Causal Trees, Causal Forests, Generalized Random Forests, Local Linear Forests

Susan Athey – Stanford University

Athey (The Impact of Machine Learning on Economics, forth.)
Athey and Imbens (Recursive Partitioning for Heterogeneous Treatment Effects, PNAS, 2016)
Wager and Athey (Estimation and Inference of Causal Effects with Random Forests, JASA, 2018)
Athey, Tibshirani and Wager (Generalized Random Forests, AOS, forth.)
Friedberg, Athey, Tibshirani, Wager (Local Linear Forests, 2018)
Athey and Wager (Efficient Policy Learning, 2016)
Zhou, Athey, and Wager (Multi-Arm Policy Estimation, 2018)

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Holland, 1986, Imbens and Rubin, 2015, Rosenbaum and Rubin, 1983, Rubin, 1974), we posit the existence of quantities $Y_i^{(0)}$ and $Y_i^{(1)}$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment ($W_i = 1$) or no treatment ($W_i = 0$).

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Goal is to estimate the **conditional average treatment effect**

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,\middle|\, X = x\right].$$

NB: In experiments, we only get to see $Y_i = Y_i^{(W_i)}$.

The potential outcomes framework

If we make no further assumptions, estimating $\tau(x)$ is not possible.

- Literature often assumes **unconfoundedness** (Rosenbaum and Rubin, 1983)

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp W_i \mid X_i.$$

- When this assumption holds, methods based on matching or propensity score estimation are usually consistent.

Low-Dimensional Representations v. Fully Nonparametric Estimation

Causal Trees

- Move the goalpost, but get guaranteed coverage
- Easy to interpret, easy to mis-interpret
- Can be many trees
- Leaves differ in many ways if covariates correlated; describe leaves by means in all covariates

Causal Forests

- Attempt to estimate $\tau(x)$
- Can estimate partial effects
- In high dimensions, still can have omitted variable issues
- Confidence intervals lose coverage in high dimensions (bias)

Baseline method: k-NN matching

Consider the k-**NN matching** estimator for $\tau(x)$:

$$\hat{\tau}(x) = \frac{1}{k} \sum_{\mathcal{S}_1(x)} Y_i - \frac{1}{k} \sum_{\mathcal{S}_0(x)} Y_i,$$

where $\mathcal{S}_{0/1}(x)$ is the set of k-nearest cases/controls to x. This is consistent given **unconfoundedness** and regularity conditions.

- **Pro:** Transparent asymptotics and good, robust performance when p is small.
- **Con:** Acute curse of dimensionality, even when $p = 20$ and $n = 20k$.

NB: Kernels have similar qualitative issues as k-NN.

Adaptive nearest neighbor matching

Random forests are a a popular heuristic for adaptive nearest neighbors estimation introduced by Breiman (2001).

- **Pro:** Excellent empirical track record.
- **Con:** Often used as a black box, without statistical discussion.

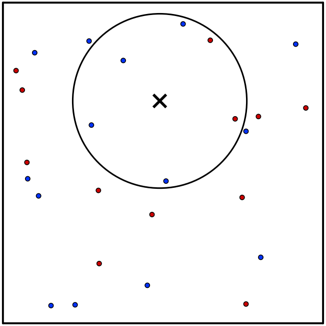
There has been considerable interest in using forest-like methods for treatment effect estimation, but without formal theory.

- Green and Kern (2012) and Hill (2011) have considered using **Bayesian forest algorithms** (BART, Chipman et al., 2010).
- Several authors have also studied related **tree-based methods**: Athey and Imbens (2016), Su et al. (2009), Taddy et al. (2014), Wang and Rudin (2015), Zeilis et al. (2008), ...

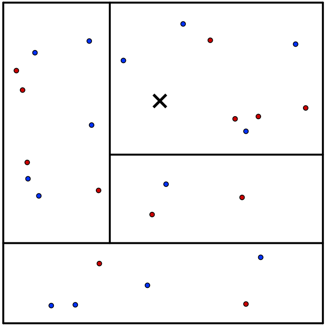
Wager and Athey (2018) provide the first formal results allowing random forest to be used for provably valid **asymptotic inference**.

Making k-NN matching adaptive

Athey and Imbens (2016) introduce **causal tree**: defines neighborhoods for matching based on **recursive partitioning** (Breiman, Friedman, Olshen, and Stone, 1984), advocate sample splitting (w/ modified splitting rule) to get assumption-free confidence intervals for treatment effects in each leaf.



Euclidean neighborhood, for k-NN matching.



Tree-based neighborhood.

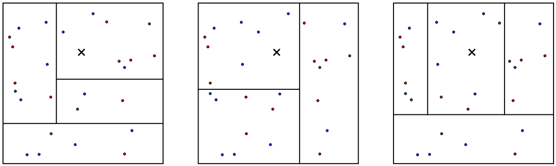
From trees to random forests (Breiman, 2001)

Suppose we have a training set $\{(X_i, Y_i, W_i)\}_{i=1}^n$, a test point x, and a tree predictor

$$\hat{\tau}(x) = T\left(x; \{(X_i, Y_i, W_i)\}_{i=1}^n\right).$$

Random forest idea: build and average many different trees T^*:

$$\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b^*\left(x; \{(X_i, Y_i, W_i)\}_{i=1}^n\right).$$



...

