Heterogeneous Treatment Effect using Random Forest[1]

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

[1]This section is based on [Wager and Athey, 2018]

Outline

Motivation

- **Causal Inference:** Utilizing data to infer **causal relationships** in various fields.
- **Data Limitations:** Historical challenges with **small datasets** limiting heterogeneity analysis.
- **Data Availability:** Emergence of **large datasets** enabling detailed individual-level analyses.
- **Heterogeneity Challenges:** Developing protocols to prevent **spurious results** in subgroup analyses.
- **Nonparametric Methods:** Handling complex relationships without imposing **parametric forms**.
- **Machine Learning:** Using **random forests** for high-dimensional data challenges.
- **Random Forests in Causal Inference:** Ensuring **consistency** and interpretable **asymptotic distributions**.

Heterogeneous Treatment Effect Estimation

Problem: Fear of spurious heterogeneity in treatment effects due to iterative search for high treatment levels in subgroups [Assmann et al. 2000; Cook et al. 2004].

Solution: Developing a nonparametric method for estimating heterogeneous treatment effects that provides valid asymptotic confidence intervals.

Classical Approaches: Nearest-neighbor matching, kernel methods, and series estimation [Crump et al. 2008, Lee 2009, Willke et al. 2012].

Our Approach: Utilizing random forest algorithms for improved performance in high-dimensional settings [Breiman 2001].

Challenges: Ensuring consistency and understanding asymptotic sampling distribution for causal inference.

Causal Forest

- **Causal Forest Algorithm Development:** A tractable method allowing for valid statistical inference.
- **Theoretical Foundation:** Establishment of consistency and asymptotic normality with honest trees.
- **Extending to Treatment Effects:** Application to potential outcomes framework.
- **Applications Beyond Causal Inference:** Predictive contexts such as healthcare resource allocation.
- **Performance Evaluation:** Superior performance of causal forest algorithm compared to k-nearest neighbors.
- **Confidence Interval Validation:** Evaluation of confidence intervals for heterogeneous treatment effects.

Outline

Setting and Challenges I

Setting:

- n i.i.d. training examples (X_i, Y_i, W_i), where $X_i \in [0,1]^d$, $Y_i \in \mathbb{R}$, $W_i \in \{0,1\}$.
- Framework of potential outcomes with $Y_i(1)$ and $Y_i(0)$ [Neyman, 1923; Rubin, 1974; Imbens and Rubin, 2015].
- Treatment effect at x: $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$.

Challenges:

- Only one potential outcome is observable; cannot directly observe $Y_i(1) - Y_i(0)$.
- Unconfoundedness assumption: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i|X_i$ [Rosenbaum and Rubin, 1983].

Setting and Challenges II

Implications of Unconfoundedness:

- Treatment assignment W_i is conditionally independent of potential outcomes given X_i.
- Nearest-neighbor and other local methods can be consistent for $\tau(x)$.
- Propensity score $e(x) = \mathbb{E}[W_i|X_i = x]$ helps to create an unbiased estimator for $\tau(x)$.

Unconfoundedness and Its Consequences I

Key Formulas:

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$$
$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i|X_i$$
$$\mathbb{E}\left[\frac{Y_i W_i}{e(x)} - \frac{Y_i(1-W_i)}{1-e(x)}|X_i = x\right] = \tau(x)$$

Unconfoundedness:

- Assumption: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i|X_i$ [Rosenbaum and Rubin, 1983].
- Implies that nearby observations in x-space can be treated as coming from a randomized experiment.
- Leads to consistency of nearest-neighbor matching and other local methods for estimating $\tau(x)$.

Unconfoundedness and Its Consequences II

Implications:

- The propensity score $e(x) = \mathbb{E}[W_i|X_i = x]$ enables an unbiased estimator for $\tau(x)$.
- Methods based on propensity weighting [Hirano et al., 2003] utilize $e(x)$ to estimate $\tau(x)$.
- Machine learning applications to causal inference often focus on estimating $e(x)$ [McCaffrey et al., 2004; Westreich et al., 2010].

Unconfoundedness and Its Consequences III

Our Approach:

- We utilize causal forests to achieve consistency under the unconfoundedness assumption without explicitly estimating the propensity $e(x)$.

