# On the Pointwise Behavior of Recursive Partitioning and Its Implications for Heterogeneous Causal Effect Estimation

Matias D. Cattaneo[1] Jason M. Klusowski[2] Peter M. Tian[3]

September 2024

[1]Princeton University
[2]Princeton University
[3]Two Sigma

# Outline

# Introduction

**Adaptive Decision Trees** are widely used in academia and industry.

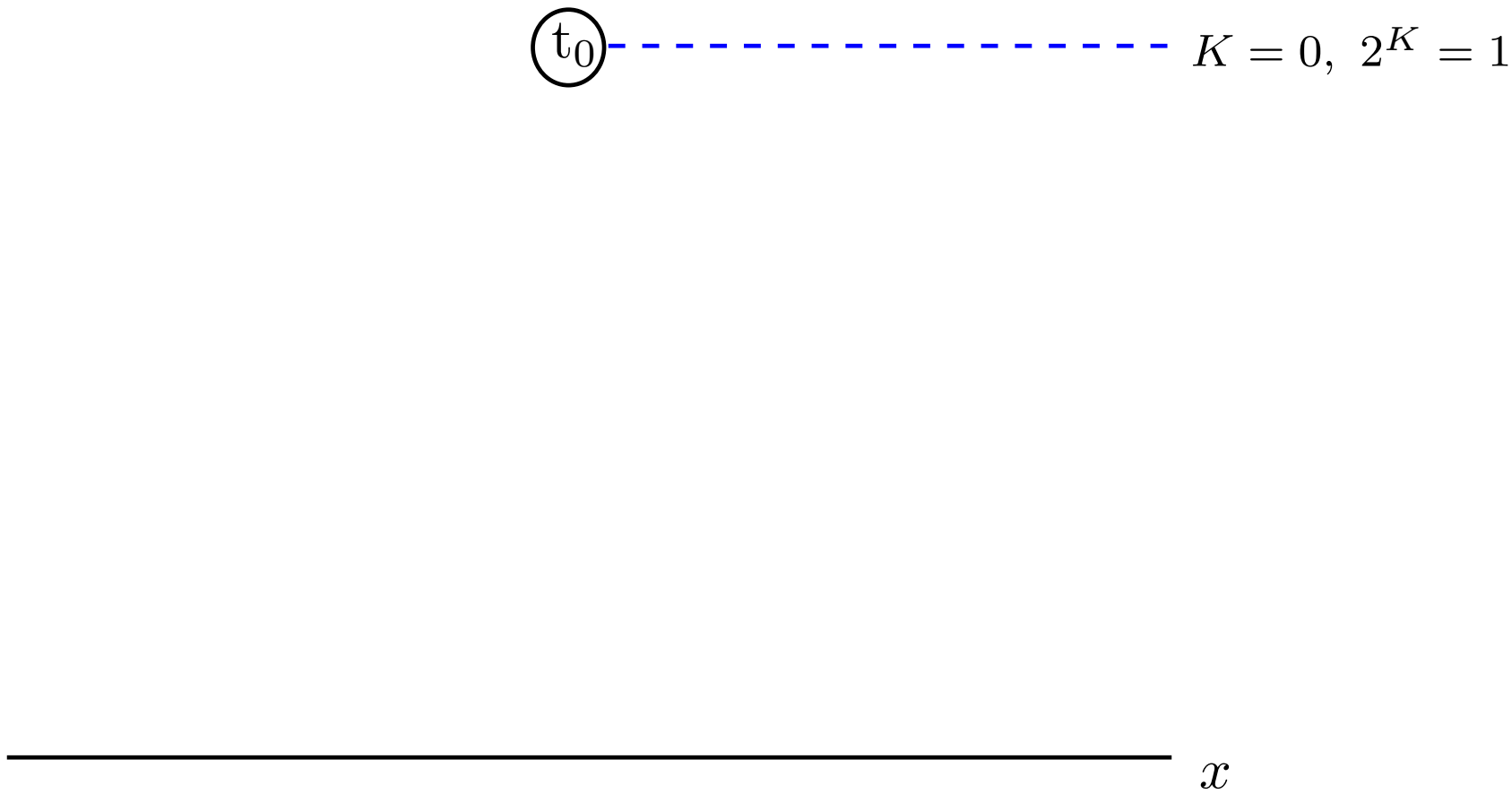
- CART: Breiman, Friedman, Olshen & Stone (1984).
- Adaptivity: incorporate data features in their construction.
- Popularity: prime example of "modern" machine learning toolkit.
- Preferred for interpretability or pointwise learning:

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

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

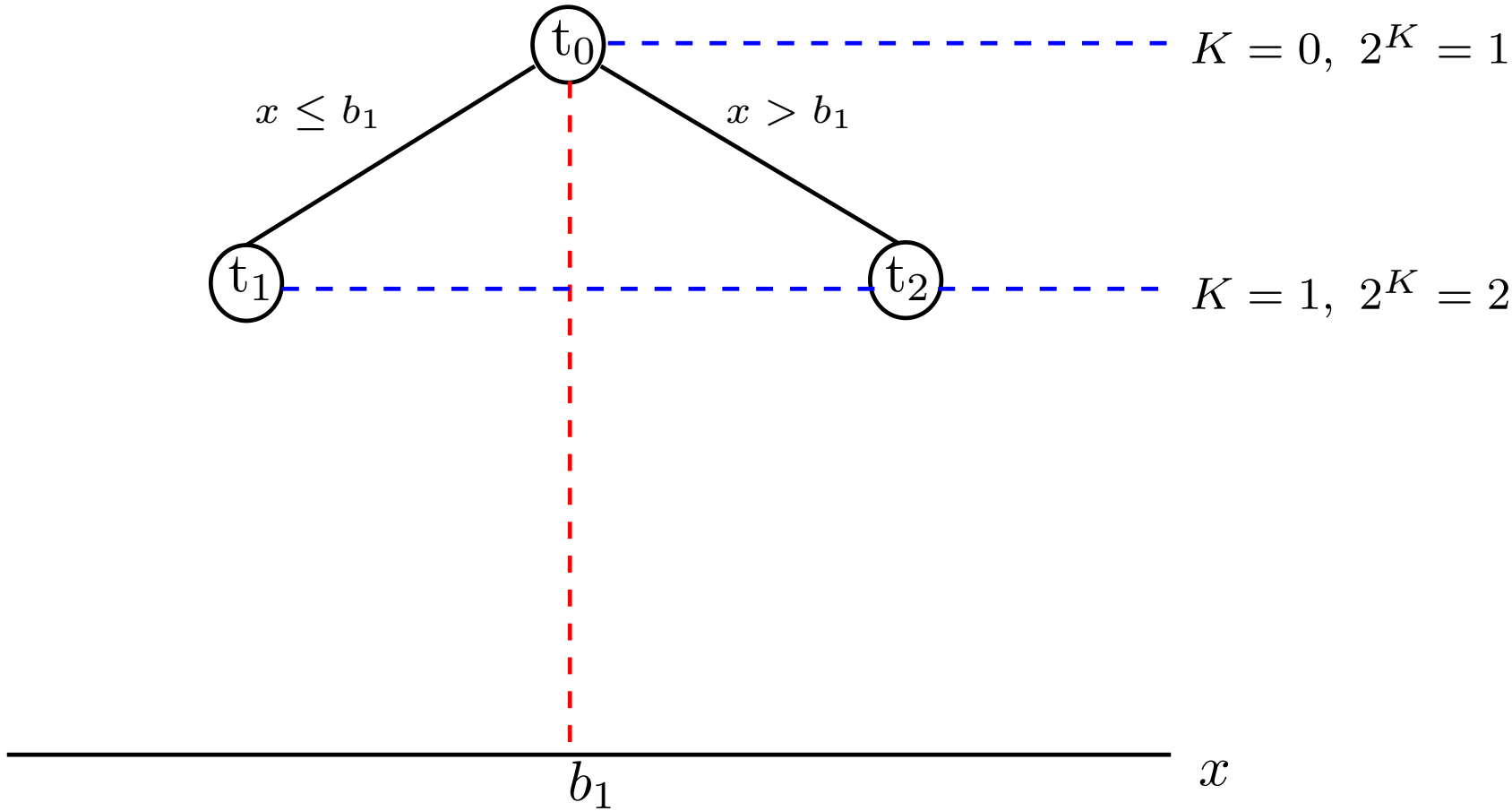
- Today: a foundational result for Adaptive Decision Trees.
  - Axis-aligned: pointwise inconsistent $\implies$ uniformly inconsistent.

