CSC2541: Introduction to Causality

Lecture 7 - Double Machine Learning

Instructor: Rahul G. Krishnan
TA: Vahid Balazadeh-Meresht

November 21, 2022

Back to ML for Causality

- Last lecture: Use-case of the invariance assumption in causal inference for machine learning (ML)
- Today: How to use ML predictive models to get *unbiased* causal effect estimations with *fast convergence rate* and *confidence intervals*?
- TAR-Net also used ML for causal effect estimation. However:
  - Not flexible in using different ML models,
  - No convergence rate guarantees,
  - No uncertainty regions,
  - Only for binary treatments.
- We will assume ignorability, i.e., covariates X block all the backdoor paths from treatment T to outcome Y

Where can we use ML for causal estimation?

- ML methods are effective in prediction contexts, but this does not translate into good performance for estimation of ”causal” parameters
    1. Overfitting bias: Capturing more than the relationship of T and Y
    2. Regularization bias: Slower convergence rate
- Often, covariates X are high-dimensional while T is low-dimensional
- The relationship between Y and X is more complex than the relationship between Y and T
- Idea: Use ML methods to model $Y \sim X$ and linear models for $Y \sim T$

A canonical example - Partially Linear Model

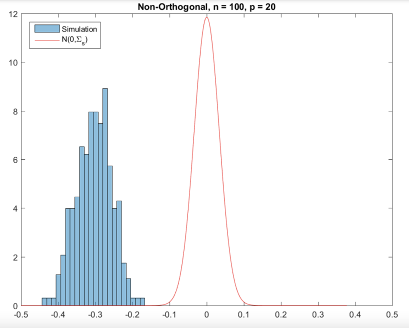
Assume the following data generating process:

$$\begin{aligned} &Y = \alpha_0 T + g_0(X) + U \\ &T = m_0(X) + V \\ &\text{with } \mathbb{E}\left[U|T,X\right] = 0,\ \mathbb{E}\left[V|X\right] = 0 \end{aligned}$$

- $Y, T \in \mathbb{R}$
- α_0 is the target parameter of interest (ATE)
- X is a high-dimensional vector
- We call $\eta_0 = (g_0, m_0)$ nuisance parameters - We do not care about their estimation as long as it results in correct α_0

Naive prediction-based ML approach is Bad

- Predict Y using X and T:
$$\hat{Y} = \hat{\alpha}_0 T + \hat{g}_0(X)$$
- For example, we can fit the model by alternating minimization
  - Given initial parameters, run a Random Forest on $Y - \hat{\alpha}_0 T$ to fit $\hat{g}_0(X)$
  - Run Ordinary Least Squares (OLS) on $Y - \hat{g}_0(X)$ to fit $\hat{\alpha}_0$
  - Repeat until convergence
- Good prediction performance $\|\hat{Y} - Y\|_2^2$. But, the distribution of $\alpha_0 - \hat{\alpha}_0$ looks like this



Why is the naive approach bad?

- Assume the minimization is converged and we learned $\hat{g}_0(X)$
- $\hat{\alpha}_0$ is the OLS solution to $Y = \alpha T + \hat{g}_0(X)$:

$$\hat{\alpha}_0 = (\frac{1}{n}\sum_i T_i^2)^{-1}\frac{1}{n}\sum_i T_i(Y_i - \hat{g}_0(X_i))$$

$$\text{assuming } \mathbb{E}[g_0(X)] = \mathbb{E}[m_0(X)] = 0$$

- Let's look at the error:

