Adaptive Decision Tree Methods

Matias D. Cattaneo

Princeton University

March 2024

Talk based on:

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

• Cattaneo, Chandak & Klusowski (2024, CCK): "Convergence Rates of Oblique Regression Trees for Flexible Function Libraries", *Annals of Statistics*, forthcoming.

Outline

Introduction

Adaptive Decision Trees are widely used in academia and industry.

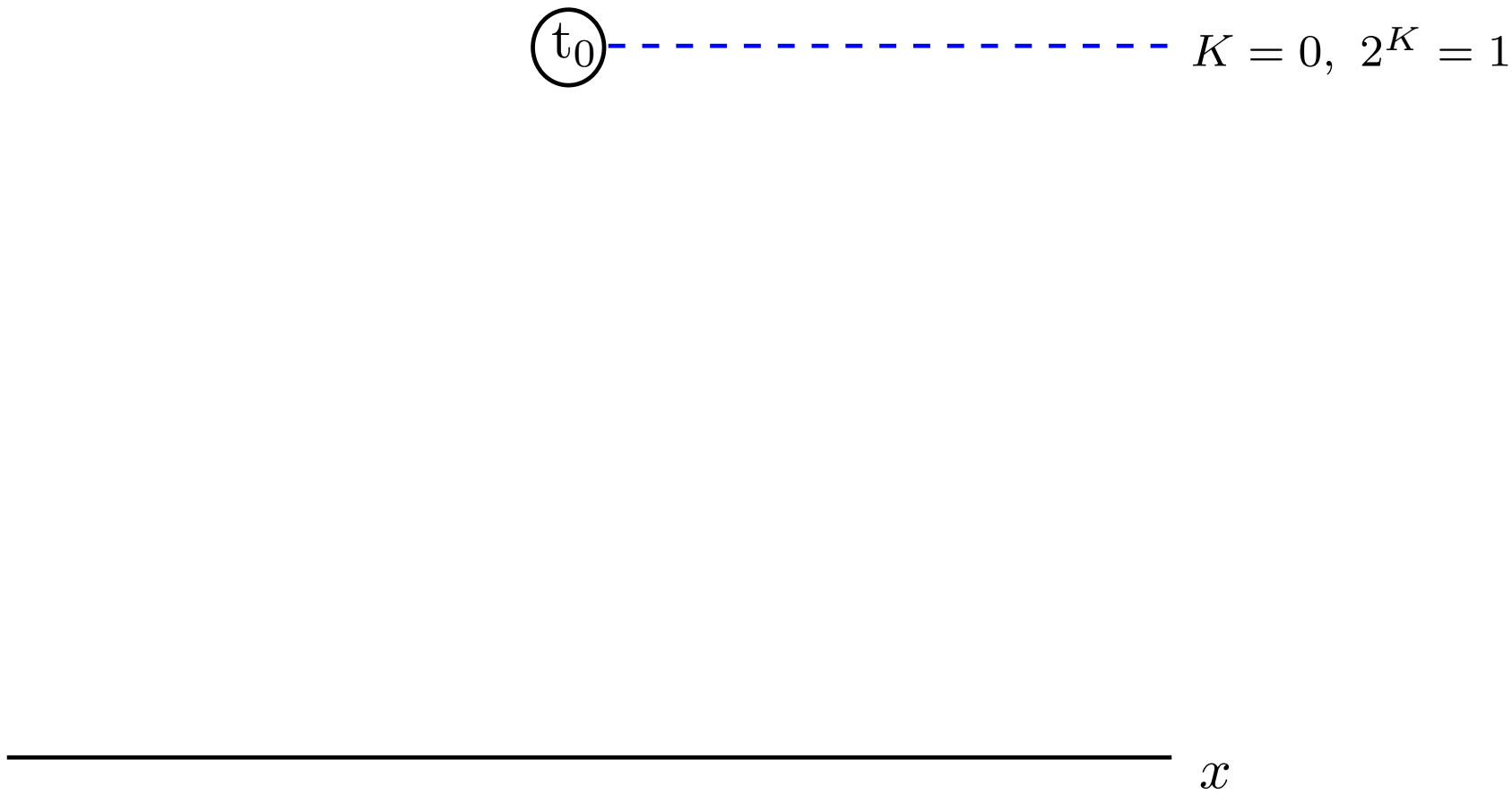
- CART: Breiman, Friedman, Olshen & Stone (1984).
- Adaptivity: incorporate data features in their construction.
- Popularity: prime example of "modern" machine learning toolkit.
- Preferred for interpretability or pointwise learning:

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

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

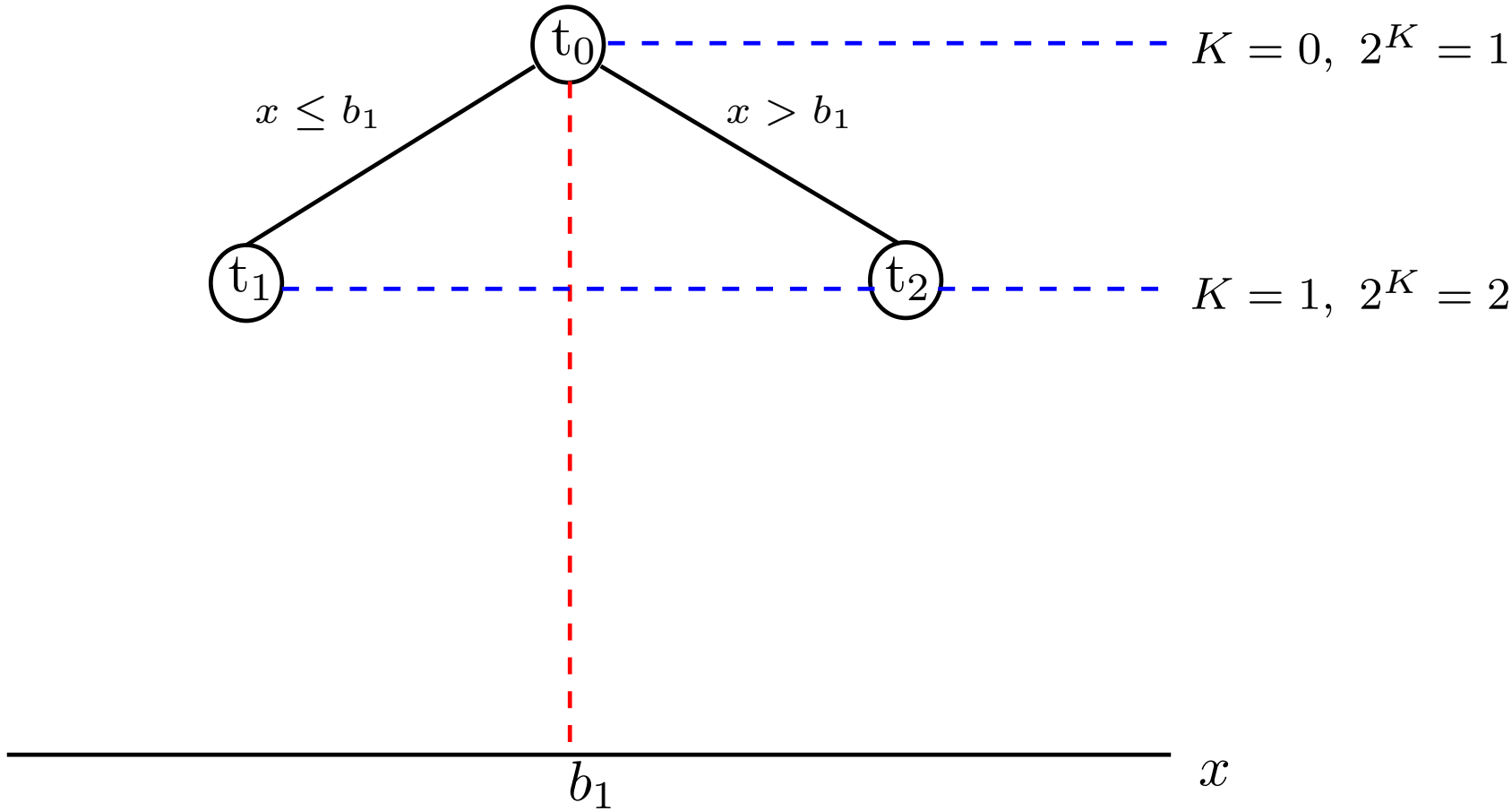
- Today: two foundational results for Adaptive Decision Trees.
  - Axis-aligned: pointwise inconsistent $\implies$ uniformly inconsistent.
  - Oblique: mean square consistent $\iff$ Single-hidden layer NN performance.

