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

Matias D. Cattaneo
Princeton University

Jason M. Klusowski
Princeton University

Ruiqi (Rae) Yu
Princeton University

November 2025

Outline

Introduction

Adaptive Decision Trees are widely used in academia and industry.

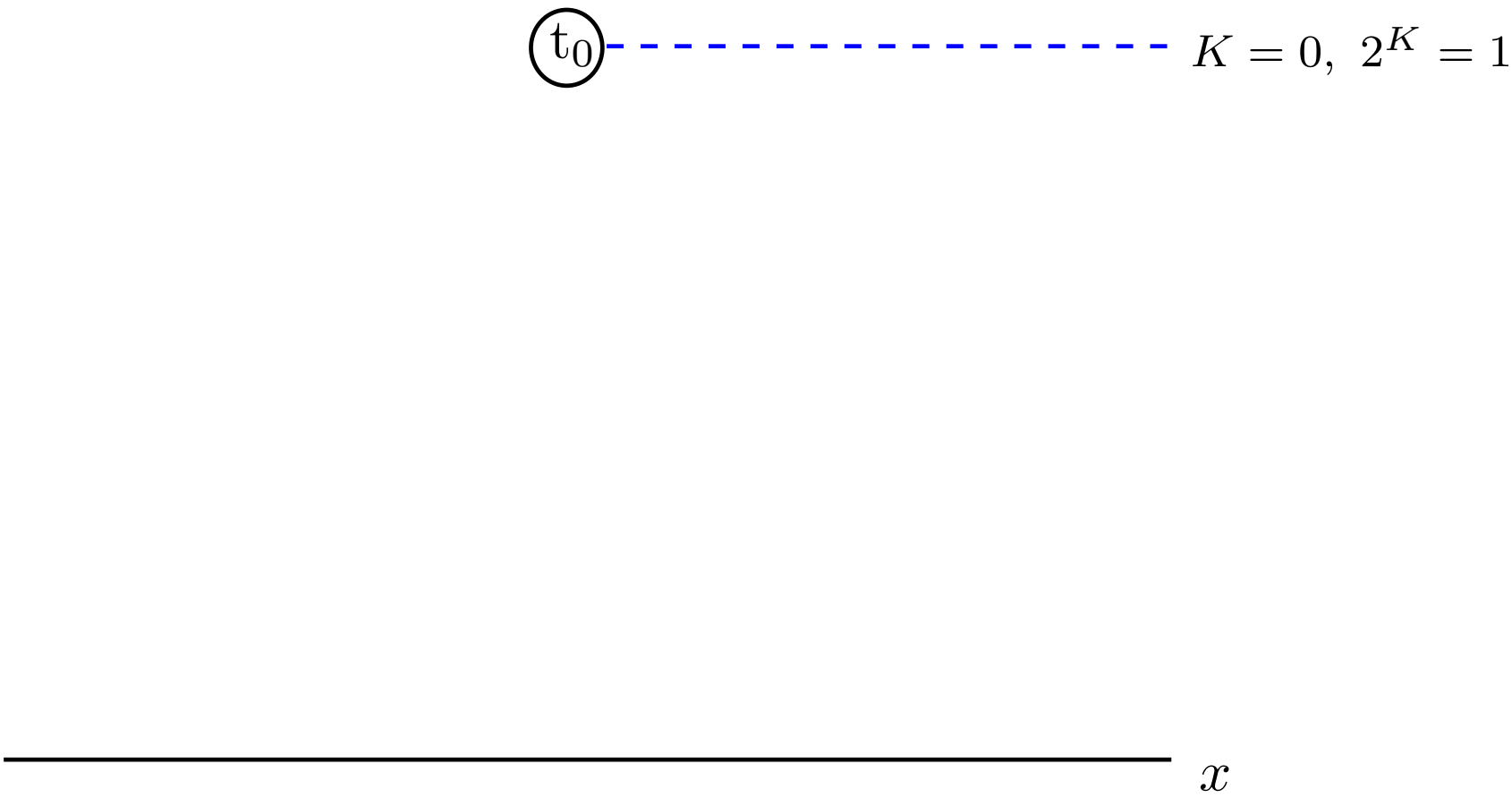
- Also known as *Recursive Partitioning* methods.
- CART: Breiman, Friedman, Olshen & Stone (1984).
- Adaptivity: incorporate data features in their construction.
- Popularity: prime example of "modern" machine learning toolkit.
- Prediction vs. Causality: **Causal Decision Trees** (Athey and Imbens, 2016, PNAS).
- Sometimes preferred for interpretability or pointwise learning:

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

  where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ covariates supported on $\mathcal{X}$.

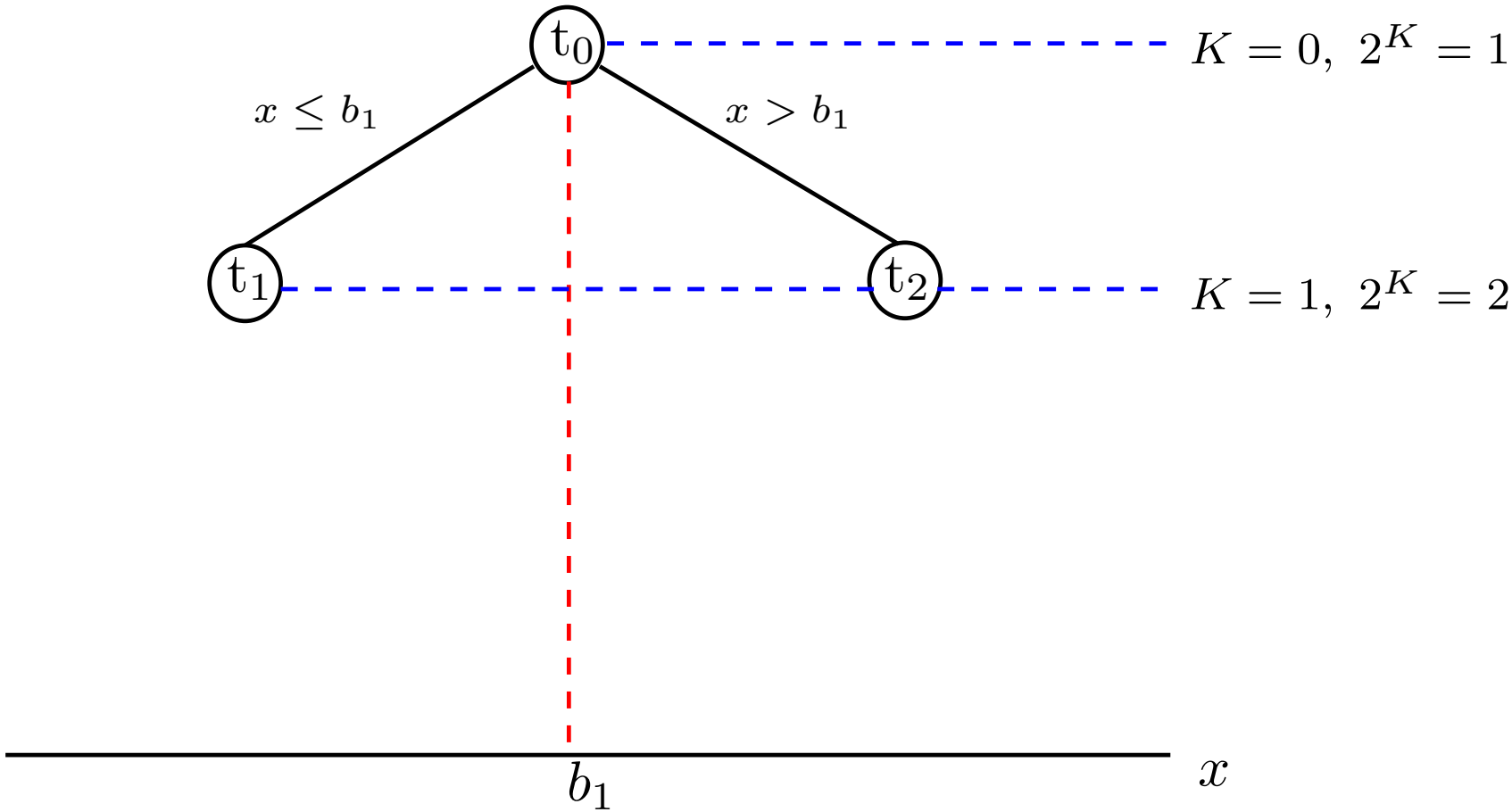
- Today: three foundational results.
  1. Highly inaccurate pointwise (hence uniform) convergence, possibly inconsistent.
  2. Sample splitting (so-called "honesty") does not help.
  3. Near-optimal mean square convergence (special case), but sample splitting does not help.

