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

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

March 17, 2026

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

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 $\|\mathbf{x}\|_\infty = \max_{i \le d} |x_i|$. 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 $\|A\|_{\max} = \max_{1 \le i \le m, 1 \le j \le n} |A_{ij}|$. 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 $\mathscr{D} = \{X_1, \cdots, X_n\}$, denote the empirical L_2 norm by $\|X\|_{\mathscr{D}} = (n^{-1} \sum_{i=1}^n \|X_i\|^2)^{1/2}$.

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

Other. $\mathbb{1}(\cdot)$ denotes the indicator function. For two random variables X and Y, $X \perp\!\!\!\perp Y$ 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. $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathsf{Beta}(\alpha, \beta)$ denotes the Beta distribution with parameter (α, β). A stochastic process $\{B(t), 0 \le t \le 1\}$ 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.
- **Proof of Lemma 4**: See Section SA-4.1.

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

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0, \qquad \mathbb{E}\big[\varepsilon_i^2 \mid \mathbf{x}_i\big] = \sigma^2(\mathbf{x}_i), \tag{SA-1}$$

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). *$\mathscr{D} = \{(y_i, \mathbf{x}_i^T) : 1 \le i \le n\}$ is a random sample such that the following conditions hold for all $i = 1, 2, \cdots, n$, 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 $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i$.*
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 $\mathscr{D}_\mu = \{(y_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\mu\}$, with $n_\mu \le n$, is the dataset used. Let $\mathtt{t}$ be the unique terminal node in T containing $\mathbf{x} \in \mathcal{X}$. The CART estimator is*

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

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

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

$$\min_{1 \le j \le p} \min_{\beta_\mathtt{L}, \beta_\mathtt{R}, \varsigma \in \mathbb{R}} \sum_{\mathbf{x}_i \in \mathtt{t}} \big(y_i - \beta_\mathtt{L} \mathbb{1}(x_{ij} \le \varsigma) - \beta_\mathtt{R} \mathbb{1}(x_{ij} > \varsigma)\big)^2, \tag{SA-2}$$

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

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

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

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

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

$$\hat{\mu}^{\texttt{HON}}(\mathbf{x}) = \hat{\mu}(\mathbf{x}; \mathsf{T}(\mathscr{D}_{\mathsf{T}}), \mathscr{D}_{\mu}).$$

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

1. *The dataset $\mathscr{D}$ is divided into $K+1$ datasets $(\mathscr{D}_{\mathsf{T}_1}, \ldots, \mathscr{D}_{\mathsf{T}_K}, \mathscr{D}_{\mu})$, with sample sizes $(n_{\mathsf{T}_1}, \ldots, n_{\mathsf{T}_K}, n_{\mu})$, respectively, and satisfying $n_{\mathsf{T}_1} = \cdots = n_{\mathsf{T}_K} = n_{\mu}$ (possibly after dropping $n \mod K$ data points at random). For each of the datasets $\mathscr{D}_{\mathsf{T}_j} = \{(y_i, \mathbf{x}_i^\top) : i = 1, \ldots, n_{\mathsf{T}_j}\}$, $j = 1, \ldots, K$, replace $\{y_i : i = 1, \ldots, n_{\mathsf{T}_j}\}$ with independent copies $\{\tilde{y}_i : i = 1, \ldots, n_{\mathsf{T}_j}\}$, while keeping the same $\{\mathbf{x}_i : i = 1, \ldots, n_{\mathsf{T}_j}\}$.*

2. *The maximal decision tree of depth K, $\mathsf{T}_K(\mathscr{D}_{\mathsf{T}_1}, \cdots, \mathscr{D}_{\mathsf{T}_K})$, is obtained by iterating K times the $l \in \{\texttt{DIM}, \texttt{IPW}, \texttt{SSE}\}$ 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*

$$\hat{\mu}^{\texttt{X}}(\mathbf{x}; K) = \hat{\mu}(\mathbf{x}; \mathsf{T}_K(\mathscr{D}_{\mathsf{T}_1}, \ldots, \mathscr{D}_{\mathsf{T}_K}), \mathscr{D}_{\mu}).$$

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

$$\begin{aligned} &\sum_{\mathbf{x}_i \in \mathrm{t}} \left(y_i - \overline{y}_{\mathrm{t}}\right)^2 - \sum_{\mathbf{x}_i \in \mathrm{t}} \left(y_i - \overline{y}_{\mathrm{t}_L} \mathbb{1}(\mathbf{x}_i \in \mathrm{t}_L) - \overline{y}_{\mathrm{t}_R} \mathbb{1}(\mathbf{x}_i \in \mathrm{t}_R)\right)^2 \\ &\quad = \frac{\left(\frac{1}{\sqrt{n(\mathrm{t})}} \sum_{\mathbf{x}_i \in \mathrm{t}_L} (y_i - \mu) - \frac{n(\mathrm{t}_L)}{n(\mathrm{t})} \frac{1}{\sqrt{n(\mathrm{t})}} \sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - \mu)\right)^2}{(n(\mathrm{t}_L)/n(\mathrm{t}))(1 - n(\mathrm{t}_L)/n(\mathrm{t}))}, \end{aligned} \tag{SA-3}$$

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

$$\frac{n(t_L) n(t_R)}{n(\mathrm{t})} \left(\overline{y}_{\mathrm{t}_L} - \overline{y}_{\mathrm{t}_R}\right)^2. \tag{SA-4}$$

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 $\hat{\jmath}$, and a splitting index $\hat{\imath}$ such that

$$\mathrm{t}_L = \{\mathbf{u} \in X : \mathbf{u}_{\hat{\jmath}} \leq x_{\pi_j(\imath),j}\}, \qquad \mathrm{t}_R = \{\mathbf{u} \in X : \mathbf{u}_{\hat{\jmath}} > x_{\pi_j(\imath),j}\}.$$

The tree output can then be written as

$$\hat{\mu}^{\texttt{NSS}}(\mathbf{x}) = \begin{cases} \bar{y}_{\mathrm{t}_L}, & \mathbf{x} \in \mathrm{t}_L \\ \bar{y}_{\mathrm{t}_R}, & \mathbf{x} \in \mathrm{t}_R \end{cases}, \tag{SA-5}$$

where $x_{\hat{\jmath}}$ denotes the value of the $\hat{\jmath}$-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 $\hat{\imath}$, 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 $\hat{\mu} = x_{i\hat{\jmath}}$ and hence $\hat{\imath} = \#\{\mathbf{x}_i : x_{i\hat{\jmath}} \leq \hat{\mu}\}$).

Theorem SA-1 (Imbalanced Splits)**.** *Suppose Assumption SA–1 holds, and let $(\hat{\imath}, \hat{\jmath})$ 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*

$$\liminf_{n\to\infty} \mathbb{P}\big(n^a \leq \hat{\imath} \leq n^b, \hat{\jmath} = \ell\big) = \liminf_{n\to\infty} \mathbb{P}\big(n - n^b \leq \hat{\imath} \leq n - n^a, \hat{\jmath} = \ell\big) \geq \frac{b-a}{2pe}, \tag{SA-6}$$

which implies

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

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*

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

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

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

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

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

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 $\mathbb{P}_X$.

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

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

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

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

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

SA-2.2 Sample Splitting

For 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*

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

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

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

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

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

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

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 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 $\liminf_{n\to\infty}\frac{K_n}{\log\log(n)} > 0$, then there exists a positive constant C not depending on n such that*

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

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*

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

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

$$\mathbb{E}\left[\int_{\mathscr{X}} (\hat{\mu}^{\mathtt{X}}(\mathbf{x}; K) - \mu)^2 dF_{\mathbf{X}}(\mathbf{x})\right] \leq C\frac{K2^K \log(n)^5}{n},$$

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

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 $\mathbf{x} \in \mathscr{X}$, where c_d is some constant, and $\mathscr{X}$ is the support of $\mathbf{x}_i$.*
4. *$x_{i,1}, \ldots, x_{i,p}$ 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 T is the tree used, and $\mathscr{D}_\tau = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n_\tau\}$, with $n_\tau \leq n$, is the dataset used. Let $\mathtt{t}$ be the unique terminal node in T containing $\mathbf{x} \in \mathscr{X}$.*

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

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

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

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

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

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

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

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

$$\frac{n(\mathtt{t}_{\mathtt{L}})n(\mathtt{t}_{\mathtt{R}})}{n(\mathtt{t})}\Big(\hat{\tau}_l(\mathtt{t}_{\mathtt{L}}; \mathsf{T}, \mathscr{D}_{\mathsf{T}}) - \hat{\tau}_l(\mathtt{t}_{\mathtt{R}}; \mathsf{T}, \mathscr{D}_{\mathsf{T}})\Big)^2, \qquad l \in \{\mathtt{DIM}, \mathtt{IPW}\}. \tag{SA-9}$$

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

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

$$\min_{a_{\mathtt{L}}, b_{\mathtt{L}}, a_{\mathtt{R}}, b_{\mathtt{R}} \in \mathbb{R}} \sum_{\mathbf{x}_i \in \mathtt{t}_{\mathtt{L}}} (y_i - a_{\mathtt{L}} - b_{\mathtt{L}} d_i)^2 + \sum_{\mathbf{x}_i \in \mathtt{t}_{\mathtt{R}}} (y_i - a_{\mathtt{R}} - b_{\mathtt{R}} d_i)^2, \tag{SA-10}$$

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

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

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

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

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

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

While the estimators $\hat{\tau}^{\mathtt{NSS}}_l(\mathbf{x})$ and $\hat{\tau}^{\mathtt{HON}}_l(\mathbf{x})$, $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ depend on the depth of the tree construction used, our the 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* $\mathscr{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \ldots, n\}$ *is the available random sample.*

1. *The dataset $\mathscr{D}$ is divided into $K+1$ datasets $(\mathscr{D}_{\mathsf{T}_1},\dots,\mathscr{D}_{\mathsf{T}_K},\mathscr{D}_\tau)$, with sample sizes $(n_{\mathsf{T}_1},\dots,n_{\mathsf{T}_K},n_\tau)$, respectively, and satisfying $n_{\mathsf{T}_1}=\cdots=n_{\mathsf{T}_K}=n_\tau$ (possibly after dropping $n \bmod K$ data points at random). For each of the datasets $\mathscr{D}_j=\{(y_i,d_i,\mathbf{x}_i^\top): i=1,\dots,n_{\mathsf{T}_j}\}$, $j=1,\dots,K$, replace $\{(y_i,d_i): i=1,\dots,n_{\mathsf{T}_j}\}$ with independent copies $\{(\tilde{y}_i,\tilde{d}_i): i=1,\dots,n_{\mathsf{T}_j}\}$, while keeping the same $\{\mathbf{x}_i: i=1,\dots,n_{\mathsf{T}_j}\}$.*

2. *The maximal decision tree of depth K, $\mathsf{T}_K^l(\mathscr{D}_{\mathsf{T}_1},\cdots,\mathscr{D}_{\mathsf{T}_K})$, is obtained by iterating K times the $l\in\{\mathtt{DIM},\mathtt{IPW},\mathtt{SSE}\}$ 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 $\mathbf{X}$-adaptive estimators are*

$$\begin{aligned}
\hat{\tau}^{\mathtt{X}}_{\mathtt{DIM}}(\mathbf{x};K)&=\hat{\tau}_{\mathtt{DIM}}(\mathbf{x};\mathsf{T}_K^{\mathtt{DIM}}(\mathscr{D}_{\mathsf{T}_1},\dots,\mathscr{D}_{\mathsf{T}_K}),\mathscr{D}_\tau),\\
\hat{\tau}^{\mathtt{X}}_{\mathtt{IPW}}(\mathbf{x};K)&=\hat{\tau}_{\mathtt{IPW}}(\mathbf{x};\mathsf{T}_K^{\mathtt{IPW}}(\mathscr{D}_{\mathsf{T}_1},\dots,\mathscr{D}_{\mathsf{T}_K}),\mathscr{D}_\tau), \quad \textit{and}\\
\hat{\tau}^{\mathtt{X}}_{\mathtt{SSE}}(\mathbf{x};K)&=\hat{\tau}_{\mathtt{DIM}}(\mathbf{x};\mathsf{T}_K^{\mathtt{SSE}}(\mathscr{D}_{\mathsf{T}_1},\dots,\mathscr{D}_{\mathsf{T}_K}),\mathscr{D}_\tau).
\end{aligned}$$