Adaptive Axis-Aligned Decision Tree (CART)



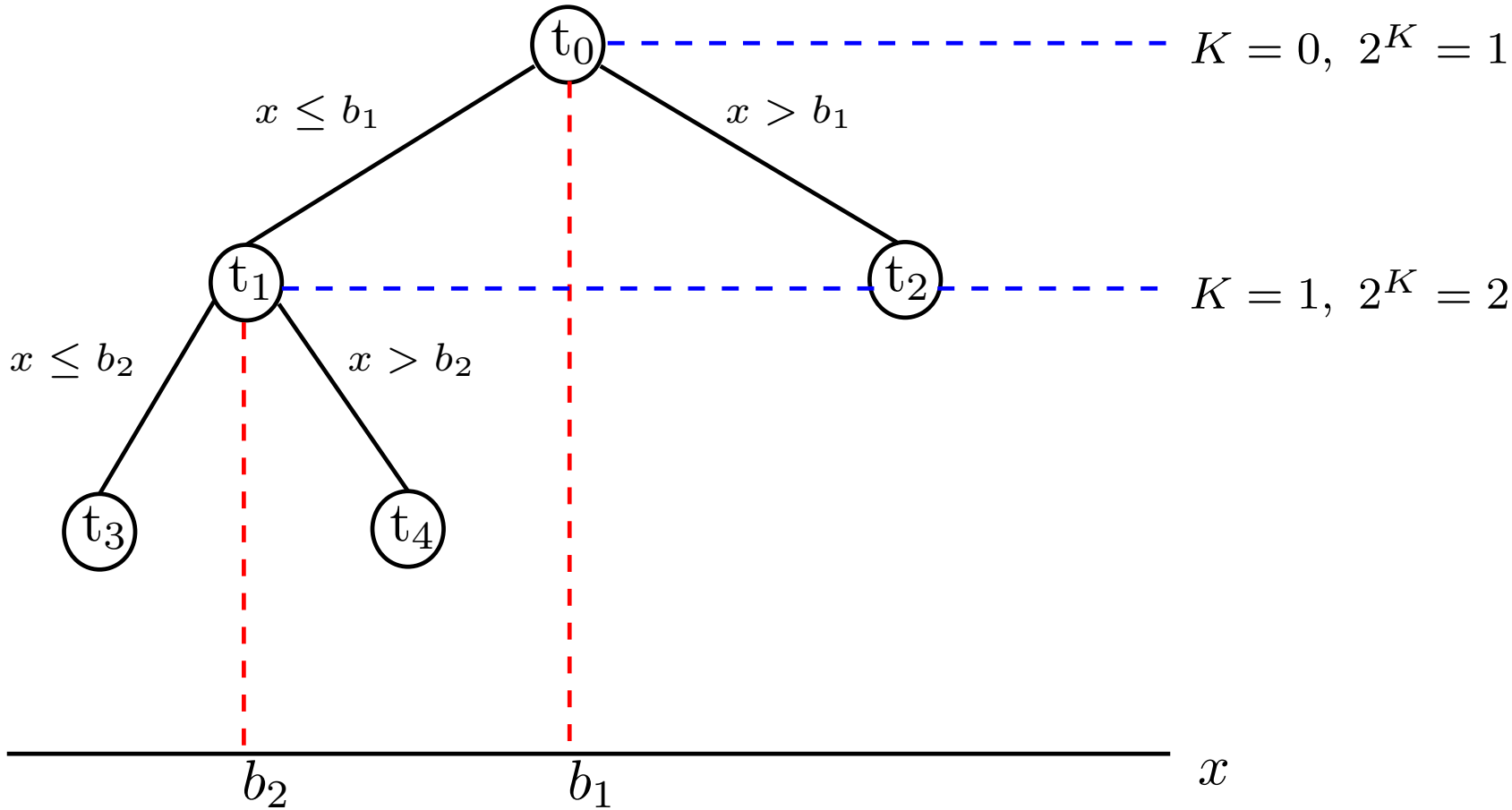
for each K : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathbf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \leq \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

Adaptive Axis-Aligned Decision Tree (CART)



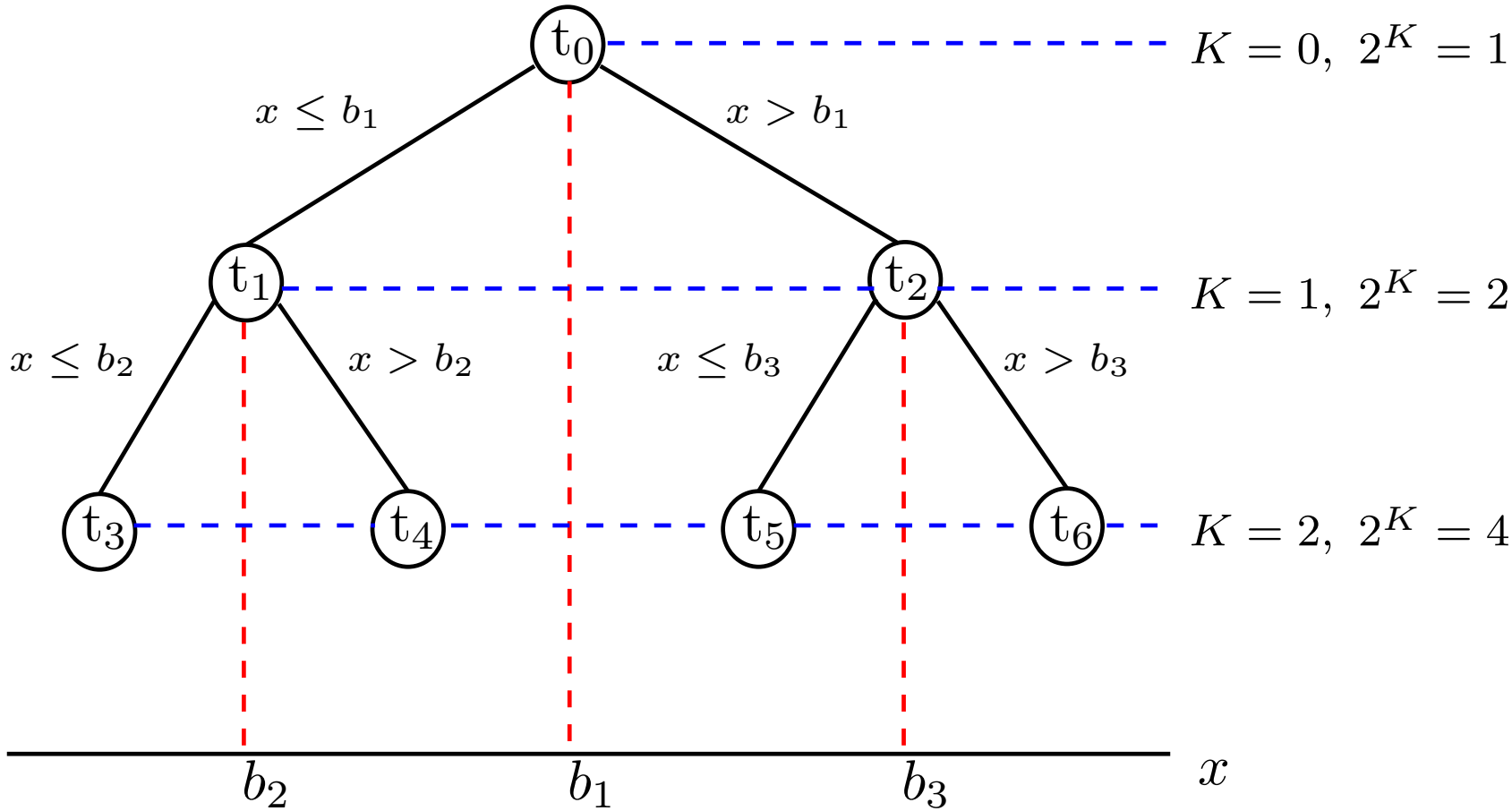
for each K : $$\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathbf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \le \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$$

Adaptive Axis-Aligned Decision Tree (CART)



