$ from Lemma SA-17, we have

$$\begin{aligned}
& \mathbb{P}\left(\exists \ell \in [p] : \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{DIM}}(k, \ell)\right) \\
& \geq p \sup_{z \in \mathbb{R}} \mathbb{P}\left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z - v_n, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z - v_n\right) \\
& \quad - p\mathbb{P}\left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z + v_n, \max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) > z - v_n\right),
\end{aligned}$$

where $v_n = O_{\mathbb{P}}(\log \log(n) s_n^{-1/2})$.

Step 2: IPw-score approximation by Gaussian approximation

Observe that the choice $s_n = \exp(\log(n)^{\rho_n})$ for $\log \log \log \log(n) / \log \log(n) \ll \rho_n \ll 1$ implies $v_n =$

$o_{\mathbb{P}}((\log \log(n))^{-1/2})$. Let $\epsilon > 0$. Then

$$\begin{aligned}
& \sup_{z \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z - v_n, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z - v_n \right) \\
& - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z + v_n, \max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) > z - v_n \right) \\
& \geq \sup_{z \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z - \frac{\epsilon}{\sqrt{2 \log \log(n)}}, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z - \frac{\epsilon}{\sqrt{2 \log \log(n)}} \right) \\
& - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z + \frac{\epsilon}{\sqrt{2 \log \log(n)}}, \max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) > z - \frac{\epsilon}{\sqrt{2 \log \log(n)}} \right) \\
& - \mathbb{P}(|v_n| > \frac{\epsilon}{\sqrt{2 \log \log n}}).
\end{aligned}$$

Choosing $z_n(u) = \frac{2 \log \log(n) + 1/2 \log \log \log(n) + u - 1/2 \log(\pi)}{\sqrt{2 \log \log(n)}}$, and from the proof of Theorem SA-1, we have

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \sup_{z \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z - \frac{\epsilon}{\sqrt{2 \log \log n}}, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z - \frac{\epsilon}{\sqrt{2 \log \log n}} \right) \\
& - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z + \frac{\epsilon}{\sqrt{2 \log \log n}}, \max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) > z - \frac{\epsilon}{\sqrt{2 \log \log n}} \right) \\
& - \mathbb{P}(|v_n| > \frac{\epsilon}{\sqrt{2 \log \log n}}) \\
& \geq \liminf_{n \rightarrow \infty} \sup_{u \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z_n(u) - \frac{\epsilon}{\sqrt{2 \log \log n}}, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z_n(u) - \frac{\epsilon}{\sqrt{2 \log \log n}} \right) \\
& - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, j) < z_n(u) + \frac{\epsilon}{\sqrt{2 \log \log n}}, \max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) > z_n(u) - \frac{\epsilon}{\sqrt{2 \log \log n}} \right) \\
& - \mathbb{P}(|v_n| > \frac{\epsilon}{\sqrt{2 \log \log n}}), \\
& \geq \liminf_{n \rightarrow \infty} \sup_{u \in \mathbb{R}} \mathbb{P} \left(\max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z_n(u) - \frac{\epsilon}{\sqrt{2 \log \log n}} \right)^{p-1} \mathbb{P} \left(\max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z_n(u) - \frac{\epsilon}{\sqrt{2 \log \log n}} \right) \\
& - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) < z_n(u) + \frac{\epsilon}{\sqrt{2 \log \log n}} \right)^{p-1} \mathbb{P} \left(\max_{k \in [s_n, n-s_n]} \bar{\mathcal{J}}^{\text{IPW}}(k, 1) > z_n(u) - \frac{\epsilon}{\sqrt{2 \log \log n}} \right) \\
& \geq \sup_{u \in \mathbb{R}} \exp \left(-2(p-1)e^{-(u-\epsilon-\log(2))} \right) \exp \left(-2e^{-(u-\epsilon-\log(2-(b-a)))} \right) \\
& - \exp \left(-2(p-1)e^{-(u+\epsilon-\log(2))} \right) \exp \left(-2e^{-(u-\epsilon-\log(2))} \right).
\end{aligned}$$

Now let $\epsilon \downarrow 0$, and then all previous steps together implies

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \mathbb{P} \left(\exists \ell \in [p] : \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k \notin [n^a, n^b]} \mathcal{J}^{\text{DIM}}(k, \ell) \right) \\
& \geq \sup_{u \in \mathbb{R}} \exp \left(-2(p-1)e^{-(u-\log(2))} \right) \left(\exp \left(-2e^{-(u-\log(2-(b-a)))} \right) - \exp \left(-2e^{-(u-\log(2))} \right) \right) \\
& \geq \frac{b-a}{2e}.
\end{aligned}$$

SA-4.21 Proof of Theorem SA-20

The proofs follow the essentially same logic as the proof for Theorem SA-2, with some tricks for the random numerator in $\frac{\sum_{1 \leq i \leq k} d_i \varepsilon_i(1)}{\sum_{1 \leq i \leq k} d_i}$.

Part 1: Inconsistency for Uniform Estimation Rates

Denote the optimal index for splitting based on the ℓ 's coordinate by

$$\hat{i}_{\text{DIM},\ell} = \arg \max_{k \in [n]} \mathcal{J}^{\text{DIM}}(k, \ell), \quad \ell \in [p].$$

For notational simplicity, denote

$$\begin{aligned} \bar{\tau}_L^{\text{DIM}}(k, \ell) &= \tau_L^{\text{DIM}}(k, \ell) - \tau = \frac{\sum_{1 \leq i \leq k} d_{\pi_\ell(i)} \varepsilon_{\pi_\ell(i)}(1)}{\sum_{1 \leq i \leq k} d_{\pi_\ell(i)}} - \frac{\sum_{1 \leq i \leq k} (1 - d_{\pi_\ell(i)}) \varepsilon_{\pi_\ell(i)}(0)}{\sum_{1 \leq i \leq k} (1 - d_{\pi_\ell(i)})}, \\ \bar{\tau}_L^{\text{DIM}}(\ell) &= \bar{\tau}_L^{\text{DIM}}(\hat{i}_{\text{DIM},\ell}, \ell), \\ \bar{\tau}_R^{\text{DIM}}(k, \ell) &= \tau_R^{\text{DIM}}(k, \ell) - \tau = \frac{\sum_{k < i \leq n} d_{\pi_\ell(i)} \varepsilon_{\pi_\ell(i)}(1)}{\sum_{k < i \leq n} d_{\pi_\ell(i)}} - \frac{\sum_{k < i \leq n} (1 - d_{\pi_\ell(i)}) \varepsilon_{\pi_\ell(i)}(0)}{\sum_{k < i \leq n} (1 - d_{\pi_\ell(i)})}, \\ \bar{\tau}_R^{\text{DIM}}(\ell) &= \bar{\tau}_R^{\text{DIM}}(\hat{i}_{\text{DIM},\ell}, \ell), \end{aligned}$$

and consider the event

$$\text{Imblce}_\ell^{\text{DIM}} = \{\max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}^{\text{DIM}}(k, j), \max_k \mathcal{J}^{\text{DIM}}(k, \ell) > \max_{k \notin [n^b, n-n^b]} \mathcal{J}^{\text{DIM}}(k, \ell)\}, \quad \ell \in [p].$$

Since we assume $\mu_0 \equiv c_0$ and $\mu_1 \equiv c_1$ with $c_1 - c_0 = \tau$, we have on $\text{Imblce}_\ell^{\text{DIM}} \cap \{\hat{i}_\ell^{\text{DIM}} \leq n/2\}$,

$$\begin{aligned} \sup_{x \in \mathcal{X}} |\hat{\tau}(x) - \tau|^2 &\geq \bar{\tau}_L^{\text{DIM}}(\ell)^2 \\ &\geq \frac{1}{\min\{\hat{i}_{\text{DIM},\ell}, n - \hat{i}_{\text{DIM},\ell}\}} \left(\hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 - (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} \leq n/2) \right). \end{aligned} \tag{SA-49}$$

Take $\bar{\tau}^{\text{DIM}} = \frac{\hat{i}_{\text{DIM},\ell}}{n} \bar{\tau}_L^{\text{DIM}} + \frac{n - \hat{i}_{\text{DIM},\ell}}{n} \bar{\tau}_R^{\text{DIM}}$. Then

$$\begin{aligned} \hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 &\geq \hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 - n \bar{\tau}^{\text{DIM}} \\ &= \frac{\hat{i}_{\text{DIM},\ell} (n - \hat{i}_{\text{DIM},\ell})}{n} \left(\bar{\tau}_L^{\text{DIM}} - \bar{\tau}_R^{\text{DIM}} \right)^2 \end{aligned}$$

By Lemma SA-17 and Lemma SA-18 with $r_n = s_n = \exp((\log n)^{\rho_n})$ for $\log \log \log \log(n) / \log \log(n) \ll \rho_n \ll 1$,

$$\begin{aligned} \frac{\hat{i}_{\text{DIM},\ell} (n - \hat{i}_{\text{DIM},\ell})}{n} \left(\bar{\tau}_L^{\text{DIM}} - \bar{\tau}_R^{\text{DIM}} \right)^2 &= \frac{\hat{i}^{\text{IPW}}(n - \hat{i}^{\text{IPW}})}{n} \left(\bar{\tau}_L^{\text{IPW}} - \bar{\tau}_R^{\text{IPW}} \right)^2 + o_{\mathbb{P}}(\log \log(n)) \\ &= \max_{1 \leq k \leq n} \mathcal{J}^{\text{IPW}}(k) + o_{\mathbb{P}}(\log \log(n)). \end{aligned}$$

By Theorem A.4.1 in Csörgö and Horváth [1997], $\max_{1 \leq k \leq n} \bar{\mathcal{J}}^{\text{IPW}}(k) = 2 \log \log(n)(1 + o_{\mathbb{P}}(1))$. Moreover,

$$\begin{aligned} \hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} > n/2) &\leq \max_{k > n/2} k \cdot \left(\frac{\sum_{1 \leq i \leq k} d_{\pi_\ell(i)} \varepsilon_{\pi_\ell(i)}(1)}{\sum_{1 \leq i \leq k} d_{\pi_\ell(i)}} - \frac{\sum_{1 \leq i \leq k} (1 - d_{\pi_\ell(i)}) \varepsilon_{\pi_\ell(i)}(0)}{\sum_{1 \leq i \leq k} (1 - d_{\pi_\ell(i)})} \right)^2 \\ &\leq \max_{k > n/2} 2k \cdot \left(\frac{\sum_{1 \leq i \leq k} d_{\pi_\ell(i)} \varepsilon_{\pi_\ell(i)}(1)}{\sum_{1 \leq i \leq k} d_{\pi_\ell(i)}} \right)^2 + 2k \cdot \left(\frac{\sum_{1 \leq i \leq k} (1 - d_{\pi_\ell(i)}) \varepsilon_{\pi_\ell(i)}(0)}{\sum_{1 \leq i \leq k} (1 - d_{\pi_\ell(i)})} \right)^2. \end{aligned}$$