From Regression Trees to Causal Trees

Adaptive Nearest Neighbor Methods:

- Trees and forests as nearest neighbor methods with an adaptive metric.
- Classical k-nearest neighbors vs. tree-based methods:
  - k-nearest: fixed pre-specified distance (e.g., Euclidean).
  - Trees: closeness defined within the same leaf of a decision tree.
- Advantage of trees: adaptively narrow or wider leaves based on signal variation.

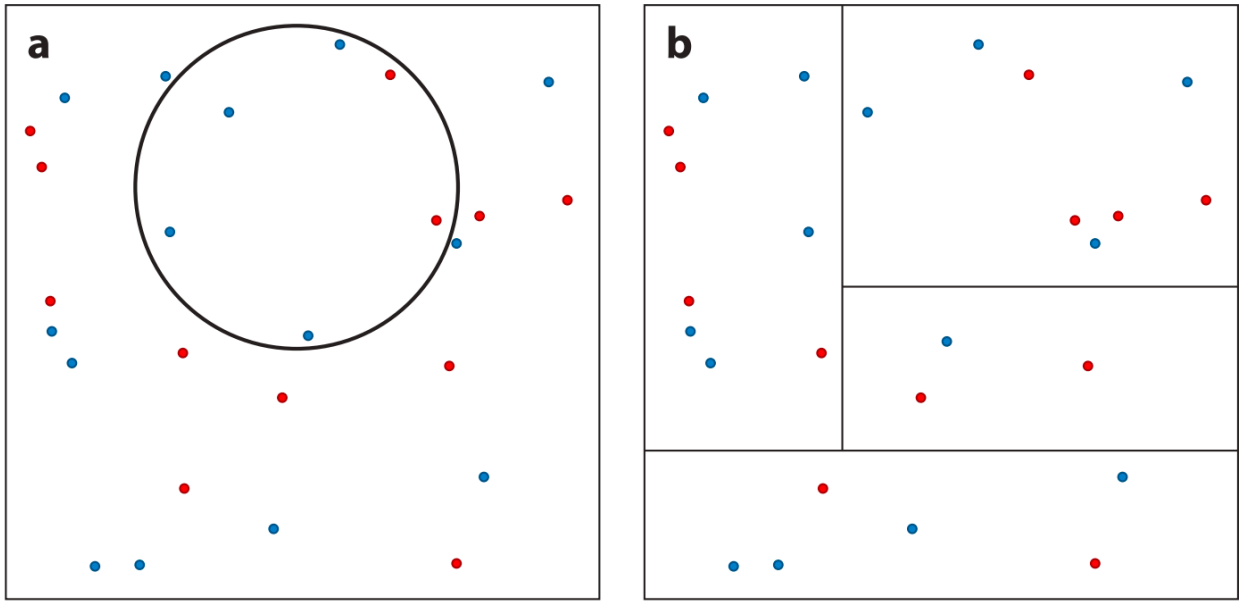

Figure 1

(*a*) Euclidean neighborhood for *k*-nearest neighbor (KNN) matching. (*b*) Tree-based neighborhood.

Figure: Enter Caption

Building Causal Trees

Objective:

- Construct causal trees closely resembling regression trees.

CART Regression Trees:

- Start with independent samples (X_i, Y_i).
- Recursively split the feature space to form leaves L containing few training samples.
- For a test point x, find the leaf $L(x)$ containing x.
- Evaluate the prediction $\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i : X_i \in L(x)\}} Y_i$.

Heuristic Justification:

- Assume responses Y_i within a leaf $L(x)$ are roughly identically distributed.
- Several procedures exist for placing splits in the decision tree [Hastie et al., 2009].

From Regression Trees to Causal Trees and Forests I

Estimating Treatment Effects with Causal Trees:

- Leaves where (Y_i, W_i) pairs mimic a randomized experiment.
- Treatment effect estimate $\hat{\tau}(x)$ for $x \in L$:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i: W_i=1, X_i \in L\}} Y_i - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i: W_i=0, X_i \in L\}} Y_i$$

Causal Forests:

- Ensemble of causal trees, each providing an estimate $\hat{\tau}_b(x)$.
- Forest estimate: $\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x)$.
- Trees built using random subsamples of training examples.
- Forests reduce variance and smooth decision boundaries compared to a single tree.

From Regression Trees to Causal Trees and Forests II

Key Advantages:

- Flexibility in handling high-dimensional feature spaces.
- Ability to capture complex interactions and heterogeneity in treatment effects.
- Robustness through aggregation, reducing overfitting and improving generalization.