Adaptive Axis-Aligned Decision Tree (CART)



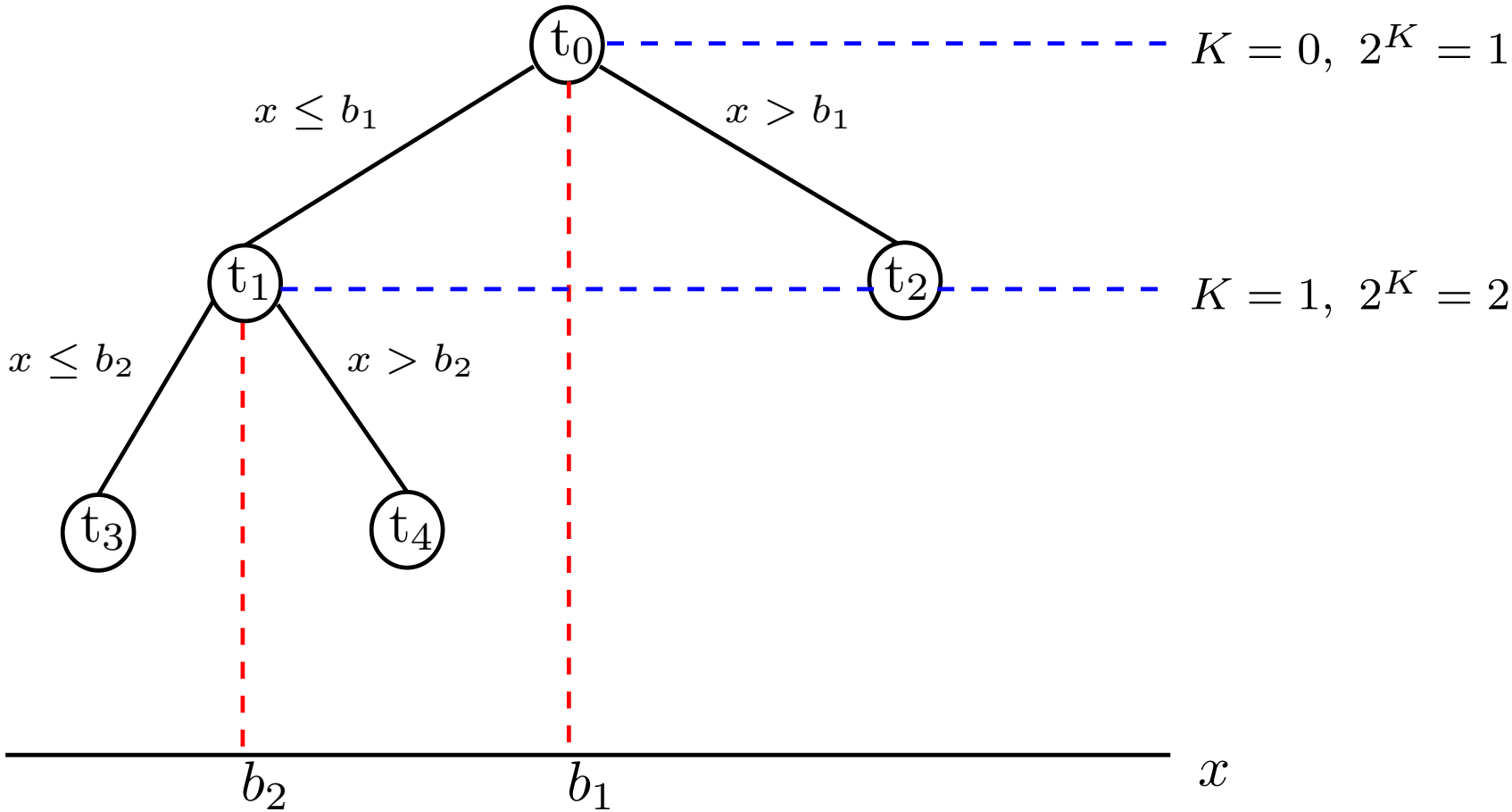
for each K : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathsf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \leq \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

Adaptive Axis-Aligned Decision Tree (CART)



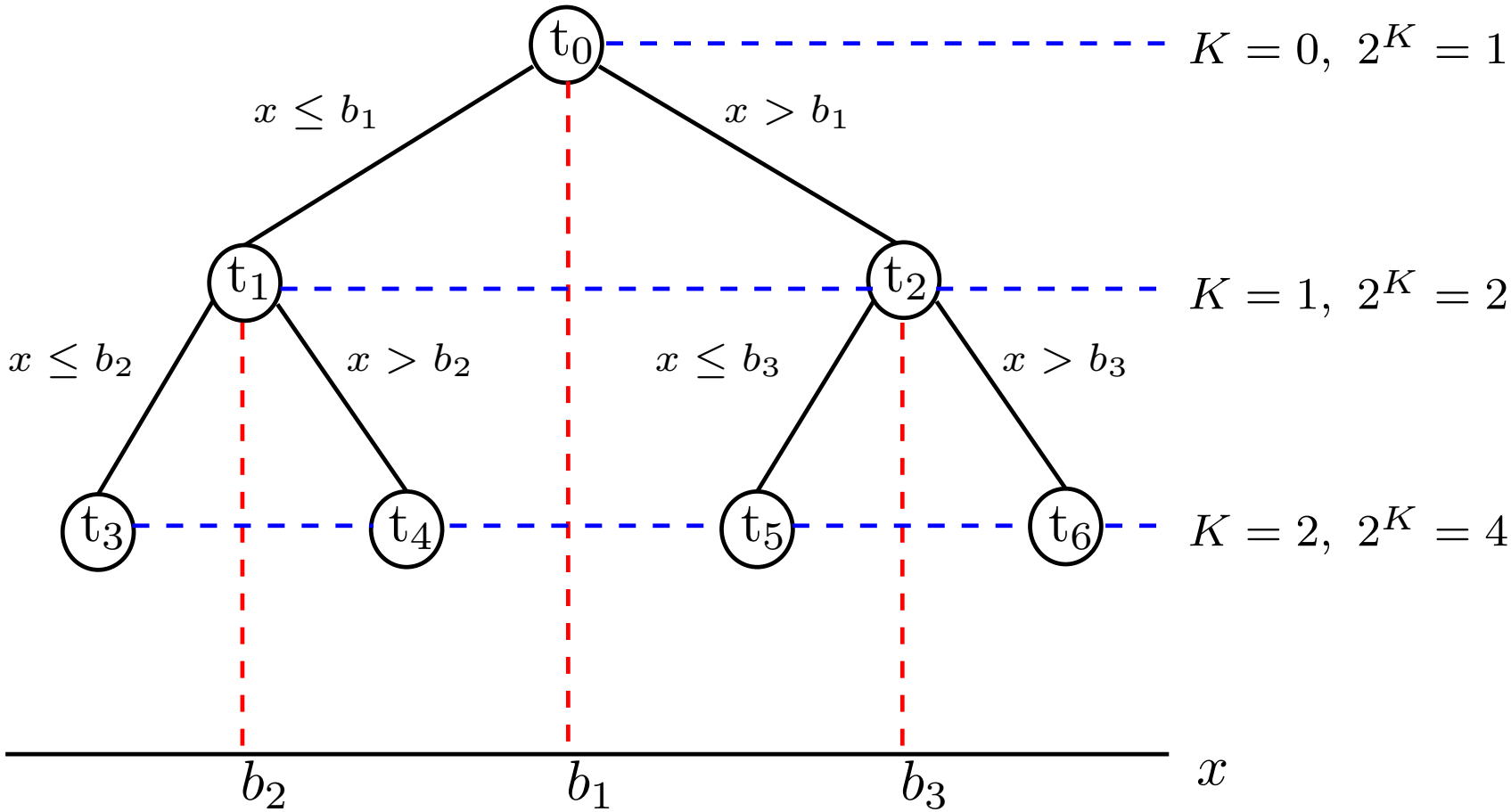
for each K : $$\min_{j=1,2,\cdots,p}\ \min_{\beta_1,\beta_2,\tau}\ \sum_{\mathbf{x}_i \in \mathsf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau)\Big)^2$$