From trees to random forests (Breiman, 2001)

Suppose we have a training set $\{(X_i,\ Y_i,\ W_i)\}_{i=1}^n$, a test point x, and a tree predictor

$$\hat{\tau}(x) = T\left(x;\ \{(X_i,\ Y_i,\ W_i)\}_{i=1}^n\right).$$

Random forest idea: build and average many different trees T^*:

$$\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b^*\left(x;\ \{(X_i,\ Y_i,\ W_i)\}_{i=1}^n\right).$$

We turn T into T^* by:

- Bagging / subsampling the training set (Breiman, 1996); this helps smooth over discontinuities (Bühlmann and Yu, 2002).
- Selecting the splitting variable at each step from m out of p randomly drawn features (Amit and Geman, 1997).

Statistical inference with regression forests

Honest trees do not use the same data to select partition (splits) and make predictions. Ex: Split-sample trees, propensity trees.

Theorem. (Wager and Athey, JASA, 2018) Regression forests are asymptotically **Gaussian and centered**,

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1), \quad \sigma_n^2(x) \to_p 0,$$

given the following assumptions (+ technical conditions):

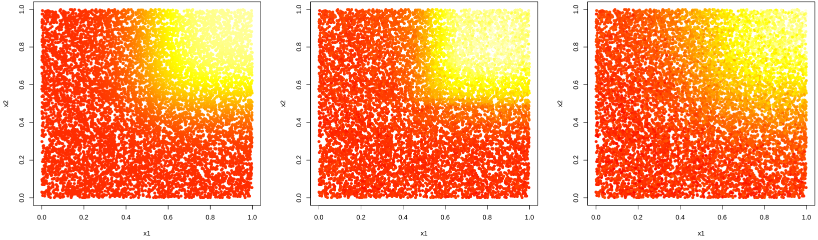
1. **Honesty.** Individual trees are honest.
2. **Subsampling.** Individual trees are built on random subsamples of size $s \asymp n^\beta$, where $\beta_{\min} < \beta < 1$.
3. **Continuous features.** The features X_i have a density that is bounded away from 0 and ∞.
4. **Lipschitz response.** The conditional mean function $\mu(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$ is Lipschitz continuous.

Causal forest example

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 6$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

true effect $\tau(x)$ **causal forest** ***k*-NN estimate**



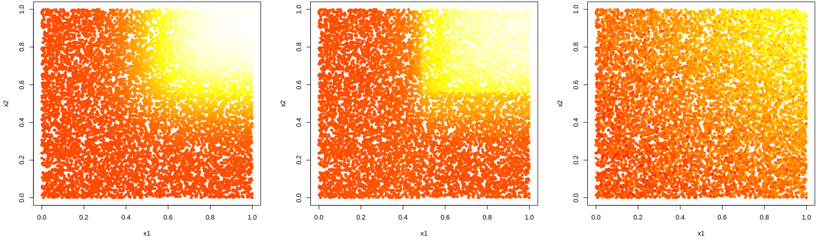
Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available at `github: susanathey/causalTree`

Causal forest example

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 20$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

true effect $\tau(x)$ | **causal forest** | ***k*-NN estimate**



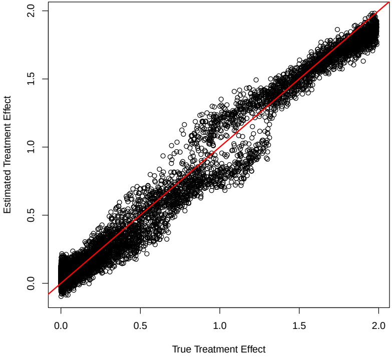
Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available at `github`: `susanathey/causalTree`

Causal forest example

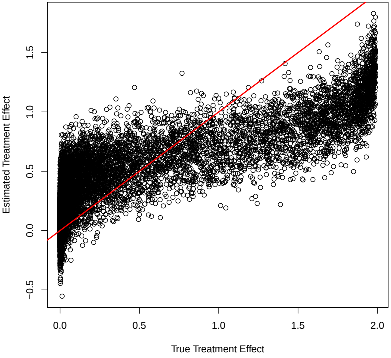
The causal forest dominates k-NN for both bias and variance.
With $p = 20$, the relative mean-squared error (MSE) for τ is

$$\frac{\text{MSE for } k\text{-NN (tuned on test set)}}{\text{MSE for forest (heuristically tuned)}} = 19.2.$$

causal forest



k-NN estimate



For $p = 6$, the corresponding MSE ratio for τ is 2.2.

Application: General Social Survey

The General Social Survey is an extensive survey, collected since 1972, that seeks to measure demographics, political views, social attitudes, etc. of the U.S. population.

Of particular interest to us is a **randomized experiment**, for which we have data between 1986 and 2010.

- **Question A:** Are we spending too much, too little, or about the right amount on **welfare**?
- **Question B:** Are we spending too much, too little, or about the right amount on **assistance to the poor**?