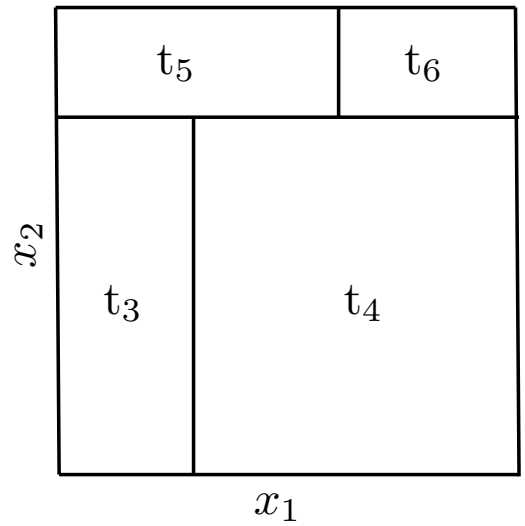
for each K : $\min_{j=1,2,\cdots,p} \min_{\beta_1,\beta_2,\tau} \sum_{\mathbf{x}_i \in \mathbf{t}} \Big(y_i - \beta_1 \mathbb{1}(x_{ij} \leq \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau) \Big)^2$

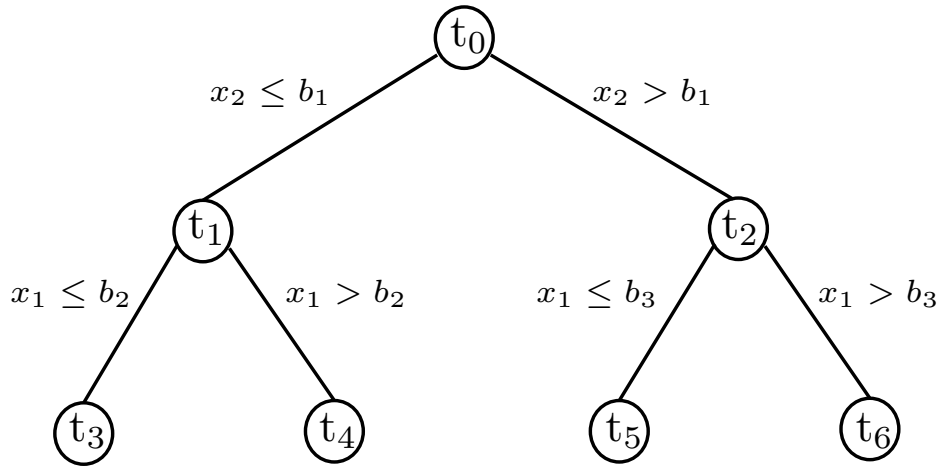
Adaptive Axis-Aligned Decision Tree (CART)



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)



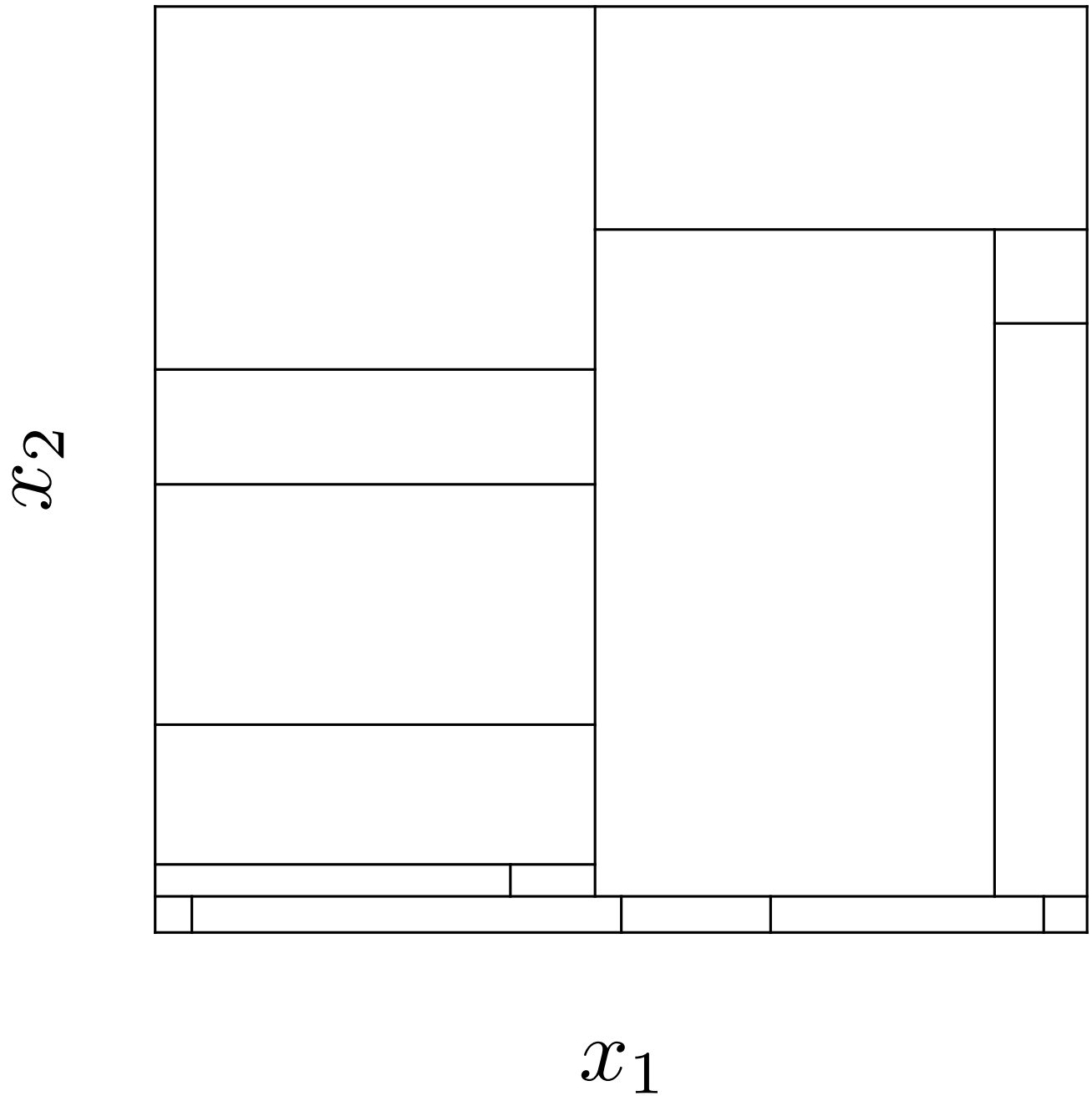


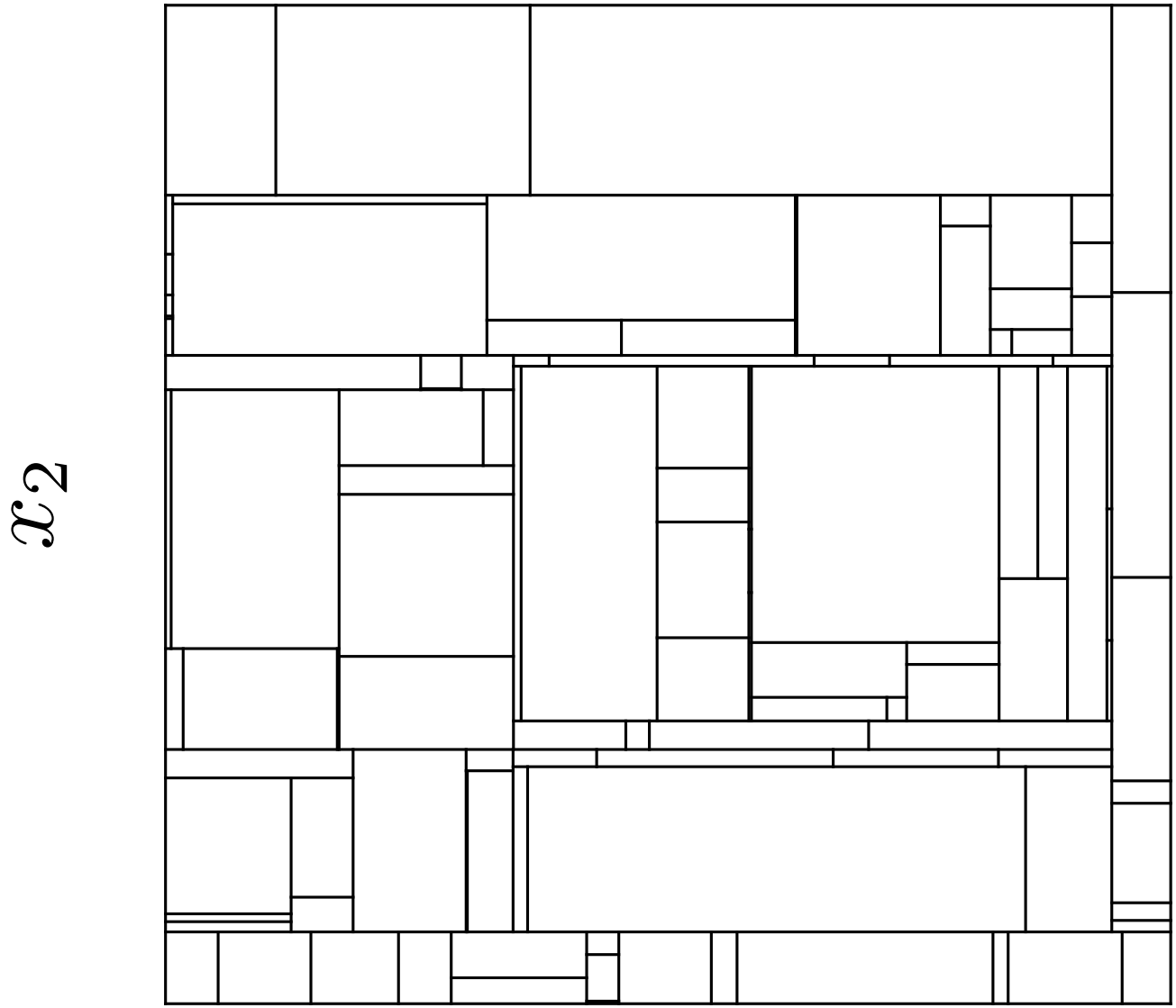
$$\widehat{\mu}(T_K)(\mathbf{x}) = \overline{y}_{\mathbf{t}} = \frac{1}{n(\mathbf{t})} \sum_{\mathbf{x}_i \in \mathbf{t}} y_i, \qquad n(\mathbf{t}) = \sum_{\mathbf{x}_i \in \mathbf{t}} \mathbb{1}(\mathbf{x}_i \in \mathbf{t}).$$