Adaptive Axis-Aligned Decision Tree (CART)



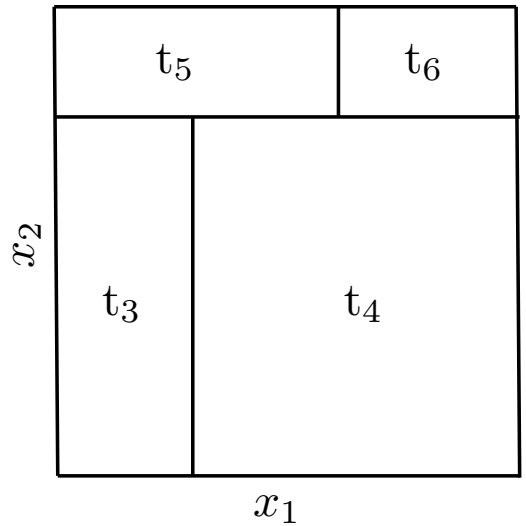
for each K : $$\min_{j=1,2,\cdots,p}\ \min_{\beta_1,\beta_2,\tau}\ \sum_{\mathbf{x}_i \in \mathsf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau)\Big)^2$$

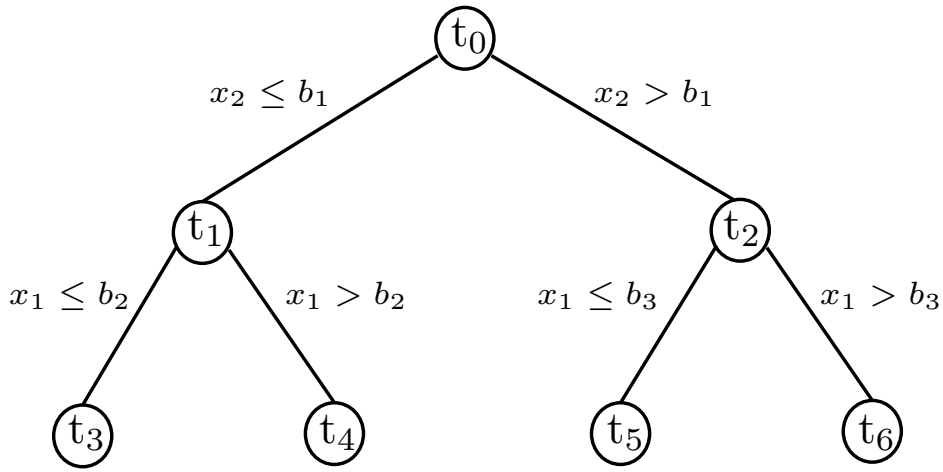
Adaptive Axis-Aligned Decision Tree (CART)



for each K : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathsf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

Adaptive Axis-Aligned Decision Tree (CART vs. "Honesty")



