

# Counterfactual Inference

---

SUSAN ATHEY

STANFORD UNIVERSITY



**Matt Ford** @fordm · 23h

2008: Twitter is a fun microblogging service you can use to keep track of Ashton Kutcher

2018: Twitter is the president's preferred tool for witness tampering in a federal criminal investigation



87



6.4K



31K



## Stability of Black-Box ML

# Artificial Intelligence/Machine Learning Desired Properties for Applications

---

## DESIRED PROPERTIES

Interpretability

Stability/Robustness

Transferability

Fairness/Non-discrimination

“Human-like” AI

- Reasonable decisions in never-experienced situations

## CAUSAL INFERENCE FRAMEWORK

Goal: learn model of how the world works

- Impact of interventions can be context-specific
- Model maps contexts and interventions to outcomes
- Formal language to separate out correlates and causes

Ideal causal model is by definition stable, interpretable

Transferability: straightforward for new context dist'n

Fairness: Many aspects of discrimination relate to correlation v. causation

- Performance may depend on physical and mental ability, psychological factors (e.g. risk taking)
- Gender and race may be correlated with factors that shift these distributions, relatively limited direct causal effects

# Artificial Intelligence/Machine Learning Desired Properties for Applications

---

## DESIRED PROPERTIES

Interpretability

Stability/Robustness

Transferability

Fairness

## CAUSAL INFERENCE FRAMEWORK

Goal: learn model of how the world works

- Impact of interventions can be context-specific
- Model maps contexts and interventions to outcomes
- Formal language to separate out correlates and causes

In practice, challenges remain, e.g. due to:

Lack of quasi-experimental data for estimation;

Unobserved contexts/confounders or insufficient data  
to control for observed confounders;

Analyst's lack of knowledge about model

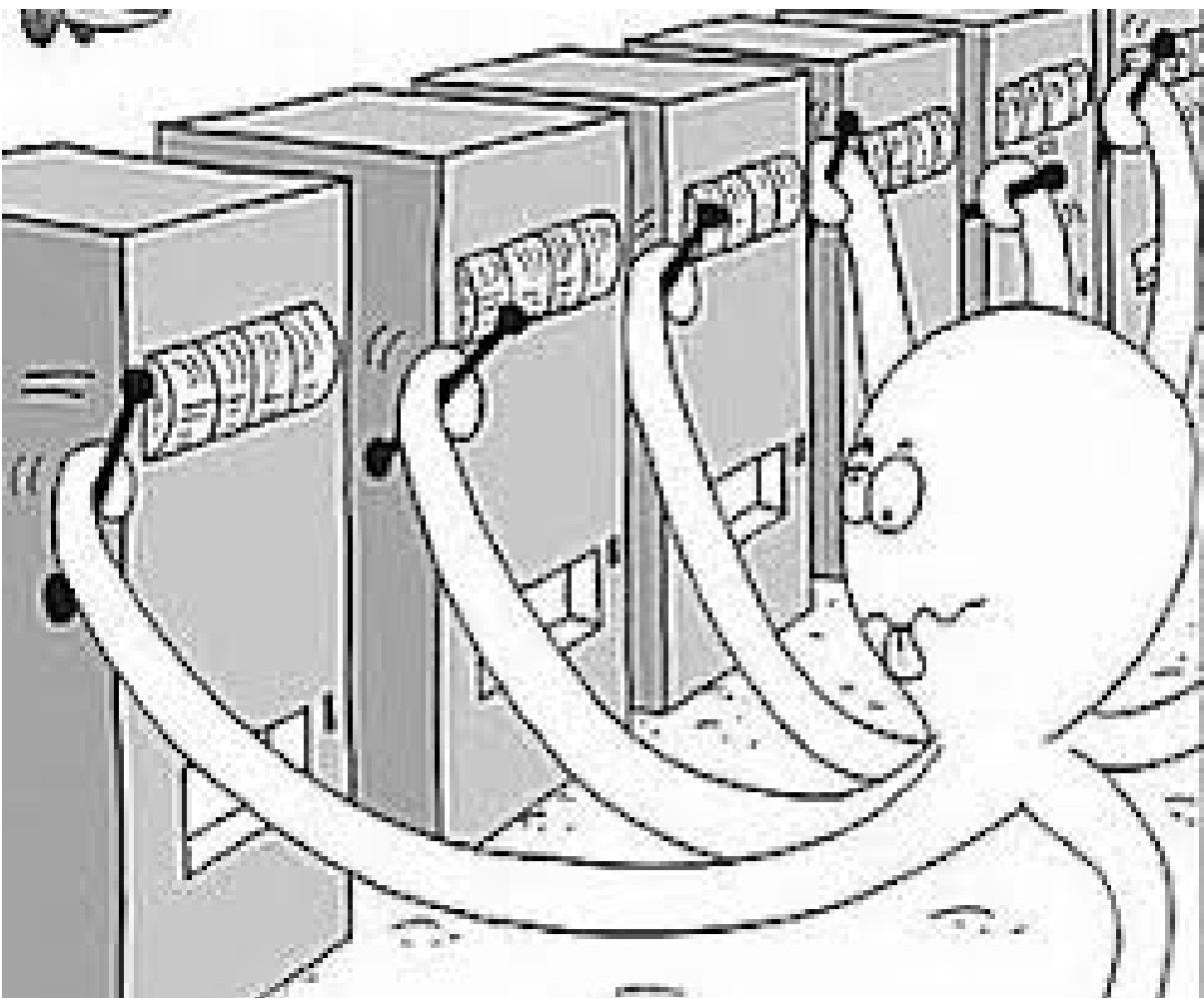
interpretable

context dist'n

elate to

mental ability,

ctors that shift  
t causal effects



# Artificial Intelligence and Counterfactual Estimation

---

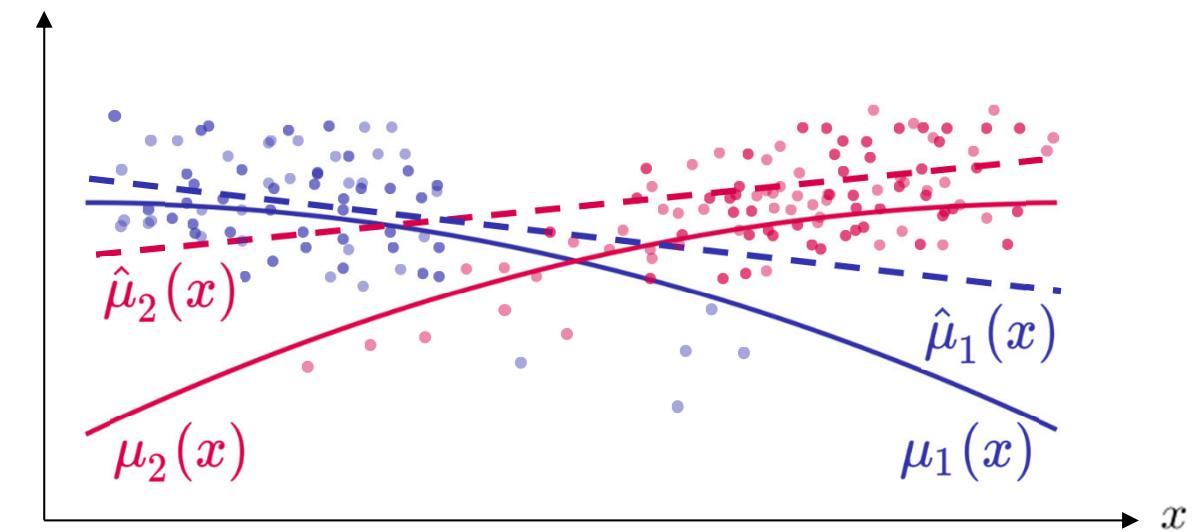
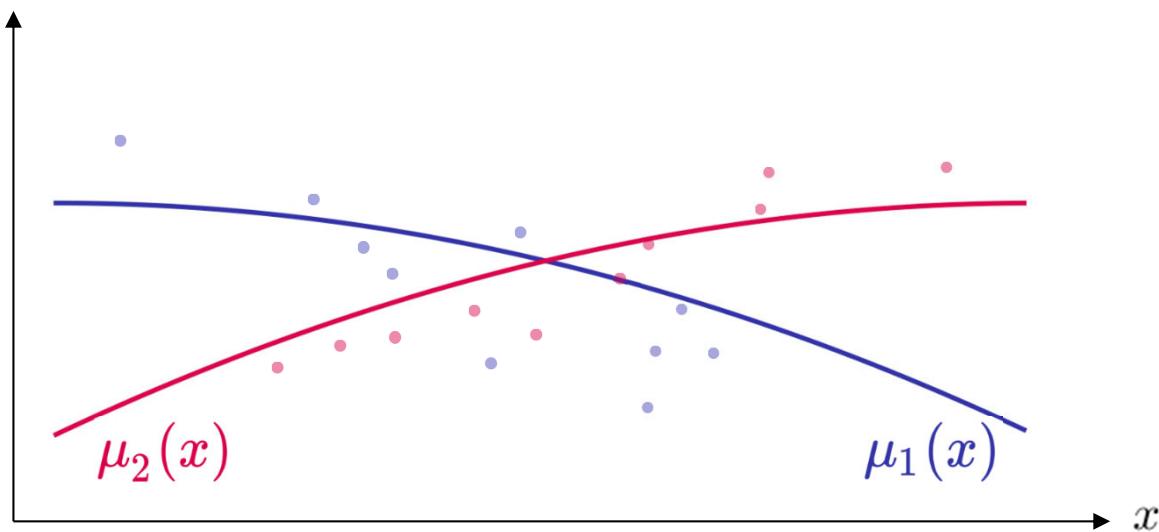
## Artificial intelligence

- Select among alternative choices
- Explicit or implicit model of payoffs from alternatives
- Learn from past data
- Initial stages of learning have limited data
- Inside the AI is a statistician performing counterfactual reasoning
- Statistician should use best performing techniques (efficiency, bias)

Simple example: contextual bandit

# Estimation is challenging: Contextual Bandit example

- Inherent bias in estimation due to adaptive assignment of contexts to arms.
  - context assigned to arm with highest reward sample or confidence bound
  - creates systematically unbalanced data



# Counterfactual Inference Approaches

“Program  
evaluation”,  
“treatment effect  
estimation”

What was the impact of the policy?

- Minimum wage, training program, class size change, etc.

Did the advertising campaign work?  
What was the ROI?

Do get-out-the vote campaigns work?

What is an optimal policy assigning workers to training programs?

# Counterfactual Inference Approaches

“Program  
evaluation”,  
“treatment effect  
estimation”

**Goal: estimate the impact of interventions or treatment assignment policies**

- Low dimensional intervention

Estimands

- Average effect
- Heterogeneous effects
- Optimal policy

Confidence intervals

Designs that enable identification and estimation of these effects

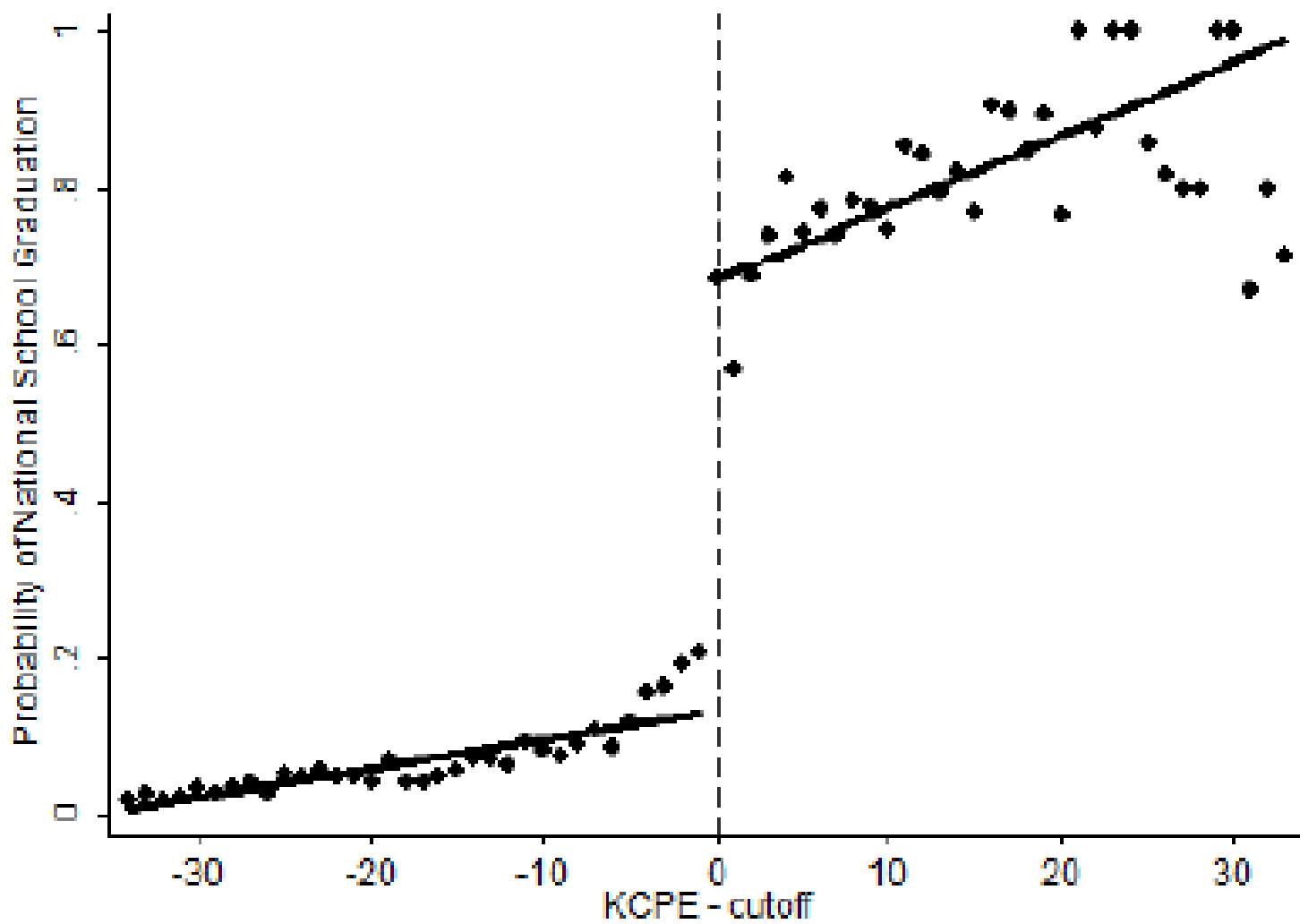
- (Alternative treatments observed in historical data in relevant contexts)
- Randomized experiments
- “Natural” experiments (Unconf., IV)
- Regression discontinuity
- Difference-in-difference
- Longitudinal data
- Randomized and natural experiments in social network/settings w/ interference

# Treatment Effect Estimation: Designs

Regression Discontinuity Design

Mbiti & Lucas (2013) estimate impact of secondary school quality on student achievement in Kenya.

Discontinuity: cut-off on the primary exit exam required to get into better secondary schools



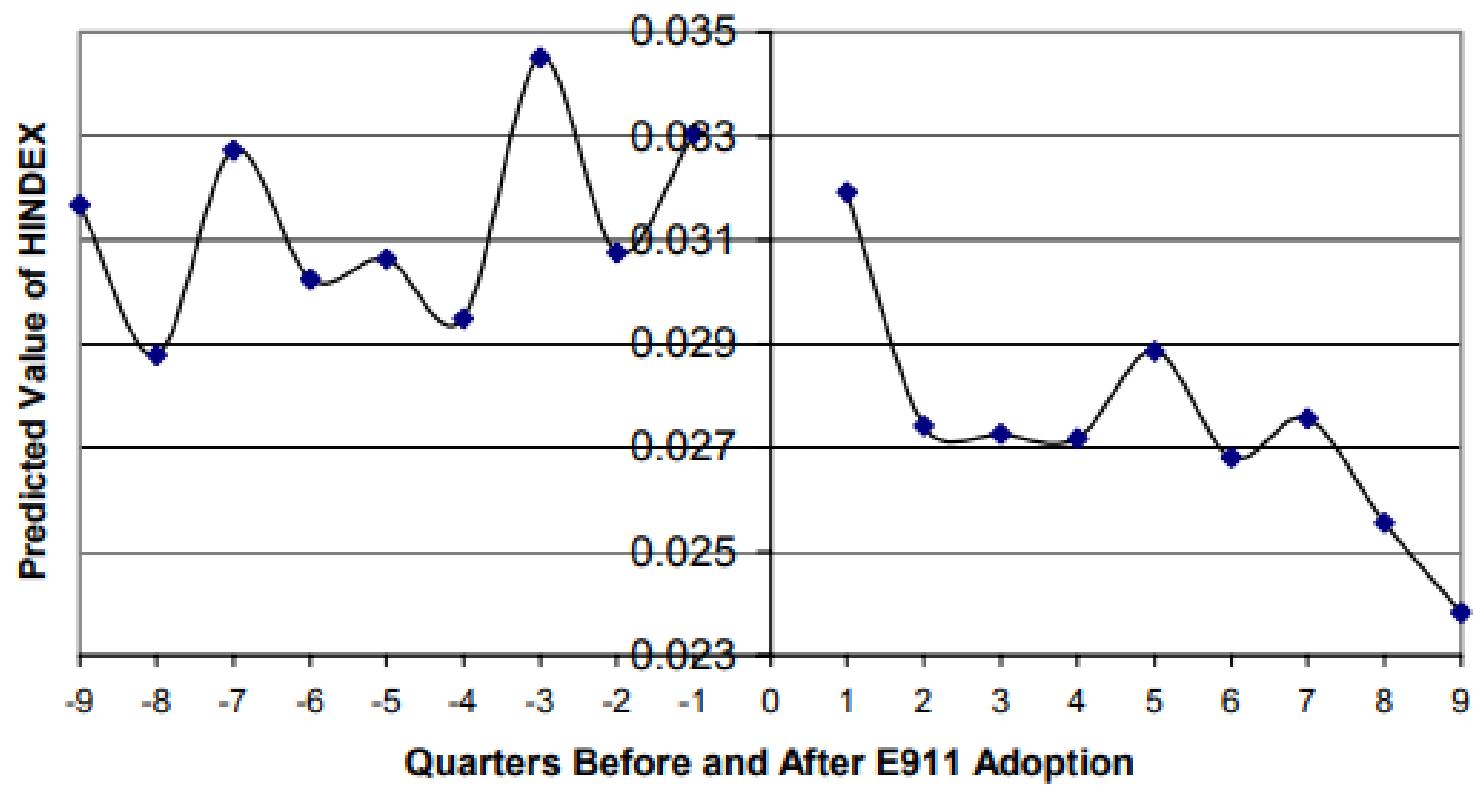
# Treatment Effect Estimation: Designs

Difference-in-Difference Designs

Athey and Stern (2002) look at the impact of Enhanced 911 (automated address lookup) on health outcomes for cardiac patients

Counties adopt at different times; estimate time trend using other counties to determine counterfactual outcomes in the absence of adoption

**Figure C: Effect of Time Before and After E911 Adoption on HINDEX**



# Counterfactual Inference Approaches

“Structural  
estimation”,  
“Generative  
Models” &  
Counterfactuals

What would happen to firm demand if price increases?

