# Online Appendix

## Honesty in Causal Forests: When It Helps and When It Hurts

Yanfang Hou and Carlos Fernández-Loría

Hong Kong University of Science and Technology

### Appendix A: Splitting Criterion

**Proposition.** For any fixed candidate split, the splitting criterion in Equation (9) is a consistent estimator of the MSE reduction presented in Equation (8).

*Proof.* The node MSE in Equation (7) is:

$$\text{MSE}(\ell) = \mathbb{E}\left[(\beta(\mathbf{X}) - \tau(\ell))^2 \mid \mathbf{X} \in \ell\right] \tag{40}$$

$$= \mathbb{E}[\beta(\mathbf{X})^2 \mid \mathbf{X} \in \ell] - 2\tau(\ell) \cdot \mathbb{E}[\beta(\mathbf{X}) \mid \mathbf{X} \in \ell] + \tau(\ell)^2 \tag{41}$$

$$= \mathbb{E}[\beta(\mathbf{X})^2 \mid \mathbf{X} \in \ell] - \tau(\ell)^2, \tag{42}$$

where we use the fact that $\tau(\ell) = \mathbb{E}[\beta(\mathbf{X}) \mid \mathbf{X} \in \ell]$.

Now consider a split of a parent node $\ell_0$ into child nodes $\ell_1$ and $\ell_2$. Throughout, $\mathbb{P}(\mathbf{X} \in \ell_i)$ abbreviates $\mathbb{P}(\mathbf{X} \in \ell_i \mid \mathbf{X} \in \ell_0)$, so that $\mathbb{P}(\mathbf{X} \in \ell_1) + \mathbb{P}(\mathbf{X} \in \ell_2) = 1$. The reduction in MSE is:

$$\text{Reduction in MSE} = \text{MSE}(\ell_0) - \mathbb{P}(\mathbf{X} \in \ell_1) \cdot \text{MSE}(\ell_1) - \mathbb{P}(\mathbf{X} \in \ell_2) \cdot \text{MSE}(\ell_2) \tag{43}$$

$$= \mathbb{E}[\beta(\mathbf{X})^2 \mid \mathbf{X} \in \ell_0] - \tau(\ell_0)^2 \tag{44}$$

$$- \mathbb{P}(\mathbf{X} \in \ell_1)\left(\mathbb{E}[\beta(\mathbf{X})^2 \mid \mathbf{X} \in \ell_1] - \tau(\ell_1)^2\right)$$

$$- \mathbb{P}(\mathbf{X} \in \ell_2)\left(\mathbb{E}[\beta(\mathbf{X})^2 \mid \mathbf{X} \in \ell_2] - \tau(\ell_2)^2\right)$$

$$= -\tau(\ell_0)^2 + \mathbb{P}(\mathbf{X} \in \ell_1) \cdot \tau(\ell_1)^2 + \mathbb{P}(\mathbf{X} \in \ell_2) \cdot \tau(\ell_2)^2. \tag{45}$$

Substitute $\tau(\ell_0) = \mathbb{P}(\mathbf{X} \in \ell_1) \cdot \tau(\ell_1) + \mathbb{P}(\mathbf{X} \in \ell_2) \cdot \tau(\ell_2)$ into the expression:

$$\text{Reduction in MSE} = \mathbb{P}(\mathbf{X} \in \ell_1) \cdot \tau(\ell_1)^2 + \mathbb{P}(\mathbf{X} \in \ell_2) \cdot \tau(\ell_2)^2 \tag{46}$$

$$- \left(\mathbb{P}(\mathbf{X} \in \ell_1) \cdot \tau(\ell_1) + \mathbb{P}(\mathbf{X} \in \ell_2) \cdot \tau(\ell_2)\right)^2$$

$$= \mathbb{P}(\mathbf{X} \in \ell_1)(1 - \mathbb{P}(\mathbf{X} \in \ell_1)) \cdot \tau(\ell_1)^2 \tag{47}$$

$$+\mathbb{P}(\mathbf{X}\in\ell_2)(1-\mathbb{P}(\mathbf{X}\in\ell_2))\cdot\tau(\ell_2)^2$$

$$-2\mathbb{P}(\mathbf{X}\in\ell_1)\mathbb{P}(\mathbf{X}\in\ell_2)\cdot\tau(\ell_1)\tau(\ell_2)$$

$$=\mathbb{P}(\mathbf{X}\in\ell_1)\mathbb{P}(\mathbf{X}\in\ell_2)\cdot(\tau(\ell_1)-\tau(\ell_2))^2\,. \tag{48}$$

Under ignorability, $\hat{\tau}(\ell_i;\mathcal{S}_{\text{sp}})$ is consistent for $\tau(\ell_i)$ and $n_{\ell_i}/n_{\ell_0}$ for $\mathbb{P}(\mathbf{X}\in\ell_i\mid\mathbf{X}\in\ell_0)$, so by the law of large numbers and the continuous mapping theorem the sample analog

$$\text{Split criterion}=\frac{n_{\ell_1}n_{\ell_2}}{(n_{\ell_1}+n_{\ell_2})^2}\left(\hat{\tau}(\ell_1;\mathcal{S}_{\text{sp}})-\hat{\tau}(\ell_2;\mathcal{S}_{\text{sp}})\right)^2 \tag{49}$$

converges in probability to the reduction in MSE as $n_{\ell_0}$ grows.

## Appendix B: Proofs for Analytical Example

All subsequent proofs pertain to adaptive estimation (AE). Therefore, for simplicity in notation, we will use $\hat{\beta}$ and $\hat{\ell}$ instead of $\hat{\beta}^{\text{A}}$ and $\hat{\ell}^{\text{A}}$.

### B.1. Common Setup and Notation

Let $X_j\sim\text{Bern}(0.5)$ for $j=1,\cdots,J$. We assume that $X_1$ is informative about treatment effects, while the remaining features are uninformative about the potential outcomes, and that

$$X_1\perp\{X_j\}_{j\neq1}. \tag{50}$$

Treatment is randomly assigned with $\mathbb{P}(T=1)=0.5$. We also assume homoscedasticity:

$$\text{Var}(Y\mid\mathbf{X},T)=\sigma^2. \tag{51}$$

Splitting on $X_j$ yields estimated SPATEs

$$\hat{\tau}_j(0),\hat{\tau}_j(1), \tag{52}$$

defined as the difference in mean outcomes between treated and control units among individuals with $X_j=0$ and $X_j=1$, respectively, using the full training sample $\mathcal{S}_{\text{tr}}$.

Define:

$$\Delta_j:=\hat{\tau}_j(1)-\hat{\tau}_j(0), \tag{53}$$

$$\eta_j:=\hat{\tau}_j(1)+\hat{\tau}_j(0), \tag{54}$$

so that:

$$\hat{\tau}_j(1) = \frac{1}{2}(\eta_j + \Delta_j), \tag{55}$$

$$\hat{\tau}_j(0) = \frac{1}{2}(\eta_j - \Delta_j). \tag{56}$$

Under random treatment assignment, the difference-in-means estimators are unbiased for the corresponding SPATEs. For sufficiently large samples, applying the multivariate Central Limit Theorem jointly to

$$\big(\hat{\tau}_1(0), \hat{\tau}_1(1), \ldots, \hat{\tau}_J(0), \hat{\tau}_J(1)\big)$$

implies that this vector is approximately jointly normal. Because $\eta_j$ and $\Delta_j$ are linear combinations of the SPATE estimators, the vector containing all $\eta_j$'s and $\Delta_j$'s is also approximately jointly normal.

Under this joint-normal approximation, the feature-specific marginal distributions depend on whether $X_j$ is informative. Here, $\sigma_j^2$ denotes the sampling variance of either child-node SPATE estimator for feature $j$.

- **Informative feature** ($j = 1$): $\tau_1(1) = \theta$, $\tau_1(0) = -\theta$, so

$$\hat{\tau}_1(1) \sim \mathcal{N}(\theta, \sigma_1^2), \quad \hat{\tau}_1(0) \sim \mathcal{N}(-\theta, \sigma_1^2), \tag{57}$$

$$\Delta_1 \sim \mathcal{N}(2\theta, 2\sigma_1^2), \quad \eta_1 \sim \mathcal{N}(0, 2\sigma_1^2). \tag{58}$$

- **Uninformative feature** ($j > 1$): $\tau_j(0) = \tau_j(1) = 0$, so

$$\hat{\tau}_j(0), \hat{\tau}_j(1) \sim \mathcal{N}(0, \sigma_j^2), \tag{59}$$

$$\Delta_j, \eta_j \sim \mathcal{N}(0, 2\sigma_j^2), \tag{60}$$

$$\sigma_j^2 > \sigma_1^2 \text{ as a result of ignoring heterogeneity explained by } X_1. \tag{61}$$

We next characterize the dependence among these quantities. For any features $j$ and $l$,

$$\begin{aligned}\mathrm{Cov}(\eta_j, \Delta_l) = \mathrm{Cov}\big(\hat{\tau}_j(1), \hat{\tau}_l(1)\big) + \mathrm{Cov}\big(\hat{\tau}_j(0), \hat{\tau}_l(1)\big) \\ - \mathrm{Cov}\big(\hat{\tau}_j(1), \hat{\tau}_l(0)\big) - \mathrm{Cov}\big(\hat{\tau}_j(0), \hat{\tau}_l(0)\big).\end{aligned} \tag{62}$$

Under the maintained analytical DGP, the covariance terms cancel, yielding

$$\mathrm{Cov}(\eta_j, \Delta_l) = 0, \qquad \forall j, l. \tag{63}$$

Because the vector is approximately jointly normal, zero covariance with every component of $(\Delta_1, \ldots, \Delta_J)$ implies

$$\eta_j \perp (\Delta_1, \ldots, \Delta_J), \qquad \forall j. \tag{64}$$

The same covariance calculation, together with independence between $X_1$ and the uninformative features, gives

$$\mathrm{Cov}(\Delta_1, \Delta_j) = 0, \qquad j > 1. \tag{65}$$

Joint normality then implies

$$\Delta_1 \perp (\Delta_2, \ldots, \Delta_J). \tag{66}$$

The algorithm selects the splitting feature $\hat{\ell}$ to maximize the absolute difference between child node estimates:

$$\hat{\ell} = \arg\max_j |\Delta_j|. \tag{67}$$

With balanced binary features, this is analogous to the splitting rule in Equation (9).

For each candidate feature $k$, define the largest competing contrast as:

$$M_{-k} = \max_{j \neq k} |\Delta_j|. \tag{68}$$

Conditioning on $\hat{\ell} = k$ implies that $|\Delta_k|$ was the largest among all candidate features:

$$\hat{\ell} = k \Leftrightarrow |\Delta_k| > M_{-k}. \tag{69}$$

Because the subsequent derivations mainly concern the informative feature, we write

$$M := M_{-1} = \max_{j>1} |\Delta_j|. \tag{70}$$

By Equation (66), $\Delta_1 \perp M$.

The AE estimate is:

$$\hat{\beta}(\mathbf{x}) = \hat{\tau}_{\hat{\ell}}(x_{\hat{\ell}}). \tag{71}$$

Throughout the appendix, we focus on estimation for an individual $\mathbf{x} = (1, x_2, \ldots, x_J)$ with $\beta(\mathbf{x}) = \theta$.

