Confounding Feature Acquisition for Causal Effect Estimation

Shirly Wang[*]   SHIRLY@LAYER6.AI
Seung Eun Yi[*]   SEUNGEUN@LAYER6.AI
Layer 6 AI, TD Bank Group
Shalmali Joshi[†]   SHALMALI@SEAS.HARVARD.EDU
Harvard University (SEAS)
Marzyeh Ghassemi   MARZYEH@CS.TORONTO.EDU
University of Toronto, Vector Institute

Abstract

Reliable treatment effect estimation from observational data depends on the availability of all confounding information. While much work has targeted treatment effect estimation from observational data, there is relatively little work in the setting of confounding variable missingness, where collecting more information on confounders is often costly or time-consuming. In this work, we frame this challenge as a problem of feature acquisition of confounding features for causal inference. Our goal is to prioritize acquiring values for a fixed and known subset of missing confounders in samples that lead to efficient average treatment effect estimation. We propose two acquisition strategies based on i) covariate balancing (`CB`), and ii) reducing statistical estimation error on observed factual outcome error (`OE`). We compare `CB` and `OE` on five common causal effect estimation methods, and demonstrate improved sample efficiency of `OE` over baseline methods under various settings. We also provide visualizations for further analysis on the difference between our proposed methods.

---

[*] Equal contribution
[†] Work done while at the Vector Institute

Keywords: Treatment effect, Casual inference, Active learning, Feature acquisition

1. Introduction

Reliable causal effect estimation (CEE) from observational data is an important step toward advancing healthcare and science, and much recent work has targeted this problem (Alaa and van der Schaar, 2017; Johansson et al., 2016; Shalit et al., 2017). However, there are several practical challenges in observational health settings. First, reliable CEE hinges on observing all confounding attributes, e.g., attributes like race that may affect both treatment and outcome. (Pearl, 2000; Madras et al., 2019; Zhang and Bareinboim, 2018) Nevertheless, in many cases not all confounders may be known (Miao et al., 2018; Pearl, 2000; Rubin, 1974), making the effect estimate unidentifiable without additional information and/or constraints (Pearl, 2000). Further, even if all confounders are known, it may be difficult to collect their values due to high costs to the institution, and/or potential loss of confidentiality (Zhang and Bareinboim, 2019). For example, while collecting everyday life variables like diet, and physical exercise is costly, they are likely to affect patients with Alzheimer's disease and

can influence the evaluation of drug-disease interactions among these patients (Liyanage et al., 2018). When such variables cannot be collected, a proxy model can be built to impute missing values, albeit at the expense of statistical biases and instabilites. (Chen et al., 2019).

Our work focuses on accurately estimating treatment effect through feature acquisition in the presence of missing confounding attributes. In our case, values for a known and fixed subset of confounding variables are missing in most samples, but a costly mechanism can be deployed to acquire them. This is a relevant setting in healthcare as some observational clinical datasets used for causal studies recruit a cohort, and then acquire additional data for samples in this cohort. For example, the UK Biobank (Biobank, 2014) recently collected and released COVID-related data from a subset of their established cohort of participants. Our work is useful for prioritizing the acquisition of data values (from fixed cohorts) that are most beneficial for CEE. Note that our formulation is distinct from latent confounding (Pearl, 2000), where confounding variables are never observed. It also differs from conventional active learning or feature acquisition (Settles, 2009), where data is sampled for supervised learning. While past work has tried to address data acquisition for CEE, their focus is to acquire *counterfactual* outcomes for an observed sample (from an expert) under a treatment different from what exists in the observational data (without missingness) (Sundin et al., 2019). In contrast, our work focuses on acquisition of missing values in known confounders as opposed to counterfactual labels to obtain reliable estimates of average treatment effects (ATE)[1], when values of some confounding variables is available for only a small subset of samples.

We propose two acquisition strategies based on i) covariate balancing (CB) between treated and control groups, and ii) statistical estimation error to obtain missing values of known confounders for efficient CEE (OE). Using a semi-synthetic dataset, we compare the effectiveness of these strategies in reducing the need for labelling confounding values to estimate accurate ATE. Our experiments involve multiple commonly used CEE models, and multiple simulation settings where information from the missing confounder can be partially recovered through other covariates. We analyze the explore-exploit trade-off of both methods, visualizing samples acquired using PCA, and find that OE have more exploratory bias than CB. We further quantify treated vs control sample preference, finding that OE prefers to acquire samples to minimize errors for the control group as they have a more complex response function, while CB maintains a more balanced feature acquisition ratio over groups.