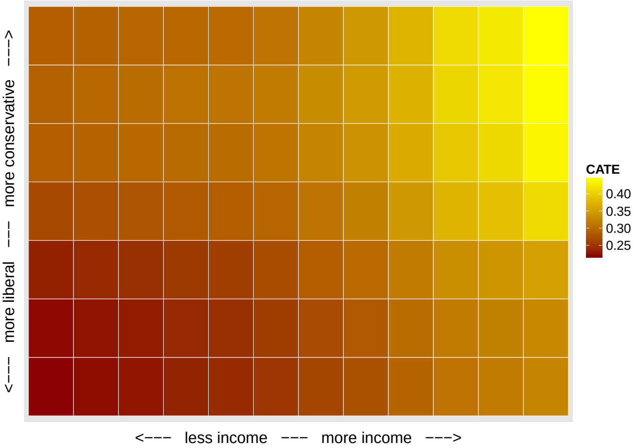
Treatment effect: how much less likely are people to answer **too much** to question B than to question A.

- We want to understand how the treatment effect depends on **covariates**: political views, income, age, hours worked, ...

NB: This dataset has also been analyzed by Green and Kern (2012) using Bayesian additive regression trees (Chipman, George, and McCulloch, 2010).

Application: General Social Survey

A causal forest analysis uncovers **strong treatment heterogeneity** ($n = 28,686, \ p = 12$).



Applications in Economics and Marketing

- Hitsch and Misra (2017): Use causal forests to target catalog mailings. Causal forest detects significant heterogeneity, performs better than alternatives including LASSO and off-the-shelf random forest
- Davis and Heller (2017): Analyze heterogeneous impacts of summer jobs using causal forest
- Athey, Campbell, Chyn, Hastings, and White (2018): Use causal forest to show that re-employment services didn't benefit in ATE, but targeted policy can have substantial benefits

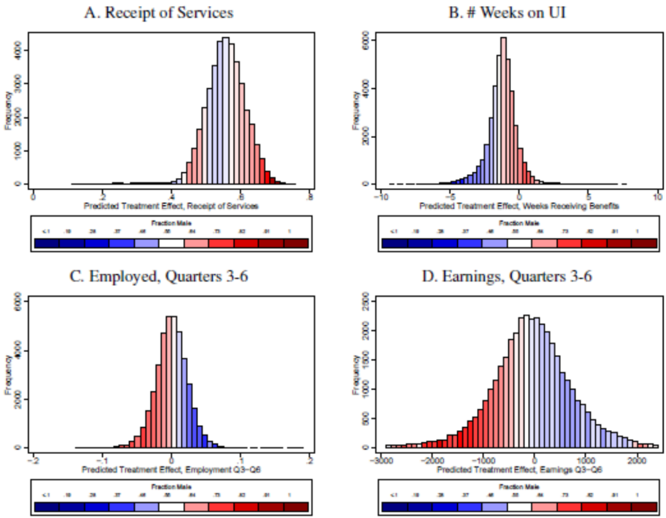
Labor Market - Reemployment Services

Athey, Campbell, Chyn, Hastings, and White (2018)

- Goal: Increase job skills and employment for all Rhode Islanders (efficiently)
- Measure impact of employment service programs
- Take advantage of a field experiment run by US Department of Labor to measure the impact of employment services on UI and subsequent employment
- From 2005-2015, states were asked to randomly send letters to UI claimants requiring employment services for continued UI receipt
- Basic evaluation of 4 states finds mixed evidence on decrease in UI and earnings impacts
- In RI we find that nudge decreased weeks on UI by 1.4/21, no impact on earnings
- Measure impact using new administrative data and causal forest (Wager and Athey 2018) to understand who benefits
- Use causal forest estimates to simulate benefits of targeted letters

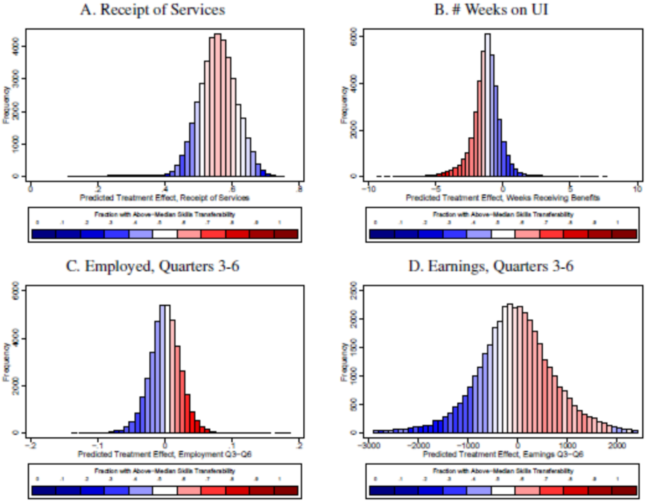
HTE in Rhode Island Re-employment Services Example

Figure 2: Distribution of Predicted Treatment Effects with Gender



HTE in Rhode Island Re-employment Services Example

Figure 5: Distribution of Predicted Treatment Effects with Occupational Skills Transferability