$$\begin{aligned}
\hat{\alpha}_0 &= (\frac{1}{n}\sum_i T_i^2)^{-1}\frac{1}{n}\sum_i T_i(Y_i - \hat{g}_0(X_i)) \\
&= (\frac{1}{n}\sum_i T_i^2)^{-1}\frac{1}{n}\sum_i T_i(\alpha_0 T_i + g_0(X_i) + U_i - \hat{g}_0(X_i)) \\
&= (\frac{1}{n}\sum_i T_i^2)^{-1}\left[(\frac{1}{n}\sum_i T_i^2)\alpha_0 + (\frac{1}{n}\sum_i T_i U_i) + \left(\frac{1}{n}\sum_i T_i(g_0(X_i) - \hat{g}_0(X_i)\right)\right] \\
&= \alpha_0 + (\frac{1}{n}\sum_i T_i^2)^{-1}\left[\frac{1}{n}\sum_i T_i U_i + \frac{1}{n}\sum_i T_i(g_0(X_i) - \hat{g}_0(X_i)\right] \\
&= \alpha_0 + \underbrace{(\frac{1}{n}\sum_i T_i^2)^{-1}}_{\mathbb{E}[T_i^2]^{-1}}\left[\frac{1}{n}\sum_i T_i U_i + \frac{1}{n}\sum_i (m_0(X_i) + V_i)(g_0(X_i) - \hat{g}_0(X_i)\right]
\end{aligned}$$

Why is the naive approach bad?

$$\sqrt{n}(\hat{\alpha}_0 - \alpha_0) = \frac{\left[\overbrace{\frac{1}{\sqrt{n}}\sum_i T_i U_i}^{A} + \overbrace{\frac{1}{\sqrt{n}}\sum_i m_0(X_i)\left(g_0(X_i) - \hat{g}_0(X_i)\right)}^{B} + \overbrace{\frac{1}{\sqrt{n}}\sum_i V_i\left(g_0(X_i) - \hat{g}_0(X_i)\right)}^{C}\right]}{\mathbb{E}\left[T_i^2\right]}$$

The goal is to find a root-n consistent and asymptotically normal estimate of α_0, i.e., $\sqrt{n}(\hat{\alpha}_0 - \alpha_0) \to \mathcal{N}(0, \sigma^2)$

- $A \to \mathcal{N}(0, \sigma_A^2)$ by Central Limit Theorem. It can be seen as sample average of random variables $T_i U_i$
- What about term B? Does $B \to \mathcal{N}(0, \sigma_B^2)$ for some σ_B^2?

Regularization Bias - Term B

- Machine learning methods employ regularization (e.g., L^2 regularization) to reduce variance. However, this often induces bias and lower convergence rate:

  $$g_0(X_i) - \hat{g}_0(X_i) \propto n^{-\phi_g}, \quad \text{for some } \phi_g < \frac{1}{2} \qquad \text{(slow convergence)}$$

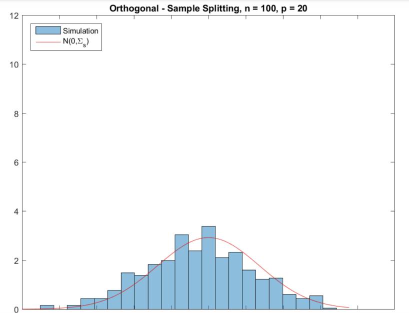
  Therefore, term B will be

  $$B = \frac{1}{\sqrt{n}} \sum_i m_0(X_i) \left(g_0(X_i) - \hat{g}_0(X_i)\right) \propto \frac{1}{\sqrt{n}} \cdot n \cdot n^{-\phi_g} \propto n^{\frac{1}{2} - \phi_g} \to \infty$$

- How to make this term vanish?

Double Machine Learning

- The naive approach was the OLS solution to $Y = \alpha T + \hat{g}_0(X)$
- Idea: Partial out the effect of covariate X on treatment T
  - Train an ML algorithm to predict T from X: $\hat{T} = \hat{m}_0(X)$
  - Consider the residual $\hat{V} = T - \hat{m}_0(X)$
  - Find the OLS solution $\hat{\beta}$ to $Y = \beta\hat{V} + \hat{g}_0(X)$
- This approach is called Double Machine Learning (DML) as we use machine learning twice: to learn $\hat{g}_0(X)$ and to learn $\hat{m}_0(X)$
- $\hat{\beta}$ is a root-n consistent estimate of α_0. $(\alpha_0 - \hat{\beta})$ looks like this