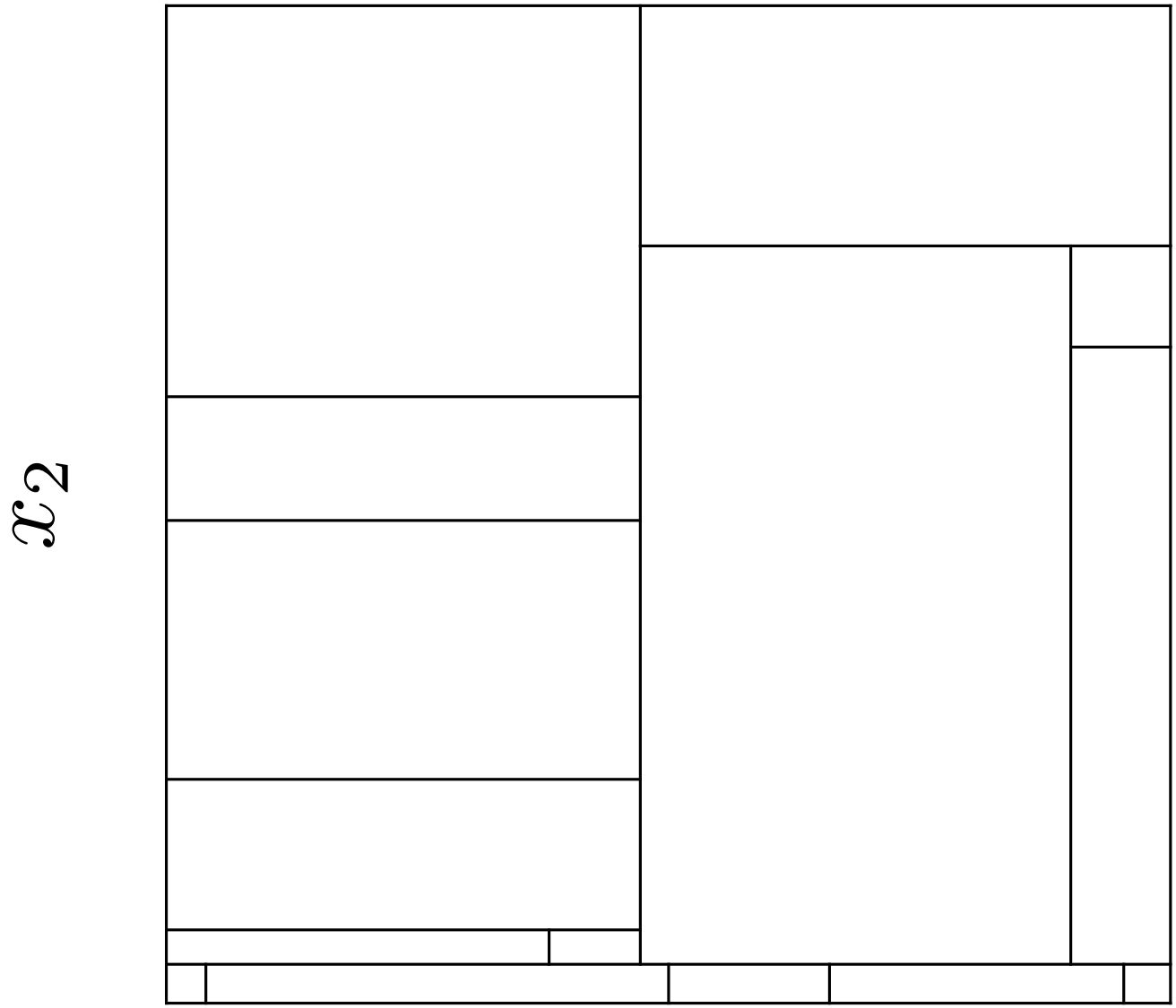


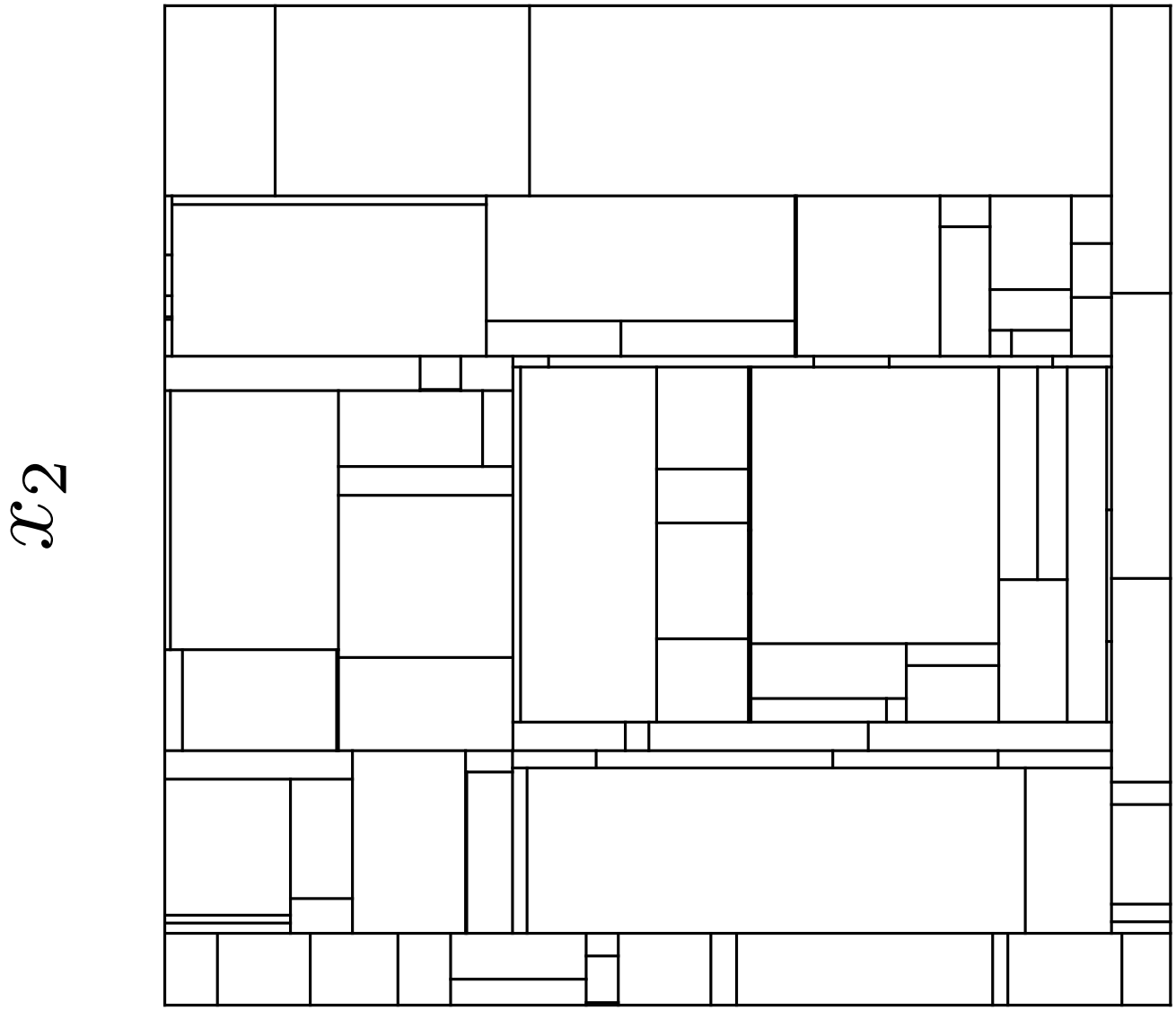
- Full sample (or CART):

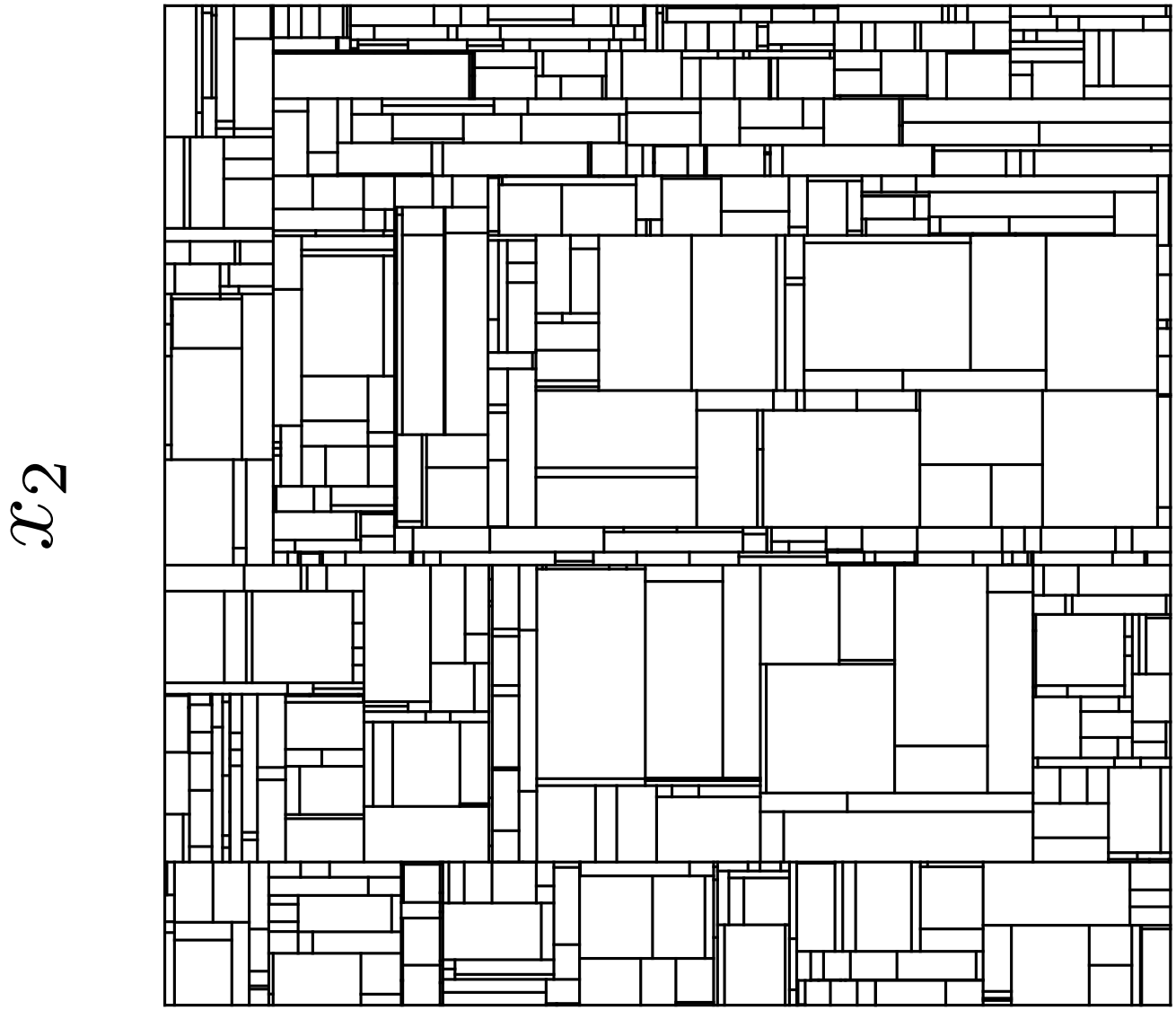
$$\widehat{\mu}(\mathbf{x};\mathsf{T}_K) = \overline{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\mathbf{x}_i\in\mathsf{t}} y_i, \qquad n(\mathsf{t}) = \sum_{\mathbf{x}_i\in\mathsf{t}} \mathbb{1}(\mathbf{x}_i\in\mathsf{t})$$

- Sample splitting (or "honesty"): $\mathsf{T}_K \perp\!\!\!\perp (\widetilde{y}_i, \widetilde{\mathbf{x}}_i : i = 1,\dots,n_\mu)$, and

