HOMEWORK 4
CAUSAL MODELING AND INFERENCE[1]

10-708 PROBABILISTIC GRAPHICAL MODELS (SPRING 2022)
https://andrejristeski.github.io/10708-22/

OUT: March 28, 2022
DUE: April 11, 2022 at 11:59 PM
TAs: Aaron, Yuchen

- **Collaboration policy:** The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes (including code) are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved. See the Academic Integrity Section on the course site for more information: https://andrejristeski.github.io/10708-22/#:~:text=Academic%20Integrity%20Policies
- **Late Submission Policy:** See the late submission policy here: https://andrejristeski.github.io/10708-22/#:~:text=Grace%20Day/Late%20Homework%20Policy
- **Submitting your work to Gradescope:** We use Gradescope (https://www.gradescope.com/courses/349316/assignments) to collect PDF submissions of open-ended questions on the homework (e.g. mathematical derivations, plots, short answers). The course staff will manually grade your submission, and you'll receive personalized feedback explaining your final marks. The homework template must be used and can be completed in Latex or by hand. Handwritten submissions must be legible otherwise we will not be able to give credit to your solutions. No changes should be made to the template, boxes and choices **MUST** remain the same size and in the same locations between the template and your completed submission, the document has 21 pages so your submission must contain no more and no less than 21 pages.
- **Programming Code:** You will also submit your code for the programming questions on the homework to Gradescope, specifically the 'Homework 4 Code' submission slot. All code written must be submitted in order for you to get any credit for the written components of the programming section.
- For **multiple choice** or **select all that apply questions**, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeX users, replace `\choice` with `\CorrectChoice` to obtain a shaded box/circle, and don't change anything else.

[1]Compiled on Tuesday 29th March, 2022 at 01:08

A Written Questions [100 pts]

Answer the following questions in the template provided. Then upload your solutions to Gradescope. You may use LaTeX or print the template and hand-write your answers then scan it in. Failure to use the template may result in a penalty. There are 100 points and 15 questions.

A.1 Multiple-choice and Short-answer Questions

1. Answer the following questions:

   (a) (1 point) The *Fundamental Problem of Causal Inference* states that because we cannot observe all potential outcomes for every unit, the average treatment effect can never be estimated using observational data.

   ◯ True

   ◯ False

   (b) (1 point) Consider a binary treatment variable X and an outcome of interest Y. Which of the following conditions guarantee that the Average Causal Effect (ACE) of X on Y (i.e., $\mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$) is equal to the associational difference ($\mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$)?

   ◯ Identifiability

   ◯ Exchangeability

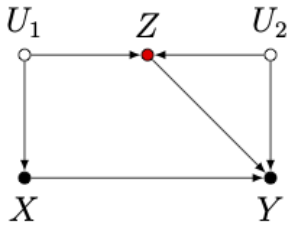
   ◯ Positivity

   ◯ No interference

   (c) (1 point) Consider a binary treatment variable T, observable covariates X, and and an outcome of interest Y. Provide a graphical counterexample to the claim that "*conditioning on observable pre-treatment variables always helps with identifying the ACE of T on Y*". Briefly explain your example.

(d) (1 point) Consider a binary treatment variable T, observable covariates X, and and an outcome of interest Y. Provide a graphical counterexample to the claim that "*conditioning on post-treatment variables always hinders identifying the ACE of T on Y*".

(e) (4 points) Consider a binary treatment variable T, observable covariates X, and and an outcome of interest Y. Let U_1, U_2 denote the unobserved exogenous variables impacting X, T, and Y as follows:



Which of the following sets is a sufficient adjustment set for identifying the ACE of X on Y? Provide a brief justification.

