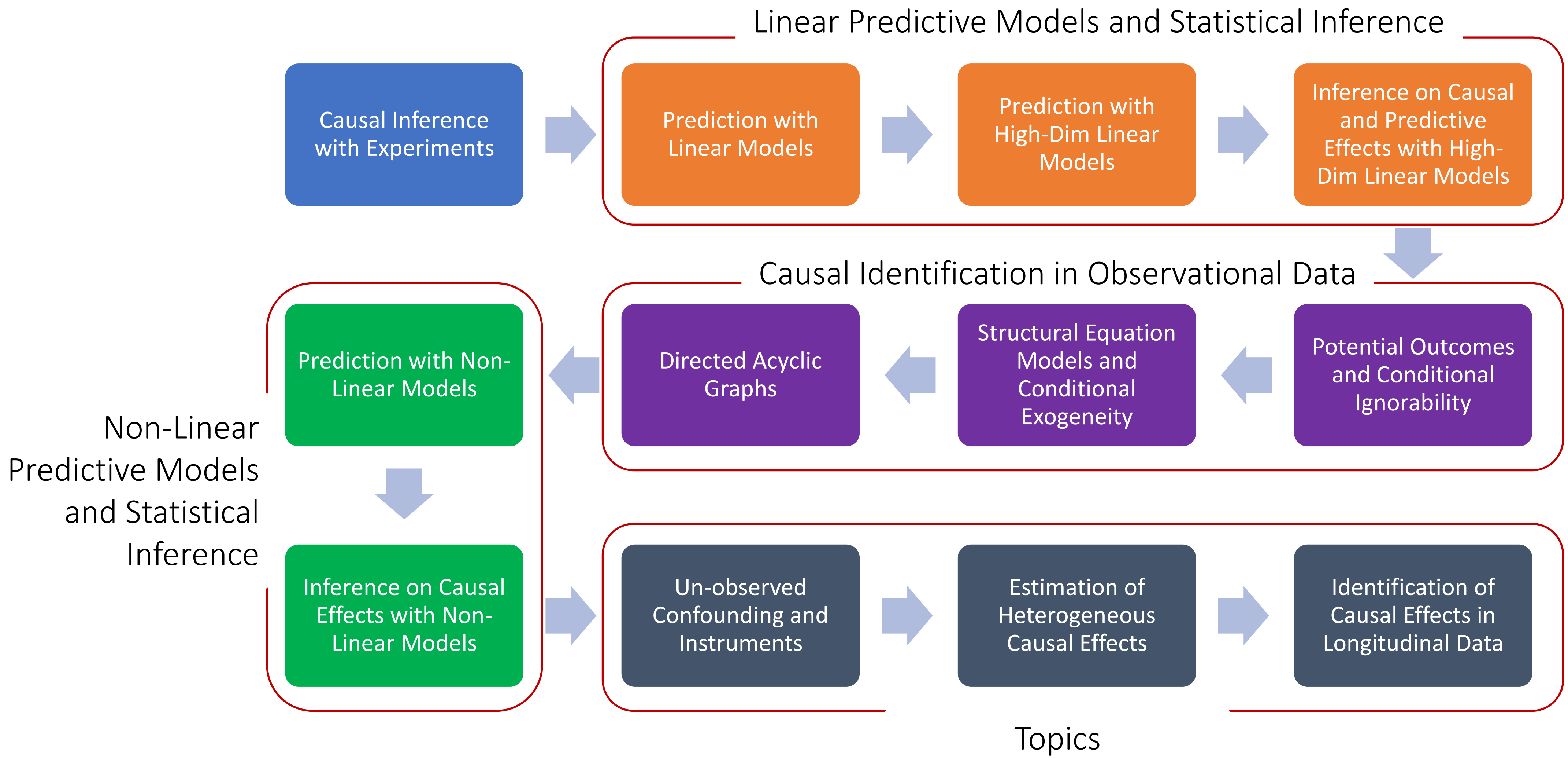
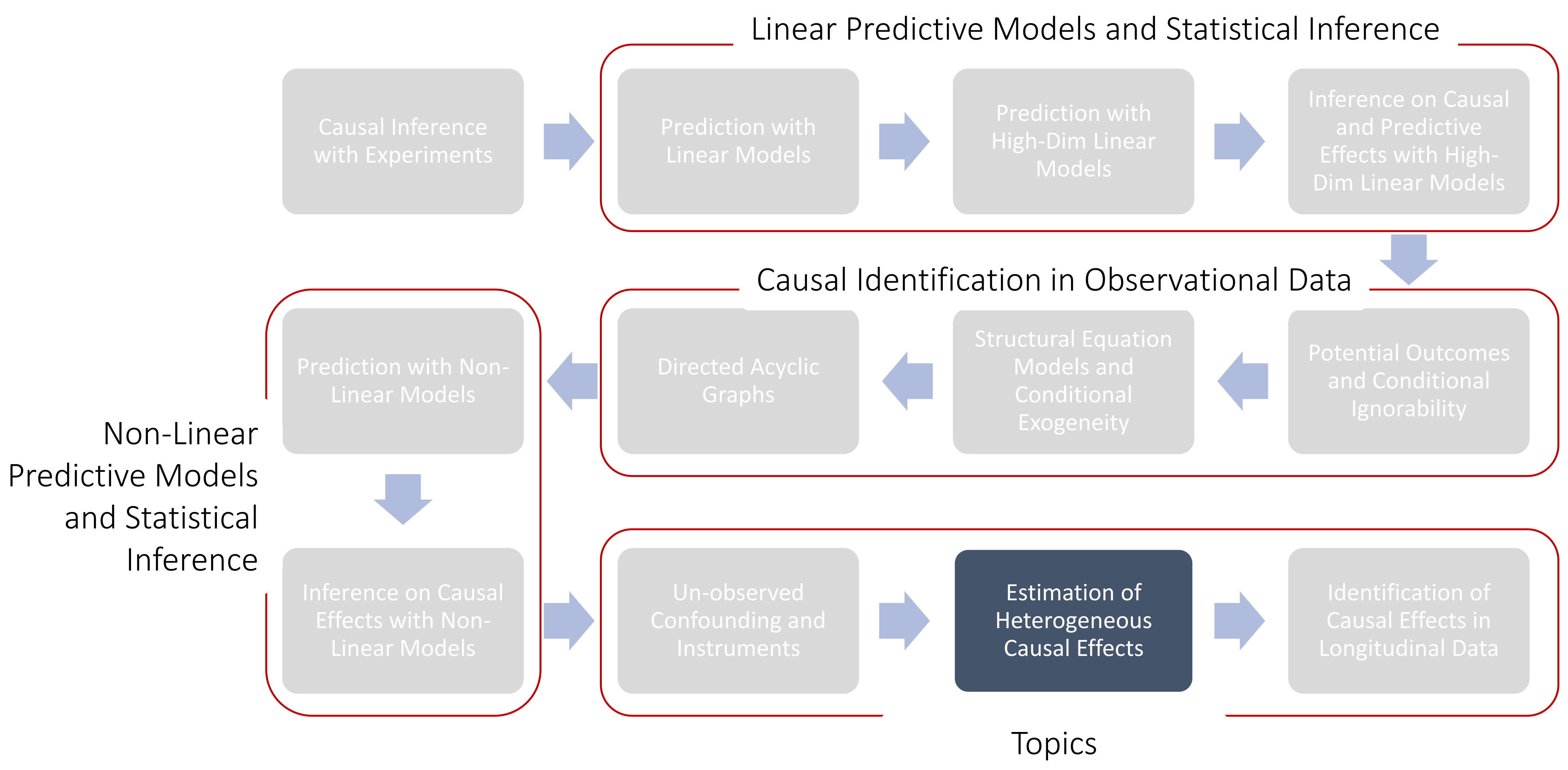
MS&E 228: Heterogeneous Treatment Effects