What would happen to prices, consumption, consumer welfare, and firm profits if two firms merge?

What would happen to platform revenue, advertiser profits and consumer welfare if Google switched from a generalized second price auction to a Vickrey auction?

# Counterfactual Inference Approaches

“Structural  
estimation”,  
“Generative  
Models” &  
Counterfactuals

Goal: estimate **impact on welfare/profits of participants** in alternative counterfactual regimes

- Counterfactual regimes **may not have ever been observed** in relevant contexts
- Need behavioral model of participants

Still need designs that enable identification and estimation, now of preference parameters

- E.g. need to see changes in prices to understand price sensitivity

Use “revealed preference” to uncover preference parameters

Rely on behavioral model to estimate behavior in different circumstances

- Also may need to specify equilibrium selection

Dynamic structural models

- Learn about value function from agent choices in different states
- See Igami (2018) who relates to AI

# Counterfactual Inference Approaches

“Structural  
estimation”,  
“Generative  
Models” &  
Counterfactuals

## Advertiser Profit Maximization Example

- Bidder in search advertising auctions has value-per-click  $v$
- $Q(b)$  is the share of available ad clicks per search from bidding  $b$  per click; upward sloping
- Bidder profit per search:

$$Q(b) \cdot (v - b)$$

- Bidder first order condition:

$$v = b + \frac{Q(b)}{Q'(b)}$$

## Inferring preferences (value per click) from data

- Analyst estimates  $Q()$  from historical log data
- For each advertiser, can infer the value  $v$  that rationalizes bid (satisfies FOC)

## Counterfactuals

- With knowledge of advertiser values and behavioral model, can solve for new equilibria
- Changing auction format
- Changing quality scores

See: Athey and Nekipelov (2012)

# Counterfactual Inference Approaches

## Dynamic Structural Estimation

## Inverse Reinforcement Learning

### Single Agent Decision Problem

- Rust (1987) studies problem of a decision-maker replacing bus engines
  - Analogous to a grand master playing chess
- Agent maximizes discounted sum of profits
- Using principles of dynamic programming, Bellman equation is:

$$V(s) = \max_{s' \in F(s)} \pi(s', s; \theta, \epsilon) + \delta V(s')$$

- Policy function :

$$\sigma(s; \theta) = \arg \max_{s' \in F(s)} \pi(s', s; \theta, \epsilon) + \delta V(s')$$

- Assume stochastic shock  $\epsilon$  to flow profits

### Solution: Nested fixed point

- Outer loop: Optimize likelihood function for  $\theta$ , where data are (state,action) pairs and model predicts optimal actions as function of  $\theta$
- Inner loop:
  - Given  $\theta$  , solve for value function by iterating over Bellman's equation
  - Evaluate policy function given value function, and evaluate likelihood

See: Igami (2018) who develops relationship between this and Bonanza algorithm; also analysis of AlphaGo algorithm relative to Hotz and Miller (1993)

# Counterfactual Inference Approaches

## Dynamic Structural Estimation

## Inverse Reinforcement Learning

What can we learn from decades of methodological and empirical work in economics, that is relevant for AI?

- Applications to human or firm behavior are challenging
- Conceptual framework has been clear from 80s and 90s
- Big problem: not enough training data, and not enough knowledge about game payoffs to create artificial training data
- Economics has some insights to help in data-poor environments...
  - Use as much structure as is known, carefully examine functional forms for how they extrapolate
  - Think about independence assumptions and biases that might arise from diff't training data
  - Take behavioral models seriously to draw better inference from agent behavior
  - See Igami (2018) for some more discussion

How can recent advances in AI help solve economic problems?

- New algorithms of past 10-15 years in ML/AI focus on computational performance and problems with large state spaces
  - Coupled with games that can be played by computers with large number of repetitions, generating very large datasets
  - The rules are clear, so possible to test different strategies against one another
  - The analyst knows the mapping from final state to payoffs, just doesn't know the value function at intermediate states
- In economic problems
  - Computational advances definitely help in problems with large state spaces...
  - But the analyst doesn't know the per-period payoff function, and thus doesn't know enough about the game to simulate play and know what the final payoffs are.
  - Can only do that given parameter values.

# Counterfactual Inference Approaches

“Causal discovery”,  
“Learning the causal  
graph”

Goal: uncover the causal structure of a system

- Many observed variables
- Analyst believes that there is an underlying structure where some variables are causes of others, e.g. a physical stimulus leads to biological responses

Focus on ways to test for causal relationships

Applications

- Understanding software systems
- Biological systems

# Counterfactual Inference Approaches

Recently, literatures  
have started coming  
together

Multiple literatures on causality within  
economics, statistics, and computer science

Different ways to represent equivalent  
concepts

Common themes: very important to have  
formal language to represent concepts

Recent literatures: **Bring causal reasoning,  
statistical theory and modern machine  
learning algorithms together to solve  
important problems**

# Preview of Themes

---

## Causal inference v. supervised learning

- Supervised learning: can evaluate in test set in model-free way
- Causal inference
  - Parameter estimation-parameter not observed in test set
  - Change objective function, e.g. consistent parameter estimation
  - Can estimate objective (MSE of parameter), but often requires maintained assumptions
  - Often sampling variation matters even in large data sets
  - Requires theoretical assumptions and domain knowledge
  - Tune for counterfactuals: distinct from tuning for fit, also different counterfactuals select different models

## Insights from statistics/econometrics

- Consider identification, then estimation
  - Could you solve problem with infinite data?
  - Design-based approach
  - Estimation: scaled up with many experiments
- Regularization induces omitted variable bias
- Omitted variables challenge causal inference, interpretability, fairness
- Semi-parametric efficiency theory can be helpful, brings insights not commonly exploited in ML
  - Cross-fitting/out of bag estimation of nuisance parameters
  - Orthogonal moments/double robustness
  - Use best possible statistician inside bandits/AI agents
- Exploit structure of problem carefully for better counterfactual predictions
  - Black-box algorithms reserved for nuisance parameters

# Estimating ATE under Unconfoundedness

---

SOLVING CORRELATION V. CAUSALITY BY CONTROLLING FOR  
CONFOUNDERS

# Setting

---

Only observational data is available

Analyst has access to data that is sufficient for the part of the information used to assign units to treatments that is related to potential outcomes

Analyst doesn't know exact assignment rule and there was some randomness in assignment

Conditional on observables, we have random assignment

Lots of small randomized experiments

Application: logged tech company data, contextual bandit data

# Example: Effect of an Online Ad

---

Ads are targeted using cookies

User sees car ads because advertiser knows that user visited car review websites

Cannot simply relate purchases for users who saw an ad and those who did not:

- Interest in cars is unobserved confounder

Analyst can see the history of websites visited by user

- This is the main source of information for advertiser about user interests

# Setup

---

Assume unconfoundedness/ignorability:

- $Y_i(1), Y_i(0) \perp W_i | X_i$

Assume overlap of the propensity score:

$$p(x) = \Pr(W_i = 1 | X_i = x) \in (0,1)$$

Then Rubin shows:

- Sufficient to control for propensity score:
- $Y_i(1), Y_i(0) \perp W_i | p(X_i)$
- If control for X well, can estimate ATE
- $E[Y_i(1) - Y_i(0)]$ .

# Intuition for Most Popular Methods

---

Control group and treatment group are different in terms of observables

Need to **predict cf outcomes for treatment group if they had not been treated**

**Weighting/Matching:** Since assignment is random conditional on X, solve problem by reweighting control group to look like treatment group in terms of distribution of X

- P.S. weighting/matching: need to estimate p.s., cannot perfectly balance in high dimensions

**Outcome models:** Build a model of  $Y|X=x$  for the control group, and use the model to predict outcomes for x's in treatment group

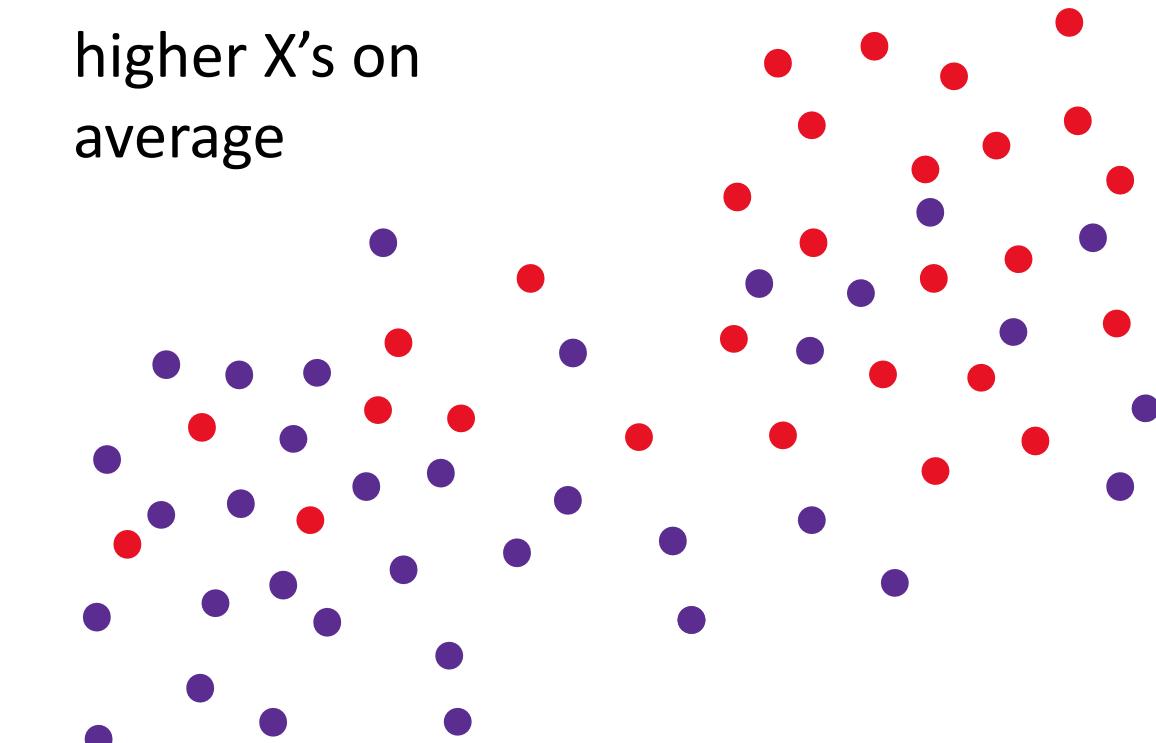
- If your model is wrong, you will predict incorrectly