- ◯ $\emptyset$
- ◯ $\{Z\}$
- ◯ $\{U_2\}$
- ◯ Neither of the above options

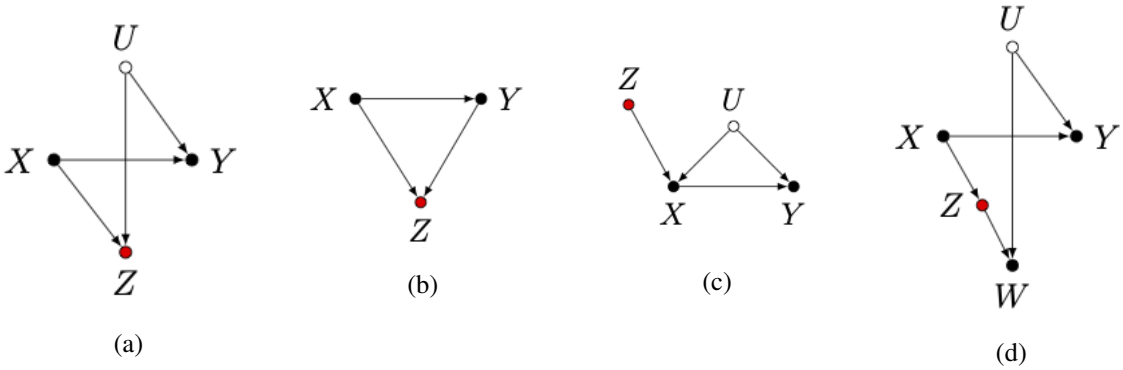


Figure A.1

(f) (2 points) Consider the causal structures in Figure A.1. In which case conditioning on Z de-identifies the ACE of X on Y? Briefly justify your answer.

- ◯ Model (a)
- ◯ Model (b)
- ◯ Model (c)
- ◯ Model (d)

A.2 Causal Modeling: The Monty Hall problem

The classic Monty Hall problem is stated as follows: Suppose you are on a game show, and you are given the choice of three doors: Behind one door is a 10K cash prize; behind the others, goats. You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, behind which is a goat. He then gives you the choice to switch your pick. Is it to your advantage to switch to door No. 2?

Let's generalize the Monty Hall problem to a setting in which there are $n \geq 3$ doors, w of them with the cash prize and $n - w$ with a goat behind them. (Assume that $1 < w < n$). You choose one door and the host reveals ℓ doors with goats behind them (where $1 < \ell < n - w$). (Note that the host deliberately chooses to reveal the doors with goats behind them) You will then be given the option to switch to an unopened door. The question is whether it is to your advantage to switch.

This problem asks you to causally model the above scenario and use it to reason about your best course of action (that is, to switch or not switch). Please show your work, but make sure your solution are clear and concise.

2. (5 points) Describe the structural causal model corresponding to the generalized Monty Hall problem.

3. (2 points) Using your answer to the previous question, describe the joint distribution of all variables in your structural causal model.

4. (5 points) Describe the conditions on $n, w,$ and ℓ under which it is to your advantage to switch to an unopened door after the host's reveals ℓ doors with goats behind them.

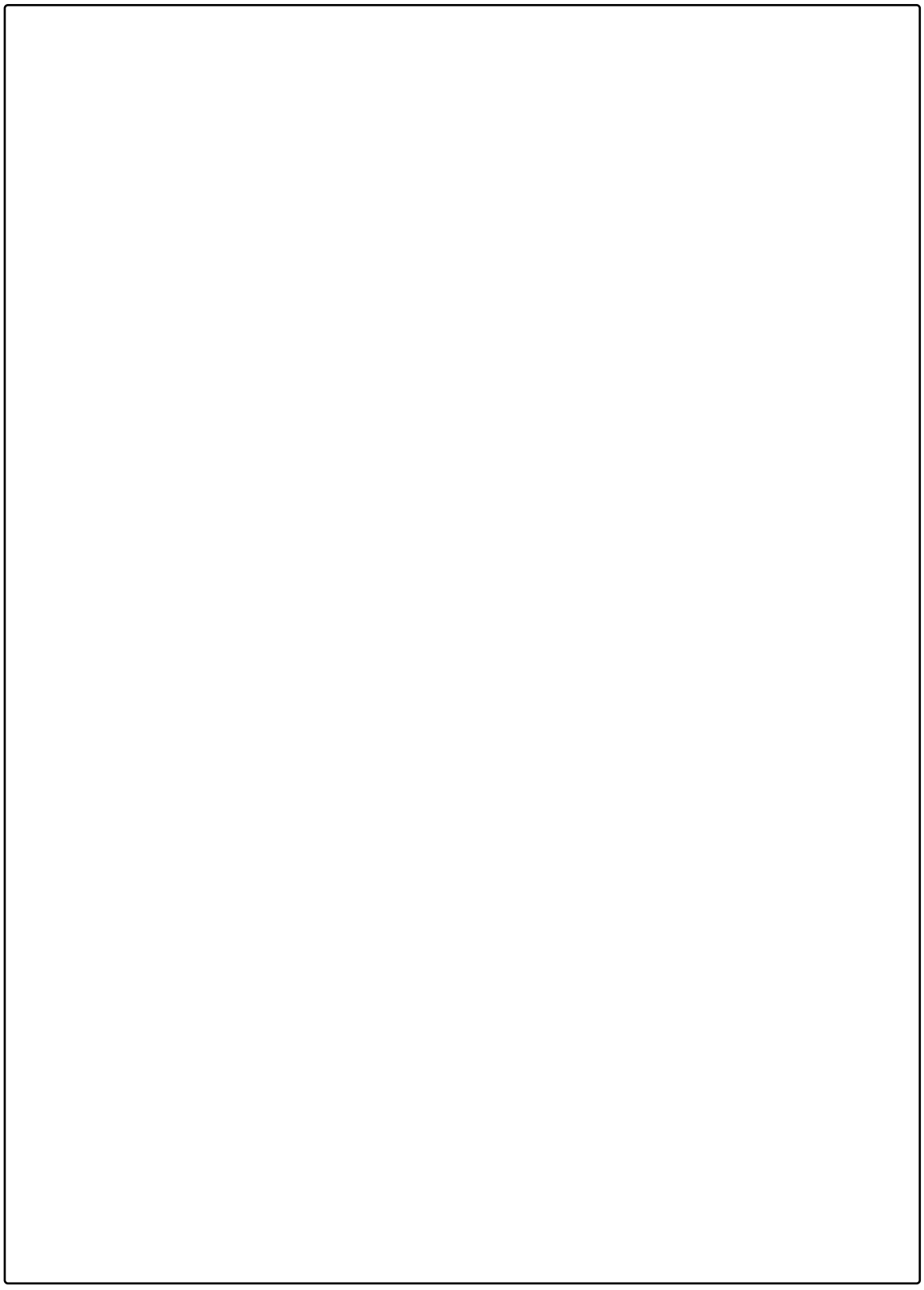
A.3 Backdoor and Front-door Adjustments via Do-Calculus