Vasilis Syrgkanis

MS&E, Stanford

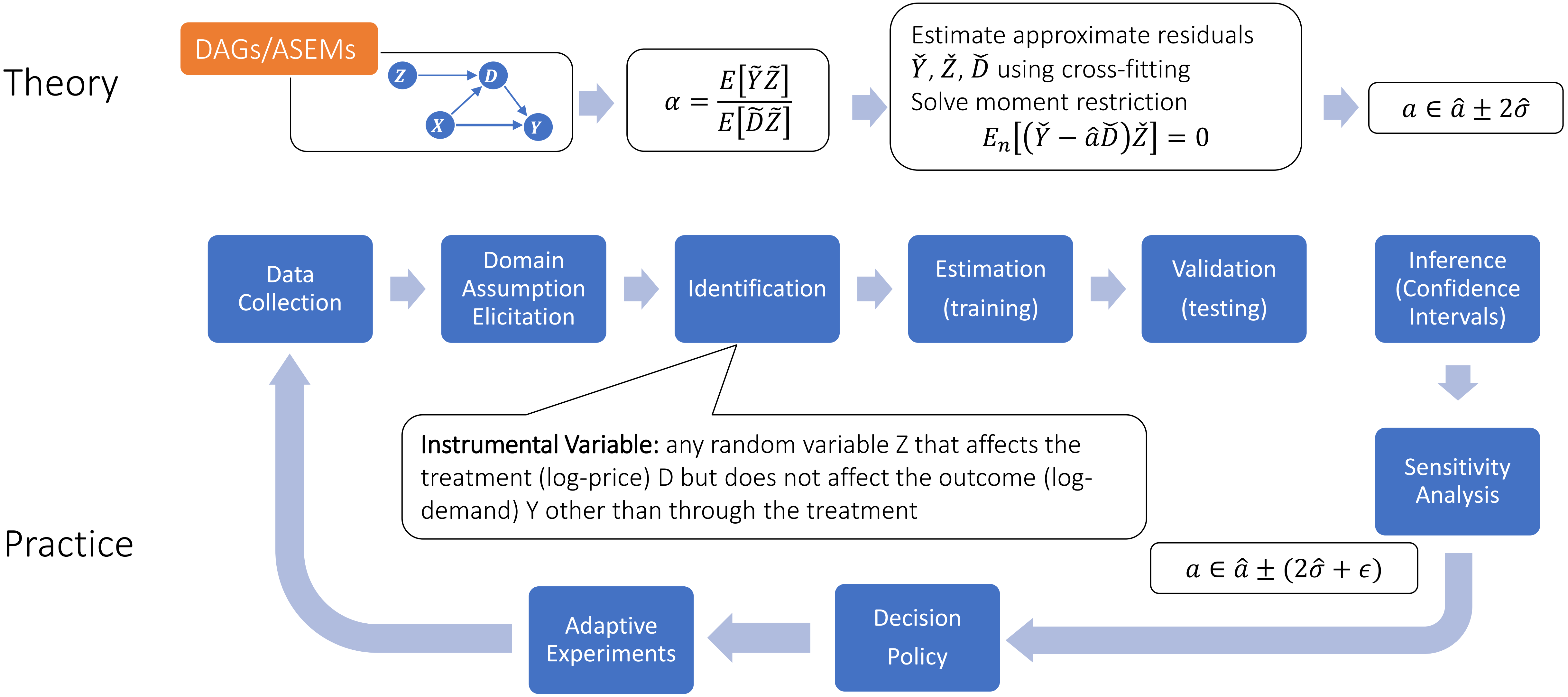

Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Non-Linear Predictive Models and Statistical Inference
Prediction with Non-Linear Models
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Estimation of Heterogeneous Causal Effects
Identification of Causal Effects in Longitudinal Data
Topics


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Non-Linear Predictive Models and Statistical Inference
Prediction with Non-Linear Models
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Estimation of Heterogeneous Causal Effects
Identification of Causal Effects in Longitudinal Data
Topics

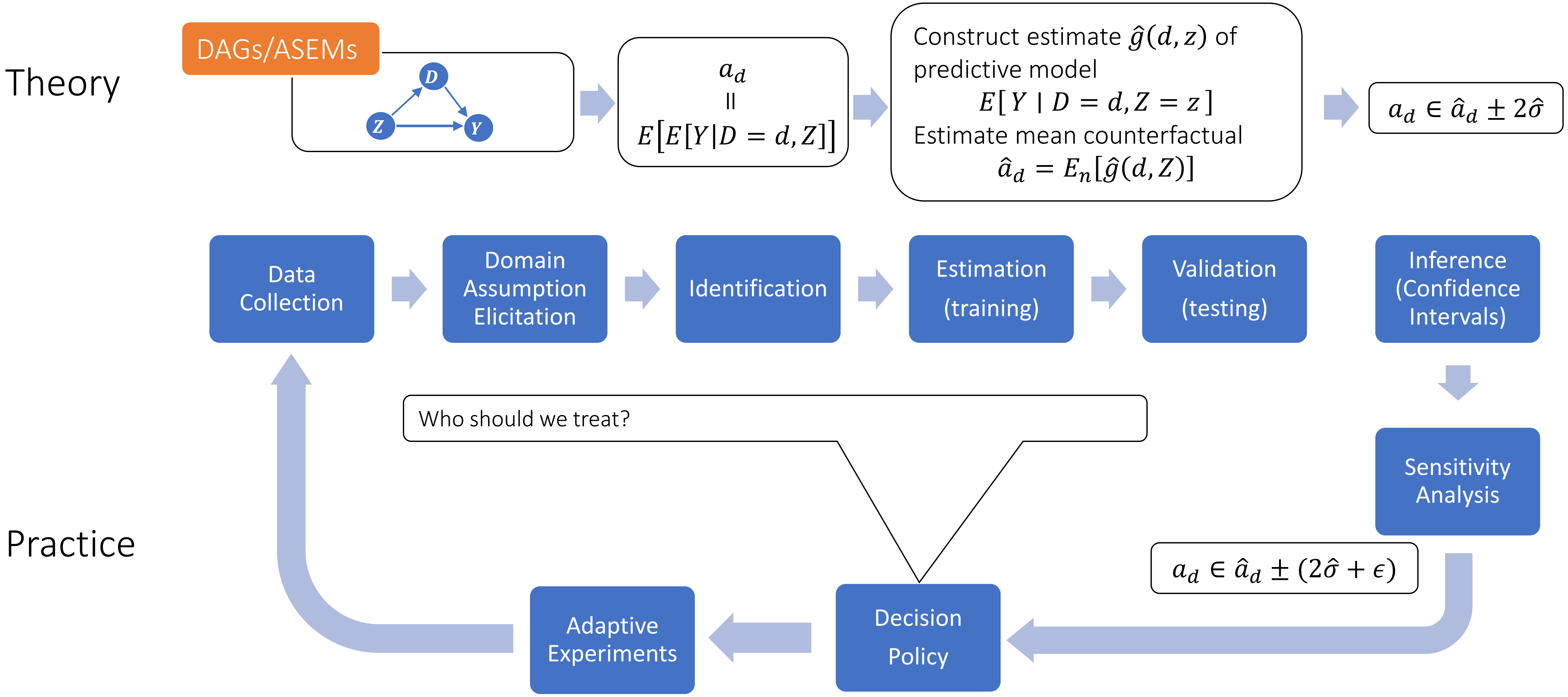
Goals for Today

- Heterogeneous Treatment Effects
- Statement of the problem
- A basic solution

Causal Inference Pipeline



Causal Inference Pipeline



Conditional Average Treatment Effects (CATE)

aka Heterogeneous Treatment Effects

Problem with Average Treatment Effect

- So far, we mostly focused on understanding average treatment effects

$$\theta = E[Y(1) - Y(0)]$$

- This quantity is not informative of who to treat
- At best we can use it to make a uniform decision for the population

$$\text{treat everyone if } \theta > 0 \text{ and don't treat otherwise}$$

- Such uniform policies can lead to severe adverse effects
- Such uniform analyses can lead us to miss on “responder subgroups”

Personalized (Refined) Policies

- To understand who to treat, we need to learn how effect varies
- Conditional Average Treatment Effect

$$\theta(x) = E[\,Y(1) - Y(0) \mid X = x\,]$$

- Allows us to understand differences (heterogeneities) in the response to treatment for different parts of the population
- We can deploy more refined “personalized” policies
- For every person that comes, we observe an $\mathrm{X} = x$ and decide

$$\text{treat if } \theta(x) > 0 \text{ else don't treat}$$

The intrinsic hardness of CATE

- The CATE quantity is not just a parameter
- It is a whole function…
- Learning such conditional expectation functions is inherently harder than learning parameters
- For instance: we might never have seen in our data other samples with the exact same x
- Such quantities are known as statistically “irregular” quantities
- We have seen such quantities when were solving the best prediction rule $E[Y|X]$

The intrinsic hardness of CATE

- Estimating CATE at least as hard as estimating the best prediction rule
- Inherently harder than estimating an “average”
- So far for our target causal quantities we wanted fast estimation rates and confidence intervals
- We were only ok with “decent” estimation rates for the auxiliary (nuisance) predictive models that entered our analysis