**Doubly robust:** methods that work if either p.s. model OR model  $Y|X=x$  is correct

Y

Treated  
observations have  
higher X's on  
average

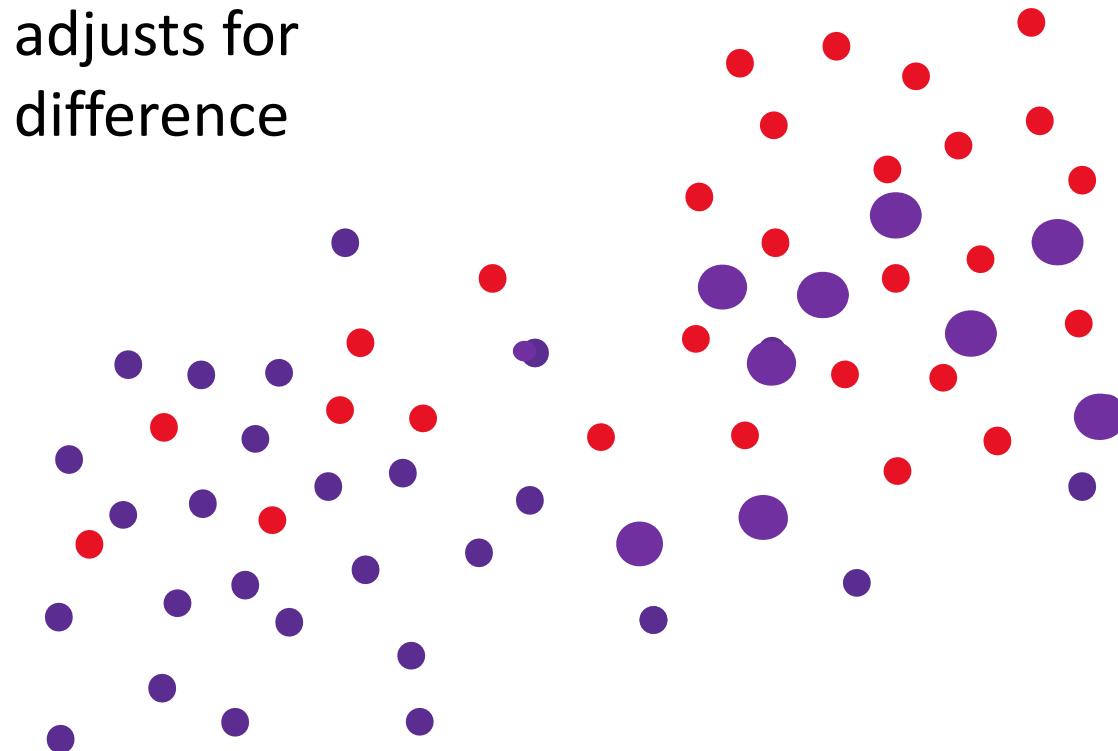
X



Y

Reweighting control  
observations with  
high X's  
adjusts for  
difference

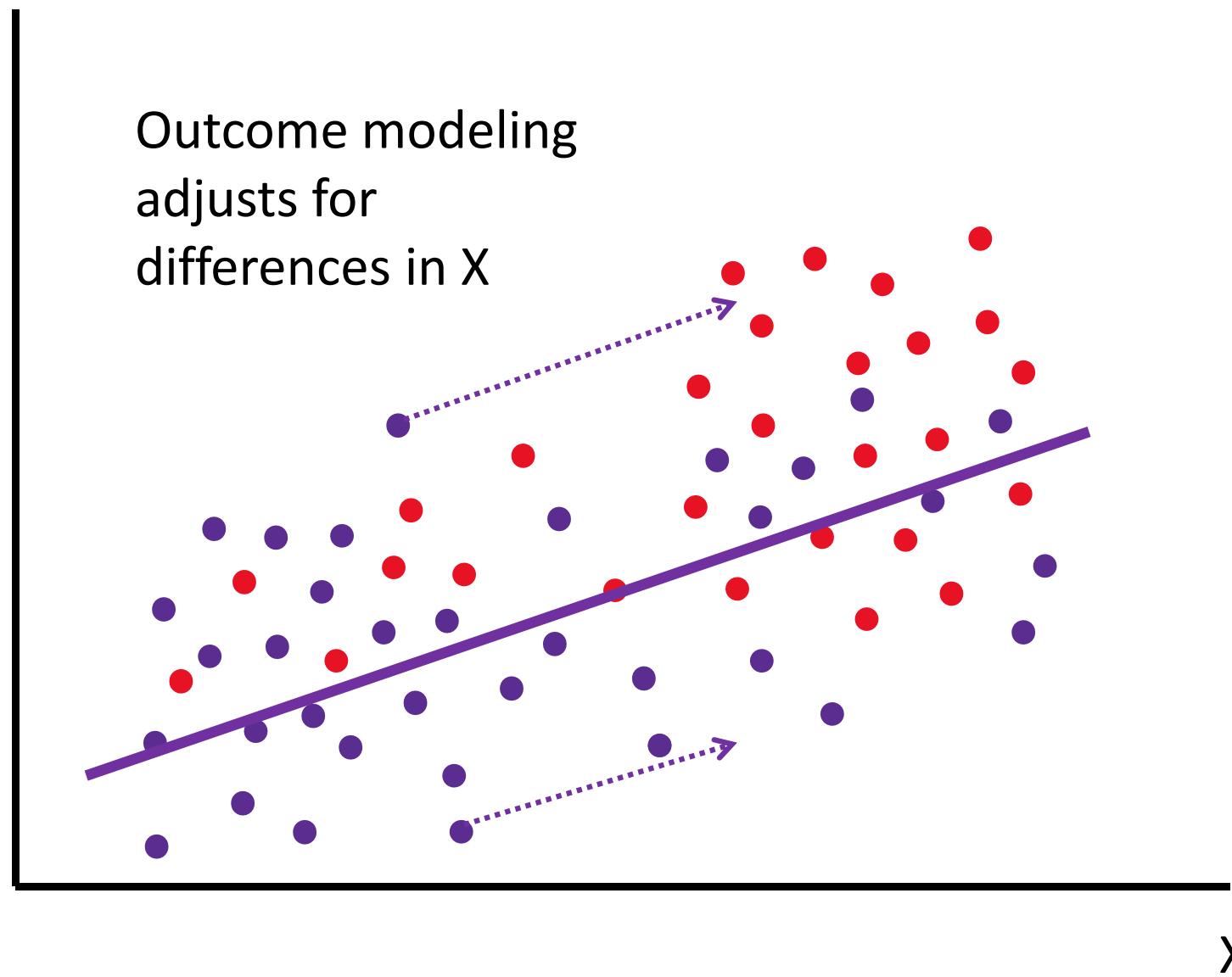
X



Y

Outcome modeling  
adjusts for  
differences in X

X

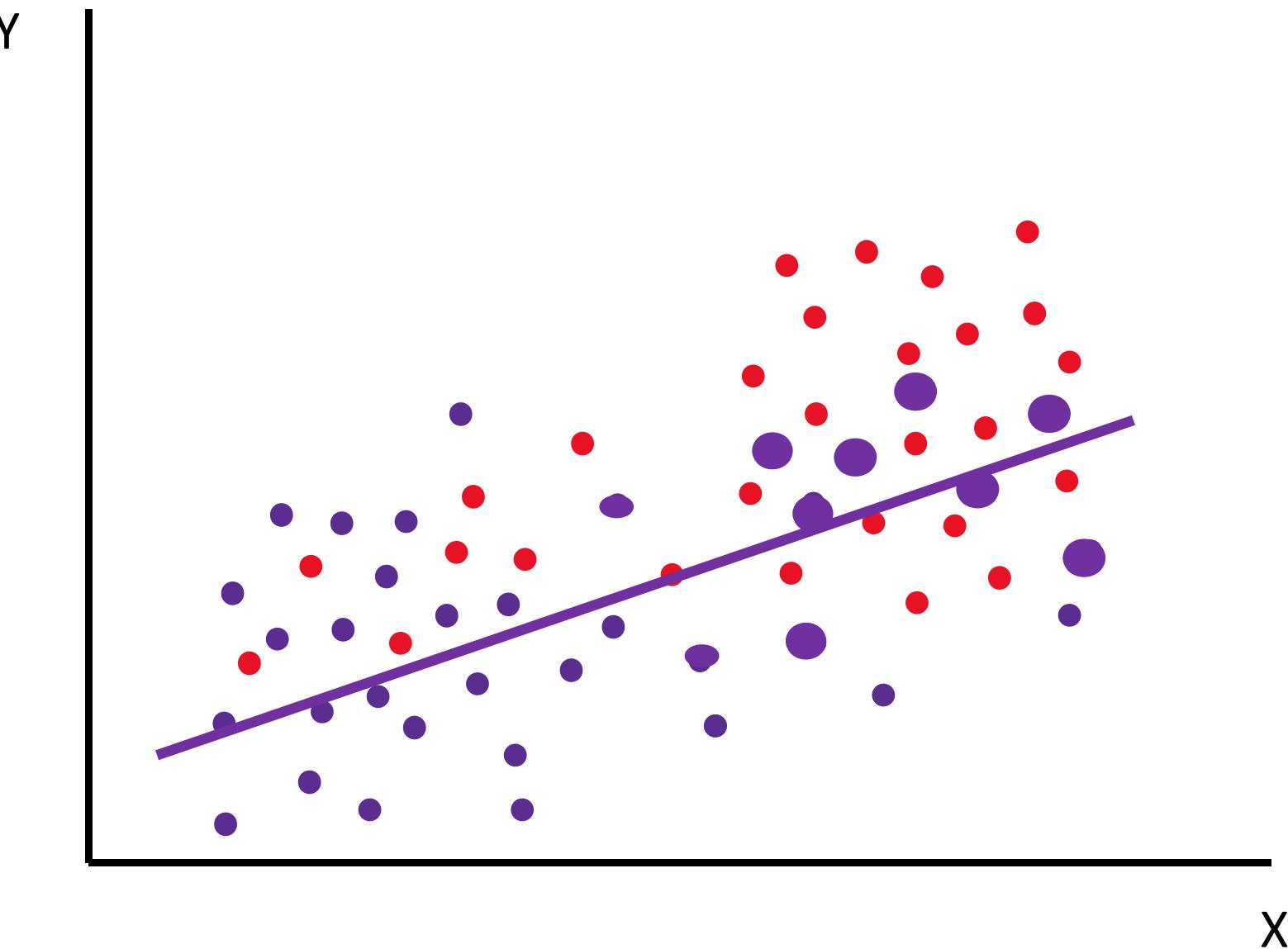


Reweighting control  
observations with high  
 $X$ 's

AND using outcome  
modeling is doubly  
robust

With correct  
reweighting, don't  
need to adjust  
outcomes

With outcome  
adjustments, don't  
need to reweight



## Using Supervised ML to Estimate ATE Under Unconfoundedness

Method I:  
Propensity score  
weighting or KNN on  
propensity score

- LASSO to estimate propensity score; e.g. McCaffrey et al. (2004); Hill, Weiss, Zhai (2011)

## Using Supervised ML to Estimate ATE Under Unconfoundedness

## Method II: Regression adjustment

- Belloni, Chernozukov, Hansen (2014):
  - LASSO of  $W \sim X; Y \sim X$
  - Regress  $Y \sim W$ , union selected  $X$
  - Sacrifice predictive power (for  $Y$ ) for causal effect of  $W$  on  $Y$
- **Contrast w/ off-the-shelf supervised learning**
  - Off-the-shelf LASSO  $Y \sim X, W$  does not select all  $X$ 's that are confounders
  - Omitting confounders leads to biased estimates
  - Prioritize getting the answer right about treatment effects

## Using Supervised ML to Estimate ATE Under Unconfoundedness

## Method III: Estimate CATE and take averages

- Hill (2011) uses BART (Chipman, 2008) or other flexible method to estimate
$$\mu(x; w) = E[Y_i | X_i = x, W_i = w]$$
  - Estimate ATE as  $E[\hat{\mu}(X_i; 1) - \hat{\mu}(X_i; 0)]$
  - See further papers by Hill and coauthors
- 
- Performs well in contests, can use propensity adjustments in estimating conditional mean function
  - Performance relies on doing a good job estimating this outcome model—depends on DGP, signal-to-noise

# Using Supervised ML to Estimate ATE Under Unconfoundedness

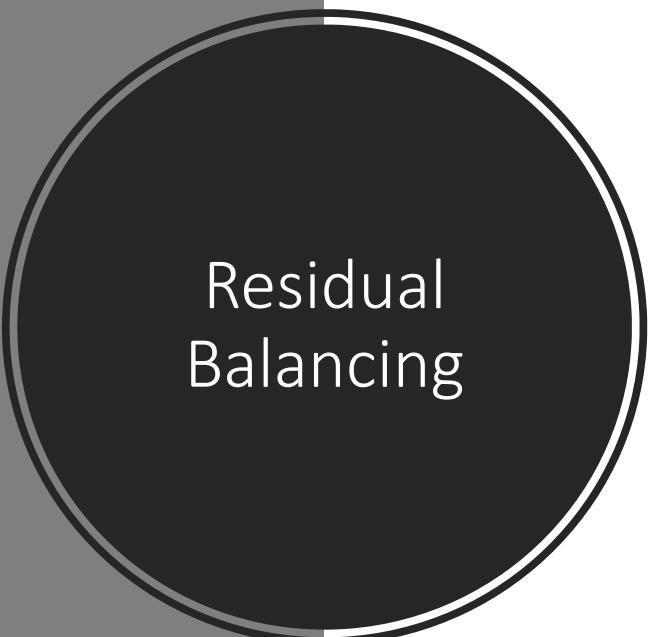
## Method IV: Double robust/double machine learning

- Cross-fitted augmented inverse propensity scores
  - These are the efficient scores (see literature on semi-parametric efficiency)
  - Orthogonal moments
  - Cross-fitted nuisance parameters:  $\hat{\tau}_{-i}(X_i)$ ,  $\hat{e}_{-i}(X_i)$ ,  $\hat{\mu}_{-i}(X_i; W_i)$ , e.g. OOB random forest
  - Score given by
  - $$\Gamma_i = \hat{\tau}_{-i}(X_i) + \frac{W_i - \hat{e}_{-i}(X_i)}{(1 - \hat{e}_{-i}(X_i))\hat{e}_{-i}(X_i)} (Y_i - \hat{\mu}_{-i}(X_i; W_i))$$
  - ATE is average of  $\Gamma_i$
  - DR: consistent estimates if either propensity score OR outcome correct
  - Can get  $\sqrt{n}$  convergence even if nuisance parameters converge more slowly, at rate  $n^{1/4}$ , which helps in high dimensions

## Using Supervised ML to Estimate ATE Under Unconfoundedness

### Method V: Residual Balancing

- Athey, Imbens and Wager (JRSS-B, 2018)
- Avoids assuming a sparse model of  $W \sim X$ , thus allowing applications with complex assignment
  - Not just slow convergence of assignment model—assignment model does not need to be estimated at all!
- LASSO  $Y \sim X$
- Solve a programming problem to find weights that minimize difference in  $X$  between groups
- Maintains the orthogonal moment form



## Residual Balancing

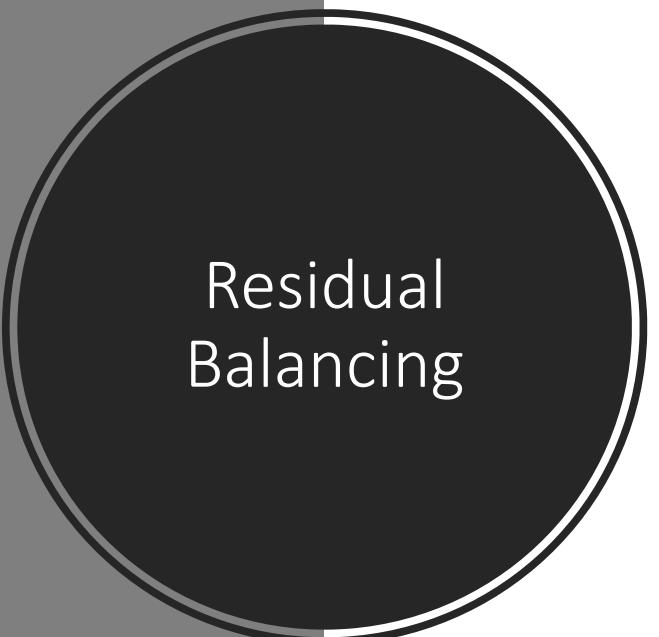
Consider the general class of estimators  $\hat{\tau} = \hat{m}^{(1)} - \hat{m}^{(0)}$ ,

$$\hat{m}^{(1)} = \bar{X} \cdot \hat{\beta}^{(1)} + \sum_{\{i: W_i=1\}} \gamma_i^{(1)} (Y_i - X_i \cdot \hat{\beta}^{(1)}).$$

**Proposition.** (Athey, Imbens, Wager; 2016) Suppose that  $Y_i = X_i \cdot \beta^{(W_i)} + \varepsilon_i$ . Writing  $\text{err}^{(1)} = \hat{m}^{(1)} - \bar{X} \cdot \beta^{(1)}$ , we have

$$|\text{err}^{(1)}| \leq \left\| \bar{X} - X(1)^\top \gamma^{(1)} \right\|_\infty \left\| \hat{\beta}^{(1)} - \beta^{(1)} \right\|_1 + \left| \sum_{\{i: W_i=1\}} \gamma_i^{(1)} \varepsilon_i \right|.$$

Here  $X(1)$  is a subset of  $X$  corresponding to the treated cases.



## Residual Balancing

Motivated by this proposition, we estimate  $\hat{\tau}$  as follows.

1. Estimate  $\hat{\beta}^{(1)}$  using a lasso or elastic net (Hastie et al., 2015) on the treated cases.
2. Estimate weights  $\gamma^{(1)}$  by quadratic programming:

$$\gamma^{(1)} = \operatorname{argmin}_{\tilde{\gamma}^{(1)}} \left\{ \zeta \left\| \tilde{\gamma}^{(1)} \right\|_2^2 + (1 - \zeta) \left\| \bar{X} - X(1)^\top \tilde{\gamma}^{(1)} \right\|_\infty^2 \right\},$$

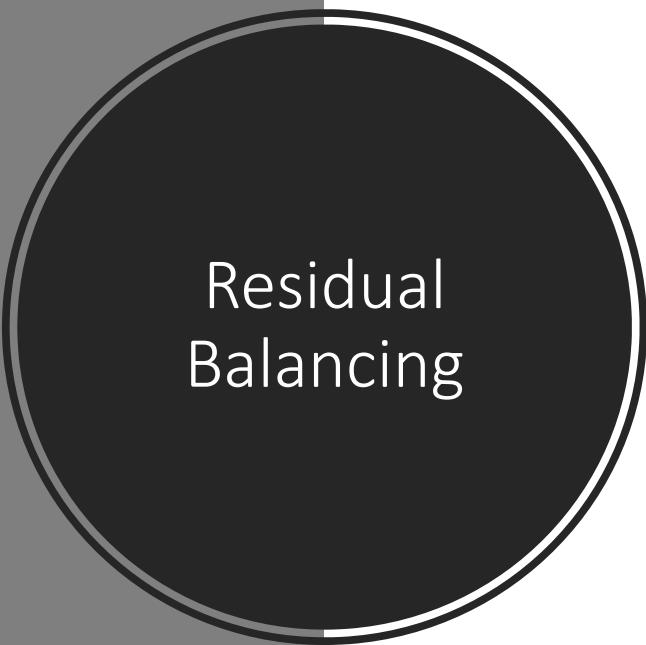
subject to constraints  $\gamma_i^{(1)} \geq 0$  and  $\sum_i \gamma_i^{(1)} = 1$ , where  $\zeta \in (0, 1)$  is a tuning parameter; see also Zubizarreta (2015).

3. Finally, our treatment effect estimate is  $\hat{\tau} = \hat{m}^{(1)} - \hat{m}^{(0)}$ ,

$$\hat{m}^{(1)} = \bar{X} \cdot \hat{\beta}^{(1)} + \sum_{\{i: W_i=1\}} \gamma_i^{(1)} (Y_i - X_i \cdot \hat{\beta}^{(1)}),$$

and  $\hat{m}^{(0)}$  is estimated analogously.

Software for R is available in the package `balanceHD`.



Residual  
Balancing

**Theorem.** (Athey, Imbens, Wager; 2016) Suppose that the following two conditions hold, along with standard assumptions (including the restricted eigenvalue condition):

Overlap:  $\eta < \mathbb{P} [W = 1 \mid X = x] < 1 - \eta$ , and

Sparsity:  $\|\beta^{(0/1)}\|_0 \ll \sqrt{n} / \log(p)$ .

Then approximate residual balancing is semiparametrically efficient:

$$(\hat{\tau} - \bar{\tau}) / \sqrt{\|\gamma^{(0)}\|_2^2 + \|\gamma^{(1)}\|_2^2} \Rightarrow \mathcal{N}(0, \sigma^2),$$

$$\limsup_{n \rightarrow \infty} n \left( \|\gamma^{(0)}\|_2^2 + \|\gamma^{(1)}\|_2^2 \right) \leq \mathbb{E} \left[ \frac{1}{e(X)} + \frac{1}{1 - e(X)} \right].$$

This formula also yields asymptotic confidence intervals for  $\bar{\tau}$ .

# Instrumental Variables

---

# What if unconfoundedness fails?

Alternate assumption: there exists an instrumental variable  $Z_i$  that is correlated with  $W_i$  (“relevance”) and where:

$$(Y_i(0), Y_i(1)) \perp Z_i | X_i$$

Treatment $W_i$	Instrument $Z_i$	Outcome $Y_i$
Military service	Draft Lottery Number	Earnings
Price	Fuel cost	Sales
Having 3 or more kids	First 2 kids same sex	Mom's wages
Education	Quarter of birth	Wage
Taking a drug	Assigned to treatment group	Health
Seeing an ad	Assigned to group of users advertiser bids on in experiment	Purchases at advertiser's web site

# Instrumental Variables: Binary Experiment Case

	Assigned to Treatment	Not Assigned to Treatment
Compliers	Treated	Not treated
Always-Takers	Treated	Treated
Never-Takers	Not treated	Not treated
Defiers	Not treated	Treated

# Different Estimands

Why not look at who was actually treated?

- Those who complied or defied were probably not random

Intention-to-treat (ITT)

- Compare average outcomes of those assigned to treatment with those assigned to control
- This may be interesting object if compliance will be similar when you actually implement the treatment, e.g. recommend patients for a drug

Local Average Treatment Effect (effect of treatment on compliers)

- Calculated as  $\text{ITT}/\Pr(\text{treat} \mid \text{assigned treatment}) = \text{ITT}/\Pr(W_i=1 \mid Z_i=1)$
- This clearly works if you can't get the treatment without being assigned to treatment group (no always-takers, no defiers)
- This also works as long as there are no defiers
- LATE is always larger than ITT

# Local Average Treatment Effects

Special case:  $W_i, Z_i$  both binary

Relevance:  $Z_i$  is correlated with  $W_i$

Exclusion:  $(Y_i(0), Y_i(1)) \perp Z_i$

Monotonicity: No defiers

Then the LATE is:

$$\frac{\mathbb{E}[Y_i | Z_i = 1] - \mathbb{E}[Y_i | Z_i = 0]}{\mathbb{E}[W_i | Z_i = 1] - \mathbb{E}[W_i | Z_i = 0]}$$

# Local Average Treatment Effects: Including Covariates

Special case:  $W_i, Z_i$  both binary

Relevance:  $Z_i$  is correlated with  $W_i$

Exclusion:  $(Y_i(0), Y_i(1)) \perp Z_i | X_i$

Monotonicity: No defiers

Then the LATE conditional on  $X_i = x$  is:

$$\frac{\mathbb{E}[Y_i | X_i = x, Z_i = 1] - \mathbb{E}[Y_i | X_i = x, Z_i = 0]}{\mathbb{E}[W_i | X_i = x, Z_i = 1] - \mathbb{E}[W_i | X_i = x, Z_i = 0]}$$

# IV Approaches: Including Covariates

Two-stage least squares approach

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2' X_i + \varepsilon_i$$

$$W_i = \gamma_0 + \gamma_1 Z_i + \gamma_2' X_i + \varepsilon_i$$

Chernozhukov et al:

- Use LASSO to select which  $X$ 's to include and partial them out
- If there are many instruments, use LASSO to construct the optimal instrument, which is the predicted value of  $W_i$
- Formally, estimate first stage using Post-LASSO
- In second stage, run 2SLS using predicted value of treatment as instrument
- Theorem: if model is sparse and instruments are strong, estimator is semi-parametrically efficient

Note: doesn't consider observable or unobservable heterogeneity of treatment effects

See also Peysakhovich & Eckles (2018)

# IV Approaches: Including Covariates

Two-stage least squares approach

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2' X_i + \varepsilon_i$$

$$W_i = \gamma_0 + \gamma_1 Z_i + \gamma_2' X_i + \varepsilon_i$$

Chernozhukov et al example:

- Angrist and Krueger quarter of birth paper
- Instruments: quarter of birth, and interactions with controls
- Using few instruments gives large standard errors

Estimator	Instruments	Schooling Coef	Rob Std Error
2SLS (3 IVs)	3	.10	.020
2SLS (All IVs)	1530	.10	.042
2SLS (LASSO IVs)	12	.10	.014

## Clicks as a Fraction of Top Position 1 Clicks

<i>Search phrase:</i>	iphone		viagra	
<i>Model:</i>	OLS	IV	OLS	IV
Top Position 2	0.66	0.67	0.28	0.66
Top Position 3	0.40	0.55	0.14	0.15
Side Position 1	0.04	0.39	0.04	0.13

User Model of Clicks:  
Results from Historical  
Experiments  
(Athey, 2010)

---

OLS Regression:

- Features: advertiser effects and position effects

IV Regression

- Project position indicators on A/B testid's.
- Regress clicks on predicted position indicators.