References:

- Breiman, L. (2001a). Random Forests.
- Bühlmann, P., & Yu, B. (2002). Analyzing bagging.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.

Causal Forests

Generation of Causal Forests:

- Ensemble of B causal trees, each providing an estimate $\hat{\tau}_b(x)$.
- Aggregate predictions: $\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x)$.
- Trees built using random subsamples of s training examples, where $s/n \approx 1$ and $s \leq n^{\beta}$ for some $\beta < 1$.

Advantages of Causal Forests:

- Uncertainty in identifying the "best" causal tree is mitigated.
- Aggregation of diverse trees reduces variance and smooths decision boundaries [Breiman, 2001a; Bühlmann and Yu, 2002].
- Offers a balance between model complexity and generalization.

Outline

Asymptotic Inference with Causal Forests

Our analysis of causal forests hinges on specific conditions related to the forest-growing scheme. We highlight the necessity of subsampling and a non-outcome-influenced splitting rule.

Consistency of Causal Forests:

- Causal forests are consistent for the true treatment effect $\tau(x)$.
- Pointwise consistency requires Lipschitz continuity of conditional mean functions:
  - $E[Y(0) \mid X = x]$ and $E[Y(1) \mid X = x]$.
- Overlap condition: $\epsilon < P(W = 1 \mid X = x) < 1 - \epsilon$ ensures sufficient treatment and control units.

Variance Estimation for Causal Forests

Infinitesimal Jackknife for Random Forests:[2]

- Assumes large number of trees B to neglect Monte Carlo variability.
- Focuses on randomness in $\hat{\tau}(x)$ due to the training sample.

Variance Estimates:

- Let $\hat{\tau}_b^*(x)$ be the treatment effect estimate by the b-th tree.
- Let $N_{ib}^* \in \{0, 1\}$ indicate if the i-th training example was used for the b-th tree.
- Variance estimate: $V_{\text{IJ}}(x) = \frac{n-1}{n} \left(\frac{n}{n-s}\right)^2 \sum_{i=1}^n \text{Cov}^*[\hat{\tau}_b^*(x), N_{ib}^*]^2$.

Consistency:

- The variance estimate $V_{\text{IJ}}(x)$ is consistent: $V_{\text{IJ}}(x)/\text{Var}[\hat{\tau}(x)] \xrightarrow{p} 1$.

[2] Developed by Efron [2014] and Wager et al. [2014], based on Jaeckel [1972].

Outline

Summary of Causal Forest Results

Flexible Results:

- Wide variety of causal forests can be tailored to the application area.
- Achieve consistency and centered asymptotic normality.
- Sub-sample size s must scale at an appropriate rate.

Requirement for Individual Trees:

- Trees must satisfy a condition called honesty.
- A tree is honest if it uses the response Y_i for either estimating within-leaf treatment effect τ or deciding where to place splits, but not both.

Causal Forest Algorithms:

- Two causal forest algorithms satisfy the honesty condition.

Double-Sample Tree Algorithm

Achieving Honesty:

- Training subsample divided into two halves I and J.
- J-sample used to place splits, I-sample held out for within-leaf estimation.
- Minimum leaf size set to $k = 1$.

Related Work:

- Similar algorithms discussed by Denil et al. [2014].
- Related ideas in semiparametric inference literature go back to Schick [1986].

Efficiency and Performance:

- Sample splitting criticized as inefficient, but forest subsampling achieves honesty without wasting data.
- Each data point participates in both I and J samples of some trees.
- Double-sample trees can improve mean-squared error compared to standard random forests.

Propensity Trees and Splitting Rule

Another way to build honest trees is to ignore the outcome data Y_i when placing splits, and instead first train a classification tree for the treatment assignments W_i.

Propensity Trees:

- Train a classification tree for the treatment assignments W_i.
- Useful in observational studies to minimize bias due to variation in $e(x)$.
- Concept of matching training examples based on estimated propensity goes back to Rosenbaum and Rubin [1983].

Splitting Rule Motivation:

- Motivated by minimizing squared-error loss in regression trees [Athey and Imbens, 2016].
- Regression trees compute predictions $\hat{\mu}$ by averaging training responses over leaves.
- Squared-error minimizing split equivalent to maximizing variance of $\hat{\mu}(X_i)$ for $i \in J$.
- In double-sample trees, splits maximize the variance of $\hat{\tau}(X_i)$ for $i \in J$.