- We might want to relax our goals…

Different Approaches to Relaxing our Goals

- Goal 1: Maybe estimate a simpler projection (e.g. analogue of BLP)
- Goal 2: Confidence intervals for predictions of this simple projection
- Goal 3: Simultaneous confidence bands for predictions of this simple projection
- Goal 4: Estimation error rate for the true CATE
- Goal 5: Confidence intervals for the prediction of a CATE model
- Goal 6: Simultaneous confidence bands for joint predictions of CATE model

Linear Doubly Robust Learner

Meta-learner approaches: S-Learner, T-Learner, X-Learner, R-Learner, DR-Learner
Neural Network approaches: TARNet, CFR
Random Forest approaches: BART

Modified (honest) ML methods: Generalized Random Forest, Orthogonal Random Forest, Sub-sampled Nearest Neighbor Regression

?? (only classical non-parametric statistic results on confidence bands of non-parametric functions)

Policy Learning

- Goal 7: Go after optimal simple treatment policies; give me a policy with value close to the best
- Goal 8: Inference on value of candidate treatment policies
- Goal 9: Inference on value of optimal policy
- Goal 10: Identify responder or heterogeneous sub-groups; policies with statistical significance;

Doubly Robust Policy Evaluation

Doubly Robust Policy Learning

Different Approaches to Relaxing our Goals

- Goal 1: Maybe estimate a simpler projection (e.g. analogue of BLP)
- Goal 2: Confidence intervals for predictions of this simple projection
- Goal 3: Simultaneous confidence bands for predictions of this simple projection
- Goal 4: Estimation error rate for the true CATE
- Goal 5: Confidence intervals for the prediction of a CATE model
- Goal 6: Simultaneous confidence bands for joint predictions of CATE model

Linear Doubly Robust Learner

Meta-learner approaches: S-Learner, T-Learner, X-Learner, R-Learner, DR-Learner
Neural Network approaches: TARNet, CFR
Random Forest approaches: BART

Modified (honest) ML methods: Generalized Random Forest, Orthogonal Random Forest, Sub-sampled Nearest Neighbor Regression

?? (only classical non-parametric statistic results on confidence bands of non-parametric functions)

Policy Learning

- Goal 7: Go after optimal simple treatment policies; give me a policy with value close to the best
- Goal 8: Inference on value of candidate treatment policies
- Goal 9: Inference on value of optimal policy
- Goal 10: Identify responder or heterogeneous sub-groups; policies with statistical significance;

Doubly Robust Policy Evaluation

Doubly Robust Policy Learning

Best Linear Projection of CATE

Identification by Conditioning

- Under conditional ignorability

$$Y(1), Y(0) \perp\!\!\!\perp D \mid Z$$

- CATE can be identified by conditioning

$$\alpha(Z) \coloneqq E[Y(1) - Y(0)|Z] = E[Y|D = 1, Z] - E[Y|D = 0, Z] = \pi(Z)$$

- If we want a CATE on some subset of variables X

$$\theta(X) = E[\,\alpha(Z) \mid X\,] = E[\pi(Z) \mid X]$$

Identification with Propensity Scores

- Under conditional ignorability

$$Y(1), Y(0) \perp\!\!\!\perp D \mid Z$$

- CATE can be identified by propensity scores

$$\alpha(Z) \coloneqq E[Y(1) - Y(0)|Z] = E[Y\, H(D,Z)|Z] = \pi(Z)$$

$$H(D,Z) = \frac{D}{\Pr(D=1|Z)} - \frac{1-D}{1-\Pr(D=1|Z)}$$

- If we want a CATE on some subset of variables X

$$\theta(X) = E[\,\alpha(Z) \mid X\,] = E[\pi(Z) \mid X]$$

Doubly Robust Identification

- Under conditional ignorability
$$Y(1), Y(0) \perp\!\!\!\perp D \mid Z$$
- CATE can be identified by combination of conditioning and propensity scores
$$a(Z) \coloneqq E\big[\, g(1, Z) - g(0, Z) + H(D, Z)\,\big(Y - g(D, Z)\big) \mid Z \,\big] = \pi(Z)$$

$$H(D, Z) = \frac{D}{p(Z)} - \frac{1 - D}{1 - p(Z)}$$

$$g(D, Z) \coloneqq E[Y|D, Z], \qquad p(Z) \coloneqq \Pr(D = 1|Z)$$

- If we want a CATE on some subset of variables X
$$\theta(X) = E[\pi(Z) \mid X] = E\big[\, g(1, Z) - g(0, Z) + H(D, Z)\,\big(Y - g(D, Z)\big) \mid X \,\big]$$

From Identification to Estimation

- If we knew the propensity or regression, we have a random variable
$$Y_{DR}(g,p) \coloneqq g(1,Z) - g(0,Z) + H(D,Z)\left(Y - g(D,Z)\right)$$

- Such that what we are looking for is the CEF
$$\theta(X) \coloneqq E[Y_{DR}(g,p)|X]$$

- In the non-linear prediction section, we saw that this is the solution to the Best Prediction rule problem!

Blast from the Past: Best Prediction Rule

- Given n samples $(Z_1, Y_1), \dots, (Z_n, Y_n)$ drawn iid from a distribution D
- Want an estimate $\hat{g}$ that approximates the Best Prediction

$$g \coloneqq \operatorname*{argmin}_{\tilde{g}} E\left[\left(Y\ - \tilde{g}(Z)\right)^2\right]$$

- Best Prediction rule is Conditional Expectation Function (CEF)

$$g(Z) = E[Y|Z]$$

- We want our estimate $\tilde{g}$ to be close to g in RMSE

$$\|\hat{g} - g\| = \sqrt{E_Z\left(\hat{g}(S) - g(Z)\right)^2} \to 0, \qquad \text{as } n \to \infty$$

Blast from the Past: Linear CEF

- If CEF is assumed linear with respect to known engineered features
$$E[\,Y \mid Z\,] = \beta' \psi(Z)$$

- Then the Best Prediction rule (CEF) coincides with the Best Linear Prediction rule (BLP)

- We can use OLS if $\psi(Z)$ is low-dimensional (p≪n) or the multitude of approaches we learned if $\psi(Z)$ is high-dimensional (Lasso, ElasticNet, Ridge, Lava)

From Identification to Estimation

- If we knew the propensity or regression, we have a random variable
$$Y_{DR}(g,p) \coloneqq g(1,Z) - g(0,Z) - H(D,Z)\left(Y - g(D,Z)\right)$$
- Such that what we are looking for is the CEF
$$\theta(X) \coloneqq E[Y_{DR}(g,p)|X]$$
- We can reduce CATE estimation to a Best Prediction rule problem!
$$\theta \coloneqq \underset{g}{\operatorname{argmin}} E\left[\left(Y_{DR}(g,p) - g(X)\right)^2\right]$$
- ML techniques can be used to solve this problem and provide RMSE rates
$$\sqrt{E\left[\left(\theta(X) - \hat{\theta}(X)\right)^2\right]} \approx 0$$

Doubly Robust Learning

[Foster, Syrgkanis, '19 Orthogonal Statistical Learning]

- Split your data in half
  - Train ML model $\hat{g}$ for $g_0(D,Z) \triangleq E[Y|D,Z]$ on the first, predict on the second and calculate regression estimate of each potential outcome

$$\tilde{Y}_i^{(d)} = \hat{g}(d, Z_i)$$

  and vice versa

  - Train ML classification model $\hat{p}_d$ for $p_d(Z) \triangleq Pr[D = d\ |Z]$ on the first, predict on the second, calculate propensity $\hat{p}_{d,i} = \Pr[D = d|Z_i]$ and vice versa
- Calculate doubly robust values:

$$\tilde{Y}_{i,DR}^{(d)} = \tilde{Y}_i^{(d)} + \frac{\left(Y_i - \tilde{Y}_i^{(D_i)}\right) 1\{D_i = d\}}{\hat{p}_{d,i}}$$

- Any ML algorithm to solve the regression:

$$\tilde{Y}_{i,DR}^{(1)} - \tilde{Y}_{i,DR}^{(0)} \quad \sim \quad X$$

Blast from the Past: Best Linear Prediction (BLP) Problem

- The BLP minimizes the MSE

$$\min_{b \in \mathbb{R}^p} E\left[\left(Y - b'\psi(X)\right)^2\right]$$

- Since by the variance decomposition

$$E\left[\left(Y - b'\psi(X)\right)^2\right] = E\left[(Y - E[Y|X])^2\right] + E\left[\left(E[Y|X] - b'\psi(X)\right)^2\right]$$

