Learning Treatment Effects from Observational and Experimental Data

Sofia Triantafillou
University of Crete

Fattaneh Jabbari
Microsoft

Gregory F Cooper
University of Pittsburgh

Abstract

Decision making often depends on causal effect estimation. For example, clinical decisions are often based on estimates of the probability of post-treatment outcomes. Experimental data from randomized controlled trials allow for unbiased estimation of these probabilities. However, such data are usually limited in the number of samples and the set of measured covariates. Observational data, such as electronic medical records, contain many more samples and a richer set of measured covariates, which can be used to estimate more personalized treatment effects; however, these estimates may be biased due to latent confounding. In this work, we propose a Bayesian method for combining observational and experimental data for unbiased conditional treatment effect estimation. Our method addresses the following question: Given observational data D_o measuring a set of covariates $\mathbf{V}$, and experimental data D_e measuring a possibly smaller set of covariates $\mathbf{V_b} \subseteq \mathbf{V}$, which set of covariates $\mathbf{Z}$ leads to the optimal, unbiased prediction of the post-intervention outcome $P(Y|do(X), \mathbf{Z})$, and when can we use observational data for this estimation? In simulated data, we show that our method improves the prediction of post-intervention outcomes..

1 INTRODUCTION

In decision making, we are often interested in finding the optimal predictive model for the post-intervention distribution of an outcome Y after we intervene on a variable X. Ideally, we would like to include in our model a set of pre-intervention covariates $\mathbf{Z}$ that allow us to predict the post-intervention outcomes $Y|do(X)$ as well as possible. An unbiased estimate for $P(Y|do(X), \mathbf{Z})$ can be obtained from experimental data D_e where the intervention X has been randomized, and covariates $\mathbf{Z}$ are measured. However, randomized trials are usually very limited in sample size and may not be powered to identify conditional distributions. Moreover, they may be missing important covariates that are helpful in predicting the post-intervention outcome. Observational data on the other hand are often plentiful in sample size and number of measured covariates, but may be biased for the estimation of post-intervention distributions due to confounding or other types of bias: Under causal insufficiency, $P(Y|do(X), \mathbf{Z})$ may not be identifiable for some or all covariate sets $\mathbf{Z}$. The condition for unbiased causal estimation of $P(Y|do(X), \mathbf{Z})$ from observational data is known as conditional ignorability. Ideally, we would like to use D_o for causal estimation if ignorability holds, and D_e when it does not hold. However, this condition is frequently untestable. In this work, we examine how we can combine large observational and limited experimental data to get the best of both worlds, when possible.

Our methods are heavily motivated by clinical settings, where we may be interested in identifying heterogeneous treatment effects for different subgroups of patients. Hence, we want to estimate $P(Y|do(X), \mathbf{Z})$, for an optimal set $\mathbf{Z}$. We want the set $\mathbf{Z}$ to be optimal in the sense that it includes all the necessary information required for optimal prediction of $Y|do(X)$, and at the same time it keeps the set as small as possible to reduce the variance of our estimator, $\hat{P}(Y|do(X), \mathbf{Z})$. In such cases, we often have access to two types of data.

1. Observational data $D_o = \{x_i^o, y_i^o, \mathbf{v}_i^o\}_{i=1}^{N_o}$, measuring a large set of pre-treatment covariates $\mathbf{V}$ in a large number N_o of patients.

2. Experimental data $D_e = \{x_i^e, y_i^e, \mathbf{v}_{\mathbf{b}i}^e\}_{i=1}^{N_e}$, measuring a possibly smaller set of covariates $\mathbf{V}_b \subseteq \mathbf{V}$ in a smaller number N_e of patients, where we assume that $N_e << N_o$.

Observational data are typically plentiful but we cannot use them to estimate a post-intervention distribution $P(Y|do(X), \mathbf{Z})$ unless the treatment assignment is independent of the outcome given $\mathbf{Z}$. Set $\mathbf{Z}$ is then said to be a

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

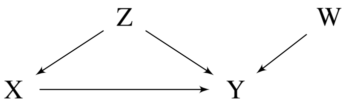

Figure 1: An example causal graph showing the causal structure among treatment X, outcome Y, and pre-treatment covariates Z and W.

valid adjustment set (Shpitser et al., 2010). This condition is typically untestable from observational data. Experimental data on the other hand may be missing important covariates, or may be lacking in sample size. Combining observational and experimental data has the potential to better predict the most effective treatment for each patient.

Our work addresses the following question: Given observational and experimental data as described above, can we identify a set of covariates $\mathbf{Z}$ that will lead to optimal prediction of $Y|do(X)$, according to some measure of predictive performance? To answer this question, we split the problem into two parts: In the first part, we determine if we can use the observational data in our estimation of $P(Y|do(X), \mathbf{Z})$, for different possible $\mathbf{Z}$'s, by determining if $\mathbf{Z}$ is an adjustment set. In the second part, we pick the optimal set $\mathbf{Z}$ by choosing the set that maximizes the expected performance. Notice that $\mathbf{Z}$ may include variables that are not included in D_e.

To illustrate our method, consider two scenarios related to the graph shown in Fig. 1. Let X be a treatment, Y an outcome, and variables Z, W be pre-treatment covariates. For the first scenario, assume that an RCT takes place, randomizing X and recording Y. Assume that the RCT measures W for the patients, but not Z. However, X, Y, W and Z are all observed in a large EHR data set from the same population. Our method can identify that $\{Z, W\}$ forms an adjustment set, and that the set is optimal for predicting $Y|do(X)$, for some proper criterion. $P(Y|do(X), Z, W)$ can then be estimated from the observational data, achieving an unbiased, low-variance estimator.

For the second scenario, assume that the neither the trial nor the observational data include Z. In that case, the optimal predictive model for $Y|do(X)$ includes the singleton $\{W\}$. However, $\{W\}$ is not an adjustment set, and thus, we cannot use the EHR data in the estimation of $P(Y|do(X), W)$. The proposed method can again identify that this is the case, and return an estimator based on the RCT data alone.

Compared to existing approaches for combining data for causal effect estimation, our contributions are:

1. We propose a method for deriving the probability that $\mathbf{Z}$ is an adjustment set, in which case $P(Y|do(X), \mathbf{Z})$ can be estimated from observational data. This allows us to take a fully Bayesian approach in estimating $P(Y|do(X), \mathbf{Z})$ given all the data.

2. Our proposed method works in cases where some of the variables are not measured in the experimental data, allowing it to add important personalization covariates to the conditional post-intervention distribution.

3. We allow for different criteria in selecting the optimal set for prediction of $Y|do(X)$. This is particularly important in clinical domains, where, for example, a patient's utility function needs to be taken into account.

The rest of this paper is as follows: In Section 2 we discuss preliminaries. Section 3 describes how we can decide which effects are identifiable from observational data (3.1), and how we select the optimal set (3.6). Section 4 discusses related literature, and in Section 5 we evaluate the performance of our method using simulated data.

2 PRELIMINARIES

Our method applies to the following setting: We assume that we have observational data D_o measuring variables $\mathbf{V}$ and experimental data D_e measuring variables $\mathbf{V_e} \subset \mathbf{V}$. We are interested in deriving the conditional probability distribution $P(Y|do(X), \mathbf{V}, D_e, D_o)$, and use it for predicting the outcomes $Y|do(X)$ for different samples (e.g., patients). Since we are interested in prediction, we want to select the minimal set of maximally informative features for $Y|do(X)$, in the interest of avoiding over-fitting.

We assume the reader is familiar with causal graphical models and related terminology. We use bold to denote variable sets, uppercase letters to denote single variables, and lowercase letters to denote variable values. We assume that there exists a causal Bayesian network $\langle \mathcal{G}', \mathcal{P}' \rangle$ over the set of observed variables $\mathbf{V}$ and a possibly empty set of latent variables $\mathbf{L}$. Let $\langle \mathcal{G}, \mathcal{P} \rangle$ be the Acyclic Directed Mixed Graph (ADMG) and joint probability distribution (jpd) stemming from marginalizing out the latent variables $\mathbf{L}$ from $\mathcal{G}'$ and jpd P', respectively (Richardson, 2003) If we know the causal ADMG $\mathcal{G}$, a hard intervention where a treatment X is set to x is denoted with the do-operator, $do(X{=}x)$. This operation corresponds to removing all incoming edges into X in the graph. The resulting ADMG is denoted $\mathcal{G}_{\overline{X}}$. The post-intervention distribution is denoted $P_{\overline{X}} := P(do(X), Y, \mathbf{V})$. For brevity, we call the post-interventional distribution of the outcome Y given some subset $\mathbf{Z}$ of $\mathbf{V}$, $P(Y|do(X), \mathbf{Z})$, a causal effect.

For a given ADMG $\mathcal{G}$, a causal effect $P(Y|do(X), \mathbf{Z})$ may or may not be identifiable from the observational distribution $\mathcal{P}$. If $\mathcal{G}$ is known, IDC (Shpitser and Pearl, 2006) returns an expression for $P(Y|do(X), \mathbf{Z})$ based on observational estimands, if one exists, and N/A if the effect is non-identifiable. For pre-treatment covariates $\mathbf{Z}$, an effect $P(Y|do(X), \mathbf{Z})$ is identifiable if and only if $\mathbf{Z}$ satisfies the adjustment criterion (Triantafillou et al., 2021). The adjustment criterion consists of a set of graphical conditions that can be checked in the graph $\mathcal{G}$. For pre-treatment covariates, the adjustment criterion is identical to the backdoor criterion (Pearl, 2000), which is satisfied when $\mathbf{Z}$ blocks all paths connecting X and Y that are into $\mathbf{X}$ (backdoor paths). We say that $\mathbf{Z}$ is an adjustment set for X and Y if it satisfies the adjustment criterion. When $\mathbf{Z}$ is an adjustment set, the conditional pre- and post-interventional distributions are identical:

$$P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z}). \tag{1}$$

Hence, we can use observational data for causal effect estimation. Moreover, for any subset $\mathbf{Z}_b$ of $\mathbf{Z}$, the conditional causal effect can be estimated by marginalizing over the remaining variables $\mathbf{Z}_o \coloneqq \mathbf{Z} \setminus \mathbf{Z}_b$:

$$P(Y|do(X), \mathbf{Z}_b) = \sum_{\mathbf{z}_o} P(Y|X, \mathbf{Z}_b, \mathbf{z}_o)P(\mathbf{z}_o|\mathbf{Z}_b). \tag{2}$$

For $\mathbf{Z}_b = \emptyset$, Eq. 2 is the well-known *adjustment formula* (Shpitser and Pearl, 2006). The adjustment criterion is shown to be complete for adjustment, i.e., if $\mathbf{Z}$ does not satisfy the adjustment criterion, there exists at least one distribution $\mathcal{P}$ induced by $\mathcal{G}$ where the adjustment formula does not hold. For our method, we require a stronger condition, which we call *adjustment faithfulness:*

Assumption 1 (Adjustment faithfulness)**.** *Eq. 2 holds only if* $\mathbf{Z}_b \cup \mathbf{Z}_o$ *is an adjustment set for* X *and* Y.

Given the causal graph $\mathcal{G}$, we can test the graphical conditions of the adjustment criterion, but when $\mathcal{G}$ is unknown these conditions are often untestable. For example, in the graph of Fig. 1, we cannot test if Z blocks all backdoor paths between X and Y using observational data. In Section 3.1 we show how we can use Eq. 2 and Assumption 1 to test if a set $\mathbf{Z}$ is an adjustment set using a smaller experimental data set.

3 THE OVERLAP ALGORITHM

We assume that we have observational data D_o measuring a set of variables $\mathbf{V}$, and experimental data D_e measuring a subset of variables $\mathbf{V}_b \subseteq \mathbf{V}$. We assume that the two data sets are sampled from the same population. This assumption is not always realistic, but it holds in an imporatant and growing set of RCTs called embedded trials (Angus et al., 2020). We discuss this assumption and the robustness of our method to it in the Supplementary. Since we are interested in causal estimation with the purpose of optimizing treatment assignment, all variables $\mathbf{V}$ are assumed to be pre-treatment variables. Our objective is to select a subset $\mathbf{Z}$ of $\mathbf{V}$ that optimizes our prediction for $Y|do(X)$. This involves two separate tasks: (a) Estimating $P(Y|do(X), \mathbf{Z})$ from D_o and D_e for different sets $\mathbf{Z}$ and (b) Evaluating the performance of these estimators with respect to a selected performance criterion such as, e.g., accuracy or log loss (for binary outcomes).