### B.2. No Estimation Bias for Uninformative Splits

**Proof for Equation** (21),

$$\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} \neq 1] = 0.$$

*Proof.* Fix $k > 1$ and suppose $\hat{\ell} = k$. We first consider the case $x_k = 1$.

$$\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = k] = \mathbb{E}[\hat{\tau}_k(1) \mid \hat{\ell} = k] \tag{72}$$

Using the decomposition in Equation (55), we get:

$$\mathbb{E}[\hat{\tau}_k(1) \mid \hat{\ell} = k] = \frac{1}{2}\mathbb{E}[\eta_k \mid \hat{\ell} = k] + \frac{1}{2}\mathbb{E}[\Delta_k \mid \hat{\ell} = k] \tag{73}$$

Because the event $\{\hat{\ell} = k\}$ is a function of $(\Delta_1, \ldots, \Delta_J)$ alone and $\eta_k$ is independent of this vector,

$$\mathbb{E}[\eta_k \mid \hat{\ell} = k] = \mathbb{E}[\eta_k] = 0 \tag{74}$$

Let $\mathbf{\Delta}_{-1} = (\Delta_2, \ldots, \Delta_J)$. Because $\mathbf{\Delta}_{-1}$ is a centered Gaussian vector,

$$\mathbf{\Delta}_{-1} \overset{d}{=} -\mathbf{\Delta}_{-1}, \tag{75}$$

regardless of the correlations among its components. Moreover, because $\Delta_1 \perp \mathbf{\Delta}_{-1}$,

$$(\Delta_1, \mathbf{\Delta}_{-1}) \overset{d}{=} (\Delta_1, -\mathbf{\Delta}_{-1}). \tag{76}$$

Recall that $\{\hat{\ell} = k\} = \{|\Delta_k| > M_{-k}\}$. Because this event depends only on the absolute values of the $\Delta_j$'s, replacing $\mathbf{\Delta}_{-1}$ with $-\mathbf{\Delta}_{-1}$ leaves the event unchanged while changing the sign of $\Delta_k$. By the distributional symmetry,

$$\mathbb{E}\left[\Delta_k \mathbf{1}\{\hat{\ell} = k\}\right] = \mathbb{E}\left[-\Delta_k \mathbf{1}\{\hat{\ell} = k\}\right] = -\mathbb{E}\left[\Delta_k \mathbf{1}\{\hat{\ell} = k\}\right]. \tag{77}$$

Therefore,

$$\mathbb{E}\left[\Delta_k \mathbf{1}\{\hat{\ell} = k\}\right] = 0. \tag{78}$$

Dividing by $\mathbb{P}(\hat{\ell} = k) > 0$ gives

$$\mathbb{E}[\Delta_k \mid \hat{\ell} = k] = 0. \tag{79}$$

Combining the two terms gives $\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = k] = 0$ when $x_k = 1$. By the decomposition in Equation (56), the same argument applies when $x_k = 0$.

Thus, $\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = k] = 0$ for either value of $x_k$ and every $k > 1$, implying $\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} \neq 1] = 0$.

### B.3. Upward Estimation Bias for Informative Splits

**Proof for Equation** (22),

$$\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = 1] > \theta.$$

*Proof.* If $\hat{\ell} = 1$, the estimate for an individual with $x_1 = 1$ is $\hat{\beta}(\mathbf{x}) = \hat{\tau}_1(1)$. Using Equations (55) and (69) we get:

$$\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = 1] = \frac{1}{2}\mathbb{E}\left[\eta_1 + \Delta_1 \mid |\Delta_1| > M\right] \tag{80}$$

Because $\eta_1$ is independent of $(\Delta_1, \cdots, \Delta_J)$,

$$\mathbb{E}\left[\eta_1 \mid |\Delta_1| > M\right] = \mathbb{E}[\eta_1] = \mathbb{E}[\hat{\tau}_1(1)] + \mathbb{E}[\hat{\tau}_1(0)] = \theta + (-\theta) = 0, \tag{81}$$

so:

$$\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = 1] = \frac{1}{2}\mathbb{E}\left[\Delta_1 \mid |\Delta_1| > M\right]. \tag{82}$$

Let $\Delta_1 = 2\theta + \epsilon_1$, where $\epsilon_1 \sim \mathcal{N}(0, 2\sigma_1^2)$. Then:

$$\mathbb{E}[\Delta_1 \mid \hat{\ell} = 1] = \mathbb{E}[2\theta + \epsilon_1 \mid |\Delta_1| > M] = 2\theta + \mathbb{E}[\epsilon_1 \mid \epsilon_1 > M - 2\theta \text{ or } \epsilon_1 < -M - 2\theta]. \tag{83}$$

To prove the inequality, it suffices to show:

$$\mathbb{E}[\epsilon_1 \mid \epsilon_1 > M - 2\theta \text{ or } \epsilon_1 < -M - 2\theta] > 0. \tag{84}$$

We can write:

$$\mathbb{E}[\epsilon_1 \mid \epsilon_1 > M - 2\theta \text{ or } \epsilon_1 < -M - 2\theta] = \frac{\mathbb{E}[\epsilon_1 \cdot \mathbf{1}_{\epsilon_1 > M-2\theta \text{ or } \epsilon_1 < -M-2\theta}]}{\mathbb{P}(\epsilon_1 > M - 2\theta \text{ or } \epsilon_1 < -M - 2\theta)}. \tag{85}$$

The denominator is positive, so we analyze the numerator. Using the law of total expectation:

$$\mathbb{E}[\epsilon_1 \cdot \mathbf{1}_{\epsilon_1 > M-2\theta \text{ or } \epsilon_1 < -M-2\theta}] = \mathbb{E}_M\left[\int_{M-2\theta}^{\infty} t f(t)dt + \int_{-\infty}^{-M-2\theta} t f(t)dt\right], \tag{86}$$

where $f$ is the PDF of $\epsilon_1$. By the symmetry around 0 of the normal distribution, we have:

$$\int_{M-2\theta}^{\infty} t f(t)dt + \int_{-\infty}^{-M-2\theta} t f(t)dt = \int_{|M-2\theta|}^{M+2\theta} t f(t)dt. \tag{87}$$

Therefore, the numerator in Equation (85) is positive, and we conclude:

$$\mathbb{E}[\Delta_1 \mid \hat{\ell} = 1] > 2\theta, \quad \text{and so} \quad \mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = 1] > \theta. \tag{88}$$

### B.4. AE Has Higher Probability of Selecting the Informative Feature

**Proof for Equation** (23),

$$\mathbb{P}(\hat{\ell}^{\mathrm{A}} = 1) \geq \mathbb{P}(\hat{\ell}^{\mathrm{H}} = 1).$$

Recall that

$$\mathbb{P}(\hat{\ell} = 1) = \mathbb{P}(|\Delta_1| > M).$$

We show that this probability increases with the sample size $n$, making AE more likely to identify the informative feature. This is because AE has lower standard errors $\sigma_j$ than HE, which uses only half the data for splitting.

*Proof.* Let $\lambda = \sqrt{n}$. Given that standard errors $\sigma_j$ scale with $1/\sqrt{n}$, we write:

$$\sqrt{2}\sigma_j = \frac{c_j}{\lambda}, \tag{89}$$

for constants $c_j$. Then:

$$\Delta_1 = 2\theta + \frac{c_1 Z_1}{\lambda}, \quad Z_1 \sim \mathcal{N}(0,1), \tag{90}$$

$$\Delta_j = \frac{c_j Z_j}{\lambda}, \quad j > 1, \quad Z_j \sim \mathcal{N}(0,1). \tag{91}$$

It follows that

$$M = \max_{j>1} |\Delta_j| = \frac{1}{\lambda} \max_{j>1} |c_j Z_j| = \frac{V}{\lambda}, \quad \text{where } V = \max_{j>1} |c_j Z_j|. \tag{92}$$

Thus, the probability of selecting the informative feature is:

$$\mathbb{P}(\hat{\ell} = 1) = \mathbb{P}\left(\left|2\theta + \frac{c_1 Z_1}{\lambda}\right| > \frac{V}{\lambda}\right) = \mathbb{P}\left(|2\theta\lambda + c_1 Z_1| > V\right). \tag{93}$$

Taking the expectation over $V$, we get:

$$\mathbb{P}(\hat{\ell} = 1) = \mathbb{E}_V\left[\mathbb{P}(|2\theta\lambda + c_1 Z_1| > V \mid V)\right]. \tag{94}$$

It suffices to show that for each $v > 0$, the conditional probability $\mathbb{P}(|2\theta\lambda + c_1 Z_1| > v)$ increases with $\lambda$. Define this probability as $g(\lambda)$:

$$g(\lambda) = \mathbb{P}(|2\theta\lambda + c_1 Z_1| > v) \tag{95}$$

$$= \mathbb{P}(2\theta\lambda + c_1 Z_1 > v) + \mathbb{P}(2\theta\lambda + c_1 Z_1 < -v) \tag{96}$$

$$= 1 - \Phi\left(\frac{v - 2\theta\lambda}{c_1}\right) + \Phi\left(\frac{-v - 2\theta\lambda}{c_1}\right) \tag{97}$$

and compute its derivative:

$$g'(\lambda) = \frac{2\theta}{c_1} \cdot \left[\phi\left(\frac{v - 2\theta\lambda}{c_1}\right) - \phi\left(\frac{-v - 2\theta\lambda}{c_1}\right)\right] \tag{98}$$

$$= \frac{2\theta}{c_1} \cdot \left[\phi\left(\frac{v - 2\theta\lambda}{c_1}\right) - \phi\left(\frac{v + 2\theta\lambda}{c_1}\right)\right] \tag{99}$$

where $\phi$ is the standard normal PDF. Because $\phi$ is symmetric and strictly decreasing on $(0, \infty)$, for any $v > 0$, we have:

$$\left|\frac{v - 2\theta\lambda}{c_1}\right| < \frac{v + 2\theta\lambda}{c_1}, \quad \text{so } \phi\left(\frac{v - 2\theta\lambda}{c_1}\right) > \phi\left(\frac{v + 2\theta\lambda}{c_1}\right) \tag{100}$$

so $g'(\lambda) > 0$. Thus, $\mathbb{P}(|2\theta\lambda + c_1 Z_1| > v)$ increases with $\lambda$, and therefore $\mathbb{P}(\hat{\ell} = 1)$ increases with $n$.

### B.5. Adaptive Estimation Has Lower Bias

**Proof for Equation** (24),

$$\left|\mathbb{E}[\hat{\beta}^{\mathrm{A}}(\mathbf{x})] - \beta(\mathbf{x})\right| \leq \left|\mathbb{E}[\hat{\beta}^{\mathrm{H}}(\mathbf{x})] - \beta(\mathbf{x})\right|.$$

We already established that

$$\mathbb{E}[\hat{\beta}^{\mathrm{A}}] > \mathbb{E}[\hat{\beta}^{\mathrm{H}}]$$

because AE overestimates the SPATE when it splits on the informative feature and is unbiased when it splits on a non-informative feature. It therefore suffices to show that

$$\mathbb{E}[\hat{\beta}^{\mathrm{A}}(\mathbf{x})] < \theta, \tag{101}$$

which we establish next.