Partialling out the effect of covariates. Frisch–Waugh–Lovell theorem

- But why does partialling out the effect of X on T results in a valid estimate?
- Let's make everything linear. Consider the following linear equation:

$$Y = T\beta_1 + X\beta_2$$

  for $T, Y, \beta_1 \in \mathbb{R}$ and $\beta_2, X \in \mathbb{R}^d$. Assume $\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{X}$ are data matrices

- To estimate β_1, one can use OLS by concatenating $\boldsymbol{T}$ and $\boldsymbol{X}$
- Frisch–Waugh–Lovell (FWL) theorem says we can estimate β_1 in another way. Residuals-on-residuals:
  - Regress (linear) $\boldsymbol{Y}$ on $\boldsymbol{X}$ and let $\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$
  - Regress (linear) $\boldsymbol{T}$ on $\boldsymbol{X}$ and let $\hat{\boldsymbol{V}} = \boldsymbol{T} - \hat{\boldsymbol{T}}$
  - Regress (linear) $\hat{\boldsymbol{U}}$ on $\hat{\boldsymbol{V}}$ to estimate β_1
- FWL is a simpler version of DML. Instead of arbitrary ML methods, it uses linear regression

FWL theorem - Proof

- Define the prediction matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$
  - E.g., the OLS solution for $Y \sim X$: $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}$
- Define the residual matrix $\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{P}$
  - E.g., $\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{Y}$
- Note that residuals are **orthogonal** to predicted values

$$\boldsymbol{R}\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P} = \boldsymbol{P} - \boldsymbol{P}^2 = \boldsymbol{0}$$

- Let's apply the residual matrix on $\boldsymbol{Y} = \boldsymbol{T}\beta_1 + \boldsymbol{X}\beta_2$:

$$\boldsymbol{R}\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{T}\beta_1 + \boldsymbol{R}\boldsymbol{X}\beta_2$$

- However,

$$\boldsymbol{R}\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{0}$$

- Therefore,

$$\begin{aligned} \boldsymbol{R}\boldsymbol{Y} &= \boldsymbol{R}\boldsymbol{T}\beta_1 \\ \boldsymbol{Y} - \hat{\boldsymbol{Y}} &= (\boldsymbol{T} - \hat{\boldsymbol{T}})\beta_1 \end{aligned}$$

Back to DML - Overcoming regularization bias

- FWL shows that partialling out X does not affect the relationship between Y and T. It essentially gives the **same** answer
- But why does the estimation from DML ($\hat{\beta}$) converges **better** than the naive solution $\hat{\alpha}_0$?
- The key is the regularization bias (term B)

$$\sqrt{n}(\hat{\alpha}_0 - \alpha_0) = \frac{\left[\overbrace{\frac{1}{\sqrt{n}}\sum_i T_i U_i}^{A} + \overbrace{\frac{1}{\sqrt{n}}\sum_i m_0(X_i)\left(g_0(X_i) - \hat{g}_0(X_i)\right)}^{B} + \overbrace{\frac{1}{\sqrt{n}}\sum_i V_i\left(g_0(X_i) - \hat{g}_0(X_i)\right)}^{C}\right]}{\mathbb{E}\left[T_i^2\right]}$$

- Let's write a similar estimation error for the DML solution $\hat{\beta}$

Why is the DML approach good?

- $\hat{\beta}_0$ is the OLS solution to $Y = \beta\hat{V} + \hat{g}_0(X)$, where $\hat{V} = T - \hat{m}_0(X)$

$$\hat{\beta}_0 = \left(\frac{1}{n}\sum_i \hat{V}_i^2\right)^{-1} \frac{1}{n}\sum_i \hat{V}_i(Y_i - \hat{g}_0(X_i))$$

- For a simpler analysis, we consider a slightly different estimator

$$\hat{\beta} = \left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1} \frac{1}{n}\sum_i \hat{V}_i(Y_i - \hat{g}_0(X_i))$$

- In finite samples, $\hat{\beta} \neq \hat{\beta}_0$. However, they both will have similar asymptotic properties as $\mathbb{E}[\hat{V}^2] = \mathbb{E}[\hat{V}T]$ for infinite samples

Why is the DML approach good?