SA-3.1 IPW Estimator

The transformed outcomes $y_i\frac{d_i-\xi}{\xi(1-\xi)}, 1\le i\le n$, 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

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

Assumption SA–2 implies $\mathbb{E}[\exp(\lambda\tilde{\varepsilon}_i)]<\infty$ for all $|\lambda|\le 1/\beta$ 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 $\ell\in[p]$,*

$$\liminf_{n\to\infty}\mathbb{P}\big(n^a\le\hat{\imath}\le n^b,\hat{\jmath}=\ell\big)=\liminf_{n\to\infty}\mathbb{P}\big(n-n^b\le\hat{\imath}\le n-n^a,\hat{\jmath}=\ell\big)\ge\frac{b-a}{2pe}.$$

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*

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

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

then

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

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

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

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

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

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,*

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

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

SA-3.1.2 Sample Splitting

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

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

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

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

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

provided $\rho^{-1} \leq \frac{n_{\mathcal{I}}}{n_\tau} \leq \rho$ *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,*

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

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* $\liminf_{n\to\infty} \frac{K_n}{\log\log(n)} > 0$. *Then, there exists a positive constant* C *not depending on* n *such that*

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

Corollary SA-16 (L2 Consistency – X)**.** *Suppose Assumption SA–2 holds. Then*

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

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 $x_{\pi_\ell(1),\ell} \leq x_{\pi_\ell(2),\ell} \leq \cdots \leq x_{\pi_\ell(n),\ell}$, $1 \leq \ell \leq p$. 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

$$\begin{aligned}
\mathscr{I}^{\mathtt{DIM}}(k,\ell) &= \frac{k(n-k)}{n}\Big(\hat{\tau}^{\mathtt{DIM}}_L(k,\ell) - \hat{\tau}^{\mathtt{DIM}}_R(k,\ell)\Big)^2, \\
\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell) &= \frac{k(n-k)}{n}\Big(\bar{\tau}^{\mathtt{IPW}}_L(k,\ell) - \bar{\tau}^{\mathtt{IPW}}_R(k,\ell)\Big)^2,
\end{aligned}$$

where

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

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

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

$$\mathrm{t}_L = \{\mathbf{u} \in \mathscr{X} : \mathbf{u}_{\hat{\jmath}_{\mathrm{DIM}}} \leq x_{\pi_{\hat{\jmath}_{\mathrm{DIM}}}(\hat{\imath}_{\mathrm{DIM}}),\hat{\jmath}_{\mathrm{DIM}}}\}, \qquad \mathrm{t}_R = \{\mathbf{u} \in \mathscr{X} : \mathbf{u}_{\hat{\jmath}_{\mathrm{DIM}}} > x_{\pi_{\hat{\jmath}_{\mathrm{DIM}}}(\hat{\imath}_{\mathrm{DIM}}),\hat{\jmath}_{\mathrm{DIM}}}\},$$

that maximizes

$$\frac{n(t_L)n(t_R)}{n(\mathrm{t})}\Big(\hat{\tau}_{\mathrm{DIM}}(t_L) - \hat{\tau}_{\mathrm{DIM}}(t_R)\Big)^2,$$

that is,

$$(\hat{\imath}_{\mathrm{DIM}}, \hat{\jmath}_{\mathrm{DIM}}) = \arg\max_{k,\ell} \mathscr{I}^{\mathtt{DIM}}(k,\ell).$$

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 $r_n \to \infty$. 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 $r_n \to \infty$. Then*

$$\max_{1\leq\ell\leq p}\max_{r_n\leq k<n-r_n}\Big|\mathscr{I}^{\mathtt{DIM}}(k,\ell) - \bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)\Big| = O_{\mathbb{P}}\left(\frac{\log\log(n)}{\sqrt{r_n}}\right).$$

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 $\limsup_{n\to\infty}\rho_n\log\log(n) = 0$, and take $s_n = \exp((\log n)^{\rho_n})$. Then*

$$\max_{1\leq\ell\leq p}\max_{1\leq k\leq s_n, n-s_n\leq k\leq n}\Big|\mathscr{I}^{\mathtt{DIM}}(k,\ell) - \bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)\Big| = O_{\mathbb{P}}\left(\rho_n\log\log(n) + \frac{s_n}{n-s_n}\log\log(n)\right).$$

Rates for Decision Stumps.

The previous two lemmas imply that we can study arg max of $\mathscr{I}^{\mathtt{DIM}}$ in terms of arg max of $\bar{\mathscr{I}}^{\mathtt{IPW}}$. 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 $\ell \in [p]$,*

$$\liminf_{n\to\infty}\mathbb{P}\big(n^a \leq \hat{\imath}_{\mathtt{DIM}} \leq n^b, \hat{\jmath}_{\mathtt{DIM}} = \ell\big) = \liminf_{n\to\infty}\mathbb{P}\big(n - n^b \leq \hat{\imath}_{\mathtt{DIM}} \leq n - n^a, \hat{\jmath}_{\mathtt{DIM}} = \ell\big) \geq \frac{b-a}{2pe}.$$

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 $a < b$,*

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

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* $\mathbf{x}_i \sim \mathsf{Uniform}([0,1]^p)$, *then*

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

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* $b \in (0,1)$,

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

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,*

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

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

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

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 Sample Splitting

With the 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* $b \in (0,1)$,

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

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

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

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

provided $\rho^{-1} \leq \frac{n_{\mathcal{I}}}{n_\tau} \leq \rho$ *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,*

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

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* $\liminf_{n\to\infty} \frac{K_n}{\log\log(K_n)} > 0$. *Then*

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

where C *is some positive constant not depending on* n.

A direct variance calculation gives

Theorem SA-26 (L2 Consistency)**.** *Suppose Assumption SA–2 holds. Then*

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

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

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

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

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

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

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

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

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

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

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

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

with

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

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 $r_n \to \infty$. Then*

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

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

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

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 $\ell \in [p]$,*

$$\liminf_{n\to\infty}\mathbb{P}\big(n^a \le \hat{\imath}_{SSE} \le n^b, \hat{\jmath}_{SSE} = \ell\big) = \liminf_{n\to\infty}\mathbb{P}\big(n-n^b \le \hat{\imath}_{SSE} \le n-n^a, \hat{\jmath}_{SSE} = \ell\big) \ge \frac{b-a}{2pe}.$$

Remark SA-1. *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*

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

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

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

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* $\mathbb{N}$,

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

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,*

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

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

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

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 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* $\mathbb{N}$*,*

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

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

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,*

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

provided $\rho^{-1} \leq \frac{n_{\mathtt{T}}}{n_\tau} \leq \rho$ *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,*

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

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* $\liminf_{n\to\infty} \frac{K_n}{\log\log(n)} > 0$*. Then*

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

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

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

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 $(\hat{\imath}, \hat{\jmath})$ are chosen so that the *squared T-statistics metric* is maximized, that is,

$$\hat{\imath}, \hat{\jmath} = \underset{k\in[n], \ell\in[p]}{\arg\max}\ n \frac{(\hat{\tau}_L(k,\ell) - \hat{\tau}_R(k,\ell))^2}{S(k,\ell)^2/k + S(k,\ell)^2/(n-k)},$$

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,

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

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

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

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*

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

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

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

SA-4 Proofs

SA-4.1 Proof of Lemma 4

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

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

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

$$L = \log\log n, \qquad A = \log c, \qquad L \to \infty \ \ (n\to\infty).$$

First, we present some elementary expansions,

$$\sqrt{2(A+L)} = \sqrt{2L}\sqrt{1+\frac{A}{L}} = \sqrt{2L}\Big(1+\frac{A}{2L}-\frac{A^2}{8L^2}+\mathrm{O}\big(L^{-3}\big)\Big),$$
$$\frac{1}{\sqrt{2(A+L)}} = \frac{1}{\sqrt{2L}}\Big(1-\frac{A}{2L}+\frac{3A^2}{8L^2}+\mathrm{O}\big(L^{-3}\big)\Big),$$
$$\log(L+A) = \log L + \frac{A}{L} - \frac{A^2}{2L^2} + \mathrm{O}\big(L^{-3}\big).$$

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

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

Then

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

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

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

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

SA-4.2 Proof of Theorem SA-1

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

$\ell \in [p]$ by

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

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

$$(\hat{\imath}, \hat{\jmath}) = \operatorname*{arg\,max}_{(i,j)\in[n]\times[p]} \mathscr{I}(i,j).$$

where

$$\mathscr{I}(k,\ell) = \frac{k(n-k)}{n}\Big(\hat{\mu}_L(k,\ell) - \hat{\mu}_R(k,\ell)\Big)^2, \qquad k \in [n], \quad \ell \in [p].$$

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,

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

The rest of the proof is organized as follows. In Section SA-4.2.1, we prove the results under $p = 1$, showing a strong approximation of the split criterion $(\mathscr{I}(k,1) : k \in [n])$ 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-4.2.2, we generalize to allow for $p \geq 1$. We show that the split criterion over different coordinates, that is, $(\mathscr{I}(k,\ell) : k \in [n])$ 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-4.2.1 can be used.

SA-4.2.1 Univariate Case

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

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

By Csörgö 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

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

where $\epsilon_n = o_{\mathbb{P}}\big((\log\log(n))^{-1/2}\big)$. We note that while Csörgö 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 $1 \le k < n^a$, $n^b < k < n$. Thus,

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

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

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

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

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

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 $\left[0,\, 2\log(n-1)\right]$ 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

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

where the last equality follows from stationarity.

Remark SA-2. *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 [Csörgö 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

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\sup_{0\le t\le (c+o(1))\log(n)} |U(t)| < \frac{2\log\log(n) + (1/2)\log\log\log(n) + z - (1/2)\log(\pi)}{\sqrt{2\log\log(n)}} \Bigg) = \exp\Big(- e^{-(z-\log(c))} \Big). \tag{SA-17}$$

For a detailed proof of a generalized result on multidimensional O-U process, see Section SA-4.1.

Let z^* maximize $z \mapsto \exp\big(- 2e^{-(z-\log(2-(b-a)))}\big) - \exp\big(- 2e^{-(z-\log(2))}\big)$, 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 $\epsilon_n = o_{\mathbb{P}}\big((\log\log(n))^{-1/2}\big)$. We have that

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

Remark SA-3. *Alternatively, for any* $0 < A < B < C$*, we have*

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

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* $\mathbb{P}(\max_{1\le k\le n} \mathscr{I}(k,1) > \max_{1\le k<n^a, n^b<k<n} \mathscr{I}(k,1))$ *by*

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

SA-4.2.2 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, $(\mathscr{I}(k,\ell) : k \in [n])$ for different ℓ's, are asymptotically independent, which will imply

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}\big(n^a \leq \hat{\imath} \leq n^b, \hat{\jmath} = \ell\big) \\
&= \liminf_{n} \mathbb{P}\Big(\max_k \mathscr{I}(k,1) > \max_{k, j\neq 1} \mathscr{I}(k,j),\ \max_k \mathscr{I}(k,1) > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1)\Big) \\
&\geq \liminf_{n} \mathbb{P}\Big(\max_k \mathscr{I}(k,1) > z_n > \max_{k, j\neq 1} \mathscr{I}(k,j),\ \max_k \mathscr{I}(k,1) > z_n > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1)\Big) \\
&\overset{(*)}{=} \Big(\liminf_{n} \mathbb{P}\Big(\max_k \mathscr{I}(k,1) < z_n\Big)\Big)^{p-1} \liminf_{n} \mathbb{P}\Big(\max_k \mathscr{I}(k,1) > z_n > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1)\Big).
\end{aligned}$$

where in equality (*) we use asymptotic independence between $\mathscr{I}(\cdot,\ell)$ 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-4.2.1.

