CounterFactual Regression with Importance Sampling Weights

Negar Hassanpour* and **Russell Greiner**
Department of Computing Science, University of Alberta, Canada
{hassanpo, rgreiner}@ualberta.ca

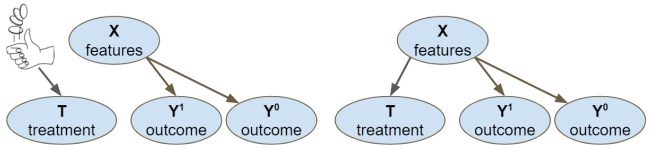
Abstract

Perhaps the most pressing concern of a patient diagnosed with cancer is her life expectancy under various treatment options. For a binary-treatment case, this translates into estimating the difference between the outcomes (*e.g.*, survival time) of the two available treatment options – *i.e.*, her Individual Treatment Effect (ITE). This is especially challenging to estimate from observational data, as that data has selection bias: the treatment assigned to a patient depends on that patient's attributes. In this work, we borrow ideas from domain adaptation to address the distributional shift between the source (outcome of the administered treatment, appearing in the observed training data) and target (outcome of the alternative treatment) that exists due to selection bias. We propose a *context-aware* importance sampling re-weighing scheme, built on top of a representation learning module, for estimating ITEs. Empirical results on two publicly available benchmarks demonstrate that the proposed method significantly outperforms state-of-the-art.

1 Introduction

To identify the appropriate action to take, an intelligent agent must infer the causal effects of its every possible action choice. A prominent example is precision medicine – *i.e.*, the customization of health-care tailored to each individual patient – that attempts to identify which medical procedure $t \in \mathcal{T}$ will benefit each specific patient x the most. Learning such models requires answering counterfactual questions [Rubin, 1974; Pearl, 2009] such as: *"Would this patient have lived longer, had she received an alternative treatment?"*. This type of counterfactual analysis is not limited to health-care; it can be used in any field where personalized action selection is of value, including intelligent tutoring systems [Rollinson and Brunskill, 2015], news article recommender systems [Li *et al.*, 2010], ad-placement systems [Bottou *et al.*, 2013], and webpage recommendation by search engines [Li *et al.*, 2015].

Pearl [2009] demonstrates that, in general, causal relationships can only be learned by experimentation (on-line exploration), or running a Randomized Controlled Trial (RCT). In RCTs, the treatment assignment does not depend on the individual X – see Fig. 1(a). In many cases, however, this is expensive, unethical, or even infeasible [Pearl, 2009]. As a result, we are forced to approximate causal effects from off-line datasets collected through observational studies. Such datasets, however, often exhibit *selection bias* [Imbens and Rubin, 2015], – *i.e.*, where $\Pr(T \mid X) \neq \Pr(T)$. In other words, the administered treatment T depends on some attribute values of the individual X – see Fig. 1(b). Fig. 2 illustrates selection bias in an example of a synthetic observational dataset.



(a) Randomized Controlled Trial (b) Observational Study

Figure 1: Belief net structure for randomized controlled trials and observational studies. Here, Y^0 (Y^1) is the outcome of applying $T =$ treatment#0 (treatment#1) to the individual represented by X.

For notation: a dataset $\mathcal{D} = \{ [x_i, t_i, y_i] \}_{i=1}^{N}$ used for causal effect estimation has the following format: for the i^{th} instance (*e.g.*, patient), we have some context information $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ (*e.g.*, age, BMI, blood work, etc.), the administered treatment t_i chosen from a set of treatment options $\mathcal{T}$ (*e.g.*, $\{0$: medication, 1: surgery$\}$), and the respective observed outcome $y_i \in \mathcal{Y}$ (*e.g.*, survival time; $\mathcal{Y} \subseteq \mathbb{R}$) as a result of receiving treatment t_i. Note that $\mathcal{D}$ only contains the outcome of the administered treatment (*observed* outcome: y_i), but not the outcome(s) of the alternative treatment(s) (*counterfactual* outcome(s): y_i^t for $t \in \mathcal{T} \setminus \{t_i\}$), which is(are) inherently unobservable. For a binary-treatment case, we denote the alternative treatment as $\neg t_i = 1 - t_i$.

In this paper, we are interested in finding the Individual Treatment Effect (ITE) for each instance i – *i.e.*, we want to estimate $\mathrm{e}_i = y_i^1 - y_i^0$. To do so, we frame the solution as learning the function $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ that can accurately predict the outcomes (both observed $\hat{y_i}^{t_i} = f(x_i, t_i)$ as well as counterfactuals $\hat{y_i}^{\neg t_i} = f(x_i, \neg t_i)$) given the context information x_i for each individual.

*Contact Author

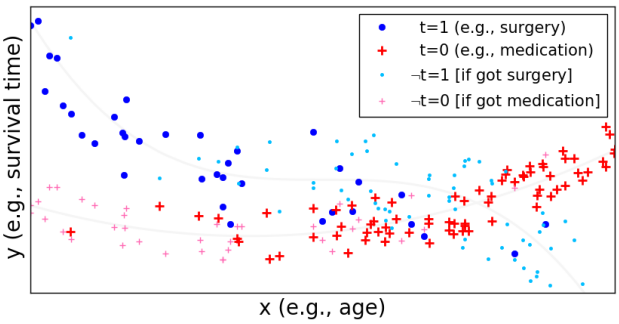


Figure 2: An example observational dataset (best viewed in color). Here, to treat heart disease, a doctor typically prescribes surgery ($t = 1$, •) to younger patients and medication ($t = 0$, +) to older ones. Note that instances with larger (smaller) x values have had a higher chance to be assigned to the $t = 0$ (1) treatment arm; hence we have selection bias. The counterfactual outcomes – only used for evaluation purpose – are illustrated by small • (+) for $\neg t = 1$ (0).