This question asks you to verify that the backdoor and front-door adjustment formulas are special cases of the do-calculus, and that do-calculus is strictly more powerful than these two adjustments formulas in identifying causal effects.

5. (2 points) Show that the Backdoor adjustment formula can be derived using the rules of do-calculus.

6. (10 points) Show that the frontdoor adjustment formula can be derived using the rules of do-calculus.

7. (6 points) Design a causal graph in which neither backdoor nor the front-door adjustment can be utilized to identify the causal effect of a treatment T on an outcome Y, but the rules of do-calculus can lead to successful identification.



A.4 Estimation of Causal Effects

Doubly robust estimation combines outcome modeling (e.g., regression) and exposure modeling (e.g., propensity scores) to estimate the causal effect of a treatment on an outcome, and it offers an unbiased estimator even if only one of the two models are correct (hence the name "doubly robust"). This question asks you to verify this advantage of doubly robust estimation in a simple example.

Consider a binary treatment T, a continuous outcome Y, and confounding covariates X. We are interested in estimating the average treatment effect of T on Y.

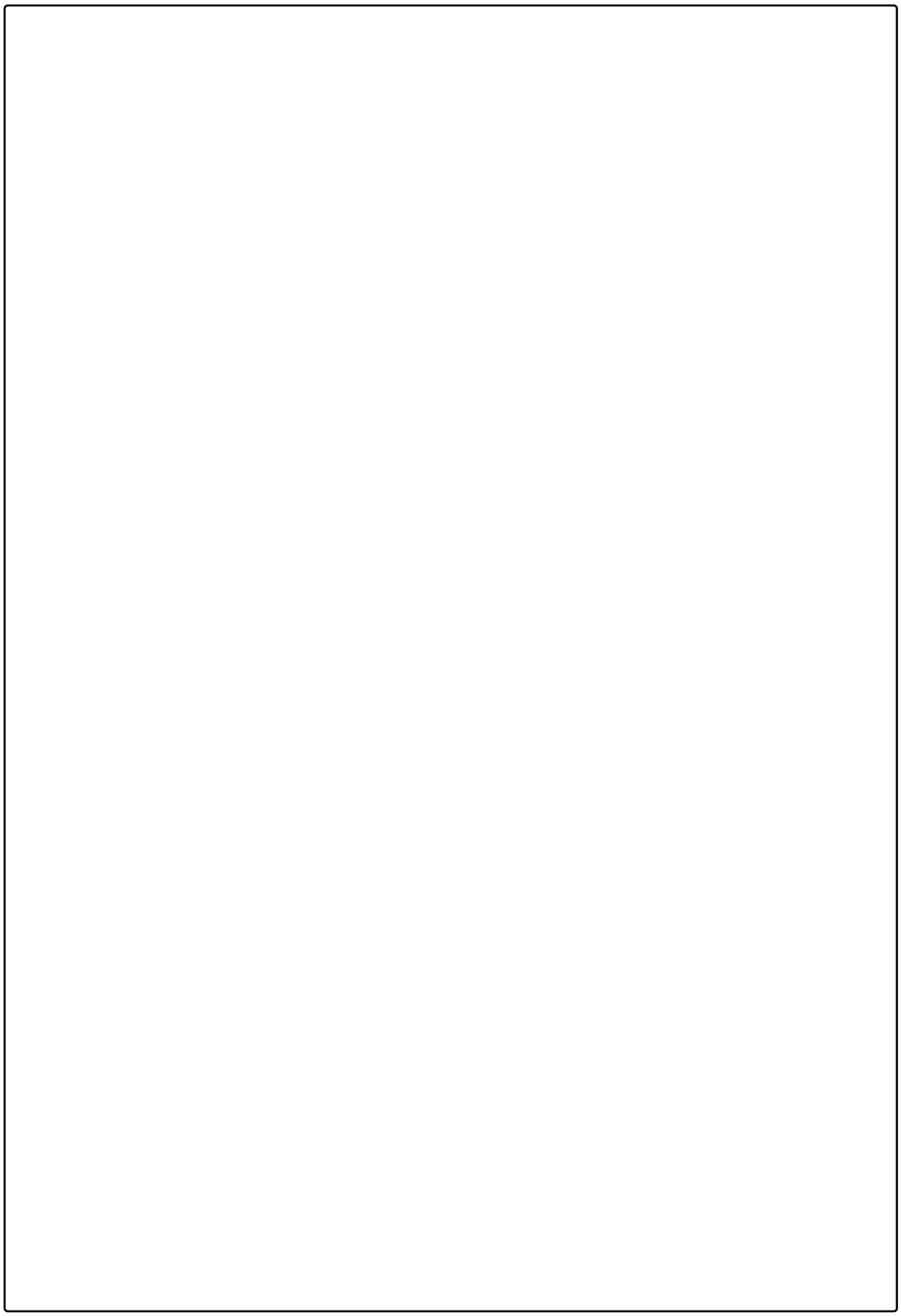
8. (2 points) Suppose we have access to an oracle $\mu(t, x) = \mathbb{E}[Y|T = t, X = x]$. Show that we can use this oracle to provide an unbiased estimator for $\mathbb{E}[Y^1 - Y^0]$.