- First part does not depend on b. The BLP minimizes

$$\min_{b \in \mathbb{R}^p} E\left[\left(E[Y|X] - b'\psi(X)\right)^2\right]$$

- The BLP is the **best linear approximation of the CEF**

From Identification to Estimation

- If we knew the propensity or regression, we have a random variable
$$Y_{DR}(g,p) \coloneqq g(1,Z) - g(0,Z) + H(D,Z)\left(Y - g(D,Z)\right)$$
- Such that what we are looking for is the CEF
$$\theta(X) \coloneqq E[Y_{DR}(g,p)|X]$$
- Estimate best linear approximation to the CATE via the BLP problem:
$$\beta \coloneqq \underset{b}{\operatorname{argmin}}\, E\left[\left(Y_{DR}(g,p) - b'\psi(X)\right)^2\right]$$

$$\theta_{BLP}(X) = \beta'\psi(X)$$

Normal Equations

- Equivalently, the solution to the normal equations

$$E\left[\left(Y_{DR}(g,p) - \beta'\psi(X)\right)\psi(X)\right] = 0$$

- Falls into the moment equation framework with nuisance components
- Nuisance components are g, p and target parameter is β
- Moment is Neyman orthogonal with respect to g, p (why?)
- Local insensitivity (orthogonality) holds even conditional on X

$$\lim_{\epsilon \to 0} \frac{E\left[\, Y_{DR}\left(g + \epsilon\, \nu_g, p + \epsilon\, \nu_p\right) \mid X \,\right] - E[\, Y_{DR}(g,p) \mid X \,]}{\epsilon} = 0$$

Asymptotic Normality

- *For conciseness, define* $\hat{Y} = Y_{DR}(\hat{g}, \hat{p})$ and $\Psi = \psi(X)$

$$E_n\left[\left(\hat{Y} - \hat{\beta}'\Psi\right)\Psi\right] = 0 \Leftrightarrow \hat{\beta} = E_n[\Psi\Psi']^{-1}E_n\left[\Psi\hat{Y}\right]$$

- *Assume* that $E[\Psi\Psi'] \succcurlyeq 0$

- *Assume* $\hat{g}, \hat{p}$ estimated on separate sample (or cross-fitting), are consistent and:

$$\sqrt{n} \cdot \mathrm{RMSE}(\hat{g}) \cdot \mathrm{RMSE}(\hat{p}) \rightarrow_p 0$$

- *Assume* random variables Y, D, Ψ have bounded fourth moments

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow_d N(0, V),$$

Estimate asymptotically normal

$$\hat{V} = E_n[\Psi\Psi']^{-1}E_n\left[\left(\hat{Y} - \hat{\beta}'\Psi\right)^2\Psi\Psi'\right]E_n[\Psi\Psi']^{-1}$$

"Sandwich" estimate of variance
(HC0 in OLS packages)

Asymptotic Linearity

- *For conciseness, define* $\hat{Y} = Y_{DR}(\hat{g}, \hat{p})$, $Y_0 = Y_{DR}(g_0, p_0)$ and $\Psi = \psi(X)$

$$E_n\left[\left(\hat{Y} - \hat{\beta}'\Psi\right)\Psi\right] = 0 \Leftrightarrow \hat{\beta} = E_n[\Psi\Psi']^{-1} E_n\left[\Psi\hat{Y}\right]$$

- *Assume* that $E[\Psi\Psi'] \succcurlyeq 0$

- *Assume* $\hat{g}, \hat{p}$ estimated on separate sample (or cross-fitting), are consistent and:

$$\sqrt{n} \cdot \mathrm{RMSE}(\hat{g}) \cdot \mathrm{RMSE}(\hat{p}) \to_p 0$$

- *Assume* random variables Y, D, Ψ have bounded fourth moments

$$\sqrt{n}(\hat{\beta} - \beta_0) \approx \sqrt{n} E_n[\Phi_0],$$

Estimate asymptotically behaves as an average of a random variable Φ_0

$$\Phi_0 = E[\Psi\Psi']^{-1}\Psi\left(Y_0 - \beta_0'\Psi\right),$$

Influence function of estimate $\hat{\beta}$

$$\hat{\Phi} = E_n[\Psi\Psi']^{-1}\Psi\left(\hat{Y} - \hat{\beta}'\Psi\right)$$

Estimate of influence function

Confidence Bands

- Any CATE BLP prediction is also asymptotically linear

$$\sqrt{n}\left(\hat{\theta}_{BLP}(x) - \theta_{BLP}(x)\right) = \sqrt{n} \cdot x'\left(\hat{\beta} - \beta_0\right) \approx \sqrt{n}\, E_n[x'\Phi_0]$$

- Holds jointly for all $x \in X$ (as long as $|X|$ not growing exponential in n)

$$\max_{x \in X} \left| \sqrt{n}\left(\hat{\theta}_{BLP}(x) - \theta_{BLP}(x)\right) - \sqrt{n}\, E_n[x'\Phi_0] \right| \approx 0$$

- High-dimensional CLT theorems also imply that jointly:

$$\left\{\sqrt{n}\left(\hat{\theta}_{BLP}(x) - \theta_{BLP}(x)\right)\right\}_{x \in X} \sim_a N(0, V), \qquad V_{x_1 x_2} = x_1' E[\Phi_0 \Phi_0'] x_2$$

Confidence Bands

- Analogous to inference on many coefficients
- Now the many predictions take the role of the many coefficients
- Confidence band: construct intervals

$$CI(x) := \left[\hat{\theta}(x) \pm c\sqrt{\hat{V}_{xx}/n}\right]$$

- Such that

$$\Pr\left(\forall x\colon\ \theta(x) \in CI(x)\right) \to 1 - \alpha$$

Confidence Bands

- Confidence band: construct intervals

$$CI(x) := \left[\hat{\theta}(x) \pm c\sqrt{\frac{\hat{V}_{xx}}{n}}\right], \qquad \Pr\big(\forall x\colon\ \theta(x) \in CI(x)\big) \to 1 - \alpha$$

- Note that

$$\Pr\big(\forall x\colon\ \theta(x) \in CI(x)\big) = \Pr\left(\max_{x\in X}\left|\frac{\sqrt{n}\left(\theta(x) - \hat{\theta}(x)\right)}{\sqrt{\hat{V}_{xx}}}\right| \le c\right)$$

- By Gaussian approximation, for $D = \mathrm{diag}(V)$

$$\Pr\left(\max_{x\in X}\left|\frac{\sqrt{n}\left(\theta(x) - \hat{\theta}(x)\right)}{\sqrt{\hat{V}_{xx}}}\right| \le c\right) \approx \Pr\left(\left\|N\left(0, D^{-1/2} V D^{-1/2}\right)\right\|_{\infty} \le c\right)$$

By Gaussian approximation, choose c as the $1 - \alpha$ quantile of the maximum entry in a gaussian vector drawn with covariance $D^{-1/2} V D^{-1/2}$

$$\mathrm{D} \coloneqq \mathrm{diag(V)} = \begin{bmatrix} V_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & V_{mm} \end{bmatrix}$$

For 95% confidence band, c slightly larger than 1.96

Computationally Friendlier Version: Multiplier Bootstrap

- By asymptotic linearity we know that:

$$\frac{\sqrt{n}\left(\theta(x) - \hat{\theta}(x)\right)}{\sqrt{\hat{V}_{xx}}} \approx \sqrt{n}\, E_n\left[\frac{x'\Phi_0}{\sqrt{V_{xx}}}\right]$$

- For every sample $i = 1 \dots n$, draw independent Gaussian, i.e. $\epsilon = (\epsilon_1, \dots, \epsilon_n) \sim N(0, I)$. Consider

$$Q_\epsilon(x) \coloneqq \frac{1}{\sqrt{n}} \sum_i \frac{x'\Phi_0}{\sqrt{V_{xx}}} \epsilon_i$$

- **Observation.** The vector of random variables $Q_\epsilon(x_1), \dots, Q_\epsilon\left(x_{|X|}\right)$, over the randomness of ϵ follows asymptotically the desired distribution $N\left(0, D^{-1/2} V D^{-1/2}\right)$
- Approximately the same holds for $\left(\hat{Q}(x_1), \dots, \hat{Q}\left(x_{|X|}\right)\right)$ with $\hat{Q}_\epsilon(x) \coloneqq \frac{1}{\sqrt{n}} \sum_i \frac{x'\hat{\Phi}}{\sqrt{\hat{V}_{xx}}} \epsilon_i$