For simplicity in showing the upper bound for $\hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} > n/2)$, we assume π is the identity permutation. Take $b_i = \sum_{1 \leq j \leq i} d_j$. By Equation (8) from Shorack and Smythe [1976], for any $\lambda > 0$,

$$\begin{aligned} \mathbb{P} \left(\max_{k > n/2} \left| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right| \geq \lambda \middle| (d_i)_{1 \leq i \leq n} \right) &\leq 16 \sum_{i > n/2} \frac{d_i \mathbb{V}[\varepsilon_i(1)]}{b_i^2} \lambda^{-2} \\ &\leq 16 \sum_{i > b_{n/2}} \frac{1}{i^2} \lambda^{-2} \mathbb{V}[\varepsilon_i(1)] \\ &\leq \frac{8}{3} \pi^2 \lambda^{-2} \mathbb{V}[\varepsilon_i(1)] \frac{1}{b_{n/2}}, \end{aligned}$$

And since d_i 's are i.i.d with $\mathbb{E}[d_i] = \xi > 0$, we have

$$(b_{n/2})^{-1} = (n/2)^{-1} \left(\xi + \frac{2}{n} \sum_{i=1}^{n/2} (d_i - \xi) \right)^{-1} = O_{\mathbb{P}}(n^{-1}).$$

Hence uncondition on $(d_i)_{1 \leq i \leq n}$, and we have

$$\max_{k \geq n/2} k \cdot \left(\frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \right)^2 = O_{\mathbb{P}}(1) = o_{\mathbb{P}}(\log \log(n)).$$

By a similar term for control, and a symmetric argument for the right node,

$$\hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} > n/2) + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} \leq n/2) = o_{\mathbb{P}}(\log \log(n)).$$

Fix $\epsilon > 0$. Consider the events

$$\begin{aligned} A_\ell^\epsilon &= \left\{ \hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 \geq (2 - \epsilon) \log \log(n) \right\}, \\ B_\ell^\epsilon &= \left\{ \hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} > n/2) + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 \mathbf{1}(\hat{i}_{\text{DIM},\ell} \leq n/2) \leq 2\epsilon \log \log(n) \right\}. \end{aligned}$$

The above arguments show that $\liminf_{n \rightarrow \infty} \mathbb{P}(A_\ell^\epsilon) = \liminf_{n \rightarrow \infty} \mathbb{P}(B_\ell^\epsilon) = 1$. From Theorem SA-19,

$$\mathbb{P}(\text{Imblce}_\ell^{\text{DIM}}) \geq \frac{b}{pe}.$$

It then follows from a union bound argument that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b}{e}.$$

Part 2: Inconsistency for Points near the Boundary

Fix $\mathbf{z} \in \mathcal{X}$ such that $z_\ell \leq n^{a-1}$. Since the order statistics $x_{(n^a),\ell} = n^{a-1}(1 + o_{\mathbb{P}}(1))$, on the event $n^a \leq \hat{i}_{\text{DIM},\ell} \leq n^b$, if $z_\ell \leq (1 + o_{\mathbb{P}}(1))n^{a-1}$, then $z_\ell \leq x_{(n^a)} \leq x_{(\hat{i}_{\text{DIM},\ell}),\ell}$, and on the event $\text{Imblce}_\ell^{\text{DIM}}$,

$$\begin{aligned} |\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{z}) - \tau|^2 &= \bar{\tau}_L^{\text{DIM}}(\ell)^2 \geq \frac{1}{\hat{i}_{\text{DIM},\ell}} \left(\hat{i}_{\text{DIM},\ell} \bar{\tau}_L^{\text{DIM}}(\ell)^2 + (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 - (n - \hat{i}_{\text{DIM},\ell}) \bar{\tau}_R^{\text{DIM}}(\ell)^2 \right) \\ &\geq \frac{1}{\hat{i}_{\text{DIM},\ell}} \left(\max_{1 \leq k \leq n} k \bar{\tau}_L^{\text{DIM}}(k, \ell)^2 + (n - k) \bar{\tau}_R^{\text{DIM}}(k, \ell)^2 - \max_{k \leq n^b} (n - k) \bar{\tau}_R^{\text{DIM}}(k, \ell)^2 \right) \\ &\geq \frac{(2 + o_{\mathbb{P}}(1)) \log \log(n)}{\hat{i}_{\text{DIM},\ell}} \\ &\geq \frac{(2 + o_{\mathbb{P}}(1)) \log \log(n)}{n^b}, \end{aligned}$$

where the second to last line is due to a similar argument as in the proof of part 1. By a symmetry argument for the event $\{n - n^b \leq \hat{i}_{\text{DIM},\ell} \leq n - n^a\}$, we have

$$\liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in \mathcal{X}_n} \mathbb{P} \left(|\hat{\tau}_{\text{DIM}}^{\text{NSS}}(\mathbf{x}) - \tau| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log \log(n)} \right) \geq \frac{b - a}{2e},$$

where $\mathcal{X}_n = \{\mathbf{x} \in [0, 1]^p : x_j = o(1)n^{a-1} \text{ or } 1 - x_j = o(1)n^{a-1} \text{ for some } j \in [p]\}$, and $\sigma^2 = \mathbb{V}[\frac{d_i y_i(1)}{\xi} + \frac{(1-d_i)y_i(0)}{1-\xi}]$.

SA-4.22 Proof of Theorem SA-21

Due to the recursive splitting and Theorem SA-19, the optimal split index \hat{i}_{DIM} at the k -th split ($k \geq 1$) also satisfies

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\hat{i}_{\text{DIM}} \leq n^b) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i}_{\text{DIM}}) \geq \frac{b}{2e}.$$

Hence the same argument as Part 1 in the proof of Theorem SA-20 leads to the result.

SA-4.23 Proof of Theorem SA-22

For notational simplicity, denote $\hat{\tau}_{\text{DIM}}^{\text{NSS}}$ by $\hat{\tau}$, the data-driven partition \mathcal{D}_{T} by \mathcal{P} .

Reduction to least square prediction error.

Observe that the leaf nodes value coincide with a least square projection given \mathcal{P} : For $\mathbf{t} \in \mathcal{P}$, we have $\hat{\tau}(\mathbf{t}) = \hat{b}_{\mathbf{t}}$, where

$$\hat{a}_{\mathbf{t}}, \hat{b}_{\mathbf{t}} = \begin{cases} \arg \min_{a,b} \sum_{i=1}^n \mathbf{1}(\mathbf{x}_i \in \mathbf{t}) (y_i - a - b d_i)^2 & \text{if } \sum_{i=1}^n \mathbf{1}(\mathbf{x}_i \in \mathbf{t}) > 0, \\ 0, 0 & \text{otherwise.} \end{cases}$$

Consider the outcome prediction model based on partition \mathcal{P} :

$$\begin{aligned} \hat{g}(\mathbf{x}, d) &= \sum_{\mathbf{t} \in \mathcal{P}} \mathbf{1}(\mathbf{x} \in \mathbf{t}) (\hat{a}_{\mathbf{t}} + \hat{b}_{\mathbf{t}} d) \\ &= \hat{A}(\mathbf{x}) + \hat{B}(\mathbf{x}) d, \end{aligned} \tag{SA-50}$$

where

$$\hat{A}(\mathbf{x}) = \sum_{\mathbf{t} \in \mathcal{P}} \mathbf{1}(\mathbf{x} \in \mathbf{t}) \hat{a}_{\mathbf{t}}, \quad \hat{B}(\mathbf{x}) = \sum_{\mathbf{t} \in \mathcal{P}} \mathbf{1}(\mathbf{x} \in \mathbf{t}) \hat{b}_{\mathbf{t}}.$$

First, we show that for L_2 -consistency of treatment effect estimation, it is enough to look at the L_2 loss for outcome prediction. Denote by $P_{X,d}$ the joint distribution of (\mathbf{x}_i, d_i) . Since we assumed \mathbf{x}_i and d_i are independent, we have $P_{X,d} = P_X \times P_d$, where P_X and P_d are the marginal distributions of X and d . Given Assumption SA-2, the target outcome prediction model is

$$g^*(\mathbf{x}_i, d_i) = \mathbb{E}[y_i | \mathbf{x}_i, d_i] = \mu + \tau d_i, \quad \mu = \mathbb{E}[y_i(0)], \quad \tau = \mathbb{E}[y_i(1) - y_i(0)].$$

Hence

$$\begin{aligned} & \mathbb{E}[\|\hat{g} - g^*\|^2] \\ &= \mathbb{E} \left[\int_{\mathcal{X} \times \{0,1\}} (\hat{g}(\mathbf{x}, d) - \mu - \tau d)^2 dP_{X,d}(\mathbf{x}, d) \right] \\ &= \mathbb{E} \left[\int_{\mathcal{X} \times \{0,1\}} (\hat{A}(\mathbf{x}) + \hat{B}(\mathbf{x})d - \mu - \tau d)^2 dP_X(\mathbf{x}) \times P_d(d) \right] \\ &= \mathbb{E} \left[\int_{\mathcal{X} \times \{0,1\}} (d (\hat{A}(\mathbf{x}) + \hat{B}(\mathbf{x}) - \mu - \tau) + (1-d) (\hat{A}(\mathbf{x}) - \mu))^2 dP_X(\mathbf{x}) \times P_d(d) \right] \\ &= \mathbb{E} \left[\int_{\mathcal{X} \times \{0,1\}} d (\hat{A}(\mathbf{x}) + \hat{B}(\mathbf{x}) - \mu - \tau)^2 + (1-d) (\hat{A}(\mathbf{x}) - \mu)^2 dP_X(\mathbf{x}) \times P_d(d) \right] \\ &= \mathbb{E} \left[\xi \int_{\mathcal{X}} (\hat{A}(\mathbf{x}) + \hat{B}(\mathbf{x}) - \mu - \tau)^2 dP_X(\mathbf{x}) + (1-\xi) \int_{\mathcal{X}} (\hat{A}(\mathbf{x}) - \mu)^2 dP_X(\mathbf{x}) \right] \\ &= \xi \mathbb{E}[\|\hat{A} + \hat{B} - \mu - \tau\|^2] + (1-\xi) \mathbb{E}[\|\hat{A} - \mu\|^2]. \end{aligned} \tag{SA-51}$$

It follows that

$$\mathbb{E}[\|\hat{\tau} - \tau\|^2] = \mathbb{E}[\|\hat{B} - \tau\|^2] \leq \frac{4}{\min\{\xi, 1-\xi\}} \mathbb{E}[\|\hat{g} - g^*\|^2].$$

Error Bound for Least Square Prediction.