9. (2 points) Suppose we have access to an oracle $e(x)$ that outputs the propensity score for a unit with covariates x. Show that we can use this oracle to provide an unbiased estimator for $\mathbb{E}[Y^1 - Y^0]$.

10. (6 points) Suppose $\hat{\mu}(t, x)$ and $\hat{e}(x)$ are approximations of the above oracles. Consider the following "doubly-robust" estimator.

$$\text{DRE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{I}[T_i = 1](y_i - \hat{\mu}(1, x_i))}{\hat{e}(x_i)} + \hat{\mu}(1, x_i) - \frac{\mathbb{I}[T_i = 0](y_i - \hat{\mu}(0, x_i))}{1 - \hat{e}(x_i)} - \hat{\mu}(0, x_i)\right)$$

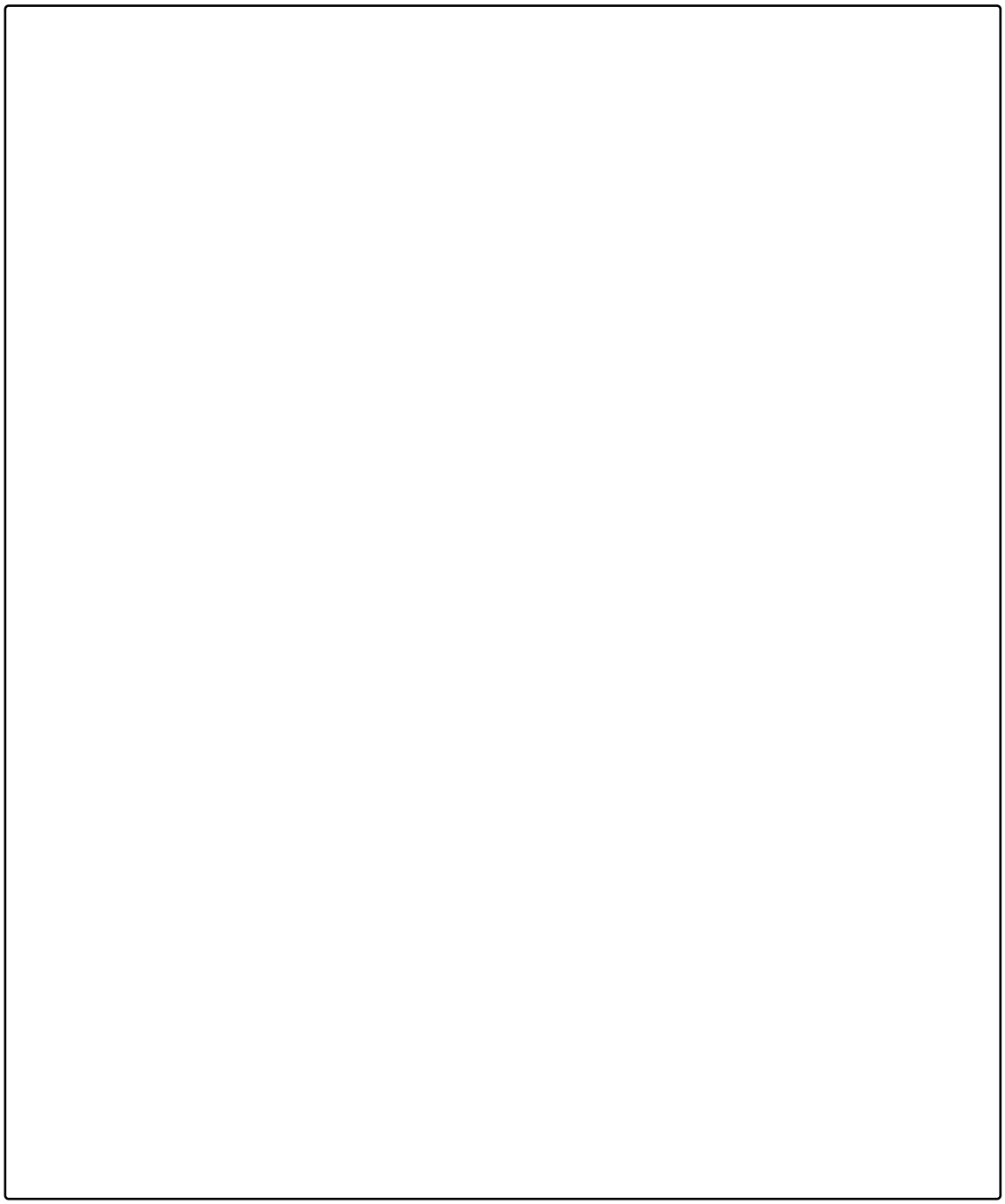
Show that the above estimator is unbiased if either $\hat{\mu}$ or $\hat{e}$ are unbiased.



A.5 Causal Discovery

This question asks you to prove several facts underlying the causal discovery algorithm we saw in the class.

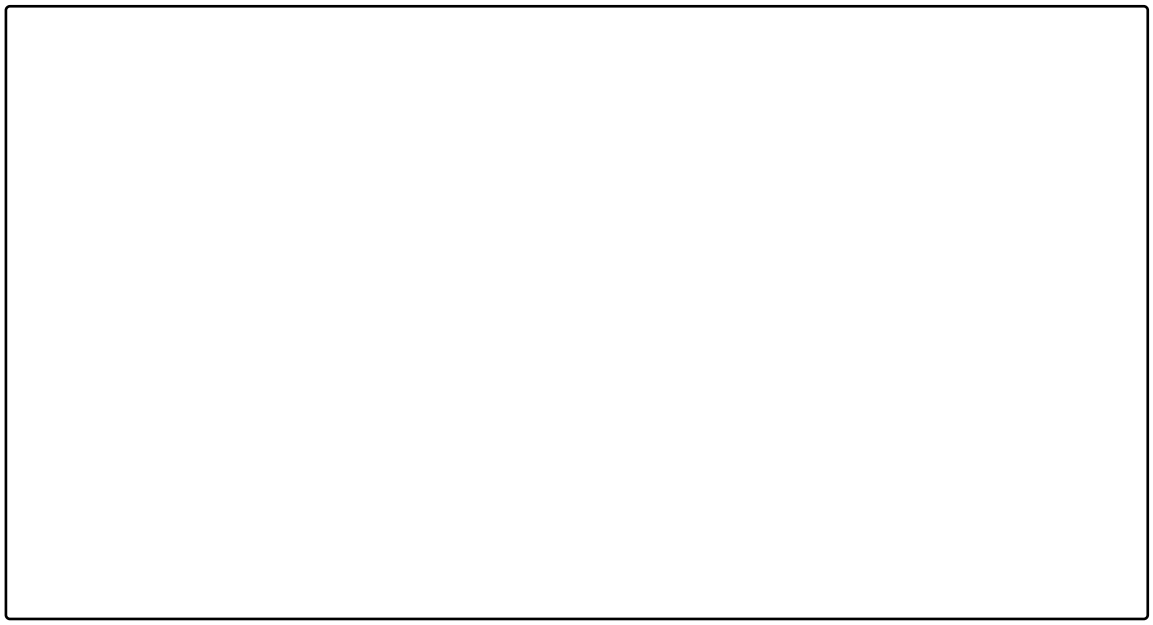
11. (5 points) Prove that, assuming faithfulness, a three-variable chain and a three-variable fork imply exactly the same set of dependence and independence relations, but that these are different from those implied by a three-variable collider.