Estimates show smaller position impact than OLS, as expected.

Position discounts important for disentangling advertiser quality scores

## IV: Heterogeneous Treatment Effects

What if we want to learn about conditional average treatment effects (conditional on features?)

For simplicity, assume treatment effects are constant conditional on  $X$ .

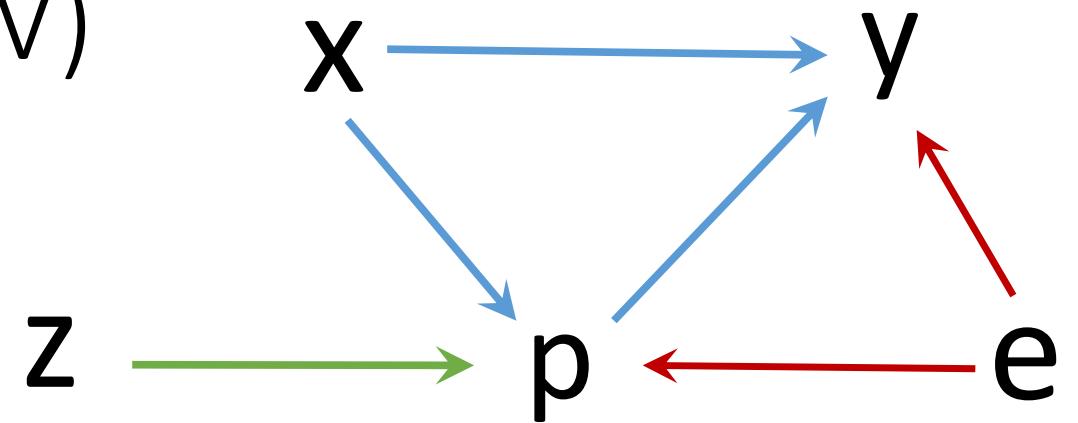
Illustrate with two approaches:

- Generalized random forests (Athey, Tibshirani, and Wager, Annals of Statistics, 2018)
  - Asymptotic normality and confidence intervals
- Deep Instrumental Variables (Taddy, Lewis, Hartford, Leyton-Brown (UBC))

Then apply to optimal policy estimation

- Athey and Wager (2016), Zhou, Athey and Wager (2018)

# Instrumental Variables (IV)



The *exclusion structure* implies

$$\mathbb{E}[y|x, z] = \int g(p, x)dF(p|x, z)$$

You can observe and estimate  $\widehat{\mathbb{E}}[y|x, z]$  and  $\widehat{F}(p|x, z)$

⇒ to solve for *structural*  $g(p, x)$  we have an inverse problem.

cf Newey+Powell 2003

$$\min_{g \in G} \sum \left( y_i - \int g(p, x_i) dF(p|x_i, z_i) \right)^2$$

**2SLS:**  $p = \beta z + \nu$  and  $g(p) = \tau p$  so that  $\int g(p) dF(p|z) = \tau \mathbb{E}[p|z]$

So you first regress  $p$  on  $z$  then regress  $y$  on  $\hat{p}$  to recover  $\hat{\tau}$ .

$$\min_{g \in G} \sum \left( y_i - \int g(p, x_i) dF(p|x_i, z_i) \right)^2$$

Or nonparametric sieves where  $g(p, x_i) \approx \sum_k \gamma_k \varphi_k(p, x_i)$  and

$$\mathbb{E}_F[\varphi_k(p, x_i)] \approx \sum_j \alpha_{kj} \beta_j(x_i, z_i) \text{ (Newey+Powell)}$$

or

$$\mathbb{E}_F[y_i - \sum_k \gamma_k \varphi_k(p, x_i)] \approx \sum_j \alpha_j \beta_j(x_i, z_i) \text{ (BCK, Chen+Pouzo)}$$

*Also Darolles et al (2011) and Hall+Horowitz (2005) for kernel methods.*

But this requires careful crafting and will not scale with  $\dim(x)$

$$\min_{g \in G} \sum \left( y_i - \int g(p, x_i) dF(p|x_i, z_i) \right)^2$$

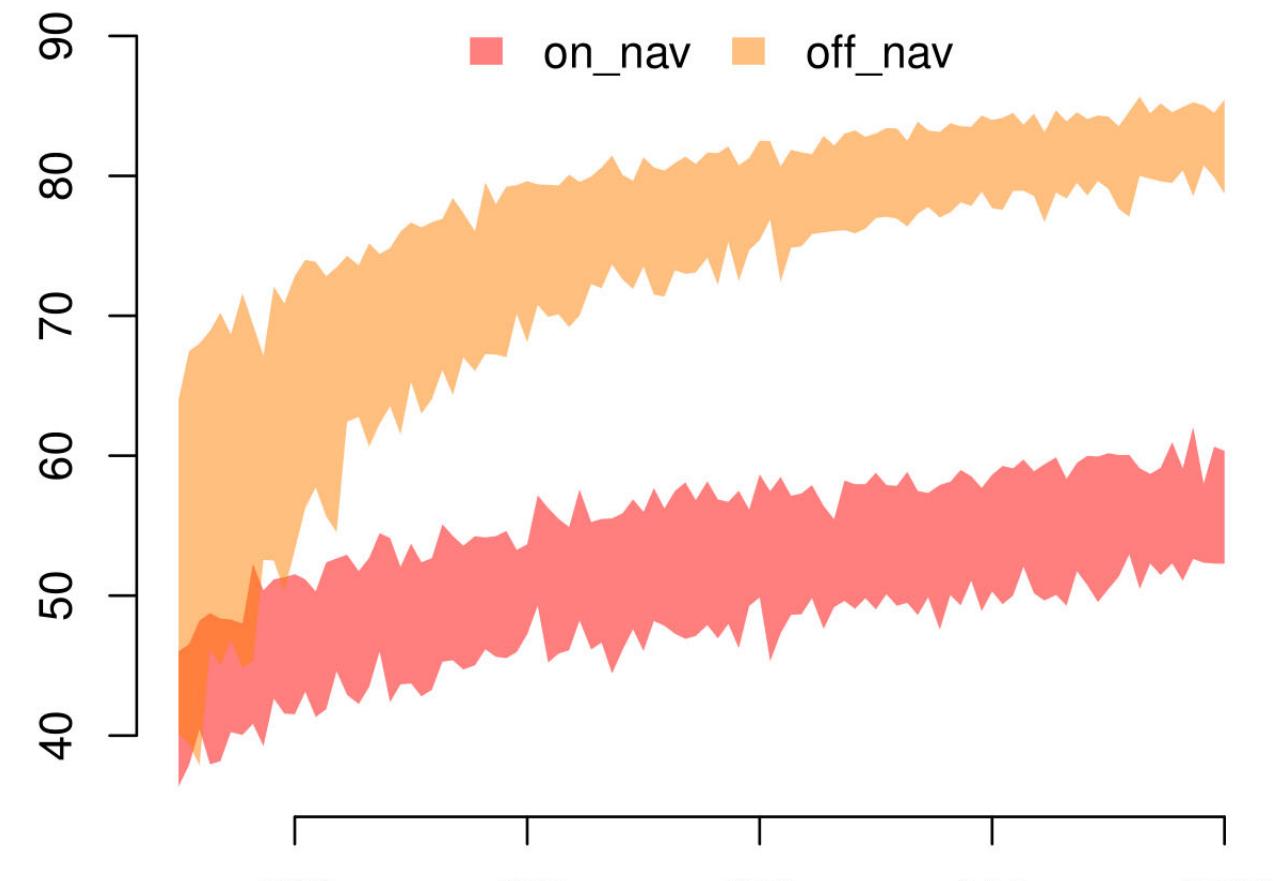
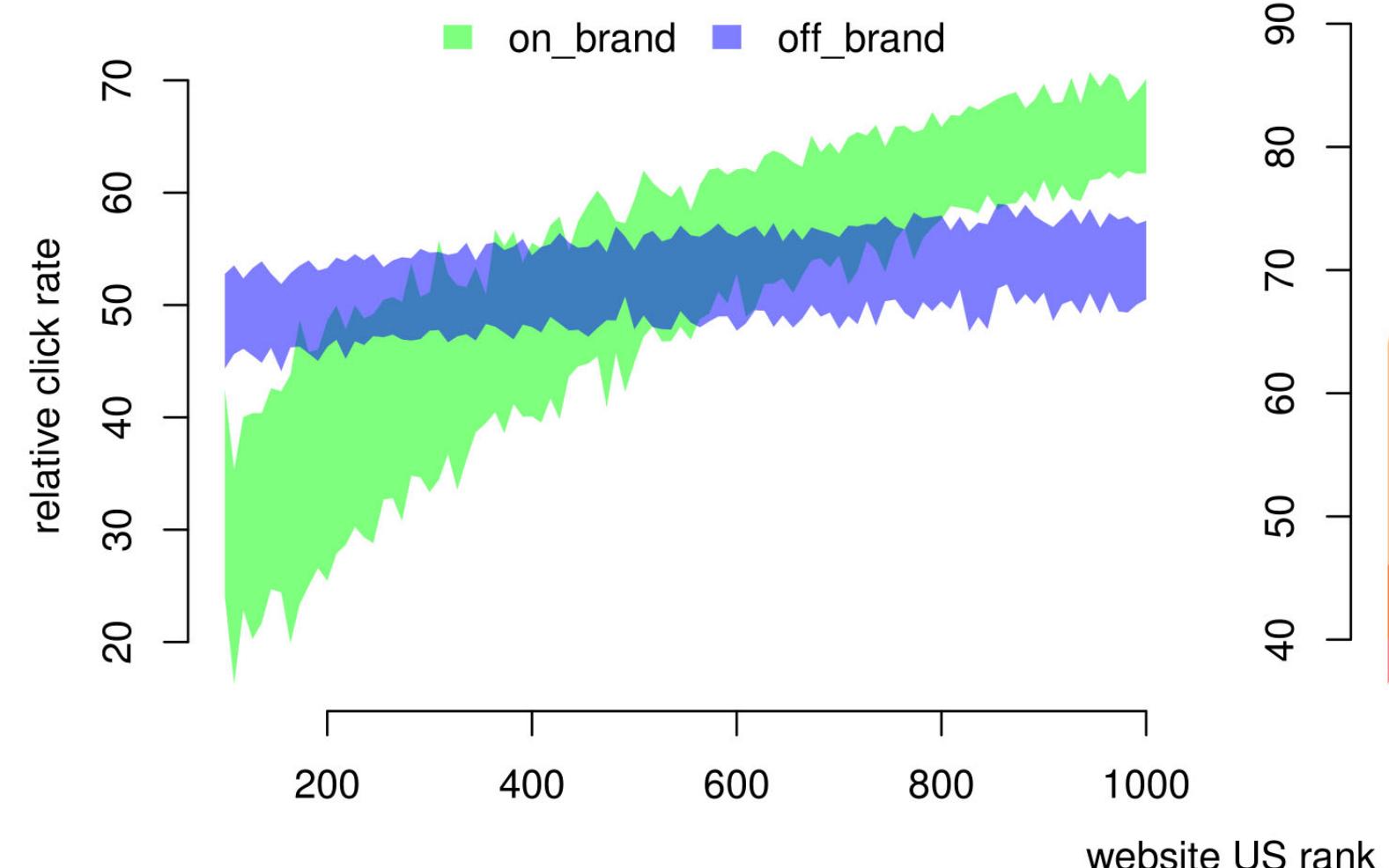
Instead, Deep IV **targets the integral loss function directly**

For discrete (or discretized) treatment

- Fit distributions  $\hat{F}(p|x_i, z_i)$  with probability masses  $\hat{f}(p_b|x_i, z_i)$
- Train  $\hat{g}$  to minimize  $[y_i - \sum_b g(\hat{p}_b, x_i) \hat{f}(p_b|x_i, z_i)]^2$

And you've turned IV into two *generic* machine learning tasks

# Search Ads Application of Deep IV: Relative Click Rate



Heterogeneity across advertiser and search

# Generalized Random Forests: Tailored Forests as Weighting Functions

---

- ▶ Local GMM/ML uses kernel weighting to estimate personalized model for each individual, weighting nearby observations more.
  - ▶ Problem: curse of dimensionality
- ▶ We propose forest methods to determine what dimensions matter for “nearby” metric, reducing curse of dimensionality.
  - ▶ Estimate model for each point using “forest-based” weights: the fraction of trees in which an observation appears in the same leaf as the target
- ▶ We derive splitting rules optimized for objective
- ▶ Computational trick:
  - ▶ Use approximation to gradient to construct pseudo-outcomes
  - ▶ Then apply a splitting rule inspired by regression trees to these pseudo-outcomes

Our parameter of interest,  $\theta(x)$ , is characterized by

$$\mathbb{E} [\psi_{\theta(x), \nu(x)}(O_i) \mid X_i = x] = 0 \text{ for all } x \in \mathcal{X},$$

where  $\nu(x)$  is an optional **nuisance parameter**.

- ▶ **Quantile regression**, where  $\theta(x) = F_x^{-1}(q)$  for  $q \in (0, 1)$ :

$$\psi_{\theta(x)}(Y_i) = q \mathbf{1}(\{Y_i > \theta(x)\}) - (1 - q) \mathbf{1}(\{Y_i \leq \theta(x)\})$$

- ▶ **IV regression**, with treatment assignment  $W$  and instrument  $Z$ . We care about the treatment effect  $\tau(x)$ :

$$\psi_{\tau(x), \mu(x)} = \begin{pmatrix} Z_i(Y_i - W_i \tau(x) - \mu(x)) \\ Y_i - W_i \tau(x) - \mu(x) \end{pmatrix}.$$

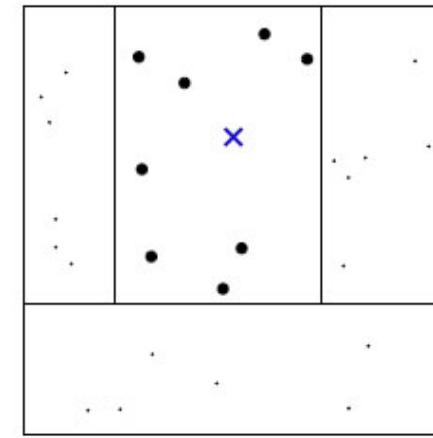
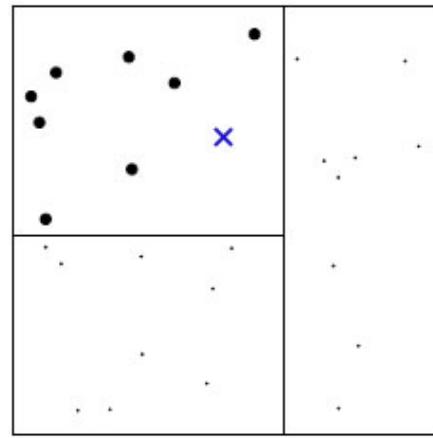
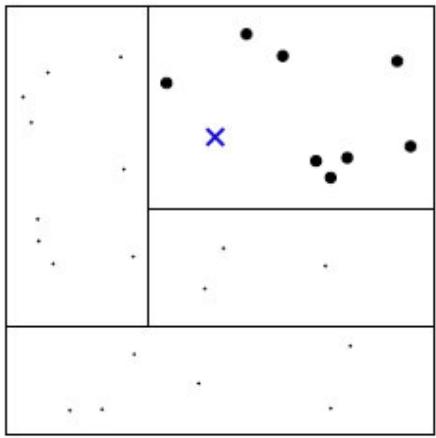
The classical approach is to rely on **local solutions** (Fan and Gijbels, 1996; Hastie and Tibshirani, 1990; Loader, 1999).

$$\sum_{i=1}^n \alpha(x; X_i) \psi_{\hat{\theta}(x), \hat{\nu}(x)}(O_i) = 0,$$

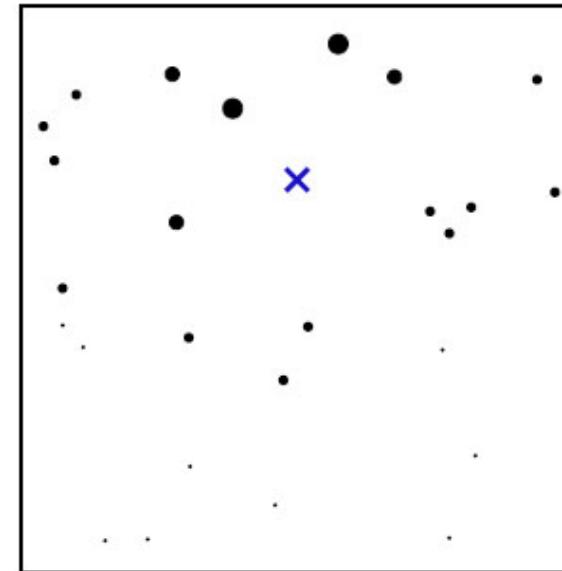
where the weights  $\alpha(x; X_i)$  are obtained from, e.g., a **kernel**.

We use random forests to get good **data-adaptive** weights. Has potential to help mitigate the **curse of dimensionality**.

- ▶ Building many trees with small leaves, then solving the estimating equation in each leaf, and finally **averaging the results** is a bad idea. Quantile and IV regression are badly **biased** in very small samples.
- ▶ Using RF as an “adaptive kernel” protects against this effect.



...



Forests induce a kernel via **averaging tree-based neighborhoods**.