Now, we bound the least square error $\mathbb{E}[\|\hat{g} - g^*\|^2]$ following the strategy for [Klusowski and Tian, 2024, Theorem 4.3]. First, assume $|y_i(t)| \leq U$, $i = 1, 2, \dots, n$, $t = 0, 1$, for some $U \geq 0$. Decompose by

$$\|\hat{g} - g^*\|^2 = E_1 + E_2,$$

where

$$\begin{aligned} E_1 &= \|\hat{g} - g^*\|^2 - 2(\|y - \hat{g}\|_{\mathcal{D}}^2 - \|y - g^*\|_{\mathcal{D}}^2) - \alpha - \beta, \\ E_2 &= 2(\|y - \hat{g}\|_{\mathcal{D}}^2 - \|y - g^*\|_{\mathcal{D}}^2) + \alpha + \beta. \end{aligned}$$

The least square representation (SA-50) implies that

$$\|y - \hat{g}\|_{\mathcal{D}}^2 \leq \min_{a \in \mathbb{R}, b \in \mathbb{R}} \sum_{i=1}^n (y_i - a - b d_i)^2 \leq \|y - \mu - \tau d\|_{\mathcal{D}}^2 = \|y - g^*\|_{\mathcal{D}}^2, \quad (\text{SA-52})$$

which implies

$$E_2 \leq \alpha + \beta.$$

We control E_1 using uniform law of large number arguments. Notice that \hat{g} is one member of the class $\mathcal{G}_n = \{A(\mathbf{x}) + d B(\mathbf{x}) : A, B \in \mathcal{H}_n\}$, where \mathcal{H}_n is the class of piecewise constant functions (bounded by U) on partitions $\mathbb{P} \in \Pi_n$. Here

$$\Pi_n = \{\mathcal{P}(\{(\mathbf{x}_1, d_1, y_1), \dots, (\mathbf{x}_n, d_n, y_n)\}) : (\mathbf{x}_i, d_i, y_i) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}\},$$

is the family of all achievable partitions \mathcal{P} by growing a depth K binary tree on n points by iteratively splitting in \mathbf{x} -space based on any criterion. By [Klusowski and Tian, 2024, Equation B.33],

$$N\left(\frac{\beta}{40U}, \mathcal{H}_n, \|\cdot\|_{P_{X^n,1}}\right) \leq (np)^{2^\kappa} \left(\frac{417eU^2}{\beta}\right)^{2^{\kappa+1}}.$$

A union bound then gives

$$N\left(\frac{\beta}{80U}, \mathcal{G}_n, \|\cdot\|_{P_{X^n,1}}\right) \leq 2(np)^{2^\kappa} \left(\frac{417eU^2}{\beta}\right)^{2^{\kappa+1}},$$

where P_{X^n} is the empirical measure based on $X^n = (X_1, \dots, X_n)$, $X_i \in \mathbb{R}^p$ for all i . Since $\hat{g} \in \mathcal{G}_n$, we can then use [Györfi et al., 2002, Theorem 11.4] to get

$$\begin{aligned} \mathbb{P}(E_1 \geq 0) &\leq \mathbb{P}(\exists g \in \mathcal{G}_n : \|\hat{g} - g^*\|^2 \geq 2(\|y - \hat{g}\|_{\mathcal{D}}^2 - \|y - g^*\|_{\mathcal{D}}^2) + \alpha + \beta) \\ &\leq 14 \sup_{X^n} N\left(\frac{\beta}{80U}, \mathcal{G}_n, \|\cdot\|_{P_{X^n,1}}\right) \exp\left(-\frac{\alpha n}{2568U^4}\right) \\ &\leq 28(np)^{2^\kappa} \left(\frac{417eU^2}{\beta}\right)^{2^{\kappa+1}} \exp\left(-\frac{\alpha n}{2568U^4}\right). \end{aligned}$$

Choosing $\alpha \propto \frac{U^4 2^K \log(np)}{n}$, and $\beta \propto \frac{U^2}{n}$, then we have

$$\mathbb{E}[\|\hat{g} - g^*\|^2] \leq C \left(\frac{U^4 2^K \log(np)}{n} + \frac{U^2}{n} \right),$$

where C is a positive universal constant.

Now we relax the condition that $|y_i(t)| \leq U$. Take $A = \{|y_i(t)| \leq U, \forall i = 1, \dots, n, t = 0, 1\}$. Then

$$\begin{aligned} \mathbb{E}[\|\hat{g} - g^*\|^2] &= \mathbb{E}[\|\hat{g} - g^*\|^2 \mathbf{1}(A)] + \mathbb{E}[\|\hat{g} - g^*\|^2 \mathbf{1}(A^c)] \\ &\leq C \left(\frac{U^4 2^K \log(np)}{n} + \frac{U^2}{n} \right) + \mathbb{E}[\|\hat{g} - g^*\|^2 \mathbf{1}(A^c)]. \end{aligned} \quad (\text{SA-53})$$

A union bound gives

$$\begin{aligned}\mathbb{P}(A^c) &\leq n\mathbb{P}(|y_i(0)| \geq U) + n\mathbb{P}(|y_i(1)| \geq U) \\ &\leq n\exp(-|U - \mu_0|) + n\exp(-|U - \mu_1|).\end{aligned}$$

Using Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[\|\hat{g} - g^*\|^2 \mathbf{1}(A^c)] &\leq \sqrt{\mathbb{E}[\|\hat{g} - g^*\|^4] \mathbb{P}(A^c)} \\ &\leq \sqrt{8n \max_{t=0,1}(\mu_t^4 + \mathbb{E}[\varepsilon_i(t)^4])} n \max_{t=0,1} \exp(-|U - \mu_t|).\end{aligned}$$

Choosing $U = \max\{\mu_0, \mu_1\} + 4\log(n)$, we have

$$\mathbb{E}[\|\hat{g} - g^*\|^2 \mathbf{1}(A^c)] \leq \frac{C}{n},$$

for some absolute constant C . Putting it back to Equation (SA-53), we get the desired conclusion.

For the high probability bound, the same analysis as Equation (SA-51) in almost sure sense gives

$$\|\hat{\tau} - \tau\|^2 \leq \mathbb{E}[\|\hat{B} - \tau\|^2] \leq \frac{4}{\min\{\xi, 1 - \xi\}} \|\hat{g} - g^*\|^2,$$

almost surely. Using sub-exponentianity of $\varepsilon_i(t)$,

$$\begin{aligned}\|g - g^*\| \mathbf{1}(A) &\leq E_1 \mathbf{1}(A) + E_2 \mathbf{1}(A) \\ &\leq C_1 \left(\frac{U^4 2^K \log(np)}{n} + \frac{U^2}{n} \right),\end{aligned}$$

with probability at least n^{-C_2} , where C_1 and C_2 are some positive absolute constants. Sub-exponentianity of $\varepsilon_i(t)$, $1 \leq i \leq n$, $t = 0, 1$, implies that $\mathbb{P}(A^c) = n^{-C_3}$ if we choose $U = C_4 \log(n)$, where C_3 and C_4 are positive constants only depending on the distribution of $(\varepsilon_i(0), \varepsilon_i(1))$. Combining with the previous two inequalities, we get the second conclusion.

SA-4.24 Proof of Theorem SA-23

Recall (\hat{i}, \hat{j}) denotes the optimal splitting index and coordinate for the decision stump. Denote $\hat{\tau}_{\text{DIM}}^{\text{HON}}(\mathbf{x})$ by $\check{\tau}(\mathbf{x})$ for simplicity. We use $(y_i, \mathbf{x}_i^\top)_{i=1}^M$ to denote $\mathcal{D}_{\text{HON},1}$, which we used to construct the causal tree. Denote by (\hat{i}, \hat{j}) the splitting index and coordinate at the K_n -th step, based on $\mathcal{D}_{\text{HON},1}$.

Use $(\tilde{y}_i, \tilde{\mathbf{x}}_i^\top)_{i=1}^N$ to denote $\mathcal{D}_{\text{HON},2}$. Then

$$\begin{aligned}\sup_{\mathbf{x} \in \mathcal{X}} |\check{\tau}(\mathbf{x}) - \tau| &\geq |\check{\tau}(\mathbf{0}) - \tau| \\ &= \left| \frac{\sum_{i=1}^N \tilde{d}_i \tilde{\varepsilon}_i(1) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}, \hat{j})})}{\sum_{i=1}^N \tilde{d}_i \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}, \hat{j})})} - \frac{\sum_{i=1}^N (1 - \tilde{d}_i) \tilde{\varepsilon}_i(0) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}, \hat{j})})}{\sum_{i=1}^N (1 - \tilde{d}_i) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}, \hat{j})})} \right|.\end{aligned}$$

Since $(\tilde{\varepsilon}_i(0), \tilde{\varepsilon}_i(1)) \perp \tilde{\mathbf{x}}_i$, condition on \hat{i}, \hat{j} and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)$, we have

$$\begin{aligned} & \frac{\sum_{i=1}^N \tilde{d}_i \tilde{\varepsilon}_i(1) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}),\hat{j}})}{\sum_{i=1}^N \tilde{d}_i \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}),\hat{j}})} - \frac{\sum_{i=1}^N (1 - \tilde{d}_i) \tilde{\varepsilon}_i(0) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}),\hat{j}})}{\sum_{i=1}^N (1 - \tilde{d}_i) \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}),\hat{j}})} \\ & \stackrel{d}{=} \frac{\sum_{i=1}^{\tilde{l}} d_i \varepsilon_i(1)}{\sum_{i=1}^{\tilde{l}} d_i} - \frac{\sum_{i=1}^{\tilde{l}} (1 - d_i) \varepsilon_i(0)}{\sum_{i=1}^{\tilde{l}} (1 - d_i)}, \end{aligned}$$

where

$$\tilde{l} = \sum_{i=1}^N \mathbf{1}(\tilde{x}_{i,\hat{j}} \leq x_{\pi_{\hat{j}}(\hat{i}),\hat{j}}).$$

Call

$$Z = \left| \frac{\sum_{i=1}^{\tilde{l}} d_i \varepsilon_i(1)}{\sum_{i=1}^{\tilde{l}} d_i} - \frac{\sum_{i=1}^{\tilde{l}} (1 - d_i) \varepsilon_i(0)}{\sum_{i=1}^{\tilde{l}} (1 - d_i)} \right|.$$