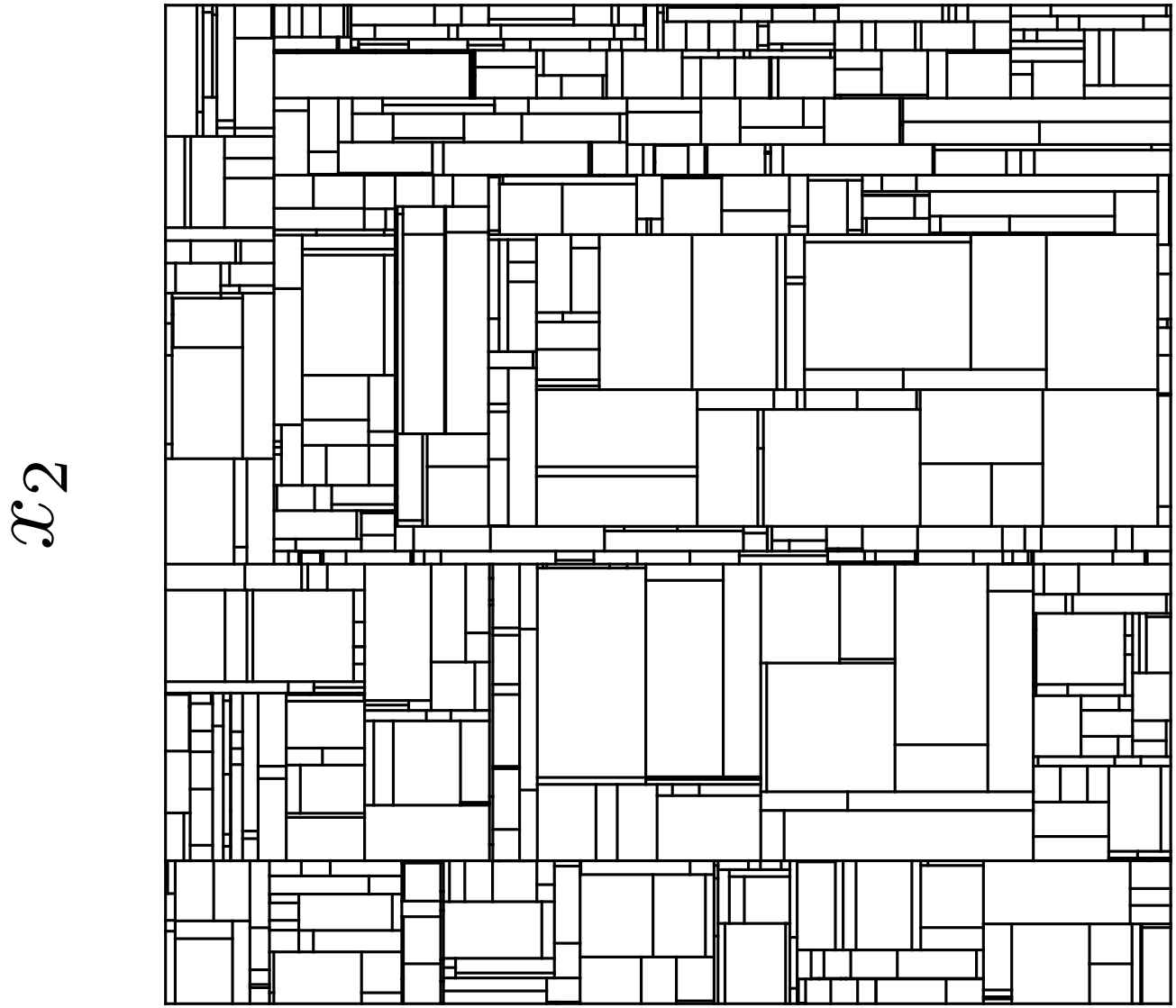
CKT (2024): for "honest" trees and $\mu(\mathbf{x}) = \mu$,

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

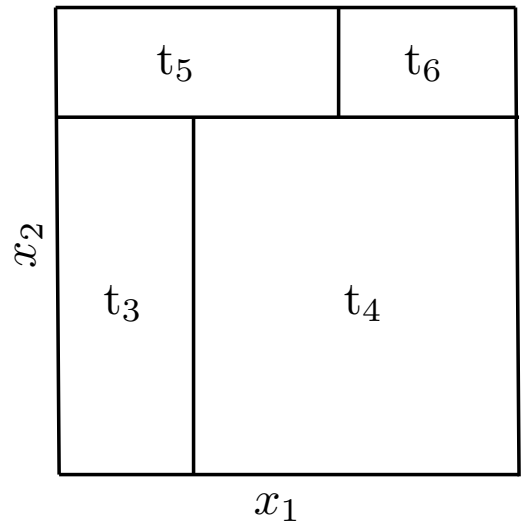
$$\mathbb{E}\big[\|\widehat{\mu}(T_K) - \mu\|^2 \big] = \mathbb{E}\Big[\int_{\mathcal{X}} (\widehat{\mu}(T_K)(\mathbf{x}) - \mu)^2 \mathbb{P}_{\mathbf{x}}(d\mathbf{x}) \Big] \le \frac{2^{K+1}\sigma^2}{n+1}.$$

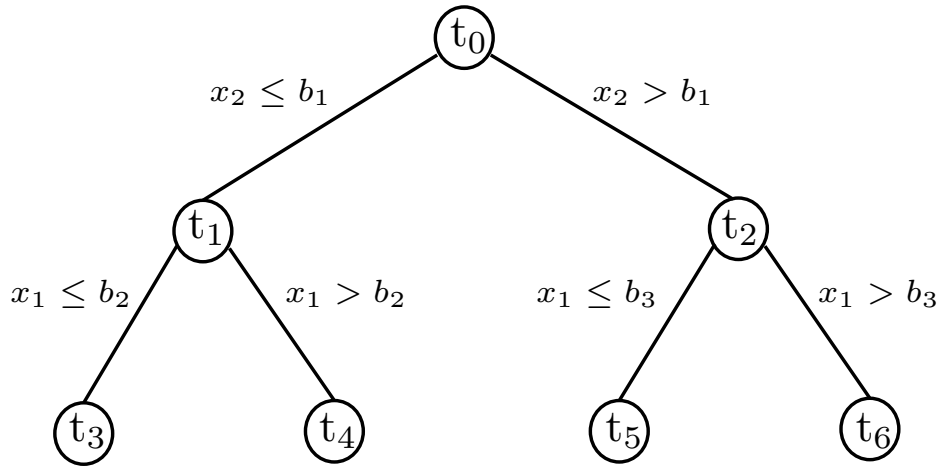

x_2
x_1


x_2
x_1


x_2
x_1

Adaptive Axis-Aligned Decision Tree (CART)