Let's look at the error:

$$\hat{\beta} = \left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1} \frac{1}{n}\sum_i \hat{V}_i(Y_i - \hat{g}_0(X_i))$$

$$= \left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1} \frac{1}{n}\sum_i \hat{V}_i(\alpha_0 T_i + g_0(X_i) + U_i - \hat{g}_0(X_i))$$

$$= \alpha_0 + \left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1} \left[(\frac{1}{n}\sum_i \hat{V}_i U_i) + \left(\frac{1}{n}\sum_i \hat{V}_i(g_0(X_i) - \hat{g}_0(X_i)\right)\right]$$

$$= \alpha_0 + \left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1} \left[(\frac{1}{n}\sum_i \hat{V}_i U_i) + \left(\frac{1}{n}\sum_i (T_i - \hat{m}_0(X_i))(g_0(X_i) - \hat{g}_0(X_i)\right)\right]$$

$$= \alpha_0 + \frac{\left[(\frac{1}{n}\sum_i \hat{V}_i U_i) + (\frac{1}{n}\sum_i (m_0(X_i) + V_i - \hat{m}_0(X_i))(g_0(X_i) - \hat{g}_0(X_i))\right]}{\left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1}}$$

Why is the DML approach good?

- Therefore,

$$\sqrt{n}(\hat{\beta} - \alpha_0) = \frac{\left[\overbrace{\frac{1}{\sqrt{n}}\sum_i \hat{V}_i U_i}^{A'} + \overbrace{\frac{1}{\sqrt{n}}\sum_i (m_0(X_i) - \hat{m}_0(X_i))\,(g_0(X_i) - \hat{g}_0(X_i))}^{B'} + \overbrace{\frac{1}{\sqrt{n}}\sum_i V_i\,(g_0(X_i) - \hat{g}_0(X_i))}^{C}\right]}{\left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1}}$$

- Compare it to

$$\sqrt{n}(\hat{\alpha}_0 - \alpha_0) = \frac{\left[\overbrace{\frac{1}{\sqrt{n}}\sum_i T_i U_i}^{A} + \overbrace{\frac{1}{\sqrt{n}}\sum_i m_0(X_i)\,(g_0(X_i) - \hat{g}_0(X_i))}^{B} + \overbrace{\frac{1}{\sqrt{n}}\sum_i V_i\,(g_0(X_i) - \hat{g}_0(X_i))}^{C}\right]}{\mathbb{E}\left[T_i^2\right]}$$

- A' behaves similarly to A. Term C is exactly the same. The difference is in the regularization terms B and B'

DML overcomes the regularization bias

$$B' = \frac{1}{\sqrt{n}} \sum_i (m_0(X_i) - \hat{m}_0(X_i)) \left(g_0(X_i) - \hat{g}_0(X_i) \right)$$

- Again, since we are using (regularized) ML methods, the convergence rates of g_0 and m_0 are slow

$$g_0(X_i) - \hat{g}_0(X_i) \propto n^{-\phi_g}, \quad \text{for some } \phi_g < \frac{1}{2}$$
$$m_0(X_i) - \hat{m}_0(X_i) \propto n^{-\phi_m}, \quad \text{for some } \phi_m < \frac{1}{2}$$

  Therefore, term B' will be

$$\begin{aligned} B' &= \frac{1}{\sqrt{n}} \sum_i (m_0(X_i) - \hat{m}_0(X_i)) \left(g_0(X_i) - \hat{g}_0(X_i) \right) \\ &\propto \frac{1}{\sqrt{n}} \cdot n \cdot n^{-\phi_m} \cdot n^{-\phi_g} \propto n^{\frac{1}{2} - \phi_g - \phi_m} \end{aligned}$$