*Proof.* To simplify the notation, let $\hat{\beta} = \hat{\beta}^{\mathrm{A}}$. By the law of total expectation,

$$\mathbb{E}[\hat{\beta}(\mathbf{x})] = \mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = 1] \cdot \mathbb{P}(\hat{\ell} = 1) + \mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} \neq 1] \cdot \mathbb{P}(\hat{\ell} \neq 1). \tag{102}$$

From previous results (Appendices B.2 and B.3),

- $\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} \neq 1] = 0$,
- $\mathbb{E}[\hat{\beta}(\mathbf{x}) \mid \hat{\ell} = 1] = \frac{1}{2}\mathbb{E}[\Delta_1 \mid |\Delta_1| > M]$,

so

$$\mathbb{E}[\hat{\beta}(\mathbf{x})] = \frac{1}{2} \cdot \mathbb{P}(|\Delta_1| > M) \cdot \mathbb{E}[\Delta_1 \mid |\Delta_1| > M]. \tag{103}$$

Now consider the decomposition of the unconditional expectation:

$$\theta = \frac{1}{2}\mathbb{E}[\Delta_1] = \frac{1}{2}\left(\mathbb{P}(|\Delta_1| > M) \cdot \mathbb{E}[\Delta_1 \mid |\Delta_1| > M] + \mathbb{P}(|\Delta_1| \leq M) \cdot \mathbb{E}[\Delta_1 \mid |\Delta_1| \leq M]\right). \tag{104}$$

Subtracting Equation (103) from Equation (104) gives

$$\theta - \mathbb{E}[\hat{\beta}(\mathbf{x})] = \frac{1}{2} \cdot \mathbb{P}(|\Delta_1| \leq M) \cdot \mathbb{E}[\Delta_1 \mid |\Delta_1| \leq M]. \tag{105}$$

We now show that $\mathbb{E}[\Delta_1 \mid |\Delta_1| \leq M] > 0$. Because $\Delta_1 \sim \mathcal{N}(2\theta, 2\sigma_1^2)$, the density $f_{\Delta_1}(t)$ is symmetric around $2\theta > 0$. For any $t > 0$,

$$|t - 2\theta| < |-t - 2\theta| \quad \Rightarrow \quad f_{\Delta_1}(t) > f_{\Delta_1}(-t). \tag{106}$$

Thus,

$$\begin{aligned}\mathbb{E}[\Delta_1 \mid |\Delta_1| \leq z] &= \frac{1}{\mathbb{P}(|\Delta_1| \leq z)} \int_{-z}^{z} t f_{\Delta_1}(t)\, dt \\ &= \frac{1}{\mathbb{P}(|\Delta_1| \leq z)} \int_{0}^{z} t\left[f_{\Delta_1}(t) - f_{\Delta_1}(-t)\right] dt > 0.\end{aligned} \tag{107}$$

This holds pointwise for every $z \geq 0$, so

$$\mathbb{E}[\Delta_1 \mid |\Delta_1| \leq M] > 0, \tag{108}$$

and therefore $\mathbb{E}[\hat{\beta}(\mathbf{x})] < \theta$ holds.

## Appendix C: Analytical Derivations for Variance Analysis

This appendix provides analytical derivations that clarify how honesty impacts each of the sources of tree–tree dependence introduced in Equation (28) of Section 3.3,

$$\mathrm{Cov}\big(\hat{\beta}_b^{\text{tree}}, \hat{\beta}_{b'}^{\text{tree}}\big) = \underbrace{\mathrm{Cov}\big(\tau_b, \tau_{b'}\big)}_{\text{Target coupling}} + \underbrace{\mathrm{Cov}\big(\varepsilon_b, \varepsilon_{b'}\big)}_{\text{Noise overlap}} + \underbrace{\mathrm{Cov}\big(\tau_b, \varepsilon_{b'}\big) + \mathrm{Cov}\big(\varepsilon_b, \tau_{b'}\big)}_{\text{Target–estimation spillover}}.$$

### C.1. Implications of Honesty for Target Coupling

Honesty affects target coupling through its impact on split selection. Target coupling is the covariance between the approximation targets of two trees evaluated at the same individual $\mathbf{x}$,

$$\mathrm{Cov}(\tau_b, \tau_{b'}), \qquad \tau_b = \tau\Big(\hat{\ell}(\mathbf{x}; \mathcal{S}_{\text{sp}}^{(b)})\Big),$$

where randomness arises from sampling variation in the splitting samples $\mathcal{S}_{\text{sp}}^{(b)}$.

Because $\tau_b$ and $\tau_{b'}$ are identically distributed, target coupling can be expressed as

$$\mathrm{Cov}(\tau_b, \tau_{b'}) = \mathrm{Corr}(\tau_b, \tau_{b'}) \cdot \mathrm{Var}(\tau_b). \tag{109}$$

Honesty affects target coupling by reducing the size of the splitting sample, denoted by

$$m = |\mathcal{S}_{\text{sp}}^{(b)}|. \tag{110}$$

Differentiating Equation (109) with respect to $m$ yields

$$\frac{\partial}{\partial m}\mathrm{Cov}(\tau_b, \tau_{b'}) = \mathrm{Corr}(\tau_b, \tau_{b'})\,\frac{\partial}{\partial m}\mathrm{Var}(\tau_b) + \mathrm{Var}(\tau_b)\,\frac{\partial}{\partial m}\mathrm{Corr}(\tau_b, \tau_{b'}). \tag{111}$$

We focus on the first term. The correlation between $\tau_b$ and $\tau_{b'}$ arises from the fact that trees are grown on overlapping subsamples of the same training sample, which induces shared successes and failures in split selection. While the degree of overlap determines the level of this dependence, the size of the splitting sample primarily affects the dispersion of approximation targets across trees. Accordingly, we assume that either (i) the derivatives of the variance and correlation terms with respect to $m$ have the same sign, or (ii) the effect of $m$ on the correlation term is not large enough to overturn the sign of the variance term. Under this assumption, the sign of $\partial\,\mathrm{Cov}(\tau_b, \tau_{b'})/\partial m$ is governed by the behavior of $\partial\,\mathrm{Var}(\tau_b)/\partial m$. The remainder of this appendix therefore analyzes $\mathrm{Var}(\tau_b)$ as a function of the splitting-sample size.

Fix an individual $\mathbf{x}$ with CATE $\beta$ that may be assigned to candidate leaves $\ell_1, \ldots, \ell_J$, each associated with a SPATE $\tau_j = \tau(\ell_j)$. Let

$$p_j(m) = \Pr(\hat{\ell} = \ell_j) \tag{112}$$

denote the probability that $\mathbf{x}$ is assigned to leaf $\ell_j$ when the splitting sample has size $m$. Among the candidate leaves, define the best available approximation target as

$$\tau^\star = \arg\min_{\tau_j} |\tau_j - \beta|. \tag{113}$$

Define the mean approximation target

$$\mu(m) := \mathbb{E}[\tau_b] = \sum_{j=1}^{J} p_j(m)\tau_j, \tag{114}$$

and the variance of approximation targets

$$V(m) := \mathrm{Var}(\tau_b) = \sum_{j=1}^{J} p_j(m)\big(\tau_j - \mu(m)\big)^2. \tag{115}$$

Differentiating with respect to $m$ yields

$$V'(m) = \sum_{j=1}^{J} p_j'(m)\,(\tau_j - \mu(m))^2, \tag{116}$$

where $\sum_{j=1}^{J} p_j(m) = 1$ implies $\sum_{j=1}^{J} p_j'(m) = 0$. Thus, changes in $V(m)$ arise entirely from reallocation of probability mass across candidate approximation targets. Equation (116) shows that the sign of $V'(m)$ depends on where probability mass shifts, weighted by squared deviations from the current mean target $\mu(m)$.

To interpret Equation (116), we define the *correctable approximation bias* as the gap between the expected approximation target and the best available target, $|\mu(m) - \tau^\star|$. The effect of the splitting-sample size $m$ on the variance of the approximation target—and, through it, on target coupling—depends on the magnitude of this bias. When correctable approximation bias is large, increasing $m$ raises the variance of the approximation target; when it is small, increasing $m$ reduces that variance. Because honesty reduces the effective splitting-sample size, it therefore dampens target coupling when correctable approximation bias is large and amplifies it when the bias is small. We now explain the mechanism underlying this result.

When correctable approximation bias is large, split selection frequently assigns $\mathbf{x}$ to suboptimal leaves even though substantially better targets exist. Increasing the splitting-sample size improves split selection and shifts probability mass toward those better targets. Because $\mu(m)$ lies near suboptimal values in this regime, the targets receiving additional probability mass are far from the current mean. As a result, the squared deviations $(\tau_j - \mu(m))^2$ are large precisely

for those $j$ with $p'_j(m) > 0$. Equation (116) therefore implies that increasing $m$ can increase $\mathrm{Var}(\tau_b)$: approximation bias falls as split selection improves, but approximation variance rises because probability mass shifts toward better yet more extreme targets.

In contrast, when correctable approximation bias is small, the expected approximation target $\mu(m)$ is already close to $\tau^\star$. Further improvements in split selection then primarily reallocate probability mass among targets that are already close to $\mu(m)$. This may occur either because split selection is highly reliable, or because target selection is intrinsically uninformative for $\mathbf{x}$—for example, when the features provide little leverage to predict effect heterogeneity or when the individual effect lies near the population average. In either case, the targets receiving additional probability mass have similar values to $\mu(m)$, so the squared deviations $(\tau_j - \mu(m))^2$ are small. Equation (116) therefore implies that increasing $m$ reduces $\mathrm{Var}(\tau_b)$ through further concentration.

As a final example, we return to the stylized setting Section 3.2, where $\tau^\star = \tau_1 = \theta$ and $\tau_j = 0$ for all $j > 1$. In this case, split selection reduces to a binary choice between $\tau^\star$ and a common alternative. When $p_1(m) < 1/2$, increasing $m$ shifts probability mass toward $\tau^\star = \theta$, which lies far from the current mean $\mu(m) = p_1(m)\theta$, so Equation (116) implies that $V(m)$ increases. Once $p_1(m) > 1/2$, the mean approximation target lies closer to $\tau^\star$, and further increases in $m$ primarily concentrate probability mass on that same target, causing $V(m)$ to decrease. This behavior explains Figure 2.

### C.2. Implications of Honesty for Target–Estimation Spillover

We analyze next the target–estimation spillover terms $\mathrm{Cov}(\tau_b, \varepsilon_{b'})$ and $\mathrm{Cov}(\varepsilon_b, \tau_{b'})$, which capture dependence between the approximation target of one tree and the estimation error of another. We show that both terms equal zero under forest honesty. Forest honesty implies that

$$\mathbb{E}[\varepsilon_b \mid \hat{\ell}_b] = 0 \quad \text{and} \quad \mathbb{E}[\varepsilon_b \mid \hat{\ell}_b, \hat{\ell}_{b'}] = 0.$$

We have that

$$\mathrm{Cov}(\varepsilon_b, \tau_{b'}) = \mathbb{E}[\tau_{b'}\,\varepsilon_b] - \mathbb{E}[\tau_{b'}]\mathbb{E}[\varepsilon_b]. \tag{117}$$

Applying the law of iterated expectations,

$$\mathbb{E}[\varepsilon_b] = \mathbb{E}\left[\mathbb{E}[\varepsilon_b \mid \hat{\ell}_b]\right] = 0, \tag{118}$$

so it suffices to show that $\mathbb{E}[\tau_{b'}\,\varepsilon_b] = 0$. If we apply the law of iterated expectations again,

$$\mathbb{E}[\tau_{b'}\,\varepsilon_b] = \mathbb{E}\Big[\tau_{b'}\,\mathbb{E}\Big[\varepsilon_b \mid \hat{\ell}_b, \hat{\ell}_{b'}\Big]\Big] = 0, \tag{119}$$

because $\tau_{b'}$ is deterministic given $\hat{\ell}_{b'}$. Therefore

$$\mathrm{Cov}\big(\varepsilon_b, \tau_{b'}\big) = 0. \tag{120}$$

An identical argument yields

$$\mathrm{Cov}\big(\tau_b, \varepsilon_{b'}\big) = 0. \tag{121}$$

### C.3. Implications of Honesty for Noise Overlap

Finally, we analyze how honesty affects noise overlap, decomposed as

$$\mathrm{Cov}(\varepsilon_b, \varepsilon_{b'}) = \mathbb{E}\big[\mathrm{Cov}(\varepsilon_b, \varepsilon_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'})\big] + \mathrm{Cov}\Big(\mathbb{E}[\varepsilon_b \mid \hat{\ell}_b, \hat{\ell}_{b'}], \mathbb{E}[\varepsilon_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}]\Big).$$

Under forest honesty, the second term vanishes because $\mathbb{E}[\varepsilon_b \mid \hat{\ell}_b, \hat{\ell}_{b'}] = 0$ for all $b \neq b'$. This appendix therefore focuses on the remaining term.