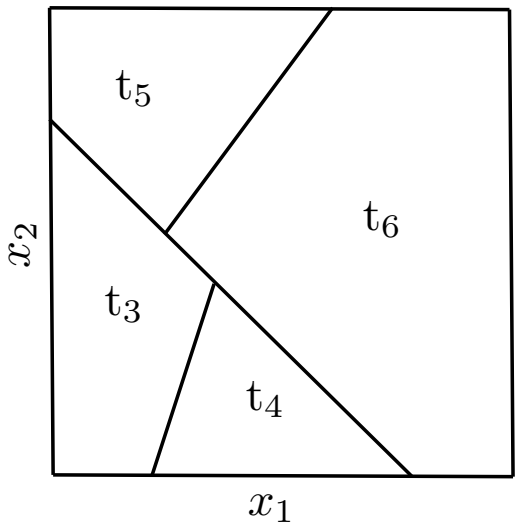
$$\widehat{\mu}(T_K)(\mathbf{x}) = \overline{y}_{\mathbf{t}} = \frac{1}{n(\mathbf{t})} \sum_{\mathbf{x}_i \in \mathbf{t}} y_i, \qquad n(\mathbf{t}) = \sum_{\mathbf{x}_i \in \mathbf{t}} \mathbb{1}(\mathbf{x}_i \in \mathbf{t}).$$

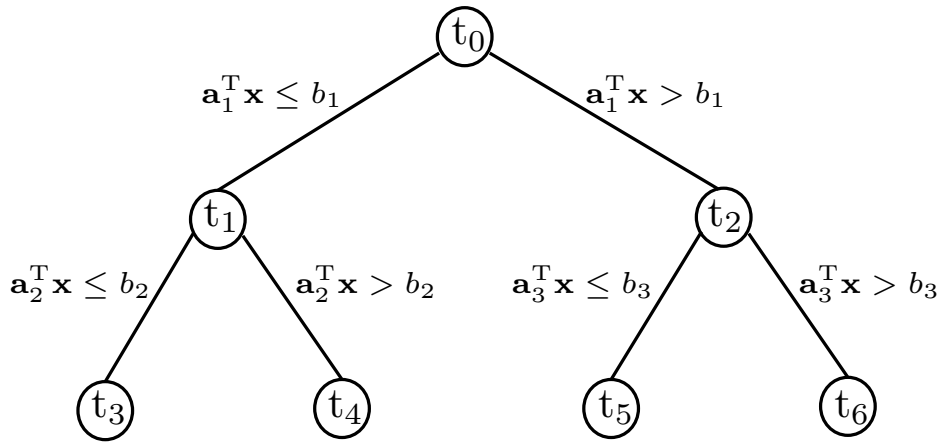
CKT (2024): for "honest" trees and $\mu(\mathbf{x}) = \mu$,

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

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

Adaptive Oblique Decision Tree (OCART)





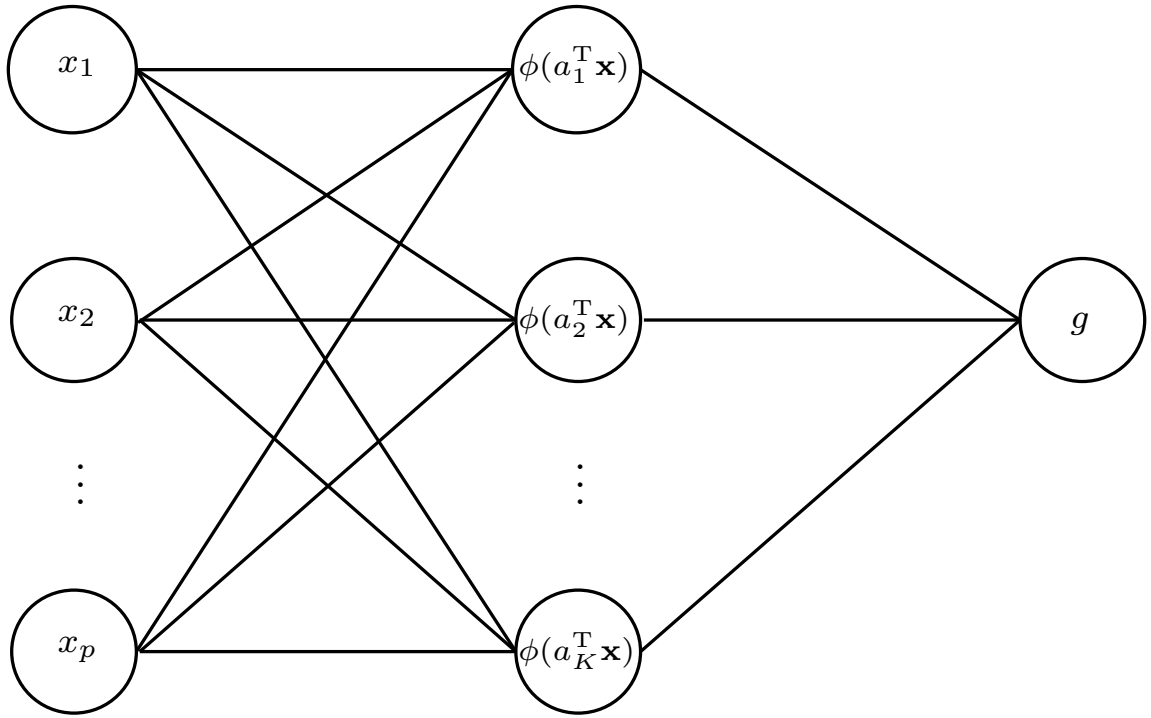
$$\widehat{\mu}(T_K)(\mathbf{x}) = \overline{y}_{\mathbf{t}} = \frac{1}{n(\mathbf{t})} \sum_{\mathbf{x}_i \in \mathbf{t}} y_i, \qquad n(\mathbf{t}) = \sum_{\mathbf{x}_i \in \mathbf{t}} \mathbb{1}(\mathbf{x}_i \in \mathbf{t}).$$

CCK (2024): for "full-sample" trees and $\mu \in$ Barron class,

$$\mathbb{E}\big[\|\widehat{\mu}(T_K) - \mu\|^2\big] \lesssim \frac{\|\mu\|_{\mathcal{L}_1}^2 \mathbb{E}\big[\max_{\mathbf{t} \in [T_K]} P_{\mathcal{A}_{\mathbf{t}}}^{-1}(\kappa)\big]}{\kappa K} + \frac{2^K d \log(np/d) \log^2(n)}{n},$$

$$\mathbb{E}\big[\|\widehat{\mu}(T_{\mathrm{opt}}) - \mu\|^2\big] \lesssim \left(\frac{p \log^3(n)}{n}\right)^{2/(2+q)} \approx \text{Optimal rate 1-HL NN.}$$

Single-Hidden Layer Neural Network with K Hidden Nodes



$$\text{OCART} \iff \phi(\cdot) = \text{ReLU}$$

- More generally, from the optimization community, feed-forward neural networks with Heaviside activations can be transformed into oblique decision trees with the same training error. See Bertsimas et al. (2018, 2021).

Outline

Decision Trees for Heterogeneous Causal Effects

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

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

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

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

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

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

- 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: 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 μ.