Forest Weaknesses

- Many economic datasets have smooth relationships
- Many relationships are monotonic or U-shaped
- Forests fit a line as a step function; very inefficient
- A variety of ML methods might improve but little theory
- Solution: Local Linear Forests + theory

Step Functions (Forests) v. Locally Linear Forest

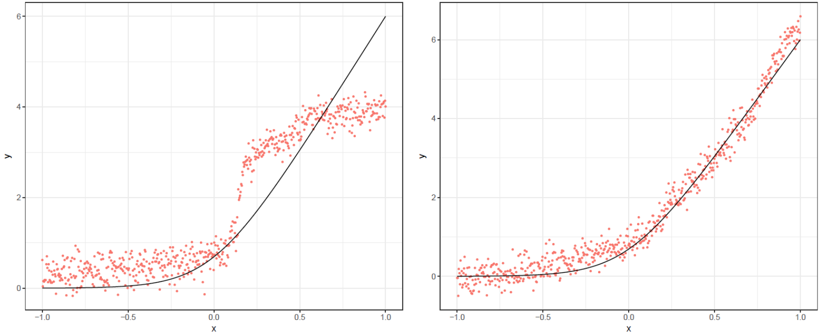
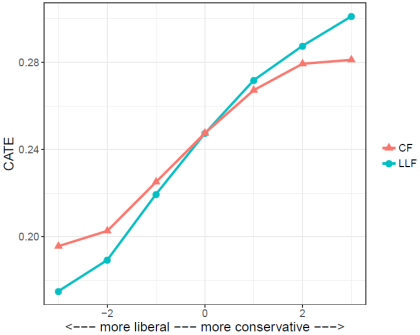
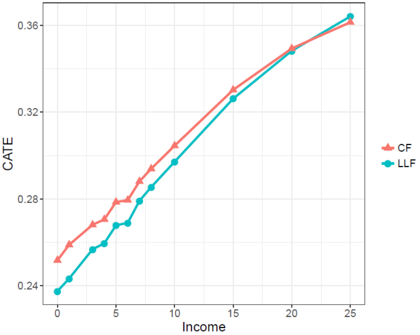


Figure 1: Predictions from random forests (left) and locally linear forests (right) on 600 test points. Training and test data were simulated from equation (1), with dimension $d = 20$ and errors $\epsilon \sim N(0, 20)$. Forests were trained also on $n = 600$ training points and tuned via cross-validation. Here the true conditional mean signal $\mu(x)$ is in black, and predictions are shown in red.

Causal Forest v. Locally Linear Causal Forest





Locally Linear Forest

Locally linear regression with ridge penalty:

$$\begin{pmatrix}\hat{\mu}(x)\\ \hat{\theta}(x)\end{pmatrix} = \arg\min_{\mu,\theta} \sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu(x) - (X_i - x)\theta(x))^2 + \lambda||\theta(x)||_2^2$$

In matrix form:

$$\begin{pmatrix}\hat{\mu}(x)\\ \hat{\theta}(x)\end{pmatrix} = (X^T A X + \lambda J)^{-1} X^T A Y$$

Weights are determine from forest a la GRF, accounting for regression in splitting for efficiency.

Theorem (Friedberg, Athey, Tibshirani and Wager (2018): Assuming $\mu(x)$ is twice continuously differentiable, estimates are asymptotically normal. Faster rate of convergence than GRF; result exploits smoothness assumptions. Extends to causal LLF.

Using Heterogeneous Treatment Effect Estimates for Optimal Targeted Policies

- Problem: find a policy $\pi : \mathcal{X} \rightarrow \mathcal{W}$ to maximize $\mathbb{E}[Y_i(\pi(X_i))]$.
- First approach: non-parametric treatment assigment
  - Treat if $\hat{\tau}(X_i) > 0$
  - Hirano and Porter (2009) show efficient (under conditions)
- How to evaluate success?
- If $\hat{\tau}(X_i)$ OOB estimate from random forest, then the implied $\hat{\pi}(X_i)$ is independent of Y_i.
  - Define Group $G^w = \{i : \hat{\pi}(X_i) = w\}$, proportion in G^w is q^w.
  - Define $\hat{\gamma}^w$ as sample average treatment effect in G^w.
  - Improvement of $\hat{\pi}(\cdot)$ over treating no one: $q^1 \cdot \hat{\gamma}^1$
  - ... over random policy: $\frac{1}{2}(q^1\hat{\gamma}^1 - q^0\hat{\gamma}^0)$.
  - Standard errors straightforward

Policy Learning

The utilitarian **value** of a policy $\pi : \mathcal{X} \to \{0, 1\}$ is

$$V(\pi) = \mathbb{E}\left[Y_i(\pi(X_i))\right] = \mathbb{E}\left[Y_i(0)\right] + \mathbb{E}\left[\tau(X)\pi(X)\right].$$

In the abstract, we maximize utility by treating according to a **thresholding rule** $\tau(X_i) > c$.

But estimating the conditional average **treatment effect function** $\tau(\cdot)$ and learning a good **policy** $\pi(\cdot)$ are different problems.