For the first step, the challenges we are faced with are that (a) some of the variables in $\mathbf{Z}$ may be missing from D_e, and (b) $P(Y|do(X), \mathbf{Z})$ may not be identifiable from the observational data D_o available to us. To address these challenges, we first estimate the probability that $P(Y|do(X), \mathbf{Z})$ is identifiable from observational data. If it is, we can use D_o to estimate it. If it is not, $P(Y|do(X), \mathbf{Z})$ is not estimable with the data available to us. Instead, we estimate $P(Y|do(X), \mathbf{Z}_b)$ using D_e alone, where $\mathbf{Z}_b \subset \mathbf{Z}$ is restricted to be variables measured in the experimental data. Our method returns a weighted average of these two cases, weighted by the probability of $\mathbf{Z}$ being an adjustment set.

In the second step, once we have estimated the post-interventional conditional distributions for many different $\mathbf{Z}$'s, we select the optimal set according to some criterion that may be domain-specific. In this work, we present our method for some common performance criteria: Accuracy, log loss, and a user-defined utility function.

3.1 Estimating $P(Y|do(X), \mathbf{Z}, D_e, D_o)$

Let $\mathbf{Z} = \mathbf{Z}_b \cup \mathbf{Z}_o$, where $\mathbf{Z}_b$ is a set of pre-treatment variables measured in both D_e and D_o, and $\mathbf{Z}_o$ is a set of pre-treatment variables measured in D_o only ($\mathbf{Z}_b \cap \mathbf{Z}_o = \emptyset$). The main idea underlying the estimation of $P(Y|do(X), \mathbf{Z}, D_e, D_o)$ is the following: If $P(Y|do(X), \mathbf{Z})$ is identifiable from observational data, we can use D_o to estimate it. If not, we can only use D_e and we can only condition on $\mathbf{Z}_b$, since $\mathbf{Z}_o$ are missing in D_e. Since $P(Y|do(X), \mathbf{Z})$ is only identifiable from observational data if and only if $\mathbf{Z}$ is an adjustment set, and we do not know the causal graph, we are interested in estimating the probability that $\mathbf{Z}$ is an adjustment set based on the data available to us.

We now present a method that estimates the probability that a set is an adjustment set using D_e and D_o. We use a binary variable $\mathcal{H}^a_{\mathbf{Z}}$ to denote that $\mathbf{Z}$ is an adjustment set for X and Y, and $\mathcal{H}^{\overline{a}}_{\mathbf{Z}}$ to denote its complementary hypothesis that $\mathbf{Z}$ is not an adjustment set. We want to estimate $P(\mathcal{H}^a_{\mathbf{Z}}|D_e, D_o)$, which is the probability that $\mathbf{Z}$ is an adjustment set given the observational and experimental data, and thus can be computed as

$$P(\mathcal{H}^a_{\mathbf{Z}}|D_e, D_o) = \frac{P(D_e|\mathcal{H}^a_{\mathbf{Z}}, D_o)P(\mathcal{H}^a_{\mathbf{Z}}|D_o)}{\sum_{A=a,\overline{a}} P(D_e|\mathcal{H}^A_{\mathbf{Z}}, D_o)P(\mathcal{H}^A_{\mathbf{Z}}|D_o)} \tag{3}$$

Notice that $P(\mathcal{H}^{\overline{a}}_{\mathbf{Z}}|D_e, D_o) = 1 - P(\mathcal{H}^a_{\mathbf{Z}}|D_e, D_o)$. The terms in Eq. 3 are: (i) $P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}})$, which is the probability of observing the experimental data D_e when the set $\mathbf{Z}$ is an adjustment set, (ii) $P(\mathcal{H}^a_{\mathbf{Z}}|D_o)$, which is the probability that $\mathbf{Z}$ is an adjustment set based only on the observational data and (iii) the same two probabilities for the event that $\mathbf{Z}$ is not an adjustment set.

3.2 Estimating $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o), P(\mathcal{H}^{\overline{a}}_{\mathbf{Z}}|D_o)$

This represents the probability that $\mathbf{Z}$ is (or is not) an adjustment set for Y w.r.t. X given only the observational data. We view this as a prior for $\mathcal{H}^{a}_{\mathbf{Z}}$ given just the observational data. Several approaches are possible to quantify this probability, for example, we could reason on the space of possible ADMGs, similar to an approach presented in Claassen and Heskes (2012). Let $\mathcal{G}$ be an ADMG over $\{X, Y, \mathbf{Z}\}$ and $\mathcal{G} \vdash \mathcal{H}^{a}_{\mathbf{Z}}$ denote that $\mathbf{Z}$ satisfies the adjustment criterion for X, Y in $\mathcal{G}$. Then we can compute $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o)$ as

$$P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o) = \frac{\sum_{\mathcal{G}\vdash\mathcal{H}^{a}_{\mathbf{Z}}} P(D_o|\mathcal{G})P(\mathcal{G})}{\sum_{\mathcal{G}} P(D_o|\mathcal{G})P(\mathcal{G})} \tag{4}$$

Eq. 4 requires exhaustively enumerating and scoring all possible ADMGs, both of which are very challenging. For a similar problem, Triantafillou and Cooper (2021) suggest approximating Eq. 4 by by finding the most probable Markov equivalence class $[\mathcal{G}]$ of graphs and restricting enumeration of graphs to those in $[\mathcal{G}]$. This is approximation is reasonable because in large sample sizes (which we assume for D_o), graphs in the true Markov equivalence class dominates the score $P(D_o|\mathcal{G})$. $[\mathcal{G}]$ can be learnt with a sound and complete algorithm like FCI. We can then use two different approaches: We can consider all ADMGs within $[\mathcal{G}]$ to be equally probable, so Eq. 4 amounts to enumerating the members of $[\mathcal{G}]$ (for the denominator) and checking in how many of them $\mathbf{Z}$ satisfies the adjustment criterion (for the numerator). This approach is very expensive computationally. Another approach is to assume that every set considered in our method may or may not be an adjustment set with equal probability, hence $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o) = P(\mathcal{H}^{\overline{a}}_{\mathbf{Z}}|D_o) = 0.5$. Therefore, to a first approximation, we consider that the observational data do not provide a lot of information about whether a set is an adjustment set or not.

In practice, $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o)$ does not significantly affect the behavior of the method, because its impact fades quickly with increasing experimental samples. This is true even for relatively small sample sizes. We illustrate this point with an example: We use a simple structure $X \to Y, X \leftarrow Z \to Y$. We then compute Eq. 3 with two quite different $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o)$: 0.1 and 0.9. We use $P_{0.1}, P_{0.9}$ to denote Eq. 3 computed with $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o) = 0.1, 0.9$, respectively. We simulated data from a graph with a single observed confounder Z and computed $P_{0.1}, P_{0.9}$, for $\mathcal{H}^{a}_{\mathbf{Z}}$. We used $N_o = 5000$ and $N_e = 10, 50, 100$ and 500. We plot the distribution of $P_{0.1} - P_{0.9}$. Fig. 2 illustrates the distribution of the absolute difference $|P_{0.1} - P_{0.9}|$ over 500 randomly simulated parameters. As we can see, the difference in estimated $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o, D_e)$ using very different priors $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o)$ diminishes rapidly with increasing experimental sample size. Similar results are reported in Triantafillou et al. (2021). For this reason, we use the uninformative $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o) = P(\mathcal{H}^{\overline{a}}_{\mathbf{Z}}|D_o) = 0.5$ in the remainder of this work, which has the advantage of having no computational overhead.

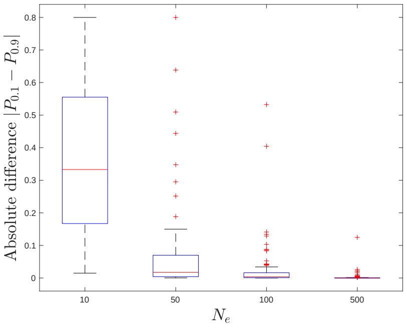

Figure 2: Effect of $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o)$ on Eq. 3. $P_{0.1} = P(\mathcal{H}^{a}_{\mathbf{Z}}|D_e, D_o)$ computed using Eq. 3 with $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o) = 0.1$. Similarly, $P_{0.9} = P(\mathcal{H}^{a}_{\mathbf{Z}}|D_e, D_o)$ is computed using Eq. 3 with $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o) = 0.9$. The effect of $P(\mathcal{H}^{a}_{\mathbf{Z}}|D_o)$ diminishes rapidly with increasing experimental samples.

3.3 Estimating $P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}})$

This represents how likely it is that we observe the experimental data, if we have already seen the observational data and $\mathbf{Z}$ is an adjustment set. This probability is computed on the basis that, when $\mathbf{Z}$ is an adjustment set, $P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z})$, hence, the experimental and observational parameters are the same. For each set of covariates $\mathbf{Z}$, some of the variables in $\mathbf{Z}$ may only be observed in D_o but not D_e, and some may be observed in both. We use the following notation: $\theta_{y_x|\mathbf{z}}$ are the parameters for $P(y|do(x), \mathbf{z})$. $\theta_{y|x,\mathbf{z}}$ are the parameters for $P(y|x, \mathbf{z})$. $\theta_{\mathbf{z}_o|\mathbf{z}_b}$ are the parameters for $P(\mathbf{z}_o|\mathbf{z}_b)$.

Case 1: $\mathbf{Z}_o = \emptyset$. This is the case where all variables are measured both in observational and experimental data. By integrating over all $\theta_{y_x|\mathbf{z}}$, we get

$$P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}) = \int_{\theta_{y_x|\mathbf{z}}} P(D_e|\theta_{y_x|\mathbf{z}}) f(\theta_{y_x|\mathbf{z}}|D_o, \mathcal{H}^{a}_{\mathbf{Z}}) d\theta_{y_x|\mathbf{z}}, \tag{5}$$

where $f(\theta_{y_x|\mathbf{z}}|D_o, \mathcal{H}^{a}_{\mathbf{Z}})$ is the posterior for $\theta_{y_x|\mathbf{z}}$ given the observational data. When $\mathcal{H}^{a}_{\mathbf{Z}}$ is true, $\mathbf{Z}$ is an adjustment set and therefore $f(\theta_{y_x|\mathbf{z}}|D_o, \mathcal{H}^{a}_{\mathbf{Z}}) = f(\theta_{y|x,\mathbf{Z}}|D_o)$, and Eq. 5 can be rewritten using observational parameters, and computed in closed form for families of distributions with closed form marginal likelihoods. Notice that $\mathbf{Z} = \mathbf{Z}_b$ in this case.

Case 2: $\mathbf{Z}_o \neq \emptyset$. In this case, we have some of the variables in $\mathbf{Z}$ not measured in the experimental data. Under $\mathcal{H}^{a}_{\mathbf{Z}}$, $\mathbf{Z}$

is an adjustment set and therefore

$$P(D_e|D_o,\mathcal{H}^a_{\mathbf{Z}}) = \int_{\theta_{y_x|\mathbf{z_b}}} P(D_e|\theta_{y_x|\mathbf{z_b}}) f(\theta_{y_x|\mathbf{z_b}}|D_o,\mathcal{H}^a_{\mathbf{Z}}) d\theta_{y_x|\mathbf{z_b}}, \tag{6}$$

Under $\mathcal{H}^a_{\mathbf{Z}}$, $P(y|do(x),\mathbf{z}_b) = \sum_{\mathbf{z}_o} P(y|x,\mathbf{z}_b,\mathbf{z}_o)P(\mathbf{z}_o|\mathbf{z}_b)$ for all $x, y, \mathbf{z}_b, \mathbf{z}_o$. Hence we can recast Eq.5 using observational parameters as follows:

$$P(D_e|D_o,\mathcal{H}^a_{\mathbf{Z}}) = \int_{\theta_{y|x,\mathbf{z}}} \int_{\theta_{\mathbf{z}_o|\mathbf{z}_b}} P(D_e|\theta_{y|x,\mathbf{z}}) f(\theta_{y|x,\mathbf{z}}, \theta_{\mathbf{z}_o|\mathbf{z}_b}|D_o) \theta_{y|x,\mathbf{z}} d\theta_{\mathbf{z}_o|\mathbf{z}_b}, \tag{7}$$

Eq. 7 includes multiple integrals and cannot be computed in closed form. However, we can estimate it using a sampling procedure, described in Alg. 1. First, we learn a Bayesian network $\mathcal{B}$ over variables $\mathbf{Z}, X, Y$ using D_o (Line 1). $\mathcal{B}$ consists of a DAG graph $\mathcal{G}_{\mathcal{B}}$ and the posterior distributions for its parameters $f(\theta_{i|pa_i}|\mathcal{G}_{\mathcal{B}}, D_o)$ (Line 2). We can then use $\mathcal{B}$ to sample from the posterior observational parameters and compute the predicted post-intervention parameters under $\mathcal{H}^a_{\mathbf{Z}}$. Notice that $\mathcal{B}$ is not a causal graphical model, but rather represents the joint distribution of $\mathbf{Z}, X, Y$. Then, for every configuration $x, \mathbf{z}_b$ in D_e and every iteration i, we can sample the posteriors $\theta^i_{v|pa_v}$ and then compute $\theta^i_{y_x|\mathbf{z}_b} = \sum_{\mathbf{z}_o} \theta^i_{y|x,\mathbf{z}} \theta^i_{\mathbf{z}_o|\mathbf{z}_b}$ (Line 6). We then score D_e by computing the likelihood given these estimated parameters $P(D_e|\boldsymbol{\theta}^i_{y_x|\mathbf{z}_b})$ (Line 7). We repeat the process over I samples, and take the average over all samples.