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

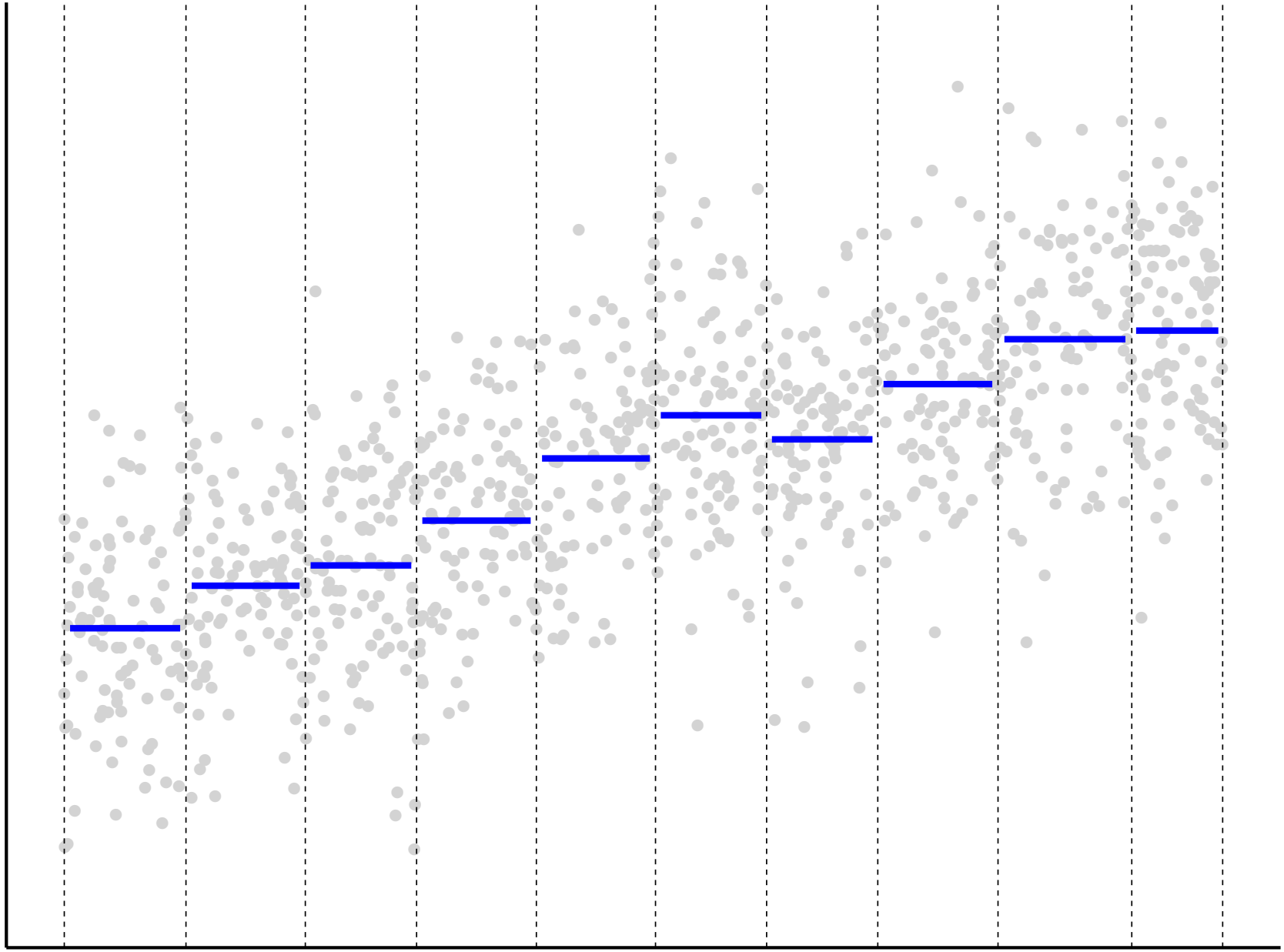
Iterating nearly inconsistent decision stumps can only make things worse... Thus, employing "honesty" (i.e., sample splitting), we have:

Theorem

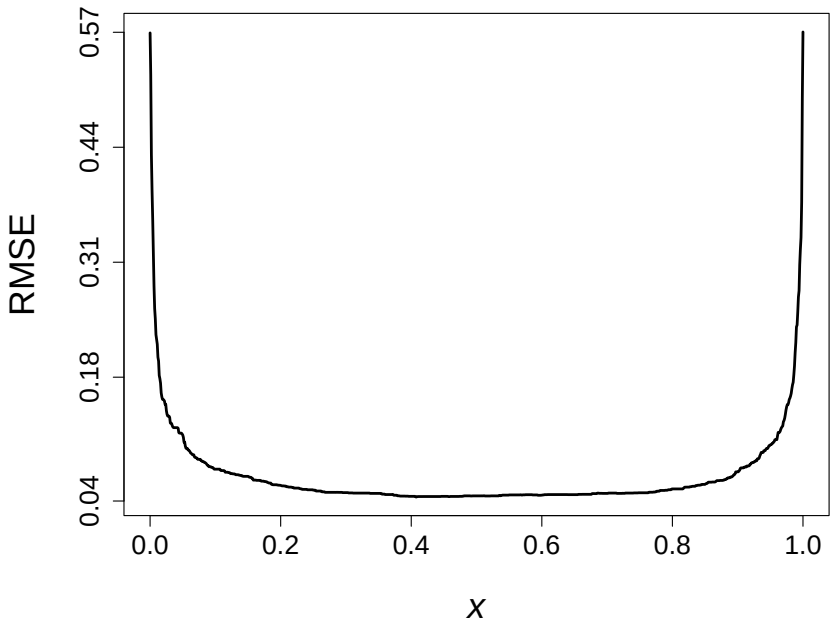
Suppose $p = 1$. Consider a maximal depth $K_n \gtrsim \log\log(n)$ tree T_{K_n} constructed with CART+ methodology. Then, there exists a positive constant C such that

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

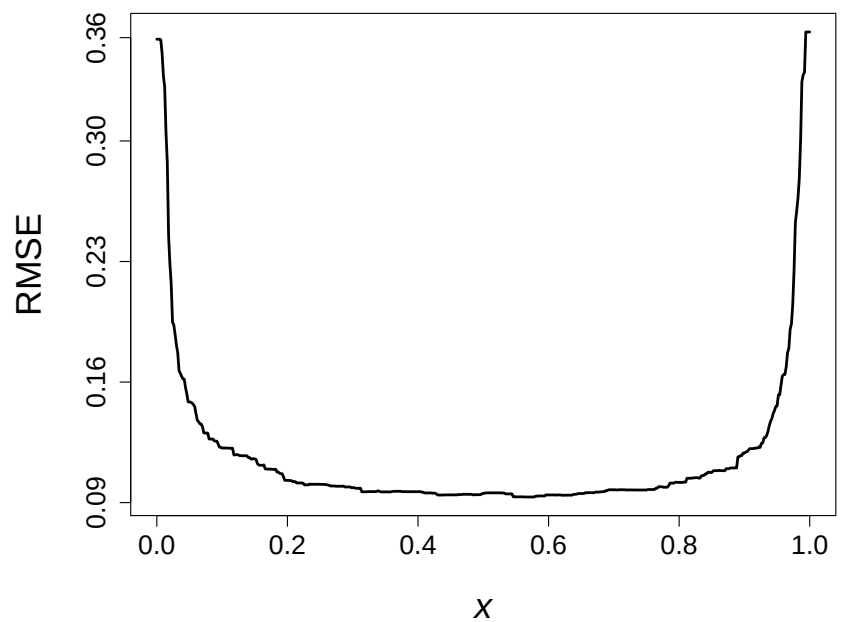
- Shallow "Honest" decision/causal trees are uniformly inconsistent.
- Inconsistency due to variance issue, not to boundary/misspecification bias.
- Inconsistency can occur at *countable* many points on the *entire* support $\mathcal{X}$.
- Pruning does not mitigate the inconsistency.
- Non-constant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



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

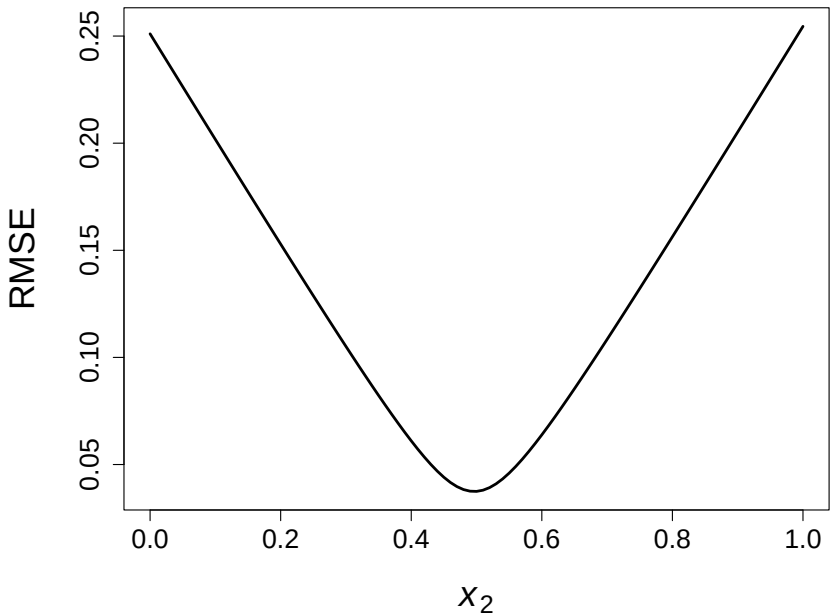


(a) Pointwise RMSE of decision stump.

