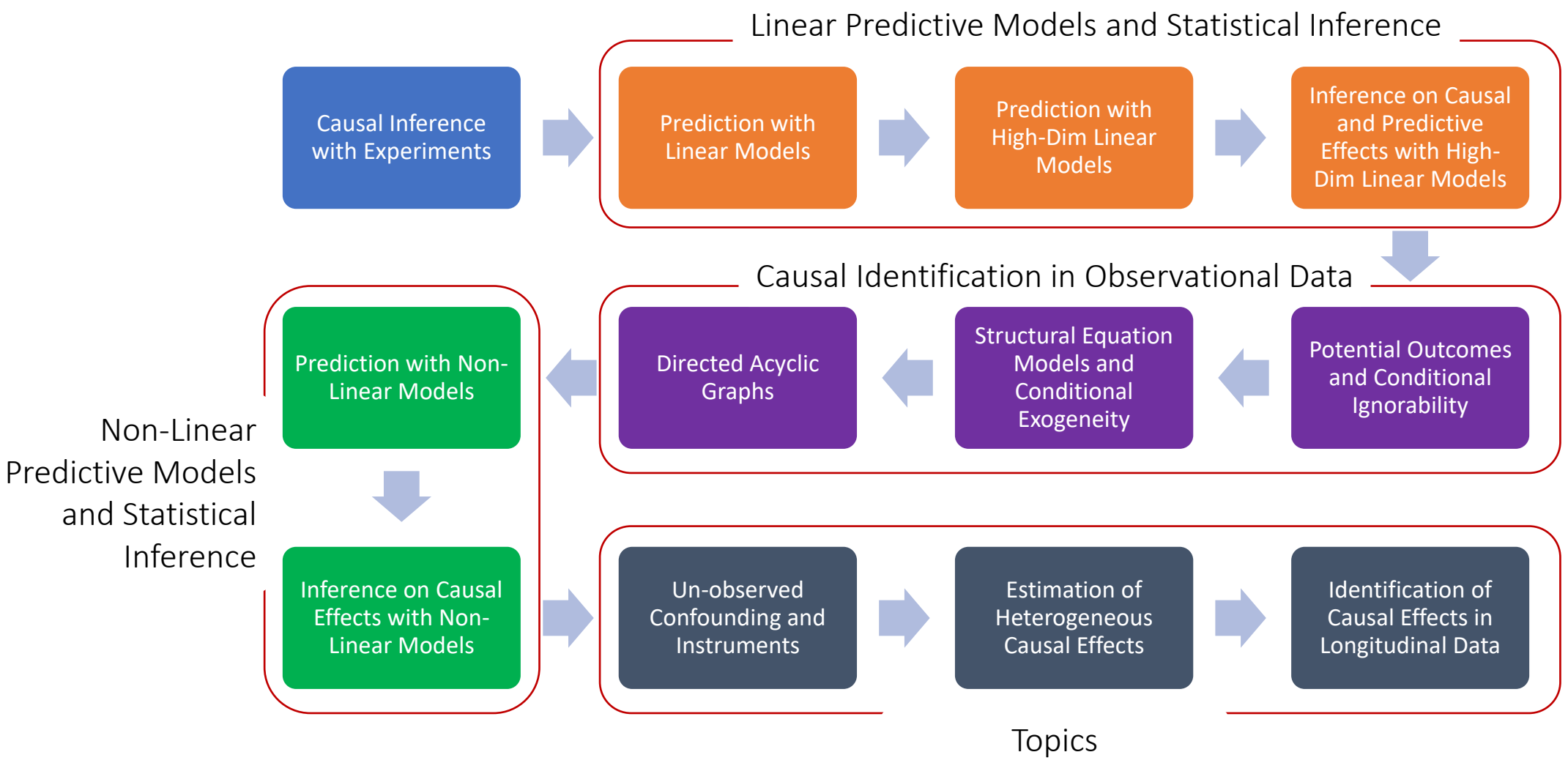
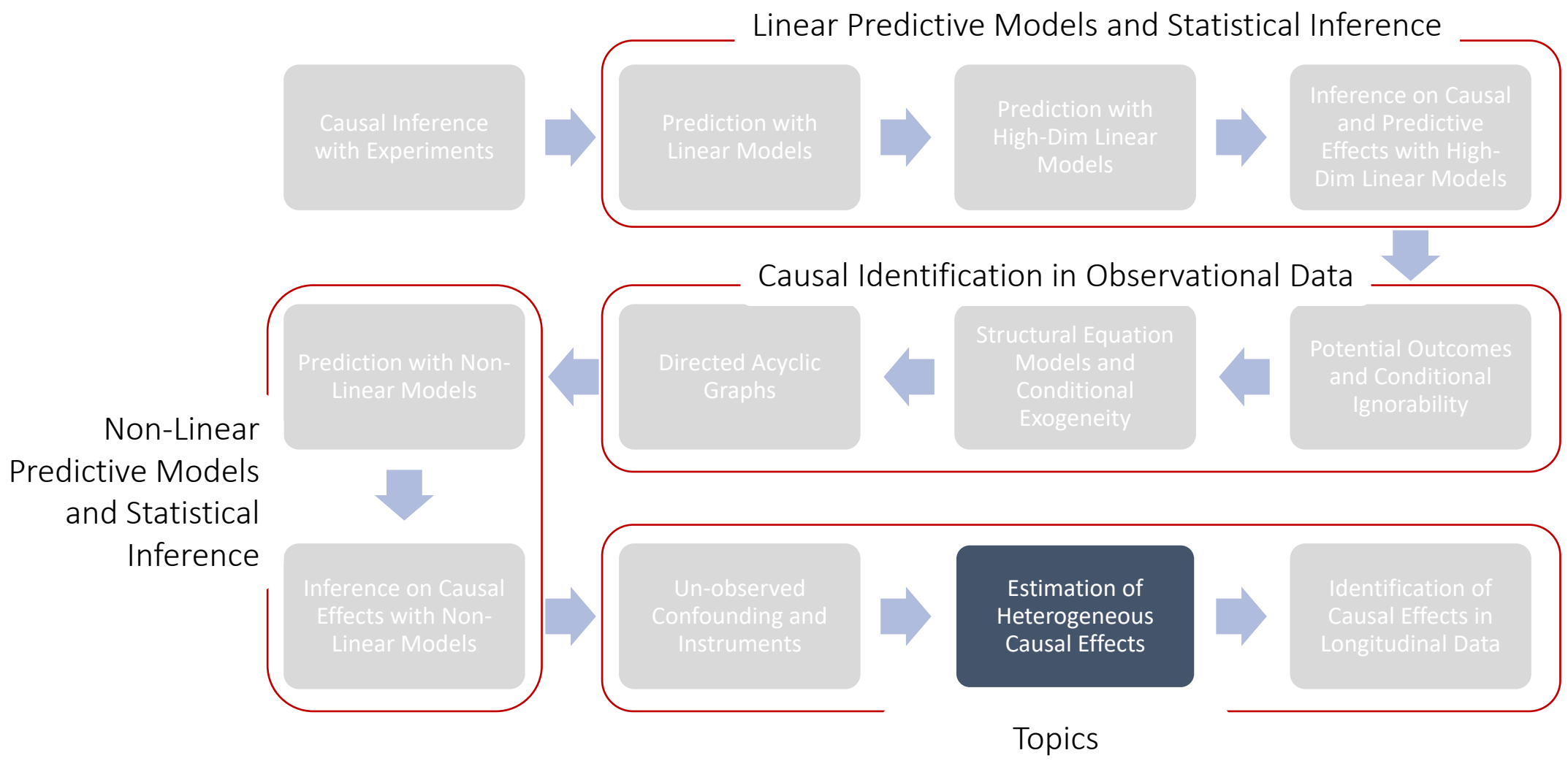
MS&E 228: Heterogeneous Treatment Effects