Our contributions are summarized as follows[2]:

- We propose two feature acquisition strategies to acquire missing values of known confounders for efficient CEE. Both strategies are compatible with common CEE models.

- We demonstrate that when the missing feature is independent of other covariates, OE achieves the best sample efficiency when used with most CEE models, and CB offers some initial benefits in reducing ATE errors. The benefits of OE persist even when the missing feature becomes more correlated with other covariates.

- We provide insights into the impact of different strategies on samples acquired and show that focusing on statistical estimation error prioritizes early exploration during

---

1. See Definition 2.

2. Code at https://github.com/MLforHealth/confounder-acquisition

acquisition and adapts to different complexities of the outcome generating function for the treated and control group.

2. Related Work

Reliable CEE from observational data is a classical statistical challenge addressed under two main frameworks of potential outcomes and causal graphical models (Pearl, 2000; Rubin, 1974). Recent works in modern machine learning focus on CEE using parametric assumptions or balancing approaches to improve effect estimates from observational data (Alaa and van der Schaar, 2017; Johansson et al., 2016; Shalit et al., 2017). A majority of such methods are developed assuming no hidden confounding or relaxing them under certain conditions (Louizos et al., 2017; Wang and Blei, 2019; Miao et al., 2018). However, very few methods have considered the problem of missingness in known confounding variables from the perspective of data acquisition or active learning.

Data acquisition in machine learning is designed for supervised learning to obtain target labels, with common strategies including query by committee (Freund et al., 1997), uncertainty sampling (Lewis and Gale, 1994), and information-based loss functions (Settles, 2009). Data acquisition of features has also been explored for supervised learning (Melville et al., 2005; Saar-Tschansky et al., 2009; Shim et al., 2018; Janisch et al., 2020). However, applicability of such methods for acquiring pre-treatment confounding variables is not well studied. Particularly, while motivations in supervised learning are simply to fit the associational distribution accurately, CEE is invalid without controlling for all confounding.

In the context of causality, active learning is primarily used for experiment design for causal discovery and obtaining interventional information (He and Geng, 2008; Yan et al., 2019). In our work, we assume that the underlying causal dependencies are already known (and correspond to Figure 1) and our goal is to improve statistical estimates of causal effects by strategically acquiring costly but missing confounding features without which estimation is impossible.

3. Methodology

We build on the potential outcomes framework of Rubin (1974) for causal inference from observational data. We formalize our setup in the following.
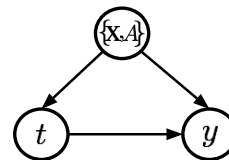
3.1. Setup

Figure 1: Causal graph of observational data considered in our setting.

We consider a setting where we have access to observational data with base covariates X, missing attribute A, treatment t, and outcome y (see Figure 1 for the causal diagram assumed throughout this work). Both A and t are considered binary variables throughout the paper for convenience. However, the proposed methods also work if the missing attribute is categorical and can be generalized in the presence of multiple missing attributes. y_0 is the potential outcome for this sample if $t = 0$ and y_1 is the outcome corresponding to $t = 1$. For each independent and identically distributed sample (X, A, t), we can only observe one potential outcome depending on the treatment assignment: $y = y_0$ if $t = 0$, and $y = y_1$ if $t = 1$. We assume the following throughout this work:

Assumption 1 (Ignorability)
Potential outcomes are independent of the treatment given covariates $\{X, A\}$:

$$y_1, y_0 \perp\!\!\!\perp t \,|\, X, A$$

Assumption 2 (Common Support)
$0 < p(t = 1|X, A) < 1$ *for all values of X, A with $p(X, A) > 0$*

In addition, A is *missing not at random* (MNAR) (i.e. the missingness of A depends on its actual values, see Appendix A). We call D_{pool} the set of samples with missing A. This hinders an accurate estimation of the average treatment effect (ATE, Definition 2) $E[y_1 - y_0]$. Our goal is to strategically query samples from D_{pool} to obtain actual values of A so that the treatment effect estimate can be improved.
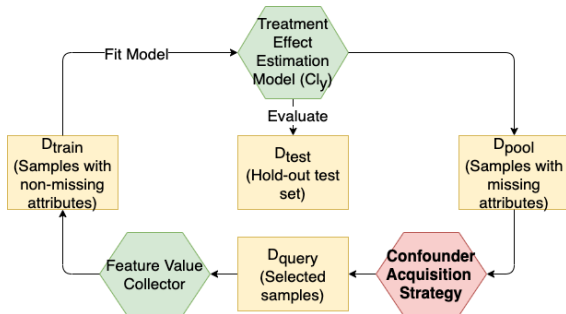

Figure 2: Pipeline for data acquisition for CEE. The focus of our work is on confounding feature acquisition (pink block).