There are two challenges associated with estimating ITEs:

(i) Training data never includes the counterfactual outcomes $y^{\neg t}$ for any training instances; which makes estimating causal effects a significantly different (and more complicated) problem than the common tasks in standard supervised machine learning.

(ii) Selection bias in observational datasets implies having fewer instances within each treatment arm at specific regions of the domain. This sparsity, in turn, would decrease the accuracy and confidence of estimating the counterfactual outcomes at those regions.

The first challenge is an inherent characteristic of this task. We focus on the following ways to mitigate the second challenge:

- **Representation learning** [Bengio *et al.*, 2013] – The idea here is to learn a representation space $\Phi(\cdot)$ in which the selection bias is reduced as much as possible but not at the expense of a decrease in accuracy of predicting the observed outcomes. In other words, assuming X is generated from three underlying (unobserved) factors as shown in Fig. 3[1], this would ideally be conducted by identifying $\{A, B, C\}$ factors and then removing A.
- **Re-weighting** – This is a common statistical method for addressing covariate shift [Shimodaira, 2000] and domain adaptation in general. It is easy to show that selection bias in observational studies translates into a domain adaptation scenario (see Appendix for details) where we want to learn a model from the "source" (observed) data distribution that will perform well in the "target" (counterfactual) one.

[1] Examples for (A) wealth: rich patients receiving the expensive treatment while poor patients receiving the cheap one, although outcomes of the possible treatments are not particularly dependent on patients' wealth status; (B) age: younger patients receiving surgery while older patients receiving medication; and (C) genetic information that determines the efficacy of various medications, however, such relationships are unknown to the attending physician.

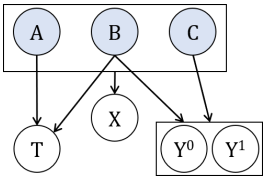


Figure 3: Underlying (latent) factors of X; A are factors that partially determine only t, but not the other variables; C are factors that partially determine y; and B are confounders (factors that partially determine both t and y). Selection bias is induced by A and B.

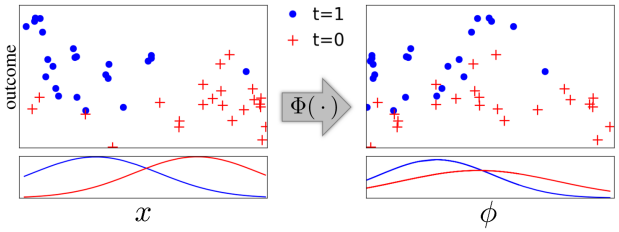


Figure 4: The learned representation has reduced the selection bias. That is, the $t=1$ and $t=0$ distributions of the transformed instances $\Phi(x)$ – here, the distribution of + versus • on the x-axis – are much closer to each other compared to those distributions in the original x space. Also note that the observed outcomes y (on the y-axis) remain unchanged through this transformation.

Main contribution: In this work, we propose a new context-aware weighting scheme based on importance sampling technique, on top of a representation learning module, to alleviate the problem of selection bias in ITE estimation.

Our analysis relies on the following assumptions:

Assumption 1: Unconfoundedness – There are no unobserved confounders (*i.e.*, covariates that contribute to both treatment selection procedure as well as determination of outcomes). Formally, $\{Y^t\}_{t\in\mathcal{T}} \perp\!\!\!\perp T \mid X$. [2]

Assumption 2: Overlap – Every individual x has a non-zero chance of being assigned to any treatment arm. That is, $\Pr(t \mid x) \neq 0 \quad \forall t \in \mathcal{T},\ \forall x \in \mathcal{X}$.

These two assumptions together are called *strong ignorability* [Rosenbaum and Rubin, 1983], which is sufficient for ITE to be identifiable [Imbens and Wooldridge, 2009].

2 Related Works

Learning treatment effects from observational studies is closely related to "off-policy learning in contextual bandits" – *cf.*, [Strehl *et al.*, 2010; Swaminathan and Joachims, 2015a], where the goal is to learn an optimal policy $\pi(t \mid x)$ that picks the best treatment for each individual. One strategy to address this task is "outcome prediction" – *i.e.*, estimating $y(x,t) \ \ \forall t \in \mathcal{T}$ for each x, then select the one that promises the best outcome $\pi(t \mid x) = \underset{t}{\operatorname{argmax}}\ y(x,t)$. This is equiv-

[2] In other words, all confounders B in Fig. 3 are either directly observed in X or discoverable by proxy from X (*e.g.*, Body Mass Index (BMI) can be considered a proxy for true body fat percentage).

alent to what is done for ITE estimation.[3] Another strategy bypasses the outcome prediction step and directly obtains the optimal policy by maximizing a utility function (similar to "expected return" in Reinforcement Learning [Sutton and Barto, 1998]). The majority of approaches under this second strategy belong to the Inverse Propensity Weighting (IPS) family of methods, which attempt to balance the source and target distributions by re-weighting certain data instances – *cf.*, [Austin, 2011; Swaminathan and Joachims, 2015b].

Atan *et al.* [2018] use an auto-encoder network to learn a representation space $\Phi(\cdot)$ that reduces the selection bias by minimizing the cross entropy loss between $\Pr(t)$ and $\Pr(t \mid \Phi(x))$. However, by training an auto-encoder, they force their network to be able to reproduce *all* the covariates in x from Φ. This could effectively neutralize the merit of representation learning when there are features (in x) that had contributed to selecting the assigned treatment, but which had no effect on determining the outcomes – see Footnote 1(A).