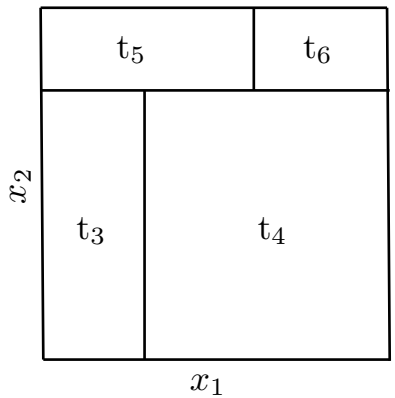
$$\widetilde{\mu}(\mathbf{x};\mathsf{T}_K) = \widetilde{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \widetilde{y}_i, \qquad n(\mathsf{t}) = \sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \mathbb{1}(\widetilde{\mathbf{x}}_i\in\mathsf{t}).$$

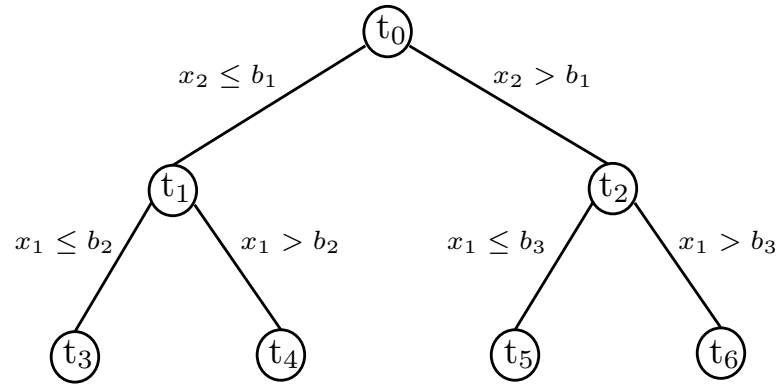

x_2
x_1


x_2
x_1


x_2
x_1

Adaptive Axis-Aligned Decision Tree (CART)





$$\textit{Full sample:} \qquad \widehat{\mu}(\mathbf{x};\mathsf{T}_K) = \overline{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\mathbf{x}_i\in\mathsf{t}} y_i, \qquad n(\mathsf{t}) = \sum_{\mathbf{x}_i\in\mathsf{t}} \mathbb{1}(\mathbf{x}_i \in \mathsf{t})$$

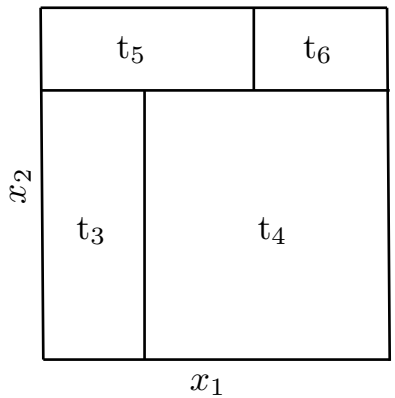
$$\textit{“Honesty”:} \qquad \widetilde{\mu}(\mathbf{x};\mathsf{T}_K) = \widetilde{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \widetilde{y}_i, \qquad n(\mathsf{t}) = \sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \mathbb{1}(\widetilde{\mathbf{x}}_i \in \mathsf{t}).$$

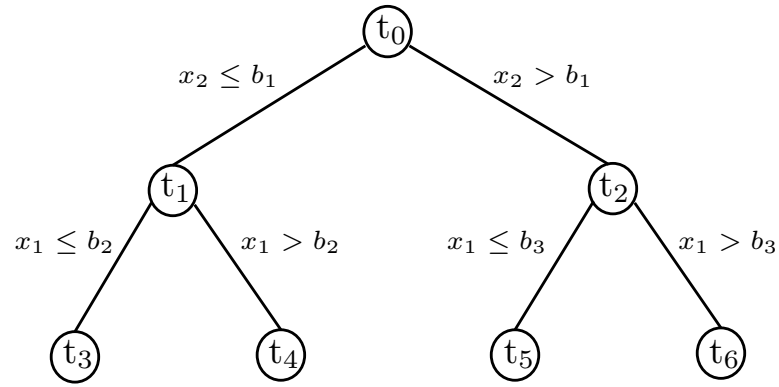
Uniform Result: for $\mu(\mathbf{x}) = \mu$, and for all $K \geq 1$ and $b \in (0,1)$,

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{\mathbf{x}\in\mathcal{X}} \big|\widehat{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 \geq C_1 \frac{\log\log(n)}{n^b} \Big) \geq C_2 b,$$

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{\mathbf{x}\in\mathcal{X}} \big|\widetilde{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 \geq C_1 \frac{1}{n^b} \Big) \geq C_2 b,$$

Adaptive Axis-Aligned Decision Tree (CART)





$$\textit{Full sample:} \qquad \widehat{\mu}(\mathbf{x};\mathsf{T}_K) = \overline{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\mathbf{x}_i\in\mathsf{t}} y_i, \qquad n(\mathsf{t}) = \sum_{\mathbf{x}_i\in\mathsf{t}} \mathbb{1}(\mathbf{x}_i \in \mathsf{t})$$

$$\textit{“Honesty”:} \qquad \widetilde{\mu}(\mathbf{x};\mathsf{T}_K) = \widetilde{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \widetilde{y}_i, \qquad n(\mathsf{t}) = \sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \mathbb{1}(\widetilde{\mathbf{x}}_i \in \mathsf{t}).$$

Mean Square Result: for $\mu(\mathbf{x}) = \mu$, and for all $K \geq 1$ and $b \in (0,1)$,

$$\mathbb{E}\Big[\int_{\mathcal{X}} \big|\widehat{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x})\Big] \leq C_1 \frac{2^K \log^5(n)}{n} + \frac{2^K \log^4(n)\log(p)}{n},$$

$$\mathbb{E}\Big[\int_{\mathcal{X}} \big|\widetilde{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x})\Big] \leq C_2 \frac{2^K \log^5(n)}{n},$$

Outline

Causal Decision Trees

Recursive partitioning for heterogeneous causal effects

Susan Athey[a,1] **and Guido Imbens**[a]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 20, 2016 (received for review June 25, 2015)

In this paper we propose methods for estimating heterogeneity in causal effects in experimental and observational studies and for conducting hypothesis tests about the magnitude of differences in treatment effects across subsets of the population. We provide a data-driven approach to partition the data into subpopulations

Within the prediction-based machine learning literature, regression trees differ from most other methods in that they produce a partition of the population according to covariates, whereby all units in a partition receive the same prediction. In this paper, we focus on the analogous goal of deriving a partition

"It enables researchers to let the data discover relevant subgroups while preserving the validity of confidence intervals constructed on treatment effects within subgroups."

"Honesty has the implication that the asymptotic properties of treatment effect estimates within the partitions are the same as if the partition had been exogenously given. Although there is a loss of precision due to sample splitting (which reduces sample size in each step of estimation), there is a benefit in terms of eliminating bias that offsets at least part of the cost."

Causal Decision Trees

PNAS

Recursive partitioning for heterogeneous causal effects

Susan Athey[a,1] **and Guido Imbens**[a]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 20, 2016 (received for review June 25, 2015)

In this paper we propose methods for estimating heterogeneity in causal effects in experimental and observational studies and for conducting hypothesis tests about the magnitude of differences in treatment effects across subsets of the population. We provide a data-driven approach to partition the data into subpopulations

Within the prediction-based machine learning literature, regression trees differ from most other methods in that they produce a partition of the population according to covariates, whereby all units in a partition receive the same prediction. In this paper, we focus on the analogous goal of deriving a partition

"It enables researchers to let the data discover relevant subgroups while preserving the validity of confidence intervals constructed on treatment effects within subgroups."