- **Repeat process $\boldsymbol{B}$ times:** each repetition b draw vector $\epsilon_1^{(b)}, \dots, \epsilon_n^{(b)}$ and calculate maximum over x

$$Z^{(b)} \coloneqq \max_{x \in X} \left|\hat{Q}_{\epsilon^{(b)}}(x)\right|$$

- Set c to be the $1 - \alpha$ quantile of $Z^{(b)}$ over the B repetitions

Different Approaches to Relaxing our Goals

- Goal 1: Maybe estimate a simpler projection (e.g. analogue of BLP)
- Goal 2: Confidence intervals for predictions of this simple projection
- Goal 3: Simultaneous confidence bands for predictions of this simple projection
- Goal 4: Estimation error rate for the true CATE
- Goal 5: Confidence intervals for the prediction of a CATE model
- Goal 6: Simultaneous confidence bands for joint predictions of CATE model

Linear Doubly Robust Learner

Meta-learner approaches: S-Learner, T-Learner, X-Learner, R-Learner, DR-Learner
Neural Network approaches: TARNet, CFR
Random Forest approaches: BART

Modified (honest) ML methods: Generalized Random Forest, Orthogonal Random Forest, Sub-sampled Nearest Neighbor Regression

?? (only classical non-parametric statistic results on confidence bands of non-parametric functions)

Policy Learning

- Goal 7: Go after optimal simple treatment policies; give me a policy with value close to the best
- Goal 8: Inference on value of candidate treatment policies
- Goal 9: Inference on value of optimal policy
- Goal 10: Identify responder or heterogeneous sub-groups; policies with statistical significance;

Doubly Robust Policy Evaluation

Doubly Robust Policy Learning

Non-Parametric Confidence Intervals

Generalized Random Forest

- We want to estimate a solution $\theta(x)$ to a conditional moment restriction
$$E[\, m(Z;\theta) \mid X = x \,] = 0$$
- We do so by splitting constructing a tree that at each level optimizes the heterogeneity of the values of the local solution created at the resulting child nodes
- At the end we have many trees each defining a neighborhood structure
- For every candidate x we use the trees to define a set of weights with every training point and we solve the moment equation
$$\sum_i w_i(x)\, m(Z_i;\theta) = 0$$

Generalized Random Forest

- If each tree is built in an honest manner (i.e. samples used in the final weighted moment equation are separate from samples used to determine splits)
- If each tree is built in a balanced manner (at least some constant fraction on each side of the split)
- If each tree is built on a sub-sample without replacement, of an appropriate size
- Then the prediction $\theta(x)$ is asymptotically normal and we can construct confidence intervals via an appropriate bootstrap procedure

GRF for CATE

- We can do this with the residual moment:
$$E\left[\left(\tilde{Y} - \theta(x)\widetilde{D}\right)\widetilde{D} \mid X = x\right] = 0$$

- (Orthogonal Random Forest) We can also do a similar approach with the doubly robust targets
$$E[\, Y_{DR}(g,p) - \theta(x) \mid X = x\,] = 0$$

- We can also do this even when X is a subset of Z

Different Approaches to Relaxing our Goals

- Goal 1: Maybe estimate a simpler projection (e.g. analogue of BLP)
- Goal 2: Confidence intervals for predictions of this simple projection
- Goal 3: Simultaneous confidence bands for predictions of this simple projection
- Goal 4: Estimation error rate for the true CATE
- Goal 5: Confidence intervals for the prediction of a CATE model
- Goal 6: Simultaneous confidence bands for joint predictions of CATE model

Linear Doubly Robust Learner

Meta-learner approaches: S-Learner, T-Learner, X-Learner, R-Learner, DR-Learner
Neural Network approaches: TARNet, CFR
Random Forest approaches: BART

Modified (honest) ML methods: Generalized Random Forest, Orthogonal Random Forest, Sub-sampled Nearest Neighbor Regression

?? (only classical non-parametric statistic results on confidence bands of non-parametric functions)

Policy Learning

- Goal 7: Go after optimal simple treatment policies; give me a policy with value close to the best
- Goal 8: Inference on value of candidate treatment policies
- Goal 9: Inference on value of optimal policy
- Goal 10: Identify responder or heterogeneous sub-groups; policies with statistical significance;

Doubly Robust Policy Evaluation

Doubly Robust Policy Learning

Meta-Learning Approaches for CATE

Meta-Learning Idea

- We assume conditional ignorability: $Y(1), Y(0) \perp\!\!\!\perp D \mid Z$
- We want to estimate the CATE: $E[\,Y(1) - Y(0) \mid X\,], X \subseteq Z$
- If we can frame CATE as a conditional expectation function, then we can deploy any ML approach for solving the corresponding Best Prediction problem

Single Learner (S-Learner)

$$\theta(X) = E[\, g(1, Z) - g(0, Z) \mid X\,], \qquad g(D, Z) = E[Y|D, Z]$$

Meta-Algorithm:

- Run ML regression predicting Y from D, Z to learn g (preferably in a cross-fitting manner, i.e. fit on half the data and predict on the other half and vice versa)
- Run ML regression predicting $g(1, Z) - g(0, Z)$ from X

Two Learner (T-Learner)

$$\theta(X) = E[\, g(1,Z) - g(0,Z) \mid X \,], \qquad g(D,Z) = E[Y|D,Z]$$

Meta-Algorithm:

- Run ML regression predicting Y from Z on subset of data for which $D = 0$ to learn $g(0,\ \cdot)$ (preferably in a cross-fitting manner)
- Run ML regression predicting Y from Z on subset of data for which $D = 1$ to learn $g(1,\ \cdot)$ (preferably in a cross-fitting manner)
- Run an ML regression predicting $g(1,Z) - g(0,Z)$ from X

Doubly Robust Learner (DR-Learner)

$$\theta(X) = E[\,Y_{DR}(g,p) \mid X\,], \qquad Y_{DR}(g,p) := g(1,Z) - g(0,Z) + H(D,Z)\left(Y - g(D,Z)\right)$$

$$H(D,Z) = \frac{D}{p(Z)} - \frac{1-D}{1-p(Z)}, \qquad g(D,Z) := E[Y|D,Z], \qquad p(Z) := \Pr(D=1|Z)$$

Meta-Algorithm:

- Run ML regression to estimate $g(1,\ \cdot)$ and $g(0,\ \cdot)$ (either S or T Learner); preferably T-Learner and in cross-fitting manner
- Run ML classification to estimate $\Pr(D=1|Z)$ and calculate $H(D,Z)$; preferably in cross-fitting manner
- Run ML regression predicting $g(1,Z) - g(0,Z) + H(D,Z)\big(Y - g(D,X)\big)$ from X

Cross Learner (X-Learner)

$$\tau(Z) = \tau_1(Z) \coloneqq E[\,Y - E[\,Y \mid D = 0, Z\,] \mid D = 1, Z\,]$$

$$\tau(Z) = \tau_0(Z) \coloneqq E[\,E[\,Y \mid D = 1, Z\,] - Y \mid D = 0, Z\,]$$

For the **control group** I observe $Y(0) \equiv Y(D) = Y$
I can impute a counterfactual outcome $\hat{Y}(1)$, by fitting a response model $\hat{g}_1(Z) \approx E[Y|D = 1, Z]$ from the treatment group and predict on the control $\hat{Y}(1) = \hat{g}_1(Z)$

$$Y(1) - Y(0) \mid Z \quad \sim \quad \hat{g}_1(Z) - Y \mid D = 0, Z$$

For the **treated group** I observe $Y(1) \equiv Y(D) = Y$
I can impute a counterfactual outcome $\hat{Y}(0)$, by fitting a response model $\hat{g}_0(Z) \approx E[Y|D = 0, Z]$ from the control group and predict on the treated $\hat{Y}(0) = \hat{g}_0(Z)$