3.4 Estimating $P(D_e|D_o, \mathcal{H}^{\overline{a}}_{\mathbf{Z}}\ )$

This term represents how likely it is that we observe the experimental data, if we have already seen the observational data and $\mathbf{Z}$ is *not* an adjustment set. Under Assumption 1, in this case, Eq.2 does not hold. Hence, under this assumption, we cannot use the observational data to compute obtain a point estimate of the post-intervention distribution[1]. Therefore $f(\theta_{y_x|\mathbf{z}_b}|D_o, \mathcal{H}^{\overline{a}}_{\mathbf{z}}) = f(\theta_{y_x|\mathbf{z}_b})$ in Eqs. 5 and 7, and we can use a weak prior to score D_e in both cases (Line 9). The score can be computed in closed form for families of distributions with closed form marginal likelihoods.

We can now compute $P(\mathcal{H}^a_{\mathbf{Z}}\ |D_e, D_o)$ using Eq. 3. (Line 10).

Equations 5 and 7 do not assume a specific distribution, and can be computed using the process described above for any distribution where the marginal likelihood can be computed in closed form or approximated. In the Supplementary, we formulate these equations for discrete variables with Dirichlet-multinomial distributions.

For discrete variables and Case 1, i.e, when $\mathbf{Z}_o = \emptyset$, the method will asymptotically correctly identify if $\mathbf{Z}$ is an adjustment set or not, with probability 1:

Theorem 1. *Let D_o, D_e be an observational data set and an experimental data set, respectively, both measuring treatment X, outcome Y, and pre-treatment covariates $\mathbf{V}$, all discrete. Let D_o, D_e contain N_o, N_e cases respectively, sampled from distributions $\mathcal{P}, \mathcal{P}_{\overline{X}}$ respectively, both strictly positive in the sample limit. Also, let $\mathcal{P}$ be a perfect map for an ADMG $\mathcal{G}$. We assume N_o and N_e increase equally without limit. Then the proposed method converges to the data-generating model in the large sample limit:*

$$\lim_{N\to\infty} P(\mathcal{H}^a_{\mathbf{Z}}\ |D_o, D_e) = \begin{cases} 1, & \textit{if } \mathbf{Z} \textit{ is an adjustment} \\ & \textit{set for } X \textit{ and } Y \\ 0, & \textit{otherwise} \end{cases}$$

A proof can be found in the Supplementary. For the case where $\mathbf{Z}_o \neq \emptyset$, asymptotic behavior is more complex because of the sampling procedure, and the convergence of Alg. 1 is left as future work.

3.5 Estimating $P(Y|do(X), \mathbf{Z}, D_e, D_o)$

Having computed the probability that $\mathbf{Z}$ is an adjustment set, we can now estimate the conditional post-interventional distribution as a weighted average of the two complementary hypotheses:

$$P(Y|do(X), \mathbf{Z}, D_e, D_o) = \sum\nolimits_{A=a,\overline{a}} P(Y|do(X), \mathbf{Z}, D_e, D_o, \mathcal{H}^A_{\mathbf{Z}}\ ) P(\mathcal{H}^A_{\mathbf{Z}}\ |D_e, D_o). \tag{8}$$

$P(Y|do(X), \mathbf{Z}, D_e, D_o, \mathcal{H}^a_{\mathbf{Z}})$ is estimated using D_o if $\mathbf{Z}_o \neq \emptyset$ and $D_e \cup D_o$ if $\mathbf{Z}_o{=}\emptyset$. $P(Y|do(X), \mathbf{Z}, D_e, D_o, \mathcal{H}^{\overline{a}}_{\mathbf{Z}})$ is estimated based on D_e and $\mathbf{Z}_b$ alone, since including the observational data would yield a biased estimate. Eq. 8 is used in line 12 of Alg. 1 to estimate the parameters of the post-intervention distribution conditional on $\mathbf{Z}$. In all cases, we use posterior expectations as the probability estimates. Finally, Alg. 1 also estimates $\hat{P}(\mathbf{Z})$ from the BN $\mathcal{B}$ (line 12), to be used in the next step of selecting the optimal set $\mathbf{Z}$.

3.6 Finding Sets for Optimal Causal Prediction

Once we have estimated $P(Y|do(X), \mathbf{Z}, D_e, D_o)$ for different sets $\mathbf{Z}$, we want to identify the set that allows for optimal prediction of $Y|do(X)$, with respect to some criterion of optimality (e.g., optimal expected utility). We use a function g to encode this optimality criterion. Hence, the

[1] This is true for point estimates, however, it may be possible to compute bounds on the post-intervention distribution. We consider this out of the scope of the present work, however, we believe that if the bounds are very tight, our method will favor $\mathcal{H}^a_{\mathbf{Z}}$, since D_o will be very informative about D_e.

Algorithm 1: ScoreD_e

```
input  : $X, Y, \mathbf{Z}_b, \mathbf{Z}_o, D_o, D_e, I$
output: $P(\mathcal{H}^a_\mathbf{Z} \,|D_e, D_o), P(\mathcal{H}^{\overline{a}}_\mathbf{Z} \,|D_e, D_o), \theta_{y_x|\mathbf{z}}, \theta_\mathbf{z}$
$\langle \mathcal{B}, \theta_{v|pa_v} \rangle \leftarrow$ LearnBN$(D_o, X, Y, \mathbf{Z})$
foreach $i = 1, \ldots, I$ do
    Sample $\theta^i_{v|pa_v}$ from $\mathcal{B}$
    foreach configuration $x, \mathbf{z}_b$ in $D_e$ do
        Compute $\theta^i_{y|x,\mathbf{z}}, \theta^i_{\mathbf{z}_o|\mathbf{z}_b}$ using inference on $\mathcal{B}$
        Compute $\theta^i_{y_x|\mathbf{z}_b}$ using Eq. 2
    Compute the likelihood $p^i = P(D_e|\boldsymbol{\theta}^i_{y_x|\mathbf{z}_b})$
$P(D_e|D_o, \mathcal{H}^a_\mathbf{Z}\ ) \leftarrow \frac{1}{I}\sum_i p^i$
$P(D_e|D_o, \mathcal{H}^{\overline{a}}_\mathbf{Z}\ ) \leftarrow$
    $\int_{\theta_{y_x|\mathbf{z_b}}} P(D_e|\theta_{y_x|\mathbf{z_b}}) f(\theta_{y_x|\mathbf{z_b}}) d\theta_{y_x|\mathbf{z_b}}$
$P(\mathcal{H}^a_\mathbf{Z} \,|D_e, D_o) \leftarrow$
    $P(D_e|D_o, \mathcal{H}^a_\mathbf{Z}) / \sum_{A=a,\overline{a}} P(D_e|D_o, \mathcal{H}^A_\mathbf{Z})$
$P(\mathcal{H}^{\overline{a}}_\mathbf{Z} \,|D_e, D_o) \leftarrow 1 - P(\mathcal{H}^a_\mathbf{Z} \,|D_e, D_o)$
$\theta_{y_x|\mathbf{z}} \leftarrow$ Compute using Eq.8
$\theta_\mathbf{z} \leftarrow$ Posterior expectations for $P(\mathbf{Z})$ based on $\mathcal{B}$
```

purpose of function g is to evaluate the expected performance of a given set of covariates $\mathbf{Z}$ in predicting $Y|do(X)$, relative to the user's performance goal.

The expected performance of a goal function will generally be a function of the covariate set $\mathbf{Z}$, and the corresponding estimated post-intervention parameters $\hat{\theta}_{y_x,\mathbf{z}}$. In the formulae described below, we also use $\hat{\theta}_\mathbf{z}$ to denote estimated the parameters for $P(\mathbf{Z})$. In this section, we derive the expectations of three different goal functions: Expected accuracy, expected log loss, and expected utility.

Expected Accuracy: We now derive the expected accuracy of predicting with the distribution $P(Y|do(X), \mathbf{Z}, D_e, D_o)$ learnt using Alg. 1, for a specific set $\mathbf{Z}$. For every configuration $\mathbf{z}$ of $\mathbf{Z}$ and every possible treatment value x, let $y^* = \arg\max_y \hat{\theta}_{y_x|\mathbf{z}}$. Hence, y^* is the predicted value of $Y|do(x)$ according to the probability distribution $\hat{P}(Y|do(x), \mathbf{z}, D_e, D_o)$ returned by Alg. 1. The expected accuracy of this prediction is $\hat{\theta}_{y^*_x|\mathbf{z}}$. To compute the overall expected accuracy, we need to weigh this accuracy by the probability that covariates $\mathbf{Z} = \mathbf{z}$ will occur. Since we make a prediction of Y for each value of X, we take the mean expected accuracy over the $|X|$ predictions. The expected accuracy $g_{acc}(\mathbf{Z}, \hat{\theta}_\mathbf{Z}, \hat{\theta}_{y_x|\mathbf{z}})$ is then

$$g_{acc}(\mathbf{Z}, \hat{\theta}_\mathbf{z}, \hat{\theta}_{y_x|\mathbf{z}}) = \frac{1}{|X|} \sum_{x,\mathbf{z}} \hat{\theta}_\mathbf{z} \hat{\theta}_{y^*_x|\mathbf{z}} \qquad (9)$$

Expected Log Loss: Log loss is also a very popular performance metric, particularly if we are interested in selecting a proper scoring rule. The log loss of a prediction of Y in a given instance is $\log(\hat{\theta}_{y^*_x|\mathbf{z}})$, where y^* is the actual outcome of Y, and $\hat{\theta}_{y^*_x|\mathbf{z}}$ is the estimated probability $P(Y = y^* \,|do(x), \mathbf{z})$ for that outcome. We can derive the expected log loss in predicting the next instance as follows:

$$g_{ll}(\mathbf{Z}, \hat{\theta}_\mathbf{z}, \hat{\theta}_{y_x|\mathbf{z}}) = -\sum_{x,\mathbf{z}} \hat{\theta}_\mathbf{z} \sum_y \hat{\theta}_{y_x|\mathbf{z}} \log(\hat{\theta}_{y_x|\mathbf{z}}) \qquad (10)$$

Algorithm 2: Overlap

```
input  : Variables $X, Y, \mathbf{V}$, data $D_o, D_e$, optimality
         criterion $g$, sampling iterations $I$.
output: Optimal set $\mathbf{Z}^*$, estimated performance $g_{\mathbf{Z}^*}$,
         optimal prediction parameters $\theta_{y_x|\mathbf{z}^*}$
1 $\mathbf{V}' \leftarrow$ Markov Boundary of $Y$ in $D_o$;
2 foreach $\mathbf{Z} = \mathbf{Z}_b \cup \mathbf{Z}_o \subseteq \mathbf{V}'$ do
3     $\hat{P}(D_e|D_o, \mathcal{H}^a_\mathbf{Z}\ ), \hat{P}(D_e|D_o, \mathcal{H}^{\overline{a}}_\mathbf{Z}\ ), \hat{\theta}_{y_x|\mathbf{z}}, \hat{\theta}_\mathbf{z} \leftarrow$
4         $ScoreD_e\ (X, Y, \mathbf{Z}_b, \mathbf{Z}_o, D_o, D_e, I)$
5     Estimate expected performance:
          $g_\mathbf{Z} \leftarrow g(\mathbf{Z}, \hat{\theta}_{y_x|\mathbf{z}}, \hat{\theta}_\mathbf{z})$ using one of Eqs. 9, 10, 11
6 Select set $\mathbf{Z}^*$ that maximizes expected performance:
  $\mathbf{Z}^* \leftarrow \arg\max_\mathbf{Z} g_\mathbf{Z}$
```

Expected Utility: In many decision-making settings, it could be the case that different pairs of treatments and outcomes have different utilities. Let $U(x, y)$ be the utility of receiving treatment $X = x$, followed by experiencing outcome $Y = y$. Then the expected utility can be computed as follows:

$$g_U(\mathbf{Z}, \hat{\theta}_\mathbf{z}, \hat{\theta}_{y_x|\mathbf{z}}) = \sum_\mathbf{z} \hat{\theta}_\mathbf{z} \sum_y \hat{\theta}_{y_{x^*}|\mathbf{z}} U(x^*, y), \qquad (11)$$

where $x^* = \arg\max_x \sum_\mathbf{z} \hat{\theta}_\mathbf{z} \sum_y \hat{\theta}_{y_x|\mathbf{z}} U(x, y)$. Thus, x^* is the action that maximizes the expected utility for patients with $\mathbf{Z} = \mathbf{z}$. This basic problem can be readily extended.