# Adaptive Axis-Aligned Decision Tree (CART)



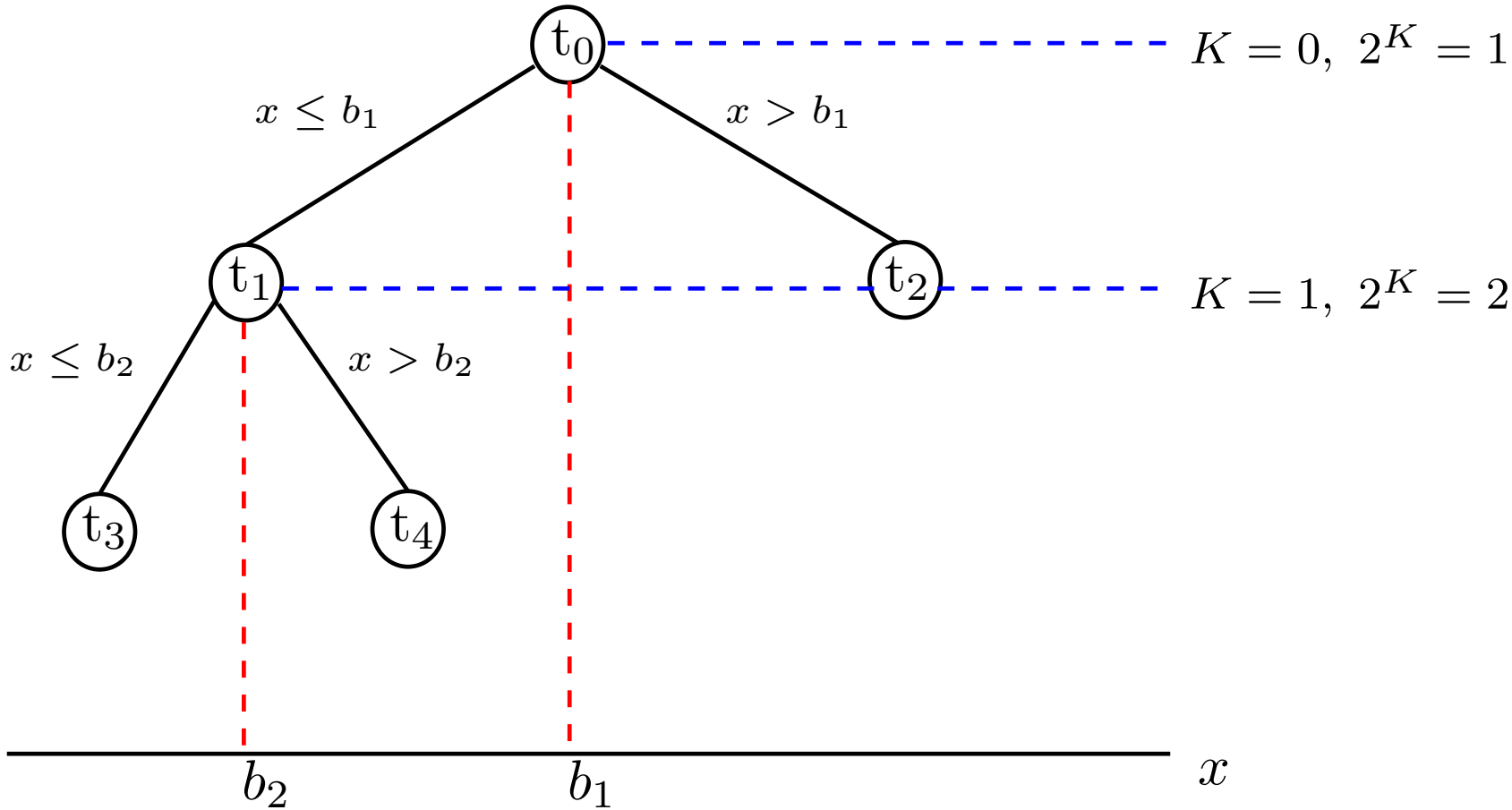

for each $K$ : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathbf{t}} \Big( y_i - \beta_1 \mathbb{1}(x_{ij} \leq \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