Recall from Equation (27) that each tree prediction can be written as

$$\hat{\beta}_b^{\text{tree}} = \tau_b + \varepsilon_b, \qquad \varepsilon_b = \hat{\tau}_b - \tau_b,$$

where $\tau_b = \tau(\hat{\ell}_b)$ is the SPATE associated with leaf $\hat{\ell}_b$ and $\hat{\tau}_b$ is its estimate based on the estimation subsample $\mathcal{S}_{\text{es}}^{(b)}$. Because $\tau_b$ and $\tau_{b'}$ are fixed given $(\hat{\ell}_b, \hat{\ell}_{b'})$,

$$\mathrm{Cov}(\varepsilon_b, \varepsilon_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}) = \mathrm{Cov}(\hat{\tau}_b, \hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}). \tag{122}$$

For analytical convenience, we assume SPATEs are estimated as sample means of unbiased, noisy signals of the CATE:

$$\hat{\tau}(\ell; \mathcal{S}_{\text{es}}) = \frac{1}{|\mathcal{I}(\ell; \mathcal{S}_{\text{es}})|} \sum_{i \in \mathcal{I}(\ell; \mathcal{S}_{\text{es}})} Z_i, \tag{123}$$

where $\mathcal{I}(\ell; \mathcal{S}_{\text{es}})$ denotes the set of estimation-sample observations falling in leaf $\ell$. In a randomized experiment, one convenient choice for the signal is the transformed outcome

$$Z = \frac{TY}{\omega(\mathbf{X})} - \frac{(1-T)Y}{1-\omega(\mathbf{X})}, \tag{124}$$

which satisfies $\mathbb{E}[Z \mid \mathbf{X}] = \beta(\mathbf{X})$. The specific form of $Z$ is not important; all arguments below rely only on $Z_i$ being i.i.d. with mean $\tau(\ell)$ within each leaf.

Under this representation, the estimator for tree $b$ is

$$\hat{\tau}_b = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} Z_i, \qquad \mathcal{I}_b = \mathcal{I}(\hat{\ell}_b; \mathcal{S}_{\text{es}}^{(b)}), \tag{125}$$

with $n_b = |\mathcal{I}_b|$. In addition, let

$$p_b = \mathbb{P}(X_i \in \hat{\ell}_b), \quad p_{b'} = \mathbb{P}(X_i \in \hat{\ell}_{b'}), \quad p_{bb'} = \mathbb{P}(X_i \in \hat{\ell}_b \cap \hat{\ell}_{b'}). \tag{126}$$

The training sample contains $N$ observations, and for each tree $b$ the estimation sample $\mathcal{S}_{\text{es}}^{(b)}$ is drawn by subsampling at rate $p_{\text{es}}$, independently of leaf assignment due to honesty.

To evaluate the covariance of the SPATE estimators, we must account for two sources of randomness: (i) noise in the signals $\{Z_i\}$ and (ii) random estimation subsamples $(\mathcal{I}_b, \mathcal{I}_{b'})$. We therefore apply the law of total covariance over the subsampling mechanism:

$$\begin{aligned}\operatorname{Cov}(\hat{\tau}_b, \hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}) = \mathbb{E}\Big[\operatorname{Cov}(\hat{\tau}_b, \hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}) \mid \hat{\ell}_b, \hat{\ell}_{b'}\Big] \\ + \operatorname{Cov}\Big(\mathbb{E}[\hat{\tau}_b \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}], \mathbb{E}[\hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}] \mid \hat{\ell}_b, \hat{\ell}_{b'}\Big).\end{aligned} \tag{127}$$

Because estimation subsamples are drawn uniformly at random under honesty, inclusion into the estimation sample is independent of the outcome signal $Z$, so

$$\mathbb{E}[\hat{\tau}_b \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}] = \tau_b, \qquad \mathbb{E}[\hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}] = \tau_{b'}. \tag{128}$$

Hence the second term in Equation (127) is zero. Using the sample-mean representation of $\hat{\tau}_b$,

$$\begin{aligned}\operatorname{Cov}(\hat{\tau}_b, \hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}) &= \frac{1}{n_b n_{b'}} \sum_{i \in \mathcal{I}_b} \sum_{j \in \mathcal{I}_{b'}} \operatorname{Cov}(Z_i, Z_j \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}) \\ &= \frac{1}{n_b n_{b'}} \sum_{i \in \mathcal{I}_b \cap \mathcal{I}_{b'}} \operatorname{Var}(Z_i \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}) \\ &= \frac{1}{n_b n_{b'}} \sum_{i \in \mathcal{I}_b \cap \mathcal{I}_{b'}} \operatorname{Var}(Z_i \mid X_i \in \hat{\ell}_b \cap \hat{\ell}_{b'}) \\ &= \frac{\sigma^2_{bb'} m_{bb'}}{n_b n_{b'}},\end{aligned} \tag{129}$$

where $m_{bb'} = |\mathcal{I}_b \cap \mathcal{I}_{b'}|$ and $\sigma^2_{bb'} = \operatorname{Var}(Z_i \mid X_i \in \hat{\ell}_b \cap \hat{\ell}_{b'})$. The second line uses the i.i.d. assumption, which implies $\operatorname{Cov}(Z_i, Z_j \mid \hat{\ell}_b, \hat{\ell}_{b'}, \mathcal{I}_b, \mathcal{I}_{b'}) = 0$ for $i \neq j$, so only observations shared across both estimation subsamples contribute to the covariance. The third line follows from independence between the outcome signal and inclusion into the estimation subsample.

An observation contributes to $m_{bb'}$ only if it belongs to both leaf populations and is selected into both estimation subsamples. Conditional on $(\hat{\ell}_b, \hat{\ell}_{b'})$, the counts $(m_{bb'}, n_b, n_{b'})$ are sums of $N$ i.i.d. indicators with means

$$\mathbb{E}[m_{bb'} \mid \hat{\ell}_b, \hat{\ell}_{b'}] = p_{bb'} p_{\text{es}}^2 N, \quad \mathbb{E}[n_b \mid \hat{\ell}_b] = p_b p_{\text{es}} N, \quad \mathbb{E}[n_{b'} \mid \hat{\ell}_{b'}] = p_{b'} p_{\text{es}} N. \tag{130}$$

Applying the delta method to the smooth map $(x, y, z) \mapsto x/(yz)$ at these means, and using that the empirical proportions $m_{bb'}/N$, $n_b/N$, $n_{b'}/N$ fluctuate around their means at rate $O(N^{-1/2})$ under standard leaf-regularity conditions, gives

$$\mathbb{E}\left[\frac{m_{bb'}}{n_b n_{b'}} \,\middle|\, \hat{\ell}_b, \hat{\ell}_{b'}\right] = \frac{p_{bb'}}{p_b p_{b'} N} + O(N^{-2}), \tag{131}$$

where the subsampling rate $p_{\text{es}}$ cancels between numerator and denominator. Substituting into Equation (129),

$$\text{Cov}(\hat{\tau}_b, \hat{\tau}_{b'} \mid \hat{\ell}_b, \hat{\ell}_{b'}) = \sigma^2_{bb'} \frac{p_{bb'}}{p_b p_{b'} N} + O(N^{-2}). \tag{132}$$

Equation (132) shows that, asymptotically, the covariance depends only on the geometric overlap between the two leaf populations. The cancellation of $p_{\text{es}}$ reflects that honesty's reduction of estimation sample sizes is exactly offset by the corresponding reduction in shared observations, so honesty does not affect asymptotic noise overlap.

## Appendix D: AE–HE Selection Rule

This appendix motivates and analyzes the rule we use to select between AE and HE when both pass the heterogeneity test. We first explain why the selection problem calls for an asymmetric rule that favors HE (D.1), then formalize the rule (D.2), discuss the choice of threshold (D.3), and present a sensitivity analysis (D.4). A final subsection discusses the implementation of the heterogeneity test used as a screening step (D.5).

### D.1. Why an Asymmetric Rule

Our theoretical analysis implies that the selection between AE and HE is a choice between a flexible method and its regularized counterpart. That means which one is better depends on the SNR: AE tends to win when SNR is high, HE when it is low. Since the SNR is unknown, a natural strategy is to compare the two methods empirically and pick whichever performs better on held-out data.

The key observation is that the reliability of this empirical comparison also depends on the SNR. Cross-validation estimates based on transformed outcomes are reliable when the SNR is high, but not so much when it is low. So when data can reliably adjudicate the choice (high SNR), it points to AE; when it cannot (low SNR), theory points to HE.

This suggests an asymmetric rule that gets the best of both worlds: treat HE as the default, and switch to AE only when the empirical evidence in its favor is strong relative to the uncertainty in the comparison. When the signal is strong, AE's advantage is large and easily clears the evidence bar; when the signal is weak, the bar is harder to clear and the selector falls back on the method that theory favors in that regime. Relying on a noisy comparison alone—selecting whichever method posts the lower estimated error—would forgo this protection, in the same way that cross-validation can overfit when uncertainty in the validation estimates is ignored (Lei 2020).

### D.2. The Decision Rule

We quantify the uncertainty in the AE–HE comparison using the standard error of the paired error difference. Because both models are evaluated on the same held-out data points, their errors are correlated; taking paired differences accounts for this shared variation and directly targets the uncertainty in the performance contrast. Paired differences are a standard tool in model comparison: Dietterich (1998) compares learning algorithms using paired differences in prediction errors across train/test splits, and Lei (2020) uses the same object to construct confidence sets of competitive models that account for the overfitting induced by CV-based selection.

Formally, let $\bar{\Delta}$ denote the mean paired error difference, defined as HE minus AE, and let $\mathrm{SE}(\bar{\Delta})$ denote its standard error. The rule is:

$$\text{select AE if } \bar{\Delta} > c\,\mathrm{SE}(\bar{\Delta}); \qquad \text{otherwise, select HE,} \tag{133}$$

where $c \geq 0$ is a prespecified threshold governing how much evidence is required before the default is abandoned. At $c = 0$, the rule reduces to selecting the model with the lower estimated MSE. As $c$ increases, AE must show a larger advantage before HE is abandoned, so the rule favors HE more often; as $c \to \infty$, the rule selects HE almost always.