Algorithm 2 describes the final algorithm, which we call Overlap. Initially, the algorithm selects a set of covariates $\mathbf{V}'$ to include in the search. We opted to chose the Markov Boundary of Y in D_o, since it is guaranteed to include an adjustment set, if one exists (Line 1). Then, for each subset of $\mathbf{V}'$, use Alg. 1 to estimate the probability that $\mathbf{Z}$ is an adjustment set or not, and the corresponding post-interventional parameters using Eq. 8 (Line 4). We also compute the expected performance of $\mathbf{Z}$ (Line 5). Finally, we select the set $\mathbf{Z}^*$ which optimizes the expected performance.

4 RELATED WORK

Our proposed methodology combines two tasks: The first is identifying adjustment sets using possibly overlapping observational and experimental data, and the second is selecting the optimal set $\mathbf{Z}$ for estimating the conditional post-intervention distribution $P(Y|do(X), \mathbf{Z})$. To the best of our knowledge, Overlap is the first algorithm to address both questions. We discuss connections to related work in each of these tasks separately.

Finding adjustment sets: One line of work tries to select an adjustment set from observational data. VanderWeele and Shpitser (2011) adjust for causes of both the treatment X and the outcome Y, where the causes are assumed to be known. The resulting set is guaranteed to be an adjustment set if one exists. Entner et al. (2013) use a set of rules for identifying valid adjustment sets from conditional (in)dependencies in the observational data, when adjustment sets are identifiable. Both methods focus on identifiability (returning a valid adjustment set), and not optimality. Some methods (Perkovic et al., 2017; Rotnitzky and Smucler, 2020; Smucler et al., 2021; Witte et al., 2020) identify adjustment sets that are optimal for Average Treatment Effect (ATE) estimation. Given a graph that is known (DAG/ADMG) or may be estimated from D_o (PDAG/PAG), they give graphical criteria for identifying adjustment sets, if it is possible. The methods apply a graphical adjustment criterion to identify a set of valid adjustment sets for estimating the ATE of X on Y, and then to identify the set that leads to the most efficient estimator. These methods are not directly comparable to ours since they focus on estimating ATEs, define optimality as the efficiency of the ATE estimator and apply to cases where adjustment sets can be identified (are amenable) from the graph. Triantafillou and Cooper (2021) use observational and experimental data to rank adjustment sets and select the best adjustment set $\mathbf{Z}$ for estimating the ATE $P(Y|do(X)$. Their approach is similar to ours, but their goal is only to select set that is most likely to be a valid adjustment set. They do not compute conditional probabilities for $\mathcal{H}^a_{\mathbf{Z}}$, and their goal is to improve the ATE estimation. They also assume that only marginal distributions of variables measured in D_e are available.

Another line of work focuses on identifiability of the post-intervention distribution from observational data, via adjustment or otherwise (Shpitser and Pearl, 2006; Tian and Shpitser, 2003; Jaber et al., 2019). These methods can answer if a causal query is identifiable from observational distributions, based on d-separation/d-connection constraints implied by the causal graph, but they are not directly comparable to our approach.

Combining observational and experimental data. There is a growing body of work for combining observational and experimental data in the field of potential outcomes, mostly focusing on improving the external validity of the RCT (See Colnet et al. (2022) and references therein). Most of these works rely on conditional ignorability, and focus on generalizability of the causal effects. Kallus et al. (2018) correct confounding bias in the observational data by assuming that it has a parametric structure that can be modeled and computed from the experimental data, measuring the same variables. There is also a body of work on combining observational and experimental data to learn causal graphs (Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Mooij et al., 2019; Andrews et al., 2020). Hyttinen et al. (2015) can also combine conditional independencies from both D_e and D_o, to answer if causal effects are identifiable. This method is not directly comparable to `Overlap` since it does not search for optimal sets for prediction. Moreover, they return expressions for post-intervention distributions based on observational data and not all available data. Finally, all methods referred to above rely on conditional independencies alone, while our method can make inferences beyond that: for example, in Fig. 1, the fact that Z is an adjustment set is not identifiable from conditional (in) dependencies if Z is measured in D_o but not in D_e. In contrast, our method can make this inference based on Eq. 2. Ilse et al. (2021) combine observational and experimental data by reducing possibly multiple latent confounders into a single latent variables, and then derive bounds on causal effects.

Optimal prediction of post-intervention target variable. Several approaches for estimating CATEs from observational data build predictive models for the post-intervention outcome, and implicitly pick variables (e.g., causal forests (Athey et al., 2019)). However, these methods rely on the untestable assumption of ignorability. Selecting the optimal conditioning set for prediction is also closely related to the notion of Markov Boundaries. The Markov Boundary MB(Y) is the minimal set of variables that, when conditioned upon, make all other variables independent from Y, and is shown to lead to optimal prediction of Y when a proper scoring rule is used (Aliferis et al., 2010). Triantafillou et al. (2021) discuss Markov boundaries for post-intervention outcomes. For post-intervention outcomes $Y|do(x)$, the corresponding set is the Markov boundary $\mathbf{Z}$ of Y in the post-intervention graph $\mathcal{G}_{\overline{X}}$, also called the Interventional Markov Boundary (IMB, denoted MB$_X(Y)$). If $\mathbf{Z}$ is the IMB of Y with respect to X, then $P(Y|do(X), \mathbf{Z} \setminus X) = P(Y|do(X), \mathbf{V})$. However, $P(Y|do(X), \mathbf{Z} \setminus X)$ may not be identifiable from observational data. To address this issue, (Triantafillou et al., 2021) propose an algorithm (`FindIMB`) for identifying interventional Markov boundaries from mixtures of observational and experimental data. `FindIMB` takes as input D_e and D_o, measuring the same set of variables $\mathbf{V}$, and returns a Bayesian estimate for $P(Y|do(X), \mathbf{V})$, computed by conditioning on possible IMBs $\mathbf{Z}$, using both observational and interventional data when appropriate. The approach is similar to `Overlap`, but it cannot include additional covariates in D_o and does not admit different criteria for optimality. It also does not compute the probability that a set is an adjustment set, and its asymptotic behavior is not discussed, while we prove the asymptotic convergence of `Overlap` in the same setting.

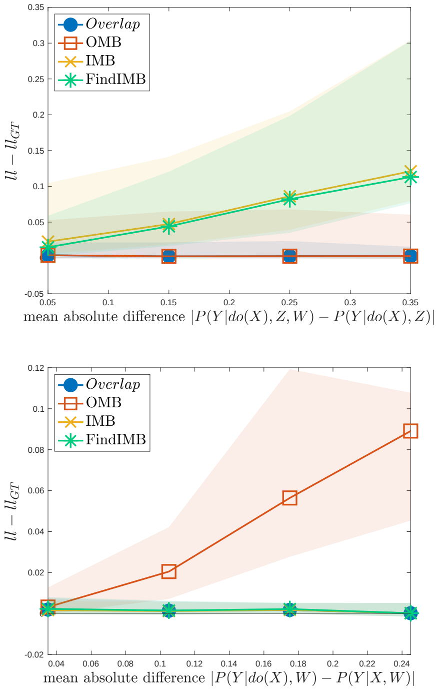

Figure 3: Simulated data based on Fig. 1 and two scenarios (Top) Scenario 1: D_o measures X, Y, Z, W, while D_e measures X, Y, Z. Hence, $P(Y|X, Z, W)$ estimated in D_o can predict $Y|do(X)$ better than $P(Y|do(X), Z)$ estimated in D_e, therefore `Overlap` performs on par with the observational data. `FindIMB` and the estimator based on D_e perform poorly as the ground truth difference of the two conditional distributions increases. (Bottom) Scenario 2: Both D_e and D_o measure X, Y, W. Conditional ignorability does not hold, and the estimator $P(Y|X, W)$ from D_o performs poorly as the ground truth bias increases. `Overlap` and `FindIMB` identify that there is bias and perform on par with the unbiased estimator from D_e. Shaded areas show the 10th and 90th percentile.

5 EXPERIMENTS

In this section, we evaluate the performance of Alg. 2 and compare against alternative approaches. Most methods for optimal prediction of $Y|do(X)$ focus on observational or experimental data alone. We note that under conditional ignorability, $P(Y|do(X), \mathbf{V}) = P(Y|X, \text{MB}_X(Y) \setminus X)$, and $\text{MB}_X(Y)$ is the minimal set for which this equation holds. We create the following baseline comparisons which learn only on observational or only on experimental data:
(a) **OMB**: Use the ground truth (observational) Markov Boundary MB(Y), assume conditional ignorability, and estimate $P(Y|do(X), \mathbf{V})$ as $P(Y|X, \text{MB}(Y) \setminus X)$ from D_o.
(b) **IMB**: Use the ground truth IMB $\text{MB}_X(Y)$ among the variables $\mathbf{V}_b$ measured in D_e, and estimate $P(Y|do(X), \mathbf{V}_b)$ as $P(Y|do(X), \text{MB}_X(Y) \setminus X)$ from D_e.

We also compare with methods combining observational and experimental data for optimal prediction. Since these methods assume that D_e and D_o measure the same variables, we apply the methods on data on $\mathbf{V}_b$ only. (c) `FindIMB`: The method proposed in Triantafillou et al. (2021) for finding the IMB by combining observational and experimental data. (d) **FCIt-IMB:** In this approach, we learn a Partial Ancestral Graph $\mathcal{P}$ over $\mathbf{V}_b$ using D_o and D_e, and then use the Markov Boundary Y in the post-intervention $\mathcal{G}_{\overline{X}}$ to be the $\text{MB}_X(Y)$. We then test if $\text{MB}_X(Y)$ is an adjustment set in the graph. If so, we use both D_e and D_o pooled together to estimate $P(Y|do(X), \text{MB}_X(Y) \setminus X)$. Otherwise, we only use D_e.

We tested our methods in a number of settings, described below. In all settings, the data are simulated from random ADMGs with a treatment X, an outcome Y with $X \to Y$, a set of covariates $\mathbf{V} = \mathbf{V}_b \cup \mathbf{V}_o$. $\mathbf{V}_b, \mathbf{V}_o$ are disjoint sets corresponding respectively to variables included in both D_e and D_o and variables included in D_o alone. We use N_o and N_e to denote the number of cases in D_o and D_e, respectively. We also simulated a data set D_{test} with 2000 samples from the post-intervention distribution, where we evaluate the performance of the methods using log-loss in the prediction of $Y|do(X)$. To make different simulations comparable, we report a metric which we call $ll - ll_{GT}$, defined as the difference of the log loss to the ground truth log loss, i.e., the log loss computed using the ground truth probabilities of $P(Y|do(X), \mathbf{V})$. Larger differences correspond to worse prediction and a zero difference corresponds to perfect estimation of the true distribution.

We first illustrate the benefits of our approach with the two very simple scenarios, based on Fig. 1. In the first scenario, D_o measures X, Y, Z, W and D_e measures X, Y, Z. Hence, conditional ignorability holds, but the observational data include an additional variable W. Asymptotically, we expect the following behavior: `Overlap` identifies that $\{Z, W\}$ is an adjustment set, and returns it as the optimal set for predicting $Y|do(X)$. It therefore uses $P(Y|do(X), Z, W) = P(Y|X, Z, W)$ estimated from D_o to predict $Y|do(X)$. OMB and methods relying on conditional ignorability will also return the same estimate. `FindIMB` will return $P(Y|do(X), Z)$ estimated from D_e, since the method does not admit additional observational variables. IMB will also use the same prediction. Hence, `Overlap` and OMB will perform best.

In the second scenario, the confounder is unobserved in both D_e and D_o. OMB will use $P(Y|X, W)$ to predict $Y|do(X)$, which is a biased estimate. In contrast, `Overlap`, `FindIMB`, and IMB will use $P(Y|do(X), W)$ estimated from D_e. Hence, `Overlap`, `FindIMB` and IMB will perform best.[2]

[2] We omit `FCItiers` for clarity of presentation in this part, as it is asymptotically performs like `FindIMB` in both scenarios.