Remark 1: Motivation for the Splitting Rule

Splitting Rule in Double-Sample Trees:

- Motivated by an algorithm for minimizing squared-error loss in regression trees [Athey and Imbens, 2016].
- Regression trees compute predictions $\hat{\mu}$ by averaging training responses over leaves.
- Minimizing squared-error loss is equivalent to maximizing the variance of $\hat{\mu}(X_i)$ for $i \in J$.

Application in Double-Sample Trees:

- Splits are chosen to maximize the variance of $\hat{\tau}(X_i)$ for $i \in J$.
- Emulates the algorithm used in regression trees for treatment effect estimation.

Remark 2: Consistency and Honesty in Forests

Challenges with Small Leaves:

- Adaptive forests with small leaves can overfit to outliers near the edges of sample space [Breiman, 2001a].
- This overfitting can lead to inconsistency in treatment effect estimation.

Honesty and Consistency:

- Honesty in trees, as used in this study, is crucial for pointwise consistency [Wasserman and Roeder, 2009].
- Some recent studies explore non-honest forests [Scornet et al., 2015; Wager and Walther, 2015].
- However, these studies either do not consider pointwise properties or do not establish centered asymptotic normality.

Outline

Procedure 1: Double-Sample Trees

Double-Sample Trees: Divide training data for estimating responses and placing splits.

Input:

- n training examples (X_i, Y_i) or (X_i, Y_i, W_i).
- Minimum leaf size k.

Algorithm:

1. Draw subsample of size s, split into sets I and J.
2. Grow tree using J for splits and I for estimation.
3. Estimate responses with I-sample observations.

Regression Trees: Predictions $\hat{\mu}(x)$ made with I-sample, minimizing mean-squared error.

Causal Trees: Estimate $\hat{\tau}(x)$ with I-sample, maximizing variance of $\hat{\tau}(X_i)$ for $i \in J$.

Procedure 2: Propensity Trees

Propensity Trees: Utilize treatment assignment W_i for splits, estimate τ with responses Y_i.
Input: Training examples (X_i, Y_i, W_i), minimum leaf size k.
Algorithm:

1. Draw random subsample I of size s.
2. Train classification tree on (X_i, W_i), ensuring k observations per leaf.
3. Estimate $\tau(x)$ in each leaf using:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1, X_i \in L\}} Y_i - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0, X_i \in L\}} Y_i$$

Split Criterion: Optimize Gini criterion for splits [Breiman et al., 1984].

Inferring Heterogeneous Treatment Effects

- **Objective:** Estimating heterogeneous treatment effects in the *potential outcomes framework* with *unconfoundedness*.
- Utilizing **random forests** to adapt regression forests for *causal inference*. (Breiman, 2001a)
- Data: Tuples $Z_i = (X_i, Y_i, W_i)$, $i = 1, \ldots, n$
  - X_i: Feature vector
  - Y_i: Response
  - W_i: Treatment assignment
- Goal: Estimate the **conditional average treatment effect** $\tau(x) = E[Y(1) - Y(0) \mid X = x]$.

Asymptotic Normality in Random Forests

- **Setup:** Training examples $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$; test point x; estimate true conditional mean $\mu(x) = \mathbb{E}[Y|X = x]$.
- **Regression tree T:** Estimates $\mu(x)$ as $T(x; \xi, Z_1, \ldots, Z_n)$, $\xi \sim \Xi$.
- **Random forest:** Average of trees over size-s subsamples, marginalizing over ξ.
- **Monte Carlo averaging for random forest:**

$$RF(x; Z_1, \ldots, Z_n) \approx \frac{1}{B} \sum_{b=1}^{B} T(x; \xi_b^*, Z_{b1}^*, \ldots, Z_{bs}^*),$$

  where $\{Z_{b1}^*, \ldots, Z_{bs}^*\}$ is drawn without replacement, $\xi_b^* \sim \Xi$, B is the number of Monte Carlo replicates.
- A high value of B is suggested for achieving accuracy (Wager et al., 2014; Mentch & Hooker, 2016), recommending selecting B approximately proportional to n.

