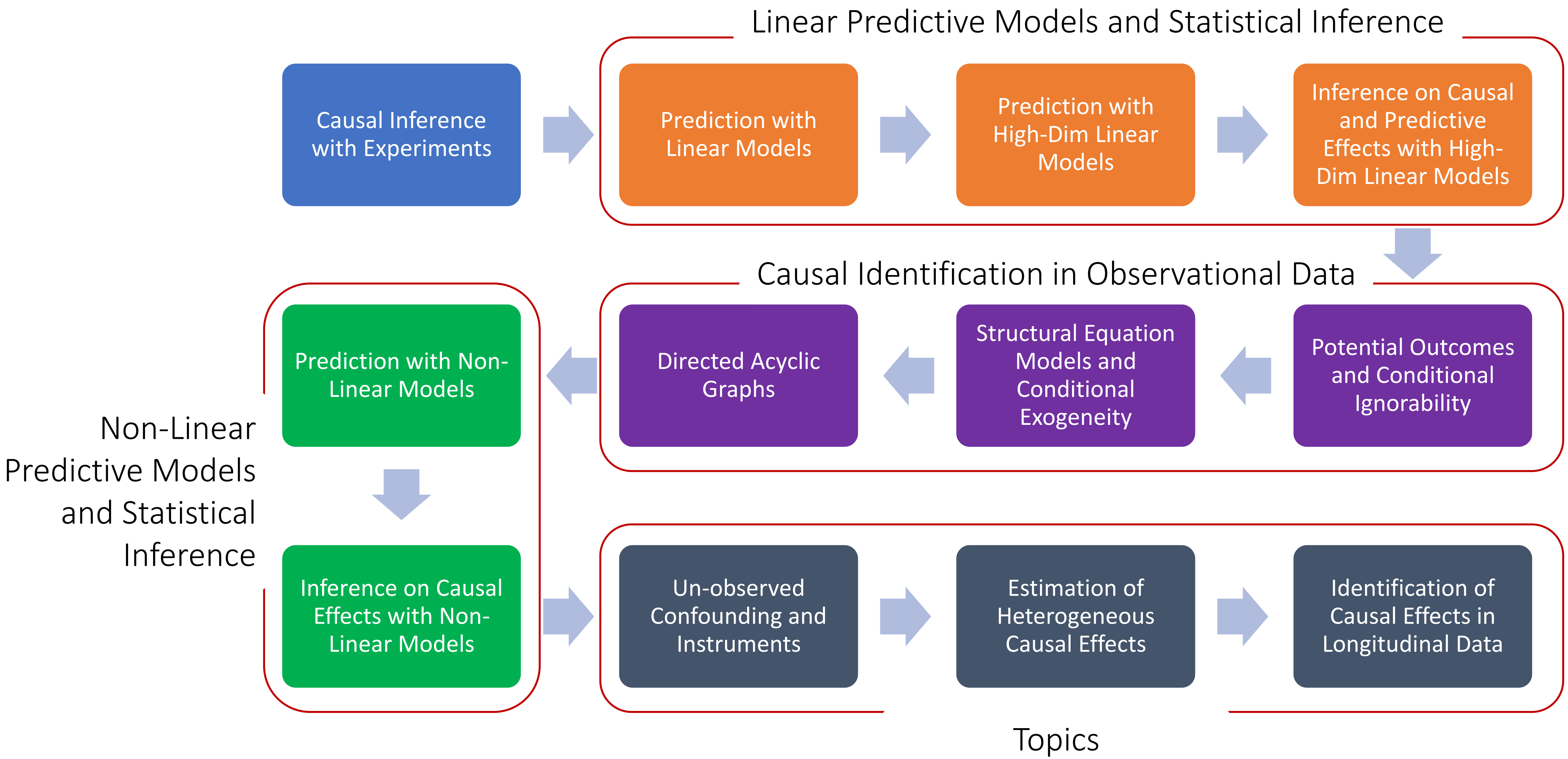
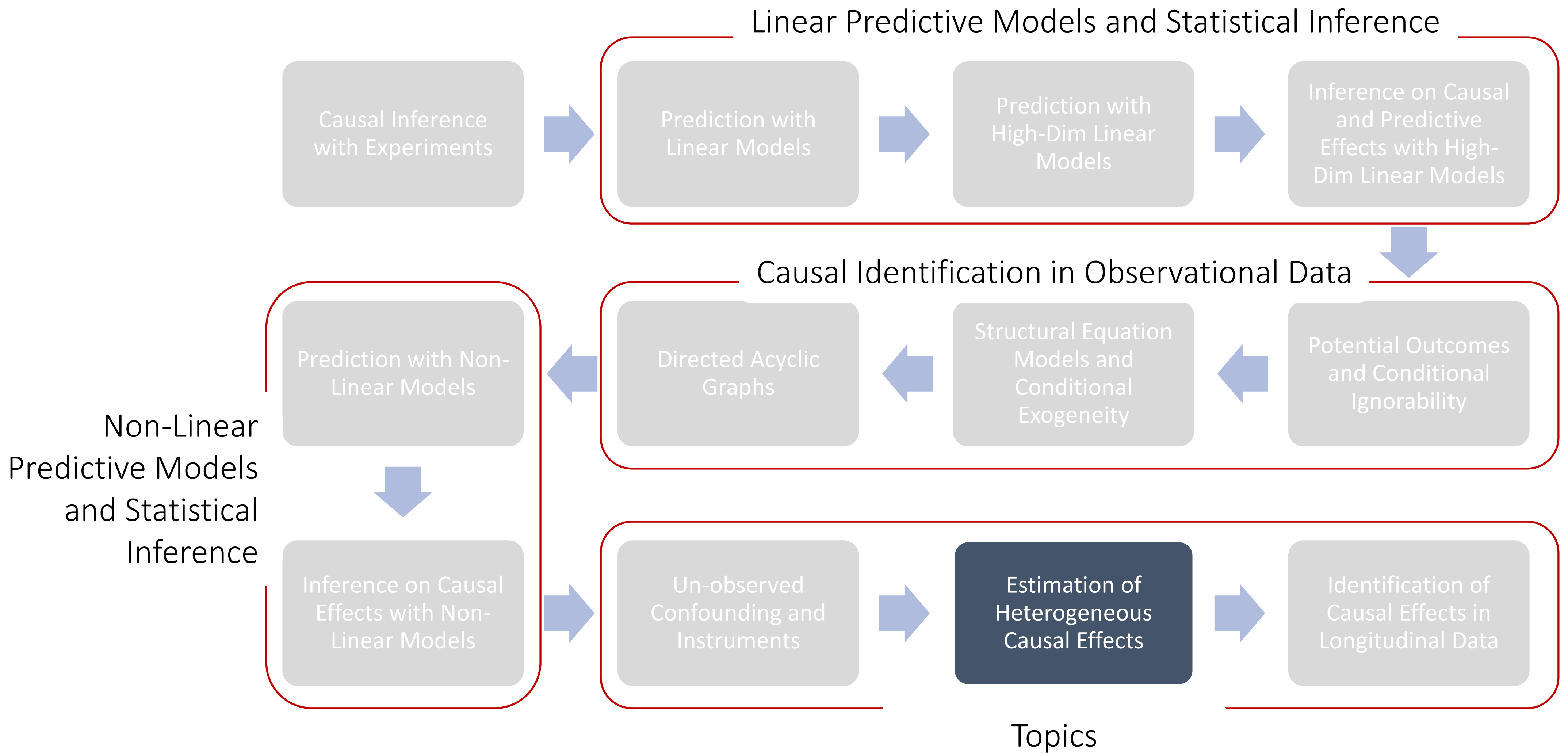
MS&E 228: Heterogeneous Treatment Effects