# Adaptive Axis-Aligned Decision Tree (CART)



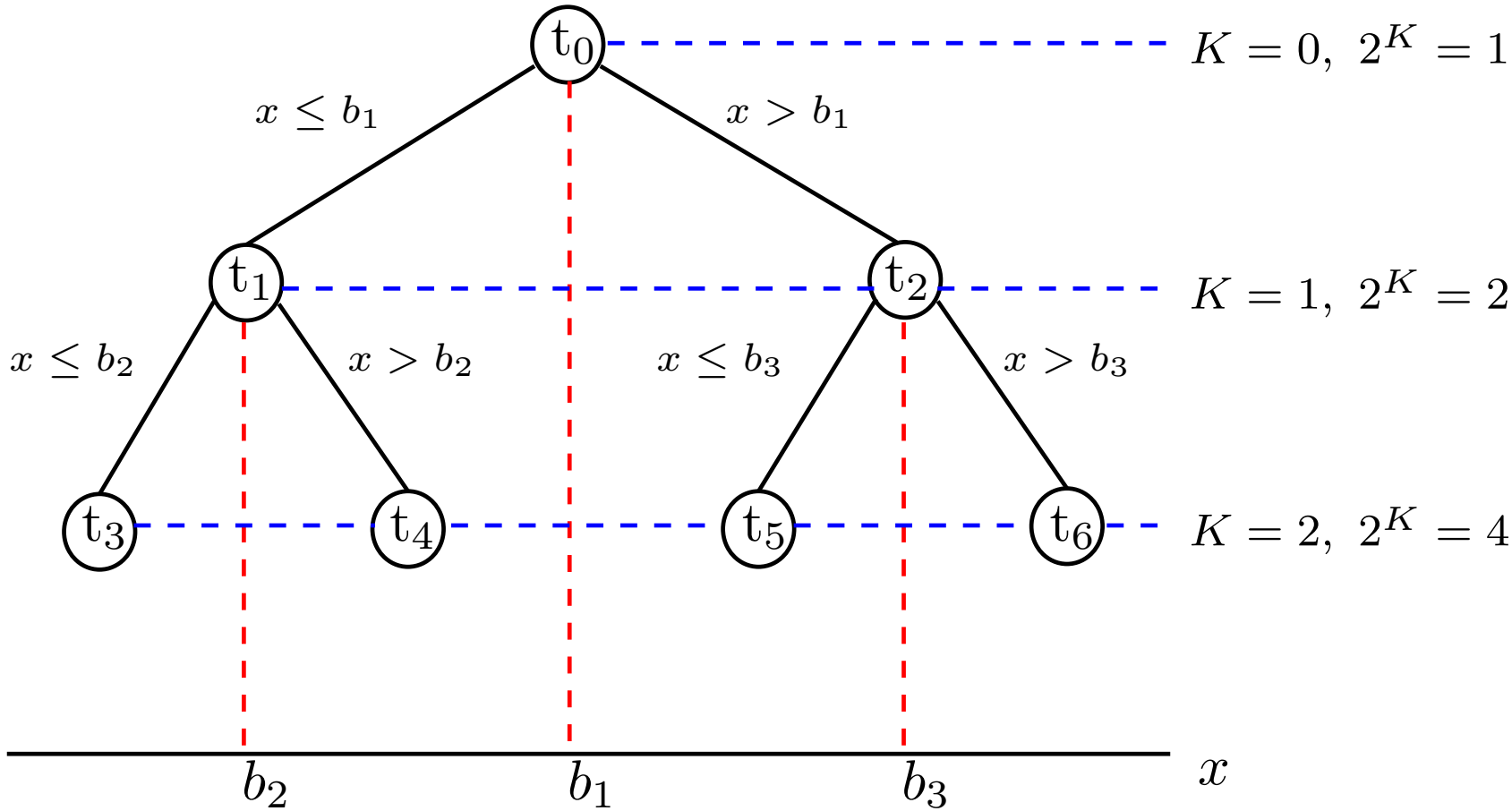

for each $K$ : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathbf{t}} \Big( y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

# Adaptive Axis-Aligned Decision Tree (CART)



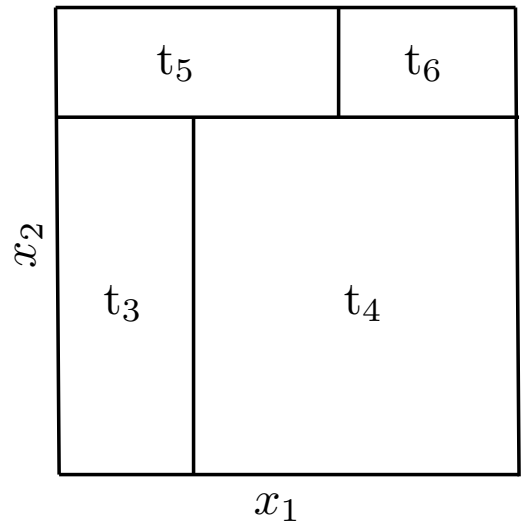

for each $K$ : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathbf{t}} \Big( y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

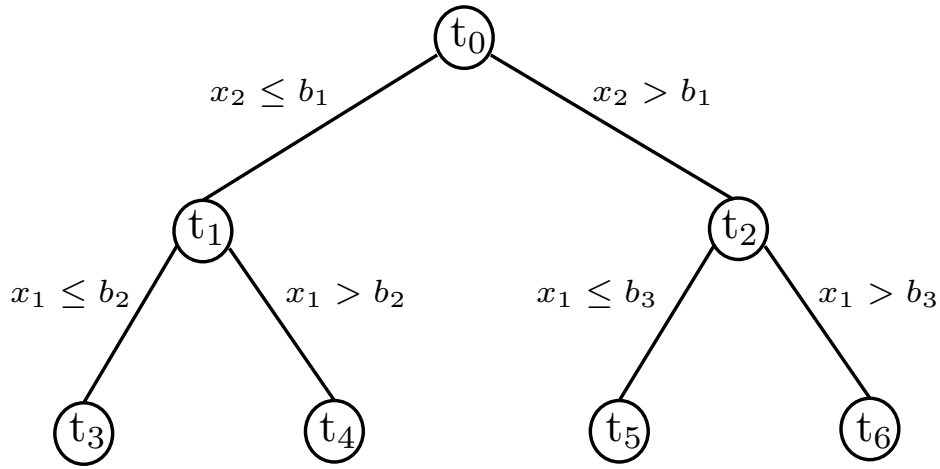
# Adaptive Axis-Aligned Decision Tree (CART)