- The correct **loss function** for policy learning is not mean-squared error on $\tau(\cdot)$.
- The $\tau(x)$ function may change with variables we cannot use for **targeting** (e.g., variables only measured after the fact).
- We may wish to impose other constraints on policy functions

Policy Learning

We seek to maximize the **utility** of the learned policy subject to **constraints** encoded via a class Π of allowed policies.

As in Manski (2004), we focus on **minimax regret** (Savage, 1951) relative to the policy class Π. We define utility regret as $R(\pi)$,

$$R(\pi) = \sup\left\{V(\pi') : \pi' \in \Pi\right\} - V(\pi),$$

and seek a policy $\hat{\pi} \in \Pi$ satisfying a high-probability **regret bound**.

We can also write policy regret in terms of $\tau(x)$,

$$R(\pi) = \sup\left\{\mathbb{E}\left[\tau(X)\pi'(X)\right] : \pi' \in \Pi\right\} - \mathbb{E}\left[\tau(X)\pi(X)\right],$$

meaning that baseline effects don't affect policy regret.

State of the Art: Empirical Welfare Maximization

Kitagawa & Tetenov (2018) propose learning policies by maximizing an **empirical estimate of value** obtained via IPW

$$\hat{\pi} = \mathrm{argmax}\left\{\widehat{V}\left(\pi\right) : \pi \in \Pi\right\},$$

$$\widehat{V}\left(\pi\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{1\left(\{W_i = \pi(X_i)\}\right)}{e(X_i)} Y_i,$$

where $e(x) = \mathbb{P}\left[W \,\middle|\, X = x\right]$ is the propensity score. Given **unconfoundedness** (Rosenbaum & Rubin, 1983),

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i,$$

they show that if $e(x)$ is known and if Π has a finite VC-dimension,

$$R\left(\hat{\pi}\right) = \mathcal{O}_P\left(\frac{\sup\left\{|Y|\right\}}{\inf\left\{1 - e(X_i),\ e(X_i)\right\}}\sqrt{\frac{\mathrm{VC}\left(\Pi\right)}{n}}\right).$$

NB: Instance of **outcome weighted learning** (Zhao et al., 2012).

State of the Art: Empirical Welfare Maximization

Given **unconfoundedness** and **known** p**-scores**, K&T show that

$$R(\hat{\pi}) = \mathcal{O}_P\left(\frac{\sup\{|Y|\}}{\inf\{1 - e(X_i),\, e(X_i)\}}\sqrt{\frac{\mathrm{VC}(\Pi)}{n}}\right).$$

In many economic applications, these conditions do not apply:

- What if we are in an **observational study** where propensity scores are not know? (K&T have a result for this case, but their rates are sub-optimal.)
- What if the treatment is **endogenous** but we have an IV?
- What if the treatment W_i is **continuous** (e.g., a price)?

Today's topic is a **unified framework** for policy learning that can address all these challenges. Our approach builds on general results for **semiparametric estimation**, e.g. Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018).

The Literature

There is an earlier literature that considers policy learning in cases where we have a **finite-dimensional model** for $\mathbb{E}\left[Y \,\middle|\, X, W\right]$.

- Manski (2004) considers discrete x, and studies asymptotics of **conditional empirical success** rules.
- Hirano and Porter (2009) has general **asymptotic results** that apply when we can estimate $\tau(x)$ at a $1/\sqrt{n}$ rate.
- Stoye (2009) derives **exact minimax** rules for discrete x.

There is a parallel literature in statistics focused on learning optimal **unrestricted policies** π at non-parametric rates:

- Zhao et al. (2012) propose **outcome weighted learning**.
- Extensions by Chen et al. (2016), Zhou et al. (2017).

The Literature

Kitagawa and Tetenov (2018) extend this line of work by pairing **structured policy classes** with unstructured models for nature.

There are also counterparts in **computer science**, especially Swaminathan and Joachims (2015).

The idea of using **doubly robust** scoring rules for policy learning is considered and empirically examined by Dudík et al. (2011) and Zhang et al. (2012).

More broadly, the idea of optimizing an empirical utility estimate has also been advocated in **operations research** (Ban and Rudin, 2018; Bertsimas and Kallus, 2014).

Motivating Example: The California GAIN Study

The California **Greater Avenues to Independence** (GAIN) program aims to reduce dependence on welfare and promote work among disadvantaged households.

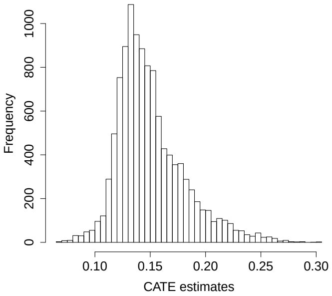
In 1988-1993, there was a **randomized evaluation** of GAIN; we want to use this to look for **heterogeneous treatment effects**. Have $p = 54$ covariates, including past income, demographics, etc.

Following Hotz, Imbens, and Klerman (2006), we focus on data from **Alameda**, **Los Angeles**, **Riverside** and **San Diego** counties.

Each county enrolled participants with a **different covariate mix**, and randomized to treatment with **different probabilities**. Once we remove county information, this is not a **randomized study**, but Hotz et al. present evidence that **unconfoundedness** holds.