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
Prediction with Non-Linear Models
Directed Acyclic Graphs
Structural Equation Models and Conditional Exogeneity
Potential Outcomes and Conditional Ignorability
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

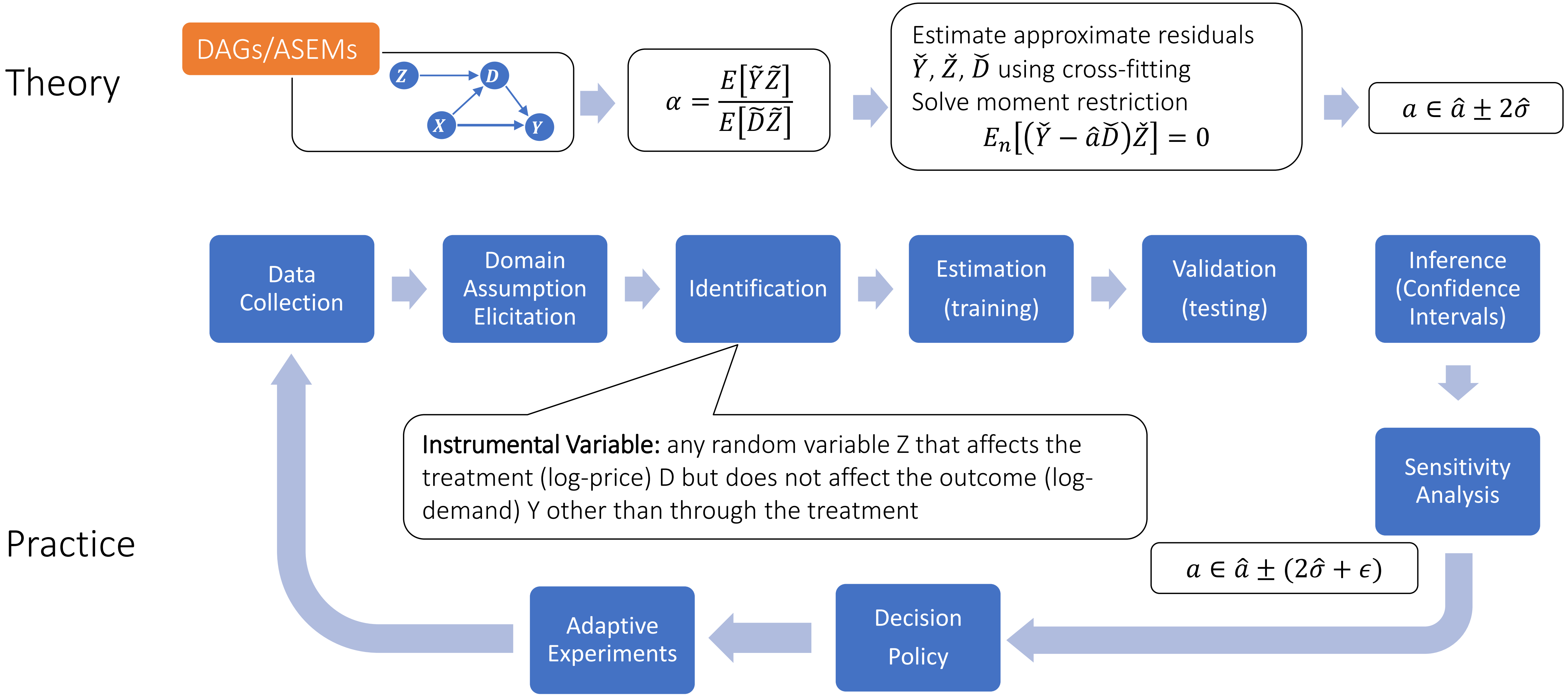


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Prediction with Non-Linear Models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