Random Forests and Honesty in Trees

Random Forest (Definition 1):

- Definition:

$$RF(x; Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_s \le n} \mathbb{E}_{\xi \sim \Xi} \left[T(x; \xi, Z_{i_1}, \ldots, Z_{i_s}) \right].$$

Honesty in Trees (Definition 2):

- A tree is **honest** if:
  - (a) (Standard case) It does not use the responses $Y_1, \ldots, Y_s$ for splits.
  - (b) (Double sample case) It does not use the I-sample responses for splits.

Definitions for Tree Consistency and Random Splits

Random-Split Trees (Definition 3):

- Each variable selected with probability at least π/d, $\pi \in (0, 1]$.
- Randomness in splitting features contained in auxiliary variable ξ.
- A tree is a **random-split tree** if the probability of splitting along the j-th feature is at least π/d, marginalizing over ξ.

α-Regularity (Definition 4):

- A tree is **α-regular** if each split leaves at least a fraction α of training examples on each side.
- In double-sample trees, regularity applies to the I sample.
- Fully grown trees have between k and $2k - 1$ observations in each terminal node.

Symmetry (Definition 5):

- A predictor is **symmetric** if its output does not depend on the order of training examples.

Asymptotic Variance and Normality of Random Forests

Asymptotic Variance Estimation:

- Using the infinitesimal jackknife of Wager et al. [2014]:

$$\hat{V}_{\mathrm{IJ}}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}^*[\hat{\mu}_b^*(x), N_{ib}^*]^2,$$

  where $\hat{\mu}_b^*(x)$ is the estimate from a single regression tree.
- Finite-sample correction is for subsampling without replacement.

Asymptotic Normality:

- Requires Lipschitz continuity of the conditional mean function $\mu(x) = \mathbb{E}[Y|X = x]$.
- Subsample size s must scale within specified bounds for asymptotic normality.
- If s grows slower, the forest may be asymptotically normal but biased.

Theorem 1: Asymptotic Normality of Random Forests

Assumptions:

- n i.i.d. training examples $(X_i, Y_i) \in [0,1]^d \times \mathbb{R}$.
- Features $X_i \sim U([0,1]^d)$, $\mu(x) = \mathbb{E}[Y|X=x]$ and $\mu^2(x) = \mathbb{E}[Y^2|X=x]$ are Lipschitz-continuous.
- $\mathrm{Var}[Y|X=x] > 0$ and $\mathbb{E}[|Y - \mathbb{E}[Y|X=x]|^{2+\delta}|X=x] \leq M$ for constants $\delta, M > 0$.

Tree Requirements:

- Honest, α-regular ($\alpha \leq 0.2$), symmetric random-split tree.

Subsample Size:

- $s_n \asymp n^\beta$ for $\beta_{\min} < \beta < 1$, where $\beta_{\min}$ is defined in the theorem.

Theorem 1: Asymptotic Normality of Random Forests (Cont'd)

Main Result:

- Random forest predictions are asymptotically Gaussian:

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow N(0,1) \text{ for a sequence } \sigma_n(x) \to 0.$$

- Asymptotic variance $\sigma_n^2(x)$ can be consistently estimated using the infinitesimal jackknife:

$$\hat{V}_{\mathrm{IJ}}(x)/\sigma_n^2(x) \xrightarrow{p} 1.$$

Remark 3 (Binary Classification):

- Theorem also holds for binary classification forests with leaf size $k = 1$.
- For $k > 1$, the theorem holds if trees are built by averaging observations within a leaf.

Outline

Bias in Regression Trees

Bounding the Bias:

- As sample size s increases, leaves get smaller due to Lipschitz-continuity and honesty.
- The diameter of a leaf $L(x)$ decreases with larger s.

Key Takeaway:

- Smaller leaves lead to more localized and accurate predictions, controlling the bias of the tree.

Lemma 2: Probabilistic Bound on Leaf Diameter

Lemma 2:

- For a regular, random-split tree T and leaf $L(x)$:
- Probability that $\text{diam}_j(L(x))$ exceeds a threshold decreases with increasing s.

Overall:

- Lemma 2 provides a theoretical foundation for understanding the behavior of random-split trees and their bias.
- Essential for ensuring the consistency and reliability of random forest predictions.

Theorem 3: Bias Bound for Random Forests

Conditions:

- Lemma 2 holds.
- $\mu(x)$ is Lipschitz continuous.
- Trees T in the random forest are honest.
- Regularity parameter $\alpha \leq 0.2$.