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,

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

where $\#\pi^\ell : [n] \to [n]$ 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

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

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 $\mathscr{B}$, the σ-algebra generateed by the p permulations $\pi^1, \cdots, \pi^p$, $\mathbf{C}_i$'s are independent, and for all $1 \leq j \leq p(n - 2r_n)$, $1 \leq \ell \leq p$, we have

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

where we assume row j in $\mathbf{C}_i$ corresponds to $\sqrt{n}\sqrt{\frac{n}{k(n-k)}}(\mathbb{1}(\#\pi^\ell(i) \leq k) - \frac{k}{n})$. 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,

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

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

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

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

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

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

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

Then for each $1\le j\le \mathtt{P}$, $1\le i\le n$, 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

$$\sup_{A\in\mathscr{A}^{re}}\left|\mathbb{P}\Big(\sum_{i=1}^{n}\mathbf{C}_i\in A\Big|\mathscr{B}\Big)-\mathbb{P}\Big(\sum_{i=1}^{n}\boldsymbol{D}_i\in A\Big|\mathscr{B}\Big)\right|\le K_1\left[\Big(\frac{\bar{L}_n^2\log^7(\mathtt{P})}{n}\Big)^{1/6}+\frac{M_{n,X}(\phi_n)+M_{n,Y}(\phi_n)}{\bar{L}_n}\right]$$
$$\lesssim\left(\frac{\log^7(\mathtt{P})}{r_n}\right)^{1/6}+\sqrt{\frac{r_n}{n}}\frac{1}{n^2}$$
$$\lesssim\left(\frac{\log^7(n)}{r_n}\right)^{1/6},\tag{SA-22}$$

where $\mathscr{A}^{\text{re}}$ is the class of all rectangles A of the form

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

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

$$\boldsymbol{D}_i=\sqrt{n}\Big(\Big(\sqrt{\frac{n}{k(n-k)}}(\mathbb{1}(\#\pi^\ell(i)\le k)-\frac{k}{n}):r_n\le k\le n-r_n\Big)^\top:1\le\ell\le p\Big)^\top u_i.$$

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 $\mathscr{B}$, we get

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

Step 2: Gaussian to Gaussian Coupling.

For $1\le\ell\le p$, 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 $\mathbf{G}_n=((G_n^1(1/n),G_n^1(2/n),\cdots,G_n^1(n/n))^\top,\cdots,(G_n^p(1/n),G_n^p(2/n),\cdots,G_n^p(n/n))^\top)^\top$. Then $\mathbf{G}_n$ is a np-dimensional Gaussian random vector, and denote by $\mathbf{\Sigma}_n$ its covariance matrix. We want to show that $\mathbf{\Sigma}_n$ 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 $\boldsymbol{\sigma}$ the sigma-algebra generated by $\pi_1, \cdots, \pi_p$. Then

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

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

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

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 $|f(\pi) - f(\pi^{s,t})| \le \frac{n}{jk}$ for all π, $s \in \{1, \cdots, j\}$, $t \in \{j+1, \cdots, n\}$, where the permutation $\pi^{s,t}$ is obtained from π by transposition of its sth and tth coordinates. We will show later that w.l.o.g. we can assume $j, k \le \lceil n/2 \rceil$. Then by Lemma 2 from El-Yaniv and Pechyony [2009], for any $t \ge 0$,

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

Since $\frac{n-1/2}{n-j}(1 - \frac{1}{2\max(j,n-j)}) \ge 1 - \frac{1}{n}$, 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 $j, k \in [n]$,

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

which implies

$$|\operatorname{Cov}[G_n^1(\frac{k}{n}), G_n^2(\frac{j}{n})]| \le C\sqrt{\frac{jk}{(n-k)(n-j)}}\frac{n}{\sqrt{jk}}\frac{1}{\sqrt{k}} \le Ck^{-1/2}. \tag{SA-23}$$

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

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

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

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

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

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

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

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

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

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

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, $1 \le \ell \le p$, then

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

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

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

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 $\mathbf{y} = (t_1 \mathbf{1}^\top, \cdots t_p \mathbf{1}^\top)$, $\mathbf{y} = -(t_1 \mathbf{1}^\top, \cdots t_p \mathbf{1}^\top)$ separately with $\mathbf{1}$ a vector of $T(n)$ 1's, we get

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

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 $\mathscr{E} = \{\exists \ell \in [p] : \arg\max_k \mathscr{I}(k, \ell) <$

r_n or $\arg\max_k \mathscr{I}(k,\ell) > n - r_n\}$. Then by [Csörgö and Horváth, 1997, proof of Theorem A.4.2], and a union bound argument, we have

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

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

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}\big(n^a \le \hat{\imath} \le n^b, \hat{\jmath} = \ell\big) \\
&= \mathbb{P}\Big(\max_{k\in[n]} \mathscr{I}(k,1) > \max_{k,j\neq 1} \mathscr{I}(k,j),\ \max_{k\in[n]} \mathscr{I}(k,1) > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1)\Big) \\
&\ge \mathbb{P}\Big(\max_{k\in[n]} \mathscr{I}(k,1) > \max_{k,j\neq 1} \mathscr{I}(k,j),\ \max_{k\in[n]} \mathscr{I}(k,1) > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1), \mathscr{E}^c\Big) - \mathbb{P}(\mathscr{E}) \\
&\ge \mathbb{P}\Big(\max_{k\in[r_n,n-r_n]} \mathscr{I}(k,1) > \max_{k,j\neq 1} \mathscr{I}(k,j),\ \max_{k\in[r_n,n-r_n]} \mathscr{I}(k,1) > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1)\Big) - 2\mathbb{P}(\mathscr{E}) \\
&\ge \mathbb{P}\Big(\max_{k\in[r_n,n-r_n]} \mathscr{I}(k,1) > \max_{k,j\neq 1} \mathscr{I}(k,j),\ \max_{k\in[r_n,n-r_n]} \mathscr{I}(k,1) > \max_{k\notin[n^a,n^b]} \mathscr{I}(k,1)\Big) + o(1).
\end{aligned}$$

Now we can using the *coupling result* developed previously. Using our notation, we have $\mathscr{I}(k,\ell) = (H_n^\ell(k/n))^2$. Hence

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

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

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

where the third line is by similar calculation as in Section SA-4.2.1. Putting together, we have

$$\liminf_{n\to\infty} \mathbb{P}\big(n^a \le \hat{\imath} \le n^b, \hat{\jmath} = \ell\big) \ge \frac{b-a}{2pe},$$

and by symmetry, we have

$$\liminf_{n\to\infty} \mathbb{P}\big(n - n^a \le \hat{\imath} \le n - n^b, \hat{\jmath} = \ell\big) \ge \frac{b-a}{2pe}.$$

SA-4.3 Proof of Theorem SA-2

For simplicity, we denote $\hat{\mu}^{\texttt{NSS}}(\mathbf{x})$ 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*,

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

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

$$\imath_\ell = \arg\max_{k\in[n]} \mathscr{I}(k,\ell), \qquad \ell \in [p].$$

Consider the event

$$\begin{aligned}\texttt{Imbalance}_\ell &= \{\hat{\jmath} = \ell, \hat{\imath} < n^b \text{ or } \hat{\imath} > n - n^b\} \\ &= \{\max_k \mathscr{I}(k,\ell) > \max_{k, j\neq\ell} \mathscr{I}(k,j),\ \max_k \mathscr{I}(k,\ell) > \max_{k\in[n^b, n-n^b]} \mathscr{I}(k,\ell)\}, \qquad \ell \in [p].\end{aligned}$$

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

$$\begin{aligned}&\sup_{\mathbf{x}\in\mathscr{X}} |\hat{\mu}(\mathbf{x}) - \mu|^2 \\ &\ge \frac{S(\imath_\ell,\ell)^2}{\imath_\ell^2} \\ &\ge \frac{1}{\imath_\ell}\left[\frac{S(\imath_\ell,\ell)^2}{\imath_\ell} + \frac{(S(n,\ell) - S(\imath_\ell,\ell))^2}{n-\imath_\ell} - \frac{(S(n,\ell) - S(\imath_\ell,\ell))^2}{n-\imath_\ell}\right] \\ &\ge \frac{1}{\min\{\imath_\ell, n-\imath_\ell\}}\left(\max_{k\in[n]}\left(\frac{S(k,\ell)^2}{k} + \frac{(n - S(k,\ell))^2}{n-k}\right) - \max_{\lfloor n/2\rfloor \le k\le n} \frac{S(k,\ell)^2}{k} - \max_{1\le k\le \lceil n/2\rceil} \frac{(n-S(k,\ell))^2}{n-k}\right).\end{aligned}$$

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

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

Fix $\epsilon>0$. Consider the events

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

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

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

where we have used the fact that $\texttt{Imbalance}_\ell$'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

$$\begin{aligned}\texttt{Off}_\ell &= \{\hat{\jmath}=\ell, \hat{\imath}\in[n^a,n^b]\} \\ &= \{\max_k \mathscr{I}(k,\ell) > \max_{k,j\neq\ell}\mathscr{I}(k,j),\ \max_k\mathscr{I}(k,\ell) > \max_{k\notin[n^a,n-n^b]}\mathscr{I}(k,\ell)\}, \qquad \ell\in[p].\end{aligned}$$

Since π_ℓ is the uniform permutation, we have

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

Together with Theorem SA-1,

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

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

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

By similar arguments as Part 1, we can show

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

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 $\hat{\imath}$ at the K_n-th split ($K_n \geq 1$) also satisfies

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

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 $\hat{\mu}^{\texttt{HON}}(\mathbf{x})$ by $\check{\mu}(\mathbf{x})$. Recall $(\hat{\imath}, \hat{\jmath})$ denotes the optimal splitting index and coordinate for the decision stump. We use $(y_i, \mathbf{x}_i^\top)_{i=1}^M$ to denote $\mathcal{D}_{\texttt{HON},1}$, which we used to construct the causal tree. Denote by $(\hat{\imath}, \hat{\jmath})$ 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 $\mathcal{D}_{\texttt{HON},2}$. By Definition SA-2, $n \lesssim M, N \lesssim n$. Then

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

Since $\tilde{y}_i \perp\!\!\!\perp \tilde{\mathbf{x}}_i$, condition on $\hat{\imath}, \hat{\jmath}$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_n)$, we have

$$\frac{\sum_{i=1}^N (\tilde{y}_i - \mu)\mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \leq x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}})}{\sum_{i=1}^N \mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \leq x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}})} \overset{d}{=} \frac{1}{\tilde{\imath}} \sum_{i=1}^{\tilde{\imath}} (y_i - \mu),$$

where

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

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

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

where in the second to last line, we have used Jensen's inequality, and in the last line we have used $\tilde{\imath}$ is measurable with resepect to the σ-algebra generated by $\hat{\imath},\hat{\jmath},\mathbf{X},\mathbf{X}'$. Then by Paley–Zygmund inequality, for any $\theta\in(0,1)$,

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

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

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

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

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

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

$$\tilde{\imath}/n \le 4M^{b-1} \le Cn^{b-1},$$

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

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

$$\liminf_{M\to\infty} \mathbb{P}(\hat{\imath} \le M^b) \ge \frac{b}{2e}.$$

Hence we have

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

This proves the conclusion.

SA-4.7 Proof of Theorem SA-6

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

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

and

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

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

$$\|y - \hat{\mu}(\mathsf{T})\|_{\mathscr{D}_\mu}^2 \le \|y - \bar{y}\|_{\mathscr{D}_\mu}^2 \le \|y - g\|_{\mathscr{D}_\mu}^2,$$

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

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

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

for any $\varepsilon \in (0,1)$,

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

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

$$\mathbb{P}_{\mathscr{D}_\mu}(E_1 \geq 0 \mid \mathscr{D}_{\mathsf{T}}) \leq 14\left(\frac{2U^2}{\beta}\right)^{2^K} \exp\left(-\frac{\alpha n_{\mathsf{T}}}{2568U^4}\right).$$

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, $\tilde{\imath}_k$ be the CART split index of this node, and $\tilde{\jmath}_k$ be the CART split coordinate of this node, with $\tilde{n}_0 = n$ and $\tilde{\imath}_0 = \hat{\imath}$ (recall that $\hat{\imath}$ is the split index for the decision stump (SA-5)). Then, the left-most cell at the k-th level can be expressed as $\mathbf{t} \cap [0, x_{\pi_{\tilde{\jmath},\tilde{\jmath}}(\tilde{\imath}_{k-1})}]$ and $\tilde{n}_k = \tilde{\imath}_{k-1}$.