### D.3. Choosing the Threshold

The threshold $c$ formalizes what "sufficiently strong evidence" means, and there is no principled, universally agreed-upon value—just as there is no first-principles justification for the conventional 5% significance level in hypothesis testing. What exist instead are conventions, and two are relevant here.

The first comes from framing the comparison as a hypothesis test. Dietterich (1998) compares learning algorithms this way, using a conventional significance level such as $\alpha = 0.05$ to decide whether to reject the null of no difference. In our setting, a one-sided 5% level corresponds approximately to $c = 1.6$. This threshold demands very strong evidence before selecting AE, which protects against costly false AE selections but sacrifices a large share of the available gains whenever AE is in fact the better choice—as it often is in the ACIC datasets. We quantify this cost in Section D.4.

The second convention comes from the model selection literature: the one-standard-error (1-SE) rule of Breiman et al. (1984). In its conventional form, the rule selects the simplest (most regularized) model whose estimated error is within one standard error of the best model's error (Hastie et al. 2009). This principle has been widely used to select regularization parameters (Breiman et al. 1984, Hastie et al. 2009), is the default behavior offered by the R package `glmnet` (Friedman et al. 2010), and has been adopted across domains: Villanueva et al. (2024) use it to select parsimonious Lasso models for RNA and protein localization, and Windmeijer et al. (2019) apply it in Mendelian

randomization to identify invalid instrumental variables. Applying the 1-SE threshold to paired performance differences also has precedent: Hammond et al. (2024) use it to compare eclipse-mapping models in astronomy, and Piironen et al. (2020) apply the same logic in projection predictive model selection. Our rule is this heuristic applied to the AE–HE choice, with HE playing the role of the more regularized model: setting $c = 1$ selects HE whenever its estimated error is within one standard error of AE's.

We adopt $c = 1$ as our default threshold, without claiming it is optimal. The optimal threshold depends on quantities that are not known in practice, such as the relative cost and probability of the two possible mistakes—overfitting with AE versus underfitting with HE. A different threshold could reasonably be chosen instead. What matters for our purposes is that (i) the rule embeds the correct qualitative asymmetry, and (ii) its performance is robust across thresholds, as we show next in Section D.4.

### D.4. Sensitivity Analysis

Figure 8 presents a sensitivity analysis over $c \in [0, 1.6]$ using the ACIC datasets. As $c$ increases, the selector favors HE more often, and both mean regret and the error rate increase; this reflects that AE is the better choice in most of these datasets. The empirical selector outperforms both pure strategies in mean regret for all thresholds up to roughly $c = 1.2$; beyond that point, it falls behind defaulting to AE. Even so, it outperforms defaulting to HE across the entire range—and since defaulting to HE is the prevailing practice, the choice of $c$ affects how much of AE's potential gains are captured, not whether the procedure improves on the status quo.

Figure 4 of the main text puts this threshold choice in context. When the SNR is high, AE tends to be as good as or better than HE, so a lower threshold—which makes the selector behave more like AE—would perform better in that regime; when the SNR is low, the reverse holds. But this is not actionable advice: the rule exists precisely because the SNR is unknown in practice. The threshold is therefore best understood as a compromise that makes the selection robust across regimes: at $c = 1$, the empirical selector slightly trails AE in the highest-SNR deciles—the price of conservatism—while outperforming it where the signal is weak, tracking the better method across the full range without knowing which regime it is in.

### D.5. Implementation of the Heterogeneity Test

The selection procedure uses the heterogeneity test of Imai and Li (2025) as a screening step: datasets in which neither tuned estimator detects heterogeneity are excluded, since the AE–HE comparison is moot when the signal is too weak for either method to model heterogeneity reliably.

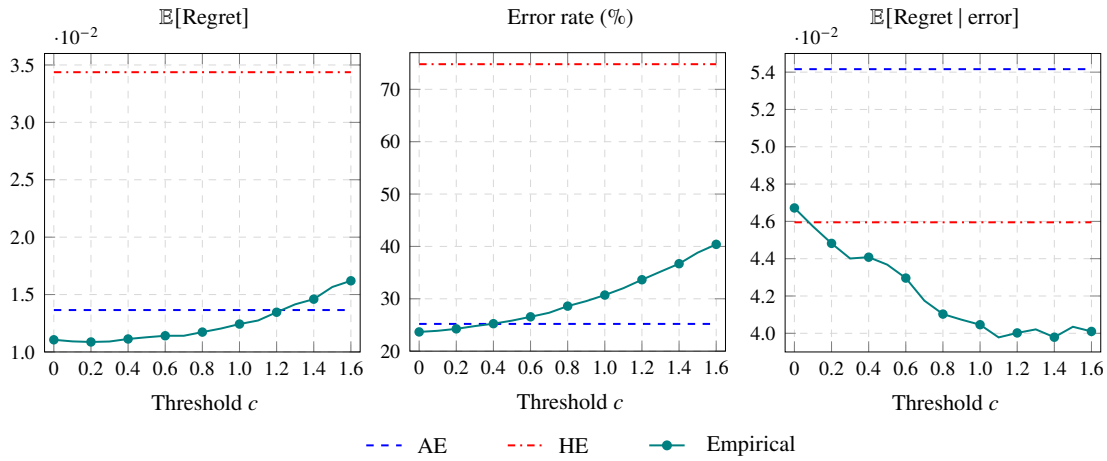


**Figure 8** **Performance of the empirical selector across selection thresholds. As $c$ increases, the selector chooses HE more often. Both mean regret and the error rate increase, while the cost of incorrect selections first decreases and then levels off.**

Our implementation follows the cross-fitting generalization of the test developed by Imai and Li (2025). We create a 5-fold partition of the training sample—distinct from the partition used for hyperparameter tuning—refit each forest on four folds to generate predictions for the held-out fold, and conduct the test on the resulting out-of-fold predictions. No observation is therefore predicted by a model that was trained on it.

One deviation from a fully sample-split protocol is that the observations used for testing were previously used, under a different fold partition, to tune the `min_samples_leaf` hyperparameter. This is a deliberate choice, for two reasons. First, the test serves as a screening device within a model-selection pipeline whose end-to-end performance is what we evaluate, not as a confirmatory inference procedure; if the goal were formal hypothesis testing about heterogeneity, the sample-splitting protocol should be followed as prescribed. Second, reserving a dedicated holdout untouched by tuning would substantially reduce the data available for tuning and for the AE–HE comparison, weakening both estimators in a way that runs counter to our goal of mirroring applied workflows.

The resulting leakage channel is narrow—a single scalar hyperparameter selected from a coarse grid—and it affects AE and HE symmetrically, since both are tuned on the same data and tested under the same partition. As an empirical check, we applied the full procedure to the 200 ACIC datasets with no true effect heterogeneity: the false positive rate was 1.5% for both AE and HE, below the nominal 5% level, indicating that the reuse does not make the test anti-conservative in practice.

## Appendix E: Additional Data Requirements

This appendix describes the procedure used to calculate the additional data requirements reported in Figure 5.

For each dataset, we compare the test MSEs of AE and HE, both trained on the full sample ($N = 4{,}000$). If the difference is not significant according to a two-sided test at the 5% level, both models are assigned an effective sample size of 4,000, and the additional data requirement is zero.

If the difference is significant, we determine the minimum training sample size at which the superior model outperforms the inferior model trained on all 4,000 observations. Holding the inferior model's MSE fixed, we use a binary search over the superior model's training sample size. We initialize the search bounds at $N_{\text{low}} = 0$ and $N_{\text{high}} = 4{,}000$. At each step, we evaluate the candidate size given by the midpoint of the current bounds: we tune and retrain the superior model at this size and evaluate its test MSE. If this MSE is lower than the full-sample inferior model's MSE and the difference is significant, we lower the upper bound to the candidate; otherwise, we raise the lower bound to the candidate. We stop once the bounds are within 250 observations.

The superior model's effective sample size is the final upper bound—the smallest sample size found to yield a statistically significant improvement, up to a resolution of 250 observations. The inferior model's effective sample size remains fixed at 4,000. Denoting the effective sample sizes of HE and AE by $N^H$ and $N^A$, respectively, we calculate the additional-data measure for HE as $(N^H - N^A)/4{,}000$. Positive values indicate that HE requires more data than AE, whereas negative values indicate that AE requires more data than HE.

Figure 5 reports the average additional-data measure across datasets within each SNR decile.

## Appendix F: Bias–Variance Estimation

This appendix describes how we estimate the bias–variance decomposition and its subcomponents.

### F.1. Forest-level Decomposition

The main text presents the bias–variance decomposition at the tree level. For estimation, it is more convenient to rewrite the same decomposition at the forest level. Fix a test point $\mathbf{x}$. Equation (27) writes the prediction from tree $b$ as

$$\hat{\beta}_b^{\text{tree}} = \tau_b + \varepsilon_b,$$

where $\tau_b$ is the approximation target of the selected leaf containing $x$, and $\varepsilon_b$ is the estimation error.

Conditional on the training sample $\mathcal{S}$, we obtain the forest-level representation by averaging this tree-level decomposition over trees,

$$\bar{\tau}(\mathcal{S}) = \mathbb{E}_b[\tau_b \mid \mathcal{S}], \qquad \bar{\varepsilon}(\mathcal{S}) = \mathbb{E}_b[\varepsilon_b \mid \mathcal{S}]. \tag{134}$$

Then the forest prediction is

$$\hat{\beta}(\mathcal{S}) = \mathbb{E}_b[\hat{\beta}_b^{\text{tree}} \mid \mathcal{S}] = \bar{\tau}(\mathcal{S}) + \bar{\varepsilon}(\mathcal{S}). \tag{135}$$

Taking expectation over training samples gives the forest-level bias decomposition:

$$\mathbb{E}_{\mathcal{S}}[\hat{\beta}(\mathcal{S})] - \beta = \underbrace{\mathbb{E}_{\mathcal{S}}[\bar{\tau}(\mathcal{S})] - \beta}_{\text{approximation bias}} + \underbrace{\mathbb{E}_{\mathcal{S}}[\bar{\varepsilon}(\mathcal{S})]}_{\text{estimation bias}}. \tag{136}$$

This is the forest-level analogue of the tree-level decomposition in the main text, Equation (16).