We set the **cost** of treatment to match the **ATE**; thus, we need to find heterogeneity in order to get non-zero utility.

Motivating Example: The California GAIN Study



The full dataset as 19,170 samples. We trained a **causal forest** to estimate $\tau(x) - \text{cost}$, and "treat" those observations whose (out-of-bag) **value estimates** are positive.

We estimate that this results in a **utility gain** of 0.080 ± 0.028. (These numbers are computed using county information.)

Variety of potential concerns with the policy, e.g. fairness/discrimination.

Motivating Example: The California GAIN Study

	non-white	white
fraction treated	76%	81%
mean control outcome	0.79	0.90

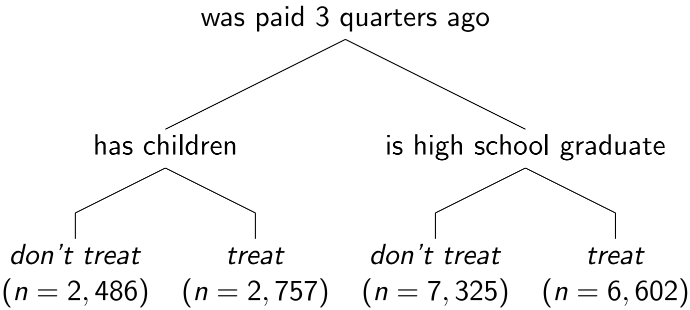
It appears that **race is a confounder**.

The full dataset as 19,170 samples. We trained a **causal forest** to estimate $\tau(x) - \text{cost}$, and "treat" those observations whose (out-of-bag) **value estimates** are positive.

We estimate that this results in a **utility gain** of 0.080 ± 0.028. (These numbers are computed using county information.)

Variety of potential concerns with the policy, e.g. fairness/discrimination.

Motivating Example: The California GAIN Study



The proposed approach will enable us to learn a **tree-shaped policy** with regret guarantees relative to the **best possible tree**.

We estimate that our method results in a **utility gain** of 0.081 ± 0.028 (again using county information).

Does not directly involve protected characteristics. (But could have other problems!)

Imposing **structure** on Π is essential in many applications. In observational studies, we use many features with a non-parametric specification to make **unconfoundedness plausible**,

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i.$$

Conversely, the policy $\pi(\cdot)$ must be **implementable in practice**. Features that should not be used in $\pi(\cdot)$ include:

- **Unreliably available features** (e.g., collected by specialist).
- **Gameable features** (e.g., self-reported preferences).
- **Legally protected classes** (e.g., religion, national origin).

Moreover, we may want Π to encode constraints on:

- **Total budget** or marginal **subgroup treatment rates** (e.g., Bhattacharya and Dupas, 2012).
- **Functional form** for easier implementation or audit.

We study policy learning in a way that is aware of such constraints.

From Evaluation to Learning

Recall that we want to pick a good **policy** $\pi : \mathcal{X} \to \{0, 1\}$ among a class Π of allowable interventions.

The **regret** from choosing π depends on the CATE function $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$:

$$R(\pi) = \sup\left\{\mathbb{E}\left[\tau(X)\pi'(X)\right] : \pi' \in \Pi\right\} - \mathbb{E}\left[\tau(X)\pi(X)\right].$$

Before discussing how to **learn** a policy, we review how how to estimate an **average effect**

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

We build on unifying results from Chernozhukov, Escanciano, Ichimura, Newey and Robins (CEINR, 2018).

From Evaluation to Learning

We have access to features X_i, an outcome Y_i, a treatment W_i, and an instrument Z_i. We assume that the exclusion restriction holds, such that potential outcomes only depend on W_i, and

$$m(x, w) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right], \quad \tau_m(x) = m(x,\, 1) - m(x,\, 0).$$

As in CEINR, suppose $\tau(x)$ can be represented via **weighting**:

$$\mathbb{E}\left[\tau_m(X) - g(X,\, Z)Y \,\middle|\, X = x\right] = 0 \text{ for all } x,\ m(\cdot).$$

CEINR then show that the **doubly robust** estimator is **efficient**,

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i,\, Z_i)\left(Y_i - \hat{m}(X_i,\, W_i)\right),$$

provided we use **cross-fitting** and have nuisance components that converge fast enough in L_2 (4th-root rates are sufficient).

From Evaluation to Learning

As in CEINR, suppose $\tau(x)$ can be represented via **weighting**:

$$\mathbb{E}\left[\tau_m(X) - g(X,\, Z)m(X,\, W) \,\middle|\, X = x\right] = 0 \text{ for all } x,\ m(\cdot).$$

CEINR then show that the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i,\, Z_i)\left(Y_i - \hat{m}(X_i,\, W_i)\right).$$

Example: Selection on observables, $\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i$. In this case, the **propensity score** can be used for weighting:

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

The corresponding doubly robust estimator is **augmented IPW** (Robins, Rotnitzky, and Zhao, 1994).