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 $\tilde{\imath}_{k-1} \mid \tilde{n}_{k-1} = m$ has the same distribution as $\tilde{\imath}_0 \mid \tilde{n}_0 = m$, because of the honest tree construction and Assumption SA–1. Therefore, we can apply (SA-6) to obtain

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

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

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

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

$$\begin{aligned}
\mathbb{P}(\tilde{n}_k \le m) &= \mathbb{P}(\tilde{n}_k \le m,\ \tilde{n}_{k-1} > m^c) + \mathbb{P}(\tilde{n}_k \le m,\ \tilde{n}_{k-1} \le m^c) && \text{(SA-28)}\\
&\ge \mathbb{P}(\tilde{n}_k \le m,\ \tilde{n}_{k-1} \le m^c) && \text{(SA-29)}\\
&= \mathbb{P}(\tilde{n}_k \le m,\ \tilde{n}_{k-1} \le m) + \mathbb{P}(\tilde{n}_k \le m,\ m < \tilde{n}_{k-1} \le m^c) && \text{(SA-30)}\\
&\ge \mathbb{P}(\tilde{n}_{k-1} \le m) + \delta \cdot \mathbb{P}(m < \tilde{n}_{k-1} \le m^c) && \text{(SA-31)}\\
&= (1-\delta) \cdot \mathbb{P}(\tilde{n}_{k-1} \le m) + \delta \cdot \mathbb{P}(\tilde{n}_{k-1} \le m^c). && \square
\end{aligned}$$

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 $\hat{\mu}^{\mathtt{X}}(\mathbf{x}; K)$ 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 $k \ge 0$ and $\ell \ge 1$ such that $n_\ell \ge M$,

$$\mathbb{P}(\tilde{n}_k \le n_\ell) \ge \sum_{k'=\ell}^{k} \binom{k'-1}{\ell-1} (1-\delta)^{k'-\ell} \delta^\ell .. \tag{SA-32}$$

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

$$\mathbb{P}(\tilde{n}_{k-1} \le n_{\ell'}) \ge \sum_{k'=\ell'}^{k-1} \binom{k'-1}{\ell'-1} (1-\delta)^{k'-\ell'} \delta^{\ell'}. \tag{SA-33}$$

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

$$\begin{aligned}
\mathbb{P}(\tilde{n}_k \le n_\ell) &\ge (1-\delta) \sum_{k'=\ell}^{k-1} \binom{k'-1}{\ell-1} (1-\delta)^{k'-\ell} \delta^\ell + \delta \sum_{k'=\ell-1}^{k-1} \binom{k'-1}{\ell-2} (1-\delta)^{k'-\ell+1} \delta^{\ell-1} && \text{(SA-34)}\\
&= \sum_{k'=\ell}^{k} \binom{k'-1}{\ell-1} (1-\delta)^{k'-\ell} \delta^\ell, && \text{(SA-35)}
\end{aligned}$$

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

Let $X \sim \mathrm{NB}(L, \delta)$, 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

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

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

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

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

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

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. □

SA-4.9 Proof of Theorem SA-8

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

Let $\mathtt{t}_1, \mathtt{t}_2, \cdots, \mathtt{t}_{2^K}$ 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 $N_1, N_2, \cdots, N_{2^K}$ and $m_1, m_2, \cdots, m_{2^K}$ denote the number of observations and the Lebesgue measure of the 2^K leaf nodes, respectively. Note that $\vec{N} = (N_1, \cdots, N_{2^K})$ are independent of the $\tilde{y}_i$ data by the honest condition and the x_i data per Assumption SA–1.

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

Thus, the IMSE can be bounded as follows:

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

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 $m_k \sim \mathsf{Beta}(N_k, N - N_k + 1)$, $k = 1, 2$.

Induction Step: $K \ge 2$. Let $\mathsf{t}_l^{\text{prev}}$ be a $(K-1)$-th level node, we annonate all relevant depth $K-1$ information with superscript prev. We already know condition on $N_1^{\text{prev}}, \cdots, N_{2^{K-1}}^{\text{prev}}$, $m_l^{\text{prev}} \sim \mathsf{Beta}(N_l^{\text{prev}}, N - N_l^{\text{prev}} + 1)$. Suppose $\mathsf{t}_l^{\text{prev}}$ is divided into $\mathsf{t}_{2l}, \mathsf{t}_{2l+1}$ 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 $\mathbf{x}_i \in \mathsf{t}_l^{\text{prev}}$, $\mathbf{x}_i \sim \mathsf{Uniform}(\mathsf{t}_l^{\text{prev}})$. Hence condition on N_l^{prev}, N_{2l} and m_l^{prev}, we have $m_{2l}/m_l^{\text{prev}} \sim \mathsf{Beta}(N_l^{\text{prev}}, N - N_l^{\text{prev}} + 1)$. Hence condition on $\vec{N} = (N_1, \cdots, N_{2^K})$, we have $m_k \sim \mathsf{Beta}(N_k, N - N_k + 1)$, $1 \le k \le 2^K$. 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.lo.g. we can assume $p = 1$ and drop the second index on the coordinate ℓ from $\mathscr{I}^{\texttt{DIM}}(k,\ell)$ and $\bar{\mathscr{I}}^{\texttt{IPW}}(k,\ell)$ everywhere. And throughout, we assume the data is already sorted so that

$$x_1 \le x_2 \le \cdots \le x_n.$$

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

$$\mathscr{I}^{\texttt{DIM}}(k) - \bar{\mathscr{I}}^{\texttt{IPW}}(k) = \frac{k(n-k)}{n} \underbrace{\left(\hat{\tau}^{\texttt{DIM}}_{\mathrm{t}_L}(k) - \hat{\tau}^{\texttt{DIM}}_{\mathrm{t}_R}(k) + \bar{\tau}^{\texttt{IPW}}_{\mathrm{t}_L}(k) - \bar{\tau}^{\texttt{IPW}}_{\mathrm{t}_R}(k)\right)}_{=:R_1(k)} \underbrace{\left(\hat{\tau}^{\texttt{DIM}}_{\mathrm{t}_L}(k) - \hat{\tau}^{\texttt{DIM}}_{\mathrm{t}_R}(k) - \bar{\tau}^{\texttt{IPW}}_{\mathrm{t}_L}(k) + \bar{\tau}^{\texttt{IPW}}_{\mathrm{t}_R}(k)\right)}_{=:R_2(k)}. \tag{SA-39}$$

We focus on the case where $1 \le k \le \frac{n}{2}$, the other case where $\frac{n}{2} < k \le n$ 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

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

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

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

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

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

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

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

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

$$\max_{r_n \le k \le n - r_n} \Big| \frac{\sum_{i=1}^k d_i \varepsilon_i(1)}{\sum_{i=1}^k d_i} \Big| = O_{\mathbb{P}}(r_n^{-1/2}). \tag{SA-41}$$

Hence

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

By similar arguments, we can show the same bound holds for other terms in the first two lines of Equation (SA-40). Hence

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

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

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

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

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

Putting together the parts for R_1 and R_2, we have

$$\max_{r_n \le k < n - r_n} |\mathscr{I}^{\mathtt{DIM}}(k) - \bar{\mathscr{I}}^{\mathtt{IPW}}(k)| = O_{\mathbb{P}}\Big(\frac{\log\log(n)}{r_n^{1/2}} \Big).$$

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.lo.g. we can assume $p = 1$ and drop the second index on the coordinate ℓ from $\mathscr{I}^{\texttt{DIM}}(k,\ell)$ and $\bar{\mathscr{I}}^{\texttt{IPW}}(k,\ell)$ everywhere. And throughout, we assume the data is already sorted so that

$$x_1 \le x_2 \le \cdots \le x_n.$$

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

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

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

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

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

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

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

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

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

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

It follows that

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

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

group,

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

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

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

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

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

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 $s_n\le n^a\le n^b\le n-s_n$. Consider the event $A_n:=\{\exists\ell\in[p]:\max_{k\in[n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell)>\max_{k\notin[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell)\}$. By Equation (A.4.18) from Csörgö and Horváth [1997],

$$\max_{1\le k\le s_n,n-s_n\le k\le n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)}=O_{\mathbb{P}}(\sqrt{\rho_n\log\log(n)}).$$

Then controlling the difference between $\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)$ and $\mathscr{I}^{\mathtt{DIM}}(k,\ell)$ by Lemma SA-18,

$$\max_{1\le k\le s_n,n-s_n\le k\le n}\mathscr{I}^{\mathtt{DIM}}(k,\ell)=O_{\mathbb{P}}\left(\rho_n\log\log(n)+\frac{s_n}{n-s_n}\log\log(n)\right) \tag{SA-47}$$

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

$$\begin{aligned}\max_{s_n<k<n-s_n}\sqrt{\mathscr{I}^{\mathtt{DIM}}(k,\ell)} &= \max_{s_n<k<n-s_n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)}+O_{\mathbb{P}}\left(\frac{\log\log(n)^{1/2}}{s_n^{1/4}}\right) \\ &\ge \max_{1\le k\le n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)}-\max_{1\le k\le s_n,n-s_n\le k\le n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)}+O_{\mathbb{P}}\left(\frac{\log\log(n)^{1/2}}{s_n^{1/4}}\right).\end{aligned}$$

Equation (A.4.20) in Csörgö and Horváth [1997] imply that $(2\log\log(n))^{-1/2}\max_{1\le k\le n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)} = 1+o_{\mathbb{P}}(1)$ and $(2\log\log(n))^{-1/2}\max_{1\le k\le s_n, n-s_n\le k\le n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)} = \rho_n(1+o_{\mathbb{P}}(1))$. Hence

$$\max_{1\le k\le n}\sqrt{\bar{\mathscr{I}}^{\mathtt{IPW}}(k,\ell)} \ge \sqrt{2\log\log(n)} + O_{\mathbb{P}}(\sqrt{\rho_n\log\log(n)}) + O_{\mathbb{P}}\left(\frac{\log\log(n)^{1/2}}{s_n^{1/4}}\right) \tag{SA-48}$$

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

$$\max_{1\le k\le s_n, n-s_n\le k\le n}\mathscr{I}^{\mathtt{DIM}}(k,\ell) = o_{\mathbb{P}}(\log\log(n)), \text{ and } \max_{s_n\le k\le n-s_n}\mathscr{I}^{\mathtt{DIM}}(k,\ell) = \sqrt{2\log\log(n)}(1+o_{\mathbb{P}}(1)).$$

Hence

$$\max_{1\le k\le s_n, n-s_n\le k\le n}\mathscr{I}^{\mathtt{DIM}}(k,\ell) = o_{\mathbb{P}}\left(\max_{s_n\le k\le n-s_n}\mathscr{I}^{\mathtt{DIM}}(k,\ell)\right), \qquad \ell\in[p],$$

which by a union bound implies

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

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

$$\begin{aligned}
&\mathbb{P}\Big(\exists\ell\in[p]: \max_k \mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{k,j\neq\ell}\mathscr{I}^{\mathtt{DIM}}(k,j),\ \max_k\mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{k\notin[n^a,n^b]}\mathscr{I}^{\mathtt{DIM}}(k,\ell)\Big)\\
&\ge \mathbb{P}\Big(\exists\ell\in[p]: \max_k \mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{k,j\neq\ell}\mathscr{I}^{\mathtt{DIM}}(k,j),\ \max_k\mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{k\notin[n^a,n^b]}\mathscr{I}^{\mathtt{DIM}}(k,\ell) \text{ and } A_n^c\Big) - \mathbb{P}(A_n)\\
&\ge \mathbb{P}\Big(\exists\ell\in[p]: \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{\substack{j\neq\ell\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,j),\\
&\qquad\qquad \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{k\notin[n^a,n^b],k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell)\Big) - 2\mathbb{P}(A_n).
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\Big(\exists\ell\in[p]: \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{\substack{j\neq\ell\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,j), \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,\ell) > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,\ell)\Big)\\
&= p\mathbb{P}\Big(\max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,1) > \max_{\substack{j\neq1\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,j), \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,1) > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,1)\Big)\\
&\ge p\sup_{z\in\mathbb{R}}\mathbb{P}\Big(\max_{\substack{j\neq1\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,j) < z, \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,1) > z > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,1)\Big)\\
&\ge p\sup_{z\in\mathbb{R}}\mathbb{P}\Big(\max_{\substack{j\neq1\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,j) < z, \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,1) < z\Big)\\
&\qquad - p\mathbb{P}\Big(\max_{\substack{j\neq1\\ k\in[s_n,n-s_n]}}\mathscr{I}^{\mathtt{DIM}}(k,j) < z, \max_{k\in[s_n,n-s_n]}\mathscr{I}^{\mathtt{DIM}}(k,1) > z\Big).
\end{aligned}$$

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

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

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

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

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

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

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

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

$$
\hat{\imath}_{\mathrm{DIM},\ell}=\arg\max_{k\in[n]}\bar{\mathscr{I}}^{\mathtt{DIM}}(k,\ell),\qquad \ell\in[p].
$$