"Honesty has the implication that the asymptotic properties of treatment effect estimates within the partitions are the same as if the partition had been exogenously given. Although there is a loss of precision due to sample splitting (which reduces sample size in each step of estimation), there is a benefit in terms of eliminating bias that offsets at least part of the cost."

Setup

- DGP: $\mathcal{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n\}$ i.i.d., and $y_i = y_i(1) \cdot d_i + y_i(0) \cdot (1 - d_i)$.
- RCT: $(y_i(0), y_i(1), \mathbf{x}_i^\top) \perp\!\!\!\perp d_i$ and $\xi = \mathbb{P}(d_i = 1) \in (0, 1)$.
- Identification:

$$\begin{aligned}
\tau_{\texttt{CATE}}(\mathbf{x}_i) &= \mathbb{E}[y_i(1) - y_i(0) \mid \mathbf{x}_i = \mathbf{x}] \\
&= \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i) \\
&= \mathbb{E}[y_i \mid \mathbf{x}_i,\ d_i = 1] - \mathbb{E}[y_i \mid \mathbf{x}_i,\ d_i = 0] \\
&= \mathbb{E}\Big[y_i \frac{d_i - \xi}{\xi(1 - \xi)} \mid \mathbf{x}_i\Big].
\end{aligned}$$

Causal Trees Methodology (Athey and Imbens, 2016, PNAS):

1. CATE estimator: DIM vs. IPW at terminal nodes (final partition).
2. Tree construction: (adaptive) recursive partitioning with causal-inference-tailored splitting criteria.
3. Data usage: full-sample vs. "honesty".

CATE Estimator

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

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

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

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

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

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

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

Tree Construction

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

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

$$\frac{n(\mathtt{t}_{\mathtt{L}})n(\mathtt{t}_{\mathtt{R}})}{n(\mathtt{t})}\Big(\widehat{\tau}_l(\mathtt{t}_{\mathtt{L}}; \mathtt{T}, \mathcal{D}_{\mathtt{T}}) - \widehat{\tau}_l(\mathtt{t}_{\mathtt{R}}; \mathtt{T}, \mathcal{D}_{\mathtt{T}})\Big)^2, \qquad l \in \{\mathtt{DIM}, \mathtt{IPW}\}.$$

  The two final causal trees are denoted by $\mathtt{T}^{\mathtt{DIM}}(\mathcal{D}_{\mathtt{T}})$ and $\mathtt{T}^{\mathtt{IPW}}(\mathcal{D}_{\mathtt{T}})$, respectively.

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

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

  where only the data $\mathcal{D}_{\mathtt{T}}$ is used.
  The final causal tree is denoted by $\mathtt{T}^{\mathtt{SSE}}(\mathcal{D}_{\mathtt{T}})$.

Data Usage

Recall that $\mathcal{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n\}$ is the available random sample.

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

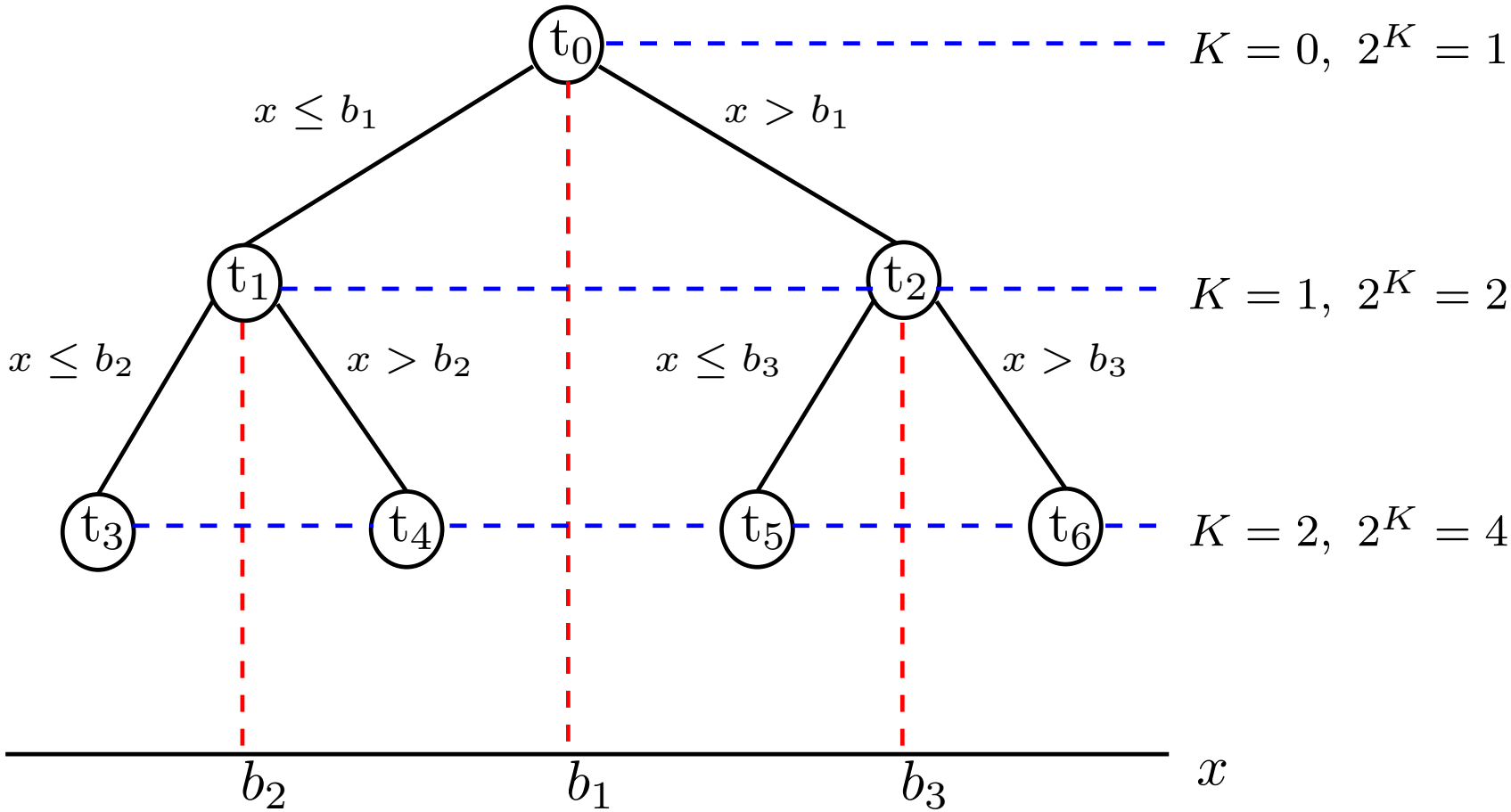
$$\begin{aligned}
\widehat{\tau}^{\texttt{NSS}}_{\texttt{DIM}}(\mathbf{x}) &= \widehat{\tau}_{\texttt{DIM}}(\mathbf{x}; \mathsf{T}^{\texttt{DIM}}(\mathcal{D}), \mathcal{D}), \\
\widehat{\tau}^{\texttt{NSS}}_{\texttt{IPW}}(\mathbf{x}) &= \widehat{\tau}_{\texttt{IPW}}(\mathbf{x}; \mathsf{T}^{\texttt{IPW}}(\mathcal{D}),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

- *Honesty* (HON): The dataset $\mathcal{D}$ is divided in two independent datasets $\mathcal{D}_{\mathsf{T}}$ and $\mathcal{D}_{\tau}$ 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