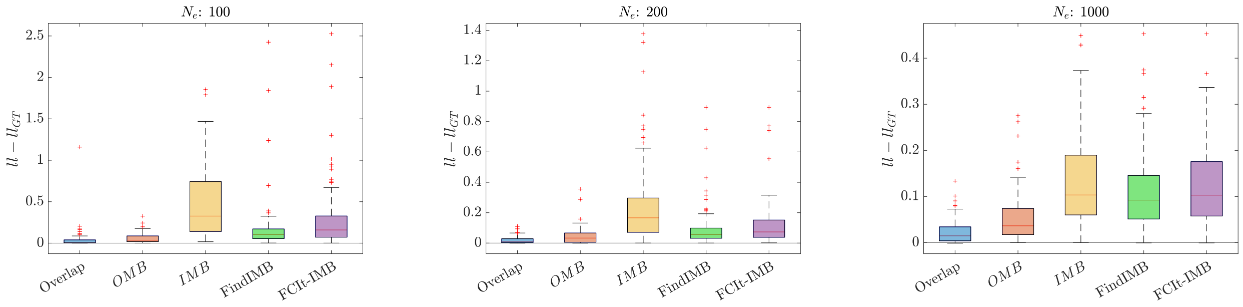

Figure 4: Boxplots of the $ll - ll_{GT}$ metric based on 100 random graphs, with $N_o = 5000$ and increasing N_e. Overlap outperforms alternative methods.

Fig. 3 shows simulations based on the scenarios described above. For the first scenario, the x-axis shows the mean difference in the true distributions $|P(y|do(x), z) - P(y|do(x))|$, averaged over all possible values x, y, z. Larger values of these difference indicate increased significance of Z in predicting $Y|do(X)$. As expected, Overlap and OMB perform best, constantly achieving an almost zero $ll - ll_{GT}$, while methods that cannot use D_o perform worse. In the second scenario, the the x-axis shows the mean difference in the true distributions $|P(y|do(x), w) - P(y|x, w)|$, averaged over all possible values x, y, w. Larger values of these difference indicate larger bias of the observational estimate. In this case, OMB performs worse, while Overlap, FindIMB and IMB perform best. Hence, Overlap performs best in both scenarios, while the ranking of the other methods depends on whether the untestable assumption of conditional ignorability holds or not. FindIMB performs similarly to Overlap when $\mathbf{V}_b = \mathbf{V}_o = \mathbf{V}$ for the log loss metric, but cannot include W in the first scenario.

We also tested our methods on random graphs with 6 covariates, 4 observed in both D_e and D_o and 2 observed only in D_o. Results can be seen in Fig. 4, showing boxplots for the $ll - ll_{GT}$ metric for every method, and for different N_o and N_e. Overlap performs best in all cases. OMB is second best; this is because random graphs often do not lead to large biases. Additional experiments can be found in the Supplementary.

6 CONCLUSIONS

In this work we discuss learning causal effects by combining observational and experimental data. The problem is split in two parts (a) determining if we can use the observational data for causal effect estimation and (b) selecting the optimal covariate set for personalized effect estimation. To our knowledge, Overlap is the first method to address both questions, when some of the variables are not measured in the experimental data. We believe our method is very relevant to the clinical domain, where it is common for RCTs to be limited in number of variables and patients measured. Overlap has some limitations: It assumes that the observational and experimental data measure the same population, and it is exhaustively scores all subsets of a set of variables relevant to the outcome, so it cannot scale up to large variable sets. In the future, we plan to address these issues and extend our implementation to continuous and mixed data sets.

Acknowledgements

This research was funded in part by grant R01HL164835 from the National Heart, Lung, and Blood Institute of the National Institutes of Health. Part of this work was done while Sofia Triantafillou was visiting the Simons Institute for the Theory of Computing.

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010). Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1).

Andrews, B., Spirtes, P., and Cooper, G. F. (2020). On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4002–4011. PMLR.

Angus, D. C., Berry, S., Lewis, R. J., Al-Beidh, F., Arabi, Y., van Bentum-Puijk, W., Bhimani, Z., Bonten, M., Broglio, K., Brunkhorst, F., et al. (2020). The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study rationale and design. *Annals of the American Thoracic Society*, 17(7):879–891.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.

Bareinboim, E. and Pearl, J. (2014). Transportability from multiple environments with limited experiments: Completeness results. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Claassen, T. and Heskes, T. (2012). A Bayesian approach to constraint based causal inference. In *Uncertainty in Artificial Intelligence*.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2022). Causal inference methods for combining randomized trials and observational studies: a review.

Cover, T. M. (1999). *Elements of Information Theory*. John Wiley & Sons.

Entner, D., Hoyer, P., and Spirtes, P. (2013). Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 256–264.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2015). Do-calculus when the true graph is unknown. In *UAI*, pages 395–404. Citeseer.

Ilse, M., Forré, P., Welling, M., and Mooij, J. M. (2021). Combining interventional and observational data using causal reductions. *arXiv*.

Jaber, A., Zhang, J., and Bareinboim, E. (2019). Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, pages 2981–2989.

Kallus, N., Puli, A. M., and Shalit, U. (2018). Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pages 10888–10897.

Mooij, J., Magliacane, S., and Claassen, T. (2019). Joint causal inference from multiple contexts. *arXiv*, (1611.10351).

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*, volume 113 of *Hardcover*. Cambridge University Press.

Perkovic, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2017). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18(1):8132–8193.

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Uncertainty in Artificial Intelligence*.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (pp. 527-536).*

Smucler, E., Sapienza, F., and Rotnitzky, A. (2021). Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65.

Tian, J. and Shpitser, I. (2003). On the identification of causal effects.

Triantafillou, S. and Cooper, G. (2021). Learning adjustment sets from observational and limited experimental data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9940–9948.

Triantafillou, S., Jabbari, F., and Cooper, G. (2021). Causal markov boundaries. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*.

Triantafillou, S. and Tsamardinos, I. (2015). Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205.

VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413.

Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. (2020). On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45.

A ADDITIONAL EXPERIMENTS

In this section, we present some additional experiments in simulated data. We simulated data as described in the main paper. In Fig. 5, we show results using data with 6 covariates, 4 observed in both D_e and D_o and 2 observed only in D_o. We simulated 1000 samples for D_o and a varying number of samples for D_e. In Fig. 6, we show results in data with 8 covariates, 6 observed in both D_e and D_o and 2 observed only in D_o. We simulated 5000 samples for D_o and a varying number of samples for D_e.

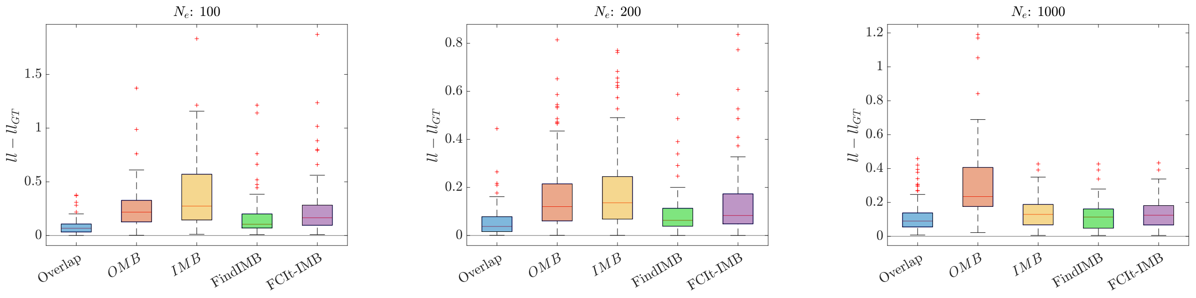

Figure 5: Boxplots of the $ll - ll_{GT}$ metric based on 100 random graphs, with $N_o = 1000$ and increasing N_e. `Overlap` outperforms alternative methods.

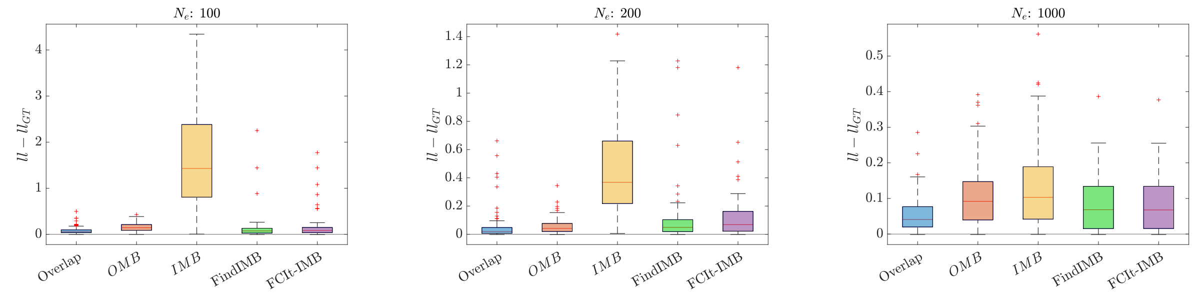

Figure 6: Boxplots of the $ll - ll_{GT}$ metric based on 100 random graphs with 8 observed and 5 hidden variables, with $N_o = 5000$ and increasing N_e. `Overlap` outperforms alternative methods.

B CLOSED-FORM FORMULAE FOR DISCRETE VARIABLES

In this section, we present the formulae for computing Eqs. 5, 8 and some terms in Alg. 1 the main paper, for multinomial distributions with Dirichlet priors. Eq. 7 is computed using sampling in Alg. 1. Subscript jk refers variable Y taking its k-th out of r configurations, and variable set $\mathbf{W} = X \cup \mathbf{Z}$ taking its j-th out of q configurations. α_{jk} is the prior for the Dirichlet distribution. We set $\alpha_{jk} = 1$ in all experiments. $N_{o,jk}, N_{e,jk}$ correspond respectively to counts in the data where $Y = k$ and $\mathbf{W} = j$ in D_o and D_e. $N_{o,j}, N_{e,j}$ correspond to counts in the data where $\mathbf{W} = j$. Also, for every configuration j let j_b be the corresponding configuration of the variables $\mathbf{W}_b = X \cup \mathbf{Z}_b$ measured in D_e. Let N_{o,j_bk}, N_{e,j_bk} be the counts in the data where $Y = k$ and $\mathbf{W}_b = j_b$, and let $N_{o,j}, N_{e,j}$ be to counts in the data where $\mathbf{W}_b = j_b$. $\alpha_{j_bk}, \alpha_{j_b}$ are additional priors for the Dirichlet distribution.

C ON THE ASSUMPTION OF SAMPLING D_e, D_o FROM THE SAME POPULATION

Our method uses the assumption that the observational and experimental populations are the same. This means that the underlying causal models are the same (apart from edge removals due to randomization). Specifically, we assume that the conditional distribution of Y given the treatment and an adjustment set $\mathbf{Z}$ remains the same between the two populations, and that the distribution of the covariates that are unobserved in D_e also remains the same between distributions. Our work is heavily motivated by embedded clinical trials Angus et al. (2020). These trials take place within usual clinical care, so the assumption that the observational and experimental populations are the same is reasonable.

While this assumption is reasonable in embedded clinical trials, it often does not hold for trials and observational data measured in different populations. In this case, Eq. 2 in the main paper will not hold, and `Overlap` will fail to identify an

Table 1: Closed-form solutions for Eq. 5 and 8 and Alg. 1 in the main paper, for multinomial distributions with Dirichlet priors.

Equation	Analytical Expression
Eq. 5	$P(D_e \vert D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \prod_{j=1}^{q} \frac{\Gamma(\alpha_j + N_{o,j})}{\Gamma(\alpha_j + N_{o,j} + N_{e,j})} \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N_{o,jk} + N_{e,jk})}{\Gamma(\alpha_{jk} + N_{o,jk})}$
Alg. 1 Line 9	$P(D_e \vert D_o, \mathcal{H}^{\overline{a}}_{\mathbf{Z}}\ ) = \prod_{j=1}^{q} \frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j + N_{e,j})} \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N_{e,jk})}{\Gamma(\alpha_{jk})}$
Terms in Eq. 8	$P(Y = k \vert \mathbf{W} = j, D_e, D_o, \mathcal{H}^{\overline{a}}_{\mathbf{Z}}\ ) = \frac{N_{e,j_b k} + \alpha_{j_b k}}{N_{e,j_b} + \alpha_{j_b}}$
Terms in Eq. 8	$P(Y = k \vert \mathbf{W} = j, D_e, D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \frac{N_{o,jk} + \alpha_{jk}}{N_{o,j} + \alpha_j}$, if $\mathbf{Z}_o \neq \emptyset$ $P(Y = k \vert \mathbf{W} = j, D_e, D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \frac{N_{o,jk} + N_{e,jk} + \alpha_{jk}}{N_{o,j} + N_{e,j} + \alpha_j}$, if $\mathbf{Z}_o = \emptyset$

adjustment set. In the future, we plan to exploit theoretical results in transporting causal effects across different populations Bareinboim and Pearl (2014) to extend our method to combine data from trials and observations that sample different populations.