The variance decomposition follows the same representation. For two distinct trees $b \neq b'$, the covariance decomposition in Equation (28) is:

$$\text{Cov}\left(\hat{\beta}_b^{\text{tree}}, \hat{\beta}_{b'}^{\text{tree}}\right) = \text{Cov}(\tau_b, \tau_{b'}) + \text{Cov}(\varepsilon_b, \varepsilon_{b'}) + \text{Cov}(\tau_b, \varepsilon_{b'}) + \text{Cov}(\varepsilon_b, \tau_{b'}).$$

We illustrate the argument using the target coupling term. By the law of total covariance,

$$\text{Cov}(\tau_b, \tau_{b'}) = \text{Cov}_{\mathcal{S}}\left(\mathbb{E}_b[\tau_b \mid \mathcal{S}], \mathbb{E}_{b'}[\tau_{b'} \mid \mathcal{S}]\right) + \mathbb{E}_{\mathcal{S}}\left[\text{Cov}(\tau_b, \tau_{b'} \mid \mathcal{S})\right]. \tag{137}$$

Conditional on $\mathcal{S}$, the two trees are independently generated, so

$$\text{Cov}(\tau_b, \tau_{b'} \mid \mathcal{S}) = 0. \tag{138}$$

Moreover,

$$\mathbb{E}_b[\tau_b \mid \mathcal{S}] = \mathbb{E}_{b'}[\tau_{b'} \mid \mathcal{S}] = \bar{\tau}(\mathcal{S}). \tag{139}$$

Therefore,

$$\text{Cov}(\tau_b, \tau_{b'}) = \text{Var}_{\mathcal{S}}\left(\bar{\tau}(\mathcal{S})\right). \tag{140}$$

Applying the same argument to the remaining terms gives

$$\text{Cov}(\varepsilon_b, \varepsilon_{b'}) = \text{Var}_{\mathcal{S}}\left(\bar{\varepsilon}(\mathcal{S})\right), \tag{141}$$

and, by symmetry,

$$\text{Cov}(\tau_b, \varepsilon_{b'}) = \text{Cov}(\varepsilon_b, \tau_{b'}) = \text{Cov}_{\mathcal{S}}\left(\bar{\tau}(\mathcal{S}), \bar{\varepsilon}(\mathcal{S})\right). \tag{142}$$

Combining these terms yields

$$\text{Cov}\left(\hat{\beta}_b^{\text{tree}}, \hat{\beta}_{b'}^{\text{tree}}\right) = \underbrace{\text{Var}_{\mathcal{S}}\left(\bar{\tau}(\mathcal{S})\right)}_{\text{target coupling}} + \underbrace{\text{Var}_{\mathcal{S}}\left(\bar{\varepsilon}(\mathcal{S})\right)}_{\text{noise overlap}} + \underbrace{2\,\text{Cov}_{\mathcal{S}}\left(\bar{\tau}(\mathcal{S}), \bar{\varepsilon}(\mathcal{S})\right)}_{\text{target–estimation spillover}}. \tag{143}$$

Thus, estimating the variance subcomponents reduces to estimating the forest-level approximation target $\bar{\tau}(\mathcal{S})$ and estimation error $\bar{\varepsilon}(\mathcal{S})$ across repeated training samples.

### F.2. Bias and Variance Estimates

This section describes the estimation procedure for the bias–variance decomposition and its subcomponents.

For each replication $r = 1, \ldots, R$, we draw an independent treatment assignment with treatment probability 0.5 and an independent train/test split of size 4000 : 802. We fit the causal forest on the training sample and generate predictions for each test observation.

After fitting the forest, we freeze the learned tree structure. For a test point $\mathbf{x}_i$, let $\hat{\ell}_b(\mathbf{x}_i)$ denote the leaf containing $\mathbf{x}_i$ in tree $b$. We then use the other test observations as an auxiliary sample, excluding unit $i$ itself. Within each tree, we average the true CATEs of auxiliary observations that fall in the same learned leaf as $\mathbf{x}_i$. The forest-level approximation target is the average of these leaf-level targets across trees:

$$\widehat{\bar{\tau}}_i^{(r)} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{j \in \mathcal{S}_{\text{te}}^{(r)},\, j \neq i} \mathbf{1}\{\mathbf{X}_j \in \hat{\ell}_b(\mathbf{x}_i)\}\beta_j}{\sum_{j \in \mathcal{S}_{\text{te}}^{(r)},\, j \neq i} \mathbf{1}\{\mathbf{X}_j \in \hat{\ell}_b(\mathbf{x}_i)\}}. \tag{144}$$

The estimation error is the difference between the forest prediction and the estimated approximation target:

$$\widehat{\bar{\varepsilon}}_i^{(r)} = \hat{\beta}_i^{(r)} - \widehat{\bar{\tau}}_i^{(r)}. \tag{145}$$

We repeat this procedure $R = 600$ times. Each unit appears in the test set approximately 100 times on average. Let $\mathcal{R}_i$ denote the set of replications in which unit $i$ appears in the test set, and let $R_i = |\mathcal{R}_i|$. All pointwise quantities are computed using only replications in $\mathcal{R}_i$.

Define the empirical means

$$\bar{\tau}_i = \mathbb{E}_{\mathcal{R}_i}\left[\widehat{\bar{\tau}}_i^{(r)}\right], \qquad \bar{\varepsilon}_i = \mathbb{E}_{\mathcal{R}_i}\left[\widehat{\bar{\varepsilon}}_i^{(r)}\right], \qquad \bar{\beta}_i = \mathbb{E}_{\mathcal{R}_i}\left[\hat{\beta}_i^{(r)}\right]. \tag{146}$$

The squared bias is estimated as

$$\widehat{\mathrm{Bias}}_i^2 = (\bar{\beta}_i - \beta_i)^2. \tag{147}$$

It decomposes into squared approximation bias, squared estimation bias, and their interaction:

$$\widehat{\mathrm{Bias}}_i^2 = \underbrace{(\bar{\tau}_i - \beta_i)^2}_{\text{squared approximation bias}} + \underbrace{\bar{\varepsilon}_i^2}_{\text{squared estimation bias}} + \underbrace{2(\bar{\tau}_i - \beta_i)\bar{\varepsilon}_i}_{\text{bias interaction}}. \tag{148}$$

The variance is estimated as

$$\widehat{\mathrm{Var}}_i = \widehat{\mathrm{Var}}_{\mathcal{R}_i}\left(\hat{\beta}_i^{(r)}\right). \tag{149}$$

It decomposes into target coupling, noise overlap, and target–estimation spillover:

$$\widehat{\mathrm{Var}}_i = \underbrace{\widehat{\mathrm{Var}}_{\mathcal{R}_i}\left(\widehat{\bar{\tau}}_i^{(r)}\right)}_{\text{target coupling}} + \underbrace{\widehat{\mathrm{Var}}_{\mathcal{R}_i}\left(\widehat{\bar{\varepsilon}}_i^{(r)}\right)}_{\text{noise overlap}} + \underbrace{2\,\widehat{\mathrm{Cov}}_{\mathcal{R}_i}\left(\widehat{\bar{\tau}}_i^{(r)}, \widehat{\bar{\varepsilon}}_i^{(r)}\right)}_{\text{target–estimation spillover}}. \tag{150}$$

Here, $\mathbb{E}_{\mathcal{R}_i}$, $\widehat{\mathrm{Var}}_{\mathcal{R}_i}$, and $\widehat{\mathrm{Cov}}_{\mathcal{R}_i}$ denote the empirical mean, variance, and covariance over replications $r \in \mathcal{R}_i$.

Together, these estimates provide a pointwise decomposition of prediction error. In the main analysis, we average the pointwise components across units and normalize them by the variance of the CATE over the full dataset.

One caveat applies to the variance subcomponents. Because $\widehat{\tilde{\tau}}_i^{(r)}$ is computed from a finite auxiliary sample, it carries sampling noise of its own. Writing $\widehat{\tilde{\tau}}_i^{(r)} = \bar{\tau}_i^{(r)} + u_i^{(r)}$, we also have that $\widehat{\tilde{\varepsilon}}_i^{(r)} = \bar{\varepsilon}_i^{(r)} - u_i^{(r)}$. If $u$ is uncorrelated with $\bar{\tau}$ and $\bar{\varepsilon}$, the estimated target coupling and noise overlap are each inflated by $\mathrm{Var}(u)$ and the estimated spillover is displaced downward by $2\,\mathrm{Var}(u)$; the estimated total variance is unaffected. The levels of the individual subcomponents should therefore not be read as estimates of the underlying population quantities. This applies in particular to the target–estimation spillover, which is displaced toward negative values.

Comparisons between AE and HE are on firmer ground. Because $\mathrm{Var}(u)$ decreases in the number of auxiliary observations per leaf, it is larger for whichever estimator produces the finer partition—in the self-optimal regime, HE. The displacement therefore works against the reported differences in target coupling and noise overlap, both of which are estimated to be lower for HE despite HE carrying the larger upward displacement, so those comparisons are conservative. Comparisons across components require more care. Target coupling and noise overlap are inflated by the same amount, so their difference is unaffected and the finding that noise overlap dominates the AE–HE variance gap does not depend on the displacement.

## Appendix G: Honest vs. Adaptive Lasso on the ACIC Datasets

Honesty is a design principle that is not exclusive to causal forest. It applies to any learning procedure that separates variable selection (or space partitioning) from parameter estimation. Lasso illustrates this idea well, as it combines variable selection with coefficient estimation. When both tasks are performed on the same sample, we obtain an adaptive Lasso, in which selection errors propagate into the estimated coefficients. In contrast, an honest Lasso decouples the two stages by using separate samples for selection and estimation, thereby mitigating overfitting.

This principle also extends to meta-learners in causal inference, whose behavior inherits the properties of their base learners. For example, a T-learner built with random forests is honest when honest forests are used as base learners and adaptive otherwise. We use Lasso as a representative example to compare adaptive and honest implementations on the ACIC datasets.

We consider the following linear interaction model (Imbens and Rubin 2015)

$$Y_i = \alpha_0 + \gamma_0 T_i + \mathbf{X}_i^\top \boldsymbol{\alpha} + T_i(\mathbf{X}_i - \bar{\mathbf{X}})^\top \boldsymbol{\gamma} + \epsilon_i, \tag{151}$$

where $\alpha_0$, $\gamma_0$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ are coefficients to be learned, and $\bar{\mathbf{X}}$ denotes the vector of covariate means. This model yields the predicted individual treatment effect

$$\hat{\beta}_i = \hat{\gamma}_0 + (\mathbf{X}_i - \bar{\mathbf{X}})^\top \hat{\boldsymbol{\gamma}}. \tag{152}$$

For adaptive Lasso, we fit a single Lasso model on the full training sample, performing variable selection and coefficient estimation simultaneously within one regression.

For honest Lasso, we separate selection and estimation using disjoint samples and implement this via 2-fold cross-fitting. The training data are randomly partitioned into two equal folds, $\mathcal{S}_{\text{sel}}$ and $\mathcal{S}_{\text{es}}$.

1. **Selection:** Fit Lasso on $\mathcal{S}_{\text{sel}}$ to identify the active set of variables.
2. **Estimation:** Fit ordinary least squares on $\mathcal{S}_{\text{es}}$ using only the selected variables to estimate coefficients.

We then swap the roles of the two folds and average predictions across the two fits. This cross-fitting scheme is essentially an ensemble of honest estimators, which recovers much of the efficiency lost in a single honest fit by ensuring that every observation contributes to both stages. For comparison, we also evaluate a single-fit honest Lasso, which does not swap the folds and uses only half of the data for each task.

For both methods, the Lasso $L_1$ regularization parameter ($\lambda$) is tuned over a log-spaced grid of 30 values from $10^{-3}$ to approximately 3, with the optimal value selected by minimizing the transformed MSE. For honest Lasso, $\lambda$ is tuned separately for the two selection models.

Figure 9 reports the average $S^2$ across SNR deciles for adaptive Lasso, honest Lasso with cross-fit, and honest Lasso with single-fit. The adaptive and cross-fitted honest versions exhibit a pattern similar to causal forests: honesty performs better in the lowest-SNR regime, whereas adaptivity becomes advantageous as the SNR increases. This pattern reinforces our core insight that honesty acts as a form of regularization.