High probability lower bound on Z . Denote $n_1 = \sum_{i=1}^{\tilde{l}} d_i$, $n_0 = \sum_{i=1}^{\tilde{l}} (1 - d_i)$, $n_0 + n_1 = \tilde{l}$. And consider the weights $w_i = \frac{d_i}{n_1} - \frac{1-d_i}{n_0}$, so that $Z = |\sum_{i=1}^{\tilde{l}} w_i \varepsilon_i|$. By Marcinkiewicz–Zygmund inequality and a Jensen’s inequality on the square root function, for some absolute constant c_{MZ} ,

$$\begin{aligned} \mathbb{E}[Z \mid \mathbf{D}, \tilde{l}] &= \mathbb{E}\left[\left|\sum_i w_i \varepsilon_i\right| \mid \mathbf{D}, \tilde{l}\right] \\ &\geq c_{\text{MZ}} \mathbb{E}\left[\left(\sum_i w_i^2 \varepsilon_i^2\right)^{1/2} \mid \mathbf{D}, \tilde{l}\right] \\ &\geq c_{\text{MZ}} \mathbb{E}\left[\left(\sum_{i=1}^{\tilde{l}} \frac{d_i}{n_1^2} \varepsilon_i(1)^2 + \sum_{i=1}^{\tilde{l}} \frac{1-d_i}{n_0^2} \varepsilon_i(0)^2\right)^{1/2} \mid \mathbf{D}, \tilde{l}\right] \\ &\geq \frac{c_{\text{MZ}}}{n_0 + n_1} \sum_{i=1}^{\tilde{l}} \left(d_i \frac{\sqrt{n_0 + n_1}}{n_1} \sqrt{\mathbb{E}[\varepsilon_i(1)^2]} + (1-d_i) \frac{\sqrt{n_0 + n_1}}{n_0} \sqrt{\mathbb{E}[\varepsilon_i(0)^2]}\right) \\ &= c_{\text{MZ}} \frac{1}{\sqrt{\tilde{l}}} \left(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\right). \end{aligned}$$

Moreover, Assumption SA-2 implies that

$$\mathbb{E}[Z^2 \mid \mathbf{D}, \tilde{l}] = \frac{\sum_{i=1}^{\tilde{l}} d_i \mathbb{E}[\varepsilon_i(1)^2]}{n_1^2} + \frac{\sum_{i=1}^{\tilde{l}} (1-d_i) \mathbb{E}[\varepsilon_i(0)^2]}{n_0^2} \geq \left(\frac{1}{n_1} + \frac{1}{n_0}\right) \min\{\mathbb{V}[\varepsilon_i(1)], \mathbb{V}[\varepsilon_i(0)]\}.$$

The Paley-Zygmund inequality implies for $\theta \in (0, 1)$,

$$\begin{aligned} \mathbb{P}(Z \geq \theta \mathbb{E}[Z \mid \mathbf{D}, \tilde{l}] \mid \mathbf{D}, \tilde{l}) &\geq (1 - \theta^2) \frac{\mathbb{E}[Z \mid \mathbf{D}, \tilde{l}]^2}{\mathbb{E}[Z^2 \mid \mathbf{D}, \tilde{l}]} \\ &\geq C(1 - \theta^2) \frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}} \frac{n_0 n_1}{\tilde{l}^2}. \end{aligned} \tag{SA-54}$$

Condition on \tilde{l} , $n_0 \sim \text{Bernoulli}(\tilde{l}, \xi)$. Hence

$$\mathbb{P}(Z \geq \theta \mathbb{E}[Z \mid \mathbf{D}, \tilde{l}] \mid \tilde{l}) \geq C(1 - \theta^2) \frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}} (\xi - \xi^2) \left(1 - \frac{1}{\tilde{l}}\right) \mathbf{1}(\tilde{l} > 0).$$

We claim that whenever $s_n \leq \hat{l} \leq n - s_n$,

$$\mathbb{E}\left[\left(1 - \frac{1}{\hat{l}}\right)\mathbf{1}(\tilde{l} > 0) \mid \hat{l}\right] = 1 + o_{\mathbb{P}}(1). \quad (\text{SA-55})$$

It then follows from Equation (SA-54) that whenever $s_n \leq \hat{l} \leq M - s_n$,

$$\mathbb{P}\left(Z \geq \theta c_{\text{MZ}} \frac{1}{\sqrt{\hat{l}}} \left(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\right) \mid \hat{l}\right) \geq C(1 - \theta^2) \frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}} (\xi - \xi^2) + o_{\mathbb{P}}(1).$$

Choose $\theta = 1/2$, and take

$$\mathbf{c} = C \frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}} \frac{\xi - \xi^2}{4}.$$

Then by Theorem SA-19, we have

$$\mathbb{P}\left(Z \geq \frac{1}{2} c_{\text{MZ}} \frac{1}{\sqrt{\hat{l}}} \left(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\right), \hat{l} \leq n^b\right) \geq \mathbf{c} \frac{b}{2e} + o_{\mathbb{P}}(1).$$

We can show via the same argument as Theorem SA-5 that $\liminf_{n \rightarrow \infty} \mathbb{P}(\tilde{l} \leq n^b/2 \mid \hat{l} \leq n^b) = 1$. Hence

$$\liminf_{n \rightarrow \infty} \mathbb{P}\left(Z \geq \frac{1}{4} c_{\text{MZ}} \frac{1}{\sqrt{n^b}} \left(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\right)\right) \geq \mathbf{c} \frac{b}{2e}.$$

Proof of Equation (SA-55). Let F be the cumulative distribution function of \mathbf{x}_i . Suppose $1 \leq k \leq n/2$. Then $F(\mathbf{x}_{(k)}) \sim \text{Beta}(k, M - k + 1)$. By a Bernstein bound for Beta variables [Skorski, 2023, Theorem 1], we have for all $\epsilon > 0$,

$$\mathbb{P}(F(\mathbf{x}_{(k)}) > k/M - \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2v}\right),$$

where for large enough n ,

$$v = \frac{k(M - k + 1)}{(M + 1)^2(M + 2)} \leq 2 \frac{k}{M^2}.$$

Hence with probability at least $1 - s_n^{-1}$,

$$F(\mathbf{x}_{(k)}) \geq k/M - 2 \frac{\sqrt{\log(s_n)k}}{M}.$$

Condition on $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{1}(\tilde{\mathbf{x}}_i \leq \mathbf{x}_{(k)})$'s are i.i.d Bernoulli($F(\mathbf{x}_{(k)})$). Hence condition on \mathbf{X} and \hat{l} , with probability at least $1 - s_n^{-1}$,

$$\tilde{l}/N = N^{-1} \sum_{i=1}^N \mathbf{1}(\tilde{\mathbf{x}}_i \geq \mathbf{x}_{(\hat{l})}) \geq F(\mathbf{x}_{(\hat{l})}) - 2 \sqrt{\frac{\log(s_n)F(\mathbf{x}_{(\hat{l})})}{N}}.$$

It follows that on $s_n \leq \hat{\iota} \leq M - s_n$, using boundedness of $(1 - \frac{1}{\hat{\iota}})\mathbf{1}(\tilde{\iota} > 0)$,

$$\begin{aligned}\mathbb{E}\left[\left(1 - \frac{1}{\hat{\iota}}\right)\mathbf{1}(\tilde{\iota} > 0) \middle| \hat{\iota}\right] &= \mathbb{E}\left[\left(1 - \frac{1}{\hat{\iota}}\right)\mathbf{1}(\tilde{\iota} > 0), \mathbf{1}(\tilde{\iota} \geq \hat{\iota}/8) \middle| \hat{\iota}\right] + \mathbb{E}\left[\left(1 - \frac{1}{\hat{\iota}}\right)\mathbf{1}(\tilde{\iota} > 0), \mathbf{1}(\tilde{\iota} < \hat{\iota}/8) \middle| \hat{\iota}\right] \\ &= 1 + O\left(\frac{8}{\hat{\iota}}\right) + O(s_n^{-1}) \\ &= 1 + O(s_n^{-1}).\end{aligned}$$

SA-4.25 Proof of Theorem SA-24

For simplicity, denote $\hat{\tau}_{\text{DIM}}^{\text{HON}}$ by $\hat{\tau}$. Since given the partition \mathcal{P} chosen by \mathcal{D}_{T} , Equation (SA-50) is still satisfied. We can use the same argument in the proof of Theorem SA-22 condition on \mathcal{D}_{T} to get

$$\mathbb{E}_{\mathcal{D}_{\text{T}}}[\|\hat{\tau} - \tau\|^2 | \mathcal{D}_{\text{T}}] \leq C \frac{2^K \log(n_{\text{T}})^5}{n_{\text{T}}},$$

where C is a positive constant that only depends on ξ , μ and the distribution of $\varepsilon_i(0), \varepsilon_i(1)$. In particular, the expectation is taken with respect to \mathcal{D}_{T} with effective sample size n_{T} .

Since condition on \mathcal{D}_{T} , the partition \mathcal{P} is fixed, we can use the same argument as in Theorem SA-6 to show that \hat{g} lies in a class $\mathcal{H}_{n_{\text{T}}}[\mathcal{P}]$ with covering number,

$$N(\varepsilon U, \mathcal{H}_{n_{\text{T}}}[\mathcal{P}], \|\cdot\|_{P_{X^{n_{\text{T}}}}}) \leq \left(\frac{2}{\varepsilon}\right)^{2^K}, \quad \varepsilon \in (0, 1),$$

when we assume $y_i(0)$ and $y_i(1)$ are bounded by U . In comparison, in the proof of Theorem SA-22, we show \hat{g} lies in $\mathcal{H}_{n_{\text{T}}}$ with covering number

$$N(\varepsilon U, \mathcal{H}_{n_{\text{T}}}, \|\cdot\|_{P_{X^{n_{\text{T}}}}}) \leq 2(n_{\text{T}}p)^{2^K} \left(\frac{417eU^2}{\beta}\right)^{2^{K+1}}, \quad \varepsilon \in (0, 1).$$

This improvement of covering number due to honesty means we can replace a $\log(n_{\text{T}}p)$ -penalty in the result of Theorem SA-22 by $\log(n_{\text{T}})$. Now uncondition over \mathcal{D}_{T} and using the fact that $\rho^{-1} \leq n_{\text{T}}/n_{\text{T}} \leq \rho$, we get the conclusion.

SA-4.26 Proof of Theorem SA-25

The conclusion follows from Theorem SA-19 and the same proof for Theorem SA-7.

SA-4.27 Proof of Theorem SA-26

For simplicity, denote $\hat{\tau}_{\text{DIM}}^{\text{X}}(\mathbf{x}; K)$ by $\tilde{\tau}(T_K)$, and $N = n/(K+1)$ denotes the sample size for each folds in the \mathbf{X} sample splitting scheme.