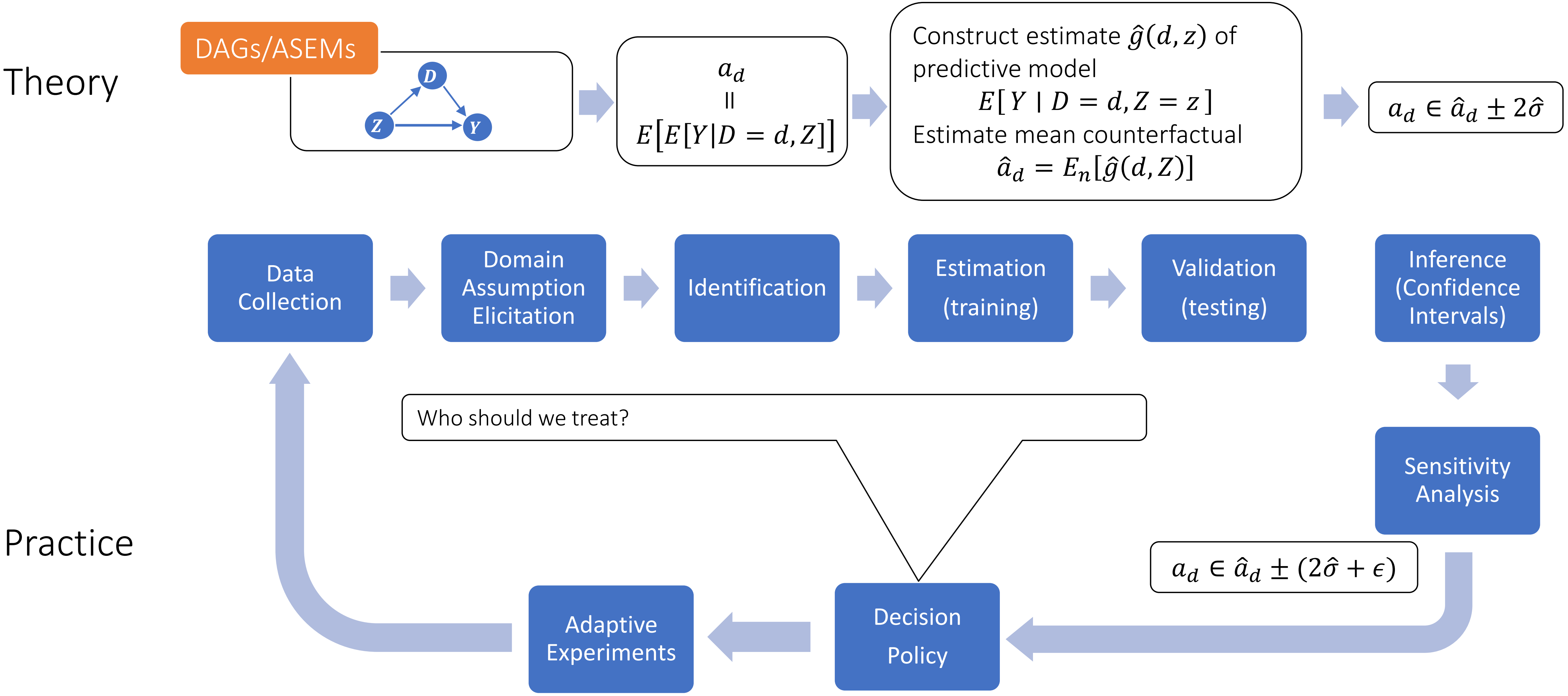
Goals for Today

- 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
$$Y_{DR}(g,p) := g(1,Z) - g(0,Z) - H(D,Z)\left(Y - g(D,Z)\right)$$

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

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

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!
$$\theta \coloneqq \operatorname*{argmin}_{g} E\left[\left(Y_{DR}(g,p) - g(X)\right)^2\right]$$
- 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
$$\lim_{\epsilon \to 0} \frac{E\left[\, Y_{DR}\left(g + \epsilon\, \nu_g, p + \epsilon\, \nu_p\right) \mid X \,\right] - E[\, Y_{DR}(g,p) \mid X \,]}{\epsilon} = 0$$

Main Theorem (linear moments)

- If moments are linear
$$m(Z;\theta,g)=\nu(Z;g)-a(Z;g)\theta$$
- Estimate is closed form:
$$\hat{\theta}=\hat{J}^{-1}E_n[\nu(Z;g)],\qquad \hat{J}=E_n[a(Z;g)]$$
- Then the estimate $\hat{\theta}$ is *asymptotically linear*
$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\approx\sqrt{n}\,E_n[\phi_0(Z)],\qquad \phi_0(Z)=-J_0^{-1}\,m(Z;\theta_0,g_0),\qquad J_0\coloneqq E[a(Z;\,g_0)]$$
- Consequently, it is *asymptotically normal*
$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\sim_a N(0,V),\qquad V\coloneqq E[\phi_0(Z)\phi_0(Z)']$$
- *Confidence intervals* for any projection based on estimate of variance are asymptotically valid
$$\ell'\theta\in\left[\ell'\hat{\theta}\pm c\sqrt{\frac{\ell'\hat{V}\ell}{n}}\right],\qquad \hat{\mathrm{V}}=\mathrm{Var_n}\left(\hat{\phi}(Z)\right),\qquad \hat{\phi}(Z)\coloneqq-\hat{J}^{-1}m\left(Z;\hat{\theta},\hat{g}\right),\qquad \hat{J}=E_n[a(Z;\hat{g})]$$

Main Theorem (linear moments)