- Now, even for slow convergence rates like $\phi_g, \phi_m = \frac{1}{4} + \epsilon$, B' will converge with root-n rate

$$n^{\frac{1}{2} - \phi_g - \phi_m} = n^{\frac{1}{2} - \frac{1}{4} - \epsilon - \frac{1}{4} - \epsilon} = n^{-2\epsilon} \to 0$$

Overfitting bias - Term C

$$\sqrt{n}(\hat{\beta} - \alpha_0) =$$

$$\frac{\left[\overbrace{\frac{1}{\sqrt{n}}\sum_i \hat{V}_i U_i}^{A'} + \overbrace{\frac{1}{\sqrt{n}}\sum_i (m_0(X_i) - \hat{m}_0(X_i))\,(g_0(X_i) - \hat{g}_0(X_i))}^{B'} + \overbrace{\frac{1}{\sqrt{n}}\sum_i V_i\,(g_0(X_i) - \hat{g}_0(X_i))}^{C}\right]}{\left(\frac{1}{n}\sum_i \hat{V}_i T_i\right)^{-1}}$$

- We saw $A' \to \mathcal{N}(0, \sigma_A^2)$
- DML used orthogonalization to overcome regularization bias B': $B' \to \mathcal{N}(0, \sigma_B^2)$
- What about term C? Does it also vanish?

Overfitting bias - Term C

- To learn $\hat{g}_0(X)$, we fitted an ML method to predict Y from X
- For example, we can (artificially) assume the estimator is as follows

$$\hat{g}_0(X_i) = g_0(X_i) + \underbrace{\frac{(Y_i - g_0(X_i))}{n^{1/2-\epsilon}}}_{\text{error}} \qquad \text{(fast but not root-n rate)}$$

- The error term is the part of Y that is unexplainable by $g_0(X)$
- Let's look at term C:

$$\begin{aligned}
C &= \frac{1}{\sqrt{n}} \sum_i V_i \left(g_0(X_i) - \hat{g}_0(X_i)\right) \\
&= \frac{1}{\sqrt{n}} \sum_i V_i \frac{(Y_i - g_0(X_i))}{n^{1/2-\epsilon}} \\
&= \frac{1}{\sqrt{n}} \sum_i V_i \frac{(g_0(X_i) + T_i + U_i - g_0(X_i))}{n^{1/2-\epsilon}} \\
&= \frac{1}{\sqrt{n}} \sum_i V_i \frac{(T_i + U_i)}{n^{1/2-\epsilon}} \\
&= \frac{1}{\sqrt{n}} \sum_i V_i \frac{(m_0(X_i) + V_i + U_i)}{n^{1/2-\epsilon}} = \frac{1}{\sqrt{n}} \sum_i \frac{V_i^2}{n^{1/2-\epsilon}} + \ldots = \frac{n}{\sqrt{n} n^{1/2-\epsilon}} \sum_i \frac{V_i^2}{n} + \ldots \\
&= n^{\epsilon} \mathrm{Var}(V) + \ldots \\
&\to \infty
\end{aligned}$$

Removing the overfitting bias with Sample Splitting

- Term C explodes since the estimated $\hat{g}_0(X)$ is overfitted: It captures more than $g_0(X)$ from Y and becomes related to noise V
- To overcome this, DML uses sample splitting
  - Use part of samples ($I \subset \{1, 2, \ldots, n\}$) to estimate $\hat{\beta}$
  - Use auxiliary samples (I^c) to estimate $\hat{g}_0(X)$
- Therefore, term C will be

$$C = \frac{1}{\sqrt{n}} \sum_{i \in I} V_i(g_0(X_i) - \hat{g}_0(X_i))$$

- This new C will vanish. Let's look at it's expectation

$$\begin{aligned}
\mathbb{E}[C] &= \frac{1}{\sqrt{n}} \sum_{i \in I} \mathbb{E}\left[V_i(g_0(X_i) - \hat{g}_0(X_i))\right] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in I} \mathbb{E}[\mathbb{E}[V_i \underbrace{(g_0(X_i) - \hat{g}_0(X_i))}_{Err_i} | X_{I^c}]] && \text{(condition on auxiliary samples)} \\
&= \frac{1}{\sqrt{n}} \sum_{i \in I} \mathbb{E}[\mathbb{E}[V_i]\mathbb{E}[Err_i | X_{I^c}]] && (Err_i \text{ only depends on auxiliary samples}) \\
&= 0 && (\mathbb{E}[V_i] = 0)
\end{aligned}$$