# Generalized Random Forests

- Athey, Tibshirani & Wager establish **asymptotic normality** of parameter estimates, **confidence intervals**
- Recommend orthogonalization
- Software: GRF (on CRAN)

# Local Linear Forests

Friedberg, Athey, Tibshirani, and Wager (2018)

- ▶ 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

# Comparing Regression Forests to Local Linear Forest: Adjusting for Large Leaves/Step Functions

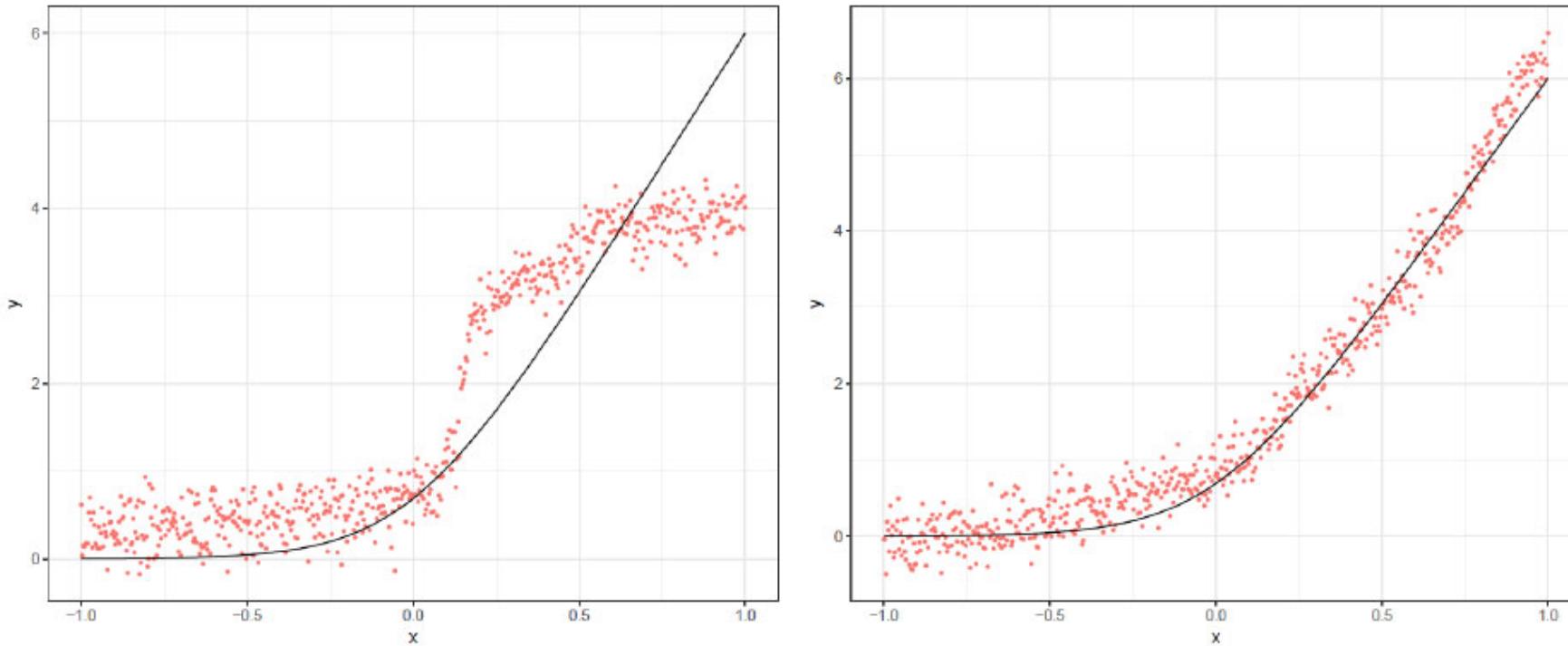
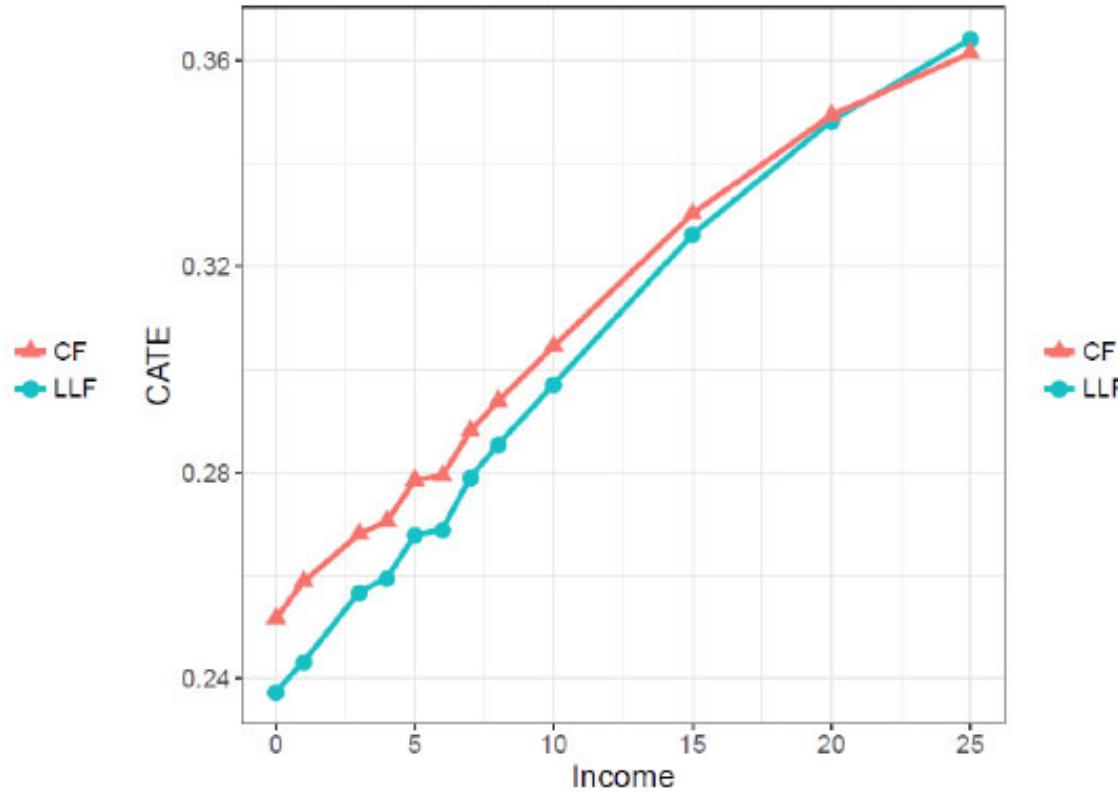
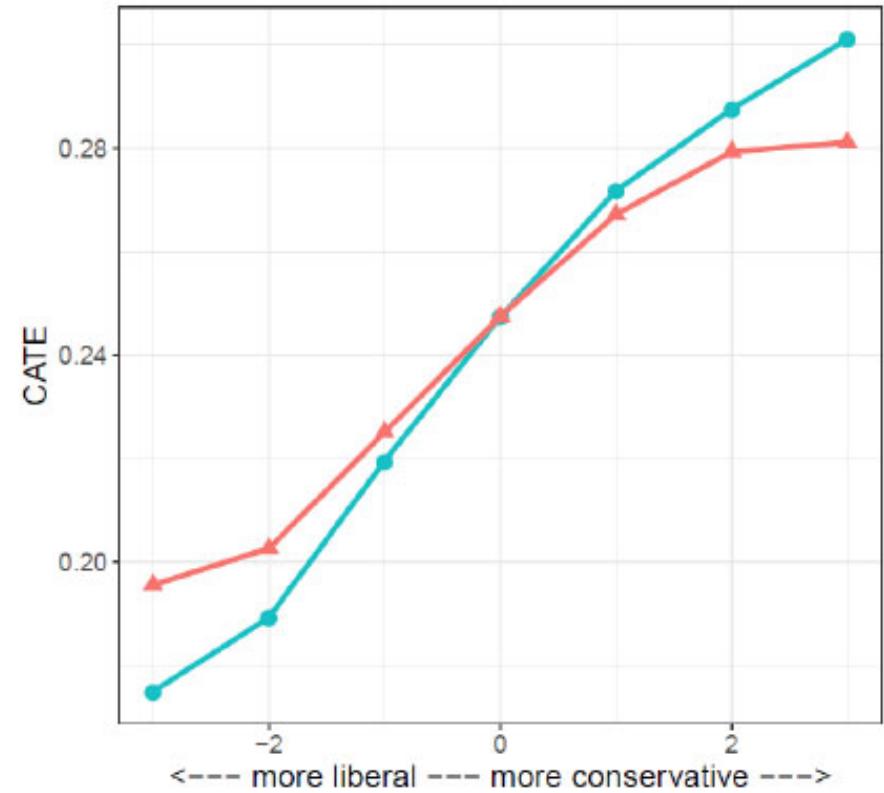


Figure 1: Predictions from random forests (left) and locally linear forests (right) on 600 test points. Training and test data were simulated from equation (1), with dimension  $d = 20$  and errors  $\epsilon \sim N(0, 20)$ . Forests were trained also on  $n = 600$  training points and tuned via cross-validation.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



Randomized Survey Experiment: Are you favor of “assistance to the poor” versus “welfare”

How does treatment effect (CATE) change with political leanings, income?

LLF has better MSE of treatment effect

# Optimal Policy Estimation

---

# Estimating Treatment Assignment Policies

---

Scenario: Analyst has Observational Data

- Historical Logged Data
- Tech firm using contextual bandit or black box algorithms
- Logged data from electronic medical records
- Historical data on worker training programs and outcomes
- Randomized Experiment with Noncompliance

Goal: Estimate Treatment Assignment Policy

- Minimize regret (v. oracle assignment)

Large Literature Spanning Multiple Disciplines

- Offline policy evaluation (e.g. Dudik et al, 2011, others...) versus efficient estimation of best policy from a set
- Two actions vs. multiple actions vs. shifting continuous treatment
- Designs
  - Randomized experiments
  - Unconfoundedness with known (logged) propensity scores
  - Unknown propensity scores
  - Instrumental Variables

Each observed (iid) sample  $i$ , with  $i = 1, \dots, n$ , has:

- ▶ **Features**  $X_i \in \mathcal{X}$ ;
- ▶ **Potential utilities**  $\{Y_i(0), Y_i(1)\} \in \mathbb{R}^2$ ; and a
- ▶ **Realized treatment**  $W_i \in \{0, 1\}$ , such that  $Y_i = Y_i(W_i)$ .
- ▶ **(Instrument**  $Z_i \in \{0, 1\}$  that may be used for identification.)

The conditional average **treatment effect**  $\tau(\cdot)$  is

$$\tau(x) = \mathbb{E} [Y_i(1) - Y_i(0) \mid X_i = x].$$

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

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

measures the expectation of  $Y$  if we **assign** treatment with  $\pi$ .

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

- ▶ **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  $x$ .

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

- ▶ See also counterparts in **computer science** and **statistics** (e.g., Swaminathan and Joachims, 2015; Zhao et al., 2014).

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).

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

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

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

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

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

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

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

## Alternative Approaches to Policy Evaluation/Estimation

Design:  
Unconfoundedness  
Literature focuses on  
this case

$$\max_{\pi \in \Pi} \sum_{i=1}^n (2\pi(X_i) - 1)\hat{\Gamma}_i$$

Different authors have proposed using different scores in the optimization problem

$$\text{CATE: } \hat{\Gamma}_i = \hat{\tau}(X_i)$$

$$\text{IPW: } \hat{\Gamma}_i = \frac{1\{W_i=\pi(X_i)\}}{\hat{e}_{-i}(X_i; W_i)} \cdot Y_i$$

$$\text{Cross-fit AIPW: } \hat{\Gamma}_i = \hat{\tau}_{-i}(X_i) + \frac{1\{W_i=\pi(X_i)\}}{\hat{e}_{-i}(X_i; W_i)} \cdot (Y_i - \hat{\mu}_{-i}(X_i, W_i))$$

- ▶ Athey and Wager (2016): Uses semi-parametric efficiency theory + complexity theory to derive efficient estimation approach for optimal personalized policies
  - ▶ Unconfoundedness; instrumental variables; continuous treatment with personalized small increase/decrease in treatment v. status quo policy
  - ▶ Policies lie in a restricted class, accomodating constraints such as budgets
  - ▶ Challenge in proof: show that results about comparing two policies extend to the case of comparing a continuum of policies within a class of limited complexity
  - ▶ Tighter bounds than prior literature using algorithm based on CATE estimation and efficient scores (semi-parametric efficiency literature)
  - ▶ First  $\sqrt{n}$  convergence results with unknown propensity scores
- ▶ Zhou, Athey and Wager (2018): Extends to multi-arm case; implement with global tree search

# Multi-Arm Generalization (Zhou, Athey and Wager, 2018)

## Step 1

Partition the data into  $K$  folds.

For each fold  $k$ : estimate  $\hat{e}_{a^j}^{-k}(\cdot)$  and  $\hat{\mu}_{a^j}^{-k}(\cdot)$  for every  $j = 1, 2, \dots, d$ .

## Step 2

$$\hat{Q}_{AIPW}(\pi) = \frac{1}{n} \sum_{i=1}^n \langle \pi(X_i), \hat{\Gamma}_i \rangle, \text{ where } \hat{\Gamma}_i = \frac{Y_i - \hat{\mu}_{A_i}^{-k(i)}(X_i)}{\hat{e}_{A_i}^{-k(i)}(X_i)} \cdot A_i + \begin{bmatrix} \hat{\mu}_{a^1}^{-k(i)}(X_i) \\ \hat{\mu}_{a^2}^{-k(i)}(X_i) \\ \vdots \\ \hat{\mu}_{a^d}^{-k(i)}(X_i) \end{bmatrix}$$

## Step 3

Take  $\hat{\pi}_{CAIPWL} = \arg \max_{\pi \in \Pi} \hat{Q}_{AIPW}(\pi)$

# Instrumental Variables Application

Build on Chernozhukov et al (2018) – “CEINR”

Framework for estimating treatment effects with orthogonal moments

**Example:** Voter mobilization  
**Treatment:** Calling voter  
**Randomized Experiment:** Voter list (not all have #s)  
**Outcome:** Did citizen vote  
**Question:** Policy for which people should be called

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

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

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

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

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

$$g(X_i, Z_i) = \frac{1}{\Delta(X_i)} \frac{Z_i - z(X_i)}{z(X_i)(1 - z(X_i))}, \quad z(x) = \mathbb{P}[Z_i \mid X_i = x],$$
$$\Delta(x) = \mathbb{P}[W \mid Z = 1, X = x] - \mathbb{P}[W \mid Z = 0, X = x],$$

to construct a doubly robust estimator.

# General Approach: Choose Policy to Assign Treatment to Units with High Scores

$$\max_{\pi \in \Pi} \sum_{i=1}^n (2\pi(X_i) - 1)\hat{\Gamma}_i$$

Key insights:

- Scores should be orthogonalized/doubly robust
- Use cross-fitting/out-of-bag for nuisance parameters
- Can solve as weighted classification problem (e.g. Beygelzimer et al; Zhou, Athey & Wager propose tree search algorithm)

# Contextual Bandits

---

# Contextual Bandits

See John Langford, Alekh Agarwal, and coauthors for surveys, tutorials, etc...

Online learning of treatment assignment policies

Issues with contexts:

- No context, small finite set of contexts: bandit for each context
- With many contexts, we need to solve a hard estimation problem (as we've been discussing)
- Best performance: state of the art causal inference methods

Most contextual bandit theory

- Assumes outcome model correct (no need for double robust, double robust can add variance)

Proposal in Dimakopoulou, Zhou, Athey and Imbens, AAAI 2019

- Use double robust estimation, shows regret bounds match existing literature

Many open questions from causal inference perspective

- Establish improvement from double robust methods with misspecification

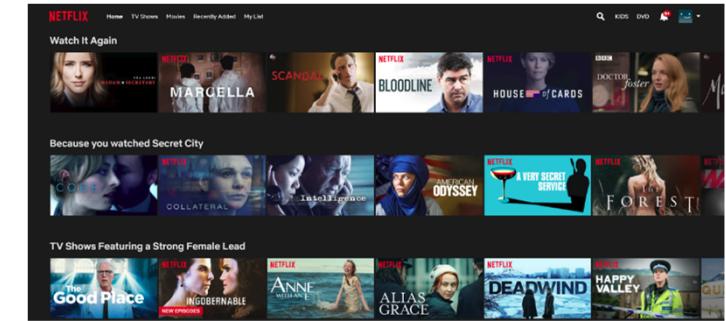
# Contextual bandits

- Arm space  $A$  with  $|A| = K$  arms.
- Context space  $X$  with dimensionality  $d$ .
- Environment generates context and rewards  $(x_t, r_t) \sim D, r_t = (r_t(1), \dots, r_t(K))$ 
  - Agent selects action  $a_t$  and observes reward only for the chosen arm,  $r_t(a_t)$
- Goal: assign each context  $x$  to the arm with the maximum expected reward
  - $\mu_a(x) = E[r_t(a) | x_t = x] = f(x; \theta_a)$  is a function of  $x$ , parameters  $\theta_a$  are unknown...
- Balance exploration (information gained for arms we are uncertain about) with exploitation (improvement in regret from assigning context to the arm viewed best).