$$Y(1) - Y(0) \mid Z \quad \sim \quad Y - \hat{g}_0(Z) \mid D = 1, Z$$

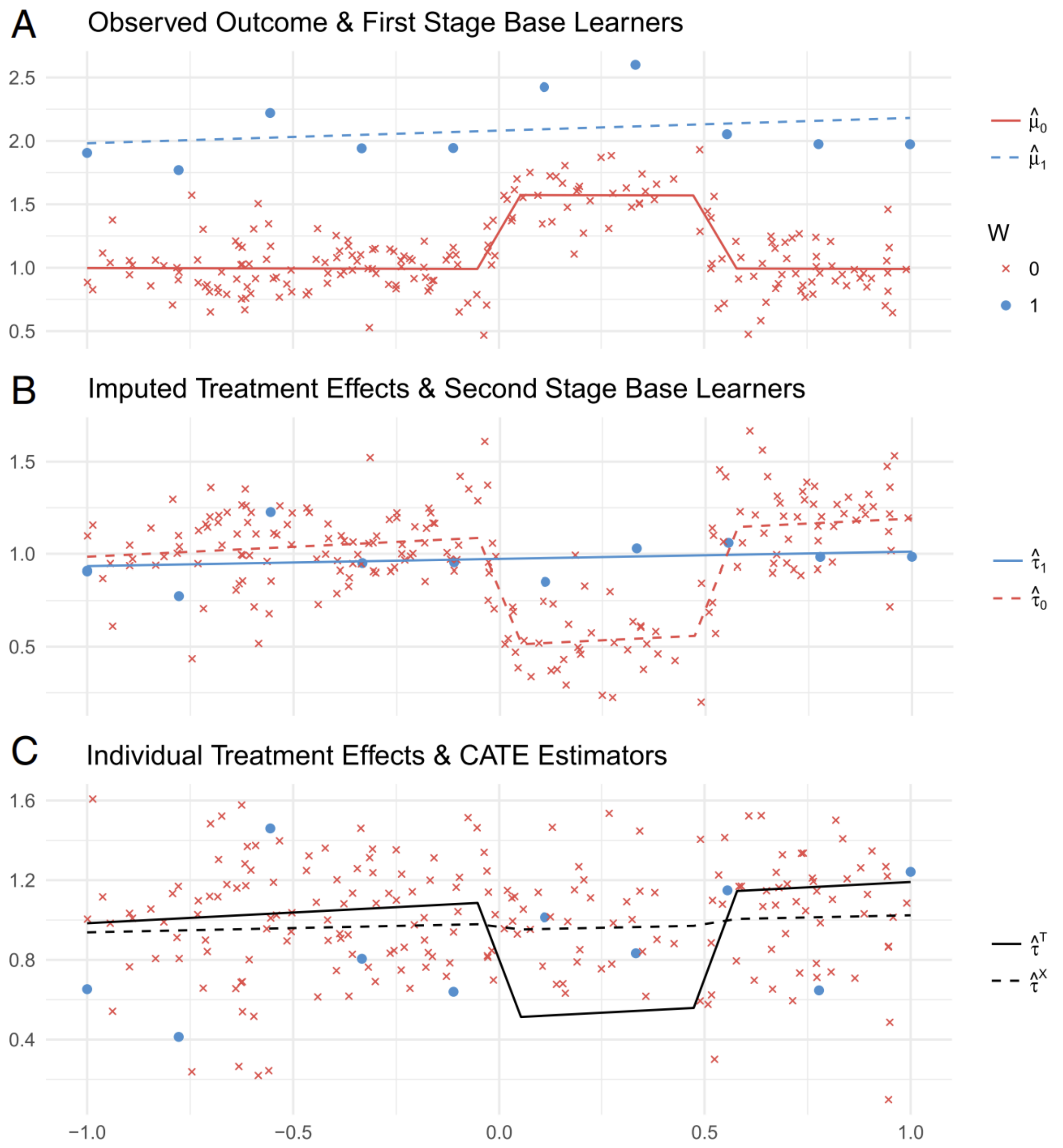
Cross Learner (X-Learner)

$$\hat{\tau}_1(Z) := E[\, Y - \hat{g}_0(Z) \mid D = 1, Z \,]$$

$$\hat{\tau}_0(Z) := E[\, \hat{g}_1(Z) - Y \mid D = 0, Z \,]$$

- Which one should we use?
- If for some Z most training data received $D = 1$, then model $\hat{g}_1$ will be a better predictor than $\hat{g}_0$; we should go with $\hat{\tau}_0$
- If for some Z most training data received $D = 0$, then model $\hat{g}_0$ will be a better predictor than $\hat{g}_1$; we should go with $\hat{\tau}_1$

$$\hat{\tau}(Z) = \Pr(D = 1|Z)\, \hat{\tau}_0(Z) + (1 - \Pr(D = 1|Z))\, \hat{\tau}_1(Z)$$



A
Observed Outcome & First Stage Base Learners
B
Imputed Treatment Effects & Second Stage Base Learners
C
Individual Treatment Effects & CATE Estimators
W
0
1

Cross Learner (X-Learner) Meta Algorithm

- Train ML regression $\hat{g}_0$ by predicting Y from Z among control samples
- Construct variables $T_i^1 \coloneqq Y - \hat{g}_0(Z)$ for all treated samples
- Train ML regression $\hat{\tau}_1$ by predicting T_i^1 from Z among treated samples
- Train ML regression $\hat{g}_1$ by predicting Y from Z among treated samples
- Construct variables $T_i^0 \coloneqq \hat{g}_1(Z) - Y$ for all control samples
- Train ML regression $\hat{\tau}_0$ by predicting T_i^0 from Z among control samples
- Train ML classifier to construct $\hat{p}(Z)$ predicting probability $D = 1$ given Z
- Train final ML regression model predicting from X the variable

$$\hat{\tau}(Z) = \hat{p}(Z)\, \hat{\tau}_0(Z) + \left(1 - \hat{p}(Z)\right) \hat{\tau}_1(Z)$$

Residual Learner (R-Learner)

- Since we have that:

$$\tau(Z) = E[Y|D = 1, Z] - E[Y|D = 0, Z]$$

- We can write:

$$E[Y|D, Z] = \tau(Z)D + f(Z)$$

- Equivalently:

$$Y = \tau(Z)D + f(Z) + \epsilon, \qquad E[\epsilon|D, Z] = 0$$

- If we further know that $\tau(Z) = \theta(X)$ (effect only depends on X)

$$E[Y|D, Z] = \theta(X)D + f(Z)$$

- We can then write:

$$Y - E[Y|Z] = \theta(X)\,(D - E[D|Z]) + \epsilon$$

Residual Learner (R-Learner)

- If we know that $\tau(Z) = \theta(X)$ (effect only depends on X), we can write
$$\tilde{Y} = \theta(X)\,\widetilde{D} + \epsilon, \qquad E[\epsilon\,|D, Z] = 0$$
- Equivalently, $\theta(\cdot)$ is the minimizer of the square loss:
$$E\left[\left(\tilde{Y} - \theta(X)\widetilde{D}\right)^2\right]$$
- Predict residual outcome $\tilde{Y}$ from residual treatment $\widetilde{D}$ and X with a model of the form $\theta(X)\widetilde{D}$
- Can also be phrased as a "weighted" square loss
$$E\left[\widetilde{D}^2\left(\tilde{Y}/\widetilde{D} - \theta(X)\right)^2\right]$$
- Predict $\tilde{Y}/\widetilde{D}$ from X with sample weights $\widetilde{D}^2$

Residual Learner (R-Learner) Meta Algorithm

- Train ML regression to predict Y from Z and calculate residual $\tilde{Y} \approx Y - E[Y|Z]$ (preferably in cross-fitting manner)
- Train ML regression to predict D from Z and calculate residual $\widetilde{D} \approx D - E[D|Z]$ (preferably in cross-fitting manner)
- Train ML regression with sample weights, to predict $\tilde{Y}/\widetilde{D}$ from X with sample weights $\widetilde{D}^2$

Residual Learner (R-Learner)

- When $\theta(X) = \alpha' \phi(X)$ for some known feature map ϕ then this is equivalent to learning heterogeneous effects with interactions

$$E\left[\left(\tilde{Y} - \alpha' \phi(X)\widetilde{D}\right)^2\right]$$

- Equivalent to OLS with outcome $\tilde{Y}$ and regressors $\phi(X)\widetilde{D}$

Residual Learner (R-Learner)

- If τ does not only depend on X then θ is a "projection"
- But it is a weighted one, it is the minimizer of the loss

$$E\left[\left(E\left[\tilde{Y} \mid Z, D\right] - \theta(X)\widetilde{D}\right)^2\right] = E\left[\left(\tau(Z)\widetilde{D} - \theta(X)\widetilde{D}\right)^2\right]$$
$$= E\left[\left(\tau(Z) - \theta(X)\right)^2 E\left[\widetilde{D}^2 \mid Z\right]\right] = E\left[\left(\tau(Z) - \theta(X)\right)^2 Var(D|Z)\right]$$

- We put more weight on regions of Z with more randomized treatment
- If some regions of the population were assigned treatments roughly deterministically, then they are ignored in the approximation

Comparing Meta-Learners

- S and T-Learners are typically poor performing as they heavily depend on outcome modelling; among them the T-Learner should be preferred
- X-Learner is a better version of S and T as it incorporates propensity knowledge
- DR-Learner and R-Learner, both possess “Neyman orthogonality” properties as they carefully combine outcome and treatment assignment modelling
- The error of the final cate model is not heavily impacted by the errors in the auxiliary models (Orthogonal Statistical Learning)
- DR-Learner estimates un-weighted projection of true CATE on model space, but can be “high-variance” due to inverse propensity
- R-Learner estimates variance weighted projection but is much more stable to extreme propensities as it never divides by propensity.