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

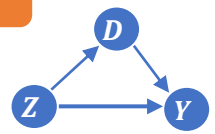

Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Prediction with Non-Linear Models
Non-Linear Predictive Models and Statistical Inference
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

Recap of Last Lecture

Causal Inference Pipeline

Theory

DAGs/ASEMs



$$a_d = E[E[Y|D=d,Z]]$$

Construct estimate $\hat{g}(d,z)$ of predictive model

$$E[Y \mid D=d, Z=z]$$

Estimate mean counterfactual

$$\hat{a}_d = E_n[\hat{g}(d,Z)]$$

$$a_d \in \hat{a}_d \pm 2\hat{\sigma}$$

Practice

Data Collection → Domain Assumption Elicitation → Identification → Estimation (training) → Validation (testing) → Inference (Confidence Intervals) → Sensitivity Analysis

$$a_d \in \hat{a}_d \pm (2\hat{\sigma} + \epsilon)$$

Who should we treat?

Decision Policy → Adaptive Experiments → Data Collection

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

Goals for Today

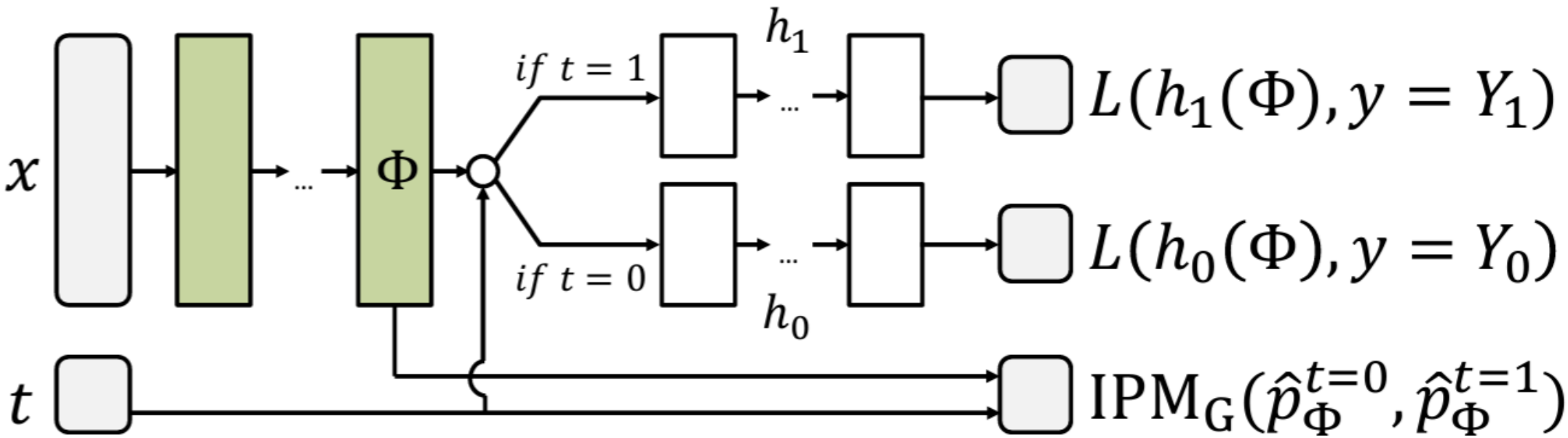
- Meta-Learners for Heterogeneous Treatment Effects

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.