Result:

- The bias of the random forest at x is bounded by:

$$|E[\hat{\mu}(x)] - \mu(x)| = O\left(s^{-\frac{1}{2}}\left(\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{d}\right)\right),$$

  where the constant in the O-bound is given in the proof.

Interpretation:

- This theorem provides a bound on the bias of random forest predictions.
- The bound decreases as the sample size s increases.
- Honesty and Lipschitz continuity are key conditions for controlling the bias.

Outline

Asymptotic Normality of Random Forests

Classical Foundations:

- Based on ideas by Hoeffding [1948] and Hájek [1968] for U-statistics.

Hájek Projection T°:

- Captures first-order effects in predictor T.
- Asymptotic normality of T implied when $\lim_{n\to\infty} \frac{\mathrm{Var}[T^\circ]}{\mathrm{Var}[T]} = 1$.

Application to Regression Trees:

- Classical theory does not directly apply.
- Analysis centered around ν-incrementality.

Regression Trees as Incremental Predictors

PNN Predictors:

- Operate by nearest-neighbor search over rectangles.
- Decision trees with axis-aligned splits and specific leaf sizes are k-PNN predictors.

k-PNN Predictors:

- Output the average of responses over a k-PNN set of x.
- Predictions written as $T(x; \xi, Z_1, \ldots, Z_s) = \sum_{i=1}^{s} S_i Y_i$.

Lemma 4: Variance Bound for k-PNN Predictors

Lemma 4:

- For symmetric k-PNN predictor T and large s:

$$s\mathrm{Var}\left[\mathbb{E}\left[S_1|Z_1\right]\right] \geq \frac{1}{k} C_{f,d} / \left(\log(s)^d\right),$$

  where $C_{f,d}$ is a constant dependent on density f and dimension d.

Interpretation:

- Provides a lower bound on the information contained in Z_1 about selection event S_1.
- Indicates that honest and regular random-split trees are incremental.

Theorem 5: Incrementality of Honest Regular Symmetric Trees

Assumptions:

- Tree T is honest, k-regular, symmetric, and meets conditions of Lemma 4.
- Conditional moments $\mu(x)$ and $\mu^2(x)$ are Lipschitz continuous.

Result:

- Tree T is $\nu(s)$-incremental at x with $\nu(s) = C_{f,d} / \left(\log(s)^d\right)$.

Extension to Double-Sample Trees:

- Theorem 5 holds for double-sample trees, treating them as honest, symmetric k-PNN predictors.

Definitions for Honesty and Regularity in Causal Trees

Honesty (Definition 2b):

- A causal tree is **honest** if:
  - (a) (Standard case) It does not use the responses Y_i for splits.
  - (b) (Double sample case) It does not use the I-sample responses for splits.

Regularity (Definition 4b):

- A causal tree is α**-regular** at x if:
  - (a) (Standard case) Splits leave at least α fraction of examples on each side, leaf containing x has at least k observations from each treatment group, and leaf has either $< 2k - 1$ observations with $W_i = 0$ or $2k - 1$ observations with $W_i = 1$.
  - (b) (Double-sample case) (a) holds for the I sample in a double-sample tree.

Key Assumptions for Causal Forests

Assumptions:

- Unconfoundedness and overlap are key for consistent estimation of $\tau(x)$.
- Honest causal trees use features X_i and treatment assignments W_i for splits, but not responses Y_i.

Consistent Estimation:

- After splitting, the expected value of the causal tree $\Gamma(x)$ can be written as:

$$\begin{aligned}\mathbb{E}[\Gamma(x)|X, W] = & \frac{\sum_{i\in I^{(1)}(x)} \mathbb{E}[Y^{(1)}|X = X_i, W = 1]}{|I^{(1)}(x)|} \\ & - \frac{\sum_{i\in I^{(0)}(x)} \mathbb{E}[Y^{(0)}|X = X_i, W = 0]}{|I^{(0)}(x)|}, \\ = & \frac{\sum_{i\in I^{(1)}(x)} \mathbb{E}[Y^{(1)}|X = X_i]}{|I^{(1)}(x)|} - \frac{\sum_{i\in I^{(0)}(x)} \mathbb{E}[Y^{(0)}|X = X_i]}{|I^{(0)}(x)|}\end{aligned}$$

where $I^{(1)}(x)$ and $I^{(0)}(x)$ are indices of treatment and control units in the leaf x.