12. (5 points) Prove that if X and Y are not parent and child, then either $X \perp\!\!\!\perp Y$, or there exists a set of variables S such that $X \perp\!\!\!\perp Y|S$.

13. (5 points) Prove that the graph produced by the edge-removal step of the PC algorithm is exactly the same as the graph produced by the edge-removal step of the SGS algorithm.

14. (5 points) Prove that if $X \perp\!\!\!\perp Y|S$ for some set of variables S, then $X \perp\!\!\!\perp Y|S'$, where every variable in S' is a neighbor of X or Y.



A.6 Transfer Learning and Causality

This question asks you to show that predicting an outcome variable Y using its causal parents Pa(Y) is optimal for ensuring robustness of predictions to covariate shift.

Consider a DAG G representing the causal associations among 10 variables $X_1, \cdots, X_{10}$ and an outcome variable Y. Suppose Pa(Y) $= \{X_1, X_2\}$ and the structural equation associated with Y is denoted by $f_Y^*(X_1, X_2) = \alpha_1 X_1 + \alpha_2 X_2 + \epsilon$ where ϵ is a mean-zero noise term independent of every other variable in the model (so $\mathbb{E}[Y|X_1, X_2] = \alpha_1 X_1 + \alpha_2 X_2$).

Suppose we are interested in predicting Y given the value of $X_1, \cdots, X_n$ and we would like our predictor to be robust to a certain form of covariate shift. Let $\mathcal{F}$ denote the family of all possible linear predictors (i.e., functions) mapping $X_1, \cdots, X_{10}$ to a predicted value for Y. Let $\mathcal{P}$ specify all joint distributions across $(X_1, \cdots, X_n, Y)$ that can be factorized according to G and are consistent with f_Y^*.

15. (10 points) Show that

$$f_Y^* \in \arg\min_{f \in \mathcal{F}} \max_{q \in \mathcal{P}} \mathbb{E}_{(X,Y)\sim q}[(Y - f(X))^2]$$

B Programming [20 pts]

In this assignment, you will implement conditional outcome modeling. Consider a binary treatment T, a continuous outcome Y, and confounding covariates X. We wish to estimate the effect of T on Y conditioned on X. This is called the conditional average treatment effect (CATE), mathematically defined as

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) | X = x\right].$$

B.1 Methods

You will implement three methods that perform this estimation. Each of the methods assumes access to a machine learning model that can estimate Y conditioned on X. In the real world, these ML models will be complex, but for this assignment we will use simple linear regression. The three methods are as follows:

Algorithm 1 S-Learner

procedure S-LEARNER(X, Y, T)
  $\hat{\mu} = M(Y \sim (X, T))$
  $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

Algorithm 2 T-Learner

procedure T-LEARNER(X, Y, T)
  $\hat{\mu}_0 = M(Y^0 \sim X^0)$
  $\hat{\mu}_1 = M(Y^1 \sim X^1)$
  $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

Algorithm 3 X-Learner

procedure X-LEARNER(X, Y, T, g)
  $\hat{\mu}_0 = M(Y^0 \sim X^0)$
  $\hat{\mu}_1 = M(Y^1 \sim X^1)$

  $\tilde{D}_i^1 = Y_i^1 - \hat{\mu}_0(X_i^1)$
  $\tilde{D}_i^0 = \hat{\mu}_1(X_i^0) - Y_i^0$

  $\hat{\tau}_1 = M(\tilde{D}^1 \sim X^1)$
  $\hat{\tau}_0 = M(\tilde{D}^0 \sim X^0)$
  $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$

For all of these methods, we take M to be simple linear regression. For the X-Learner, we set $g(x) = P(T = 1 | X = x)$.

Implement these three methods.

B.2 Simulations

You will now run these methods on simulated data to compare their performance.

For each of the simulations, we generate a 20-dimensional vector $X_i \sim \mathcal{N}(0, \Sigma)$ where each X_i is iid and Σ is a random covariance matrix.

We then generate the outcomes

$$Y_i(1) = \mu_1(X_i) + \epsilon_i(1),$$

$$Y_i(0) = \mu_0(X_i) + \epsilon_i(0),$$

where $\epsilon_i(1), \epsilon_i(0) \sim \mathcal{N}(0, 1)$ and the errors are iid.