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

# Adaptive Axis-Aligned Decision Tree (CART)



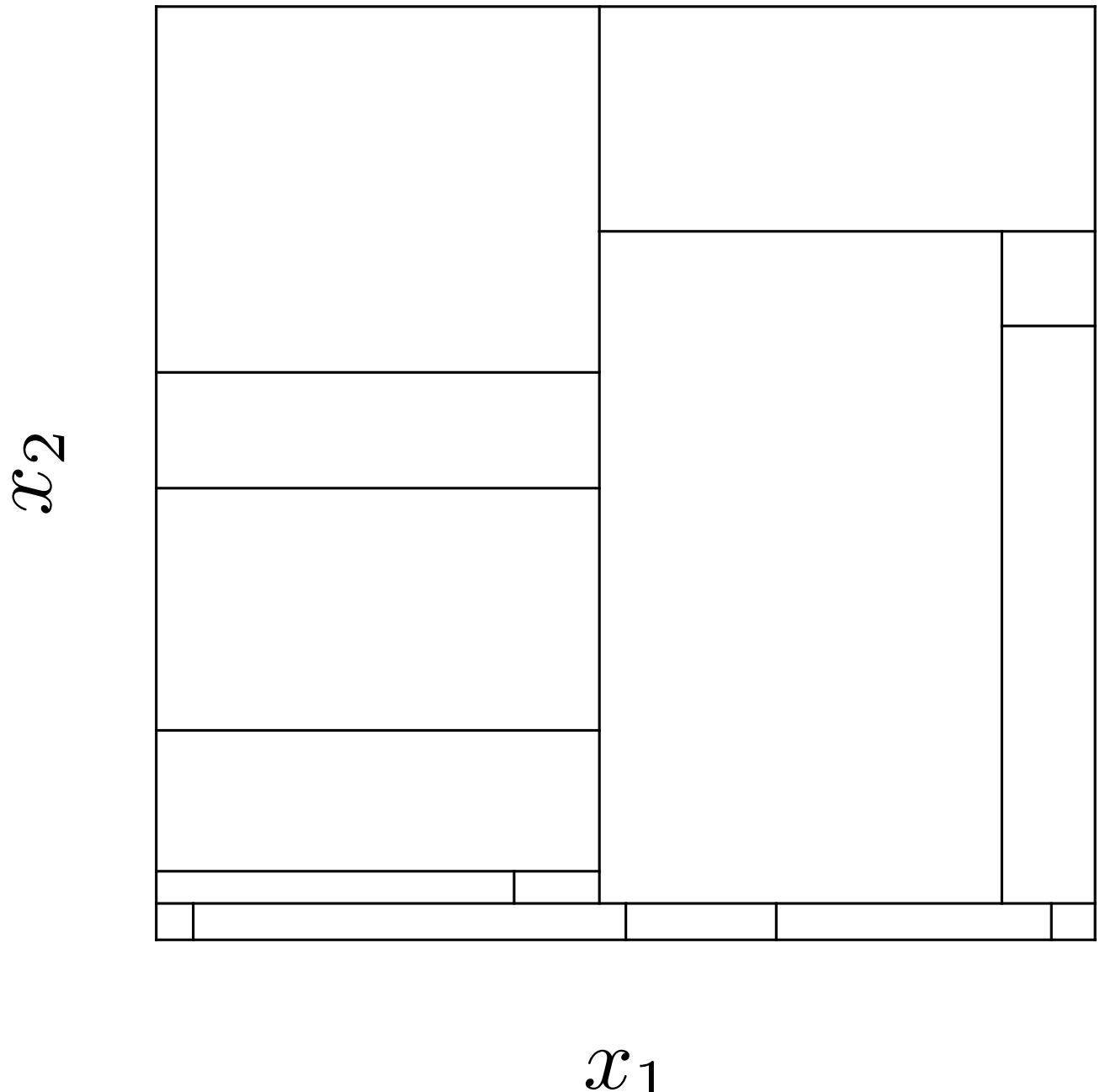



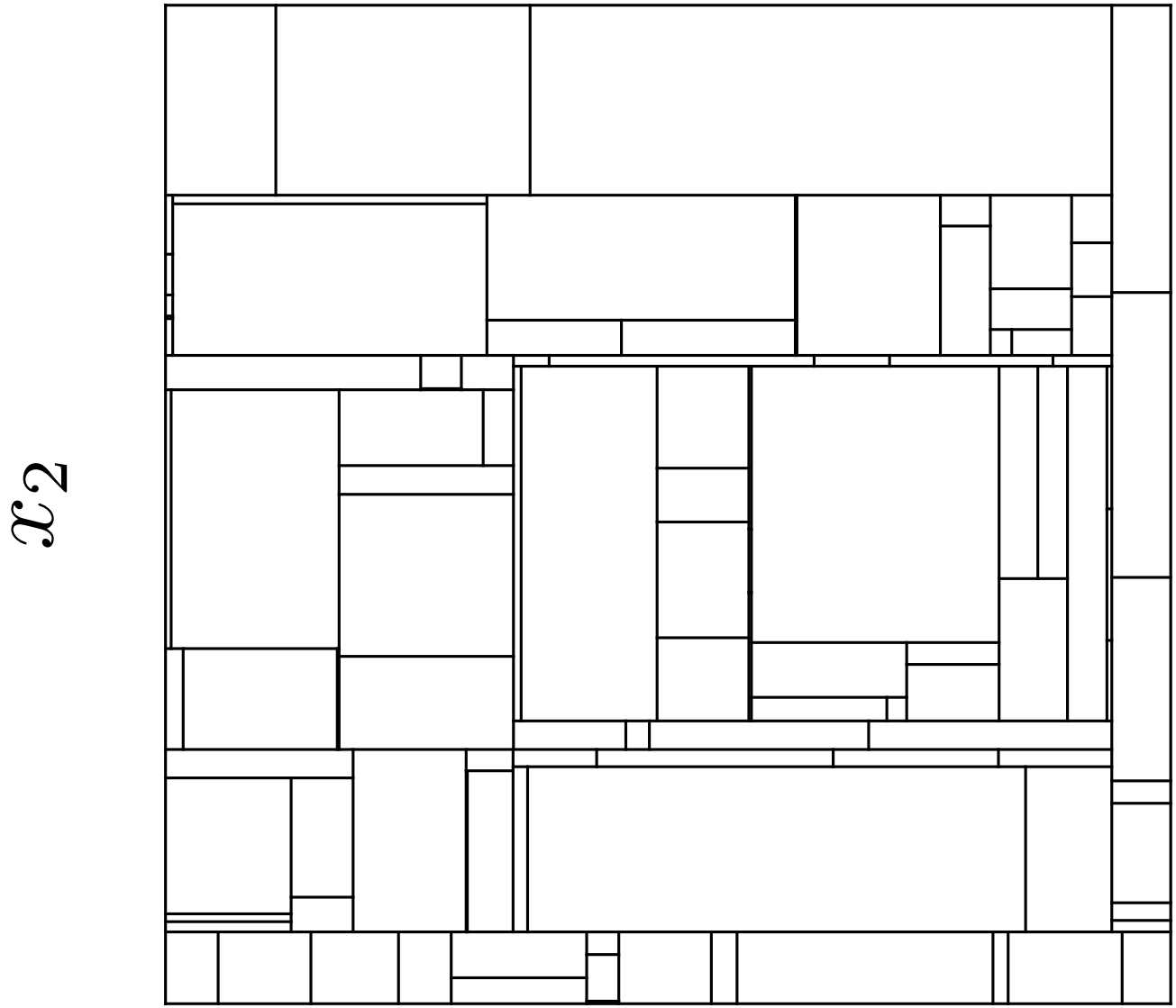

$$\widehat{\mu}(T_K)(\mathbf{x}) = \overline{y}_{\mathbf{t}} = \frac{1}{n(\mathbf{t})} \sum_{\mathbf{x}_i \in \mathbf{t}} y_i, \qquad n(\mathbf{t}) = \sum_{\mathbf{x}_i \in \mathbf{t}} \mathbb{1}(\mathbf{x}_i \in \mathbf{t}).$$

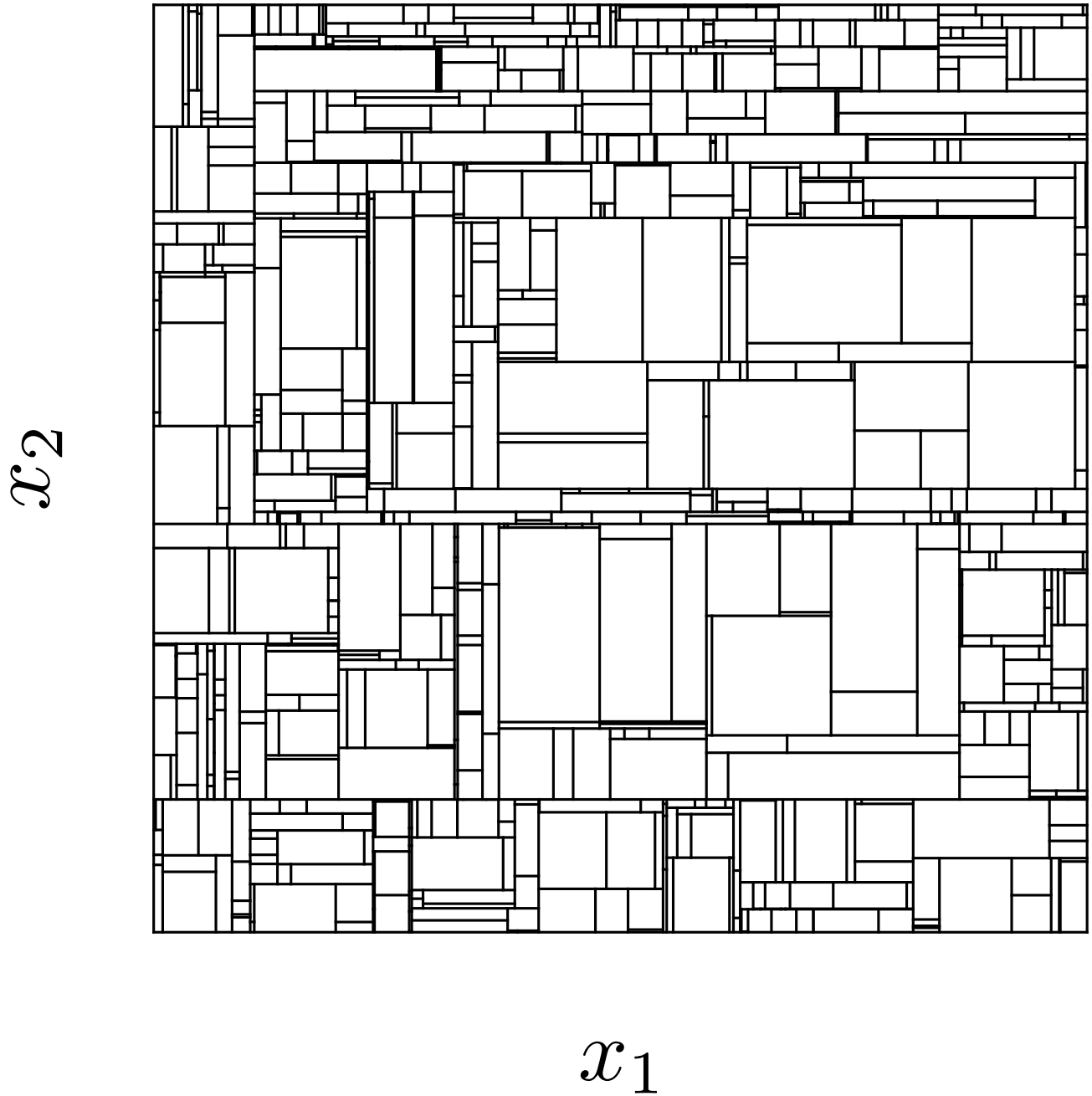