We use an iterative procedure to evaluate the effectiveness of our proposed strategies. Missing features are acquired in batches and used in subsequent model training to obtain effect sizes. This procedure is summarized in the pipeline Figure 2 and Algorithm 1.

3.2. Feature Acquisition Strategies

We propose two feature acquisition strategies. The first consists of explicitly characterizing and fixing covariate imbalance between

---

Algorithm 1 Feature acquisition for CEE

1: **Input:** $D_{\text{train}}, D_{\text{test}}, D_{\text{pool}}$, Models Cl_y, Cl_A, Acquisition Method (CB or OE), Batch size: β, ATE variance threshold (hyperparameter): σ^2_{ate}.
2: **while** $|D_{\text{pool}}| > 0$ and $Var(\epsilon_{ATE}) > \sigma^2_{ate}$ **do**
3:    $\text{Cl}_y \leftarrow \text{Cl}_y(X_{\text{train}}, A_{\text{train}}, t_{\text{train}})$
4:    $\text{Cl}_A \leftarrow \text{Cl}_A(X_{\text{train}}, t_{\text{train}})$
5:    Choose batch of size β, i.e. $(X^*, t^*, y^*)_\beta$ using Eq. 2 for CB or Eq. 3 for OE
6:    $A^*_\beta \leftarrow \text{Oracle}((X^*, t^*, y^*)_\beta)$
7:    $D_{\text{train}} \leftarrow D_{\text{train}} \cup (X^*, t^*, y^*, A^*)_\beta$
8:    $D_{\text{pool}} \leftarrow D_{\text{pool}} \setminus (X^*, t^*, y^*)_\beta$
9: **end while**
10: **return** $\epsilon_{PEHE}, \epsilon_{ATE}$

---

treated and control samples. Addressing such imbalances is critical for unbiased CEE. The second is motivated by improving estimation error of outcomes for treated and control populations. That is, we estimate the expected utility of decrease in estimation error by acquiring missing confounding value of A, and select samples entirely based on statistical errors in potential outcomes estimation. Our methods are described in the following.

Covariate Balancing (CB) The presence of confounding creates unwanted dependencies between the treatment assignment and outcome of interest, and is a deterrent to reliable CEE. Adjusting for confounding aims to account for distributional differences between the treatment and control group (i.e., *imbalance*), and is a key element in CEE (Rubin, 2005; Shalit et al., 2017). Our first proposed strategy aims to use feature acquisition to decrease imbalance. We acquire samples from D_{pool} that, when added to D_{train}, reduce the expected Maximum Mean Discrepancy (MMD)[3] between treated and control groups.

---

3. See Def 1.

We denote the population MMD estimated from samples $\texttt{U} \sim P$ and samples $\texttt{V} \sim Q$ by $\text{MMD}(\texttt{U}, \texttt{V})$.

Definition 1 (MMD)

$$MMD(P,Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} E_{X \sim P}[f(X)] - E_{X \sim Q}[f(X)] \tag{1}$$

where $\mathcal{H}$ is the Reproducing Kernel Hilbert Space (RKHS).

Since values of A are missing, for each sample in D_{pool}, we first model $p(A|X,T)$ (using Cl_A learnt with samples in D_{train}), compute the expected MMD, and choose samples that will minimize the following estimate. Let $\texttt{T}$ be the treated samples in D_{train}, and $\texttt{C}$ be the control samples in D_{train}.

$$\underset{(X,t,y) \in \text{D}_{\text{pool}}}{arg\,min} E_{A \sim p(A|X,t)}(\llbracket t=1 \rrbracket \text{MMD}(\texttt{T} \cup (X,A), \texttt{C}) +$$
$$\llbracket t=0 \rrbracket \text{MMD}(\texttt{T}, \texttt{C} \cup (X,A))) \tag{2}$$

where $\llbracket \cdot \rrbracket$ denotes the indicator function. Note that we model A only using other confounding attributes and treatment (not outcome y) so that treated and control units are balanced without observing the factual outcome.