- If moments are linear
$$m(Z;\beta,g,p) = Y_{DR}(g,p)\psi(X) - \psi(X)\psi(X)'\theta$$
- Estimate is closed form:
$$\hat{\theta} = \hat{J}^{-1}E_n[Y_{DR}(g,p)\psi(X)], \qquad \hat{J} = E_n[\psi(X)\psi(X)']$$
- Then the estimate $\hat{\beta}$ is *asymptotically linear*
$$\sqrt{n}\big(\hat{\beta} - \beta_0\big) \approx \sqrt{n}\, E_n[\phi_0(Z)], \qquad \phi_0(Z) = -J_0^{-1}\, m(Z;\beta_0,g_0,p_0), \qquad J_0 \coloneqq E[\psi(X)\psi(X)']$$
- Consequently, it is *asymptotically normal*
$$\sqrt{n}\,\big(\hat{\beta} - \beta_0\big) \sim_a N(0,V), \qquad V \coloneqq E[\phi_0(Z)\phi_0(Z)']$$
- *Confidence intervals* for any projection based on estimate of variance are asymptotically valid
$$x'\beta \in \left[x'\hat{\beta} \pm c\sqrt{\frac{x'\hat{V}x}{n}}\right], \qquad \hat{\mathrm{V}} = \mathrm{Var_n}\left(\hat{\phi}(Z)\right), \qquad \hat{\phi}(Z) \coloneqq -\hat{J}^{-1}m\big(Z;\hat{\theta},\hat{g}\big), \qquad \hat{J} = E_n[\psi(X)\psi(X)']$$

Confidence Bands

- Since $\hat{\beta}$ are asymptotically linear, predictions are asymptotically linear
- Then the estimate $\hat{\beta}$ is *asymptotically linear*

$$\sqrt{n}\big(\hat{\theta}_{BLP}(x) - \theta_{BLP}(x)\big) = \sqrt{n}\big(x'\hat{\beta} - x'\beta_0\big) \approx \sqrt{n}\, E_n[x'\phi_0(Z)]$$

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

$$\max_{x \in X} \left| \sqrt{n}\left(\hat{\theta}_{BLP}(x) - \theta_{BLP}(x)\right) - \sqrt{n}\, E_n[x'\phi_0(Z)] \right| \approx 0$$

- High-dimensional CLT theorems also imply that jointly:

$$\left\{\sqrt{n}\left(\hat{\theta}_{BLP}(x) - \theta_{BLP}(x)\right)\right\}_{x \in X} \sim_a N(0, V), \qquad V_{x_1 x_2} = E[x_1'\phi_0(Z)\phi_0(Z)x_2]$$

Confidence Bands

- Similar 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

$$\Pr\big(\forall x:\ \theta(x) \in CI(x)\big) \to 1 - \alpha$$

Confidence Bands

- Confidence band: construct intervals

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

- 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 $D = \operatorname{diag}(V)$

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

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

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

For 95% confidence band, c slightly larger than 1.96

Computationally Friendlier Version: Multiplier Bootstrap

- By asymptotic linearity we know that:

$$\frac{\sqrt{n}\left(\theta(x) - \hat{\theta}(x)\right)}{\sqrt{\hat{V}_{xx}}} \approx \sqrt{n}\, E_n\left[\frac{x'\phi_0(Z)}{\sqrt{V_{xx}}}\right]$$

- For every sample $i = 1 \dots n$, draw an independent Gaussian $\epsilon_i \sim N(0,1)$ and consider the variable

$$Q(x; \epsilon_1, \dots, \epsilon_n) \coloneqq \sqrt{n}\, E_n\left[\frac{x'\phi_0(Z)}{\sqrt{V_{xx}}}\epsilon\right] = \frac{1}{\sqrt{n}}\sum_i \frac{x'\phi_0(Z)}{\sqrt{V_{xx}}}\epsilon_i$$

- The vector of random variables $\left(Q(x_1), \dots, Q\left(x_{|X|}\right)\right) \sim_a N\left(0, D^{-1/2} V D^{-1/2}\right)$
- Approximately the same holds for $\left(\hat{Q}(x_1), \dots, \hat{Q}\left(x_{|X|}\right)\right)$ with $\hat{Q}(x; \epsilon_1, \dots, \epsilon_n) = \frac{1}{\sqrt{n}} \sum_i \frac{x'\hat{\phi}(Z)}{\sqrt{\hat{V}_{xx}}}\epsilon_i$

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

$$Z^{(b)} \coloneqq \max_{x \in X}\left|\hat{Q}(x; \epsilon_1, \dots, \epsilon_n)\right|$$

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