# Examples

- **Content recommendation in web services**

- arms: recommendations
- context: user profile and history of interactions
- reward: user engagement and user lifetime value



- **Online education platforms**

- arm: teaching method
- context: characteristics of a student
- reward: student's scores



- **Survey experiments**

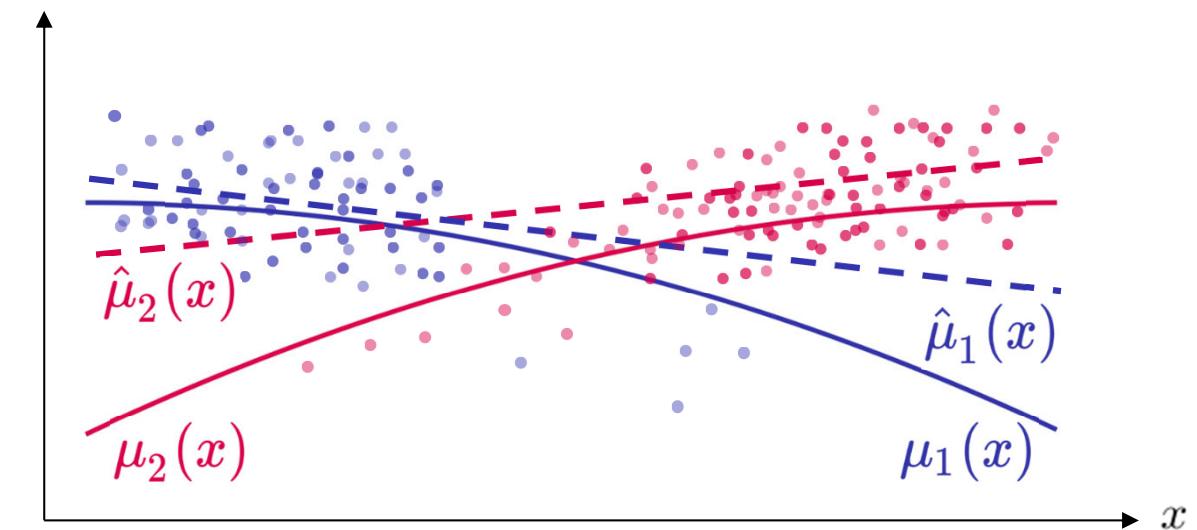
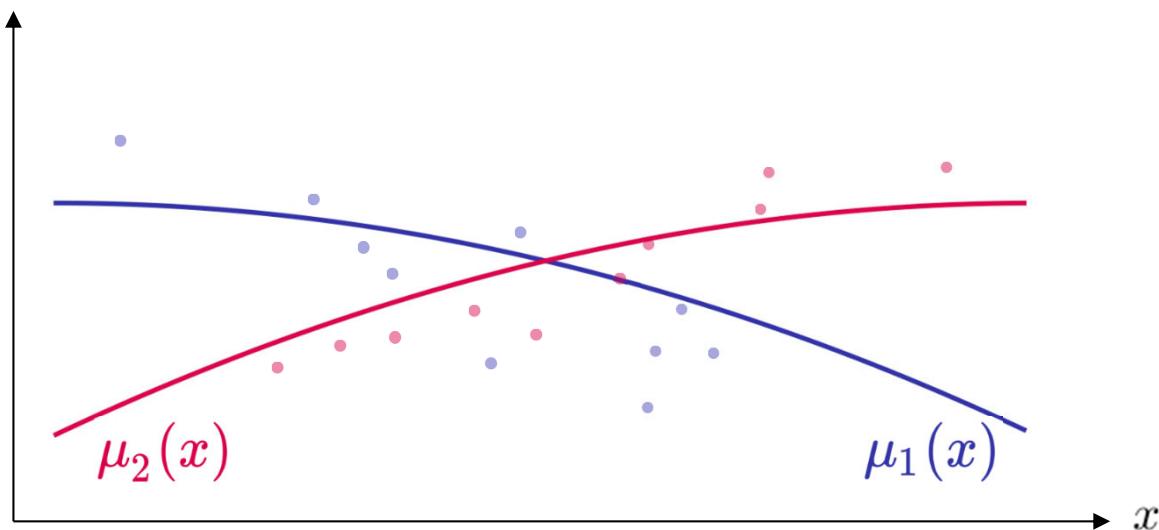
- arm: what information or persuasion to use
- context: respondent's demographics, beliefs, characteristics
- reward: response

# Linear contextual bandits

- Build parametric model for expected reward of each arm given covariates
  - linear bandit:  $E[r_t(a) | x_t = x] = \theta_a^T x$  for all  $a$
- **LinUCB** and **LinTS** have near-optimal regret bounds (requires correct specification).
- **LinUCB**
  - use ridge regression to get an estimate of  $\theta_a$  and a confidence bound of  $\theta_a^T x$
  - assign context  $x$  to arm with highest confidence bound
- **LinTS**
  - start with a Gaussian prior on parameter  $\theta_a$
  - use Bayesian ridge regression to obtain the posterior of  $\theta_a$
  - sample parameters for each arm and assign  $x$  to arm with highest sampled reward

# Estimation is challenging

- Inherent bias to the estimation due to the adaptive assignment of contexts to arms.
  - context assigned to arm with highest reward sample or confidence bound
  - creates systematically unbalanced data
  - complete randomization gives unbiased estimates, but this defeats the purpose



# Estimation is challenging

- Inherent bias to the estimation due to the adaptive assignment of contexts to arms.
  - context assigned to arm with highest reward sample or confidence bound
  - creates systematically unbalanced data
  - complete randomization gives unbiased estimates, but this defeats the purpose
- Aggravating sources of bias in practice
  - **model misspecification**
    - true generative model and functional form used by the learner differ
  - **covariate shift**
    - early adopters of an online course have different features than late adopters

# Balanced contextual bandits

- Dimakopoulou, Zhou, Athey, Imbens (AAAI, 2019)
- Propensity score  $p_t(a_t)$  the probability that context  $x_t$  is assigned to arm  $a_t$
- Balanced LinTS (**BLTS**) and balanced LinUCB (**BLUCB**)
  - Weight each observation  $(x_t, a_t, r_t)$  by  $1/p_t(a_t)$
  - Use the weighted observations in ridge regression.
- For Thompson sampling, **propensity is known**.
  - *Note: Formal Bayesian justification for weighting in Thompson sampling is not clear, similar to justification for using the propensity score in observational studies.*
- For UCB, **propensity is estimated** (e.g. via logistic regression).
  - *Note: The notion of “propensity” in UCB at a given time is contrived (either 0 or 1). Treating the arrival of a context as random, we use the context’s ex ante propensity.*

# Why balancing helps?

- In practice, balancing can help with **covariate shift** and **model mis-specification**.
- **Doubly-robust** nature of inverse propensity score weighted regression
  - accurate value estimates **either** with a well-specified model of rewards **or** with a well-specified model of arm assignment policy.
- Contextual bandits:
  - generally, do not have a well-specified model of rewards
  - even if they do, it **cannot be estimated well** with **small datasets in the beginning**
  - but, they **control arm assignment policy** conditional on observed context
  - hence, **access to accurate propensities** results in **more accurate value estimates**

---

**Algorithm 1** Balanced Linear Thompson Sampling

---

```
1: Input: Regularization parameter  $\lambda > 0$ , propensity score threshold  $\gamma \in (0, 1)$ , constant  $\alpha$  (default is 1)
2: Set  $\hat{\theta}_a \leftarrow \mathbf{null}$ ,  $B_a \leftarrow \mathbf{null}$ ,  $\forall a \in \mathcal{A}$ 
3: Set  $X_a \leftarrow$  empty matrix,  $r_a \leftarrow$  empty vector  $\forall a \in \mathcal{A}$ 
4: for  $t = 1, 2, \dots, T$  do
5:   if  $\exists a \in \mathcal{A}$  s.t.  $\hat{\theta}_a = \mathbf{null}$  or  $B_a = \mathbf{null}$  then
6:     Select  $a \sim \text{Uniform}(\mathcal{A})$ 
7:   else
8:     Draw  $\tilde{\theta}_a$  from  $\mathcal{N}\left(\hat{\theta}_a, \alpha^2 \mathbb{V}(\hat{\theta}_a)\right)$  for all  $a \in \mathcal{A}$ 
9:     Select  $a = \arg \max_{a \in \mathcal{A}} x_t^\top \tilde{\theta}_a$ 
10:  end if
11:  Observe reward  $r_t(a)$ .
12:  Set  $W_a \leftarrow$  empty matrix
13:  for  $\tau = 1, \dots, t$  do
14:    Compute  $p_a(x_\tau)$  and set  $w = \frac{1}{\max(\gamma, p_a(x_\tau))}$ 
15:     $W_a \leftarrow \text{diag}(W_a, w)$ 
16:  end for
17:   $X_a \leftarrow [X_a : x_t^\top]$ 
18:   $B_a \leftarrow X_a^\top W_a X_a + \lambda \mathbf{I}$ 
19:   $r_a \leftarrow [r_a : r_t(a)]$ 
20:   $\hat{\theta}_a \leftarrow B_a^{-1} X_a^\top W_a r_a$ 
21:   $\mathbb{V}(\hat{\theta}_a) \leftarrow B_a^{-1} (r_a - X_a^\top \hat{\theta}_a)^\top W_a (r_a - X_a^\top \hat{\theta}_a)$ 
22: end for
```

---

$$\text{Regret}(T; \text{BLTS}) = \tilde{O}\left(d\sqrt{\frac{KT^{1+\epsilon}}{\epsilon}}\right)$$

---

**Algorithm 2** Balanced Linear UCB

---

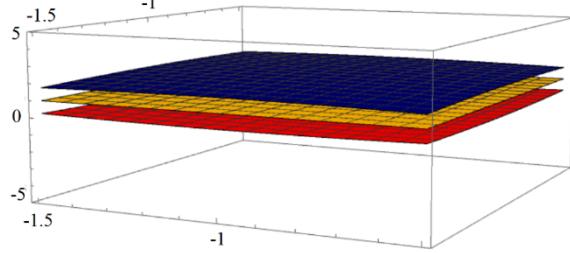
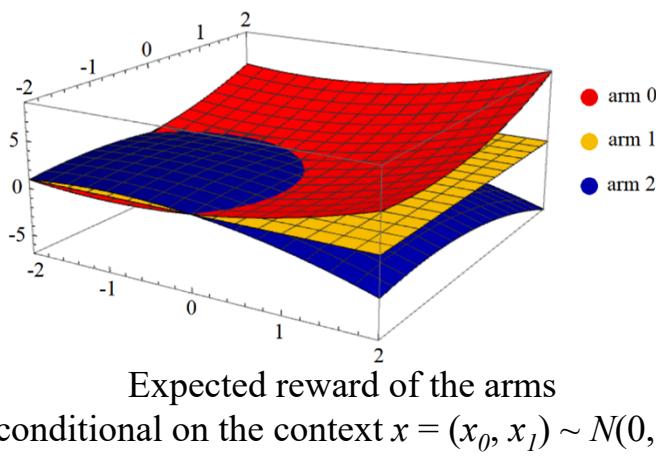
```
1: Input: Regularization parameter  $\lambda > 0$ , propensity score threshold  $\gamma \in (0, 1)$ , constant  $\alpha$ .
2: Set  $\hat{\theta}_a \leftarrow \mathbf{null}$ ,  $B_a \leftarrow \mathbf{null}$ ,  $\forall a \in \mathcal{A}$ 
3: Set  $X_a \leftarrow$  empty matrix,  $r_a \leftarrow$  empty vector  $\forall a \in \mathcal{A}$ 
4: for  $t = 1, 2, \dots, T$  do
5:   if  $\exists a \in \mathcal{A}$  s.t.  $\hat{\theta}_a = \mathbf{null}$  or  $B_a = \mathbf{null}$  then
6:     Select  $a \sim \text{Uniform}(\mathcal{A})$ 
7:   else
8:     Select  $a = \arg \max_{a \in \mathcal{A}} \left( x_t^\top \hat{\theta}_a + \alpha \sqrt{x_t^\top \mathbb{V}(\hat{\theta}_a) x_t} \right)$ 
9:   end if
10:  Observe reward  $r_t(a)$ .
11:  Set  $W_a \leftarrow$  empty matrix
12:  for  $\tau = 1, \dots, t$  do
13:    Estimate  $\hat{p}_a(x_\tau)$  and set  $w = \frac{1}{\max(\gamma, \hat{p}_a(x_\tau))}$ 
14:     $W_a \leftarrow \text{diag}(W_a, w)$ 
15:  end for
16:   $X_a \leftarrow [X_a : x_t^\top]$ 
17:   $B_a \leftarrow X_a^\top W_a X_a + \lambda \mathbf{I}$ 
18:   $r_a \leftarrow [r_a : r_t(a)]$ 
19:   $\hat{\theta}_a \leftarrow B_a^{-1} X_a^\top W_a r_a$ 
20:   $\mathbb{V}(\hat{\theta}_a) \leftarrow B_a^{-1} (r_a - X_a^\top \hat{\theta}_a)^\top W_a ((r_a - X_a^\top \hat{\theta}_a)$ 
21: end for
```

---

$$\text{Regret}(T; \text{BLUCB}) = \tilde{O}\left(\sqrt{TdK}\right)$$

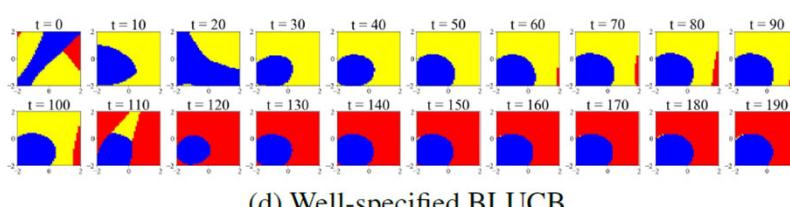
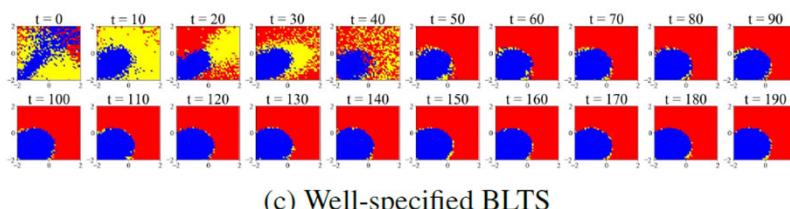
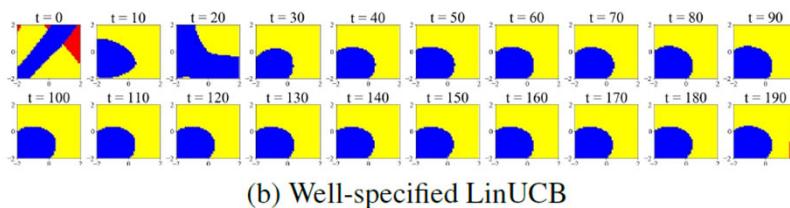
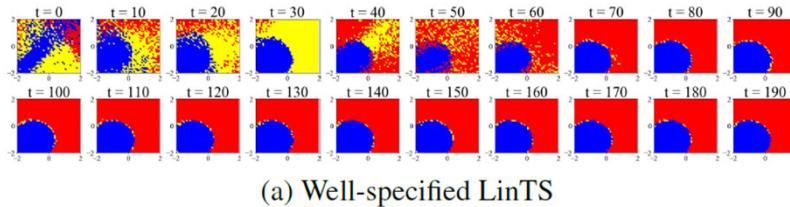
**State of the art regret guarantees, but better performance in practice.**

# A simple synthetic example

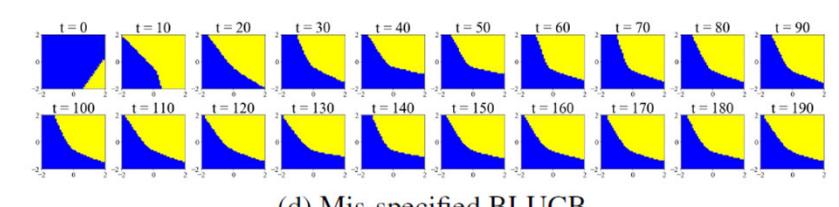
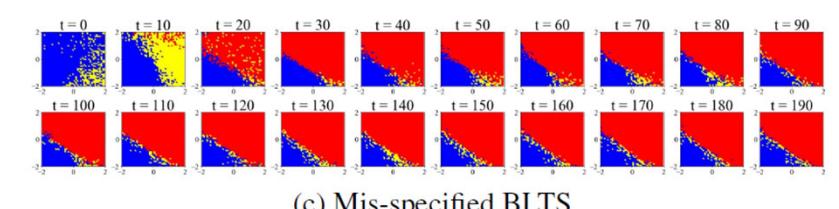
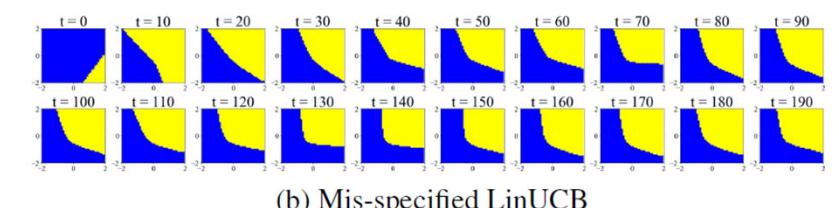
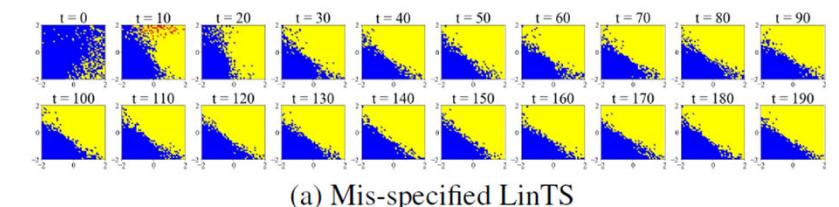


Initial contexts come from a subset of the covariate space around the global optima.