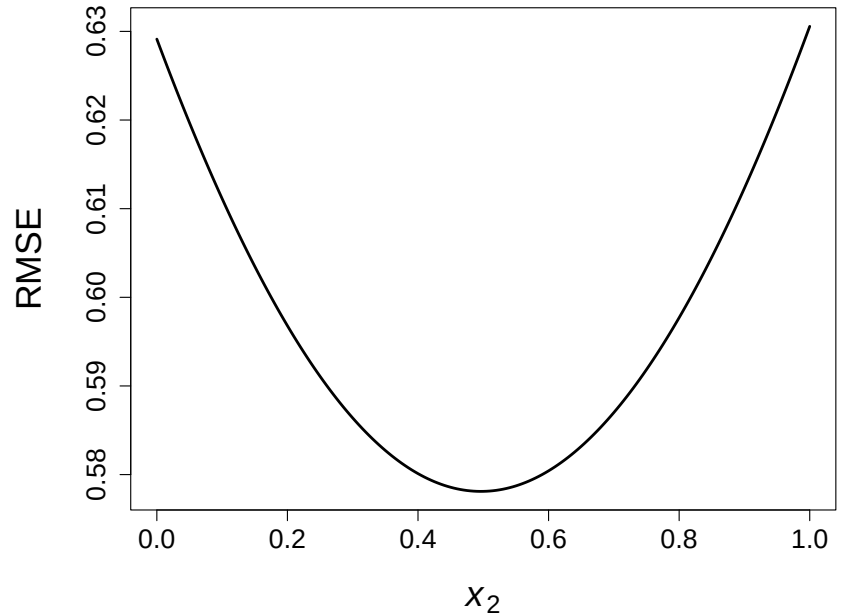


(b) Pointwise RMSE of causal decision stump.

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

Motivation

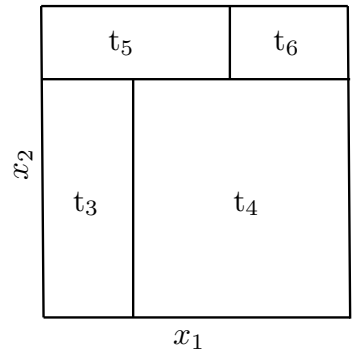
- Popular belief: decision trees compromise accuracy for being easy to use and understand, whereas neural networks are more accurate but less transparent.
- However, growing body of empirical work in optimization literature shows that certain trees are competitive with neural networks.

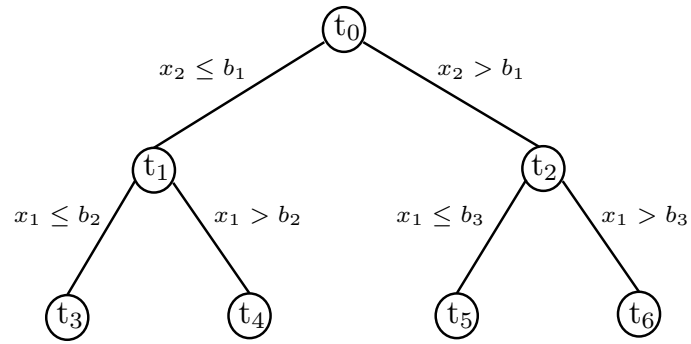
Classification Dataset	n	p	Number of Classes	Oblique Decision Tree DT depth	Oblique Decision Tree Error	2-Layer NN Width	2-Layer NN Error
Bank Marketing	45,211	17	2	3	89.6%	8	89.6%
Framingham Heart Study	3,658	15	2	2	83.3%	4	82.1%
Image Segmentation	210	18	7	4	86.0%	16	88.4%
Letter Recognition	20,000	16	26	6	72.0%	64	66.8%
Magic Gamma Telescope	19,020	10	2	5	88.6%	16	87.5%
Skin Segmentation	245,057	3	2	4	99.9%	16	99.9%
Thyroid Disease ANN	3,772	21	3	3	99.9%	8	97.7%

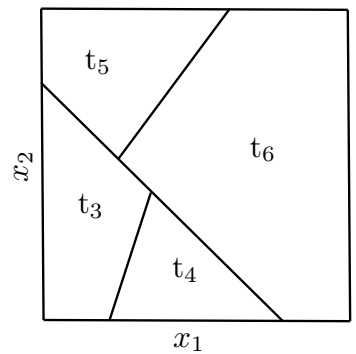
Bertsimas et al., (2018)

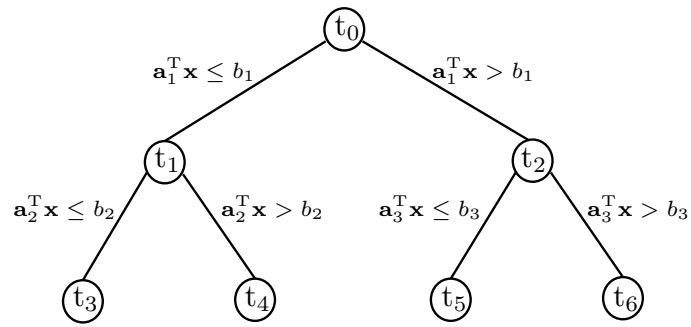
- **Question**: Is there a theoretical basis for this?
- Key advantages of binary (adaptive) decision trees:
  - Interpretability.
  - Connection to rule-based decision-making.
  - Mimics way doctor or business manager thinks.

Adaptive Axis-Aligned vs. Oblique Decision Tree (CART vs. OCART)









- Maximal decision trees with depth $K = 2$.
- OCART: splits occur along hyperplanes $\implies$ partitions are convex polytopes.

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

Oblique Tree Construction

- CART methodology: parent node $\mathbf{t}$ (region in $\mathbb{R}^p$) is divided into two child nodes, $\mathbf{t}_L$ and $\mathbf{t}_R$, by finding least squares decision stump

$$\psi(\mathbf{x}) = \beta_1 \mathbb{1}(\mathbf{a}'\mathbf{x} \leq b) + \beta_2 \mathbb{1}(\mathbf{a}'\mathbf{x} > b).$$

- Maximize decrease in sum-of-squares error

$$\widehat{\Delta}(b, \mathbf{a}, \mathbf{t}) = \sum_{\mathbf{x}_i \in \mathbf{t}} (y_i - \overline{y}_{\mathbf{t}})^2 - \sum_{\mathbf{x}_i \in \mathbf{t}} (y_i - \psi(\mathbf{x}_i))^2$$

  with respect to $(b, \mathbf{a})$.

- Greedy Refinement of Partition: Optimizers $(\hat{b}, \hat{\mathbf{a}})$ produce refinement of parent node $\mathbf{t}$ via child nodes

$$\mathbf{t}_L = \{\mathbf{x} \in \mathbf{t} : \hat{\mathbf{a}}'\mathbf{x} \leq \hat{b}\}, \quad \mathbf{t}_R = \{\mathbf{x} \in \mathbf{t} : \hat{\mathbf{a}}'\mathbf{x} > \hat{b}\}.$$

- Child nodes become new parent nodes at next level and can be further refined in same manner until desired depth D is reached.

Computational Challenges and Framework

- Challenging to find direction $\hat{\mathbf{a}}$ that minimizes squared error.
- Restrict search space to more tractable subset of candidate directions $\mathbf{a} \in \mathcal{A}_\mathbf{t}$ and allow slackness factor κ:

$$P_{\mathcal{A}_\mathbf{t}}(\kappa) = \mathbb{P}_{\mathcal{A}_\mathbf{t}}\left(\max_{(b,\mathbf{a}) \in \mathbb{R} \times \mathcal{A}_\mathbf{t}} \widehat{\Delta}(b, \mathbf{a}, \mathbf{t}) \geq \kappa \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, \mathbf{t}) \right)$$

- Choose meaningful method for generating $\mathcal{A}_\mathbf{t}$ so that $P_{\mathcal{A}_\mathbf{t}}(\kappa) \geq \rho > 0$, a.s.
  - **Deterministic**. Direct optimization, i.e., $\mathcal{A}_\mathbf{t} = \mathbb{R}^p$; solve least squares problem using mixed-integer linear optimization.
  - **Purely random**. Generate candidate directions $\mathcal{A}_\mathbf{t}$ uniformly at random (à la random forests).
  - **Data-driven**. Use dimension-reduction techniques on separate sample, e.g., $\mathcal{A}_\mathbf{t}$ defined in terms of top principle components produced by PCA or LDA, or, similarly, in terms of relevant variables selected by Lasso.