The performance gap is smaller than for forests because Lasso combines a linear structure with $L_1$ shrinkage. This constraint already limits flexibility and overfitting, so the additional regularization provided by honesty has a more modest effect. Moreover, even half the sample is often enough to identify relevant predictors, which reduces the cost of sample splitting.

Finally, the single-split honest Lasso illustrates the cost of honesty: using only half the data at each stage increases variance. Cross-fitting mitigates this loss by bagging two honest estimators, allowing all observations to contribute. In this sense, moving from a single-fit to a cross-fit honest Lasso is analogous to moving from a causal tree to a forest—aggregation offsets the variance introduced by sample splitting while preserving the regularization benefit. A key practical takeaway is that, whenever honest estimation is used, it should be paired with bagging.

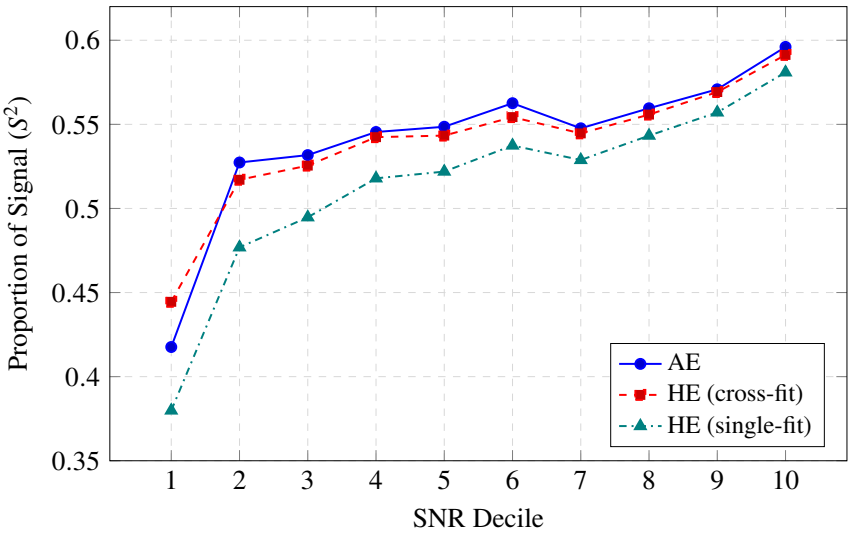


**Figure 9** **Proportion of explained variance ($S^2$) across SNR deciles for adaptive, cross-fit honest, and single-fit honest Lasso models. AE underperforms HE in the lowest SNR decile and is preferable when SNR is higher. This highlights the regularization effect of honesty. The weak performance of the single-fit HE emphasizes the efficiency loss of sample splitting, which is mitigated via cross-fitting.**

## Appendix H: Learning Curve Analysis for the MegaFon Dataset

To complement the smaller sample sizes in the ACIC datasets, we extend our comparison of honest and adaptive forests to a large scale setting. We use the MegaFon dataset from the 2021 MegaFon Uplift Competition, a synthetic dataset designed to mimic realistic telecom customer behavior.[8] It contains 600,000 observations and 50 features, which allows us to evaluate model performance across a wide range of training sample sizes. The treatment indicates whether a customer receives an offer (treatment rate 50%), and the outcome records whether the desired action is taken. The average treatment effect is 4.95%.

Our goal is to assess how adaptive and honest causal forests scale with training data size. In the ACIC datasets, different datasets differ in signal strength and therefore the SNR. In the MegaFon setting, we instead vary the SNR through the training sample size: less data leads to more sampling error—and thus a lower SNR from the learner's perspective—while larger training sets increase the SNR. This setup lets us study performance along a learning curve.

We use 5×5 nested cross-validation to obtain out-of-sample estimates for every observation.

[8] MegaFon dataset URL: `https://ods.ai/tracks/df21-megafon/competitions/megafon-df21-comp/data`

- **Outer loop (evaluation)**: We split the data into five outer folds. In each iteration, one fold is held out as a test set, and the remaining four folds form the training pool. From this pool, we draw training subsets of increasing sizes

$$N \in \{1\text{K}, 2\text{K}, 5\text{K}, 10\text{K}, 20\text{K}, 50\text{K}, 100\text{K}, 200\text{K}, 480\text{K}\}.$$

For each $N$, we train a model on the subset and evaluate it on the held-out outer test fold. This procedure measures how predictive performance scales with data availability.

- **Inner loop (hyperparameter tuning)**: For each training subset of size $N$, we perform 5-fold inner cross-validation to tune the `min_samples_leaf` hyperparameter over $\{10, 20, 50, 100, 200\}$. We select the value that minimizes the MSE on transformed outcomes within the validation set, as discussed in Section 4.2. The model was then retrained on the full subset using the optimal hyperparameter before final evaluation on the outer test fold.

Because the true CATE is unknown in the MegaFon dataset, we evaluate performance using the MSE of transformed outcomes ($Z$) on the full dataset. We adapt the $S^2$ metric as

$$S_z^2 = 1 - \frac{\text{MSE}}{\text{Var}(Z)}, \tag{153}$$

where $\text{Var}(Z)$ is the variance of the transformed outcomes across the full dataset. In this context, $S_z^2$ measures the proportion of variance in the transformed outcomes explained by the model.

Figure 10a reports the $S_z^2$ of adaptive and honest forests across training sizes, and Figure 10b shows the average difference in squared errors, alongside 95% confidence intervals. Note that $S_z^2$ values remain small due to the high variance in transformed outcomes.

The results closely mirror the patterns observed on the ACIC datasets when training size is interpreted as a proxy for the SNR. In the small-sample regime ($N \leq 2\text{K}$), HE performs better than AE, although the negative $S_z^2$ values indicate that both methods overfit the noise and underperform an ATE baseline. As $N$ grows, AE overtakes HE, expanding its lead as the SNR improves. Finally, at large sample sizes ($N \geq 50\text{K}$), this performance gap gradually narrows.

These results show that AE's advantage is not a small-sample artifact; it persists in large samples. But sample size is only one determinant of the SNR, which also depends on the data-generating process. Thus, the performance gap between AE and HE reflects differences in underlying SNR rather than sample size per se. The MegaFon learning curves confirm that AE tends to outperform in higher-SNR regimes, while HE works better when the signal is weak.

## Appendix I: Literature Review

This appendix reviews work on causal trees and causal forests, with a focus on honest estimation.

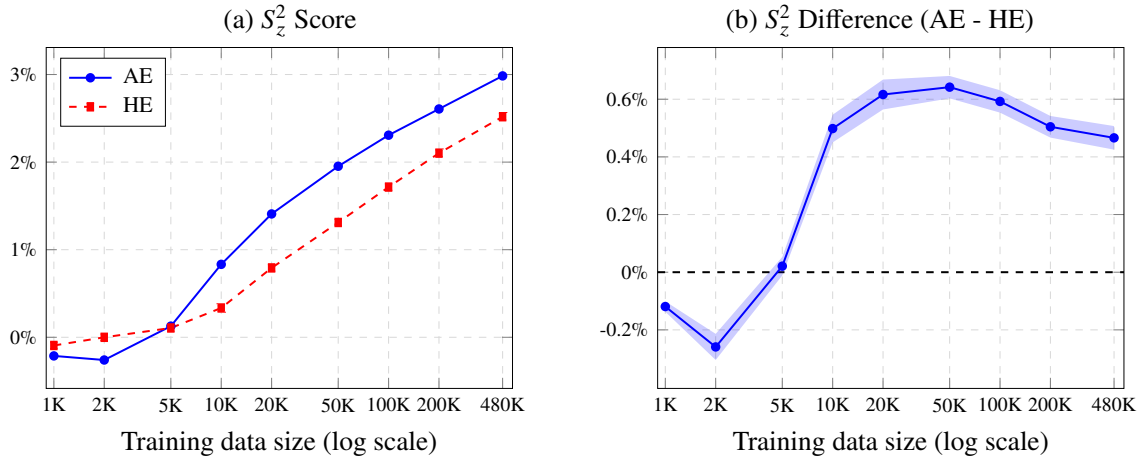


**Figure 10** **Learning curves for causal forests on the MegaFon dataset. The panels illustrate the out-of-sample $S_z^2$ (left) and the average difference in squared errors with 95% confidence intervals (right) across training sample sizes ($N$). The results closely mirror the patterns observed on the ACIC datasets when training size is interpreted as a proxy for the SNR: HE holds a slight advantage in the high-noise, small-sample regime ($N \leq 2$K), while AE achieves a significant performance lead as data size increases.**

Athey and Imbens (2016) adapt the CART algorithm to the problem of estimating heterogeneous treatment effects. Their key contributions are a splitting criterion designed to detect effect heterogeneity, and the introduction of honest estimation, in which separate samples are used to construct the tree partition and to estimate effects within leaves. The motivation for honesty is to reduce bias from adaptive overfitting and to facilitate valid statistical inference after partition selection.

Wager and Athey (2018) extend this framework to random forests, showing that ensembles of honest trees yield consistent effect estimates and enable asymptotically valid inference. This approach is further generalized in the generalized random forest framework (Athey et al. 2019), which casts forest methods as adaptive weighting schemes for a broad class of local moment conditions, allowing effect estimation to be embedded in a more flexible and unified framework.

Recent theoretical work helps explain why honesty is appealing, but also why it is not a safeguard against poor CATE estimates. When the data used to choose splits is separated from the data used to estimate outcomes within leaves, trees can achieve consistent estimation under standard regularity conditions. Specifically, Bladt and Lemvig (2026) show that honest regression trees and random forests can consistently estimate conditional expectations when leaves become

more localized in feature space—reducing approximation error—while still containing enough observations for stable within-leaf estimation.

However, this consistency does not guarantee good finite-sample performance, particularly when the resulting leaves are unstable or poorly aligned with the underlying heterogeneity. For instance, Cattaneo et al. (2022) show that CART-style greedy splitting has a persistent "end-cut preference," producing imbalanced leaves with non-vanishing probability even when pruning is allowed; ensembling via random forests can mitigate the impact of this issue on predictions, but does not eliminate it. Cattaneo et al. (2025) extend this analysis to derive accuracy limits for widely used causal-tree estimators and splitting rules, showing that honest estimation does not address these small-leaf pathologies. Taken together, these results suggest that honesty can simplify theoretical analysis and enable valid inference, but the resulting partitions may still provide poor approximations for individual CATEs.

Empirical work suggests that whether honest or adaptive estimation works better depends on the data-generating process and on how model complexity (such as tree depth) is tuned. Havelka (2022) compares the two approaches in simulations and finds that honesty tends to perform better in settings where subgroup selection is prone to overfitting—for example, when treatment effect heterogeneity is relatively mild—while adaptive estimation can perform better when effects are sharply concentrated in specific subgroups. Although not framed in these terms by Havelka, this pattern is consistent with interpreting honesty as a form of regularization, reducing variance from overfitting at the cost of potentially higher approximation bias.

Similarly, Prodan (2025) shows that whether honesty improves accuracy depends on how tree depth and leaf size are tuned, as these hyperparameters shape the bias–variance trade-off. In noisy or high-dimensional settings, pairing honesty with shallower trees and larger leaves improves performance by controlling variance. But when treatment effects are highly nonlinear and involve many interactions, deeper trees and more adaptive (non-honest) approaches can reduce approximation bias and achieve lower overall error. These findings suggest that the empirical question is not whether honest or adaptive estimation is universally preferable, but under which regimes—defined jointly by noise, the nature of effect heterogeneity, sample size, and tuning choices—one approach is better than the other.