For notational simplicity, denote

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

and consider the event

$$\texttt{Imblce}_\ell^{\texttt{DIM}} = \{\max_k \bar{\mathscr{I}}^{\texttt{DIM}}(k,\ell) > \max_{k,j\ne\ell} \bar{\mathscr{I}}^{\texttt{DIM}}(k,j),\ \max_k \bar{\mathscr{I}}^{\texttt{DIM}}(k,\ell) > \max_{k\notin[n^b,n-n^b]} \bar{\mathscr{I}}^{\texttt{DIM}}(k,\ell)\}, \qquad \ell\in[p].$$

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

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} |\hat{\tau}(x) - \tau|^2 \ge \bar{\tau}_L^{\texttt{DIM}}(\ell)^2\\
&\ge \frac{1}{\min\{\hat{\imath}_{\mathrm{DIM},\ell}, n - \hat{\imath}_{\mathrm{DIM},\ell}\}}\left(\hat{\imath}_{\mathrm{DIM},\ell}\bar{\tau}_L^{\texttt{DIM}}(\ell)^2 + (n-\hat{\imath}_{\mathrm{DIM},\ell})\bar{\tau}_R^{\texttt{DIM}}(\ell)^2 - (n-\hat{\imath}_{\mathrm{DIM},\ell})\bar{\tau}_R^{\texttt{DIM}}(\ell)^2 \mathbb{1}(\hat{\imath}_{\mathrm{DIM},\ell}\le n/2)\right).
\end{aligned} \tag{SA-49}$$

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

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

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

$$\begin{aligned}
\frac{\hat{\imath}_{\mathrm{DIM},\ell}(n-\hat{\imath}_{\mathrm{DIM},\ell})}{n}\left(\bar{\tau}_L^{\texttt{DIM}} - \bar{\tau}_R^{\texttt{DIM}}\right)^2 &= \frac{\hat{\imath}^{\mathrm{ipw}}(n-\hat{\imath}^{\mathrm{ipw}})}{n}\left(\bar{\tau}_L^{\texttt{IPW}} - \bar{\tau}_R^{\texttt{IPW}}\right)^2 + o_{\mathbb{P}}(\log\log(n))\\
&= \max_{1\le k\le n} \bar{\mathscr{I}}^{\texttt{IPW}}(k) + o_{\mathbb{P}}(\log\log n).
\end{aligned}$$

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

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

For simplicity in showing the upper bound for $\hat{\imath}_{\mathrm{DIM},\ell}\bar{\tau}_L^{\texttt{DIM}}(\ell)^2\mathbb{1}(\hat{\imath}_{\mathrm{DIM},\ell} > n/2)$, we assume π is the identity

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

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

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

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

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

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

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

$$\hat{\imath}_{\mathrm{DIM},\ell}\bar{\tau}_L^{\mathtt{DIM}}(\ell)^2\mathbb{1}(\hat{\imath}_{\mathrm{DIM},\ell} > n/2) + (n - \hat{\imath}_{\mathrm{DIM},\ell})\bar{\tau}_R^{\mathtt{DIM}}(\ell)^2\mathbb{1}(\hat{\imath}_{\mathrm{DIM},\ell} \le n/2) = o_{\mathbb{P}}(\log\log(n)).$$

Fix $\epsilon > 0$. Consider the events

$$\begin{aligned}A_\ell^\epsilon &= \Big\{\hat{\imath}_{\mathrm{DIM},\ell}\bar{\tau}_L^{\mathtt{DIM}}(\ell)^2 + (n - \hat{\imath}_{\mathrm{DIM},\ell})\bar{\tau}_R^{\mathtt{DIM}}(\ell)^2 \ge (2-\epsilon)\log\log(n)\Big\},\\ B_\ell^\epsilon &= \Big\{\hat{\imath}_{\mathrm{DIM},\ell}\bar{\tau}_L^{\mathtt{DIM}}(\ell)^2\mathbb{1}(\hat{\imath}_{\mathrm{DIM},\ell} > n/2) + (n - \hat{\imath}_{\mathrm{DIM},\ell})\bar{\tau}_R^{\mathtt{DIM}}(\ell)^2\mathbb{1}(\hat{\imath}_{\mathrm{DIM},\ell} \le n/2) \le 2\epsilon\log\log(n)\Big\}.\end{aligned}$$

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

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

It then follows from a union bound argument that

$$\liminf_{n\to\infty}\mathbb{P}\Big(\sup_{\mathbf{x}\in\mathscr{X}}|\hat{\tau}_{\mathtt{DIM}}^{\mathtt{NSS}}(\mathbf{x}) - \tau| \ge \sigma n^{-b/2}\sqrt{(2+o(1))\log\log(n)}\Big) \ge \frac{b}{e}.$$

Part 2: Inconsistency for Points near the Boundary

Fix $\mathbf{z}\in\mathscr{X}$ such that $z_\ell \le n^{a-1}$. Since the order statistics $x_{(n^a),\ell} = n^{a-1}(1+o_{\mathbb{P}}(1))$, on the event

$n^a \le \hat{\imath}_{\mathrm{DIM},\ell} \le n^b$, if $z_\ell \le (1+o_{\mathbb{P}}(1))n^{a-1}$, then $z_\ell \le x_{(n^a)} \le x_{(\hat{\imath}_{\mathrm{DIM},\ell}),\ell}$, and on the event $\mathtt{Imblce}_\ell^{\mathtt{DIM}}$,

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

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 $\{n-n^b \le \hat{\imath}_{\mathrm{DIM},\ell} \le n-n^a\}$, we have

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

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 $\hat{\imath}_{\mathtt{DIM}}$ at the k-th split ($k\ge 1$) also satisfies

$$\liminf_{n\to\infty}\mathbb{P}\big(\hat{\imath}_{\mathtt{DIM}} \le n^b\big) = \liminf_{n\to\infty}\mathbb{P}\big(n-n^b \le \hat{\imath}_{\mathtt{DIM}}\big) \ge \frac{b}{2e}.$$

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 $\hat{\tau}_{\mathtt{DIM}}^{\mathtt{NSS}}$ by $\hat{\tau}$, the data-driven partition $\mathscr{D}_{\mathsf{T}}$ by $\mathscr{P}$.

Reduction to least square prediction error.

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

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

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

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

where

$$\hat{A}(\mathbf{x}) = \sum_{\mathsf{t}\in\mathscr{P}} \mathbb{1}(\mathbf{x}\in\mathsf{t})\hat{a}_{\mathsf{t}}, \qquad \hat{B}(\mathbf{x}) = \sum_{\mathsf{t}\in\mathscr{P}} \mathbb{1}(\mathbf{x}\in\mathsf{t})\hat{b}_{\mathsf{t}}.$$

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

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

Hence

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

It follows that

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

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 $|y_i(t)| \le U$, $i = 1, 2, \cdots, n$, $t = 0, 1$, for some $U \ge 0$. Decompose by

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

where

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

The least square representation (SA-50) implies that

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

which implies

$$E_2 \le \alpha+\beta.$$

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

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

is the family of all achievable partitions $\mathscr{P}$ 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],

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

A union bound then gives

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

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

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

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

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

where C is a positive universal constant.

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

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

A union bound gives

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

Using Cauchy-Schwarz inequality,

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

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

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

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

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

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

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

with probability at least n^{-C_2}, where C_1 and C_2 are some positive absolute constants. Sub-exponentianity of $\varepsilon_i(t)$, $1 \le i \le n$, $t = 0,1$, 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 $(\hat{\imath},\hat{\jmath})$ denotes the optimal splitting index and coordinate for the decision stump. Denote $\hat{\tau}^{\texttt{HON}}_{\texttt{DIM}}(\mathbf{x})$ by $\check{\tau}(\mathbf{x})$ for simplicity. We use $(y_i,\mathbf{x}_i^\top)_{i=1}^M$ to denote $\mathscr{D}_{\texttt{HON},1}$, which we used to construct the causal tree. Denote by $(\hat{\imath},\hat{\jmath})$ the splitting index and coordinate at the K_n-th step, based on $\mathscr{D}_{\texttt{HON},1}$.

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

$$\begin{aligned}\sup_{\mathbf{x}\in\mathscr{X}}|\check{\tau}(\mathbf{x})-\tau| &\ge |\check{\tau}(\mathbf{0})-\tau| \\ &= \left|\frac{\sum_{i=1}^N \tilde{d}_i\tilde{\varepsilon}_i(1)\mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \le x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}})}{\sum_{i=1}^N \tilde{d}_i\mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \le x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}})} - \frac{\sum_{i=1}^N (1-\tilde{d}_i)\tilde{\varepsilon}_i(0)\mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \le x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}})}{\sum_{i=1}^N (1-\tilde{d}_i)\mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \le x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}})}\right|.\end{aligned}$$

Since $(\tilde{\varepsilon}_i(0), \tilde{\varepsilon}_i(1)) \perp\!\!\!\perp \tilde{\mathbf{x}}_i$, condition on $\hat{\imath}, \hat{\jmath}$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_n)$, we have

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

where

$$\tilde{\imath} = \sum_{i=1}^N \mathbb{1}(\tilde{x}_{i,\hat{\jmath}} \le x_{\pi_{\hat{\jmath}}(\hat{\imath}),\hat{\jmath}}).$$

Call

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

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

$$\begin{aligned}
\mathbb{E}[Z \mid \boldsymbol{D}, \tilde{\imath}] &= \mathbb{E}\Big[\Big| \sum_i w_i \varepsilon_i \Big| \;\Big|\; \boldsymbol{D}, \tilde{\imath} \Big] \\
&\ge c_{\mathrm{MZ}}\, \mathbb{E}\Big[\Big(\sum_i w_i^2 \varepsilon_i^2 \Big)^{1/2} \;\Big|\; \boldsymbol{D}, \tilde{\imath} \Big] \\
&\ge c_{\mathrm{MZ}}\, \mathbb{E}\Big[\Big(\sum_{i=1}^{\tilde{\imath}} \frac{d_i}{n_1^2} \varepsilon_i(1)^2 + \sum_{i=1}^{\tilde{\imath}} \frac{1-d_i}{n_0^2} \varepsilon_i(0)^2 \Big)^{1/2} \;\Big|\; \boldsymbol{D}, \tilde{\imath} \Big] \\
&\ge \frac{c_{\mathrm{MZ}}}{n_0+n_1} \sum_{i=1}^{\tilde{\imath}} \Big(d_i \frac{\sqrt{n_0+n_1}}{n_1} \sqrt{\mathbb{E}[\varepsilon_i(1)^2]} + (1-d_i) \frac{\sqrt{n_0+n_1}}{n_0} \sqrt{\mathbb{E}[\varepsilon_i(0)^2]} \Big) \\
&= c_{\mathrm{MZ}} \frac{1}{\sqrt{\tilde{\imath}}} \Big(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]} \Big).
\end{aligned}$$

Moreover, Assumption SA–2 implies that

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

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

$$\begin{aligned}
\mathbb{P}\big(Z \ge \theta \mathbb{E}[Z \mid \boldsymbol{D}, \tilde{\imath}] \mid \boldsymbol{D}, \tilde{\imath} \big) &\ge (1-\theta^2) \frac{\mathbb{E}[Z \mid \boldsymbol{D}, \tilde{\imath}]^2}{\mathbb{E}[Z^2 \mid \boldsymbol{D}, \tilde{\imath}]} \\
&\ge C(1-\theta^2) \frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}} \frac{n_0 n_1}{\tilde{\imath}^2}. \qquad \text{(SA-54)}
\end{aligned}$$

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

$$\mathbb{P}\big(Z \ge \theta \mathbb{E}[Z \mid \boldsymbol{D}, \tilde{\imath}] \mid \tilde{\imath} \big) \ge C(1-\theta^2) \frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}} (\xi - \xi^2)(1 - \frac{1}{\tilde{\imath}}) \mathbb{1}(\tilde{\imath} > 0).$$

We claim that whenever $s_n \le \hat{\iota} \le n - s_n$,

$$\mathbb{E}\Big[(1 - \frac{1}{\tilde{\iota}})\mathbb{1}(\tilde{\iota} > 0)|\hat{\iota}\Big] = 1 + o_{\mathbb{P}}(1). \tag{SA-55}$$

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

$$\mathbb{P}\bigg(Z \ge \theta c_{\mathrm{MZ}} \frac{1}{\sqrt{\tilde{\iota}}}\Big(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\Big)\bigg|\hat{\iota}\bigg) \ge C(1-\theta^2)\frac{\min\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}{\max\{\mathbb{V}[\varepsilon_i(0)], \mathbb{V}[\varepsilon_i(1)]\}}(\xi - \xi^2) + o_{\mathbb{P}}(1).$$

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

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

Then by Theorem SA-19, we have

$$\mathbb{P}\bigg(Z \ge \frac{1}{2}c_{\mathrm{MZ}} \frac{1}{\sqrt{\tilde{\iota}}}\Big(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\Big), \hat{\iota} \le n^b\bigg) \ge \mathtt{C}\frac{b}{2e} + o_{\mathbb{P}}(1).$$

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

$$\liminf_{n\to\infty} \mathbb{P}\bigg(Z \ge \frac{1}{4}c_{\mathrm{MZ}} \frac{1}{\sqrt{n^b}}\Big(\sqrt{\mathbb{V}[\varepsilon(1)]} + \sqrt{\mathbb{V}[\varepsilon(0)]}\Big)\bigg) \ge \mathtt{C}\frac{b}{2e}.$$

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

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

where for large enough n,

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

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

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

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

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

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

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

SA-4.25 Proof of Theorem SA-24

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

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

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 $\mathscr{D}_\tau$ with effective sample size n_τ.