**CKT (2024)**: for "honest" trees and $\mu(\mathbf{x}) = \mu$,

$$\mathbb{P}\Big( \sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\mu}(T_K)(\mathbf{x}) - \mu| > C \Big) > C^2 \qquad \text{if} \quad K \gtrsim \log\log(n),$$

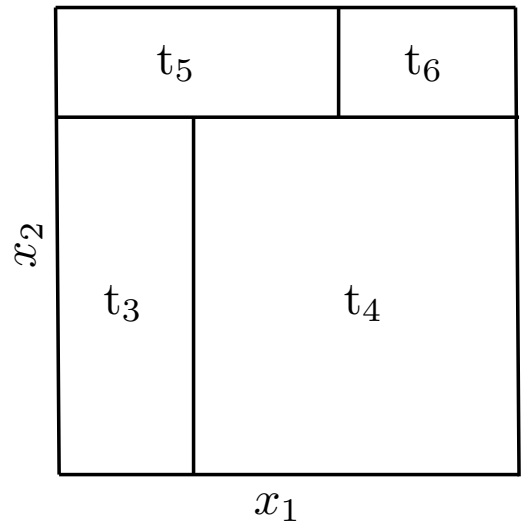
$$\mathbb{E}\big[ \|\widehat{\mu}(T_K) - \mu\|^2 \big] = \mathbb{E}\Big[ \int_{\mathcal{X}} (\widehat{\mu}(T_K)(\mathbf{x}) - \mu)^2 \mathbb{P}_{\mathbf{x}}(d\mathbf{x}) \Big] \le \frac{2^{K+1}\sigma^2}{n+1}.$$