Outcome Error (OE) Our second strategy is to focus only on overall estimation errors and acquire data that improve estimation directly. For each sample in D_{pool}, we predict the expected outcome $E_{A \sim p(A|X,t)}[\hat{y}|X,A,T]$, where the weights are $p(A|X,T)$. $\texttt{Cl}_y \triangleq p(y|X,A,t)$ is the estimator trained on D_{train} and $\texttt{Cl}_A \triangleq p(A|X,t)$ is trained as in the previous method. Since the observed factual outcome y is known, the samples that lead to the highest error between the predicted outcome and the observed outcome are picked and the information of A is queried. In other words, we choose samples using:

$$\underset{(X,t,y) \in \text{D}_{\text{pool}}}{arg\,max} |E_{A \sim p(A|X,t)}[\hat{y}|X,A,t] - y| \tag{3}$$

While this is closer in spirit to traditional active learning, note that the latter focuses on

acquiring the label Y rather than a confounding feature A with the explicit purpose of effect size estimation. Additionally, the source of the statistical errors in outcome estimates comes from confounding variables.

We compare both proposed strategies to two commonly used active learning baseline strategies:

- **Random** This method consists of taking random samples and querying their A values.

- **Uncertainty of A (Uncertainty)** At each iteration, a classifier ($\texttt{Cl}_A$) is trained to predict A from the covariates X and the treatment t. Samples from D_{pool} with the highest uncertainty on the value of A are selected.

3.3. Models

We use a random forest classifier (Breiman, 2001) to train $\texttt{Cl}_A$, used in all strategies to predict A from covariates X and treatment t at each iteration. As for $\texttt{Cl}_y$, we test five different CEE models. They range from traditional statistical approaches to more recent machine learning approaches. These models are:

- **Causal Forest (CF)**(Wager and Athey, 2018). CF is an extension of the random forest algorithm adapted to infer causal effects. As treatment effect is directly predicted in the leaf nodes without prediction of factual or counterfactual outcome, $\texttt{OE}$ strategy does not work with CF.

- **Doubly Robust Estimation (DR)** (Bang and Robins, 2005). DR combines the capacity of linear model and covariate balance from propensity scoring. DR originally combines a logistic regression for propensity scoring and a linear regression for final outcome prediction. We implement a more

complex estimator by replacing linear models with two single-layer neural networks.

- **Multiheaded Multi-Layer Perceptron (MLP_Multi)**. We fit two Multilayer Perceptron neural networks (Buitinck et al., 2013), one for the treated group and the other for the control group to estimate potential outcomes. The final treatment effect estimate is the difference between them.

- **Multiheaded Gaussian Processes (GP_Multi)**. We fit two Gaussian process regressors with RBF kernels (Buitinck et al., 2013), one for the treated group and the other for the control group to estimate potential outcomes. Final treatment effect estimate is taken as the difference between them.

- **Causal Multitask Gaussian Processes (CMGP)**. We augment the CMGP procedure introduced by Alaa and van der Schaar (2017) with our acquisition strategies to estimate treatment effect.

3.4. Data

We use a semi-synthetic benchmark dataset, Infant Health and Development Program (IHDP) dataset (Hill and McCulloch, 2011), to evaluate the proposed feature acquisition strategies. As in other work, we remove a subset of the population to create a biased dataset (Shalit et al., 2017; Madras et al., 2019; Johansson et al., 2016) and use mother's ethnicity *momwhite* as A.

To ensure that information from A is truly unavailable for most samples, we create independence between other covariates and A by randomly permuting values for A. (We later relax this independence assumption in Section 5.1). After normalizing the dataset, we adapt the generation of outcomes using the response surface type B of Hill and McCulloch (2011). Specifically, we sample the treatments

under Bernoulli distributions and potential outcomes under normal distribution. To create a partially observed dataset, we mask 95% of A values, where the probability of missing depends on unmasked A values.

Details on the simulation steps are described in Algorithms 2 - 4 in Appendix A.

3.5. Evaluation

We evaluate the effectiveness of our proposed acquisition strategies in terms of average treatment effect (ATE) and individual treatment effect (ITE) using a hold-out test set.

For ATE, we measure sample estimate of average absolute error in ATE. For ITE, we measure sample estimates of Precision in Estimation of Heterogeneous Effect (PEHE).

Let $\hat{y}_1(X, A)$ and $\hat{y}_0(X, A)$ be the potential outcomes estimated under any of the aforementioned hypothesis classes. Definition of these metrics are listed below:

Definition 2 (Sample Error in ATE)

$$
\begin{aligned}
\epsilon_{ATE} = &|\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_1(X_i, A_i) - \hat{y}_0(X_i, A_i)) \\
&- \frac{1}{n} \sum_{i=1}^{n} (y_1(X_i, A_i) - y_0(X_i, A_i))|
\end{aligned}
\tag{4}
$$

Definition 3 (Sample PEHE)

$$
\begin{aligned}
\epsilon_{PEHE} = &\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_1(X_i, A_i) - \hat{y}_0(X_i, A_i) \\
&- (y_1(X_i, A_i) - y_0(X_i, A_i)))^2
\end{aligned}
\tag{5}
$$

Note also that we use the *noiseless* outcome in data generation as $y_1(X_i, A_i)$ and $y_0(X_i, A_i)$. This is consistent to other papers using the IHDP dataset (Hill and McCulloch, 2011; Shalit et al., 2017).