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

$$N(\varepsilon U, \mathscr{H}_{n_{\mathsf{T}}}[\mathscr{P}], \|\cdot\|_{P_{X^{n_\tau}}}) \le \left(\frac{2}{\varepsilon}\right)^{2^K}, \qquad \varepsilon \in (0,1),$$

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 $\mathscr{H}_{n_{\mathsf{T}}}$ with covering number

$$N(\varepsilon U, \mathscr{H}_{n_{\mathsf{T}}}, \|\cdot\|_{P_{X^{n_\tau}}}) \le 2(n_{\mathsf{T}} p)^{2^K} \left(\frac{417 e U^2}{\beta}\right)^{2^{K+1}}, \qquad \varepsilon \in (0,1).$$

This improvement of covering number due to honesty means we can replace a $\log(n_\tau p)$-penalty in the result of Theorem SA-22 by $\log(n_\tau)$. Now uncondition over $\mathscr{D}_{\mathsf{T}}$ and using the fact that $\rho^{-1} \le n_\tau / n_{\mathsf{T}} \le \rho$, 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 $\hat{\tau}^{\mathscr{X}}_{\texttt{DIM}}(\mathbf{x}; K)$ by $\tilde{\tau}(T_K)$, and $N = n/(K+1)$ denotes the sample size for each folds in the $\texttt{X}$ sample splitting scheme.

Let $\mathtt{t}_1, \mathtt{t}_2, \cdots, \mathtt{t}_{2^K}$ 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 $N_1, N_2, \cdots, N_{2^K}$ and $m_1, m_2, \cdots, m_{2^K}$ denote the number of observations and the Lebesgue measure of the 2^K leaf nodes, respectively. Note that $\vec{N} = (N_1, \cdots, N_{2^K})$ 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 $\vec{N}$, $m_k \sim \mathsf{Beta}(N_k, N - N_k + 1)$

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

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

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

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

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

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

SA-4.28 Proof of Lemma SA-27

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

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

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

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

Now we consider the randomness induced by $n_0, n_{L,0}, n_{R,0}$. By Theorem A.4.1 in Csörgö and Horváth [1997],

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

which implies

$$\sup_{r_n \le k < n - r_n} \left| \frac{n_{L,0}(k) n_{R,0(k)}}{n_0} - (1-\xi)\frac{k(n-k)}{n} \right| = O_{\mathbb{P}}(\sqrt{r_n \log\log n}).$$

Putting together, triangle inequality implies

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

SA-4.29 Proof of Lemma SA-28

The proof of Lemma SA-18 implies that

$$\max_{1 \le \ell \le p} \max_{1 \le k \le s_n} k \left| (\hat{\mu}_{L,0}(k,\ell) - \hat{\mu}_{R,0}(k,\ell))^2 - (\bar{\mu}_{L,0}(k,\ell) - \bar{\mu}_{R,0}(k,\ell))^2 \right| = O_{\mathbb{P}}(\alpha_n),$$

where $\alpha_n = \rho_n \log\log n + \frac{s_n}{n - s_n} \log\log n$, and

$$\max_{1 \le \ell \le p} \max_{1 \le k \le s_n} k (\bar{\mu}_{L,0}(k,\ell) - \bar{\mu}_{R,0}(k,\ell))^2 = O_{\mathbb{P}}(\rho_n \log\log n).$$

Hence it also follows that

$$\max_{1 \le \ell \le p} \max_{1 \le k \le s_n} k (\hat{\mu}_{L,0}(k,\ell) - \hat{\mu}_{R,0}(k,\ell))^2 = O_{\mathbb{P}}\Big(\frac{s_n}{n - s_n} \log\log n + \alpha_n \Big) = O_{\mathbb{P}}(\alpha_n).$$

When $1 \le k \le s_n$, we have $\frac{n_{L,0}(k) n_{R,0}(k)}{n_0} \le n_{L,0}(k) \le k$. 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 $s_n \le n^a \le n^b \le n - s_n$. Consider the event

$$A_n = \{ \exists \ell \in [p] : \max_{k \in [n]} \mathscr{I}^{\mathrm{SSE}}(k,\ell) > \max_{k \notin [s_n, n - s_n]} \mathscr{I}^{\mathrm{SSE}}(k,\ell) \}.$$

Equation (A.4.18) and (A.4.20) imply that for each $\ell \in [p]$,

$$\max_{1 \le k \le s_n, n - s_n \le k \le n} \mathscr{I}^{\mathrm{prox}}(k,\ell) = O_{\mathbb{P}}(\rho_n \log\log(n)),$$
$$\max_{s_n \le k \le n - s_n} \mathscr{I}^{\mathrm{prox}}(k,\ell) = 2\log\log(n)(1 + o_{\mathbb{P}}(1)).$$

Hence

$$\max_{1\le k\le s_n, n-s_n\le k\le n} \mathscr{I}^{\text{prox}}(k,\ell) = o_{\mathbb{P}}\Big(\max_{s_n\le k\le n-s_n} \mathscr{I}^{\text{prox}}(k,\ell)\Big), \qquad \ell\in[p],$$

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

$$\max_{1\le k\le s_n, n-s_n\le k\le n} \mathscr{I}^{\text{SSE}}(k,\ell) = o_{\mathbb{P}}\Big(\max_{s_n\le k\le n-s_n} \mathscr{I}^{\text{SSE}}(k,\ell)\Big), \qquad \ell\in[p].$$

Using a union bound, we get $\mathbb{P}(A_n)\to 0$ as $n\to\infty$. Observe that on the event A_n^c, the argmax for $\mathscr{I}^{\text{SSE}}$ should be inside $[s_n, n-s_n]$. Hence

$$\begin{aligned}
&\mathbb{P}\Big(\exists \ell\in[p] : \max_k \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{k,j\neq\ell} \mathscr{I}^{\text{SSE}}(k,j),\ \max_k \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{k\notin[n^a,n^b]} \mathscr{I}^{\text{SSE}}(k,\ell)\Big)\\
&\ge \mathbb{P}\Big(\exists \ell\in[p] : \max_k \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{k,j\neq\ell} \mathscr{I}^{\text{SSE}}(k,j),\ \max_k \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{k\notin[n^a,n^b]} \mathscr{I}^{\text{SSE}}(k,\ell) \text{ and } A_n^c\Big) - \mathbb{P}(A_n)\\
&\ge \mathbb{P}\Big(\exists \ell\in[p] : \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{\substack{j\neq\ell\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j),\\
&\qquad\qquad \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,\ell)\Big) - 2\mathbb{P}(A_n).
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\Big(\exists \ell\in[p] : \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{\substack{j\neq\ell\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j), \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,\ell)\Big)\\
&= p\mathbb{P}\Big(\max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,1) > \max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j), \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,1) > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,1)\Big)\\
&\ge p\sup_{z\in\mathbb{R}} \mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j) < z, \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,1) > z > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,1)\Big)\\
&\ge p\sup_{z\in\mathbb{R}} \mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j) < z, \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,1) < z\Big)\\
&\qquad - p\mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j) < z, \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,1) < z\Big).
\end{aligned}$$

Then using the fact that $\mathscr{I}^{\text{prox}}(k,\ell)$ approximates $\mathscr{I}^{\text{SSE}}(k,\ell)$ from Lemma SA-17, we have

$$\begin{aligned}
&\mathbb{P}\Big(\exists \ell\in[p] : \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{\substack{j\neq\ell\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,j), \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{SSE}}(k,\ell) > \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{SSE}}(k,\ell)\Big)\\
&\ge p\sup_{z\in\mathbb{R}} \mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{prox}}(k,j) < z - v_n, \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{prox}}(k,1) < z - v_n\Big)\\
&\qquad - p\mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\text{prox}}(k,j) < z + v_n, \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\text{prox}}(k,1) < z + v_n\Big),
\end{aligned} \tag{SA-56}$$

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

Step 2: Gaussian approximation of IPW partial sums

Recall that

$$\mathscr{I}^{\text{prox}}(k,\ell) = (1-\xi)\frac{k(n-k)}{n}(\bar{\mu}_{L,0}(k,\ell) - \bar{\mu}_{R,0}(k,\ell))^2 + \xi\frac{k(n-k)}{n}(\bar{\mu}_{L,1}(k,\ell) - \bar{\mu}_{R,1}(k,\ell))^2. \tag{SA-57}$$

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

$$\begin{aligned}\mathbf{C}_i = &\Big(\sqrt{n}\Big(\Big(\sqrt{\frac{n}{k(n-k)}}(\mathbb{1}(\#\pi^\ell(i) \le k) - \frac{k}{n}) : r_n \le k \le n - r_n\Big)^\top : 1 \le \ell \le p\Big)^\top \frac{1-d_i}{1-\xi}\varepsilon_i(0), \\ &\sqrt{n}\Big(\Big(\sqrt{\frac{n}{k(n-k)}}(\mathbb{1}(\#\pi^\ell(i) \le k) - \frac{k}{n}) : r_n \le k \le n - r_n\Big)^\top : 1 \le \ell \le p\Big)^\top \frac{d_i}{\xi}\varepsilon_i(1)\Big)^\top,\end{aligned}$$

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 $\mathscr{J}$ be a subset of $[n] \times [p]$. Consider the class of closed convex sets $\mathscr{A}$ containing sets of the form

$$A = \{\mathbf{u} \in \mathbb{R}^{2np} : (\mathbf{e}_{0,k,\ell}^\top \mathbf{u}, \mathbf{e}_{1,k,\ell}^\top \mathbf{u}) \in B_2(s_{k,\ell}), s_{k,\ell} \in (0,n], (k,\ell) \in \mathscr{J}\}, \tag{SA-58}$$

where $B_2(r)$ denotes the Euclidean ball centered at $\mathbf{0}$ with radius r in $\mathbb{R}^2$. That is, the class $\mathscr{A}$ contains intersections of cylinders $\{\mathbf{u} \in \mathbb{R}^{2np} : \|(\mathbf{e}_{j_1}^\top \mathbf{u}, \mathbf{e}_{j_2}^\top \mathbf{u})\|_2 \le s\}$. 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 $\mathscr{A}$.