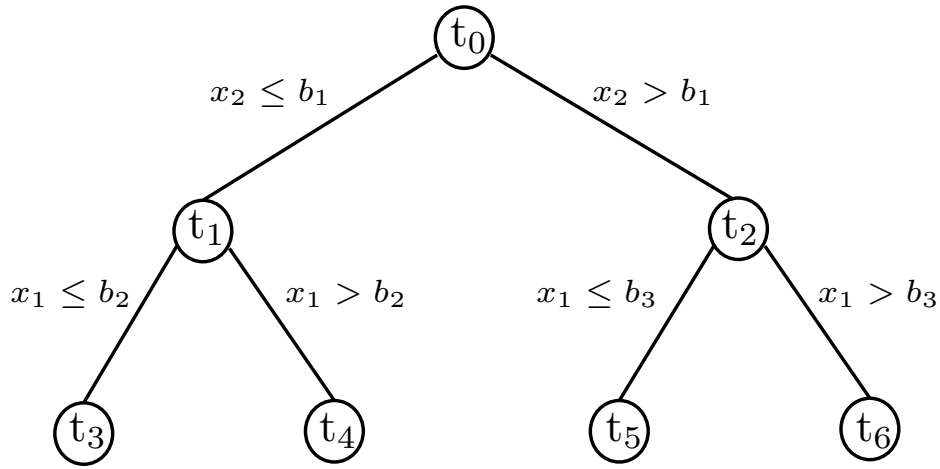

$x_2$
$x_1$


$x_2$
$x_1$


$x_2$
$x_1$

# Adaptive Axis-Aligned Decision Tree (CART)




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

**Main Result**: for "honest" trees and $\mu(\mathbf{x}) = \mu$,

$$\mathbb{P}\Big( \sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\mu}(T_K)(\mathbf{x}) - \mu| > C \Big) > C^2 \qquad \text{if} \quad K \gtrsim \log\log(n),$$

$$\mathbb{E}\big[ \|\widehat{\mu}(T_K) - \mu\|^2 \big] = \mathbb{E}\Big[ \int_{\mathcal{X}} (\widehat{\mu}(T_K)(\mathbf{x}) - \mu)^2 \mathbb{P}_{\mathbf{x}}(d\mathbf{x}) \Big] \leq \frac{2^{K+1}\sigma^2}{n+1}.$$

# Outline

# Decision Trees for Heterogeneous Causal Effects

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

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

- Our paper challenges this claim.

# Motivation: Heterogeneous TE, Policy Decisions, Design RCTs, etc.

- $\{(y_i, \mathbf{x}_i', d_i) : i = 1, 2, \ldots, n\}$ i.i.d., and $y_i = y_i(1) \cdot d_i + y_i(0) \cdot (1 - d_i)$.
- RCT: $(y_i(0), y_i(1), \mathbf{x}_i^T) \perp\!\!\!\perp d_i$ and $\xi = \mathbb{P}(d_i = 1) \in (0, 1)$, so

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

"Honest" Causal Decision Trees (Athey and Imbens, 2019):

- Regression-based heterogeneity discovery:

$$\widehat{\tau}_{\texttt{REG}}(T_K)(\mathbf{x}) = \frac{1}{\#\{\mathbf{x}_i \in \mathbf{t} : d_i = 1\}} \sum_{\mathbf{x}_i \in \mathbf{t}: d_i = 1} y_i - \frac{1}{\#\{\mathbf{x}_i \in \mathbf{t} : d_i = 0\}} \sum_{\mathbf{x}_i \in \mathbf{t}: d_i = 0} y_i$$

- IPW-based heterogeneity discovery:

$$\widehat{\tau}_{\texttt{IPW}}(T_K)(\mathbf{x}) = \frac{1}{\#\{\mathbf{x}_i \in \mathbf{t}\}} \sum_{\mathbf{x}_i \in \mathbf{t}} y_i \frac{d_i - \xi}{\xi(1-\xi)}$$

- Adaptive tree $T_K$ with sample splitting, and $\mathbf{t}$ denotes the unique (terminal) node containing $\mathbf{x} \in \mathcal{X}$.

# Setup: Constant (Treatment Effect/Regression) Model

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

The following conditions hold.

1. $(y_i, \mathbf{x}_i')$, $i = 1, 2, \ldots, n$, is a random sample.
2. $\mu(\mathbf{x}) \equiv \mu$ is constant for all $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$.
3. $\mathbf{x}_i$ has a continuous distribution.
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 $\nu > 0$.

**CKT (2024)**: axis-aligned adaptive (CART) decision trees.

1. Decision stumps ($K = 1$) split with high probability "near" the boundaries.
2. $\widehat{\mu}(T_1)(\mathbf{x})$ has at best polylog($n$) convergence rate near boundaries.
3. "Honest" $\widehat{\mu}(T_K)(\mathbf{x})$ are uniformly inconsistent as soon as $K \gtrsim \log\log(n)$.
   - $n = 1$ billion implies depth $\log\log(n) \approx 3$.
   - Inconsistency occurs at countable many points on support, not just at boundaries.
4. Pruning does not solve the inconsistency; other regularization requires care...

# Decision Stumps: polylog($n$) Convergence Rate Near Boundaries

Recall: for each level $K$, adaptive (CART) decision trees solve

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

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

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

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

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

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

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

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

# Decision Stumps: polylog($n$) Convergence Rate Near Boundaries

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

## Theorem

*Suppose $p = 1$. Let $\hat{\mu}(T_1)(x)$ be the CART estimator of the regression function at the root node. For any $a, b \in (0,1)$ with $a < b$, we have*

$$\liminf_{n\to\infty} \inf_{x\in\mathcal{X}_n} \mathbb{P}\Big(|\hat{\mu}(T_1)(x) - \mu| \geq \sigma n^{-b/2}\sqrt{(2+o(1))\log\log(n)}\Big) \geq \frac{b-a}{e},$$

*where $\mathcal{X}_n = [0,\ (1+o(1))n^{a-1}) \cup (1-(1+o(1))n^{a-1},\ 1]$.*

- Decision stumps cannot converge at a polynomial rate, i.e., its rate is slower than any polynomial-in-$n$.
- With arbitrary high probability, split index $\hat{\imath}$ will concentrate near its extremes, from the beginning of any tree construction.
- The first split generates cell containing, at most, $\log^a(n)$ observations, with probability at least $(\log(n))^{-b}$, up to constant factors.
- 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 $\mu$.