D PROOFS

D.1 Proof of Convergence

In this section, we present the proof of Theorem 1 in the main paper. The theorem states that in the large sample limit, $P(\mathcal{H}^{a}_{\mathbf{Z}}\ |D_e, D_o)$ computed using Eq. 3 in the main paper will go to 1 if and only if $\mathbf{Z}$ is an adjustment set. The proof is for discrete data, where the marginal likelihood $P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ )$ can be computed in closed form using the BDE score. The proof proceeds as follows: In the first part, we prove that the equality of conditional distributions $P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z})$ holds only for adjustment sets under faithfulness: Theorem 4 shows that for faithful distributions, $P(Y|do(X), \mathbf{Z}) \neq P(Y|X, \mathbf{Z})$ if $\mathbf{Z}$ is not an adjustment sets. The implication is that equality holds only under $\mathcal{H}^{a}_{\mathbf{Z}}$, and inequality holds under $\mathcal{H}^{\overline{a}}_{\mathbf{Z}}$. Then, use the decomposition of the BDE score into a conditional entropy term and a complexity penalty term, to obtain the large sample approximation for $P(D_e|\mathcal{H}^{a}_{\mathbf{Z}}\ , D_o)$. The conditional entropy term involves the conditional entropy of the observational, experimental, and joint data. Lemma 7 proves an inequality result among these three conditional entropies, which is then used to show that in the large sample limit, Eq. 3 in the main paper goes to 1 if $P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z})$, and to 0 otherwise.

We first state our assumptions:

Assumptions A: Let D_o be an observational dataset measuring a discrete treatment X, a discrete outcome Y, and discrete pre-treatment covariates $\mathbf{V}$ that contains N_o cases that is sampled from distribution P, which is strictly positive as $N \to \infty$, and is a perfect map for an ADMG G. Also, let D_e be an experimental dataset measuring treatment X, outcome Y, and pre-treatment covariates $\mathbf{V}$ that contains N_e cases, where we assume N_o and N_e increase equally without limit.

Lemma 2. *Let Y, $\mathbf{Z}$, $\mathbf{W}$ be sets of discrete-valued variables. Let $\mathbf{w}$ denote a configuration of $\mathbf{W}$, and $\mathbf{w^c}$ denote all the remaining possible configurations of $\mathbf{W}$. If $P(Y|\mathbf{Z}, \mathbf{w}) = P(Y|\mathbf{Z}, \mathbf{w}^c)$ for all configurations $\mathbf{w}$ of $\mathbf{W}$, then $\mathbf{Y}$ is independent of $\mathbf{W}$ given $\mathbf{Z}$.*

Proof: This is straightforward from the rules of probability, i.e.

$$P(y|\mathbf{z}) = P(y|\mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}) + P(y|\mathbf{z}, \mathbf{w}^c)P(\mathbf{w}^c|\mathbf{z}) =$$

$$= P(y|\mathbf{z}, \mathbf{w})[P(\mathbf{w}|\mathbf{z}) + P(\mathbf{w}^c|\mathbf{z})] = P(y|\mathbf{z}, \mathbf{w})$$

Lemma 3. *Let X, Y be a treatment and an outcome variable such that $X \rightarrow Y$ in $\mathcal{G}$. Let* $\mathbf{Z}$, $\mathbf{W}$ *be pre-treatment covariates such that $\mathbf{Z} \cup \mathbf{W}$ is a adjustment set and $\mathbf{Z}$ is not a adjustment set. Then $\mathbf{W} \not\perp\!\!\!\perp Y|X, \mathbf{Z}$ in $\mathcal{G}$ and $\mathbf{W} \not\perp\!\!\!\perp X|\mathbf{Z}$ in $\mathcal{G}$.*

Proof: Since $\mathbf{Z}$ is not a adjustment set, but $\mathbf{Z} \cup \mathbf{W}$ is a adjustment set, there exists a path p from X to Y that is active given $\mathbf{Z}$ and blocked given $\mathbf{Z} \cup \mathbf{W}$. Thus, some $W \in \mathbf{W}$ is a non-collider on that path, hence $p = Xp_{XW}Wp_{WY}Y$.But then p_{XW}, p_{WY} are also active given $\mathbf{Z}$ and X,$\mathbf{Z}$ respectively, so $\mathbf{W} \not\perp\!\!\!\perp_{\mathcal{G}} Y|X, \mathbf{Z}$ in $\mathcal{G}$ and $\mathbf{W} \not\perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{Z}$ in $\mathcal{G}$.

Theorem 4. *Let $\mathcal{G}$ be a DAG and $\mathcal{P}$ be a distribution faithful to the DAG. Let $\mathbf{Z}$ be a set of pre-treatment covariates such that $\mathbf{Z}$ is not an adjustment set for Y with respect to X. Then $P(Y|do(X), \mathbf{Z}) \neq P(Y|X, \mathbf{Z})$.*

Proof. We will assume that $\mathcal{P}(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z})$ and show that it leads to contradiction. Let $\mathbf{Z}$ be a set of pre-treatment covariates such that $\mathbf{Z}$ is not an adjustment set for X and Y, but for which

$$P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z}). \tag{12}$$

Let $\mathbf{W}$ be a set such that $\mathbf{Z} \cup \mathbf{W}$ is an adjustment set ($\mathbf{W}$ may be unobserved in the ADMG $\mathcal{G}$, but exist in the underlying DAG). Then by rule 2 of do-calculus,

$$P(Y|do(X), \mathbf{Z}, \mathbf{W}) = P(Y|X, \mathbf{Z}, \mathbf{W}). \tag{13}$$

Moreover, by Lemma 3,

$$P(\mathbf{W}|X, \mathbf{Z}) \neq P(\mathbf{W}|\mathbf{Z}). \tag{14}$$

Let $\mathbf{W} = \mathbf{w}$ be a configuration of $\mathbf{W}$, and $\mathbf{w}^c$ denote the event that $\mathbf{W}$ does not take configuration $\mathbf{w}$. Then by Eq. 12 and the rule of total probability,

$$P(y|do(x), \mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}) + P(y|do(x), \mathbf{z}, \mathbf{w}^c)P(\mathbf{w}^c|\mathbf{z}) = P(y|x, \mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}, x) + P(y|x, \mathbf{z}, \mathbf{w}^c)P(\mathbf{w}^c|\mathbf{z}, x) \Leftrightarrow$$

$$P(y|x, \mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}) + P(y|x, \mathbf{z}, \mathbf{w}^c)P(\mathbf{w}^c|\mathbf{z}) = P(y|x, \mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}, x) + P(y|x, \mathbf{z}, \mathbf{w}^c)P(\mathbf{w}^c|\mathbf{z}, x) \Leftrightarrow$$

$$P(y|x, \mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}) + P(y|x, \mathbf{z}, \mathbf{w}^c)(1 - P(\mathbf{w}|\mathbf{z})) = P(y|x, \mathbf{z}, \mathbf{w})P(\mathbf{w}|\mathbf{z}, x) + P(y|x, \mathbf{z}, \mathbf{w}^c)(1 - P(\mathbf{w}|\mathbf{z}, x)) \Leftrightarrow$$

$$P(y|x, \mathbf{z}, \mathbf{w})[P(\mathbf{w}|\mathbf{z}) - P(\mathbf{w}|\mathbf{z}, x)] = P(y|x, \mathbf{z}, \mathbf{w}^c)(1 - P(\mathbf{w}|\mathbf{z}, x) - 1 + P(\mathbf{w}|\mathbf{z})) \Leftrightarrow$$

$$P(y|x, \mathbf{z}, \mathbf{w}) = P(y|x, \mathbf{z}, \mathbf{w}^c)$$

But then by Lemma 2 and faithfulness, $Y \perp\!\!\!\perp \mathbf{W}|X, \mathbf{Z}$ which is a contradiction based on Lemma 3. Hence, $P(Y|do(X), \mathbf{Z}) \neq P(Y|X, \mathbf{Z})$. □

We now present the definition of conditional entropy, which we will need in our proof:

Definition 1 (Conditional Entropy)**.** *Let P be the full joint probability distribution over a set of variables $\boldsymbol{V}$, let $Y \in \boldsymbol{V}$ be a variable, and let $\mathbf{Z} \subseteq \boldsymbol{V} \setminus \{Y\}$ be a set of variables. Then, the conditional entropy of Y given $\mathbf{Z}$ is defined as follows Cover (1999):*

$$H(Y|\mathbf{Z}) = -\sum_{y}\sum_{z} P(y, z) \cdot \log P(y|z) \tag{15}$$

where y and z denote the values of Y and $\mathbf{Z}$, respectively.

We can now derive large sample approximations for $P(D_e|\mathcal{H}^{a}_{\mathbf{Z}}\ , D_o)$ and $P(D_e|\mathcal{H}^{\overline{a}}_{\mathbf{Z}}\ , D_o)$. In the rest of this document, we will use $\mathbf{W}$ to denote the union of the candidate adjustment set $\mathbf{Z}$ with the treatment X: $\mathbf{W} = \mathbf{Z} \cup X$.

Lemma 5. *Given **Assumptions A**, the BD score for $\log P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ )$ in the large sample limit is defined as follows:*

$$\begin{aligned} \lim_{N \to \infty} \log P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \\ \lim_{N \to \infty} -(N_o + N_e) \cdot H_{o,e}(Y|\mathbf{W}) + N_o \cdot H_o(Y|\mathbf{W}) \\ - \frac{q(r-1)}{2}\left[\log(N_o + N_e) - \log N_o\right] + const, \end{aligned} \tag{16}$$

where $\mathbf{W} = \mathbf{Z} \cup X$.

Proof. The BD score for $P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\ )$ is calculated as follows Heckerman et al. (1995):

$$P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\ ) = \prod_{j=1}^{q} \frac{\Gamma(\alpha_j + N_{o,j})}{\Gamma(\alpha_j + N_{o,j} + N_{e,j})} \cdot \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N_{o,jk} + N_{e,jk})}{\Gamma(\alpha_{jk} + N_{o,jk})}\,, \tag{17}$$

where q denotes instantiations of variables in $\mathbf{W}$ and r denotes values of variable Y. The term $N_{e,jk}$ is the number of cases in D_e in which variable $Y = k$ and $\mathbf{W} = j$; also, $N_{e,j} = \sum_{k=1}^{r} N_{e,jk}$. The term $N_{o,jk}$ is the number of cases in D_o in which variable $Y = k$ and $\mathbf{W} = j$; also, $N_{o,j} = \sum_{k=1}^{r} N_{o,jk}$. The term α_{jk} is a finite positive real number that is called the Dirichlet prior parameter and may be interpreted as representing "pseudo-counts", where $\alpha_j = \sum_{k=1}^{r} \alpha_{jk}$. BD can be re-written in log form as follows:

$$\begin{aligned}\log P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\ ) = \sum_{j=1}^{q} \Bigg[& \log\Gamma(\alpha_j + N_{o,j}) - \log\Gamma(\alpha_j + N_{o,j} + N_{e,j}) \\ & + \sum_{k=1}^{r} [\log\Gamma(\alpha_{jk} + N_{o,jk} + N_{e,jk}) - \log\Gamma(\alpha_{jk} + N_{o,jk})] \Bigg]. \end{aligned} \tag{18}$$

We can re-arrange the terms in Equation (18) to gather the terms as follows:

$$\begin{aligned}\log P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\ ) = & \sum_{j=1}^{q} \left[-\log\Gamma(\alpha_j + N_{o,j} + N_{e,j}) + \sum_{k=1}^{r} \log\Gamma(\alpha_{jk} + N_{o,jk} + N_{e,jk}) \right] \\ & + \sum_{j=1}^{q} \left[\log\Gamma(\alpha_j + N_{o,j}) - \sum_{k=1}^{r} \log\Gamma(\alpha_{jk} + N_{o,jk}) \right] \end{aligned} \tag{19}$$