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

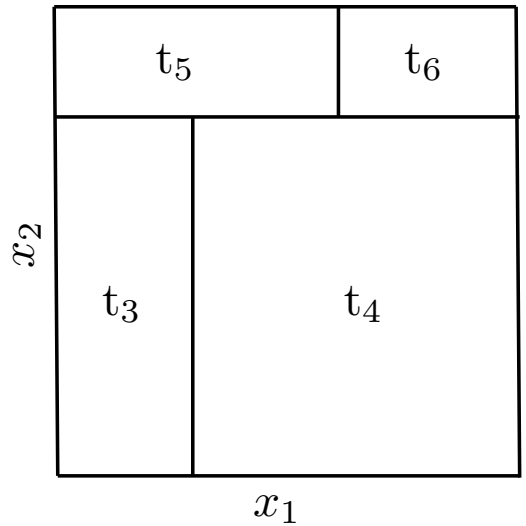
Outline

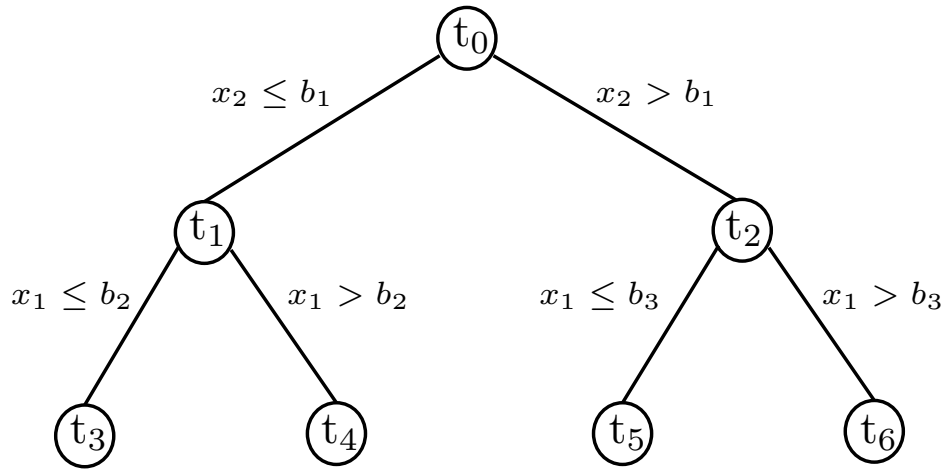
Adaptive Axis-Aligned Decision Tree (CART)



for each K : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathsf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

Adaptive Axis-Aligned Decision Tree (CART vs. "Honesty")





- Full sample (or CART):

$$\widehat{\mu}(\mathbf{x};\mathsf{T}_K) = \overline{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\mathbf{x}_i\in\mathsf{t}} y_i, \qquad n(\mathsf{t}) = \sum_{\mathbf{x}_i\in\mathsf{t}} \mathbb{1}(\mathbf{x}_i\in\mathsf{t})$$

- Sample splitting (or "honesty"): $\mathsf{T}_K \perp\!\!\!\perp (\widetilde{y}_i, \widetilde{\mathbf{x}}_i : i = 1,\ldots,n_\mu)$, and

$$\widetilde{\mu}(\mathbf{x};\mathsf{T}_K) = \widetilde{y}_{\mathsf{t}} = \frac{1}{n(\mathsf{t})}\sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \widetilde{y}_i, \qquad n(\mathsf{t}) = \sum_{\widetilde{\mathbf{x}}_i\in\mathsf{t}} \mathbb{1}(\widetilde{\mathbf{x}}_i\in\mathsf{t}).$$

Setup: Constant Regression Model

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

Assumption (DGP)

1. $(y_i, \mathbf{x}_i^\top)$, $i = 1, 2, \ldots, n$, *is a random sample.*
2. $\mu(\mathbf{x}) \equiv \mu$ *is constant for all* $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$.
3. $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ *independent and continuously distributed.*
4. $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i$ *for all* $i = 1, 2, \ldots, n$.
5. $\mathbb{E}\big[|\varepsilon_i|^{2+\nu}\big] < \infty$ *for some* $\nu > 0$.

Main takeaways: adaptive (CART or variants thereof) decision trees.

1. Decision stumps ($K = 1$) split with high probability "near" boundaries.
2. $\widehat{\mu}(\mathbf{x}; \mathsf{T}_K)$ cannot achieve n^{-b} convergence rate uniformly over $\mathcal{X}$.
3. $\widehat{\mu}(\mathbf{x}; \mathsf{T}_K)$ achieves near-optimal n^{-1} mean squared convergence rate.
4. "Honesty" does not help much (if at all) in either case.
5. $\mathbf{X}$-adaptive tree constructions are uniformly inconsistent as soon as $K \gtrsim \log\log(n)$.
6. Low accuracy/inconsistency at countable many points on $\mathcal{X}$, not just at boundaries.
7. Pruning does not help; regularization needs careful consideration.

Decision Stumps

For each level K, adaptive (CART) decision trees solve

$$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathtt{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \leq \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2,$$

which is equivalent to maximizing the so-called *impurity gain*

$$\sum_{\mathbf{x}_l \in \mathtt{t}} (y_l - \mu)^2 - \sum_{\mathbf{x}_l \in \mathtt{t}} \big(y_l - \overline{y}_{\mathtt{t}_L} \mathbb{1}(x_{lj} \leq \tau) - \overline{y}_{\mathtt{t}_R} \mathbb{1}(x_{lj} > \tau) \big)^2$$

$$= \frac{1}{i(n(\mathtt{t}) - i)} \Big(\frac{1}{\sqrt{n(\mathtt{t})}} \sum_{l=1}^{i} (y_l - \mu) - \frac{i}{n(\mathtt{t})} \frac{1}{\sqrt{n(\mathtt{t})}} \sum_{l=1}^{n(\mathtt{t})} (y_l - \mu) \Big)^2$$

with respect to index i and variable j, after reordering the data $\implies (\hat{\imath}, \hat{\jmath})$.

- Darling-Erdös (1956) limit law (Berkes & Weber, 2006): for any non-decreasing function $1 \leq h(m) \leq m$ for which $\lim_{m\to\infty} h(m) = \infty$ and any $w \in \mathbb{R}$,

$$\mathbb{P}\left(\max_{m/h(m) \leq i \leq m} \left| \frac{1}{\sqrt{i}} \sum_{l=1}^{i} (y_l - \mu) \right| < \lambda(h(m), w) \right) \to e^{e^{-w}},$$

  as $m \to \infty$, where $\lambda(\cdot,\cdot)$ is known.

Decision Stumps: Split Location and Covariate Selected

Careful study of maximum over different ranges of the split index gives:

Theorem

For each $a, b \in (0,1)$ *with* $a < b$ *and* $j \in \{1, 2, \ldots, p\}$,

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

- With positive probability, split index $\hat{\imath}$ concentrates near extremes.
- Too few observations will be available on one of the cells after the first split for CART to deliver a polynomial-in-n consistent estimator of μ.
- Decision stumps exhibit slower than any polynomial-in-n convergence rate.
- Intuitively, iterating the procedure down the tree T_K can only make things worse.

Main Results: Adaptive (Decision/Causal) Trees

Let Assumption DGP hold, and T_K have at least one split (i.e., at least two terminal nodes).

Theorem (Full Sample)

For all $b \in (0,1)$,

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{\mathbf{x}\in\mathcal{X}} \big|\widehat{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 \geq C_1 \frac{\log\log(n)}{n^b} \Big) \geq C_2 b.$$

Furthermore,

$$\mathbb{E}\Big[\int_{\mathcal{X}} \big|\widehat{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x}) \Big] \leq C_3 \frac{2^K \log^4(n)(\log(n)+\log(p))}{n}.$$

Theorem (Honesty)

For all $b \in (0,1)$,

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{\mathbf{x}\in\mathcal{X}} \big|\widetilde{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 \geq C_1 \frac{1}{n^b} \Big) \geq C_2 b.$$

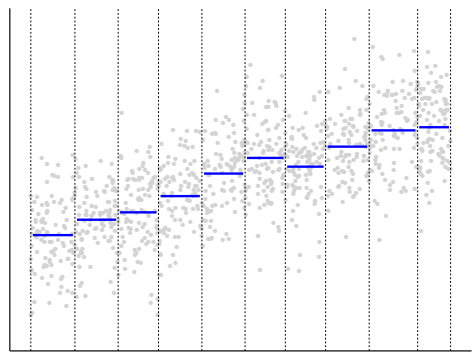
Furthermore,

$$\mathbb{E}\Big[\int_{\mathcal{X}} \big|\widetilde{\mu}(\mathbf{x};\mathsf{T}_K) - \mu(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x}) \Big] \leq C_3 \frac{2^K \log^5(n)}{n}.$$