The treatment is simulated by $T_i \sim \text{Bern}(e(X_i))$, and once assigning the treatment, we set $Y_i = Y(T_i)$. This yields the tuple (X_i, Y_i, T_i). Let $e(x)$ denote the propensity score, i.e. the proportion of units that receive treatment.

In each simulation, μ_0, μ_1, and e vary. For each simulation, a single experiment consists of training the treatment estimators on a training set of size $\{10^3, 10^4, 10^5, 10^6\}$ and tested against a set of size 10^5. Each experiment will be repeated 30 times. For each experiment, calculate the MSE of $\hat{\tau}(X_i) - \tau(X_i)$. Plot the MSE as a function of the size of the training set, one line for each of the three CATE estimators.

(a) (5 points) Simulation 1.

$$e(x) = 0.01$$

$$\mu_0(x) = x^T\beta, \quad \beta \sim \text{Unif}([-5, 5]^{20})$$

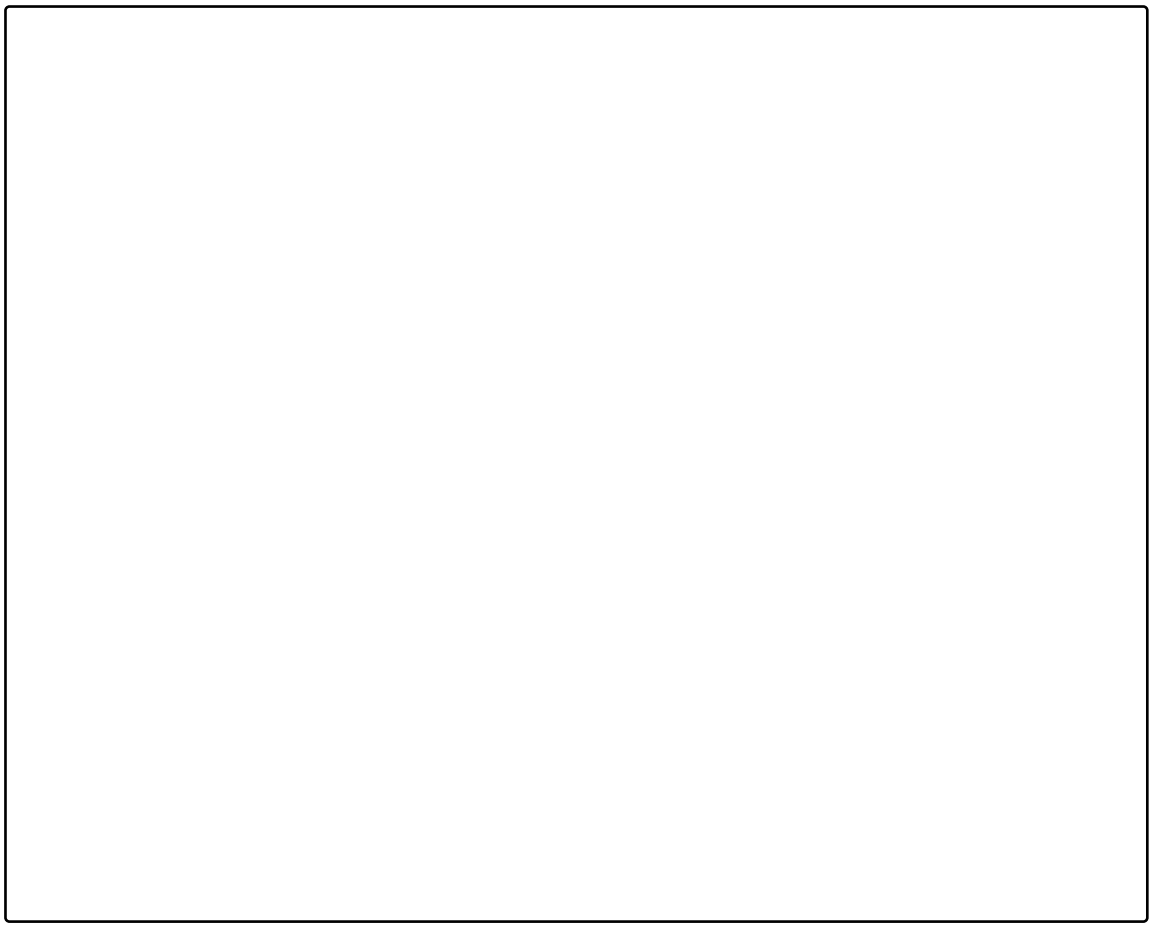
$$\mu_1(x) = \mu_0(x) + 8x_0$$

Plot the CATE MSE below. Comment on the differences in performance and explain why.