# "Honest" (Decision/Causal) Trees: Uniform Inconsistency

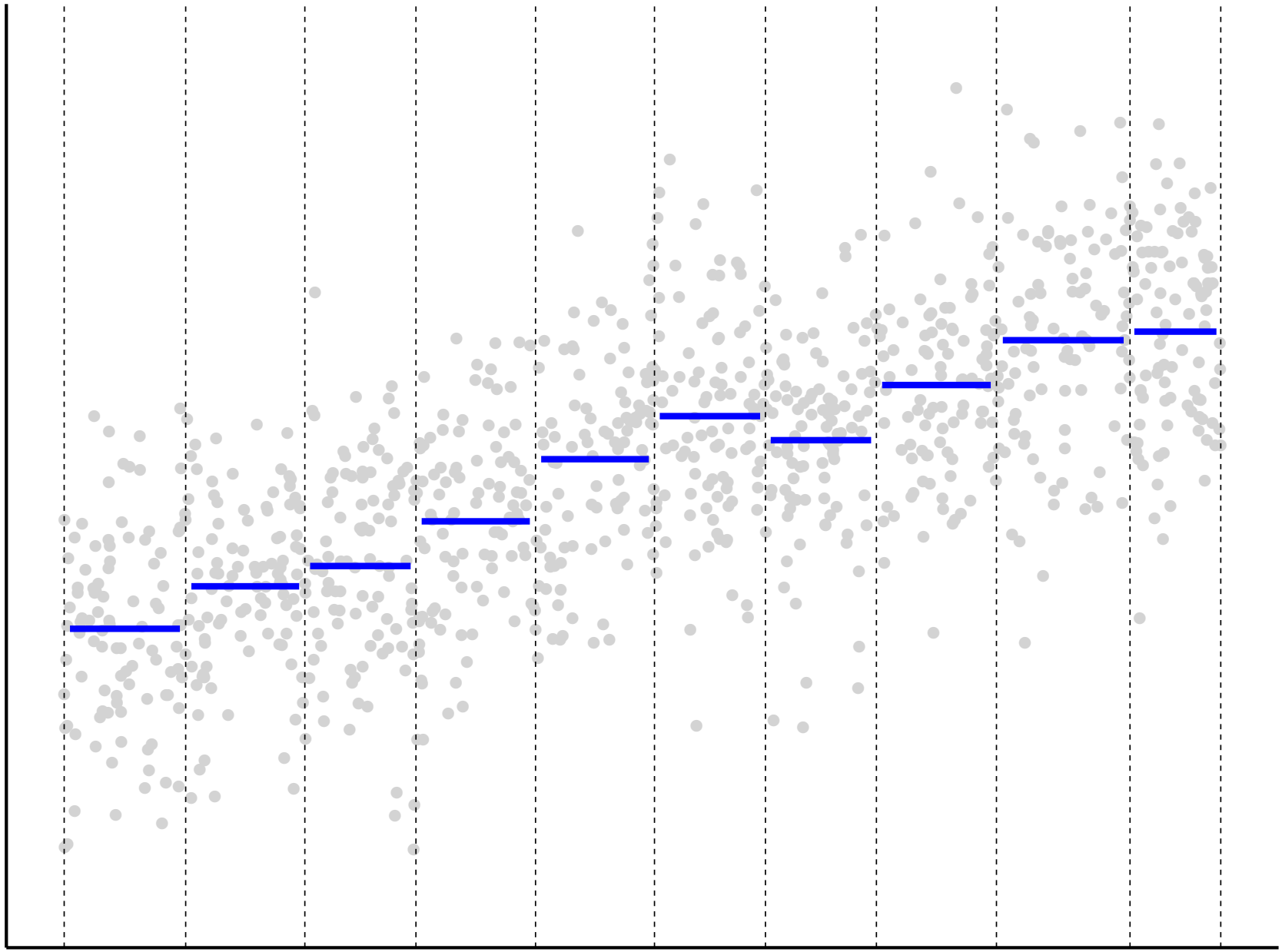
Iterating nearly inconsistent decision stumps can only make things worse... Thus, employing "honesty" (i.e., sample splitting), we have:

**Theorem**

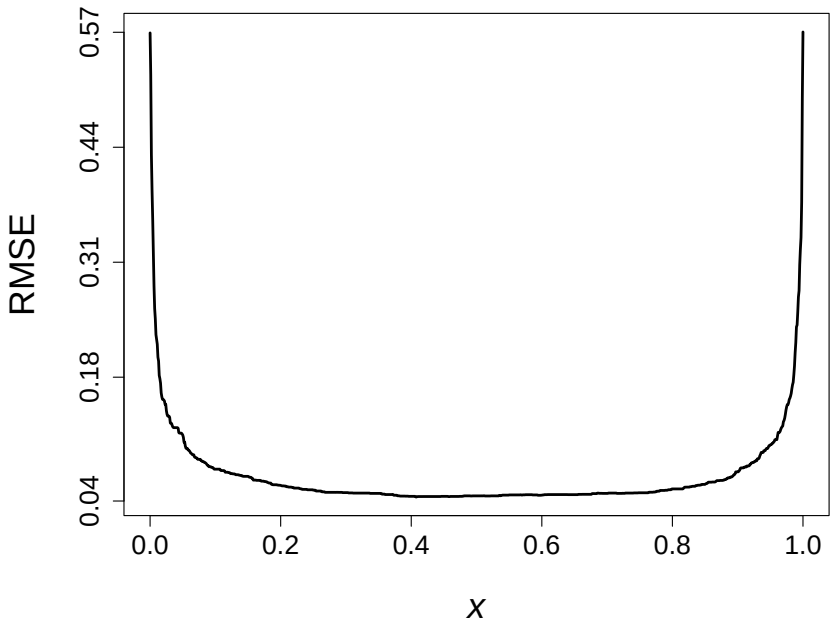
*Suppose $p = 1$. Consider a maximal depth $K_n \gtrsim \log\log(n)$ tree $T_{K_n}$ constructed with CART+ methodology. Then, there exists a positive constant $C$ such that*

$$\liminf_{n\to\infty} \mathbb{P}\left( \sup_{x\in\mathcal{X}} |\tilde{\mu}(T_{K_n})(x) - \mu| > C \right) > 0.$$