Well-specified reward model  
(include both linear and quadratic terms in context)



Mis-specified reward model  
(include only linear terms in context)



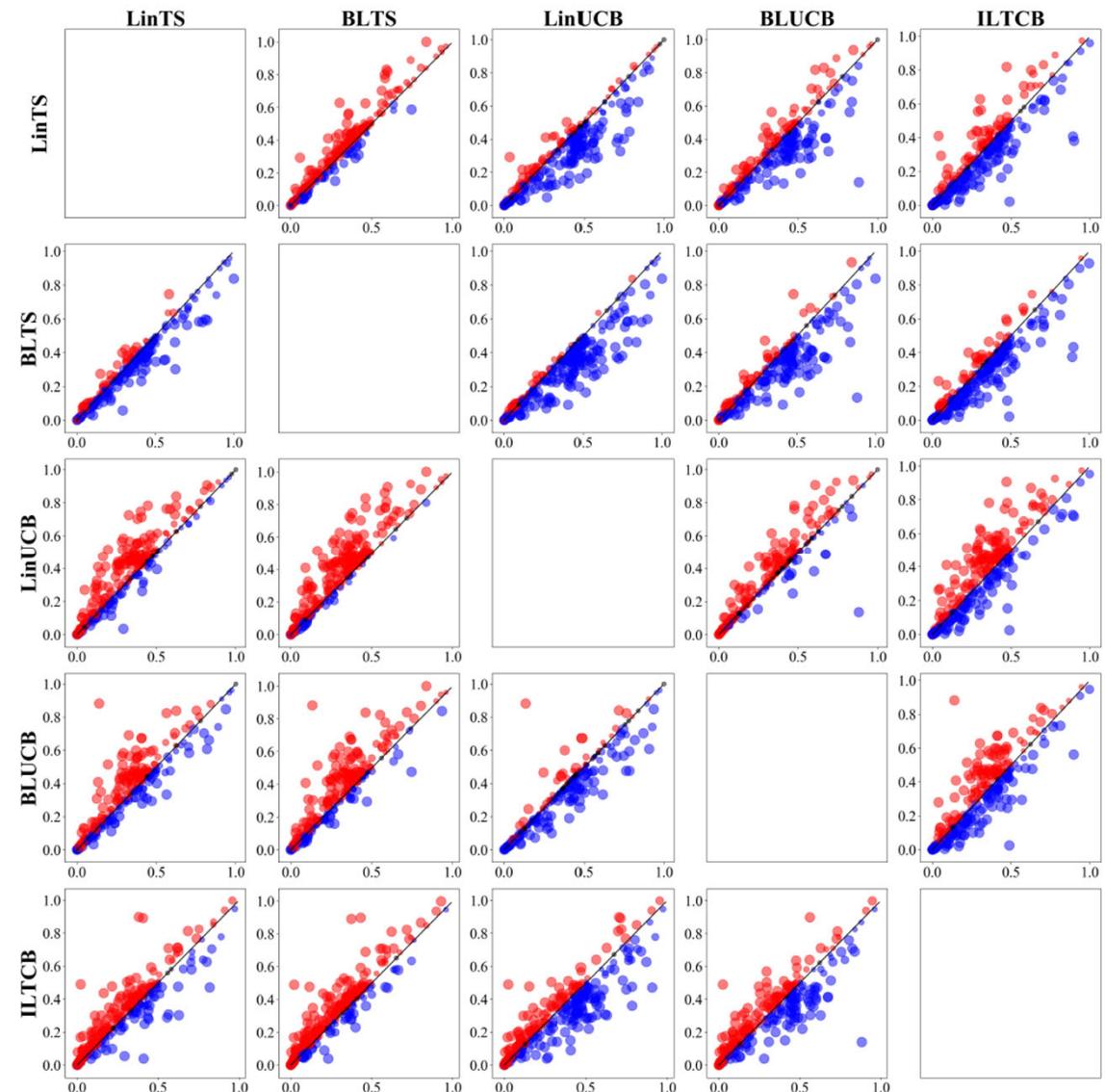
# Experiments on 300 classification datasets

- A classification dataset can be turned into a contextual bandit
  - labels → arms,
  - features → context,
  - accuracy → reward
  - reveal only accuracy of chosen label
- 300 datasets from Open Media Library

Observations	Datasets
$\leq 100$	58
$> 100$ and $\leq 1000$	152
$> 1000$ and $\leq 10000$	57
$> 10000$	33

Classes	Count
2	243
$> 2$ and 10	48
$> 10$	9

Features	Count
$\leq 10$	154
$> 10$ and $\leq 100$	106
$> 100$	40



# Structural Models

---

# Themes for ML + Structural Models

---

## FROM STRUCTURAL LITERATURE

Attention to identification, estimation using “good” exogenous variation in data

- Supermarket application: Tues-Wed comparisons when prices change Tues night; attention to holiday purchases or high seasonality items

Adding sensible structure improves performance

- Required for never-seen counterfactuals
- Increased efficiency for sparse data (e.g. longitudinal data)

Nature of structure

- Learning underlying preferences that generalize to new situations
- Incorporating nature of choice problem
- Many domains have established setups that perform well in data-poor environments

Tune models for counterfactual performance

- Focus on parameters of interest, not fit
- Get a different answer depending on CF of interest

## FROM ML LITERATURE

More efficient computational tools

- E.g. stochastic gradient descent
- E.g. variational inference

Dimension reduction for longitudinal data

- E.g. matrix factorization

Formal model tuning on validation set

- But with different objectives, e.g. counterfactual

# Discrete Choice Models

---

User  $u$ , product  $i$ , time  $t$

$$\mu_{uit} = \nu_i + \beta X_i - \alpha_u p_{it}$$

$$U_{uit} = \mu_{uit} + \epsilon_{uit}$$

If  $\epsilon_{uit}$  i.i.d. Type I extreme value, then

$$\Pr(Y_{uit} = i) = \frac{\exp(\mu_{uit})}{\sum_j \exp(\mu_{ujt})}$$

If sufficient exogenous variation in prices, can identify & estimate distribution of  $\alpha_u$ .

With longitudinal data and sufficient price variation, can estimate  $\alpha_u$  for each user. (Often Bayesian.)

Revealed preference (users' choices) allow us to understand welfare.

- Can solve for a firm's optimal price, optimal coupon
- Understand the impact on firm profits (given cost information) and consumer welfare.

Can evaluate the impact of a new product introduction or the removal of a product from choice set.

Dan McFadden (early 1970s): Counterfactual estimates of extending BART in San Francisco area.

# Combining Discrete Choice Models with Modern Machine Learning....

---

Ruiz, Athey, and Blei (2017), Athey, Blei, Donnelly, and Ruiz (2018), Athey, Blei, Donnelly, Ruiz and Schmidt (2018)

Bring in matrix factorization, and apply to shopping for many items (baskets, restaurants)

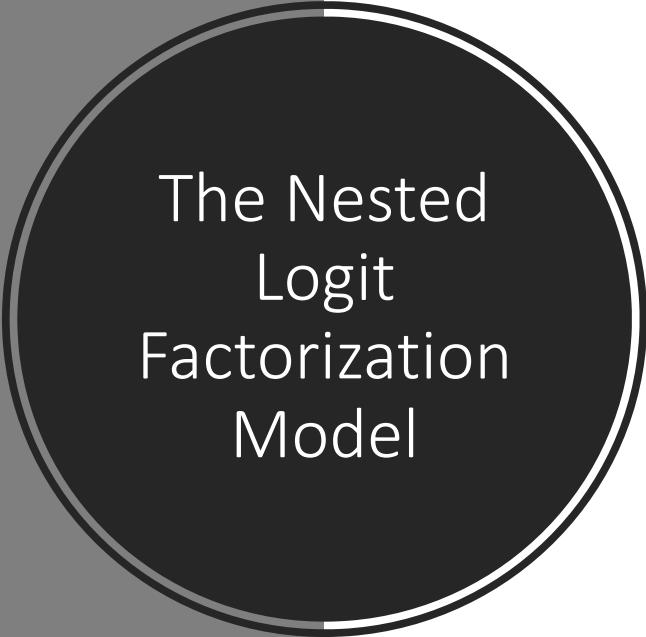
Incorporate choice to not purchase

Two approaches to product interactions

- Use information about product categories, assume products substitutes within categories
- Do not use available information about categories, estimate subs/complements

Can analyze counterfactuals

- Personalized coupons
- Restaurants opening and closing



## The Nested Logit Factorization Model

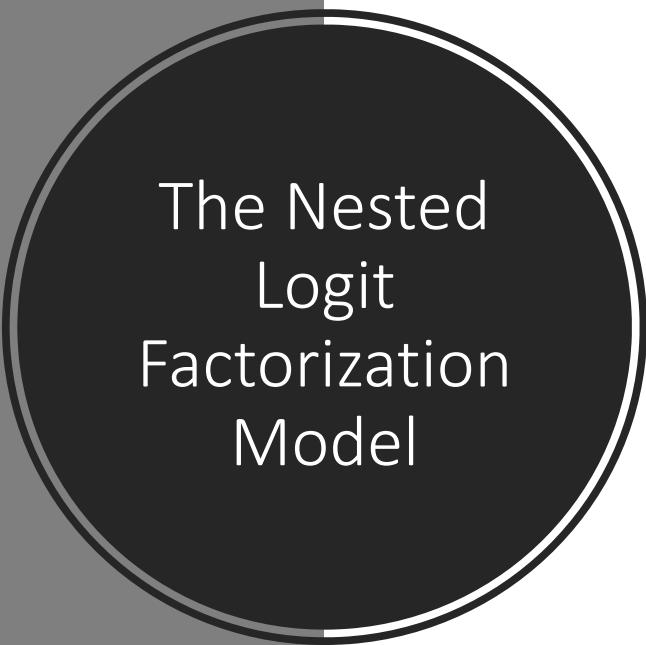
- ▶ Choices in one category are independent of other categories
- ▶ User  $u$  has  $K$ -vector of preferences  $\theta_u$  and a vector of price sensitivity parameters  $\gamma_u$
- ▶ Item  $i$  has two  $K$ -vector of attributes  $\alpha_i$  and  $\beta_i$ .
- ▶ Mean utility for item

$$\mu_{uit} = \theta_u^\top \beta_i - \gamma_u^\top \alpha_i p_{it} \quad (1)$$

- ▶ Utility is  $U_{uit} = \mu_{uit} + \epsilon_{uit}$  where  $\epsilon_{uit}$  drawn from extreme value and indep. cond. on purchasing an item within a category, implying

$$Pr(Y_{ut} = i | \text{purchase in cat for } ut) = \frac{\exp \mu_{uit}}{\sum_{j>0} \exp \mu_{ujt}} \quad (2)$$

- ▶ She chooses the highest utility item in each category or the outside option; the outside option is in its own nest



## The Nested Logit Factorization Model

- ▶ Users  $u$  independently chooses whether or not to make a purchase from each product category  $c$ .
- ▶ Mean utility for not choosing category

$$\mu_{uc_0t} = \theta_{c,u}^\top \beta_{c_0}$$

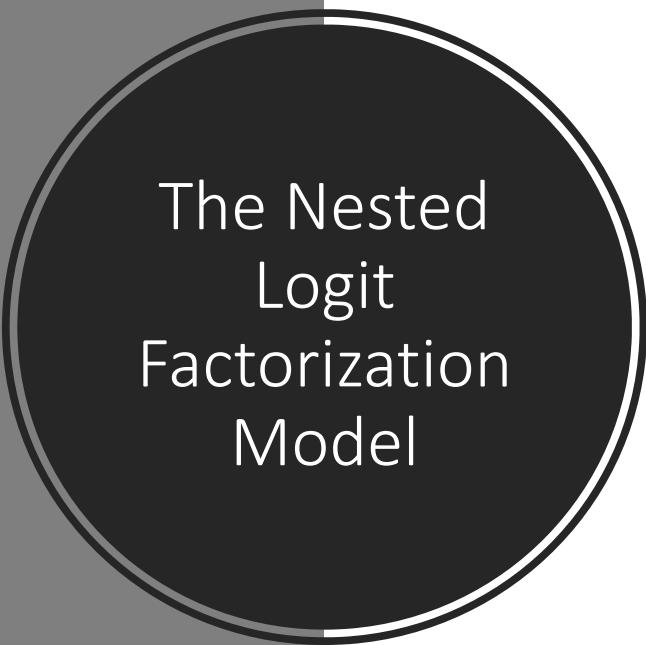
(3)

- ▶ Utility is  $\mu_{uc_0t} + \epsilon_{uc_0t}$
- ▶ Utility for choosing category

$$U_{uc_1t} = \theta_{c,u}^\top \beta_{c_1} - \alpha_u \delta_{c_1} IV_c + \epsilon_{uc_1t}$$

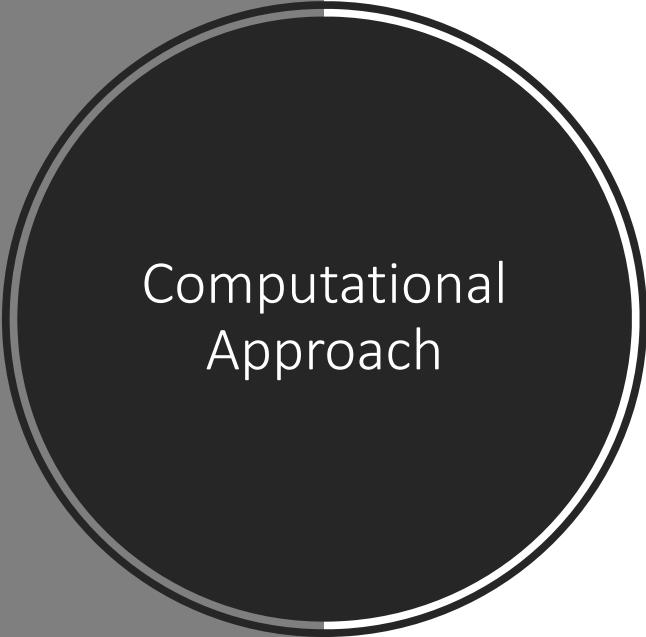
(4)

- ▶ Where  $IV_c$  is the inclusive value of the items in the category is given by  $IV_c = \log \sum_{i \in J_c} \exp \mu_{uit}$ , which is the expectation of the max of the  $U_{uit}$  prior to learning the  $\epsilon_{uit}$  for each item.



## The Nested Logit Factorization Model

- Counterfactual inference in nested logit models uses structure
  - Model specifies how user substitutes if choice set changes, e.g. product out of stock
    - Conditional on purchasing a single item in a category, choice probabilities redistributed in proportion to probabilities of other items
  - Model makes counterfactual predictions about what happens when prices change
    - Given price sensitivity for a given product, model makes sensible predictions about how purchase probabilities for other products change when the price of the given product changes



## Computational Approach

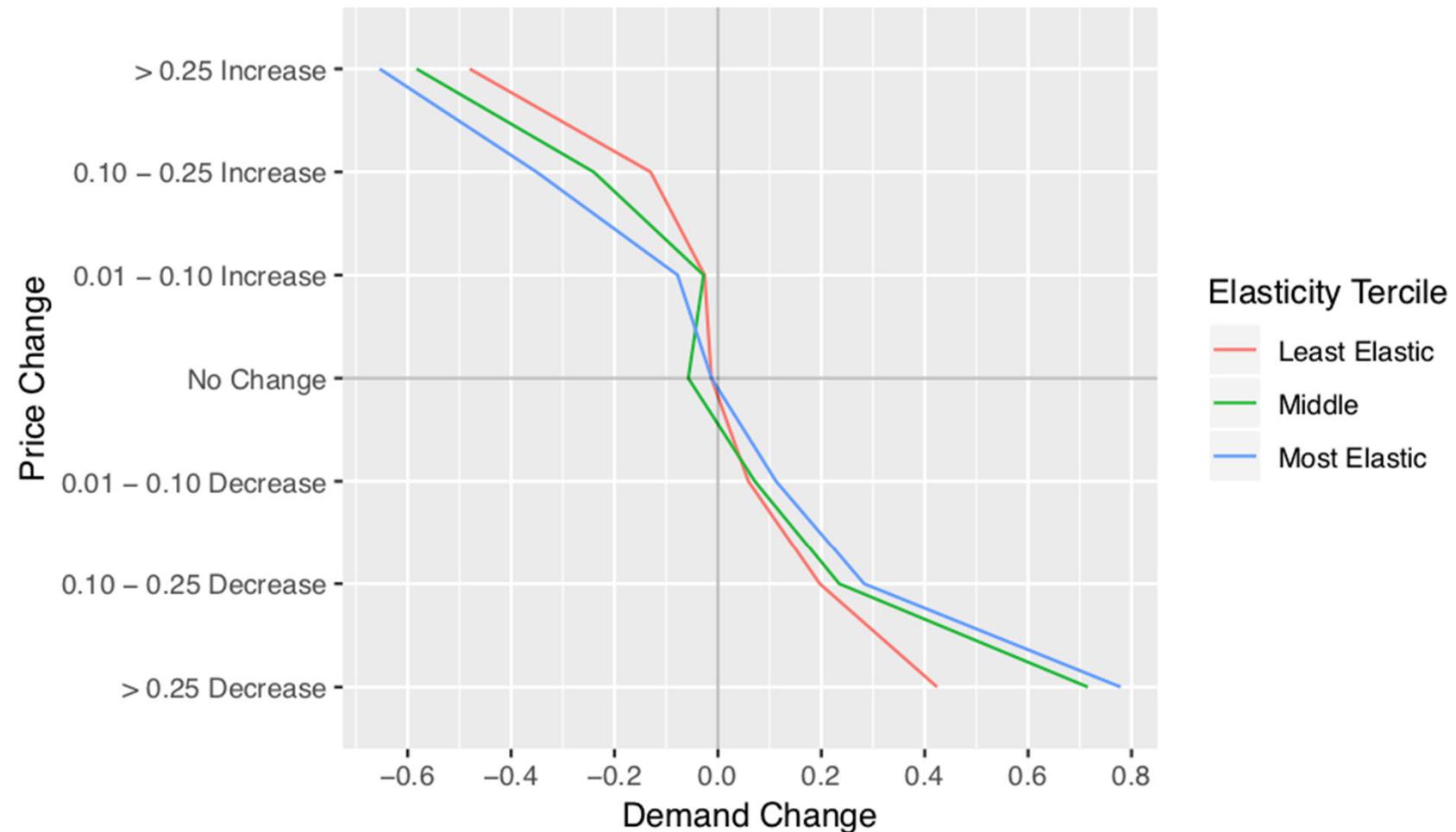
- ▶ MCMC-based Bayesian methods: Common in marketing for estimating models with heterogeneity, but computationally infeasible as data size and number of parameters grows
- ▶ This choice of functional form allows for fast and efficient estimation using variational Bayesian inference
- ▶ Variational Bayes:
  - ▶ Choose parameterized family of distributions  $q(\cdot|\eta)$  to approximate the posterior
  - ▶ Find  $\eta$  that minimizes KL-divergence to the true posterior
  - ▶ With appropriate choice of priors and  $q$ , this optimization can be done using simple coordinate ascent
  - ▶ Accuracy similar to MCMC, but 1000s of times more quickly
- ▶ Introducing price effects and time-varying price slows things down substantially (hours rather than minutes; but still feasible unlike MCMC)
- ▶ Introducing substitutability within categories requires additional computational tricks

Table: Mean Log Likelihood for Cross Price Weeks

Model	Popular UPCs		Less Common UPCs	
	Aggregate	Individual	Aggregate	Individual
Nested Factorization	-2.4527 (0.0262)	-0.0925 (0.0008)	-1.2017 (0.0084)	-0.0149 (0.0001)
Nested Logit with HPF Controls	-2.4844 (0.0277)	-0.1041 (0.0009)	-1.2177 (0.0076)	-0.0164 (0.0001)
Multinomial Logit with HPF Controls	-2.4836 (0.0276)	-0.1041 (0.0009)	-1.2165 (0.0076)	-0.0164 (0.0001)
Hierarchical Poisson Factorization (HPF)	-2.4919 (0.0294)	-0.1008 (0.0009)	-1.2343 (0.0078)	-0.0166 (0.0001)
Multinomial Logit with Demographic Controls	-2.4966 (0.0279)	-0.1162 (0.0010)	-1.2182 (0.0077)	-0.0170 (0.0001)
Nested Logit with Demographic Controls	-2.5066 (0.0292)	-0.1161 (0.0010)	-1.2194 (0.0078)	-0.0170 (0.0001)
Mixed Logit with Random Price and Random Intercepts	-2.5167 (0.0242)	-0.1166 (0.0009)	-1.3176 (0.0064)	-0.0182 (0.0001)
Mixed Logit with Random Price Effects and HPF Controls	-2.4850 (0.0253)	-0.1011 (0.0009)	-1.3168 (0.0066)	-0.0190 (0.0001)
Mixed Logit with Random Price Effects and Demographics	-2.5312 (0.0267)	-0.1136 (0.0009)	-1.3989 (0.0069)	-0.0235 (0.0001)