Let $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{2^K}$ denote the 2^K leaf nodes in the decision tree, (if a node cannot be further refined, we duplicate the split indices and values at the next level). And let N_1, N_2, \dots, N_{2^K} and m_1, m_2, \dots, m_{2^K} denote the number of observations and the Lebesgue measure of the 2^K leaf nodes, respectively. Note that $\vec{N} = (N_1, \dots, N_{2^K})$ are independent of the \tilde{y}_i data by the honest condition and the x_i data per Assumption SA-1. As in the proof of Theorem SA-8, we can show condition on \vec{N} , $m_k \sim \text{Beta}(N_k, N - N_k + 1)$

Thus, the IMSE can be bounded as follows: Since condition on \vec{N} , m_k 's are independent to the refreshed samples $\tilde{d}_i, \tilde{\varepsilon}_i(0), \tilde{\varepsilon}_i(1)$'s, we have

$$\begin{aligned}
& \mathbb{E} \left[\int_{\mathcal{X}} (\tilde{\tau}(T_K)(x) - \tau)^2 \mathbb{P}_x(dx) \right] \\
&= \sum_{k=1}^{2^K} \mathbb{E} \left[m_k \left(\frac{\sum_{\mathbf{x}_i \in \mathbf{t}_k} \tilde{d}_i \tilde{\varepsilon}_i(1)}{\sum_{\mathbf{x}_i \in \mathbf{t}_k} \tilde{d}_i} - \frac{\sum_{\mathbf{x}_i \in \mathbf{t}_k} (1 - \tilde{d}_i) \tilde{\varepsilon}_i(0)}{\sum_{\mathbf{x}_i \in \mathbf{t}_k} 1 - \tilde{d}_i} \right)^2 \right] \\
&\leq \sum_{k=1}^{2^K} \mathbb{E} \left[\mathbb{E}[m_k | \vec{N}] \mathbb{E} \left[\left(\frac{\sum_{\mathbf{x}_i \in \mathbf{t}_k} \tilde{d}_i \tilde{\varepsilon}_i(1)}{\sum_{\mathbf{x}_i \in \mathbf{t}_k} \tilde{d}_i} - \frac{\sum_{\mathbf{x}_i \in \mathbf{t}_k} (1 - \tilde{d}_i) \tilde{\varepsilon}_i(0)}{\sum_{\mathbf{x}_i \in \mathbf{t}_k} 1 - \tilde{d}_i} \right)^2 \middle| \vec{N} \right] \right] \\
&\leq \sum_{k=1}^{2^K} \mathbb{E} \left[\frac{N_k}{N} \left(\frac{\mathbb{1}(\sum_{\mathbf{x}_i \in \mathbf{t}_k} d_i > 0)}{\sum_{\mathbf{x}_i \in \mathbf{t}_k} d_i} + \frac{\mathbb{1}(\sum_{\mathbf{x}_i \in \mathbf{t}_k} 1 - d_i > 0)}{\sum_{\mathbf{x}_i \in \mathbf{t}_k} 1 - d_i} \right) \right] \max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}.
\end{aligned}$$

Notice that condition on \vec{N} , $\sum_{\mathbf{x}_i \in \mathbf{t}_k} d_i \sim \text{Bin}(N_k, \xi)$ and $\sum_{\mathbf{x}_i \in \mathbf{t}_k} 1 - d_i \sim \text{Bin}(N_k, 1 - \xi)$. Using the fact that for a binomial random variable $W \sim \text{Bin}(n, p)$, we have

$$\mathbb{E} \left[\frac{1}{W} \mathbb{1}(W > 0) \right] \leq \frac{C}{npC_p},$$

where C is an absolute constant, and C_p is some constant that only depends on p . It follows that

$$\begin{aligned}
& \mathbb{E} \left[\int_{\mathcal{X}} (\tilde{\tau}(T_K)(x) - \tau)^2 \mathbb{P}_x(dx) \right] \\
&\leq \sum_{k=1}^{2^K} \mathbb{E} \left[\frac{N_k}{N} \left(\frac{1}{N_k \xi} + \frac{1}{N_k (1 - \xi)} \right) \right] \max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\} \\
&\lesssim \frac{2^K}{N}.
\end{aligned}$$

SA-4.28 Proof of Lemma SA-27

Assume w.l.o.g. $k \leq n/2$, since the case of $k > n/2$ can be dealt with by symmetry. From the proof of Lemma SA-17,

$$\sup_{r_n \leq k < n - r_n} \frac{k(n-k)}{n} \left| (\hat{\mu}_{L,0}(k, \ell) - \hat{\mu}_{R,0}(k, \ell))^2 - (\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell))^2 \right| = O_{\mathbb{P}} \left(\frac{\log \log n}{\sqrt{r_n}} \right).$$

Moreover, the proof of the term R_1 in Lemma SA-17 implies

$$\sup_{r_n \leq k < n - r_n} ((\hat{\mu}_{L,0}(k, \ell) - \hat{\mu}_{R,0}(k, \ell))^2 + (\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell))^2) = O_{\mathbb{P}} \left(\frac{\log \log n}{r_n} \right).$$

Now we consider the randomness induced by $n_0, n_{L,0}, n_{R,0}$. By Theorem A.4.1 in Csörgő and Horváth [1997],

$$\max_{r_n \leq k < n - r_n} \sqrt{k} \cdot \left| \frac{1}{k} \sum_{i=1}^k \left(\frac{d_i}{\xi} - 1 \right) \right| = O_{\mathbb{P}}(\sqrt{\log \log n}),$$

which implies

$$\sup_{r_n \leq k < n - r_n} \left| \frac{n_{L,0}(k)n_{R,0}(k)}{n_0} - (1 - \xi) \frac{k(n - k)}{n} \right| = O_{\mathbb{P}}(\sqrt{r_n \log \log n}).$$

Putting together, triangle inequality implies

$$\max_{1 \leq \ell \leq p} \max_{r_n \leq k < n - r_n} \left| \mathcal{J}^{\text{SSE}}(k, \ell) - \mathcal{J}^{\text{prox}}(k, \ell) \right| = O_{\mathbb{P}}\left(\frac{\log \log(n)^{3/2}}{r_n^{1/2}}\right).$$

SA-4.29 Proof of Lemma SA-28

The proof of Lemma SA-18 implies that

$$\max_{1 \leq \ell \leq p} \max_{1 \leq k \leq s_n} k \left| (\hat{\mu}_{L,0}(k, \ell) - \hat{\mu}_{R,0}(k, \ell))^2 - (\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell))^2 \right| = O_{\mathbb{P}}(\alpha_n),$$

where $\alpha_n = \rho_n \log \log n + \frac{s_n}{n - s_n} \log \log n$, and

$$\max_{1 \leq \ell \leq p} \max_{1 \leq k \leq s_n} k (\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell))^2 = O_{\mathbb{P}}(\rho_n \log \log n).$$

Hence it also follows that

$$\max_{1 \leq \ell \leq p} \max_{1 \leq k \leq s_n} k (\hat{\mu}_{L,0}(k, \ell) - \hat{\mu}_{R,0}(k, \ell))^2 = O_{\mathbb{P}}\left(\frac{s_n}{n - s_n} \log \log n + \alpha_n\right) = O_{\mathbb{P}}(\alpha_n).$$

When $1 \leq k \leq s_n$, we have $\frac{n_{L,0}(k)n_{R,0}(k)}{n_0} \leq n_{L,0}(k) \leq k$. The conclusion then follows.

SA-4.30 Proof of Theorem SA-29

The proof is similar to the proof of Theorem SA-19, except that in Theorem SA-19, we approximate the split criterion by a time-transformed O-U process, while here we approximate the split criterion by the summation of *two independent* time transformed O-U processes. We divide the proofs into two steps.

Step 1: Approximation of fit-based processes by ipw-based processes

Let $0 < a < b < 1$. Let ρ_n be a sequence of real numbers taking values in $(0, 1)$ to be determined, and take $s_n = \exp((\log n)^{\rho_n})$. Then for large enough n , we have $s_n \leq n^a \leq n^b \leq n - s_n$. Consider the event

$$A_n = \{\exists \ell \in [p] : \max_{k \in [n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{k \notin [s_n, n - s_n]} \mathcal{J}^{\text{SSE}}(k, \ell)\}.$$

Equation (A.4.18) and (A.4.20) imply that for each $\ell \in [p]$,

$$\begin{aligned} \max_{1 \leq k \leq s_n, n - s_n \leq k \leq n} \mathcal{J}^{\text{prox}}(k, \ell) &= O_{\mathbb{P}}(\rho_n \log \log(n)), \\ \max_{s_n \leq k \leq n - s_n} \mathcal{J}^{\text{prox}}(k, \ell) &= 2 \log \log(n)(1 + o_{\mathbb{P}}(1)). \end{aligned}$$

Hence

$$\max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \mathcal{J}^{\text{prox}}(k, \ell) = o_{\mathbb{P}} \left(\max_{s_n \leq k \leq n-s_n} \mathcal{J}^{\text{prox}}(k, \ell) \right), \quad \ell \in [p],$$

Approximations results from Lemma SA-27 (taking $r_n = s_n$) and Lemma SA-28, using the same argument as *step 1* in the proof of Theorem SA-19, with $\log \log \log \log(n) / \log \log(n) \ll \rho_n \ll 1$, then implies

$$\max_{1 \leq k \leq s_n, n-s_n \leq k \leq n} \mathcal{J}^{\text{SSE}}(k, \ell) = o_{\mathbb{P}} \left(\max_{s_n \leq k \leq n-s_n} \mathcal{J}^{\text{SSE}}(k, \ell) \right), \quad \ell \in [p].$$

Using a union bound, we get $\mathbb{P}(A_n) \rightarrow 0$ as $n \rightarrow \infty$. Observe that on the event A_n^c , the argmax for \mathcal{J}^{SSE} should be inside $[s_n, n-s_n]$. Hence

$$\begin{aligned} & \mathbb{P} \left(\exists \ell \in [p] : \max_k \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}^{\text{SSE}}(k, j), \max_k \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{k \notin [n^a, n^b]} \mathcal{J}^{\text{SSE}}(k, \ell) \right) \\ & \geq \mathbb{P} \left(\exists \ell \in [p] : \max_k \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{k, j \neq \ell} \mathcal{J}^{\text{SSE}}(k, j), \max_k \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{k \notin [n^a, n^b]} \mathcal{J}^{\text{SSE}}(k, \ell) \text{ and } A_n^c \right) - \mathbb{P}(A_n) \\ & \geq \mathbb{P} \left(\exists \ell \in [p] : \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{\substack{j \neq \ell \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j), \right. \\ & \quad \left. \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, \ell) \right) - 2\mathbb{P}(A_n). \end{aligned}$$

Now we focus on the first term. By symmetry in the p coordinates,

$$\begin{aligned} & \mathbb{P} \left(\exists \ell \in [p] : \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{\substack{j \neq \ell \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j), \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, \ell) \right) \\ & = p \mathbb{P} \left(\max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, 1) > \max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j), \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, 1) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, 1) \right) \\ & \geq p \sup_{z \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j) < z, \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, 1) > z > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, 1) \right) \\ & \geq p \sup_{z \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j) < z, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, 1) < z \right) \\ & \quad - p \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j) < z, \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, 1) < z \right). \end{aligned}$$

Then using the fact that $\mathcal{J}^{\text{prox}}(k, \ell)$ approximates $\mathcal{J}^{\text{SSE}}(k, \ell)$ from Lemma SA-17, we have

$$\begin{aligned} & \mathbb{P} \left(\exists \ell \in [p] : \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{\substack{j \neq \ell \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, j), \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{SSE}}(k, \ell) > \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{SSE}}(k, \ell) \right) \\ & \geq p \sup_{z \in \mathbb{R}} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{prox}}(k, j) < z - v_n, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{prox}}(k, 1) < z - v_n \right) \\ & \quad - p \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{prox}}(k, j) < z + v_n, \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{prox}}(k, 1) < z + v_n \right), \end{aligned} \tag{SA-56}$$

where $v_n = O_{\mathbb{P}}(\log \log(n) s_n^{-1/2})$.