(b) (5 points) Simulation 2.

$$e(x) = 0.5$$
$$\mu_0(x) = x^T\beta_0, \;\; \beta_0 \sim \text{Unif}([-5,5]^{20})$$
$$\mu_1(x) = x^T\beta_1, \;\; \beta_1 \sim \text{Unif}([-5,5]^{20})$$

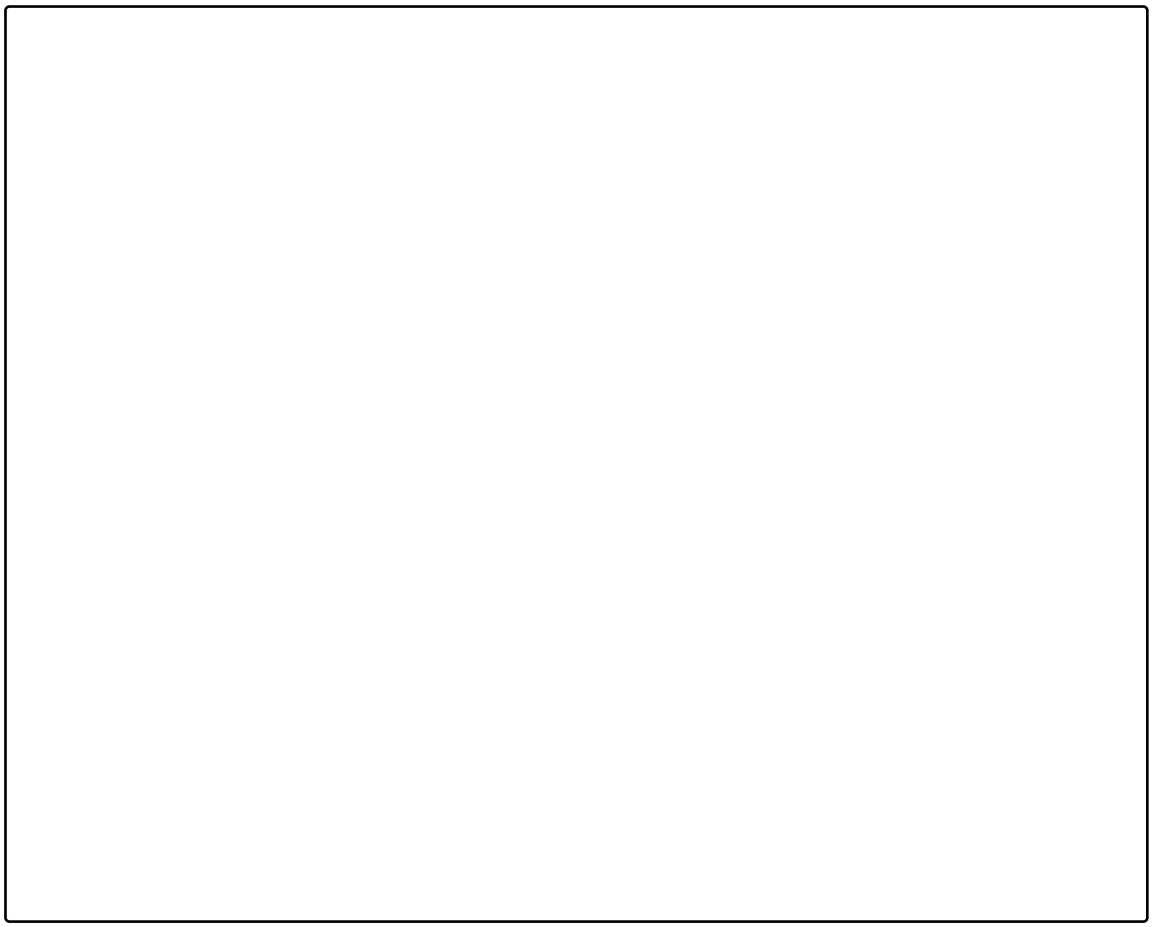
Plot the CATE MSE below. Comment on the differences in performance and explain why.



(c) (5 points) Simulation 3.

$$e(x) = 0.5$$
$$\mu_0(x) = x^T \beta_0, \;\; \beta_0 \sim \text{Unif}([-5, 5]^{20})$$
$$\mu_1(x) = \mu_0(x)$$

Plot the CATE MSE below. Comment on the differences in performance and explain why.



(d) (5 points) Simulation 4.

$$e(x) = 0.5 + 0.25 \cdot \text{sgn}(x_0)$$
$$\mu_0(x) = 2x_0 + 5$$
$$\mu_1(x) = \mu_0(x)$$

Plot the CATE MSE below. Comment on the differences in performance and explain why.

C Collaboration Policy

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies for this course.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.