From Evaluation to Learning

As in CEINR, suppose $\tau(x)$ can be represented via **weighting**:

$$\mathbb{E}\left[\tau_m(X) - g(X,\, Z)m(X,\, W) \,\middle|\, X = x\right] = 0 \text{ for all } x,\ m(\cdot).$$

CEINR then show that the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i,\, Z_i)\left(Y_i - \hat{m}(X_i,\, W_i)\right).$$

Example: Endogenous treatment with instrument and **conditional homogeneity**, $\tau(x) = \text{Cov}\left[Y,\, Z \,\middle|\, X = x\right] / \text{Cov}\left[W,\, Z \,\middle|\, X = x\right]$. Now use the **compliance score** (Aronow and Carnegie, 2013),

$$g(X_i,\, Z_i) = \frac{1}{\Delta(X_i)} \frac{Z_i - z(X_i)}{z(X_i)(1 - z(X_i))}, \quad z(x) = \mathbb{P}\left[Z_i \,\middle|\, X_i = x\right],$$

$$\Delta(x) = \mathbb{P}\left[W \,\middle|\, Z = 1,\, X = x\right] - \mathbb{P}\left[W \,\middle|\, Z = 0,\, X = x\right],$$

to construct a doubly robust estimator.

From Evaluation to Learning

In many problems, the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i)\left(Y_i - \hat{m}(X_i, W_i)\right).$$

Our main result is that we can also use the same scores of **learning**

$$\hat{\pi} = \mathrm{argmax}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\widehat{\Gamma}_i : \pi \in \Pi\right\}.$$

Regret bounds depend on n, Π, and the semiparametric efficient variance for policy evaluation.

From Evaluation to Learning

In many problems, the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i,\, Z_i)\left(Y_i - \hat{m}(X_i,\, W_i)\right).$$

Our main result is that we can also use the same scores of **learning**

$$\hat{\pi} = \operatorname{argmax}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\widehat{\Gamma}_i : \pi \in \Pi\right\}.$$

Regret bounds depend on n, Π, and the semiparametric efficient variance for policy evaluation.

NB: The policy $\pi^* = \operatorname{argmax}\left\{\mathbb{E}\left[(2\pi(X_i) - 1)\tau(X_i)\right] : \pi \in \Pi\right\}$ gets **zero regret**. Our estimator effectively replaces $\tau(X_i)$ with $\widehat{\Gamma}_i$.

Back to the California GAIN Study

Each county enrolled participants with a **different covariate mix**, and randomized to treatment with **different probabilities**. Once we remove county information, this is not a **randomized study**, but Hotz et al. present evidence that **unconfoundedness** holds.

We set the **cost** C of treatment to match the **ATE**; thus, we need to find heterogeneity in order to get non-zero utility.

We estimate nuisance components with **forests**, and then optimize over the class Π of low-depth **trees**:

$$\hat{\pi} = \operatorname{argmax}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i)-1\right)\widehat{\Gamma}_i : \pi \in \Pi\right\},$$

$$\widehat{\Gamma}_i = \hat{\tau}^{(-i)}(X_i) - C + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1-\hat{e}^{(-i)}(X_i)\right)}$$

$$\times\left(Y_i - \hat{y}^{(-i)}(X_i) - (W_i - \hat{e}^{(-i)}(X_i))\hat{\tau}^{(-i)}(X_i)\right).$$

California GAIN Study

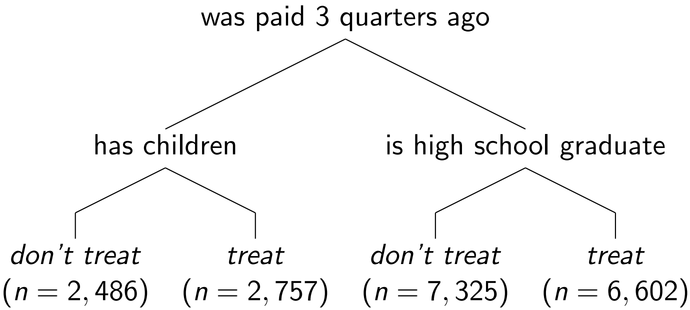
method	estimated improvement	oracle improvement
causal forest	0.095 ± 0.026	0.080 ± 0.028
IPW depth 2	0.073 ± 0.026	0.055 ± 0.028
AIPW depth 1	0.065 ± 0.026	0.050 ± 0.028
AIPW depth 2	0.098 ± 0.026	0.081 ± 0.028

We estimate **policy improvement**, i.e., gain over random assignment, by cross-validation

$$A(\pi) = \mathbb{E}\left[(2\pi(X) - 1)\tau(X)\right].$$

On the held out folds, we estimate A via either **augmented IPW** or by exploiting **within-county randomization**.

California GAIN Study



The learned **depth-2** tree is given above. During cross-validation for **depth-1** trees, 6/10 trees learned the **"has children"** rule and 4/10 trees learned the **"is high school graduate"** rule.