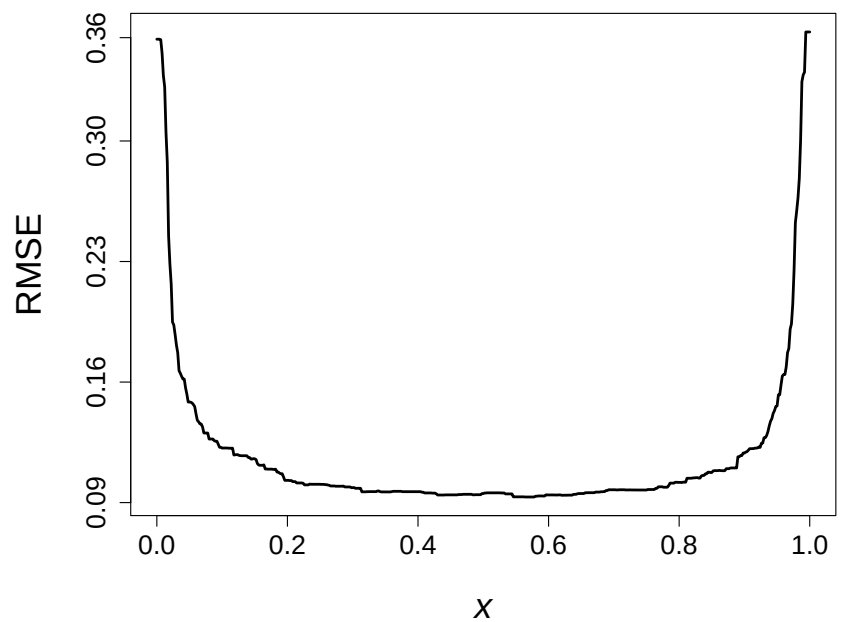
- Shallow "Honest" decision/causal trees are uniformly inconsistent.
- Inconsistency due to variance issue, not to boundary/misspecification bias.
- Inconsistency can occur at *countable* many points on the *entire* support $\mathcal{X}$.
- Pruning does not mitigate the inconsistency.
- Non-constant $\mu$ have similar problems: e.g., piecewise heterogeneity.



# Simulations: Decision Stumps ($K = 1$) for Constant (Treatment) Model

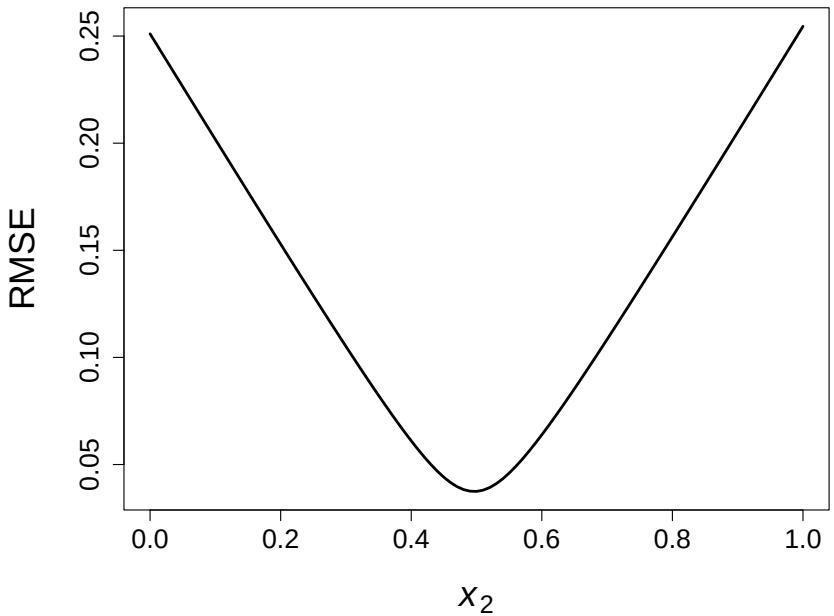



(a) Pointwise RMSE of decision stump.

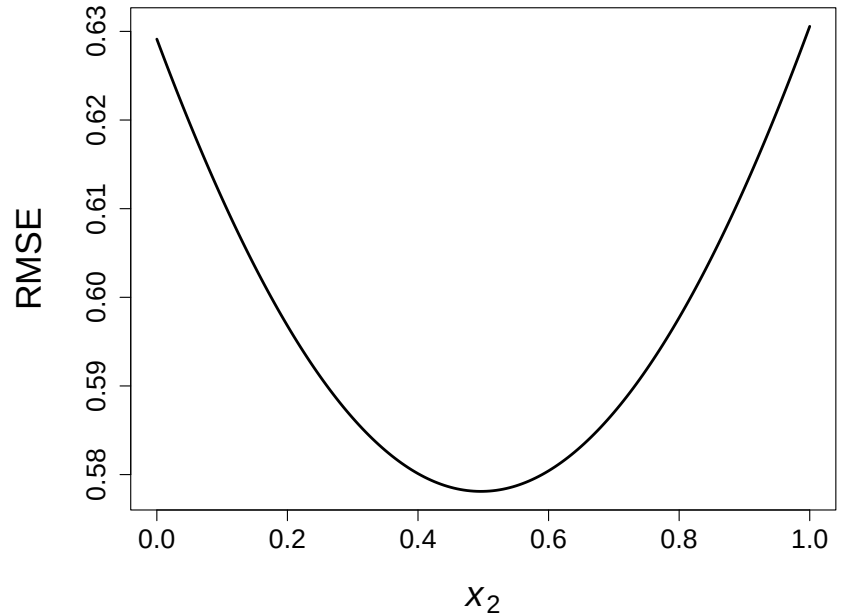



(b) Pointwise RMSE of causal decision stump.

# Simulations: Decision Stumps ($K = 1$) with Pruning




(a) Pointwise RMSE for pruned tree at $\mathbf{x} = (0, x_2)^T$.




(b) Pointwise RMSE for pruned causal tree at $\mathbf{x} = (0, x_2)^T$.

# Outline

# Takeaways

**Adaptive Decision Trees** are a leading component of the machine learning toolkit.

- Today: two foundational results for Adaptive Decision Trees.
  - Axis-aligned: pointwise inconsistent $\implies$ uniformly inconsistent.
- 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?

# Related References

## **Today**:


1. C, Klusowski & Tian (2024): "On the Pointwise Behavior of Recursive Partitioning and Its Implications for Heterogeneous Causal Effect Estimation", `arXiv:2211.10805`.


## **Related Work**:


1. C, Feng & Shigida (2024): "Uniform Estimation and Inference for Nonparametric Partitioning-Based M-Estimators", `arXiv:2409.05715`.

2. C, Crump, Farrell & Feng (2024): "Nonlinear Binscatter Methods", `arXiv:2407.15276`.

3. C, Klusowski & Underwood (2024): "Estimation and Inference using Mondrian Random Forests", `arXiv:2310.09702`.

4. C, Crump, Farrell & Feng (2024): "On Binscatter", *American Economic Review*.

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

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

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