For each $A \in \mathscr{A}$, 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 $A \in \mathscr{A}$ of the form (SA-58), take

$$A^m = \{\mathbf{u} \in \mathbb{R}^{2np} : (\mathbf{e}_{0,k,\ell}^\top \mathbf{u}, \mathbf{e}_{1,k,\ell}^\top \mathbf{u}) \in B_2^{\text{in,n}}(s_{k,\ell}), s_{k,\ell} \in (0,n], (k,\ell) \in \mathscr{J}\},$$

and

$$A^{m,\epsilon} = \{\mathbf{u} \in \mathbb{R}^{2np} : (\mathbf{e}_{0,k,\ell}^\top \mathbf{u}, \mathbf{e}_{1,k,\ell}^\top \mathbf{u}) \in B_2^{\text{out,n}}(s_{k,\ell}), s_{k,\ell} \in (0,n], (k,\ell) \in \mathscr{J}\}.$$

Then $A^m \subseteq A \subseteq A^{m,\epsilon}$. Moreover, denote by $\mathscr{V}(A^m)$ 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

$$A^m = \cup_{\mathbf{v} \in \mathscr{V}(A^m)}\{\mathbf{w} \in \mathbb{R}^{2np} : \mathbf{w}^\top \mathbf{v} \le S_A(\mathbf{v})\}, \qquad S_A(\mathbf{v}) = \sup\{\mathbf{w}^\top \mathbf{v} : \mathbf{w} \in A\}.$$

Then we can analogously characterize $A^{m,\epsilon}$ by

$$A^{m,\epsilon} = \cup_{\mathbf{v}\in\mathscr{V}(A^m)}\{\mathbf{w}\in\mathbb{R}^{2np} : \mathbf{w}^\top\mathbf{v} \le S_A(\mathbf{v}) + \epsilon_{\mathbf{v}}\}, \qquad S_A(\mathbf{v}) = \sup\{\mathbf{w}^\top\mathbf{v} : \mathbf{w}\in A\},$$

where $\epsilon_{\mathbf{v}} \le n^{-1}$ for large enough n. This shows our class $\mathscr{A}$ is a subclass of $\mathscr{A}^{\mathrm{si}}(1,3)$ (see [Chernozhukov et al., 2017, Section 3.1]). Now we check its conditions (M.1'), (M.2') and (E.1'). Let $\mathbf{v}\in\mathscr{V}(A^m)$. 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 $(k,\ell)\in\mathscr{J}$, and $v_{0,k,\ell}^2 + v_{1,k,\ell}^2 = 1$. Let $\mathbf{v}\in\mathscr{V}(A^m)$.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\mathbf{v}^\top\mathbf{C}_i|^2] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Bigg[\Bigg(v_{0,k,\ell}\frac{n}{\sqrt{k(n-k)}}(\mathbb{1}(\#\pi^\ell(i)\le k) - \frac{k}{n})\frac{1-d_i}{1-\xi}\varepsilon_i(0) \\
&\qquad\qquad + v_{1,k,\ell}\frac{n}{\sqrt{k(n-k)}}(\mathbb{1}(\#\pi^\ell(i)\le k) - \frac{k}{n})\frac{d_i}{\xi}\varepsilon_i(1)\Bigg)^2\Bigg] \\
&= \frac{1}{n}\left(\frac{n}{\sqrt{k(n-k)}}\right)^2\sum_{i=1}^n\Bigg\{v_{0,k,\ell}^2\mathbb{E}\Bigg[\Bigg((\mathbb{1}(\#\pi^\ell(i)\le k) - \frac{k}{n})\frac{1-d_i}{1-\xi}\varepsilon_i(0)\Bigg)^2\Bigg] \\
&\qquad\qquad + v_{1,k,\ell}^2\mathbb{E}\Bigg[\Bigg((\mathbb{1}(\#\pi^\ell(i)\le k) - \frac{k}{n})\frac{d_i}{\xi}\varepsilon_i(1)\Bigg)^2\Bigg]\Bigg\} \\
&\ge \min\{\mathbb{V}[(1-\xi)^{-1}(1-d_i)\varepsilon_i(0)], \mathbb{V}[\xi^{-1}d_i\varepsilon_i(1)]\},
\end{aligned}$$

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

$$n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{v}^\top\mathbf{C}_i|^3] \le 4n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{0,k,\ell}^\top\mathbf{C}_i|^3] + 4n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{1,k,\ell}^\top\mathbf{C}_i|^3] \lesssim \sqrt{n/r_n},$$

where the last inequality is from the calculation in Equation (SA-21), and this verifies (M.2') for the third moment. Moreover,

$$\begin{aligned}
n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{v}^\top\mathbf{C}_i|^4] &\le 8n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{0,k,\ell}^\top\mathbf{C}_i|^4] + 8n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{1,k,\ell}^\top\mathbf{C}_i|^4] \\
&\le \sqrt{n/r_n}8n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{0,k,\ell}^\top\mathbf{C}_i|^3] + \sqrt{n/r_n}8n^{-1}\sum_{i=1}^n \mathbb{E}[|\mathbf{e}_{1,k,\ell}^\top\mathbf{C}_i|^3] \\
&\lesssim n/r_n.
\end{aligned}$$

The same logic shows that $\mathbb{E}[\exp(|\mathbf{v}^\top\mathbf{C}_i|/(K\sqrt{n/r_n}))] \le 2$, 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 $\boldsymbol{D}_i \sim N(\mathbf{0}, \mathbb{E}[\mathbf{C}_i\mathbf{C}_i^\top])$ such that

$$\sup_{A\in\mathscr{A}} |\mathbb{P}(n^{-1/2}\sum_{i=1}^n \mathbf{C}_i \in A) - \mathbb{P}(n^{-1/2}\sum_{i=1}^n \boldsymbol{D}_i \in A)| \lesssim \left(\frac{\log^7(n)}{r_n}\right)^{1/6}. \tag{SA-59}$$

Step 4: Gaussian-to-Gaussian Approximation

Observe that for any $k_1, k_2 \in [n], \ell_1, \ell_2 \in [p]$, we have $\mathrm{Cov}[\mathbf{e}_{0,k_1,\ell_1}^\top\mathbf{C}_i, \mathbf{e}_{1,k_2,\ell_2}^\top\mathbf{C}_i] = 0$. The same calculation

as *Multivariate Case Step 2* for the proof of Theorem SA-1 implies we can replace $\boldsymbol{D}_i$ by another mean-zero Gaussian random vector $\mathbf{Z}_i$ such that

$$\mathrm{Cov}[\mathbf{e}_{t_1,k_1,\ell_1}^\top \mathbf{Z}_i, \mathbf{e}_{t_2,k_2,\ell_2}^\top \mathbf{Z}_i] = \begin{cases} \mathrm{Cov}[\mathbf{e}_{t_1,k_1,\ell_1}^\top \boldsymbol{D}_i, \mathbf{e}_{t_2,k_2,\ell_2}^\top \boldsymbol{D}_i], & \text{if } \ell_1 = \ell_2, \\ 0, & \text{otherwise.} \end{cases}$$

We want to show $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i$ is close to $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{D}_i$, measured by the probability of taking value in sets from $\mathscr{A}$ 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 $A \in \mathscr{A}$, for a reason given in [Chernozhukov et al., 2017, proof of Proposition 3.1]. Moreover, $\mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i \in A^m) = \mathbb{P}((\mathbf{v}^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i))_{\mathbf{v}\in\mathscr{V}(A^m)} \le \mathbf{t})$ for some $\mathbf{t} \in \mathbb{R}^m$. Hence we only need to show

$$\sup_{\mathbf{t}\in\mathbb{R}^m} \left| \mathbb{P}((\mathbf{v}^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i))_{\mathbf{v}\in\mathscr{V}(A^m)} \le \mathbf{t}) - \mathbb{P}((\mathbf{v}^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{D}_i))_{\mathbf{v}\in\mathscr{V}(A^m)} \le \mathbf{t}) \right| = o(1).$$

But the definition of $\mathscr{A}$ in Equation (SA-58) implies for any $A \in \mathscr{A}, \mathbf{v} \in \mathscr{V}(A^m)$, 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

$$\mathrm{Cov}\left[\mathbf{e}_k^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i), \mathbf{e}_j^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i)\right] = \mathrm{Cov}\left[\mathbf{e}_k^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{D}_i), \mathbf{e}_j^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{D}_i)\right] = 0,$$

and hence

$$\min_{\mathbf{z}\in\mathscr{V}(A^m)} \mathbb{V}\left[\mathbf{v}^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i))_{\mathbf{v}\in\mathscr{V}(A^m)}\right] \gtrsim 1.$$

Together with Equation (SA-24), we know

$$\max_{\mathbf{v}_1,\mathbf{v}_2\in\mathscr{V}(A^m)} \left| \mathrm{Cov}\left[\mathbf{v}_1^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i), \mathbf{v}_2^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i)\right] - \mathrm{Cov}\left[\mathbf{v}_1^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{D}_i), \mathbf{v}_2^\top(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{D}_i)\right] \right| = O(r_n^{-1/2}).$$

The Gaussian-to-Gaussian Comparison result [Chernozhuokov et al., 2022, Proposition 2.1] them implies

$$\sup_{A\in\mathscr{A}} |\mathbb{P}(n^{-1/2}\sum_{i=1}^n \mathbf{Z}_i \in A) - \mathbb{P}(n^{-1/2}\sum_{i=1}^n \boldsymbol{D}_i \in A)| = O(\log(n) r_n^{-1/2}). \tag{SA-60}$$

Step 5: Orstein-Uhlenbeck Process Calculations

Now we revisit Equation (SA-56). Consider

$$\mathscr{I}^{\mathrm{Gauss}}(k,\ell) = (1-\xi)\frac{k(n-k)}{n}(\tilde{\mu}_{L,0}(k,\ell) - \tilde{\mu}_{R,0}(k,\ell))^2 + \xi\frac{k(n-k)}{n}(\tilde{\mu}_{L,1}(k,\ell) - \tilde{\mu}_{R,1}(k,\ell))^2,$$

with

$$\tilde{\mu}_{L,0}(k,\ell) = \frac{1}{k}\sum_{i\le k} u_{\pi_\ell(i)}, \qquad \tilde{\mu}_{L,1}(k,\ell) = \frac{1}{k}\sum_{i\le k} v_{\pi_\ell(i)},$$
$$\tilde{\mu}_{R,0}(k,\ell) = \frac{1}{n-k}\sum_{i>k} u_{\pi_\ell(i)}, \qquad \tilde{\mu}_{R,1}(k,\ell) = \frac{1}{n-k}\sum_{i>k} v_{\pi_\ell(i)}.$$

Equations (SA-59) and (SA-60) imply that

$$\begin{aligned}
&\sup_{z\in[-n,n]} \mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\mathrm{prox}}(k,j) < z - v_n, \max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\mathrm{prox}}(k,1) < z - v_n\Big) \\
&\qquad - \mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\mathrm{prox}}(k,j) < z + v_n, \max_{k\in[s_n,n-s_n]} \mathscr{I}^{\mathrm{prox}}(k,1) < z + v_n\Big) \\
&= \sup_{z\in[-n,n]} \mathbb{P}\Big(\max_{k\in[s_n,n-s_n]} \mathscr{I}^{\mathrm{Gauss}}(k,1) < z - v_n\Big)^{p-1} \mathbb{P}\Big(\max_{\substack{k\notin[n^a,n^b]\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\mathrm{Gauss}}(k,1) < z - v_n\Big) \\
&\qquad - \mathbb{P}\Big(\max_{\substack{j\neq 1\\ k\in[s_n,n-s_n]}} \mathscr{I}^{\mathrm{Gauss}}(k,j) < z + v_n\Big)^{p-1} \mathbb{P}\Big(\max_{k\in[s_n,n-s_n]} \mathscr{I}^{\mathrm{Gauss}}(k,1) < z + v_n\Big) + o(1).
\end{aligned}$$

The same argument as [Csörgö 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

$$\Big| \max_{k\in[s_n,n-s_n]} \sqrt{\mathscr{I}^{\mathrm{Gauss}}(k,1)} - \sup_{t\in[s_n/n,1-s_n/n]} \sigma\sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}}\Big| = \epsilon_n,$$
$$\Big| \max_{k\in[s_n,n-s_n]\setminus[n^a,n^b]} \sqrt{\mathscr{I}^{\mathrm{Gauss}}(k,1)} - \sup_{t\in[s_n/n,1-s_n/n]\setminus[n^{1-a},n^{1-b}]} \sigma\sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}}\Big| = \epsilon_n,$$

with $\sigma^2 = \mathbb{V}[\varepsilon_i(0)] = \mathbb{V}[\varepsilon_i(1)]$ and $\epsilon_n = o_{\mathbb{P}}((\log\log n)^{-1/2})$. 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

$$\left\{\left(\frac{B_{n,L}}{\sqrt{t(1-t)}}, \frac{B_{n,R}}{\sqrt{t(1-t)}}\right) : t\in[0,1]\right\} \overset{d}{=} \{(U_L(\log(t/(1-t))), U_R(\log(t/(1-t)))) : t\in[0,1]\}.$$