Main Result

Goal is to show that we can use doubly robust scores of **learning**

$$\hat{\pi} = \operatorname{argmax}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\widehat{\Gamma}_i : \pi \in \Pi\right\},$$

$$\widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i,\, Z_i)\left(Y_i - \hat{m}(X_i,\, W_i)\right).$$

Theorem. (Athey and Wager, 2018) Suppose $g(X,\, Z) \leq \eta^{-1}$, and that our nuisance estimates satisfy (we use **cross-fitting**)

$$\mathbb{E}\left[\left(\hat{m}(X,\, W) - m(X,\, W)\right)^2\right]\mathbb{E}\left[\left(\hat{g}(X,\, Z) - g(X,\, Z)\right)^2\right] = o_P\left(\frac{1}{n}\right).$$

Suppose moreover that Π has finite **VC-dimension**. Then,

$$R\left(\hat{\pi}\right) = \mathcal{O}_P\left(\sqrt{S\,\mathrm{VC}\left(\Pi\right) / \, n}\right),$$

with $S = \mathbb{E}\left[\left(\tau_m(X_i) + g(X_i,\, Z_i)\left(Y_i - m(X_i,\, W_i)\right)\right)^2\right]$.

Proof Ingredients

- ATE literature looks at efficient coupling of the **efficient score**: $|\widehat{A}(\pi) - \widetilde{A}(\pi)| = o_P(1/\sqrt{n})$, where

$$\widetilde{A}(\pi) = \sum_{i=1}^{n} (2\pi(X_i) - 1)\left(\tau_m(X_i) + g(X_i,\, Z_i)\left(Y_i - m(X_i,\, W_i)\right)\right).$$

- Here, need **uniform coupling**:
  $\sup\left\{\left|\widehat{A}(\pi) - \widetilde{A}(\pi)\right| : \pi \in \Pi\right\} = o_P(1/\sqrt{n})$.
  - Our uniform coupling result is specific to the **doubly robust** construction, and may not hold for other estimators that are efficient at a single π, e.g., empirical IPW (Hirano et al., 2003).
- Next: **concentration** of $\widetilde{A}(\pi)$ over the class $\pi \in \Pi$. Defining

$$S = \mathbb{E}\left[\left(\tau_m(X_i) + g(X_i,\, Z_i)\left(Y_i - m(X_i,\, W_i)\right)\right)^2\right],$$

remix Dudley's classical **chaining argument** to verify that

$$\sup\left\{\left|\widetilde{A}(\pi) - A(\pi)\right| : \pi \in \Pi\right\} = \mathcal{O}_P\left(\sqrt{\frac{S\ VC(\Pi)}{n}}\right).$$

Bound **Rademacher complexity** via chaining.

Lower bounds

Any statement about lower bounds depends on how **general** we want to be, and how **adaptive** we want to be to problem structure. We proved that $R(\hat{\pi}) = \mathcal{O}_P(\sqrt{S\,\mathrm{VC}(\Pi)/n})$, and argue that this is optimal. We first note, however:

- If treatment effects are **smaller** than $1/\sqrt{n}$, bound is loose.
- If treatment effects are **very large**, this bound is loose as finding the optimal rule is easy (Luedtke and Chambaz, 2017).
- VC-dimension may be a loose summary of the **complexity** of Π (Bartlett and Mendelson, 2006).

We show that our bound is **sharp** when our treatment effects scale as $1/\sqrt{n}$ and we summarize complexity via VC-dimension. Similar **local asymptotics** are also used by Hirano and Porter (2009).

NB: Our upper bound allows the data-generating distribution (and Π) to change with n, so changing $\tau(\cdot)$ with n is valid.

Lower bounds

In the **unconfoundedness** setting, define a sequence of problems

$$X_i \sim \text{Uniform}\left(\mathcal{X}_s\right), \quad W_i \,\big|\, X_i \sim \text{Bernoulli}(e(X_i)),$$
$$Y_i \,\big|\, X_i,\, W_i \sim \left(y(X_i) + (W_i - e(X_i))\,\frac{\tau(X_i)}{\sqrt{n}},\, \sigma^2(X_i)\right).$$

Theorem. (Athey and W., 2018) In this setting, there is a class Π with $\text{VC}(\Pi) = d$ whose **minimax regret** satisfies

$$\liminf_{n\to\infty} \left\{ \sqrt{n} \inf_{\hat{\pi}_n} \left\{ \sup_{|\tau(x)|\leq C} \left\{ \mathbb{E}\left[R_n\left(\hat{\pi}_n\right)\right] \right\} \right\} \right\} \geq 0.33\sqrt{Sd},$$

where $S = \mathbb{E}\left[\sigma^2(X)/(e(X)(1-e(X)))\right]$.

Our method **achieves this bound** up to a universal constant. Other methods do not, e.g., for **IPW** with known propensity scores, Kitagawa & Tetenov (2018) prove a bound that depends on $\sup\left\{|Y_i|\right\} / \inf\left\{e(X_i),\, (1-e(X_i))\right\}$ instead of $\sqrt{S}$.

Conclusion

- Machine learning based methods very useful to analyze heterogeneous treatment effects and targeted policies
- Methods work in a variety of design settings (experiments, unconfounded, IV)
- Methods can give either simple or very complex policies, with statistical guarantees
- Online learning can help discover good policies
- Semi-parametric efficiency literature guides methods