Step 2: Gaussian approximation of IPW partial sums

Recall that

$$\mathcal{J}^{\text{prox}}(k, \ell) = (1 - \xi) \frac{k(n-k)}{n} (\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell))^2 + \xi \frac{k(n-k)}{n} (\bar{\mu}_{L,1}(k, \ell) - \bar{\mu}_{R,1}(k, \ell))^2. \quad (\text{SA-57})$$

and we will show that high dimensional random vector Ξ from concatenating $(\sqrt{(1-\xi)\frac{k(n-k)}{n}}(\bar{\mu}_{L,0}(k, \ell) - \bar{\mu}_{R,0}(k, \ell)) : k \in [n], \ell \in [p])$ and $(\sqrt{(1-\xi)\frac{k(n-k)}{n}}(\bar{\mu}_{L,1}(k, \ell) - \bar{\mu}_{R,1}(k, \ell)) : k \in [n], \ell \in [p])$ can be approximated by a Gaussian random vector with the same covariance structure. The proof will still be based on writing Ξ as $\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{C}_i$ where

$$\begin{aligned} \mathbf{C}_i = & \left(\sqrt{n} \left(\left(\sqrt{\frac{n}{k(n-k)}} (\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) : r_n \leq k \leq n - r_n \right)^\top : 1 \leq \ell \leq p \right)^\top \frac{1 - d_i}{1 - \xi} \varepsilon_i(0), \right. \\ & \left. \sqrt{n} \left(\left(\sqrt{\frac{n}{k(n-k)}} (\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) : r_n \leq k \leq n - r_n \right)^\top : 1 \leq \ell \leq p \right)^\top \frac{d_i}{\xi} \varepsilon_i(1) \right)^\top, \end{aligned}$$

where $\#\pi^\ell$ denotes the inverse mapping of π^ℓ , as in the proof of Theorem SA-1.

Notice that the random vectors are $2np$ dimensional. For notational simplicity, in what follows, denote by $\mathbf{e}_{t,k,\ell}$ the indicator of the position corresponding to $\sqrt{(1-\xi)\frac{k(n-k)}{n}}(\bar{\mu}_{L,t}(k, \ell) - \bar{\mu}_{R,t}(k, \ell))$, $t = 0, 1$, $k \in [n]$, $\ell \in [p]$.

However, the format of Equation (SA-57) induces a different geometry when approximating probabilities in Equation (SA-56). Instead of high dimensional CLT for hyper-rectangles, we consider the class of simple convex sets [Chernozhukov et al., 2017, Section 3.1].

Let \mathcal{J} be a subset of $[n] \times [p]$. Consider the class of closed convex sets \mathcal{A} containing sets of the form

$$A = \{\mathbf{u} \in \mathbb{R}^{2np} : (\mathbf{e}_{0,k,\ell}^\top \mathbf{u}, \mathbf{e}_{1,k,\ell}^\top \mathbf{u}) \in B_2(s_{k,\ell}), s_{k,\ell} \in (0, n], (k, \ell) \in \mathcal{J}\}, \quad (\text{SA-58})$$

where $B_2(r)$ denotes the Euclidean ball centered at $\mathbf{0}$ with radius r in \mathbb{R}^2 . That is, the class \mathcal{A} contains intersections of cylinders $\{\mathbf{u} \in \mathbb{R}^{2np} : \|(\mathbf{e}_{j_1}^\top \mathbf{u}, \mathbf{e}_{j_2}^\top \mathbf{u})\|_2 \leq s\}$. Notice that for $z \in (0, n]$, the event in Equation (SA-57) (inside $\sup z$) can be characterized as the high dimensional vector Ξ lies in a set in \mathcal{A} .

For each $A \in \mathcal{A}$, we consider its approximation by simple convex sets. For each $B_2(r)$, denote by $B_2^{\text{in},n}(r)$ and $B_2^{\text{out},n}(r)$ its inscribed and circumscribed regular n^2 -gon. Take $m = n^2|J|$. Then for each $A \in \mathcal{A}$ of the form (SA-58), take

$$A^m = \{\mathbf{u} \in \mathbb{R}^{2np} : (\mathbf{e}_{0,k,\ell}^\top \mathbf{u}, \mathbf{e}_{1,k,\ell}^\top \mathbf{u}) \in B_2^{\text{in},n}(s_{k,\ell}), s_{k,\ell} \in (0, n], (k, \ell) \in \mathcal{J}\},$$

and

$$A^{m,\epsilon} = \{\mathbf{u} \in \mathbb{R}^{2np} : (\mathbf{e}_{0,k,\ell}^\top \mathbf{u}, \mathbf{e}_{1,k,\ell}^\top \mathbf{u}) \in B_2^{\text{out},n}(s_{k,\ell}), s_{k,\ell} \in (0, n], (k, \ell) \in \mathcal{J}\}.$$

Then $A^m \subseteq A \subseteq A^{m,\epsilon}$. Moreover, denote by $\mathcal{V}(A^m)$ the set consisting of m unit vectors that are outward normal to the facets of A^m . Then A^m can be alternatively characterized by

$$A^m = \cup_{\mathbf{v} \in \mathcal{V}(A^m)} \{\mathbf{w} \in \mathbb{R}^{2np} : \mathbf{w}^\top \mathbf{v} \leq S_A(\mathbf{v})\}, \quad S_A(\mathbf{v}) = \sup\{\mathbf{w}^\top \mathbf{v} : \mathbf{w} \in A\}.$$

Then we can analogously characterize $A^{m,\epsilon}$ by

$$A^{m,\epsilon} = \cup_{\mathbf{v} \in \mathcal{V}(A^m)} \{\mathbf{w} \in \mathbb{R}^{2np} : \mathbf{w}^\top \mathbf{v} \leq S_A(\mathbf{v}) + \epsilon_{\mathbf{v}}\}, \quad S_A(\mathbf{v}) = \sup\{\mathbf{w}^\top \mathbf{v} : \mathbf{w} \in A\},$$

where $\epsilon_{\mathbf{v}} \leq n^{-1}$ for large enough n . This shows our class \mathcal{A} is a subclass of $\mathcal{A}^{\text{si}}(1, 3)$ (see [Chernozhukov et al., 2017, Section 3.1]). Now we check its conditions (M.1'), (M.2') and (E.1'). Let $\mathbf{v} \in \mathcal{V}(A^m)$. The definition of A^m implies $\mathbf{v} = v_{0,k,\ell} \mathbf{e}_{0,k,\ell} + v_{1,k,\ell} \mathbf{e}_{1,k,\ell}$ for some $(k, \ell) \in \mathcal{J}$, and $v_{0,k,\ell}^2 + v_{1,k,\ell}^2 = 1$. Let $\mathbf{v} \in \mathcal{V}(A^m)$.

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \mathbb{E}[|\mathbf{v}^\top \mathbf{C}_i|^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left(v_{0,k,\ell} \frac{n}{\sqrt{k(n-k)}} (\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) \frac{1-d_i}{1-\xi} \varepsilon_i(0) \right. \right. \\ & \quad \left. \left. + v_{1,k,\ell} \frac{n}{\sqrt{k(n-k)}} (\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) \frac{d_i}{\xi} \varepsilon_i(1) \right)^2 \right] \\ &= \frac{1}{n} \left(\frac{n}{\sqrt{k(n-k)}} \right)^2 \sum_{i=1}^n \left\{ v_{0,k,\ell}^2 \mathbb{E} \left[\left((\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) \frac{1-d_i}{1-\xi} \varepsilon_i(0) \right)^2 \right] \right. \\ & \quad \left. + v_{1,k,\ell}^2 \mathbb{E} \left[\left((\mathbf{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n}) \frac{d_i}{\xi} \varepsilon_i(1) \right)^2 \right] \right\} \\ &\geq \min\{\mathbb{V}[(1-\xi)^{-1}(1-d_i)\varepsilon_i(0)], \mathbb{V}[\xi^{-1}d_i\varepsilon_i(1)]\}, \end{aligned}$$

which verifies (M.1'). The fact that only two entries of \mathbf{v} are nonzero and $v_{0,k,\ell}^2 + v_{1,k,\ell}^2 = 1$ implies that

$$n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{v}^\top \mathbf{C}_i|^3] \leq 4n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{0,k,\ell}^\top \mathbf{C}_i|^3] + 4n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{1,k,\ell}^\top \mathbf{C}_i|^3] \lesssim \sqrt{n/r_n},$$

where the last inequality is from the calculation in Equation (SA-21), and this verifies (M.2') for the third moment. Moreover,

$$\begin{aligned} n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{v}^\top \mathbf{C}_i|^4] &\leq 8n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{0,k,\ell}^\top \mathbf{C}_i|^4] + 8n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{1,k,\ell}^\top \mathbf{C}_i|^4] \\ &\leq \sqrt{n/r_n} 8n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{0,k,\ell}^\top \mathbf{C}_i|^3] + \sqrt{n/r_n} 8n^{-1} \sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{1,k,\ell}^\top \mathbf{C}_i|^3] \\ &\lesssim n/r_n. \end{aligned}$$

The same logic shows that $\mathbb{E}[\exp(|\mathbf{v}^\top \mathbf{C}_i|/(K\sqrt{n/r_n}))] \leq 2$, where K is an absolute constant. Putting together, we verify conditions (M.2') and (E.1') with $B_n = \sqrt{n/r_n}$. Hence by [Chernozhukov et al., 2017, Proposition 3.1], there exists mean-zero random vectors $\mathbf{D}_i \sim N(\mathbf{0}, \mathbb{E}[\mathbf{C}_i \mathbf{C}_i^\top])$ such that

$$\sup_{A \in \mathcal{A}} |\mathbb{P}(n^{-1/2} \sum_{i=1}^n \mathbf{C}_i \in A) - \mathbb{P}(n^{-1/2} \sum_{i=1}^n \mathbf{D}_i \in A)| \lesssim \left(\frac{\log^7(n)}{r_n} \right)^{1/6}. \quad (\text{SA-59})$$

Step 4: Gaussian-to-Gaussian Approximation

Observe that for any $k_1, k_2 \in [n], \ell_1, \ell_2 \in [p]$, we have $\text{Cov}[\mathbf{e}_{0,k_1,\ell_1}^\top \mathbf{C}_i, \mathbf{e}_{1,k_2,\ell_2}^\top \mathbf{C}_i] = 0$. The same calculation

as *Multivariate Case Step 2* for the proof of Theorem SA-1 implies we can replace \mathbf{D}_i by another mean-zero Gaussian random vector \mathbf{Z}_i such that

$$\text{Cov}[\mathbf{e}_{t_1, k_1, \ell_1}^\top \mathbf{Z}_i, \mathbf{e}_{t_2, k_2, \ell_2}^\top \mathbf{Z}_i] = \begin{cases} \text{Cov}[\mathbf{e}_{t_1, k_1, \ell_1}^\top \mathbf{D}_i, \mathbf{e}_{t_2, k_2, \ell_2}^\top \mathbf{D}_i], & \text{if } \ell_1 = \ell_2, \\ 0, & \text{otherwise.} \end{cases}$$