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

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{t\in[1/n,1-1/n]\setminus[n^{1-a},n^{1-b}]} \sqrt{\frac{B_{n,L}^2}{t(1-t)} + \frac{B_{n,R}^2}{t(1-t)}} \le y\Big) \\
&= \mathbb{P}\Big(\sup_{-\log(n-1)\le t<\log(n^{a-1}/(1-n^{a-1})),\log(n^{b-1}/(1-n^{b-1}))<t\le\log(n-1)} |N(t)| \le y\Big) \\
&= \mathbb{P}\Big(\sup_{0\le t<\log(n^{a-1}(n-1)/(1-n^{a-1})),\log\frac{n^{b-1}(n-1)}{1-n^{b-1}})<t\le 2\log(n-1)} |N(t)| \le y\Big),
\end{aligned}$$

and

$$\mathbb{P}\Big(\sup_{t\in[1/n,1-1/n]\setminus[n^{1-a},n^{1-b}]}\sqrt{\frac{B_{n,L}^2}{t(1-t)}+\frac{B_{n,R}^2}{t(1-t)}}\le y\Big)=\mathbb{P}\Big(\sup_{0\le t<2\log(n-1)}|N(t)|\le y\Big).$$

An expansion based on [Horváth, 1993, Lemma 2.1] (Lemma TODO) gives for any $z\in\mathbb{R}$,

$$\mathbb{P}\Big(\sup_{0\le t<c\log(n)}|N(t)|\le\frac{z+2\log\log(n)+\log\log\log(n)}{\sqrt{2\log\log(n)}}+\epsilon_n\Big)=\exp(-e^{-z+\log(c)})+o(1).$$

Moreover, Gaussian correlation inequality [Latała and Matlak, 2017, Remark 3 (i)] and stationarity of O-U process implies

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{0\le t<\log(\frac{n^{a-1}(n-1)}{1-n^{a-1}}),\,\log\frac{n^{b-1}(n-1)}{1-n^{b-1}})<t\le 2\log(n-1)}|N(t)|<\frac{z+2\log\log(n)+\log\log\log(n)}{\sqrt{2\log\log(n)}}+\epsilon_n\Big)\\
&\ge\mathbb{P}\Big(\sup_{0\le t<\log(\frac{n^{a-1}(n-1)}{1-n^{a-1}})}|N(t)|<\frac{z+2\log\log(n)+\log\log\log(n)}{\sqrt{2\log\log(n)}}+\epsilon_n\Big)\\
&\qquad\cdot\mathbb{P}\Big(\sup_{0<t\le\log(n^{1-b}(n-1)(1-n^{b-1}))}|N(t)|<\frac{z+2\log\log(n)+\log\log\log(n)}{\sqrt{2\log\log(n)}}+\epsilon_n\Big)\\
&=\exp(-2e^{-z+\log(2-(b-a))})+o(1).
\end{aligned}$$

Putting together and choosing z^* that maximizes $z\mapsto\exp(-2e^{-z+\log(2-(b-a))})-\exp(-2e^{-z+\log(c)})$, we can get

$$\begin{aligned}
&\sup_{z\in[-n,n]}\mathbb{P}\Big(\max_{k\in[s_n,n-s_n]}\mathscr{I}^{\text{Gauss}}(k,1)<z-v_n\Big)^{p-1}\mathbb{P}\Big(\max_{\substack{k\notin[n^a,n^b]\\k\in[s_n,n-s_n]}}\mathscr{I}^{\text{Gauss}}(k,1)<z-v_n\Big)\\
&\qquad-\mathbb{P}\Big(\max_{\substack{j\neq 1\\k\in[s_n,n-s_n]}}\mathscr{I}^{\text{Gauss}}(k,j)<z+v_n\Big)^{p-1}\mathbb{P}\Big(\max_{k\in[s_n,n-s_n]}\mathscr{I}^{\text{Gauss}}(k,1)<z+v_n\Big)\\
&\ge\sup_z\exp\Big(-2(p-1)e^{-(z-\log(2))}\Big)\Big(\exp\Big(-2e^{-(z-\log(2-(b-a)))}\Big)-\exp\Big(-2e^{-(z-\log(2))}\Big)\Big)\\
&=\frac{b-a}{2p}\Big(1-\frac{b-a}{2p}\Big)^{\frac{2p}{b-a}-1}\\
&\ge\frac{b-a}{2pe}.
\end{aligned}$$

Symmetry then implies for any $0<a<b<1$ and $\ell\in[p]$, we have

$$\liminf_{n\to\infty}\mathbb{P}\big(n^a\le\hat{\imath}_{\text{SSE}}\le n^b,\hat{\jmath}_{\text{SSE}}=\ell\big)=\liminf_{n\to\infty}\mathbb{P}\big(n-n^b\le\hat{\imath}_{\text{SSE}}\le n-n^a,\hat{\jmath}_{\text{SSE}}=\ell\big)\ge\frac{b-a}{2pe}.$$

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 $\mathtt{X}$ under Assumption SA–2. Since (y_i, d_i)'s are from dataset $\mathscr{D}_\tau$ independent to the dataset $\mathscr{D}_{\mathsf{T}_1}$ to $\mathscr{D}_{\mathsf{T}_K}$ for tree construction, it is easy to check that

$$\mathbb{E}[\hat{\tau}_l^{\mathscr{X}}(\mathbf{x}; K)] = \mathbb{E}[\mathbb{E}[\hat{\tau}_l^{\mathscr{X}}(\mathbf{x}; K)|\mathsf{T}, (\mathbf{x}_i)_{i\in\mathscr{D}_\tau}]] = \tau, \qquad l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}.$$

Next, we consider $\mathtt{HON}$ under Assumption SA–2. Denote by $\mathsf{t}(\mathbf{x})$ the node that contains $\mathbf{x}$, and denote by $n(\mathsf{t})$ the *local sample size* in $\mathscr{D}_\tau$, where $n(\mathsf{t}) = \sum_{i\in\mathscr{D}_\tau} \mathbb{1}(\mathbf{x}_i \in \mathsf{t})$. Then

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_{\mathtt{IPW}}^{\mathtt{HON}}(\mathbf{x}; K)|\mathscr{D}_\mathsf{T}] &= \mathbb{E}[\hat{\tau}_{\mathtt{IPW}}^{\mathtt{HON}}(\mathbf{x}; K)\mathbb{1}(n(\mathsf{t}(\mathbf{x})) > 0)|\mathscr{D}_\mathsf{T}] + 0 \cdot \mathbb{P}(n(\mathsf{t}(\mathbf{x})) = 0|\mathscr{D}_\mathsf{T}) \\
&= \mathbb{E}[\hat{\tau}_{\mathtt{IPW}}^{\mathtt{HON}}(\mathbf{x}; K)|\mathscr{D}_\mathsf{T}, n(\mathsf{t}(\mathbf{x})) > 0]\mathbb{P}(n(\mathsf{t}(\mathbf{x})) > 0|\mathscr{D}_\mathsf{T}) \\
&= \tau\mathbb{P}(n(\mathsf{t}(\mathbf{x})) > 0|\mathscr{D}_\mathsf{T}),
\end{aligned}$$

where in the third line, we have used the fact that $\varepsilon_i(0)$ and $\varepsilon_i(1)$ in $\mathscr{D}_\tau$ are independent to $\mathbf{x}_i$'s in $\mathscr{D}_\tau$ and the whole dataset $\mathscr{D}_{\mathtt{T}}$, with $\mathbb{E}[\varepsilon_i(0)] = \mathbb{E}[\varepsilon_i(1)] = 0$. Unconditioning over $\mathscr{D}_{\mathtt{T}}$, then we get

$$\mathbb{E}[\hat{\tau}^{\mathtt{HON}}_{\mathtt{IPW}}(\mathbf{x}; K)] = \tau \mathbb{P}(n(\mathtt{t}(\mathbf{x})) > 0).$$

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

$$\begin{aligned}\text{DIM}: \quad & \frac{n(\mathtt{t}_{\mathtt{L}})n(\mathtt{t}_{\mathtt{R}})}{n(\mathtt{t})}\Big(\frac{1}{n_1(\mathtt{t}_L)}\sum_{i:\mathbf{x}_i\in\mathtt{t}_L} d_i\varepsilon_i(1) - \frac{1}{n_0(\mathtt{t}_L)}\sum_{i:\mathbf{x}_i\in\mathtt{t}_L}(1-d_i)\varepsilon_i(0) \\ & \qquad - \frac{1}{n_1(\mathtt{t}_R)}\sum_{i:\mathbf{x}_i\in\mathtt{t}_R} d_i\varepsilon_i(1) + \frac{1}{n_0(\mathtt{t}_R)}\sum_{i:\mathbf{x}_i\in\mathtt{t}_R}(1-d_i)\varepsilon_i(0))\Big)^2, \end{aligned} \tag{SA-61}$$

and

$$\mathtt{IPW}: \quad \frac{n(\mathtt{t}_{\mathtt{L}})n(\mathtt{t}_{\mathtt{R}})}{n(\mathtt{t}_L)}\Big(\frac{1}{n(\mathtt{t}_L)}\sum_{i:\mathbf{x}_i\in\mathtt{t}_L}(\frac{d_i}{\xi}\varepsilon_i(1) - \frac{1-d_i}{1-\xi}\varepsilon_i(0)) - \frac{1}{n(\mathtt{t}_R)}\sum_{i:\mathbf{x}_i\in\mathtt{t}_R}(\frac{d_i}{\xi}\varepsilon_i(1) - \frac{1-d_i}{1-\xi}\varepsilon_i(0))\Big)^2, \tag{SA-62}$$

and

$$\begin{aligned}\text{SSE}: \quad & \frac{n_1(\mathtt{t}_{\mathtt{L}})n_1(\mathtt{t}_{\mathtt{R}})}{n_1(\mathtt{t})}\Big(\frac{1}{n_1(\mathtt{t}_{\mathtt{L}})}\sum_{i:\mathbf{x}_i\in\mathtt{t}_{\mathtt{L}}} d_i\varepsilon_i(1) - \frac{1}{n_1(\mathtt{t}_{\mathtt{R}})}\sum_{i:\mathbf{x}_i\in\mathtt{t}_{\mathtt{R}}} d_i\varepsilon_i(1)\Big)^2 \\ & + \frac{n_0(\mathtt{t}_{\mathtt{L}})n_0(\mathtt{t}_{\mathtt{R}})}{n_0(\mathtt{t})}\Big(\frac{1}{n_0(\mathtt{t}_{\mathtt{L}})}\sum_{i:\mathbf{x}_i\in\mathtt{t}_{\mathtt{L}}}(1-d_i)\varepsilon_i(0) - \frac{1}{n_0(\mathtt{t}_{\mathtt{R}})}\sum_{i:\mathbf{x}_i\in\mathtt{t}_{\mathtt{R}}}(1-d_i)\varepsilon_i(0)\Big)^2. \end{aligned} \tag{SA-63}$$

Denote the vector $\boldsymbol{\varepsilon} = (\varepsilon_1(0), \varepsilon_1(1), \cdots, \varepsilon_n(0), \varepsilon_n(1))$. Notice that for all three criteria, for any $\mathbf{d} = (d_1, \cdots, d_n)$ and $\mathtt{t}_L, \mathtt{t}_R$, $\varepsilon = \mathbf{u}$ and $\varepsilon = -\mathbf{u}$ give the same value. Hence condition on $\mathbf{d}$ and the data-driven split region $\hat{\mathtt{t}}_L$ and $\hat{\mathtt{t}}_R$, $\boldsymbol{\varepsilon}$ 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 $\mathtt{t}$ in layer $K-1$ is further partitioned into $\mathtt{t}_L$ and $\mathtt{t}_R$ such that Equations (SA-61), (SA-62) and (SA-63) are maximized. The induction hypothesis is that condition on all leaf $\mathtt{t}$ 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 $\mathtt{t}$, for any $\mathbf{d} = (d_1, \cdots, d_n)$ and $\mathtt{t}_L, \mathtt{t}_R$, $\varepsilon = \mathbf{u}$ and $\varepsilon = -\mathbf{u}$ give the same value. Hence the resulting Kth level $\hat{\mathtt{t}}_L$ and $\hat{\mathtt{t}}_R$ are such that condition on $\mathbf{d}$ and the data-driven split region $\hat{\mathtt{t}}_L$ and $\hat{\mathtt{t}}_R$, $\boldsymbol{\varepsilon}$ 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 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.

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 bessel 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— sk—/bk where k ∈ nr. *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.