4. Results

We measure the effectiveness of different strategies by their empirical sample efficiency

Table 1: Comparison of results for different model types and acquisition strategies. Lower is better. ± shows the width of 95% confidence interval, running on 500 realizations of simulated IHDP data. Statistically significant better performing strategies for each model are bolded.

(a) Optimal performance of each ATE estimation method when all feature values are acquired and all samples are used for training.

	DR	CMGP	GP_Multi	MLP_Multi	CF
Optimal ϵ_{ATE}	0.65 ± 0.07	0.45 ± 0.05	0.72 ± 0.07	0.81 ± 0.08	0.90 ± 0.09
Optimal $\sqrt{\epsilon_{PEHE}}$	5.41 ± 0.42	4.44 ± 0.35	7.03 ± 0.48	7.90 ± 0.58	8.28 ± 0.57

(b) Average number of samples needed to reach within 1% of optimal performance.

	DR	CMGP	GP_Multi	MLP_Multi	CF
Number of Samples Needed to Reach within 1% of Optimal ϵ_{ATE}					
Random	168 ± 13	165 ± 14	174 ± 15	107 ± 10	136 ± 13
Uncertainty	160 ± 13	154 ± 14	175 ± 16	108 ± 11	129 ± 13
CB	162 ± 13	137 ± 13	128 ± 13	98 ± 11	108 ± 12
OE	$\mathbf{92 \pm 9}$	$\mathbf{111 \pm 11}$	$\mathbf{99 \pm 9}$	117 ± 10	
Number of Samples Needed to Reach within 1% of Optimal $\sqrt{\epsilon_{PEHE}}$					
Random	451 ± 12	487 ± 9	492 ± 9	435 ± 16	425 ± 13
Uncertainty	422 ± 12	478 ± 10	486 ± 10	427 ± 16	427 ± 13
CB	413 ± 12	472 ± 10	496 ± 9	428 ± 16	434 ± 13
OE	$\mathbf{170 \pm 12}$	$\mathbf{352 \pm 14}$	$\mathbf{398 \pm 9}$	$\mathbf{343 \pm 18}$	

in reaching the effect estimate when all samples are acquired or there is no missingness. Table 1b summarizes this information in terms of the number of samples needed to reach within 1% of the optimal performance (presented in Table 1a). It should also be noted that here we focus on ATE, not ITE, as our acquisition setup is not designed for individual level CEE. We show PEHE for completeness.

As shown in the table, OE is significantly more effective for DR, CMGP, and GP_Multi in reaching both optimal error in ATE and optimal PEHE. OE also provides significant benefit in reaching optimal PEHE for MLP_Multi.

As DR and CMGP are the best models in terms of optimal error in ATE and PEHE, we plot the changes in values of these metrics when running different acquisition strategies with them in Figure 3. The figures clearly demonstrate the significant benefit from using OE. Additionally, when using DR, OE offers consistent better performance regardless
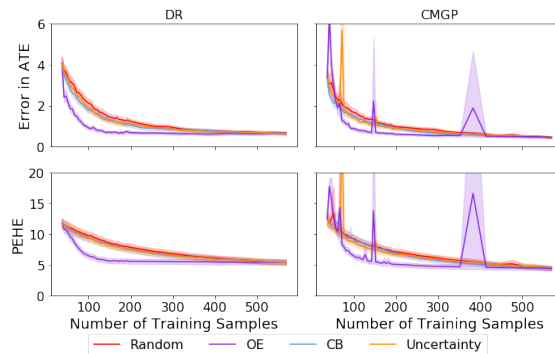

Figure 3: Error in ATE and PEHE. X-axis is the number of points in the queried pool. Each line shows the change in average value of performance metrics in 500 realizations, and the shaded region is the corresponding 95% confidence interval. Lower is better. Spikes in CMGP results are due to numerical instabilities as a result of implementing CMGP using the GPy package.

of number of samples queried, as compared to CB and baseline strategies. It should be noted that by focusing on a specific source of bias - imbalance - in effect estimates, CB provides some benefits initially for both DR and CMGP but these benefits are generally outweighed by relying on cumulative sources of statistical errors (OE) for feature acquisition.