Using the Stirling's approximation of $\lim_{n\to\infty} \log\Gamma(n) = (n - \frac{1}{2}) \log(n) - n + const.$, we can re-write Equation (19) as follows:

$$\begin{aligned} & \lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\ ) = \\ & \lim_{N\to\infty} \sum_{j=1}^{q} \Bigg[-(\alpha_j + N_{o,j} + N_{e,j} - \frac{1}{2}) \log(\alpha_j + N_{o,j} + N_{e,j}) + (\alpha_j + N_{o,j} + N_{e,j}) \\ & + \sum_{k=1}^{r} \Big((\alpha_{jk} + N_{o,jk} + N_{e,jk} - \frac{1}{2}) \log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) - (\alpha_{jk} + N_{o,jk} + N_{e,jk}) \Big) \Bigg] \\ & + \sum_{j=1}^{q} \Bigg[(\alpha_j + N_{o,j} - \frac{1}{2}) \log(\alpha_j + N_{o,j}) - (\alpha_j + N_{o,j}) \\ & + \sum_{k=1}^{r} -(\alpha_{jk} + N_{o,jk} - \frac{1}{2}) \log(\alpha_{jk} + N_{o,jk}) + (\alpha_{jk} + N_{o,jk}) \Bigg] \end{aligned} \tag{20}$$

$$
\begin{aligned}
&= \lim_{N\to\infty} \sum_{j=1}^{q} \Bigg[-\alpha_j \log(\alpha_j + N_{o,j} + N_{e,j}) - N_{o,j}\log(\alpha_j + N_{o,j} + N_{e,j}) - N_{e,j}\log(\alpha_j + N_{o,j} + N_{e,j}) \\
&+ \frac{1}{2}\log(\alpha_j + N_{o,j} + N_{e,j}) + \alpha_j + N_{o,j} + N_{e,j} \\
&+ \sum_{k=1}^{r} \Bigg(\alpha_{jk}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) + N_{o,jk}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) + N_{e,jk}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) \\
&- \frac{1}{2}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) - \alpha_{jk} - N_{o,jk} - N_{e,jk} \Bigg) \Bigg] \\
&+ \sum_{j=1}^{q} \Bigg[\alpha_j \log(\alpha_j + N_{o,j}) + N_{o,j}\log(\alpha_j + N_{o,j}) - \frac{1}{2}\log(\alpha_j + N_{o,j}) - \alpha_j - N_{o,j} \\
&+ \sum_{k=1}^{r} -\alpha_{jk}\log(\alpha_{jk} + N_{o,jk}) - N_{o,jk}\log(\alpha_{jk} + N_{o,jk}) + \frac{1}{2}\log(\alpha_{jk} + N_{o,jk}) + \alpha_{jk} + N_{o,jk} \Bigg]
\end{aligned}
$$

The terms not involving a $\log$ term cancel out; then, we used the facts that $\sum_{k=1}^{r} N_{o,jk} = N_{o,j}$, $\sum_{k=1}^{r} N_{e,jk} = N_{e,j}$, and $\sum_{k=1}^{r} \alpha_{jk} = \alpha_j$ and re-arranged the remaining terms to obtain:

$$
\begin{aligned}
&\lim_{N\to\infty} \log P(D_e | D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \\
&\lim_{N\to\infty} \sum_{j=1}^{q} \Bigg[-N_{o,j}\log(\alpha_j + N_{o,j} + N_{e,j}) + \sum_{k=1}^{r} N_{o,jk}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) \Bigg] \\
&+ \sum_{j=1}^{q} \Bigg[-N_{e,j}\log(\alpha_j + N_{o,j} + N_{e,j}) + \sum_{k=1}^{r} N_{e,jk}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) \Bigg] \\
&+ \sum_{j=1}^{q} \Bigg[-\alpha_j\log(\alpha_j + N_{o,j} + N_{e,j}) + \sum_{k=1}^{r} \alpha_{jk}\log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) \Bigg] \\
&+ \sum_{j=1}^{q} \Bigg[N_{o,j}\log(\alpha_j + N_{o,j}) - \sum_{k=1}^{r} N_{o,jk}\log(\alpha_{jk} + N_{o,jk}) \Bigg] \\
&+ \sum_{j=1}^{q} \Bigg[\alpha_j\log(\alpha_j + N_{o,j}) - \sum_{k=1}^{r} \alpha_{jk}\log(\alpha_{jk} + N_{o,jk}) \Bigg] \\
&+ \frac{1}{2}\sum_{j=1}^{q} \Bigg[\log(\alpha_j + N_{o,j} + N_{e,j}) - \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk} + N_{e,jk}) \Bigg] \\
&+ \frac{1}{2}\sum_{j=1}^{q} \Bigg[-\log(\alpha_j + N_{o,j}) + \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk}) \Bigg]
\end{aligned}
\tag{21}
$$

We can apply the identities mentioned above again to Equation (21) to obtain the following:

$$\begin{aligned}
&\lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \\
&\lim_{N\to\infty} \sum_{j=1}^{q}\sum_{k=1}^{r}\Bigg[N_{o,jk}\log(\frac{\alpha_{jk}+N_{o,jk}+N_{e,jk}}{\alpha_j+N_{o,j}+N_{e,j}}) + N_{e,jk}\log(\frac{\alpha_{jk}+N_{o,jk}+N_{e,jk}}{\alpha_j+N_{o,j}+N_{e,j}}) + \alpha_{jk}\log(\frac{\alpha_{jk}+N_{o,jk}+N_{e,jk}}{\alpha_j+N_{o,j}+N_{e,j}})\Bigg] \\
&-\lim_{N\to\infty} \sum_{j=1}^{q}\sum_{k=1}^{r}\Bigg[N_{o,jk}\log(\frac{\alpha_{jk}+N_{o,jk}}{\alpha_j+N_{o,j}}) + \alpha_{jk}\log(\frac{\alpha_{jk}+N_{o,jk}}{\alpha_j+N_{o,j}})\Bigg] \\
&+\frac{1}{2}\sum_{j=1}^{q}\Bigg[\log(\alpha_j+N_{o,j}+N_{e,j}) - \sum_{k=1}^{r}\log(\alpha_{jk}+N_{o,jk}+N_{e,jk})\Bigg] \\
&+\frac{1}{2}\sum_{j=1}^{q}\Bigg[-\log(\alpha_j+N_{o,j}) + \sum_{k=1}^{r}\log(\alpha_{jk}+N_{o,jk})\Bigg]
\end{aligned} \tag{22}$$

Given that

$$\lim_{N\to\infty} \frac{\alpha_{jk}+N_{o,jk}+N_{e,jk}}{\alpha_j+N_{o,j}+N_{e,j}} = \frac{N_{o,jk}+N_{e,jk}}{N_{o,j}+N_{e,j}},$$

$$\lim_{N\to\infty} \frac{\alpha_{jk}+N_{o,jk}}{\alpha_j+N_{o,j}} = \frac{N_{o,jk}}{N_{o,j}},$$

$$\lim_{N\to\infty} \sum_{j=1}^{q}\sum_{k=1}^{r}\alpha_{jk}\log(\frac{\alpha_{jk}+N_{o,jk}+N_{e,jk}}{\alpha_j+N_{o,j}+N_{e,j}}) = \mathit{const.},$$

and

$$\lim_{N\to\infty} \sum_{j=1}^{q}\sum_{k=1}^{r}\alpha_{jk}\log(\frac{\alpha_{jk}+N_{o,jk}}{\alpha_j+N_{o,j}}) = \mathit{const.},$$

in the limit, Equation (22) becomes:

$$\begin{aligned}
&\lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \\
&\lim_{N\to\infty} \sum_{j=1}^{q}\sum_{k=1}^{r}(N_{o,jk}+N_{e,jk})\log\frac{N_{o,jk}+N_{e,jk}}{N_{o,j}+N_{e,j}} - \sum_{j=1}^{q}\sum_{k=1}^{r}N_{o,jk}\log\frac{N_{o,jk}}{N_{o,j}} \\
&+\frac{1}{2}\sum_{j=1}^{q}\Bigg[\log(\alpha_j+N_{o,j}+N_{e,j}) - \sum_{k=1}^{r}\log(\alpha_{jk}+N_{o,jk}+N_{e,jk})\Bigg] \\
&+\frac{1}{2}\sum_{j=1}^{q}\Bigg[-\log(\alpha_j+N_{o,j}) + \sum_{k=1}^{r}\log(\alpha_{jk}+N_{o,jk})\Bigg]+\mathit{const.},
\end{aligned} \tag{23}$$

or equivalently:

$$\begin{aligned}
&\lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\,) = \\
&\lim_{N\to\infty} (N_o + N_e)\cdot \sum_{j=1}^{q}\sum_{k=1}^{r} \frac{N_{o,jk}+N_{e,jk}}{(N_o+N_e)} \log \frac{N_{o,jk}+N_{e,jk}}{N_{o,j}+N_{e,j}} - N_o \cdot \sum_{j=1}^{q}\sum_{k=1}^{r} \frac{N_{o,jk}}{N_o} \log \frac{N_{o,jk}}{N_{o,j}} \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[\log(\alpha_j + N_{o,j} + N_{e,j}) - \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk} + N_{e,jk})\right] \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[-\log(\alpha_j + N_{o,j}) + \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk})\right] + const. \\
&= \lim_{N\to\infty} -(N_o+N_e)\cdot H_{o,e}(Y|\mathbf{W}) + N_o \cdot H_o(Y|\mathbf{W}) \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[\log(\alpha_j + N_{o,j} + N_{e,j}) - \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk} + N_{e,jk})\right] \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[-\log(\alpha_j + N_{o,j}) + \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk})\right] + const.
\end{aligned} \tag{24}$$

where $H(.)$ terms denote conditional entropies.

To simplify the last two terms in Equation (24), we perform the following transformations:

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{2}\sum_{j=1}^{q}\left[\log(\alpha_j + N_{o,j} + N_{e,j}) - \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk} + N_{e,jk})\right] \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[-\log(\alpha_j + N_{o,j}) + \sum_{k=1}^{r} \log(\alpha_{jk} + N_{o,jk})\right] \\
&= \lim_{N\to\infty} \frac{1}{2}\sum_{j=1}^{q}\Bigg[\log(\frac{\alpha_j + N_{o,j} + N_{e,j}}{N_o+N_e}) + \log(N_o+N_e) \\
&- \sum_{k=1}^{r} \log(\frac{\alpha_{jk} + N_{o,jk} + N_{e,jk}}{N_o+N_e}) + \log(N_o+N_e)\Bigg] \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[-\log(\frac{\alpha_j + N_{o,j}}{N_o}) - \log N_o + \sum_{k=1}^{r} \log(\frac{\alpha_{jk} + N_{o,jk}}{N_o}) + \log N_o\right] \\
&= \lim_{N\to\infty} \frac{1}{2}\sum_{j=1}^{q}\left(\log(N_o+N_e) - \sum_{k=1}^{r}\log(N_o+N_e)\right) \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left(-\log N_o + \sum_{k=1}^{r}\log N_o\right) \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[\log(\frac{\alpha_j + N_{o,j} + N_{e,j}}{N_o+N_e}) - \sum_{k=1}^{r} \log(\frac{\alpha_{jk} + N_{o,jk} + N_{e,jk}}{N_o+N_e})\right] \\
&+ \frac{1}{2}\sum_{j=1}^{q}\left[-\log(\frac{\alpha_j + N_{o,j}}{N_o}) + \sum_{k=1}^{r} \log(\frac{\alpha_{jk} + N_{o,jk}}{N_o})\right] \\
&= -\frac{q(r-1)}{2}\log(N_o+N_e) + \frac{q(r-1)}{2}\log N_o + const. \\
&= -\frac{q(r-1)}{2}\left[\log(N_o+N_e) - \log N_o\right] + const.
\end{aligned} \tag{25}$$

Combining Equations (24) and (25), we obtain:

$$\begin{aligned}
&\lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\ ) = \\
&\qquad \lim_{N\to\infty} -(N_o + N_e)\cdot H_{o,e}(Y|\mathbf{W}) + N_o \cdot H_o(Y|\mathbf{W}) \\
&\qquad - \frac{q(r-1)}{2}\left[\log(N_o+N_e) - \log N_o\right] + const.
\end{aligned} \tag{26}$$

□

Lemma 6. *Given* ***Assumptions A****,* $\lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\ )$ *is defined as follows:*

$$\begin{aligned}
&\lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\ ) = \\
&\qquad \lim_{N\to\infty} -N_e \cdot H_e(Y|\mathbf{W}) - \frac{q(r-1)}{2}\log N_e + const.
\end{aligned} \tag{27}$$

where

$$\lim_{N\to\infty} P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\ ) = \prod_{j=1}^{q} \frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j + N_{e,j})} \cdot \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N_{e,jk})}{\Gamma(\alpha_{jk})} \tag{28}$$

Proof. Proof is similar to the proof of Lemma 5. □

Lemma 7. *Let* $P_o, P_e, P_{o,e}$ *denote the frequentist distribution in the observational, experimental, and joint data, respectively. Also, let* $\mathbf{Z} \subset \mathbf{V}$ *be a subset of variables and* $\mathbf{W} = \mathbf{Z} \cup X$*. Then, when* $N \to \infty$

$$2H_{o,e}(Y|\mathbf{W}) \geq H_o(Y|\mathbf{W}) + H_e(Y|\mathbf{W}) \tag{29}$$

or equivalently

$$2H(P_{o,e}(Y|\mathbf{W})) \geq H(P_o(Y|\mathbf{W})) + H(P_e(Y|\mathbf{W})), \tag{30}$$

where the equality in Equation (30) holds if and only if $\mathcal{H}^{a}_{\mathbf{Z}}$ *is true.*