DML Algorithm - Summary

In summary, for a given dataset $\{T^i, X^i, Y^i\}_{i=1}^n$, DML follows the following to estimate average treatment effect:

1. Split samples to two parts I and I^c s.t. $I \cup I^c = \{1, \dots, n\}$ and $I \cap I^c = \emptyset$
2. Train any (regularized) machine learning model M_t to predict T from X using auxiliary I^c
3. Train any (regularized) machine learning model M_y to predict Y from X using I^c
4. Obtain the residuals $Y_R = Y - M_y(X)$ and $T_R = T - M_t(X)$ from samples I
5. Regress (linearly) Y_R on T_R to get the estimated ATE

To increase sample efficiency, we can get another estimate by changing the role of I and I^c and take the average of the two estimations

DML properties

- It allows using any a broad range of ML or non-parametric algorithms to estimate high-dimensional nuisance parameters ($\eta_0 = (g_0, m_0)$)
- It gives a root-n consistent estimator for ATE - Fast convergence
- We can get valid confidence intervals over ATE as the estimate is asymptotically normal
- DML was published in 2016[1] and is still among the best methods in causal inference competitions[2]

[1] Chernozhukov et al., 2016.
[2] ACIC 2022 data challenge - `https://acic2022.mathematica.org/results`

Application outside of causal estimation - Identifying causal parents

- Now that we know what double machine learning actually is, how can we use it to solve practical problems?
- **Question:** How do we identify the causal parents of a variable?
- Given genetic expression data, we might want to know which are the causal parents while controlling for the effect of other genes.
- (Raj et al., 2020) use DML as a black box and devise a parallel search strategy to treat each gene as a treatment and predict the outcome (disease incidence).
- **Discuss:** What might the pros/cons of this approach be?

Recap of Introduction to Causality

- Correlation is not causation!
- No causal inference without assumptions – positivity, no unobserved confounding.
- Potential outcomes, Causal Bayesian networks, Structural causal models.
- Identifying interventions, evaluating counterfactuals and do-calculus.
- Estimation methods: G-formula, Matching, Inverse propensity weighting.
- Handling unobserved confounding - instrumental variables and local average treatment effects.
- Causal inference for ML: learning from environments.
- ML for causal inference: double machine learning for estimating treatment effects.

What we did not cover

- Sensitivity analysis - understanding how much unobserved confounding one needs to change the outcomes of your study.
- Dynamic treatment effects - causal effects with time-varying data.
- Partial identification - bounding causal effects rather than point identification.
- Causal decision making - what (among) many interventions should I make?
- Applications of causal inference to improve RL, control, planning, predictive modeling in healthcare.
- Causal representation learning - ????

General advice

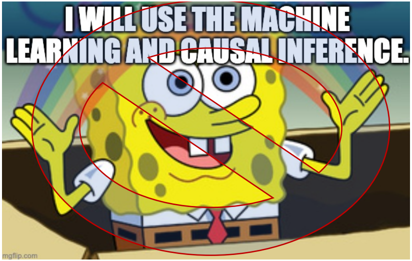


Figure: Be critical of the methods you use!

- The easiest person to fool is yourself – always question your assumptions!
- Work closely with domain experts – common sense and practical wisdom >>> any result from any algorithm.
- Always ask "where do the bits come from"?

Chernozhukov, Victor et al. (2016). “Double/debiased machine learning for treatment and causal parameters”. In: *arXiv preprint arXiv:1608.00060.*

Raj, Anant et al. (2020). “Causal feature selection via orthogonal search”. In: *arXiv preprint arXiv:2007.02938.*