Goodness of Fit (Tuned for CF)  
 Weeks where another product in category changed prices

## Average Changes in Demand in Test Set

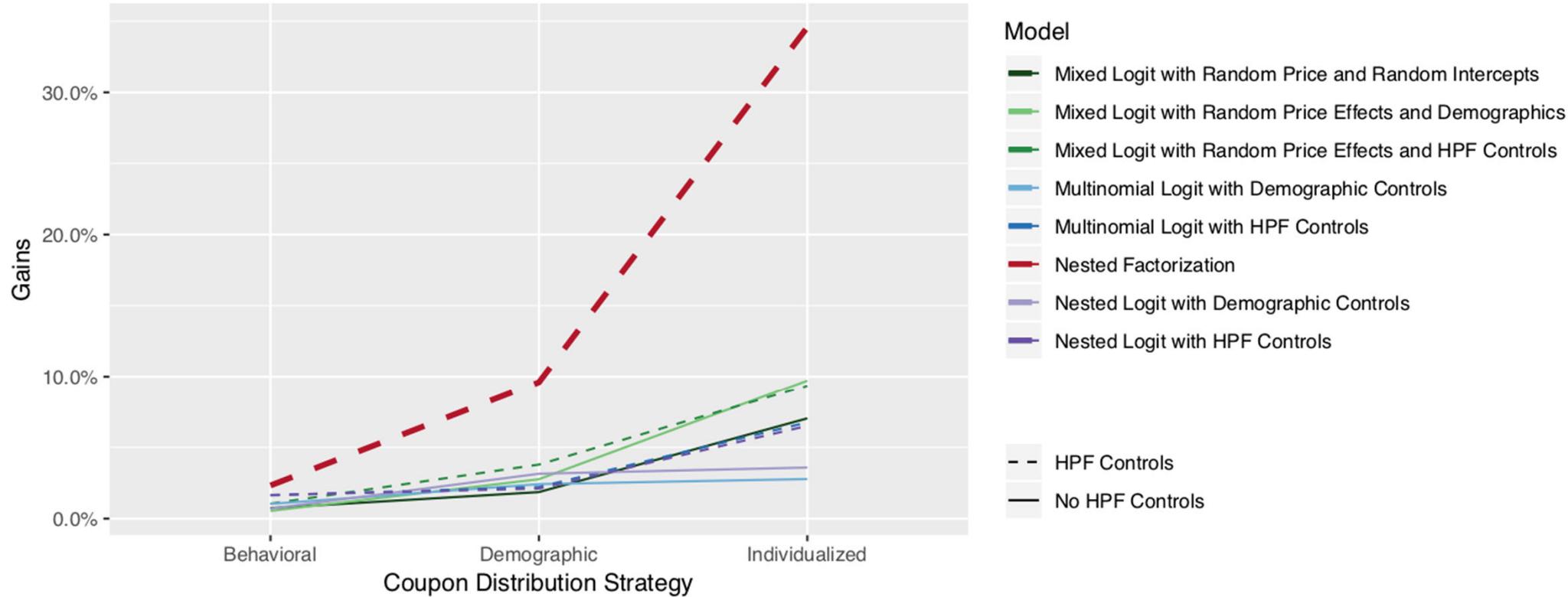


## Validation of Structural Parameter Estimates

Compare Tuesday-Wednesday change in price to Tuesday-Wednesday change in demand, in test set

Break out results by how price-sensitive (elastic) we have estimated consumers to be

### Profit Gains from Personalized Pricing Relative to Uniform Coupon Distribution



## Personalized Pricing Matrix Factorization Approach Allows Accurate Personalization

How much profit can be made by giving a 30% off coupon for a single product to a targeted selection of 30% of the shoppers in the store? Compare uniform randomization, demographic, or individual targeting policies based on structural estimates

# Conclusions

---

Causal inference is key to using machine learning and artificial intelligence to make decisions

- This is a tautological statement: but at the same time, not fully appreciated

Artificial intelligence agents will improve if they are good statisticians

AI based on causal modeling has desirable properties (stability, fairness, robustness, transferability, ....)

There is an enormous literature on theory and applications of causal inference in many settings and with many approaches

The conceptual framework is well worked out for both static and dynamic settings

Structural models enable counterfactuals for never-seen worlds

Machine learning algorithms can greatly improve practical performance, scalability

Challenges: data sufficiency, finding sufficient/useful variation in historical data

- Recent advances in computational methods in ML don't help with this
- But tech firms conducting lots of experiments, running bandits, and interacting with humans at large scale can greatly expand ability to learn about causal effects!

# References

---

# Selected References: Traditional “Program Evaluation” or Treatment Effect Estimation

---

## BOOKS

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

- Summarizes literature from stats/econometrics/biostatistics perspective in pre-machine learning era

Angrist and Pischke, 2008, Mostly Harmless Econometrics

- Informal introduction to causal inference

Cunningham, Causal Inference: The Mixtape

- Applied economics perspective; recent and accessible, and available free online  
[http://scunning.com/cunningham\\_mixtape.pdf](http://scunning.com/cunningham_mixtape.pdf)

Pearl and MacKenzie, Book of Why?

- Recent and accessible

Stephen L Morgan and Christopher Winship. Counterfactuals and causal inference. Cambridge University Press, 2014

## SURVEY AND NONTECHNICAL PAPERS

Guido Imbens and Jeffrey Wooldridge. Recent developments in the econometrics of program evaluation. Journal of Economic Literature, 47(1):5–86, 2009.

Susan Athey and Guido Imbens. “The state of applied econometrics causality and policy evaluation.” Journal of Economic Perspectives, 2017.

# Selected References: Randomization Approach to Causal Inference

---

Neyman [1923/1990] is a classic paper, reprinted in *Statistical Science*.

Fisher [1935] is another classic reference.

General statistics texts: Wu and Hamada [2011], Cook and DeMets [2007], Cox and Reid [2000], Hinkelman et al. [1996]

Athey and Imbens [2016a] is a survey focused on an economics audience.

Bruhn and McKenzie [2009], Morgan and Rubin [2015, 2012] discuss re-randomization.

Middleton and Aronow [2015], Murray [1998] discuss clustered randomized experiments.

The relation to regression is discussed in Abadie et al. [2014], Lin [2013], Freedman [2008], Samii and Aronow [2012].

Imbens and Menzel [2018] develop a version of the bootstrap focused on causal effects.

# Selected References: ATE Under Unconfoundedness

---

Rosenbaum and Rubin [1983]: Potential outcomes, theory of propensity score weighting

Imbens [2004] presents a survey.

Matching estimators: Abadie and Imbens [2006, 2008], Rubin and Thomas [1996].

Hahn [1998] derives the efficiency bound and proposes an efficient estimator.

Robins and Rotnitzky [1995], Robins et al. [1995]: Doubly robust methods.

Hirano et al. [2003]: Weighting estimators with the estimated propensity score.

Crump et al. [2009] discuss trimming to improve balance.

Yang et al. [2016], Imbens [2000], Hirano and Imbens [2004] discuss settings with treatments taking on more than two values

Hotz et al. [2005] discuss the role of external validity.

Applications to the Lalonde data: LaLonde [1986], Dehejia and Wahba [1999], Heckman and Hotz [1989].

Athey and Imbens [2016, AER], Athey, Imbens, Pham, Wager [2017], Athey and Imbens [2018, JEP] discuss robustness and supplementary analysis

# Selected References: Instrumental Variables

---

Imbens and Angrist [1994], Angrist et al. [1996]: LATE

Imbens [2014] presents a general discussion for statisticians

Classic applications: Angrist [1990], Angrist and Krueger [1991].

Staiger and Stock [1997], Moreira [2003] discuss inference with weak instruments.

Chamberlain and Imbens [2004] discuss settings with many weak instruments

# Selected References: Regression Discontinuity Designs

---

Thistlewaite and Campbell [1960]: original reference.

Imbens and Lemieux [2008], Lee and Lemieux [2010], Van Der Klaauw [2008], Skovron and Titiunik [2015], Choi and Lee [2016]: theory

Hahn et al. [2001]: fuzzy regression discontinuity

Imbens and Kalyanaraman [2012], Calonico et al. [2014]: optimal bandwidth choices.

Gelman and Imbens [2018] discuss the pitfalls of using higher order polynomials.

Bertanha and Imbens [2014], Battistin and Rettore [2008], Dong and Lewbel [2015], Angrist and Rokkanen [2015], Angrist [2004] discuss external validity of regression discontinuity designs.

Applications: Angrist and Lavy [1999], Black [1999], Lee et al. [2010], Van Der Klaauw [2002]

Regression kink designs: Card et al. [2015].

Recent work focuses on settings where instead of choosing a bandwidth directly optimal weights are calculated: Kolesar and Rothe [2018], Imbens and Wager [2017], Armstrong and Kolesar [2018].

# Selected References: Differences-in-Differences, Synthetic Controls

---

Angrist and Krueger [2000]: General discussion

Applications: Ashenfelter and Card [1985], Eissa and Liebman [1996], Meyer et al. [1995], Card [1990], Card and Krueger [1994]

Nonlinear version: Athey and Imbens [2006]

Synthetic control methods: Abadie and L'Hour [2016], Abadie et al. [2010, 2015], Abadie and Gardeazabal [2003], Doudchenko and Imbens [2016], Xu [2015], Gobillon and Magnac [2013], Ben-Michael et al. [2018], Athey and Imbens [2018].

Links between the matrix completion literature and the causal panel data literature are given in Athey, Bayati, Doudchenko, Imbens, Khosravi [2017].

# Selected References: Econometrics and ML

---

## Prediction v. Estimation

- Mullainathan, Sendhil, and Jann Spiess. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31.2 (2017): 87-106.

## Prediction policy

- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. "Prediction policy problems." *The American Economic Review* 105, no. 5 (2015): 491-495.

## Prediction v. Causal Inference

- S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355 (6324):483-485, 2017.
- A. Belloni, V. Chernozhukov, C. Hansen: "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28 (2), Spring 2014, 29-50.  
<https://www.aeaweb.org/articles?id=10.1257/jep.28.2.29>

# Selected References: Treatment Effect Estimation and Machine Learning

---

Survey: Athey, "The Impact of Machine Learning on Economics," NBER Volume, 2018

ATE

- McCaffrey et al. [2004] (propensity score)
- Athey, Imbens and Wager [2018], Belloni et al. [2013], Chernozhukov et al. [2016], Chernozhukov et al [2017], Chernozhukov et al [2018], van der Laan and Rubin [2006] focus on doubly robust methods.

Dynamic Treatment Regimes

- Chakraborty and Murphy (2014): Survey

Heterogeneous Treatment Effects

- Imai and Ratkovic, 2013: LASSO
- Zeiles et al [2008], Athey and Imbens [2016]: Subgroups
- Friedberg et al. [2018]: Local linear forests
- Wager and Athey [2018], Athey, Tibshirani, and Wager [2018]: Causal forests and generalized random forests
- Kunzel et al [2017]: Meta-learners
- Hartford et al [2017]: Deep IV
- Chernozhukov et al [2018]: Testing top CATEs

Instrumental Variables

- Chernozhukov et al (multiple papers)
- Goldman and Rao [2017], Peysakhovich & Eckles [2018]: experiments as instruments
- Athey, Tibshirani, and Wager [2018]; Hartford et al [2017]: heterogeneous treatment effects

Optimal Policy Estimation

- Dudik et al [2011], Li et al [2012], Dudik et al [2014], Li et al [2014]
- Thomas and Brunskill [2016], Kallus [2017]
- Kitagawa and Tetenov [2016], Swaminathan and Joachim [2015], Zhao et al [2014]-IPW
- Athey and Wager [2017], Zhou, Athey, and Wager [2018], CAIPW (doubly robust, efficiency with unknown propensity)

Contextual Bandits

- Li et al [2010], Chapelle and Li [2011], Li et al [2017], Bastani and Bayati [2015]
- See Agarwal et al [2016] for a survey; John Langford for many tutorials and articles
- Bakshy et al-Bayesian optimization perspective
- Dimakopoulou, Zhou, Athey and Imbens [2018]

# Selected References: Social Networks and Interference

---

Aronow [2018], Athey, Eckles and Imbens [2018]:  
Randomization Inference Approach

Kizilcec, R.F., Bakshy, E., Eckles, D., & Burke, M. [2018]:  
Social Influence

Eckles, D., Karrer, B., & Ugander, J. [2017]: Reducing Bias  
from interference

Eckles, D., Kizilcec, R. F. & Bakshy, E. [2016]

# Selected References: Structural Estimation

---

## DISCRETE CHOICE/DEMAND SYSTEMS/ SUPPLY BEHAVIOR/WELFARE ESTIMATION

McFadden [1972]

Deaton, A., and J. Muellbauer [1980]

Berry [1994]

Berry, Levinsohn, and Pakes [1995, 2004]

Nevo [2000, 2001]

Keane et al. [2013]

Elrod [1988]; Elrod and Keane, [1995];  
Chintagunta [1994] (latent variable models)

## OLIGOPOLY/EQUILIBRIUM APPLICATIONS

Porter and Zona [1999]

Nevo [2000]

Busse and Rysman [2005]

Dafny [2009]

Marshall and Marx [2012]

# Selected References: Structural Estimation and Market Design

---

## TRADITIONAL AUCTIONS

Laffont et al. (1995), Perrigne and Vuong: Identification and estimation of first price auctions

Athey, Levin and Seira (2011), Athey, Coey and Levin (2013): counterfactual analysis of auction design and small business set-asides in timber auctions

Hendricks, Pischke, and Porter: Identification and estimation with Common Values

Athey and Haile [2002]: Identification

Athey and Haile [2007]: Survey

Haile and Tamer [2003]: Bounds on counterfactuals with partial identification

## MARKET DESIGN

Sponsored search auctions

- Varian (2009)
- Athey and Nekipelov (2012)
- Bottou (2012)

Matching markets

- Agarwal (2015): Medical match
- Agarwal and Somaini (2018): School choice
- Agarwal, Ashlagi, Rees, Somaini, Waldinger (2018): Kidney allocation

# Selected References: Structural Estimation in Dynamic Settings

---

## SINGLE PLAYER DYNAMIC OPTIMIZATION

Ackerberg, Daniel, "Advertising, Learning, and Consumer Choice in Experience Good Markets:A Structural Empirical Examination," International Economic Review, 44: 1007-1040, (2003).

Aguirregabiria, Victor, "The Dynamics of Markups and Inventories in Retailing Firms," Review of Economic Studies 66(2): 275-308, (1999).

Benkard, C. Lanier, "Learning and Forgetting: The Dynamics of Aircraft Production," American Economic Review, 90(4): 1034-1054, (2000).

Hitsch, Gunter, "An Empirical Model of Optimal Dynamic Product Launch and Exit Under Demand Uncertainty," Marketing Science, 25(1): 25-50, (2006).

Hotz, Joseph and Robert Miller, "Conditional Choice Probabilities and the Estimation of Dynamic Models," Review of Economic Studies 60(3): 497-530, (1993).

Hotz, Joseph, Robert Miller, Seth Sanders and Jeffrey Smith "A Simulation Estimator for Dynamic Models of Discrete Choice," Review of Economic Studies, 61: 256-289, (1994).

Pakes, Ariel, "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," Econometrica, 54(4): 755-784, (1986).

Rust, John, "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," Econometrica, 55(5): 999-1033, (1987).

# Selected References: Structural Estimation in Dynamic Settings

---

## MULTI-PLAYER GAMES

- Ackerberg, Daniel, Steven Berry, Lanier Benkard, and Ariel Pakes, "Econometric Tools for Analyzing Market Outcomes," in *Handbook of Econometrics*. J.J. Heckman and E.E. Leamer (ed.), Elsevier. Edition 1, volume 6, (2007).
- Bajari, Patrick, Lanier Benkard, and Jonathan Levin, "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 75(5): 1331-1370, (2007). 17
- Benkard, Lanier, "Dynamic Analysis of the Market for Wide-Bodied Commercial Aircraft," *Review of Economic Studies*, 71(3): 581-611, (2004).
- Ericson, Richard and Ariel Pakes, "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies*, 62(1): 53-82, (1995).
- Gowrisankaran, Guatam and Robert Town, "Dynamic Equilibrium in The Hospital Industry," *Journal of Economics and Management Strategy*, 6(1): 45-74, (1997).
- Markovich, Sarit, "Snowball: The Evolution of Dynamic Oligopolies with Network Externalities," *Journal of Economic Dynamics and Control*, 33(3): 909-938, (2007).
- Pakes, Ariel and Paul McGuire, "Computing Markov-Perfect Nash Equilibria: Numerical Implications of a Dynamic Differentiated Product Model," *Rand Journal of Economic*, 25(4): 555-589, (1994).
- Pakes, Ariel and Richard Ericson, "Empirical Implications of Alternative Models of Firm Dynamics," *Journal of Economic Theory*, 79(1): 1-45, (1998).
- Pakes, Ariel, Michael Ostrovsky, and Steven Berry, "Simple Estimators for the Parameters of Dynamic Discrete Games (with Entry/Exit Examples)," *Rand Journal of Economics*, 38(2): 373- 399, (2007).
- Pakes Ariel and U. Doraszelski, "A Framework for Applied Dynamic Analysis in IO". In: Armstrong M, Porter R, *The Handbook of Industrial Organization*. Vol. 3. New York: Elsevier; pp. Chapter 33 2183-2162 (2007). Ryan, Stephen, "The Costs of Environmental Regulation in a Concentrated Industry," *Econometrica*, 80(3): 1019-1061, (2012).

# Selected References: Structural Estimation and Machine Learning

---

## CONSUMER CHOICE

Counterfactual Inference for Consumer Choice Across Many Product Categories (Susan Athey, David Blei, Rob Donnelly, Francisco Ruiz, in progress)

SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements (Francisco Ruiz, Susan Athey, David Blei, 2017)

Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data (Susan Athey, David Blei, Rob Donnelly, Francisco Ruiz, Tobias Schmidt, AEA Papers and Proceedings, 2018)

Wan, Mengting, et al. "Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs." *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.