Function Class Approximations: 2-Layer NN vs. Tree Expansions

- **2-Layer Neural Networks**: distributed hierarchical representations

$$\left\{ g(\mathbf{x}) = \sum_k c_k \phi(\mathbf{a}_k' \mathbf{x}),\ c_k \in \mathbb{R},\ \mathbf{a}_k \in \mathbb{R}^p \right\}$$

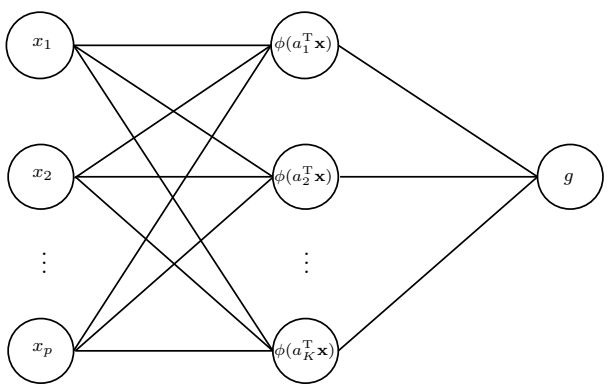
  Fixed activation function ϕ (e.g., ReLU).

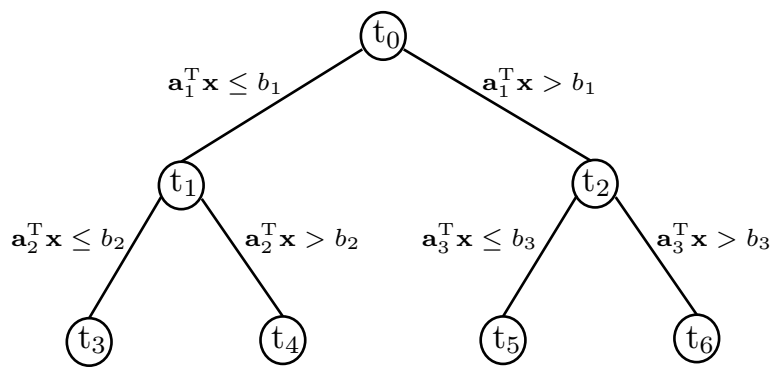
- **Decision Trees**:

$$\left\{ g(\mathbf{x}) = \sum_k c_k \mathbb{1}(\mathbf{x} \in \mathbf{t}_k) : c_k \in \mathbb{R},\ \mathbf{t}_k \text{ disjoint convex polytope} \right\}$$

  Regions $\mathbf{t}_k$ are determined by sequence of linear constraints, $\mathbf{a}'\mathbf{x} \leq b$ or $\mathbf{a}'\mathbf{x} > b$.

- Very different functional forms.





Three Key Assumptions

1. **Local Variation**: Define norm of $\mu(\mathbf{x}) = \sum_k c_k \phi(\mathbf{a}_k' \mathbf{x})$ on region $\mathbf{t}$ by

$$\|\mu\|_{\mathcal{L}_1(\mathbf{t})} = \sum_k |c_k| \mathrm{V}_k(\mathbf{t}),$$

where $\mathrm{V}_k(\mathbf{t})$ is total variation of ϕ on interval $\left[\min_{\mathbf{x}\in\mathbf{t}} \mathbf{a}_k'\mathbf{x},\ \max_{\mathbf{x}\in\mathbf{t}} \mathbf{a}_k'\mathbf{x}\right]$.

   - Measures how much μ varies on region $\mathbf{t}$.
   - Example: If $\mu(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$, then $\|\mu\|_{\mathcal{L}_1([0,1]^p)} = \|\boldsymbol{\beta}\|_{\ell_1}$.

2. **Global Variation**: There exist $V > 0$ and $q > 2$ such that

$$\mathbb{E}\left[\sum_{\mathbf{t}\in T_K} \|\mu\|^q_{\mathcal{L}_1(\mathbf{t})}\right] \leq V^q$$

   - ℓ_q constraint on total variations of μ across all terminal nodes of tree.
   - Ensures compatibility between tree and ridge expansion.

3. **Node Size**: There exist $A = \mathrm{polylog}(n)$ and $\nu \geq 1 + 2/(q-2)$ such that

$$\left(\mathbb{E}\left[\left(\max_{\mathbf{t}\in T_K} n(\mathbf{t})\right)^\nu\right]\right)^{1/\nu} \leq \frac{An}{2^K}$$

   - No region contains disproportionately more observations than average ($n/2^K$).
   - Allows for some regions to contain very few observations.

Expected Training / Prediction Error

- Training error of tree: $\|y - \widehat{\mu}(T_K)\|_n^2 = \frac{1}{n}\sum_i (y_i - \widehat{\mu}(T_K)(\mathbf{x}_i))^2$
- Prediction error of tree: $\|\mu - \widehat{\mu}(T_K)\|^2 = \int (\mu(\mathbf{x}) - \widehat{\mu}(T_K)(\mathbf{x}))^2 d\mathbb{P}_{\mathbf{x}}$

Theorem

For any depth $K \geq 1$, $\mathbb{E}\big[\|y - \widehat{\mu}(T_K)\|_n^2 - \|y - \mu\|_n^2\big] \leq 4^{-(K-1)/q} \frac{AV^2}{\rho\kappa}$.

Furthermore, if $K \approx \frac{q}{2+q}\log_2(n/p)$, *then*

$$\mathbb{E}\big[\|\widehat{\mu}(T_K) - \mu\|^2\big] \leq C\left(\frac{p}{n}\right)^{2/(2+q)}$$

Statistical Accuracy Comparisons:

- When $q \approx 2$, convergence rate

$$\left(\frac{p}{n}\right)^{2/(2+q)} \approx \left(\frac{p}{n}\right)^{1/2}$$

  same as for least squares neural network estimators (Barron, 1994).
- q plays role of effective dimension, not ambient dimension p.
- If μ is smooth, we expect $q \leq p$, and so convergence rate always at least as fast as minimax optimal rate: $(1/n)^{2/(2+p)}$.

Discussion

- **Pruned Tree**: same guarantees hold for pruned subtree that minimizes penalized risk

$$T_{\text{opt}} \in \underset{T \preceq T_{\max}}{\arg\min} \left\{ \|y - \widehat{\mu}(T)\|_n^2 + \lambda |T| \right\},$$

  where $\lambda \gtrsim p/n$. Optimal subtree T_{opt} can be found efficiently.

- **Key Technical Idea**: tree output $\widehat{\mu}(T_D)$ is orthogonal projection of y onto span of orthonormal functions $\psi_{\mathbf{t}} = \psi_{\mathbf{t}}(b, \mathbf{a})$. That is,

$$\widehat{\mu}(T_D)(\mathbf{x}) = \sum_{\mathbf{t} \in T_D} \langle y, \psi_{\mathbf{t}} \rangle \, \psi_{\mathbf{t}}(\mathbf{x}),$$

  where $\langle y, \psi_{\mathbf{t}} \rangle = \frac{1}{n} \sum_i (y_i - \overline{y}_{\mathbf{t}}) \psi_{\mathbf{t}}(\mathbf{x}_i)$ is empirical inner product, maximized at least squares solution $(\hat{b}, \hat{\mathbf{a}})$.

- **Connections to Linear Regression**: similar to forward-stepwise regression. At each current decision node $\mathbf{t}$, tree is grown by selecting "feature", $\psi_{\mathbf{t}}$, most correlated with residuals, $y_i - \overline{y}_{\mathbf{t}}$, and adding chosen feature along with coefficient, $\langle y, \psi_{\mathbf{t}} \rangle$, to tree output:

$$\widehat{\mu}(T_{D+1})(\mathbf{x}) = \widehat{\mu}(T_D)(\mathbf{x}) + \langle y, \psi_{\mathbf{t}} \rangle \, \psi_{\mathbf{t}}(\mathbf{x}).$$

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.
  - Oblique: mean square consistent $\iff$ Single-hidden layer NN performance.
- 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 (2024): "On the Pointwise Behavior of Recursive Partitioning and Its Implications for Heterogeneous Causal Effect Estimation", `arXiv:2211.10805`.

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

Related Work:

1. C, Feng & Shigida (2024): "Uniform Inference for Nonparametric Partitioning-Based M-Estimators", coming soon.

2. C, Crump, Farrell & Feng (2024): "Generalized Binscatter Methods", coming soon.

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*, forthcoming.

5. C, Farrell & Feng (2020): "Large Sample Properties of Partitioning-Based Series Estimators", *Annals of Statistics* 48(3): 1718-1741.

6. C & Farrell (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators", *Journal of Econometrics* 174(2): 127-143.