We want to show $\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i$ is close to $\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{D}_i$, measured by the probability of taking value in sets from \mathcal{A} defined at Equation (SA-58). We omit details for simplicity, but illustrate the main skeleton here. As in Step 2, Nazarov inequality implies we only need to work on the m -generated convex approximation with ϵ precision $A^m = A^m(A)$ for $A \in \mathcal{A}$, for a reason given in [Chernozhukov et al., 2017, proof of Proposition 3.1]. Moreover, $\mathbb{P}(\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i \in A^m) = \mathbb{P}((\mathbf{v}^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i))_{\mathbf{v} \in \mathcal{V}(A^m)} \leq \mathbf{t})$ for some $\mathbf{t} \in \mathbb{R}^m$. Hence we only need to show

$$\sup_{\mathbf{t} \in \mathbb{R}^m} \left| \mathbb{P}((\mathbf{v}^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i))_{\mathbf{v} \in \mathcal{V}(A^m)} \leq \mathbf{t}) - \mathbb{P}((\mathbf{v}^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{D}_i))_{\mathbf{v} \in \mathcal{V}(A^m)} \leq \mathbf{t}) \right| = o(1).$$

But the definition of \mathcal{A} in Equation (SA-58) implies for any $A \in \mathcal{A}$, $\mathbf{v} \in \mathcal{V}(A^m)$, there exists $\mathbf{e}_k, \mathbf{e}_j$ and $v_k^2 + v_j^2 = 1$ such that $\mathbf{v} = v_k \mathbf{e}_k + v_j \mathbf{e}_j$, with

$$\text{Cov} \left[\mathbf{e}_k^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i), \mathbf{e}_j^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i) \right] = \text{Cov} \left[\mathbf{e}_k^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{D}_i), \mathbf{e}_j^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{D}_i) \right] = 0,$$

and hence

$$\min_{\mathbf{z} \in \mathcal{V}(A^m)} \mathbb{V} \left[\mathbf{v}^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i)_{\mathbf{v} \in \mathcal{V}(A^m)} \right] \gtrsim 1.$$

Together with Equation (SA-24), we know

$$\max_{\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}(A^m)} \left| \text{Cov} \left[\mathbf{v}_1^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i), \mathbf{v}_2^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i) \right] - \text{Cov} \left[\mathbf{v}_1^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{D}_i), \mathbf{v}_2^\top (\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{D}_i) \right] \right| = O(r_n^{-1/2}).$$

The Gaussian-to-Gaussian Comparison result [Chernozhukov et al., 2022, Proposition 2.1] then implies

$$\sup_{A \in \mathcal{A}} |\mathbb{P}(n^{-1/2} \sum_{i=1}^n \mathbf{Z}_i \in A) - \mathbb{P}(n^{-1/2} \sum_{i=1}^n \mathbf{D}_i \in A)| = O(\log(n) r_n^{-1/2}). \quad (\text{SA-60})$$

Step 5: Orstein-Uhlenbeck Process Calculations

Now we revisit Equation (SA-56). Consider

$$\mathcal{J}^{\text{Gauss}}(k, \ell) = (1 - \xi) \frac{k(n-k)}{n} (\tilde{\mu}_{L,0}(k, \ell) - \tilde{\mu}_{R,0}(k, \ell))^2 + \xi \frac{k(n-k)}{n} (\tilde{\mu}_{L,1}(k, \ell) - \tilde{\mu}_{R,1}(k, \ell))^2,$$

with

$$\begin{aligned}\tilde{\mu}_{L,0}(k, \ell) &= \frac{1}{k} \sum_{i \leq k} u_{\pi_\ell(i)}, & \tilde{\mu}_{L,1}(k, \ell) &= \frac{1}{k} \sum_{i \leq k} v_{\pi_\ell(i)}, \\ \tilde{\mu}_{R,0}(k, \ell) &= \frac{1}{n-k} \sum_{i > k} u_{\pi_\ell(i)}, & \tilde{\mu}_{R,1}(k, \ell) &= \frac{1}{n-k} \sum_{i > k} v_{\pi_\ell(i)}.\end{aligned}$$

Equations (SA-59) and (SA-60) imply that

$$\begin{aligned}& \sup_{z \in [-n, n]} \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{prox}}(k, j) < z - v_n, \max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{prox}}(k, 1) < z - v_n \right) \\ & - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{prox}}(k, j) < z + v_n, \max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{prox}}(k, 1) < z + v_n \right) \\ & = \sup_{z \in [-n, n]} \mathbb{P} \left(\max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{Gauss}}(k, 1) < z - v_n \right)^{p-1} \mathbb{P} \left(\max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{Gauss}}(k, 1) < z - v_n \right) \\ & - \mathbb{P} \left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{Gauss}}(k, j) < z + v_n \right)^{p-1} \mathbb{P} \left(\max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{Gauss}}(k, 1) < z + v_n \right) + o(1).\end{aligned}$$

The same argument as [Csörgő and Horváth, 1997, (A.4.25) to (A.4.37)] shows that there exists two independent standard Brownian bridges over $[0, 1]$, $B_{n,L}$ and $B_{n,R}$, for each n , such that

$$\begin{aligned}& \left| \max_{k \in [s_n, n-s_n]} \sqrt{\mathcal{J}^{\text{Gauss}}(k, 1)} - \sup_{t \in [s_n/n, 1-s_n/n]} \sigma \sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}} \right| = \epsilon_n, \\ & \left| \max_{k \in [s_n, n-s_n] \setminus [n^a, n^b]} \sqrt{\mathcal{J}^{\text{Gauss}}(k, 1)} - \sup_{t \in [s_n/n, 1-s_n/n] \setminus [n^{1-a}, n^{1-b}]} \sigma \sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}} \right| = \epsilon_n,\end{aligned}$$

with $\sigma^2 = \mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$ and $\epsilon_n = o_p((\log \log n)^{-1/2})$. Let $\{U_L(t) : t \in \mathbb{R}\}$ and $\{U_R(t) : t \in \mathbb{R}\}$ be two independent O-U processes with $\mathbb{E}[U_j(t)] = 0$ and $\mathbb{E}[U_j(s)U_j(t)] = e^{-|s-t|}$, $j = L, R$. Then

$$\left\{ \left(\frac{B_{n,L}}{\sqrt{t(1-t)}}, \frac{B_{n,R}}{\sqrt{t(1-t)}} \right) : t \in [0, 1] \right\} \stackrel{d}{=} \{(U_L(\log(t/(1-t))), U_R(\log(t/(1-t)))) : t \in [0, 1]\}.$$

Take $N(t) = \|(U_L(t), U_R(t))\|_2$, $t \in \mathbb{R}$. Then a time change and stationarity of O-U process implies

$$\begin{aligned}& \mathbb{P} \left(\sup_{t \in [1/n, 1-1/n] \setminus [n^{1-a}, n^{1-b}]} \sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}} \leq y \right) \\ & = \mathbb{P} \left(\sup_{-\log(n-1) \leq t < \log(n^{a-1}/(1-n^{a-1})), \log(n^{b-1}/(1-n^{b-1})) < t \leq \log(n-1)} |N(t)| \leq y \right) \\ & = \mathbb{P} \left(\sup_{0 \leq t < \log(n^{a-1}(n-1)/(1-n^{a-1})), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}} < t \leq 2 \log(n-1)} |N(t)| \leq y \right),\end{aligned}$$

and

$$\mathbb{P}\left(\sup_{t \in [1/n, 1-1/n] \setminus [n^{1-a}, n^{1-b}]} \sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}} \leq y\right) = \mathbb{P}\left(\sup_{0 \leq t < 2 \log(n-1)} |N(t)| \leq y\right).$$

An expansion based on [Horváth, 1993, Lemma 2.1] (Lemma TODO) gives for any $z \in \mathbb{R}$,

$$\mathbb{P}\left(\sup_{0 \leq t < c \log(n)} |N(t)| \leq \frac{z + 2 \log \log(n) + \log \log \log(n)}{\sqrt{2 \log \log(n)}} + \epsilon_n\right) = \exp(-e^{-z+\log(c)}) + o(1).$$

Moreover, Gaussian correlation inequality [Latała and Matlak, 2017, Remark 3 (i)] and stationarity of O-U process implies

$$\begin{aligned} & \mathbb{P}\left(\sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}}), \log \frac{n^{b-1}(n-1)}{1-n^{b-1}}) < t \leq 2 \log(n-1)} |N(t)| < \frac{z + 2 \log \log(n) + \log \log \log(n)}{\sqrt{2 \log \log(n)}} + \epsilon_n\right) \\ & \geq \mathbb{P}\left(\sup_{0 \leq t < \log(\frac{n^{a-1}(n-1)}{1-n^{a-1}})} |N(t)| < \frac{z + 2 \log \log(n) + \log \log \log(n)}{\sqrt{2 \log \log(n)}} + \epsilon_n\right) \\ & \quad \cdot \mathbb{P}\left(\sup_{0 < t \leq \log(n^{1-b}(n-1)(1-n^{b-1}))} |N(t)| < \frac{z + 2 \log \log(n) + \log \log \log(n)}{\sqrt{2 \log \log(n)}} + \epsilon_n\right) \\ & = \exp(-2e^{-z+\log(2-(b-a))}) + o(1). \end{aligned}$$

Putting together and choosing z^* that maximizes $z \mapsto \exp(-2e^{-z+\log(2-(b-a))}) - \exp(-2e^{-z+\log(c)})$, we can get

$$\begin{aligned} & \sup_{z \in [-n, n]} \mathbb{P}\left(\max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{Gauss}}(k, 1) < z - v_n\right)^{p-1} \mathbb{P}\left(\max_{\substack{k \notin [n^a, n^b] \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{Gauss}}(k, 1) < z - v_n\right) \\ & \quad - \mathbb{P}\left(\max_{\substack{j \neq 1 \\ k \in [s_n, n-s_n]}} \mathcal{J}^{\text{Gauss}}(k, j) < z + v_n\right)^{p-1} \mathbb{P}\left(\max_{k \in [s_n, n-s_n]} \mathcal{J}^{\text{Gauss}}(k, 1) < z + v_n\right) \\ & \geq \sup_z \exp\left(-2(p-1)e^{-(z-\log(2))}\right) \left(\exp\left(-2e^{-(z-\log(2-(b-a)))}\right) - \exp\left(-2e^{-(z-\log(2))}\right)\right) \\ & = \frac{b-a}{2p} \left(1 - \frac{b-a}{2p}\right)^{\frac{2p}{b-a}-1} \\ & \geq \frac{b-a}{2pe}. \end{aligned}$$

Symmetry then implies for any $0 < a < b < 1$ and $\ell \in [p]$, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}(n^a \leq \hat{i}_{\text{SSE}} \leq n^b, \hat{j}_{\text{SSE}} = \ell) = \liminf_{n \rightarrow \infty} \mathbb{P}(n - n^b \leq \hat{i}_{\text{SSE}} \leq n - n^a, \hat{j}_{\text{SSE}} = \ell) \geq \frac{b-a}{2pe}.$$

SA-4.31 Proof of Corollary SA-30

Notice that although the splitting criteria is different from the regression tree, once cells are given the estimator given by the fit-based tree is exactly the same as the regression tree (see Section SA-3.2). Hence result can be proved based on Theorem SA-29 and the same logic as Theorem SA-20.