Neural Network CATE Learners (CFR Net)

Shalit et al. 17



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 $\tilde{Y}, \tilde{D}$ in a cross-fitting manner. For any candidate CATE model θ evaluate

$$L(\theta) \coloneqq E\left[\left(\tilde{Y} - \theta(X)\tilde{D}\right)^2\right]$$

- **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

$$m^* \coloneqq \underset{m}{\operatorname{argmin}}\, L(\theta_m)$$

Ensembling and Stacking

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

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

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

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

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

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

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

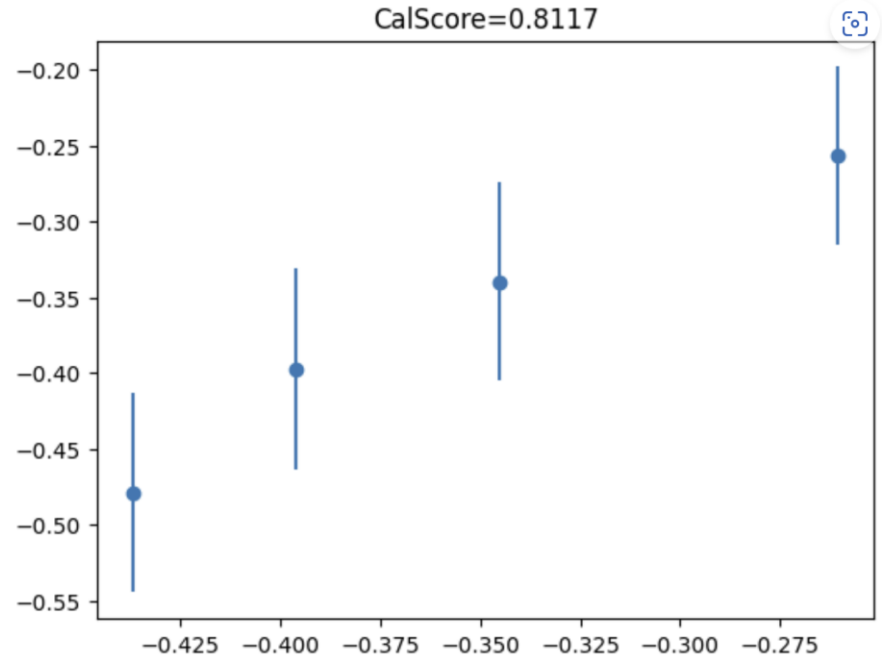
- One natural definition of groups is the “percentile groups of the CATE predictions”
- For the top 25% of the CATE predictions based on the model θ, the mean of model predictions, should match the out-of-sample GATE for that group
- Consider a set of quantiles $q_1, .., q_K$ (e.g. 0, 25, 50, 75)
- Consider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

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

- Calibration score:

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

- 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

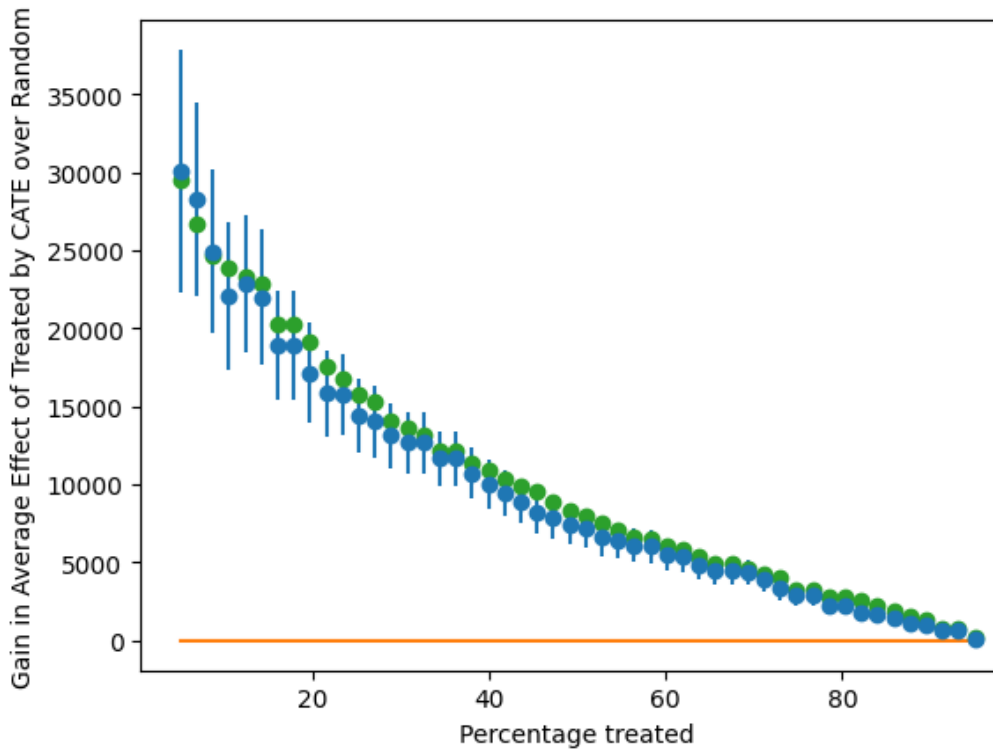
Stratification Motivated Evaluation

- 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:

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

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

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



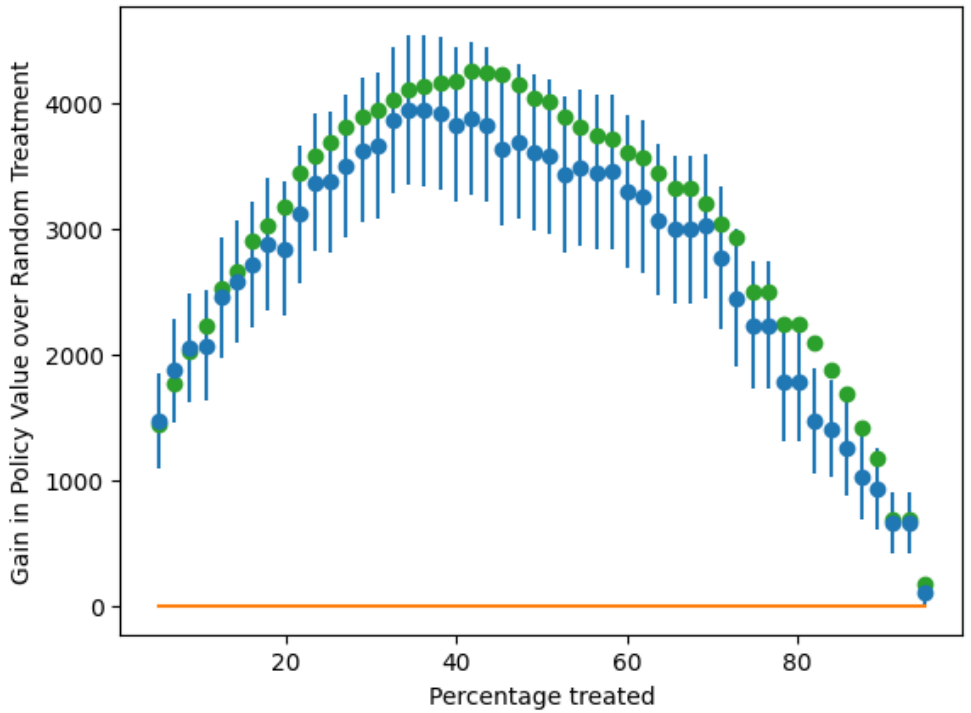
Stratification Motivated Evaluation

- 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:

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

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

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



Policy Learning

Candidate Policy

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

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

- 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

Non-Parametric Confidence Intervals

Generalized Random Forest

- We want to estimate a solution to a conditional moment restriction

$$\theta(x) \coloneqq E[\, m(Z;\theta) \mid X = x \,]$$

- 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 children 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:

$$E\left[\left(\tilde{Y} - \theta(x)\tilde{D}\right)\tilde{D} \mid X = x\right] = 0$$

- (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