5. Analysis

5.1. Analysis of Relationship between A and Other Confounders

In this section, we evaluate whether the benefits of our proposed methods persist even when the missing attribute is correlated with other confounders. DR is used as the CEE model for all experiments in this section as it is the one of the best performing models and requires a significantly shorter run time compared to CMGP. We run 100 realizations of simulated IHDP data, and vary dependency in two ways:

Original A Values We use the original values of *momwhite* in the IHDP dataset. At one extreme, we replace *momwhite* values in all samples with randomly generated values (i.e. the setting that produces our main results in Section 4). At the other extreme, we keep the original *momwhite* values that are correlated with other covariates naturally. In between, we randomly replace 20%, 40%, 60%, or 80% of original A values with a random value. When *momwhite* value is generated, it follows a Bernoulli distribution where the positive probability is equal to the sample probability in the original dataset. This evaluates benefits of each strategy with varying levels of noise affecting model fit for $p(A|X,t)$.

Bivariate Gaussian Simulation We simulate *momwhite* such that it has varying degrees of correlation with *birthweight* denoted by $X_b \subseteq X$ (one of the covariates in IHDP).

More specifically, we first generate A such that $[logit(A), X_b]$ is jointly Gaussian, with $logit(A)$ following a standard normal distribution, and X_b following a normal distribution with mean and standard deviation matching those of *birthweight*. Correlation coefficient between $logit(A)$ and X_b varies between 0 and 1. A is generated by thresholding $logit(A)$ such that the proportions of *momwhite* is the same as that in the original dataset. This evaluation assesses whether benefits of using CB and OE persist at different levels of correlation, when we explicitly model the relationship between A and other covariates $X \supseteq X_b$.

Tables 2 summarizes the number of samples needed under each acquisition strategy to reduce ATE to a level that is within 1% of the optimal performance. As shown in the table, OE consistently outperforms the baseline strategies despite varying level of correlation between A and other covariates. We also observe that as A becomes more correlated with other covariates, it generally requires fewer acquired samples to reach optimal performance. Evaluation of PEHE gives similar patterns and is presented in Appendix B.

5.2. Comparison of Acquisition Strategies

We analyze the performance difference between CB and OE by visualizing what samples are prioritized by each strategy, as well as comparing some resulting summary statistics of the acquired samples. Results shown in this section use CMGP for Cl_y because of its strong performance and are under the same setting as our main experiment (i.e. A is independent of other covariates). We have two findings:

Explore-Exploit Tradeoff One challenge in a traditional active learning problem is how explore-exploit trade-off is handled (Bondu et al., 2010). To understand how our two

Table 2: Sample efficiency to reach optimal errors in ATE while relaxing the independence assumption. Values are the average number of samples needed in training to reach within 1% of the optimal performance level, with ± showing the width of 95% confidence interval, over 100 realizations of the simulated data. Lower is better. Statistically significant better performing strategies for each experiment type are bolded.

(a) Experiments with noise added to original A values. Columns represent fraction of original A values retained.

	0	0.2	0.4	0.6	0.8	1
Random	156 ± 29	175 ± 30	161 ± 27	158 ± 27	160 ± 29	145 ± 23
Uncertainty	149 ± 28	182 ± 33	128 ± 22	142 ± 26	123 ± 23	119 ± 24
CB	144 ± 26	144 ± 25	149 ± 24	122 ± 24	125 ± 25	150 ± 31
OE	$\mathbf{87 \pm 19}$	$\mathbf{81 \pm 18}$	91 ± 20	84 ± 18	108 ± 26	80 ± 16

(b) Experiments with Bivariate gaussian simulation. Columns represent correlation coefficient between simulated A and *birthweight*.

|	0	0.2	0.4	0.6	0.8	1
Random	193 ± 33	171 ± 28	182 ± 32	159 ± 30	140 ± 28	141 ± 25
Uncertainty	156 ± 30	143 ± 27	139 ± 28	142 ± 31	109 ± 19	134 ± 24
CB	134 ± 23	184 ± 32	154 ± 28	142 ± 25	143 ± 27	142 ± 28
OE	97 ± 23	$\mathbf{76 \pm 16}$	$\mathbf{92 \pm 19}$	$\mathbf{87 \pm 18}$	87 ± 18	$\mathbf{81 \pm 15}$

strategies handle this issue, we assess the imbalance at every stage of our procedure by visualizing the first two principal components of all pre-treatment confounders (X, A). Figure 4a shows a 2D presentation of samples in the initial labeled set, as well as when samples are acquired (out of a total of 569). While CB first acquires points similar to the already queried point, OE explores more and acquires points from all possible regions in the input space. This is especially true for samples in the control group, where the representation is less clustered. When most samples are acquired, however, the input space is well explored and represented in the training samples by both strategies. Our results in Section 4 suggest that with this particular setting, the benefit of early exploration in OE outweighs the benefit of early exploitation in CB.