Proof and Simplification of Estimation

Simplification by Unconfoundedness:

- By unconfoundedness, the expected value simplifies to a difference of two terms:
- Each term is consistent for estimating $\mathbb{E}[Y^{(1)}|X = x]$ and $\mathbb{E}[Y^{(0)}|X = x]$, respectively.

Further Details:

- Detailed proof of Theorem 11 is provided in the appendix.
- Establishes the consistency of causal forest estimates under the given assumptions.

Remark 4: Testing at Many Points

Limitation of Regularity:

- It is not generally possible to construct causal trees that are regular for all x simultaneously.
- Example: For $d = 1$ and $W_i = 1$ for $X_i \geq 0$, the tree can have at most one leaf where it is regular.

Single Test Point Consideration:

- In the proof of Theorem 11, only a single test point x is considered.
- It is always possible to build a tree that is regular at a single given point x.

Predicting at Multiple Test Points:

- To predict at many test points, different trees may need to be assigned to be valid for different test points.
- When predicting at a specific x, treat the set of trees valid at that x as the relevant forest and apply Theorem 11.

Overcoming Bias in Causal Forests

Sources of Bias:

- Identify neighborhoods where the treatment effect $\tau(x)$ is stable.
- Address bias due to varying sampling propensities $e(x)$.

Comparison with k-NN:

- Causal forests compared to standard k nearest neighbors (k-NN) matching.
- k-NN estimates the treatment effect as $\hat{\tau}_{\text{KNN}}(x) = \frac{1}{k}\sum_{i \in S_1(x)} Y_i - \frac{1}{k}\sum_{i \in S_0(x)} Y_i$.
- Confidence intervals for k-NN modeled as Gaussian with specific mean and variance.

Simulation Study Goal:

- Verify that forest-based methods can build rigorous, asymptotically valid confidence intervals.
- Improve over non-adaptive methods like k-NN in finite samples.

Broader Context:

- Forest-based methods have shown promise for treatment effect estimation.
- Conceptual tools developed in this paper can help analyze various forest-based methods.

Experimental Setup I

Parameters:

- Sample size: n
- Ambient dimension: d
- Main effect: $m(x) = 2^{-1}\mathbb{E}[Y(0) + Y(1)|X = x]$
- Treatment effect: $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$
- Treatment propensity: $e(x) = P[W = 1|X = x]$

Assumptions:

- Unconfoundedness
- $X \sim U([0, 1]^d)$
- Homoscedastic noise: $Y(0/1) \sim N(\mathbb{E}[Y(0/1)|X], 1)$

Experimental Setup II

Performance Metrics:

- Expected mean-squared error for estimating $\tau(X)$
- Expected coverage of $\tau(X)$ with a target coverage rate of 0.95

First Experiment - Bias Resistance

Objective:

- Test the method's ability to resist bias due to interaction between $e(x)$ and $m(x)$.
- Emulate the problem in observational studies where treatment assignment is correlated with potential outcomes.

Settings:

- Treatment effect: $\tau(x) = 0$
- Treatment propensity: $e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1))$
- Main effect: $m(X) = 2X_1 - 1$
- Sample size: $n = 500$, Ambient dimension: d varied between 2 and 30

Method:

- Base learner: Propensity trees (Procedure 2)
- Number of trees: $B = 1000$, Subsample size: $s = 50$

Second Experiment - Adaptation to Heterogeneity

Objective:

- Evaluate the ability of causal forests to adapt to heterogeneity in $\tau(x)$.
- Given unconfoundedness and constant $e(x)$, this setting emulates a randomized experiment.

Settings:

- Treatment effect: $\tau(X) = \varsigma(X_1)\varsigma(X_2)$, where $\varsigma(x) = 1 + \frac{1}{1+e^{-20(x-1/3)}}$
- Main effect: $m(x) = 0$, Treatment propensity: $e(x) = 0.5$
- Sample size: $n = 5000$, Ambient dimension: d varied from 2 to 8

Method:

- Base learner: Double-sample trees with Athey and Imbens [2016] splitting rule (Procedure 1)
- Number of trees: $B = 2000$, Subsample size: $s = 2500$ ($|I| = 1250$)

References I


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