Model Selection and Evaluation

Model Selection within Method

- Each of the meta learners is defined based on a loss function
- We can use loss function for model selection within each meta-learning approach
- For each hyper-parameter evaluate the out-of-sample loss in a cross-validation manner and choose the best hyper-parameter for the meta-learning method
- This way we have M CATE models, $\hat{\theta}_1, \dots, \hat{\theta}_M$ from each meta-learning approach

Model Selection Across Methods

- To compare across any CATE learner, we can evaluate based on a "Neyman orthogonal loss", which is robust to nuisance estimation
- **R-Loss:** for a separate sample, calculate residuals $\tilde{Y}, \widetilde{D}$ in a cross-fitting manner. For any candidate CATE model θ evaluate
$$L(\theta) \coloneqq E\left[\left(\tilde{Y} - \theta(X)\widetilde{D}\right)^2\right]$$
- **DR-Loss:** for a separate sample, calculate regression model g (using T-Learner) and propensity model p. For any candidate CATE model θ evaluate
$$L(\theta) \coloneqq E\left[\left(Y_{DR}(g,p) - \theta(X)\right)^2\right]$$
- Given M estimated CATE models $\hat{\theta}_1, \dots, \hat{\theta}_M$, evaluate the loss out-of-sample and choose the best model
$$m^* \coloneqq \operatorname*{argmin}_m L(\theta_m)$$

Ensembling and Stacking

- We can also use these losses to construct stacked ensembles of a set of CATE models $\left(\hat{\theta}_1, \dots, \hat{\theta}_M\right)$:

$$\hat{\theta}_w(X) = \sum_{m=1}^{M} w_m \hat{\theta}_m(X)$$

- **Stacking with R-Loss:** (penalized) linear regression predicting $\tilde{Y}$ with regressors $\theta_1(X)\widetilde{D}, \dots, \theta_M(X)\widetilde{D}$

$$\min_w E_n\left[\left(\tilde{Y} - \sum_{m=1}^{M} w_m \hat{\theta}_m(X)\,\widetilde{D}\right)^2\right] + \lambda \text{Penalty}(w)$$

- **Stacking with DR-Loss:** (penalized) linear regression predicting $Y_{DR}(g, p)$ with regressors $\theta_1(X), \dots, \theta_M(X)$

$$\min_w E_n\left[\left(Y_{DR}(g, p) - \sum_{m=1}^{M} w_m \hat{\theta}_m(X)\right)^2\right] + \lambda \text{Penalty}(w)$$

Evaluation via Testing Approaches

- If CATE model $\hat{\theta}$ was good, then out-of-sample BLP of CATE, when using $\left(1, \hat{\theta}(X)\right)$ as feature map, should assign a lot of weight on $\hat{\theta}(X)$
- Run OLS regression predicting $Y_{DR}(g, p)$ using regressors $\left(1, \hat{\theta}(X)\right)$

$$E\left[\left(Y_{DR}(g, p) - \beta_0 - \beta_1 \hat{\theta}(X)\right)^2\right]$$

- Construct confidence intervals and test whether $\beta_1 \neq 0$; then $\theta(X)$ correlates with the true CATE! Ideally $(\beta_0 = 0, \beta_1 = 1)$
- The parameter β_1 is identifying the quantity (in the population limit):

$$\beta_1 := \frac{Cov\left(Y(1) - Y(0), \hat{\theta}(X)\right)}{Var\left(\hat{\theta}(X)\right)}$$

Validation via GATEs

- For any large enough group G, we can calculate out-of-sample group average effects by simply averaging $Y_{DR}(g, p)$
$$GATE(G) \coloneqq E[Y(1) - Y(0) | X \in G] = E[Y_{DR}(g, p) | X \in G]$$

- If the CATE model $\hat{\theta}$ is accurate, then if we restrict to some group G then the average of $\hat{\theta}$ over this group, should match the out-of-sample group average treatment effect
$$E\left[\hat{\theta}(X) \middle| X \in G\right] \approx GATE(G)$$

- We can measure such GATE discrepancies out-of-sample

Validation via Calibration

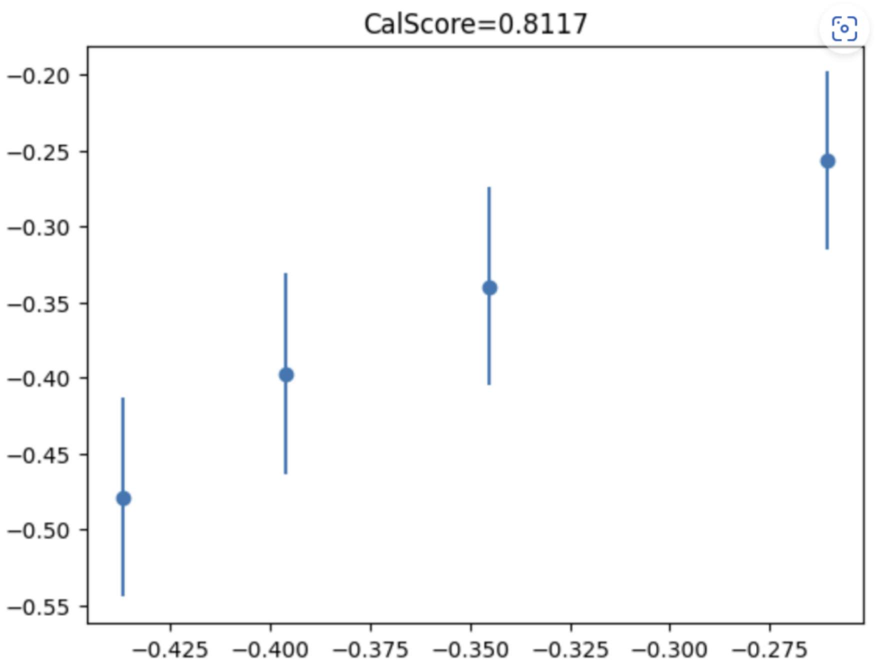
- One natural definition of groups is the "percentile groups of the CATE predictions"
- For the top 25% of the CATE predictions based on the model θ, the mean of model predictions, should match the out-of-sample GATE for that group
- Consider a set of quantiles $q_1, .., q_K$ (e.g. 0, 25, 50, 75)
- Consider the distribution D of $\hat{\theta}(X)$ over the training data X
- Let G_i be the groups defined as $\left\{X: \hat{\theta}(X) \in [q_i\ q_{i+1}]\ quantile\ of\ D\right\}$

$$\tau_i \coloneqq E\left[\hat{\theta}(X) \middle| X \in G_i\right] \approx GATE(G_i) \coloneqq E[Y_{DR}(g, p) | X \in G_i]$$

- Calibration score:

$$\mathrm{CalScore}(\theta) \coloneqq \sum_i \Pr(G_i) \cdot |\tau_i - GATE(G_i)|$$

- Normalized calibration score: $1 - \frac{\mathrm{CalScore}(\hat{\theta})}{\mathrm{CalScore}(constant\ CATE{=}E[Y_{DR}(g,p)])}$



Testing for Heterogeneity

- We can easily construct joint confidence intervals for all the GATEs
- GATEs are the coefficients in the BLP of CATE using group one-hot-encoding as features

$$E\left[\left(Y_{DR}(g,p) - \beta'(1\{X \in G_1\}, \dots, 1\{X \in G_K\})\right)^2\right]$$