Shalit *et al.* [2017] – called "SJS" below – attempt to reduce selection bias by learning a common representation space $\Phi(\cdot)$ that tries to make $\Pr(x \mid t=0)$ and $\Pr(x \mid t=1)$ as close to each other as possible (see Fig. 4), provided that $\Phi(x)$ retains enough information that all $|\mathcal{T}|$ learned regressors $h^t(\Phi)$ can generalize well on the observed outcomes. Φ and h^t are implemented as neural networks and learned by minimizing:

$$J(h,\Phi) = \frac{1}{N}\sum_{i=1}^{N} \omega_i \cdot L\big[\, y_i,\, h^{t_i}\big(\Phi(x_i)\big)\,\big] + \lambda \cdot \mathfrak{R}(h) + \alpha \cdot \text{IPM}\big(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}\big) \quad (1)$$

where $L\big[\, y_i,\, h^{t_i}\big(\Phi(x_i)\big)\,\big]$ is the loss of predicting the observed outcome for sample i, weighted by ω_i, derived via:

$$\omega_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)} \quad (2)$$

where $u = \frac{1}{N}\sum_{i=1}^{N} t_i = \Pr(t=1)$. Also, $\mathfrak{R}(h)$ in Eq. (1) is the regularization term for penalizing model complexity, and the final term $\texttt{disc} = \text{IPM}\big(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}\big)$ is the *discrepancy* – calculated by an Integral Probability Metric (IPM) – that measures the distance between the two distributions $\Pr(\Phi(x) \mid t=0)$ and $\Pr(\Phi(x) \mid t=1)$. See Fig. 5(a) for SJS's model architecture.

The SJS model is closely related to its predecessor [Johansson *et al.*, 2016], which defined `disc` between the joint distributions of Φ and t (factual) versus Φ and $\neg t$ (counterfactual) – *i.e.*, $\texttt{disc} = \text{IPM}\Big(\big\{[\,\Phi(x_i), t_i\,]\big\}_{i=1}^{N}, \big\{[\,\Phi(x_i), \neg t_i\,]\big\}_{i=1}^{N}\Big)$. This makes sense in theory: if the factual and counterfactual joint distributions are hard to distinguish, it means that the data is close to RCT. However, since the two joint distributions only differ in their treatment bit (*i.e.*, t versus $\neg t$, while $\Phi(x)$ is the same for both), the numerical value of `disc` would naturally be small. Therefore, its contribution to the objective would be negligible. Moreover, a high dimensional $\Phi(\cdot)$ can overshadow the information in the treatment bit, which results in an even smaller `disc`.

Perhaps the work most related to ours is [Johansson *et al.*, 2018], which also applies sample re-weighting on top of representation learning to balance their source and target domains by minimizing `disc` between the factual joint distribution $p_\mu(\phi, t)$ and a weighted (ω) counterfactual one $\omega \cdot p_\pi(\phi, \neg t)$, where ϕ is set as the numerical value of $\Phi(x)$. Hence, this method is also susceptible to and suffers from the same issue with small `disc` as discussed above.

3 Context-aware Importance Weighting

Observe that $J(h,\Phi)$'s first term in Eq. (1) tries to minimize a weighted sum of the factual losses – *i.e.*, a standard supervised machine learning objective. We can re-write this term as:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \omega_i \cdot L\big[\, y_i,\, h^{t_i}\big(\Phi(x_i)\big)\,\big] \\
&\quad= \frac{1}{N}\sum_{t\in\mathcal{T}} N_t \frac{1}{N_t}\sum_{j=1}^{N_t} \omega_j \cdot L\big[\, y_j,\, h^{t}\big(\Phi(x_j)\big)\,\big] \\
&\quad= \sum_{t\in\mathcal{T}} \Pr(t)\frac{1}{N_t}\sum_{j=1}^{N_t} \omega_j \cdot L\big[\, y_j,\, h^{t}\big(\Phi(x_j)\big)\,\big] \qquad (3)
\end{aligned}$$

where N_t is the number of instances assigned to the treatment arm $t \in \{0,1\}$.

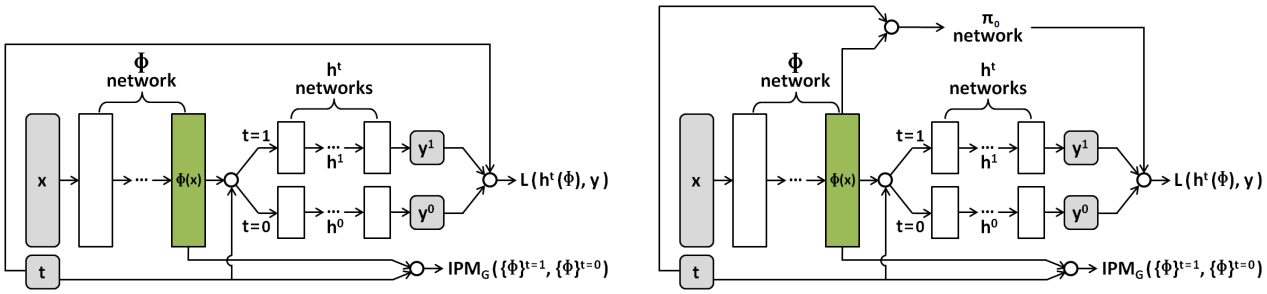
Using Eq. (2), SJS is basically setting $\omega_i = \frac{1}{2\Pr(t_i)}$, where $\Pr(t_i)$ is simply the observed probability of using the treatment $t_i \in \{0,1\}$ over the entire population. This effectively reduces the loss term in Eq. (3) to the macro-average $\frac{1}{2}\sum_{t\in\mathcal{T}}\frac{1}{N_t}\sum_{j=1}^{N_t} L\big[\, y_j,\, h^{t_j}\big(\Phi(x_j)\big)\,\big]$. In other words, different treatment arms contribute equally to the objective, irrespective of their sample size. This somewhat makes sense since, at test time, we want to estimate the outcomes of *all* possible treatments.

Such weights, however, do not account for the remainder selection bias in $\Phi(x)$ due to the presence of confounding factors B (see Fig. 3).[4] In our work, inspired by the importance sampling technique, we propose *context-aware* weights that incorporate the valuable context information of each instance $\Phi(x)$, thus further mitigating the impact of selection bias on estimating ITEs.

Importance sampling is used to compute $\mathbb{E}_{x\sim p(x)}\big[f(x)\big]$ when in fact we observe samples that are drawn from an alternative distribution $q(x)$, where p and q are called the "nominal" and "importance" distributions respectively. It is easy to show that $\mathbb{E}_{x\sim p(x)}\big[f(x)\big] = \mathbb{E}_{x\sim q(x)}\big[f(x)\frac{p(x)}{q(x)}\big]$ (see Appendix for proof). In the task of ITE estimation, we have a similar problem. Therefore, we need to first identify the

[3] While this approach is overkill – as computing an optimal policy only requires *ranking* the potential treatments – we focus on this approach as predicting these exact outcomes is valuable to both patients as well as insurance companies: knowing the margin of effect would hopefully increase compliance in the former and persuade the latter to accommodate the more expensive treatment.

[4] The `disc` term tries to balance the two distributions by pushing to eliminate factors A and B from Φ, while the factual loss term fights to keep B in Φ. Due to this trade-off, we anticipate that Φ will learn to eliminate A and keep B and C. Note it is critical that Φ includes B as it contributes to accurately predicting the outcome (y) and is critical to correctly modeling the un-removable selection bias.



(a) [Shalit *et al.*, 2017]'s model architecture (b) The proposed model architecture

Figure 5: Comparing model architectures for ITE estimation. Note the addition of the propensity network in our method versus SJS.

importance distribution that generated the data, then design a nominal distribution that helps improve the performance.

Re-visiting Eq. (3), our solution strategy is to learn an independent regression function $h^t\big(\Phi(x)\big)$ for each treatment arm $t \in \{0,1\}$ that predicts the outcome of the respective treatment t for subject x. By decoupling the weights from $J(h,\Phi)$'s parameters via setting $\phi = \Phi(x)$, we arrive at the following belief net: $t \leftarrow x \rightarrow \phi \rightarrow \{y^1, y^0\}$. The importance distribution of $L\big[y,\ h^t(\phi)\big]$ is then:

$$\Pr(y, \phi \mid t) \ = \ \Pr(y \mid \phi) \cdot \Pr(\phi \mid t)$$

We choose $\Pr(y, \phi \mid \neg t)$ as our nominal distribution in order to emphasize those instances that are important for predicting accurate **counterfactual** outcomes. This yields the likelihood ratio of $\frac{\Pr(y,\phi \mid \neg t)}{\Pr(y,\phi \mid t)} = \frac{\Pr(y \mid \phi)\cdot\Pr(\phi \mid \neg t)}{\Pr(y \mid \phi)\cdot\Pr(\phi \mid t)} = \frac{\Pr(\phi \mid \neg t)}{\Pr(\phi \mid t)}$. Moreover, to ensure that our model also performs well on the observed instances (associated with t_i), we add $\frac{\Pr(\phi_i \mid t_i)}{\Pr(\phi_i \mid t_i)} = 1$ to the derived likelihood ratio so that our objective accounts for the **factual** loss as well. Our weights would then be:

$$\omega_i \ = \ 1 \ + \ \frac{\Pr(\phi_i \mid \neg t_i)}{\Pr(\phi_i \mid t_i)} \tag{4}$$

Note these ω_i weights depend on ϕ_i whose numerical values are derived from $\Phi(x_i)$. This means that estimating these weights adds a nested optimization loop (for learning the $\omega(\cdot)$ parameters) within the main optimization loop (for learning the $\Phi(\cdot)$ and $h^t(\cdot)$ parameters). This motivates us to devise an efficient method for learning the weights. In this sense, learning the weights directly is not desirable because:

- It requires fitting two density functions: $\Pr(\phi \mid t)$ and $\Pr(\phi \mid \neg t)$ that doubles the necessary computations.
- Efficient solutions, such as fitting simple multi-variate Gaussians, are anticipated to yield inaccurate densities.
- More flexible solutions, such as fitting Gaussian mixture models, are of high computational complexity.

To circumvent these issues, we use the Bayes theorem to learn $\Pr(\phi \mid t)$ indirectly from $\pi_0(t \mid \phi)$ – *i.e.*, probability of selecting the assigned treatment t given the context ϕ – which can be efficiently obtained by fitting a Logistic Regression (LR) model. As a result, the counterfactual part of our proposed weight function can be simplified as follows:

$$\frac{\Pr(\phi_i \mid \neg t_i)}{\Pr(\phi_i \mid t_i)} \ = \ \frac{\frac{\pi_0(\neg t_i \mid \phi_i)\ \cdot\ \Pr(\phi_i)}{\Pr(\neg t_i)}}{\frac{\pi_0(t_i \mid \phi_i)\ \cdot\ \Pr(\phi_i)}{\Pr(t_i)}}$$

$$= \ \frac{\Pr(t_i)}{\Pr(\neg t_i)} \cdot \frac{\pi_0(\neg t_i \mid \phi_i)}{\pi_0(t_i \mid \phi_i)} \ = \ \frac{\Pr(t_i)}{1-\Pr(t_i)} \cdot \frac{1-\pi_0(t_i \mid \phi_i)}{\pi_0(t_i \mid \phi_i)} \tag{5}$$

where $\pi_0(t \mid \phi)$ is parametrized by LR with $[W, b]$ as:

$$\pi_0(t \mid \phi) \ = \ \frac{1}{1 + e^{-(2t-1)(\phi \cdot W + b)}}$$

and parameters $[W, b]$ are learned by minimizing:

$$C(W,b) \ = \ \frac{1}{N}\sum_{i=1}^{N} -\log\big[\pi_0(t_i \mid \phi_i)\big] \tag{6}$$

Since π_0 depends on Φ, we update $[W, b]$ with every update of the parameters of Φ and h. Hence, this is a multi-objective optimization problem with two objectives – *i.e.*, Eqs. (1) and (6) – that we try to solve alternatingly. That is, each training iteration consists of two steps:

(i) Minimizing Eq. (1) using stochastic gradient descent to update the parameters of the representation and hypothesis networks – *i.e.*, U and V. Note that ω_is in the factual loss term are calculated based on Eqs. (4) and (5), with parameters W and b held fixed during optimization.

(ii) Minimizing Eq. (6) to update parameters of the propensity score function $\pi_0(t \mid \phi)$ – *i.e.*, W and b – with parameters U and V held fixed.

Algorithm 1 describes this procedure in more details. Note that both objective functions are computed for one mini-batch at a time. Fig. 5(b) illustrates our network architecture.

4 Experiments

As mentioned earlier, an inherent characteristic of causal inference datasets is that counterfactual outcomes are unobservable, which makes it difficult to evaluate any proposed algorithm. The common solution in the literature is to synthesize datasets where the outcomes of all possible treatments are available.

Algorithm 1 CFR-ISW: CounterFactual Regression with Importance Sampling Weights

Input: Factual samples $\{[x_1, t_1, y_1], ..., [x_N, t_N, y_N]\}$, batch size m, scaling parameter $\alpha > 0$, regularization parameter $\lambda > 0$, loss function $L(\cdot, \cdot)$, representation network Φ_U with initial weights $[U]$, outcome network h_V with initial weights $[V]$, function family for IPM, propensity network π with initial weights $[W, b]$, and limit on the total number of iterations I.
Estimate probabilities $\Pr(t)$ for $t \in \{0, 1\}$
for $iter = 1$ **to** I **do**
  Sample mini-batch $\{i_1, i_2, ..., i_m\} \subset \{1, 2, ..., N\}$
  Calculate the gradient of the discrepancy term:
    $g_d = \nabla_U \text{IPM}(\{\Phi_U(x_{i_j})\}_{t_{i_j}=0}, \{\Phi_U(x_{i_j})\}_{t_{i_j}=1})$
  Calculate the proposed importance sampling weights ω_{i_j} from W and $\Pr(t)$ following Eq. (5)
  Calculate the gradients of the empirical loss:
    $g_U = \nabla_U \frac{1}{m} \sum_j \omega_{i_j} \cdot L\big[h_V^{t_{i_j}}\big(\Phi_U(x_{i_j})\big), y_{i_j}\big]$
    $g_V = \nabla_V \frac{1}{m} \sum_j \omega_{i_j} \cdot L\big[h_V^{t_{i_j}}\big(\Phi_U(x_{i_j})\big), y_{i_j}\big]$
  Obtain step size scalar or matrix η_1 with standard neural net methods (*e.g.*, `Adam` [Kingma and Ba, 2015])
  Update weights of the representation and hypothesis networks:
    $[U, V] \leftarrow \big[U - \eta_1(\alpha g_d + g_U),\ V - \eta_1(g_V + 2\lambda V)\big]$
  Calculate gradients of the propensity network's cost function:
    $g_W = \nabla_W \frac{1}{m} \sum_j \log\big[1 + e^{-(2t_{i_j}-1)(\Phi_U(x_{i_j})\cdot W + b)}\big]$
    $g_b = \nabla_b \frac{1}{m} \sum_j \log\big[1 + e^{-(2t_{i_j}-1)(\Phi_U(x_{i_j})\cdot W + b)}\big]$
  Obtain $\eta_2 \in \mathbb{R}^+$ % `distance to move`
  Update the propensity network's weights:
    $[W, b] \leftarrow [W, b] - \eta_2[g_W, g_b]$
end for
Output: $[U, V]$

Some entries are then discarded in order to create a proper observational dataset with characteristics (such as selection bias) similar to a real-world one – see for example [Hassanpour and Greiner, 2018] and [Beygelzimer and Langford, 2009]. To make performance comparison easier, however, we do not synthesize our own datasets here. Instead, we use two publicly available benchmarks – see Sec. 4.3.

4.1 Evaluation Criteria

There are two categories of performance measures for evaluating causal effect estimation algorithms: individual-based and population-based. Our main focus here is producing models with high individual-based performance, as measured by "Precision in Estimation of Heterogeneous Effect" (PEHE) [Hill, 2011] and "Effect-Normalized Root Mean Squared Error" (ENoRMSE) [Shimoni *et al.*, 2018; Karavani *et al.*, 2018]:

$$\text{PEHE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\text{e}}_i - \text{e}_i)^2}$$

$$\text{ENoRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\hat{\text{e}}_i}{\text{e}_i}\right)^2}$$

where $\hat{\text{e}}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the predicted effect and $\text{e}_i = y_i^1 - y_i^0$ is the true effect. We also consider a population-based performance measure, namely, bias of the "Average Treatment Effect (ATE)": $\epsilon_{\text{ATE}} = \left|\text{ATE} - \widehat{\text{ATE}}\right|$ where $\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} y_i^1 - \frac{1}{N}\sum_{j=1}^{N} y_j^0$ in which y_i^1 and y_j^0 are the true outcomes for the treatment and control arms respectively[5] and $\widehat{\text{ATE}}$ is calculated based on the estimated outcomes.

4.2 Hyperparameter Selection

As counterfactuals are unobserved, it is impossible for our learning algorithm to perform standard internal cross-validation, to set the hyperparameters. Therefore, our learner needs to obtain some estimate $\tilde{\text{e}}_i$ of the true effect $\text{e}_i = y_i^1 - y_i^0$, so that it can calculate a surrogate for its desired performance measure. SJS estimated the outcome of $y(x_i, \neg t_i)$ as the observed outcome $y_{j(i)}^{\neg t_i}$, where $j(i)$ is the nearest neighbor of x_i who received treatment $\neg t_i$ (*i.e.*, 1-NN based on a distance metric defined on the original x space). The surrogate effect would then be $\tilde{\text{e}}_{\text{1-NN}} = (2t_i - 1)(y_i^{t_i} - y_{j(i)}^{\neg t_i})$.

However, as our empirical results also confirm, this method is quite unlikely to select good hyperparameters. This is expected since, due to selection bias, the nearest neighbor $j(i)$ in the alternative treatment arm might not be a good enough representative of the counterfactual outcome. Hence, its estimated surrogate effect might not be reliable for finding the best set of hyperparameters.

A better solution is to employ a stronger counterfactual regression method – such as Bayesian Additive Regression Trees (BART) [Chipman *et al.*, 2010]. This is interesting because, even though our empirical results (see Sec. 4.3) show that BART's performance is not as good as either CFR or CFR-ISW, $\tilde{\text{e}}_{\text{BART}}$ identifies much better set of hyperparameters (via $\text{PEHE}_{\text{BART}}$ or $\text{ENoRMSE}_{\text{BART}}$) compared to $\tilde{\text{e}}_{\text{1-NN}}$.

4.3 Results and Discussion

In this paper, we empirically compare the proposed CFR-ISW with the following ITE estimation methods[6]:

- 1-NN: One Nearest Neighbor method (as described in Sec. 4.2) – the baseline.
- BART: Bayesian Additive Regression Trees method [Chipman *et al.*, 2010].
- CFR: CounterFactual Regression method (*i.e.*, SJS).
- RCFR: Re-weighted CFR [Johansson *et al.*, 2018].[7]

[5] We can calculate ATE here since we work with a synthetic dataset and so have access to both observed and counterfactual outcomes. In RCTs, the Sample Average Treatment Effect (SATE) $= \frac{1}{N_1}\sum_{i=1}^{N_1} y_i^1 - \frac{1}{N_0}\sum_{j=1}^{N_0} y_j^0$ is used as a proxy for the true ATE, where N_1 (N_0) is the number of treated (controlled) subjects and y_i^1 (y_j^0) is the outcome of subject i (j) upon receiving treatment (control).

[6] While these are only a subset of the many methods in the literature, Table 1 of [Shalit *et al.*, 2017] establishes that CFR significantly outperforms several notable ones such as Random Forest [Breiman, 2001], Causal Forest [Wager and Athey, 2018], and Targeted Maximum Likelihood Estimation [Gruber and van der Laan, 2011].

[7] As RCFR's code is unavailable, we are limited in comparing its performance against contenders to what is reported in their paper.

	METHODS	ENORMSE	PEHE	ϵ_{ATE}
	1-NN	24.6 (189)	4.85 (6.29)	0.67 (1.27)
	BART	2.13 (11.3)	1.57 (2.41)	0.22 (0.30)
	CFR†		0.78 (0.0)	0.31 (0.01)
	RCFR‡		0.65 (0.04)	
P1	**CFR**	2.65 (1.67)	0.88 (0.10)	0.20 (0.03)
P1	**CFR-ISW**	3.82 (3.17)	**0.77** (0.10)	**0.19** (0.03)
PB	**CFR**	1.87 (1.29)	0.65 (0.05)	0.21 (0.03)
PB	**CFR-ISW**	2.50 (2.05)	**0.55** (0.05)	**0.20** (0.03)
EB	**CFR**	1.18 (0.29)	0.84 (0.07)	0.23 (0.03)
EB	**CFR-ISW**	**0.88** (0.29)	**0.66** (0.05)	**0.16** (0.02)

Table 1: ENoRMSE, PEHE, and ϵ_{ATE} performance measures (lower is better), each of the form "mean (standard deviation)" on the **IHDP** benchmark. Symbols † and ‡ indicate results reported in [Shalit *et al.*, 2017] and [Johansson *et al.*, 2018] respectively. Rows P1, PB, and EB report results of our runs for CFR and CFR-ISW whose hyperparameters were selected based on $\text{PEHE}_{\text{1-NN}}$, $\text{PEHE}_{\text{BART}}$, and $\text{ENoRMSE}_{\text{BART}}$ respectively. Comparing CFR-ISW with CFR, entries in **bold** indicate the best performance in each category (statistically significant based on the Welch's unpaired t-test with $\alpha = 0.05$).

Below, we explain the characteristics of the two benchmarks used for evaluation. We also discuss the performance of the proposed method and compare it with its contenders.

Infant Health and Development Program (IHDP)

IHDP is a synthetic binary-treatment dataset, designed to evaluate the effect of specialist home visits on future cognitive test scores of premature infants. Hill [2011] induced selection bias by removing a non-random subset of the treated population from the original RCT data in order to create a realistic observational dataset. The resulting dataset contains 747 instances (608 control, 139 treated) with 25 covariates that measure different attributes of infants and their mothers.

We worked with the same dataset provided by and used in [Shalit *et al.*, 2017; Johansson *et al.*, 2016; Johansson *et al.*, 2018], in which outcomes are simulated as setting "A" of the Non-Parametric Causal Inference (NPCI) package [Dorie, 2016]. The noiseless outcomes are used to compute the true individual effects (available for evaluation purpose only). We report the methods' performances by averaging over 100 realizations of outcomes with 63/27/10 train/validation/test splits.

Table 1 reports ENoRMSE, PEHE, and ϵ_{ATE} performances of the considered methods on the IHDP dataset. Our results show that CFR-ISW significantly outperforms CFR and RCFR in all three evaluation measures. Note that $\tilde{\text{e}}_{\text{BART}}$ selects better hyperparameters than $\tilde{\text{e}}_{\text{1-NN}}$ – compare P1 and PB rows. Also note that we should use a proper surrogate measure for hyperparameter selection depending on the performance measure that we would like to optimize – compare PB and EB rows. This is expected, since, there is no way to encode such a criterion in the objective function that is being optimized.

Atlantic Causal Inference Conference 2018 (ACIC'18)

ACIC'18 is a collection of 24 synthetic binary-treatment datasets released for a data challenge; with number of instances $n_m \in \{1, 2.5, 5, 10, 25, 50\} \times 10^3$ (four datasets in each category) for $m \in \{1, ..., 24\}$, each comprised of 177 features. The covariates matrix for each of these datasets are sub-sampled from a covariates table of real-world medical measurements taken from the Linked Birth and Infant Death Data (LBIDD) [MacDorman and Atkinson, 1998], that contains information corresponding to 100,000 subjects.

	DATASETS	1-NN	BART	CFR	CFR-ISW
	ALL	54.56	9.35	5.43 (5.78)	**1.03** (0.27)
# INSTANCES	**1** k	66.70	73.66	7.08 (8.97)	1.54 (0.87)
# INSTANCES	**2.5** k	33.31	15.12	8.33 (14.78)	0.68 (0.31)
# INSTANCES	**5** k	31.89	8.15	2.00 (2.28)	0.88 (0.35)
# INSTANCES	**10** k	31.46	2.60	0.86 (1.00)	0.74 (0.39)
# INSTANCES	**25** k	19.47	1.27	0.85 (0.30)	1.00 (0.28)
# INSTANCES	**50** k	75.43	12.27	8.23 (8.63)	1.13 (0.23)

Table 2: Aggregated ENoRMSE (lower is better) on the **ACIC'18** benchmark. Model hyperparameters for both CFR and CFR-ISW methods are selected according to $\text{ENoRMSE}_{\text{BART}}$. Comparing CFR-ISW with CFR, entry in **bold** indicates significantly better performance (Welch's unpaired t-test with $\alpha = 0.05$).

For each of the 24 datasets, we have access to both factual and counterfactual tables. For each subject, factual tables contain the treatment bit and the respective observed outcome. Counterfactual tables (only to be used for evaluation purpose) contain the true outcomes $\{y^0,\ y^1\}$ for treatments 0 and 1 respectively. For each synthetic dataset, a Data Generating Process (DGP) determines t, y^0, and y^1 for each sampled x instance. The challenge organizers have not revealed the used DGPs. Here, we look at two evaluation measures: (i) the aggregated ENoRMSE for datasets with the same number of instances (*i.e.*, A_n for $n \in S = \{1, 2.5, 5, 10, 25, 50\} \times 10^3$), where S is the set of different dataset sizes; and (ii) the aggregated ENoRMSE of all the 24 datasets (*i.e.*, A). A_n and A respectively are calculated as follows:

$$A_n = \sqrt{\frac{1}{|D_n|} \sum_{i \in D_n} \left[\text{ENoRMSE}(i)\right]^2}\ ,\ A = \sqrt{\frac{1}{\sum_{n \in S} n} \sum_{n \in S} n A_n^2}$$

where D_n is set of all datasets that have n instances.

Table 2 summarizes the macro-average performances of the four methods on the ACIC'18 datasets in terms of aggregated ENoRMSE. Our empirical results indicate that incorporating the proposed context-aware importance sampling weights into the network's objective function improves the aggregated ENoRMSE on all datasets significantly and by a large margin. We also computed the micro-average performances (not shown) which confirms that, as expected, CFR-ISW significantly outperforms CFR in all categories as well.

5 Future Works and Conclusion

Currently, this approach can only be applied to *binary*-treatment datasets. We plan to explore ways to facilitate counterfactual regression when multiple (categorical) treatments are available; or even real-valued treatment options – such as predicting the right dosage of insulin for diabetic patients.

In this work, we proposed a context-aware importance sampling weighting scheme that helps mitigate the negative effect of selection bias on the accuracy of models that estimate Individual Treatment Effects (ITEs). Additionally, we proposed a hyperparameter selection procedure, which plays an important role in determining the model performance. The proposed improvements were applied to the CounterFactual Regression (CFR) framework [Shalit *et al.*, 2017], leading to our method: CFR with Importance Sampling Weights (CFR-ISW).

We evaluated CFR-ISW against 1-NN (baseline), Bayesian Additive Regression Trees (BART), and the state-of-the-art methods CFR and Re-weighted CFR on two publicly available synthetic benchmarks: (i) Infant Health and Development Program (IHDP) and (ii) Atlantic Causal Inference Conference 2018 (ACIC'18) data challenge. The empirical results demonstrated that CFR-ISW significantly ($p < \alpha = 0.05$) outperforms all the contender methods in terms of three common measures of performance for estimating causal effects, namely: Precision in Estimation of Heterogeneous Effect (PEHE), Effect-Normalized Root Mean Squared Error (ENoRMSE), and bias of the Average Treatment Effect (ϵ_{ATE}).

Appendix

A Selection Bias Entails Covariate Shift

Here, we want to prove that existence of selection bias in data $\Pr(T \mid X) \neq \Pr(T)$ entails covariate shift: $\Pr(X, T) \neq \Pr(X, \neg T)$.

Proof by contraposition:

Assume $\Pr(X, T) = \Pr(X, \neg T)$, then:

$$\begin{aligned} \Pr(T \mid X) \cdot \Pr(X) &= \Pr(\neg T \mid X) \cdot \Pr(X) \\ \implies \quad \Pr(T \mid X) &= \Pr(\neg T \mid X) \\ \implies \quad T &\perp\!\!\!\perp X \\ \implies \quad \Pr(T \mid X) &= \Pr(T) \end{aligned}$$

Having proved the contrapositive $\Pr(T \mid X) = \Pr(T)$, we infer the original statement $\Pr(T \mid X) \neq \Pr(T) \implies \Pr(X, T) \neq \Pr(X, \neg T)$ to hold. □

B Importance Sampling

Here, we want to show $\mathbb{E}_{x\sim p(x)}\big[f(x)\big] = \mathbb{E}_{x\sim q(x)}\big[f(x)\frac{p(x)}{q(x)}\big]$, where p and q are probability density functions defined on $\mathbb{R}^d$, with $p(x) \neq 0 \quad \forall\, x \in \mathcal{D}$ and $p(x) = 0$ otherwise, and $q(x) > 0$ for $x \in \mathcal{Q}$ where $f(x)p(x) \neq 0$, then:

$$\begin{aligned} &\mathbb{E}_{x\sim q(x)}\big[f(x)\frac{p(x)}{q(x)}\big] \\ &= \int_{\mathcal{Q}} \frac{f(x)p(x)}{q(x)} q(x)\mathrm{d}x \quad = \quad \int_{\mathcal{Q}} f(x)p(x)\mathrm{d}x \\ &= \int_{\mathcal{D}} f(x)p(x)\mathrm{d}x + \int_{\mathcal{D}^c\cap\mathcal{Q}} f(x)p(x)\mathrm{d}x - \int_{\mathcal{D}\cap\mathcal{Q}^c} f(x)p(x)\mathrm{d}x \\ &= \int_{\mathcal{D}} f(x)p(x)\mathrm{d}x \quad = \quad \mathbb{E}_{x\sim p(x)}\big[f(x)\big] \end{aligned}$$

since $p(x) = 0$ for $x \in \mathcal{D}^c \cap \mathcal{Q}$ and $f(x) = 0$ for $x \in \mathcal{D} \cap \mathcal{Q}^c$. □

Parameter name	Range
Imbalance parameter α	1E{-2, -1, 0, 1}
Num. of representation layers	{3, 5}
Num. of hypothesis layers	{3, 5}
Dim. of representation layers	{50, 100, 200}
Dim. of hypothesis layers	{50, 100, 200}
Batch size	{100, 300}

Table 3: Hyperparameters and ranges

C Proposed Weighting Scheme: Intuition

To illustrate the idea (in a trivialized fashion), imagine subject S received treatment $T0$, but his 10 clones $\{S_1, \ldots, S_{10}\}$ were each observed to receive treatment $T1$. How much should we weight our estimate of $h^{T0}(S)$? One component is based on the fact that we observed $[S,\ T0]$, which should contribute $\Pr(\Phi(S) \mid T0)$. But later, to estimate the ITE for each clone S_i, our algorithm will want to know what-would-have-happened had S_i received $T0$. In this situation, that would also be $h^{T0}(S)$. Hence, the weight should also include the density of instances that look like S, but received the other treatment – *i.e.*, $\Pr(\Phi(S) \mid T1)$ – which here would be based on the 10 clones S_i. Of course, the real situation is much more complicated, as we will not typically have exact clones. In general, this suggests that the weight associated with observing $[\phi_i,\ t_i]$ should be $\Pr(\phi_i \mid t_i) + \Pr(\phi_i \mid \neg t_i)$, normalized in the expectation by dividing by $\Pr(\phi_i \mid t_i)$.

D Hyperparameters

We trained CFR-ISW's π_0 logistic regression function with gradient descent optimizer and a learning rate of 1E-3.

For both CFR and CFR-ISW, we trained the Φ and h^t networks with regularization coefficient λ=1E-3, `elu` as the non-linear activation function, `Adam` optimizer [Kingma and Ba, 2015], learning rate of 1E-3, and maximum number of iterations of 3000. We used the Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012] as our IPM to calculate `disc` between the $\Pr(\Phi \mid t=1)$ and $\Pr(\Phi \mid t=0)$ distributions. See Table 3 for details on our hyperparameter search space.

Acknowledgements

The authors gratefully acknowledge financial support from Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Machine Intelligence Institute (Amii). We wish to thank Dr. Martha White and Junfeng Wen for fruitful conversations, and Dr. Fredrik Johansson for publishing/maintaining the code-base for the CFR method online.

References

[Atan *et al.*, 2018] Onur Atan, James Jordon, and Mihaela van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *AAAI*, pages 2071–2078, 2018.

[Austin, 2011] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[Beygelzimer and Langford, 2009] Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *15th ACM SIGKDD*. ACM, 2009.

[Bottou *et al.*, 2013] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *JMLR*, 14(1), 2013.

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[Chipman *et al.*, 2010] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 2010.

[Dorie, 2016] Vincent Dorie. NPCI: Non-parametrics for causal inference, 2016. https://github.com/vdorie/npci.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(Mar):723–773, 2012.

[Gruber and van der Laan, 2011] Susan Gruber and Mark J van der Laan. tmle: An R package for targeted maximum likelihood estimation. 2011.

[Hassanpour and Greiner, 2018] Negar Hassanpour and Russell Greiner. A novel evaluation methodology for assessing off-policy learning methods in contextual bandits. In *Canadian AI*, pages 31–44, 2018.

[Hill, 2011] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[Imbens and Rubin, 2015] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[Imbens and Wooldridge, 2009] Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

[Johansson *et al.*, 2016] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, pages 3020–3029, 2016.

[Johansson *et al.*, 2018] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

[Karavani *et al.*, 2018] Ehud Karavani, Yishai Shimoni, and Chen Yanover. IBM causal inference benchmarking framework, 2018. https://github.com/IBM-HRL-MLHLS/.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *19th International Conference on World Wide Web*. ACM, 2010.

[Li *et al.*, 2015] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *24th International Conference on World Wide Web*. ACM, 2015.

[MacDorman and Atkinson, 1998] Marian F MacDorman and Jonnae O Atkinson. Infant mortality statistics from the 1996 period linked birth/infant death data set. *Monthly vital statistics report*, 46(12):1980–92, 1998.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[Rollinson and Brunskill, 2015] Joseph Rollinson and Emma Brunskill. From predictive models to instructional policies. *International Educational Data Mining Society*, 2015.

[Rosenbaum and Rubin, 1983] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.

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

[Shalit *et al.*, 2017] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2), 2000.

[Shimoni *et al.*, 2018] Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

[Strehl *et al.*, 2010] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *NeurIPS*, pages 2217–2225. 2010.

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT Press Cambridge, 1998.

[Swaminathan and Joachims, 2015a] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *JMLR*, 16, 2015.

[Swaminathan and Joachims, 2015b] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *NeurIPS*, 2015.

[Wager and Athey, 2018] 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.