SA-4.32 Proof of Corollary SA-31

Notice that although the splitting criteria is different from the regression tree, once cells are given the estimator given by the fit-based tree is exactly the same as the regression tree (see Section SA-3.2). Hence result can be proved based on Theorem SA-29 and the same logic as Theorem SA-21.

SA-4.33 Proof of Corollary SA-32

Since the tree is constructed by minimizing the objective Equation (SA-10) iteratively. The empirical risk minimization property Equation (SA-52) still holds. Hence the result follows from the same argument as the proof of Theorem SA-22.

SA-4.34 Proof of Corollary SA-33

Notice that although the splitting criteria is different from the regression tree, once cells are given the estimator given by the fit-based tree is exactly the same as the regression tree (see Section SA-3.2). Hence result can be proved based on Theorem SA-29 and the same logic as Theorem SA-23.

SA-4.35 Proof of Corollary SA-34

Since the tree is constructed by minimizing the objective Equation (SA-10) iteratively. The empirical risk minimization property Equation (SA-52) still holds. Hence the result follows from the same argument as the proof of Theorem SA-24.

SA-4.36 Proof of Corollary SA-35

Notice that although the splitting criteria is different from the regression tree, once cells are given the estimator given by the fit-based tree is exactly the same as the regression tree (see Section SA-3.2). Hence result can be proved based on Theorem SA-29 and the same logic as Theorem SA-25.

SA-4.37 Proof of Corollary SA-36

The result follows from the same argument as Theorem SA-26.

SA-4.38 Proof of Lemma SA-37

First, we consider **X** under Assumption SA-2. Since (y_i, d_i) 's are from dataset \mathcal{D}_τ independent to the dataset \mathcal{D}_{T_1} to \mathcal{D}_{T_K} for tree construction, it is easy to check that

$$\mathbb{E}[\hat{\tau}_l^{\mathcal{X}}(\mathbf{x}; K)] = \mathbb{E}[\mathbb{E}[\hat{\tau}_l^{\mathcal{X}}(\mathbf{x}; K) | \mathbf{T}, (\mathbf{x}_i)_{i \in \mathcal{D}_\tau}]] = \tau, \quad l \in \{\text{DIM}, \text{IPW}, \text{SSE}\}.$$

Next, we consider **HON** under Assumption SA-2. Denote by $\mathbf{t}(\mathbf{x})$ the node that contains \mathbf{x} , and denote by $n(\mathbf{t})$ the *local sample size* in \mathcal{D}_τ , where $n(\mathbf{t}) = \sum_{i \in \mathcal{D}_\tau} \mathbf{1}(\mathbf{x}_i \in \mathbf{t})$. Then

$$\begin{aligned} \mathbb{E}[\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}; K) | \mathcal{D}_\tau] &= \mathbb{E}[\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}; K) \mathbf{1}(n(\mathbf{t}(\mathbf{x})) > 0) | \mathcal{D}_\tau] + 0 \cdot \mathbb{P}(n(\mathbf{t}(\mathbf{x})) = 0 | \mathcal{D}_\tau) \\ &= \mathbb{E}[\hat{\tau}_{\text{IPW}}^{\text{HON}}(\mathbf{x}; K) | \mathcal{D}_\tau, n(\mathbf{t}(\mathbf{x})) > 0] \mathbb{P}(n(\mathbf{t}(\mathbf{x})) > 0 | \mathcal{D}_\tau) \\ &= \tau \mathbb{P}(n(\mathbf{t}(\mathbf{x})) > 0 | \mathcal{D}_\tau), \end{aligned}$$

where in the third line, we have used the fact that $\varepsilon_i(0)$ and $\varepsilon_i(1)$ in \mathcal{D}_τ are independent to \mathbf{x}_i 's in \mathcal{D}_τ and the whole dataset \mathcal{D}_T , with $\mathbb{E}[\varepsilon_i(0)] = \mathbb{E}[\varepsilon_i(1)] = 0$. Unconditioning over \mathcal{D}_T , then we get

$$\mathbb{E}[\hat{\tau}_{IPW}^{\text{HON}}(\mathbf{x}; K)] = \tau \mathbb{P}(n(\mathbf{t}(\mathbf{x})) > 0).$$

The results for DIM and SSE can be obtained by similar arguments.

Finally, we consider NSS under Assumption SA-2 and the additional symmetric error $\varepsilon_i(0)$, and $\varepsilon_i(1)$ assumption. We will use an induction assumption.

Base case: $K = 1$. Due to the assumption that μ_0 and μ_1 are constant, we can rewrite the splitting criteria from Definition 2 in the main paper as

$$\begin{aligned} \text{DIM : } \quad & \frac{n(\mathbf{t}_L)n(\mathbf{t}_R)}{n(\mathbf{t})} \left(\frac{1}{n_1(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} d_i \varepsilon_i(1) - \frac{1}{n_0(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} (1 - d_i) \varepsilon_i(0) \right. \\ & \left. - \frac{1}{n_1(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} d_i \varepsilon_i(1) + \frac{1}{n_0(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} (1 - d_i) \varepsilon_i(0) \right)^2, \end{aligned} \quad (\text{SA-61})$$

and

$$\begin{aligned} \text{IPW : } \quad & \frac{n(\mathbf{t}_L)n(\mathbf{t}_R)}{n(\mathbf{t}_L)} \left(\frac{1}{n(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} \left(\frac{d_i}{\xi} \varepsilon_i(1) - \frac{1 - d_i}{1 - \xi} \varepsilon_i(0) \right) - \frac{1}{n(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} \left(\frac{d_i}{\xi} \varepsilon_i(1) - \frac{1 - d_i}{1 - \xi} \varepsilon_i(0) \right) \right)^2, \end{aligned} \quad (\text{SA-62})$$

and

$$\begin{aligned} \text{SSE : } \quad & \frac{n_1(\mathbf{t}_L)n_1(\mathbf{t}_R)}{n_1(\mathbf{t})} \left(\frac{1}{n_1(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} d_i \varepsilon_i(1) - \frac{1}{n_1(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} d_i \varepsilon_i(1) \right)^2 \\ & + \frac{n_0(\mathbf{t}_L)n_0(\mathbf{t}_R)}{n_0(\mathbf{t})} \left(\frac{1}{n_0(\mathbf{t}_L)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_L} (1 - d_i) \varepsilon_i(0) - \frac{1}{n_0(\mathbf{t}_R)} \sum_{i:\mathbf{x}_i \in \mathbf{t}_R} (1 - d_i) \varepsilon_i(0) \right)^2. \end{aligned} \quad (\text{SA-63})$$

Denote the vector $\boldsymbol{\varepsilon} = (\varepsilon_1(0), \varepsilon_1(1), \dots, \varepsilon_n(0), \varepsilon_n(1))$. Notice that for all three criteria, for any $\mathbf{d} = (d_1, \dots, d_n)$ and $\mathbf{t}_L, \mathbf{t}_R$, $\boldsymbol{\varepsilon} = \mathbf{u}$ and $\boldsymbol{\varepsilon} = -\mathbf{u}$ give the same value. Hence condition on \mathbf{d} and the data-driven split region $\hat{\mathbf{t}}_L$ and $\hat{\mathbf{t}}_R$, $\boldsymbol{\varepsilon}$ is symmetrically distributed around zero. It then follows from the form of the three estimators that all of them are unbiased.

Induction step: $K \geq 2$. Each leaf node \mathbf{t} in layer $K - 1$ is further partitioned into \mathbf{t}_L and \mathbf{t}_R such that Equations (SA-61), (SA-62) and (SA-63) are maximized. The induction hypothesis is that condition on all leaf \mathbf{t} in the $K - 1$ th layer and $\mathbf{d}, \boldsymbol{\varepsilon}$ is symmetrically distributed around zero. Again for all three criteria, given $K - 1$ th leaf node \mathbf{t} , for any $\mathbf{d} = (d_1, \dots, d_n)$ and $\mathbf{t}_L, \mathbf{t}_R$, $\boldsymbol{\varepsilon} = \mathbf{u}$ and $\boldsymbol{\varepsilon} = -\mathbf{u}$ give the same value. Hence the resulting K th level $\hat{\mathbf{t}}_L$ and $\hat{\mathbf{t}}_R$ are such that condition on \mathbf{d} and the data-driven split region $\hat{\mathbf{t}}_L$ and $\hat{\mathbf{t}}_R$, $\boldsymbol{\varepsilon}$ is symmetrically distributed around zero, making the estimators unbiased.

References

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

Matias D Cattaneo, Jason M Klusowski, and Peter M Tian. On the pointwise behavior of recursive parti-

- tioning and its implications for heterogeneous causal effect estimation. *Technical report, arXiv preprint arXiv:2211.10805*, 2022.
- Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309 – 2352, 2017.
- Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *Annals of Statistics*, 50(5):2562–2586, 2022.
- M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley, 1997.
- M. Csörgö and P. Révész. *Strong Approximations in Probability and Statistics*. Probability and Mathematical Statistics : a series of monographs and textbooks. Academic Press, 1981.
- F. Eicker. The asymptotic distribution of the suprema of the standardized empirical processes. *Annals of Statistics*, 7(1):116 – 138, 1979.
- Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.
- Anja Göing-Jaeschke and Marc Yor. A survey and some generalizations of besell processes. *Bernoulli*, 9(2): 313 – 349, 2003.
- László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, 2002.
- Lajos Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of statistics*, 21(2):671–680, 1993.
- Jason M Klusowski and Peter M Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024.
- Rafał Latała and Dariusz Matlak. *Royen’s Proof of the Gaussian Correlation Inequality*, pages 265–275. Springer International Publishing, 2017.
- Valentin V. Petrov. On lower bounds for tail probabilities. *Journal of Statistical Planning and Inference*, 137(8):2703–2705, 2007.
- Galen R Shorack and RT Smythe. Inequalities for $\max_{k \in \mathbb{N}} \frac{s_k}{b_k}$ where $k \in \mathbb{N}$. *Proceedings of the American Mathematical Society*, pages 331–336, 1976.
- Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.