Beyond treatment effect estimation, related work also examines honesty in forest-based decision making, including stochastic optimization and policy learning. In this literature, sample splitting between tree construction and decision estimation is often imposed to make asymptotic optimality guarantees tractable. Kallus and Mao (2023) note that this requirement is largely technical—introduced to simplify theoretical analysis—and can reduce performance in finite samples, a pattern they corroborate with empirical comparisons.

Taken together, these findings—including ours—imply that the impact of honest estimation on accuracy depends on context and should be evaluated within each specific application. Yet in practice, it is often treated as a default rather than a deliberate choice. In our review of 45 studies published in INFORMS journals (Appendix J), fewer than half explicitly reported using honest estimation (22 studies), fewer than a third justified that choice (14 studies), and only two studies considered potential accuracy trade-offs or compared it to adaptive estimation (Fernández-Loría and Provost 2025, Kallus and Mao 2023). In both cases, honest estimation often reduced performance. Among the 14 studies that provided a justification, most focused on avoiding overfitting or ensuring consistency, rather than on whether honesty improved performance in the specific application. This pattern suggests that honest estimation is frequently adopted without careful evaluation. We hope our study encourages researchers to treat honesty as a design choice—one that merits explicit reasoning and empirical assessment.

## Appendix J: Applications of Causal Tree and Forest Methods

Table 2 presents 45 papers published in INFORMS journals (2019–2025) that use causal trees or causal forests. The majority employ these methods to estimate individual-level effects, which are then used for other analyses.

Table 2: Research Papers Using Causal Tree/Forest Methods

| **Paper** | **Journal** | **Application** |
|---|---|---|
| Bertsimas et al. (2019): Optimal Prescriptive Trees | INFORMS Journal on Optimization | Treatment assignment |
| Luo et al. (2019): When and How to Leverage E-commerce Cart Targeting: The Relative and Moderated Effects of Scarcity and Price Incentives with a Two-Stage Field Experiment and Causal Forest Optimization | Information Systems Research | Effect heterogeneity analysis |
| Ge et al. (2021): Human–Robot Interaction: When Investors Adjust the Usage of Robo-Advisors in Peer-to-Peer Lending | Information Systems Research | ATE estimation |
| Hermosilla (2021): Rushed Innovation: Evidence from Drug Licensing | Management Science | Effect heterogeneity analysis |
| Kallus and Zhou (2021): Minimax-Optimal Policy Learning Under Unobserved Confounding | Management Science | Treatment assignment |

| Paper | Journal | Application |
|---|---|---|
| Fernández-Loría and Provost (2022): Causal Decision Making and Causal Effect Estimation Are Not the Same...and Why It Matters | INFORMS Journal on Data Science | Treatment assignment |
| Choudhary et al. (2022): Nudging Drivers to Safety: Evidence from a Field Experiment | Management Science | Effect heterogeneity analysis |
| Cui and Davis (2022): Tax-Induced Inequalities in the Sharing Economy | Management Science | Effect heterogeneity analysis |
| Godinho De Matos and Adjerid (2022): Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider | Management Science | Effect heterogeneity analysis |
| Pattabhiramaiah et al. (2022): Spillovers from Online Engagement: How a Newspaper Subscriber's Activation of Digital Paywall Access Affects Her Retention and Subscription Revenue | Management Science | Effect heterogeneity analysis |
| Wang (2022): The Effect of Medicaid Expansion on Wait Time in the Emergency Department | Management Science | Causal forest-based method development |
| Wang et al. (2022): An Instrumental Variable Forest Approach for Detecting Heterogeneous Treatment Effects in Observational Studies | Management Science | Causal tree-based method development |
| Fernández-Loría et al. (2023): A Comparison of Methods for Treatment Assignment with an Application to Playlist Generation | Information Systems Research | Treatment assignment |
| Cohen et al. (2023): Managing Airfares Under Competition: Insights from a Field Experiment | Management Science | Effect heterogeneity analysis |
| Kallus and Mao (2023): Stochastic Optimization Forests | Management Science | Causal forest-based method development |
| Serra-Garcia and Szech (2023): Incentives and Defaults Can Increase COVID-19 Vaccine Intentions and Test Demand | Management Science | Effect heterogeneity analysis |
| Unal and Park (2023): Fewer Clicks, More Purchases | Management Science | Effect heterogeneity analysis |

| Paper | Journal | Application |
|---|---|---|
| Yoganarasimhan et al. (2023): Design and Evaluation of Optimal Free Trials | Management Science | Treatment assignment |
| Zhang and Luo (2023): Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp | Management Science | ATE estimation |
| Zhang et al. (2023): Uncovering Synergy and Dysergy in Consumer Reviews: A Machine Learning Approach | Management Science | Effect heterogeneity analysis |
| Borenstein et al. (2023): Ancillary Services in Targeted Advertising: From Prediction to Prescription | Manufacturing & Service Operations Management | Effect heterogeneity analysis |
| Wang et al. (2023): Personalized Healthcare Outcome Analysis of Cardiovascular Surgical Procedures | Manufacturing & Service Operations Management | Causal tree-based method development |
| Zhou et al. (2023): Offline Multi-Action Policy Learning: Generalization and Optimization | Operations Research | Treatment assignment |
| Ayabakan et al. (2024): Impact of Telehealth and Process Virtualization on Healthcare Utilization | Information Systems Research | ATE estimation |
| Leng and Dimmery (2024): Calibration of Heterogeneous Treatment Effects in Randomized Experiments | Information Systems Research | Correction for causal forest effect estimates |
| Bundorf et al. (2024): How Do Consumers Interact with Digital Expert Advice? Experimental Evidence from Health Insurance | Management Science | Effect heterogeneity analysis |
| Kraus et al. (2024): Data-Driven Allocation of Preventive Care with Application to Diabetes Mellitus Type II | Manufacturing & Service Operations Management | Treatment prioritization |
| Pauphilet (2024): Robust and Heterogeneous Odds Ratio: Estimating Price Sensitivity for Unbought Items | Manufacturing & Service Operations Management | Effect heterogeneity analysis |

| Paper | Journal | Application |
|---|---|---|
| Berman and Feit (2024): Latent Stratification for Incrementality Experiments | Marketing Science | ATE estimation |
| Huang and Ascarza (2024): Doing More with Less: Overcoming Ineffective Long-Term Targeting Using Short-Term Signals | Marketing Science | Treatment prioritization |
| Turjeman and Feinberg (2024): When the Data Are Out: Measuring Behavioral Changes Following a Data Breach | Marketing Science | Causal forest-based method development |
| Fernández-Loría and Provost (2025): Observational vs. Experimental Data When Making Automated Decisions Using Machine Learning | INFORMS Journal on Data Science | Treatment assignment |
| Kim et al. (2025): Working Daily, Paid Monthly? Effects of On-Demand Wage Access on the Financial Engagement of Low-Wage Workers | Information Systems Research | ATE estimation |
| Alyakoob and Rahman (2025): Market Design Choices, Racial Discrimination, and Equitable Microentrepreneurship in Digital Marketplaces | Management Science | Effect heterogeneity analysis |
| Cao et al. (2025): Antisocial Responses to the "Coal to Gas" Regulation: An Unintended Consequence of a Residential Energy Policy | Management Science | Effect heterogeneity analysis |
| Gulen et al. (2025): Balancing External vs. Internal Validity: An Application of Causal Forest in Finance | Management Science | Effect heterogeneity analysis |
| Opitz et al. (2025): The Algorithmic Assignment of Incentive Schemes | Management Science | Treatment assignment |
| Long et al. (2025): The Choice Overload Effect in Online Recommender Systems | Manufacturing & Service Operations Management | Effect heterogeneity analysis |
| Luo et al. (2025): From Stopping to Shopping: A Field Experiment on Free Return and Free Shipping Retargeting Policies in Online Retail Operations | Manufacturing & Service Operations Management | Effect heterogeneity analysis |

| Paper | Journal | Application |
|---|---|---|
| Pourghannad and Wang (2025): Matching Patients with Surgeons: Heterogeneous Effects of Surgical Volume on Surgery Duration | Manufacturing & Service Operations Management | Effect heterogeneity analysis |
| Huang and Morozov (2025): The Promotional Effects of Live Streams by Twitch Influencers | Marketing Science | Effect heterogeneity analysis |
| Kaul et al. (2025): Call Me Maybe: Does Customer Feedback Seeking Impact Nonsolicited Customers? | Marketing Science | Effect heterogeneity analysis |
| Lin and Strulov-Shlain (2025): Choice Architecture, Privacy Valuations, and Selection Bias in Consumer Data | Marketing Science | Effect heterogeneity analysis |
| Rafieian et al. (2025): Multiobjective Personalization of Marketing Interventions | Marketing Science | Effect heterogeneity analysis |
| Von Zahn et al. (2025): Smart Green Nudging: Reducing Product Returns Through Digital Footprints and Causal Machine Learning | Marketing Science | Treatment assignment |

## References

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

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *The Annals of Statistics* 47(2):1148–1178.

Bladt M, Lemvig RF (2026) Consistency of honest decision trees and random forests. *arXiv preprint arXiv:2601.14991* .

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees* (Belmont, CA: Wadsworth International Group).

Cattaneo MD, Klusowski JM, Tian PM (2022) On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.10805* .

Cattaneo MD, Klusowski JM, Yu RR (2025) The honest truth about causal trees: Accuracy limits for heterogeneous treatment effect estimation. *arXiv preprint arXiv:2509.11381* .

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10(7):1895–1923.

Fernández-Loría C, Provost F (2025) Observational vs. experimental data when making automated decisions using machine learning. *INFORMS Journal on Data Science* 4(3):197–229.

Friedman JH, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1):1–22.

Hammond M, Lewis NT, Boone S, Chen X, Mendonça JM, Parmentier V, Taylor J, Bell T, dos Santos L, Crouzet N, et al. (2024) Identifying and fitting eclipse maps of exoplanets with cross-validation. *Monthly Notices of the Royal Astronomical Society* 532(4):4350–4368.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer), 2nd edition.

Havelka M (2022) Honesty in causal forests, is it worth it? https://repository.tudelft.nl/file/File_86d8e7a1-8951-4b8a-8714-63a01ba9c805, bachelor's thesis, Delft University of Technology.

Imai K, Li ML (2025) Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. *Journal of Business & Economic Statistics* 43(1):256–268.

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge, UK: Cambridge University Press).

Kallus N, Mao X (2023) Stochastic optimization forests. *Management Science* 69(4):1975–1994.

Lei J (2020) Cross-validation with confidence. *Journal of the American Statistical Association* 115(532):1978–1997.

Piironen J, Paasiniemi M, Vehtari A (2020) Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics* 14(1):2155–2197.

Prodan RM (2025) Analyzing the impact of depth and leaf size on cate estimation in honest causal trees. https://repository.tudelft.nl/file/File_edf8b751-f408-4186-b343-6bcfb2aa6765, bachelor's thesis, Delft University of Technology.

Villanueva E, Smith T, Pizzinga M, Elzek M, Queiroz RM, Harvey RF, Breckels LM, Crook OM, Monti M, Dezi V, et al. (2024) System-wide analysis of rna and protein subcellular localization dynamics. *Nature Methods* 21(1):60–71.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Windmeijer F, Farbmacher H, Davies N, Davey Smith G (2019) On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association* 114(527):1339–1350.