Proof. Let $P_o(\mathbf{W}=j), P_e(\mathbf{W}=j), P_{o,e}(\mathbf{W}=j)$ denote the frequentist probabilities of $\mathbf{W}=j$ in the observational, experimental, and joint data, respectively. Also, let $P_o(Y|\mathbf{W}=j), P_e(Y|\mathbf{W}=j), P_{o,e}(Y|\mathbf{W}=j)$ be the frequentist conditional probabilities of Y given $\mathbf{W}=j$ in the observational, experimental and joint data, respectively. We use $N_{o,jk}$, $N_{e,jk}$ and $N_{o,e,jk}$ to denote the number of cases where $Y=k$ and $\mathbf{W}=j$ in the observational, experimental, and joint data, respectively. We use $N_{o,j}$, $N_{e,j}$ and $N_{o,e,j}$ denote the number of cases where $\mathbf{W}=j$ in the observational, experimental, and joint data, respectively. Hence the following hold:

$$P_o(\mathbf{W}=j) = \frac{N_{o,j}}{N_o}, \quad P_e(\mathbf{W}=j) = \frac{N_{e,j}}{N_e}, \quad P_{o,e}(\mathbf{W}=j) = \frac{N_{o,e,j}}{N_{o,e}}$$

$$P_o(Y=k|\mathbf{W}=j) = \frac{N_{o,jk}}{N_{o,j}}, \quad P_e(Y=k|\mathbf{W}=j) = \frac{N_{e,jk}}{N_{e,j}}, \quad P_{o,e}(Y=k|\mathbf{W}=j) = \frac{N_{o,e,jk}}{N_{o,e,j}}$$

By Assumptions A, we assume that in the limit, $N_o = N_e \coloneqq N$. Then, for each $\mathbf{W}=j$,

$$\begin{aligned}
&lim_{N\to\infty} P_{o,e}(\mathbf{W}=j) = lim_{N\to\infty} \frac{N_{o,j}+N_{e,j}}{N_o+N_e} = \\
&lim_{N\to\infty} (\frac{N_{o,j}}{2N} + \frac{N_{o,j}}{2N}) = lim_{N\to\infty} \frac{1}{2}\left(P_o(\mathbf{W}=j) + P_e(\mathbf{W}=j)\right),
\end{aligned} \tag{31}$$

Additionally, we can derive $P_{o,e}(Y|\mathbf{W}=j)$ as follows

$$\begin{aligned}
&P_{o,e}(Y|\mathbf{W}=j) = \frac{N_{o,jk}+N_{e,jk}}{N_{o,j}+N_{e,j}} = \frac{N_{o,jk}}{N_{o,j}+N_{e,j}} + \frac{N_{e,jk}}{N_{o,j}+N_{e,j}} = \\
&\frac{N_{o,j}}{N_{o,j}}\frac{N_{o,jk}}{N_{o,j}+N_{e,j}} + \frac{N_{e,j}}{N_{e,j}}\frac{N_{e,jk}}{N_{o,j}+N_{e,j}} = \frac{N_{o,j}}{N_{o,j}+N_{e,j}}\frac{N_{o,jk}}{N_{o,j}} + \frac{N_{e,j}}{N_{o,j}+N_{e,j}}\frac{N_{e,jk}}{N_{e,j}},
\end{aligned} \tag{32}$$

where in line 2 we first multiply each fraction with either $\frac{N_{o,j}}{N_{o,j}}$ or $\frac{N_{e,j}}{N_{e,j}}$ and then switch the order in the denominators in each part of the sum. By dividing numerators and denominators in Eq. 32 with N we get

$$P_{o,e}(Y|\mathbf{W}=j) = \frac{P_o(\mathbf{W}=j)}{P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j)} P_o(Y|\mathbf{W}=j) + \frac{P_e(\mathbf{W}=j)}{P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j)} P_e(Y|\mathbf{W}=j), \tag{33}$$

for every j. The final formula reflects the fact that $P_{o,e}(Y|\mathbf{W}=j)$ is a mixture of $P_o(Y|\mathbf{W}=j)$ and $P_e(Y|\mathbf{W}=j)$, with proportions $\frac{P_o(\mathbf{W}=j)}{P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j)}$ and $\frac{P_e(\mathbf{W}=j)}{P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j)}$.

Given that entropy is a concave function, the following hold for probability mass functions p_1, p_2

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2), \tag{34}$$

where inequality is strict if $p_1 \neq p_2$Cover (1999) (page 32).

For each $\mathbf{W}=j$, we define λ_j to be the proportion of observational data where $\mathbf{W}=j$ to the proportion of the joint data where $\mathbf{W}=j : \lambda_j = \frac{P_o(\mathbf{W}=j)}{P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j)}, 1-\lambda_j = \frac{P_e(\mathbf{W}=j)}{P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j)}$. Using $p_1 = P_o(Y|\mathbf{W}=j), p_2 = P_e(Y|\mathbf{W}=j)$, and λ_j as defined above in Equation (34), we can write

$$\begin{split} H(\lambda_j P_o(Y|\mathbf{W}=j) + (1-\lambda_j)P_e(Y|\mathbf{W}=j)) \geq \\ \lambda_j H(P_o(Y|\mathbf{W}=j)) + (1-\lambda_j)H(P_e(Y|\mathbf{W}=j)) \end{split} \tag{35}$$

where the right-hand size is equal to $H(P_{o,e}(Y|\mathbf{W}=j))$. We can multiply both sides with $\frac{1}{2}(P_o(\mathbf{W}=j)+P_e(\mathbf{W}=j))$ to obtain the following

$$\begin{split} &P_{o,e}(\mathbf{W}=j)H(P_{o,e}(Y|\mathbf{W}=j)) \geq \\ &\frac{1}{2}P_o(\mathbf{W}=j)H(P_o(Y|\mathbf{W}=j)) + \frac{1}{2}P_e(\mathbf{W}=j)H(P_e(Y|\mathbf{W}=j)), \end{split} \tag{36}$$

which can be re-written as follows by multiplying both sides by 2

$$\begin{split} &2P_{o,e}(\mathbf{W}=j)H(P_{o,e}(Y|\mathbf{W}=j)) \geq \\ &P_o(\mathbf{W}=j)H(P_o(Y|\mathbf{W}=j)) + P_e(\mathbf{W}=j)H(P_e(Y|\mathbf{W}=j)). \end{split} \tag{37}$$

We then sum over all possible j's to obtain

$$\begin{split} &2\sum_j P_{o,e}(\mathbf{W}=j)H(P_{o,e}(Y|\mathbf{W}=j)) \geq \\ &\sum_j P_o(\mathbf{W}=j)H(P_o(Y|\mathbf{W}=j)) + \sum_j P_e(\mathbf{W}=j)H(P_e(Y|\mathbf{W}=j)), \end{split} \tag{38}$$

where each sum term is the definition of the conditional entropy as given in the following equations:

$$H(P_{o,e}(Y|\mathbf{W})) = \sum_j P_{o,e}(\mathbf{W}=j)H(P_{o,e}(Y|\mathbf{W}=j))$$

$$H(P_o(Y|\mathbf{W})) = \sum_j P_o(\mathbf{W}=j)H(P_o(Y|\mathbf{W}=j))$$

$$H(P_e(Y|\mathbf{W})) = \sum_j P_e(\mathbf{W}=j)H(P_e(Y|\mathbf{W}=j)).$$

Therefore, we can re-write Equation (38) as follows:

$$2H(P_{o,e}(Y|\mathbf{W})) \geq H(P_o(Y|\mathbf{W})) + H(P_e(Y|\mathbf{W})). \tag{39}$$

Moreover, under $\mathcal{H}^a_{\mathbf{z}}$, $P_o(Y|\mathbf{W}) = P_e(Y|\mathbf{W}) = P_{o,e}(Y|\mathbf{W})$ when $N \to \infty$ and equality holds, while under $\mathcal{H}^{\overline{a}}_{\mathbf{z}}$ $P_o(Y|\mathbf{W}) \neq P_e(Y|\mathbf{W})$ and the inequality is strict. □

We can now prove our main theorem:

Theorem 1. *Let D_o, D_e be an observational data set and an experimental data set, respectively, both measuring treatment X, outcome Y, and pre-treatment covariates $\mathbf{V}$, all discrete. Let D_o, D_e contain N_o, N_e cases respectively, sampled from distributions $\mathcal{P}, \mathcal{P}_{\overline{X}}$ respectively, both strictly positive in the sample limit. Also, let $\mathcal{P}$ be a perfect map for an ADMG $\mathcal{G}$. We assume N_o and N_e increase equally without limit. Then the proposed method converges to the data-generating model in the large sample limit:*

$$\lim_{N\to\infty} P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o, D_e) = \begin{cases} 1, & \textit{if } \mathbf{Z} \textit{ is an adjustment} \\ & \textit{set for } X \textit{ and } Y \\ 0, & \textit{otherwise} \end{cases}$$

Proof. For a set $\mathbf{Z}$, let $\mathbf{W} = \mathbf{Z} \cup X$. We have that

$$\lim_{N\to\infty} P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o, D_e) \lim_{N\to\infty} = \frac{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o) + P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)P(\mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,|D_o)}. \tag{40}$$

By inverting Equation (40), and for $P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o) = 1/2$ we obtain the following:

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o, D_e)} = \lim_{N\to\infty} \frac{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,) + P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)} &= \\ 1 + \lim_{N\to\infty} (\frac{P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)}) &= \\ 1 + \lim_{N\to\infty} \exp(\log \frac{P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)}) & \end{aligned} \tag{41}$$

Using Equations (26) and (27), we obtain $log(\frac{P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)})$ in the large sample limit as follows:

$$\begin{aligned} &\lim_{N\to\infty} log(\frac{P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)}) = \lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,) - \lim_{N\to\infty} \log P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,) \\ &= \lim_{N\to\infty} -N_e \cdot H_e(Y|\mathbf{W}) + (N_o + N_e) \cdot H_{o,e}(Y|\mathbf{W}) - N_o \cdot H_o(Y|\mathbf{W}) \\ &- \frac{q(r-1)}{2} \log N_e + \frac{q(r-1)}{2} [\log(N_o + N_e) - \log N_o] + const. \\ &= \lim_{N\to\infty} N \cdot [-H_e(Y|\mathbf{W}) + 2H_{o,e}(Y|\mathbf{W}) - H_o(Y|\mathbf{W})] \\ &- \frac{(r-1)}{2}(q \log N - q \log 2) + const \\ &= \lim_{N\to\infty} N \cdot [-H_e(Y|\mathbf{W}) + 2H_{o,e}(Y|\mathbf{W}) - H_o(Y|\mathbf{W})] \\ &- \frac{q(r-1)}{2}(\log \frac{N}{2}) + const. \end{aligned} \tag{42}$$

where the last step is possible since $N_e = N_o \coloneqq N$.

If $\mathbf{Z}$ is an adjustment set, it follows from Lemma 7 that

$$H_{o,e}(Y|\mathbf{W}) = H_o(Y|\mathbf{W}) = H_e(Y|\mathbf{W});$$

therefore

$$\lim_{N\to\infty} log(\frac{P(D_e|D_o, \mathcal{H}^{\bar{a}}_{\mathbf{Z}}\,)}{P(D_e|D_o, \mathcal{H}^{a}_{\mathbf{Z}}\,)}) = \lim_{N\to\infty} -\frac{q(r-1)}{2}(\log \frac{N}{2}) + const = -\infty \tag{43}$$

Hence by Eq. 41,

$$\lim_{N\to\infty} \frac{1}{P(\mathcal{H}^{a}_{\mathbf{Z}}\,|D_o, D_e)} \to 1$$

and therefore $P(\mathcal{H}^a_{\mathbf{Z}}\ |D_o, D_e)$ goes to 1 as N goes to infinity.

If $\mathbf{Z}$ is not an adjustment set, then by Lemma 7, when $N \to \infty$

$$-H_e(Y|\mathbf{W}) + 2H_{o,e}(Y|\mathbf{W}) - H_o(Y|\mathbf{W}) > 0$$

and therefore

$$\lim_{N\to\infty} N \cdot [-H_e(Y|\mathbf{W}) + 2H_{o,e}(Y|\mathbf{W}) - H_o(Y|\mathbf{W})] = \infty.$$

Notice that this term is $O(N)$ and will dominate the second term, $-\frac{q(r-1)}{2}\log\frac{N}{2}$. Therefore

$$\lim_{N\to\infty} log(\frac{P(D_e|D_o, \mathcal{H}^{\overline{a}}_{\mathbf{Z}}\ )}{P(D_e|D_o, \mathcal{H}^a_{\mathbf{Z}}\ )}) = \infty, \tag{44}$$

and by Eq. 40

$$\lim_{N\to\infty} \frac{1}{P(\mathcal{H}^a_{\mathbf{Z}}\ |D_o, D_e)} = \infty,$$

thus $P(\mathcal{H}^a_{\mathbf{Z}}\ |D_o, D_e)$ goes to 0 as N goes to infinity.

□