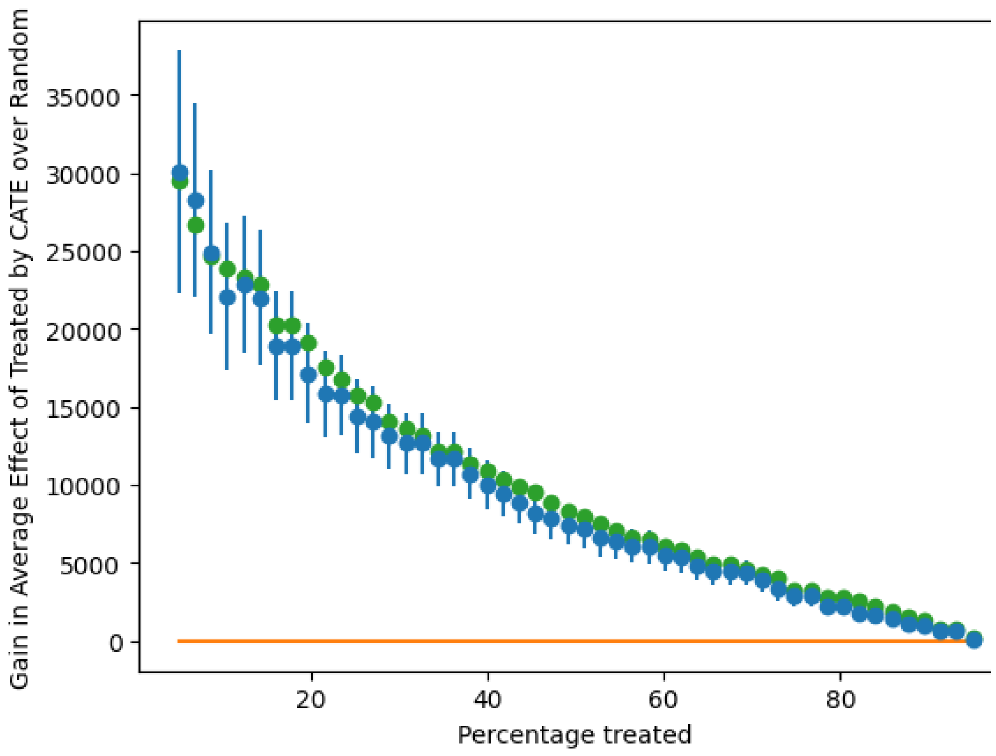
- We can use joint confidence intervals for BLP via the DR-Learner
- If there was heterogeneity, then we should have that there are GATEs whose confidence intervals are non-overlapping

Stratification Motivated Evaluation

- If we were to “prioritize” into treatment based on $\hat{\theta}$ with a target to treat around q-percent of population then what would be the GATE of the treated group
- Consider distribution D_n of $\theta(X)$ over training data X
- We can define the groups:

$$G_q \coloneqq \{X: \theta(X) \geq (1-q) - th\ quantile\ of\ D_n\}$$
$$\tau(q) = E\left[Y_{DR}(g,p) \mid X \in G_q \right] - E[Y_{DR}(g,p)]$$

- Ideally, $\tau(q)$ should be always positive and increasing!
- AUTOC $\approx$ the area under the curve $\tau(q)$



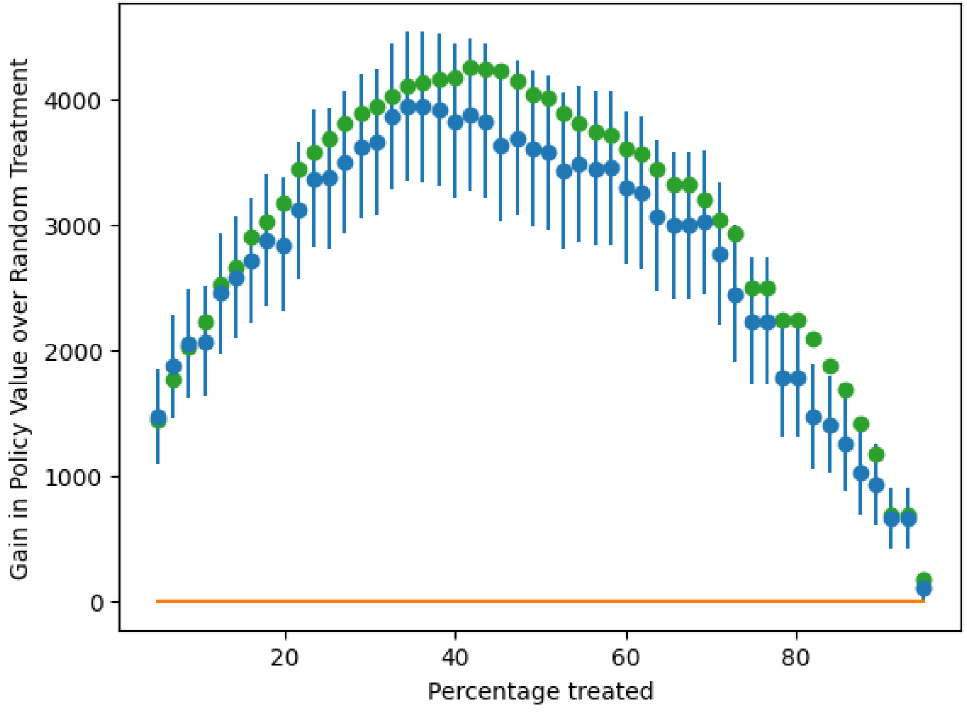
Stratification Motivated Evaluation

- If we were "prioritize" into treatment based on $\hat{\theta}$ with a target to treat around q-percent of the population then what would be the policy value we would get over treating q percentage at random
- Consider distribution D_n of $\hat{\theta}(X)$ over the training data X
- We can define the group:

$$G_q \coloneqq \{X: \hat{\theta}(X) \geq (1-q) - th\ quantile\ of\ D_n\}$$

$$\tau_Q(q) = \Pr(X \in G_q)\ (E[\,Y_{DR}(g,p) \mid X \in G_q\,] - E[Y_{DR}(g,p)])$$

- Ideally, $\tau_Q(q)$ should be large positive for some values!
- QINI ≈ the area under the curve $\tau_Q(q)$



Different Approaches to Relaxing our Goals

- Goal 1: Maybe estimate a simpler projection (e.g. analogue of BLP)
- Goal 2: Confidence intervals for predictions of this simple projection
- Goal 3: Simultaneous confidence bands for predictions of this simple projection

Linear Doubly Robust Learner

- Goal 4: Estimation error rate for the true CATE
- Goal 5: Confidence intervals for the prediction of a CATE model
- Goal 6: Simultaneous confidence bands for joint predictions of CATE model

Meta-learner approaches: S-Learner, T-Learner, X-Learner, R-Learner, DR-Learner
Neural Network approaches: TARNet, CFR
Random Forest approaches: BART

Modified (honest) ML methods: Generalized Random Forest, Orthogonal Random Forest, Sub-sampled Nearest Neighbor Regression

?? (only classical non-parametric statistic results on confidence bands of non-parametric functions)

Policy Learning

- Goal 7: Go after optimal simple treatment policies; give me a policy with value close to the best
- Goal 8: Inference on value of candidate treatment policies
- Goal 9: Inference on value of optimal policy
- Goal 10: Identify responder or heterogeneous sub-groups; policies with statistical significance;

Doubly Robust Policy Evaluation

Doubly Robust Policy Learning

Policy Learning

Candidate Policy

- What if I have a candidate policy π on who to treat
- The average policy effect is of the form:

$$V(\pi) = E\left[\pi(X)\left(Y(1) - Y(0)\right)\right]$$

- Under conditional ignorability:

$$V(\pi) = E[\pi(X)(E[Y|D = 1, Z] - E[Y|D = 0, Z])]$$

- We can also measure performance via the doubly robust outcome

$$V(\pi) = E[\pi(X)\, Y_{DR}(g, p)]$$

- Also falls in the Neyman orthogonal moment estimation framework

$$E[\pi(X) Y_{DR}(g, p) - \theta] = 0$$

- We can easily construct confidence intervals

Policy Optimization

- We can optimize over a space of policies Π on the samples

$$\hat{V}(\pi) = E_n[\pi(X) Y_{DR}(\hat{g}, \hat{p})]$$

- Regret:

$$\max_{\pi \in \Pi} V(\pi) - V(\hat{\pi})$$

- Regret not impacted a lot by errors in $\hat{g}$ or $\hat{p}$
- Performance as if true g, p (assuming estimation rates of $n^{-\frac{1}{4}}$)
- Maximizing $V(\pi)$ can be viewed as sample-weighted classification, with labels $\mathbf{sign}\big(Y_{DR}(g, p)\big)$ and sample weights $|Y_{DR}(g, p)|$
- Any classification method can be deployed