Treatment vs Control Samples We next look at how CB and OE differ in terms of number of treated and control samples acquired. As shown in Figure 4b, OE strategy queries more points in the control group in the early stage. This is because the expected outcome for the control group is an exponential func-

tion of the covariates and has a larger range as compared to the mean outcome for the treated group. OE strategy learns that it is more difficult to predict the a control outcome and therefore queries more control samples. On the other hand, as CB strategy focuses on reducing imbalance, it maintains a more balanced treatment ratio in the acquired samples. OE strategy has a performance advantage in this setting as the outcome generating processes have different complexities between treatment and control group. This advantage of OE empirically outweighs imbalance issues and is an interesting statistical trade-off that should be explored for feature acquisition strategies intended for CEE.

6. Conclusion

Confounding feature acquisition strategies for CEE and down-stream decision making need special attention, particularly in healthcare, where sensitive and/or costly but critical pre-treatment confounders may be unavailable. In this work, we propose two strategies for efficiently selecting missing values in pre-treatment confounders to acquire which con-
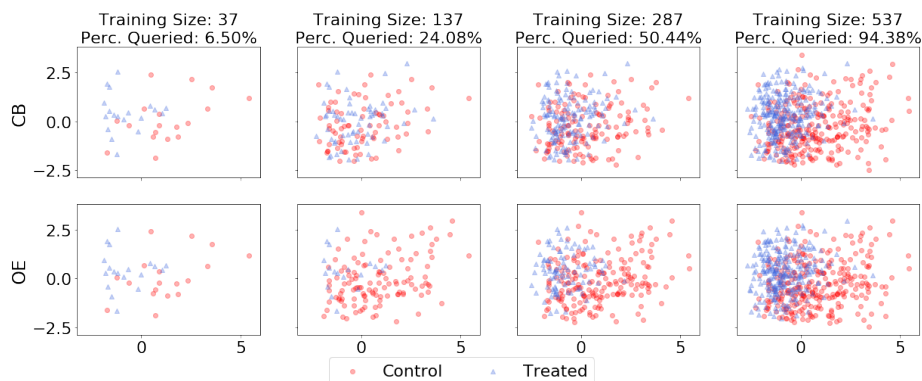
(*a*) Visualization of training samples using first two principal component of all covariates (including *A*) as more feature values are acquired. Results shown are obtained on a single realization of simulated IHDP data. Results on other realizations show the same pattern.
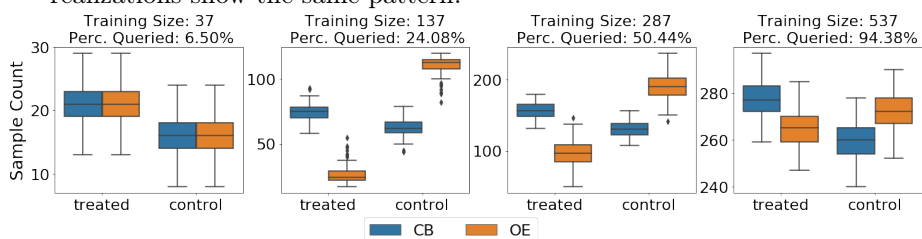
(*b*) Average number of treated and control observations as more feature values are acquired, across 500 realizations of simulated IHDP data.

Figure 4: Comparison of `CB` and `OE` strategies using CMGP as the CEE model.

sist of: i) explicitly characterizing amount of imbalance between treated and control groups and ii) relying on estimation errors of outcome estimates. We observe benefits of relying on statistical errors versus imbalance in acquisition strategies, and provide insights on their explore-exploit properties and sample choice based on the complexity of outcome response functions. Our results highlight the importance of addressing the trade-off between estimation bias and covariate imbalance for this task. Other datasets providing ground truth of the treatment effect could further validate our proposed methods, which we leave for future work.

Acknowledgments

References

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *31st Conference on Neural Information Processing Systems (NIPS)*. NIPS, 2017.

---

4. www.vectorinstitute.ai/partners

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4): 962–973, 2005.

UK Biobank. About uk biobank. *Available at h ttps://www. ukbiobank. ac. uk/a bout-biobank-uk*, 2014.

Alexis Bondu, Vincent Lemaire, and Marc Boullé. Exploration vs. exploitation in active learning: A bayesian approach. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2010.

Leo Breiman. Random forests. *Machine Learning 45*, 2001.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 339–348. ACM, 2019.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

Jennifer L. Hill and Robert E. McCulloch. Bayesian nonparametric modeling for causal inference. In *Journal of Computational and Graphical Statistics-Volume 20*, pages 217–240, 2011.

Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. Classification with costly features as a sequential decision-making problem. *Machine Learning*, pages 1–29, 2020.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

S Imindu Liyanage, Clarissa Santos, and Donald F Weaver. The hidden variables problem in alzheimer's disease clinical trial design. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 4: 628–635, 2018.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358. ACM, 2019.

Prem Melville, Maytal Saar-Tschansky, Foster Provost, and Raymond Mooney. An

expected utility approach to active feature-value acquisition. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 2009.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in neural information processing systems*, pages 1368–1378, 2018.

Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. Active learning for decision-making from im-

balanced observational data. *arXiv preprint arXiv:1904.05268*, 2019.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. The label complexity of active learning from observational data. In *Advances in Neural Information Processing Systems*, pages 1810–1819, 2019.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making - the causal explanation formula. 2018.

Junzhe Zhang and Elias Bareinboim. Fair transfer learning with missing protected attributes. pages 91–98, 2019.

Appendix A. Simulation Algorithms

Algorithm 2 to generate missing A values

For each data point i
- Sample u_i from $\mathcal{N}((0,1)$
- Compute $p_i = (2 - A_i) \times 0.2 + u_i \times 0.5$

Rank data point based in descending order of p_i and mask A values for the top 95%.

Algorithm 3 Algorithm to generate the treatments

Input: a subset of features from X (X_{sub}), and ξ ($\xi \in \mathbb{R}^n$ where n is the number of features in $Xsub$)

    Compute $p = Clip(X_{sub}\xi^T, 0.005, 0.995)$

        Sample t following $Bernoulli(p)$

Output: t.

Algorithm 4 Algorithm to generate the outcomes

Input: X, and β, and W. ($\beta, W \in \mathbb{R}^n$ where n is the number of covariates.)

Compute β and W.

- β
  - Specified values for *b.marr*, *mom.scoll*, *work.dur*, *momwhite*, *cig*, *drugs*
  - For other continuous covariates: generate a vector of coefficients [0, 0.1, 0.2, 0.3, 0.4], sampled with probabilities [0.5, 0.125, 0.125, 0.125, 0.125].
  - For other binary covariates: generate a vector of coefficients [0, 0.1, 0.2, 0.3, 0.4], sampled with probabilities [0.6, 0.1, 0.1, 0.1, 0.1].
- W 0 for the 6 features with specified β values; 0.5 for other covariates

Sample y_0 following $\mathcal{N}(\exp{((X+W)\beta^T)}, 1)$

Sample y_1 following $\mathcal{N}((X+W)\beta^T, 1)$

Output: y_0, y_1.

Appendix B. PEHE Results for Section 5.1

Table 3: Speed of reaching optimal PEHE across acquisition strategies while relaxing the independence assumption. Values are the average number of samples needed in training to reach within 1% of the optimal performance level, with $\pm$ showing the width of 95% confidence interval, over 100 realizations of the simulated data. Lower is better. Statistically significant better performing strategies for each experiment type are bolded.

(a) Experiments involving original A values. Columns represent fraction of original A values remaining.

	0	0.2	0.4	0.6	0.8	1
Random	436 ± 28	439 ± 25	439 ± 26	425 ± 27	413 ± 28	392 ± 34
Uncertainty	420 ± 31	413 ± 27	389 ± 27	389 ± 31	342 ± 30	341 ± 32
CB	396 ± 28	357 ± 28	410 ± 28	379 ± 28	395 ± 29	406 ± 30
OE	$\mathbf{160 \pm 26}$	$\mathbf{159 \pm 24}$	$\mathbf{151 \pm 23}$	$\mathbf{163 \pm 27}$	$\mathbf{161 \pm 29}$	$\mathbf{152 \pm 25}$

(b) Experiments involving bivariate Gaussian simulation. Columns represent correlation coefficient between simulated A and *birthweight*.

	0	0.2	0.4	0.6	0.8	1
Random	453 ± 26	457 ± 22	444 ± 28	422 ± 27	439 ± 27	431 ± 26
Uncertainty	421 ± 32	432 ± 30	434 ± 29	434 ± 32	456 ± 24	460 ± 23
CB	372 ± 30	424 ± 27	440 ± 25	421 ± 27	460 ± 23	459 ± 23
OE	$\mathbf{179 \pm 27}$	$\mathbf{168 \pm 29}$	$\mathbf{164 \pm 25}$	$\mathbf{181 \pm 30}$	$\mathbf{178 \pm 28}$	$\mathbf{158 \pm 25}$