Outline

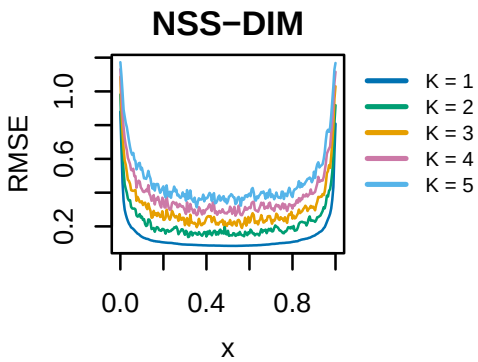
Main Results: Adaptive (Decision/Causal) Trees

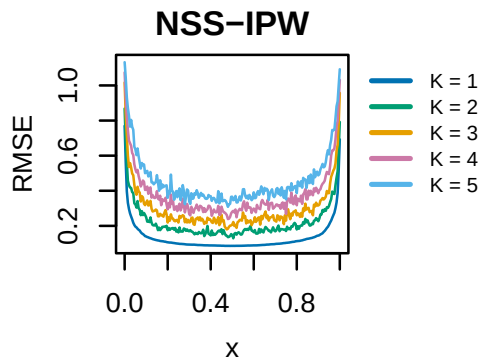
- Full-sample and "honest" decision/causal trees are highly inaccurate uniformly.
- Inconsistency due to variance issue, not to boundary/misspecification bias.
- Inconsistency can occur at *countable* many points on the *entire* support $\mathcal{X}$.
- Tension between prediction (mean square) and causality (pointwise/uniform).
- Non-constant μ have similar problems: e.g., piecewise heterogeneity.

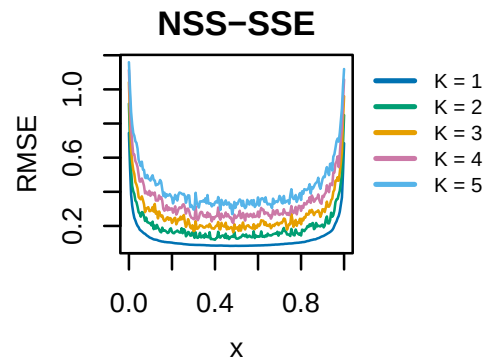


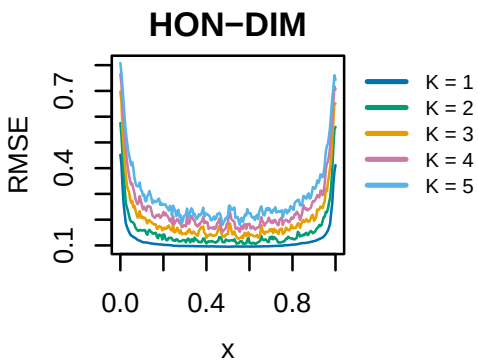
- So-called α**-regularity** fails: cf., random causal forest and related methods.
- Regularization of "small nodes" can lead to large misspecification bias.
- Standard inference methods based on large sample approximations are invalid.

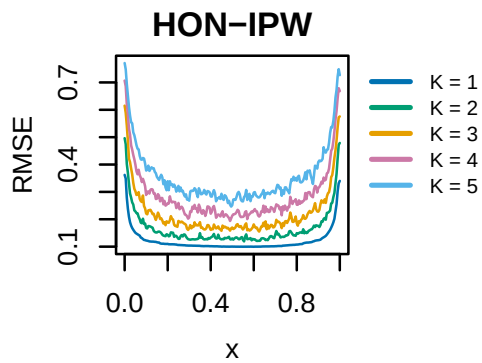
Simulations: Causal Tree Estimators ($p = 1$)

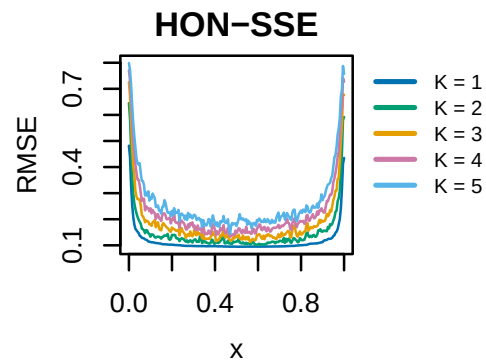


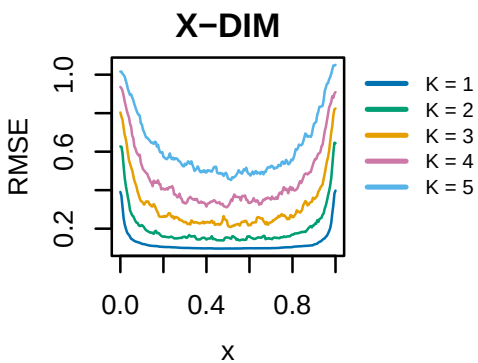


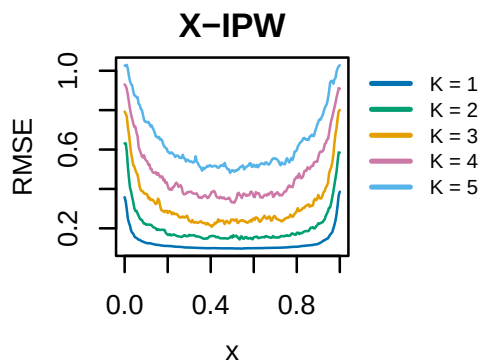


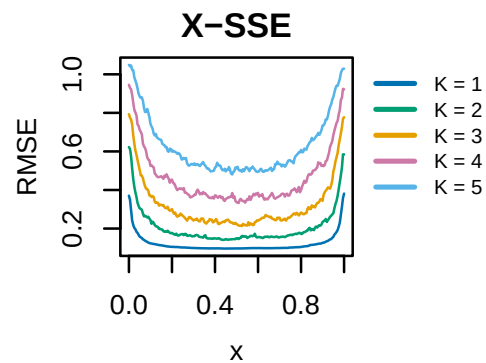












Outline

Conclusion

Recursive partitioning methods are a leading component of the machine learning toolkit.

- Today: three foundational results for Adaptive Decision/Causal Trees.
  1. Highly inaccurate pointwise (hence uniform) convergence, possibly inconsistent.
  2. Sample splitting (so-called "honesty") does not help.
  3. Near-optimal mean square convergence (special case), but sample splitting does not help.
- Adaptive ML methods have advantages and disadvantages.
- Statistical and algorithmic implementations must be studied together.
- Mechanical implementations of machine learning can be detrimental.
- Open question: do other machine learning methods have similar problems?

References

Today:

1. C, Klusowski, Tian & Yu (2025, `arXiv:2509.11381`): "The Honest Truth About Causal Trees: Accuracy Limits for Heterogeneous Treatment Effect Estimation".

Related Work:

1. C, Feng & Shigida (2025, `arXiv:2409.05715`): "Uniform Estimation and Inference for Nonparametric Partitioning-Based M-Estimators".

2. C, Crump, Farrell & Feng (2025, `arXiv:2407.15276`): "Nonlinear Binscatter Methods".

3. C, Klusowski & Underwood (2026, JRRSB): "Estimation and Inference using Mondrian Random Forests".

4. C, Crump, Farrell & Feng (2024, AER): "On Binscatter".

5. C, Chandak & Klusowski (2024, AOS): "Convergence Rates of Oblique Regression Trees for Flexible Function Libraries".

6. C, Farrell & Feng (2020, AOS): "Large Sample Properties of Partitioning-Based Series Estimators".

7. C & Farrell (2013, JOE): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators".