CounterFactual Regression with Importance Sampling Weights

Negar Hassanpour [1] **Russell Greiner** [1]

Abstract

Perhaps the most pressing concern of a patient recently diagnosed with cancer is her life expectancy for various treatment options. Answering such questions requires estimating the unobserved (*i.e.*, counterfactual) outcomes of the treatments that were not administered for each patient in the training data. This "counterfactual challenge" applies not only to healthcare, but also to other fields such as education, economics, etc. This paper extends the work of Shalit *et al.* (2017) for estimating Individualized Treatment Effect (ITE) in two directions: modifying (i) the objective function by adding importance sampling weights, and (ii) the procedure for finding the best set of model hyperparameters. Our evaluation on the synthetic datasets from the 2018 Atlantic Causal Inference Conference Data Challenge[1] demonstrated significantly better performance of the proposed weighting scheme over that of (Shalit et al., 2017).

1. Introduction

Precision medicine – *i.e.*, the customization of healthcare tailored to each individual patient – requires making predictions about causal effects of various treatments. This requires identifying which medical procedure would benefit each individual patient the most. Such analysis is not limited to healthcare, as it can be used in any field where personalization would be of value, including education, economics, public welfare, etc.

In datasets that we work with, for each instance i (*e.g.*, a patient), we have some context information x_i (*e.g.*, her age, blood work, etc.), the administered treatment t_i chosen from a set of treatment options T (*e.g.*, $\{0\!:$ surgery, $1\!:$ medication$\}$) and the respective observed outcome y_i (*e.g.*, her survival time) after receiving treatment t_i. These are *observational studies* – *i.e.*, the principal investigator who collects the data has no control over the treatment assignment procedure (which might be biased by the clinician's decisions) and merely records the values of interest.

In such datasets, both outcome and the treatment assignment are dependent on some or all of the context information. [2] This is why sample selection bias is an inherent characteristic of observational datasets. The probabilistic graphical model of observational studies is illustrated in Figure 1(a). Figure 1(b) shows an example observational dataset, where a doctor prescribes surgery $(t = 0)$ to younger patients and medication $(t = 1)$ to older ones[3]. Note that, for each patient i, we only observe the outcome y_i of the received treatment t_i (*i.e.*, either surgery or medication, but not both). Figure 1(b) uses a **+** to denote a patient who received surgery $(t_i = 0)$ or a ● for a patient who received medication $(t_i = 1)$. The performance system can never observe her counterfactual outcome – *i.e.*, the outcome of the alternative treatment that was not administered; however, in synthetic datasets, counterfactuals are available for evaluation purposes. These are shown as ● (+) for patients who – in reality– received treatment 0 (1).

Estimating causal effects from observational data is different from standard supervised machine learning in that training data never contains the counterfactual outcomes. This is closely related to "learning from logged bandit feedback" in the literature (*cf.* (Strehl et al., 2010; Li et al., 2010; 2011; 2015; Swaminathan & Joachims, 2015)). The only distinction here is that, unlike applications such as ad-placement (Bottou et al., 2013), we do not have access to the underlying mechanism for treatment assignment.

2. Method

This paper extends the work of Shalit *et al.* (2017) in the following two directions: modifying (i) the objective func-

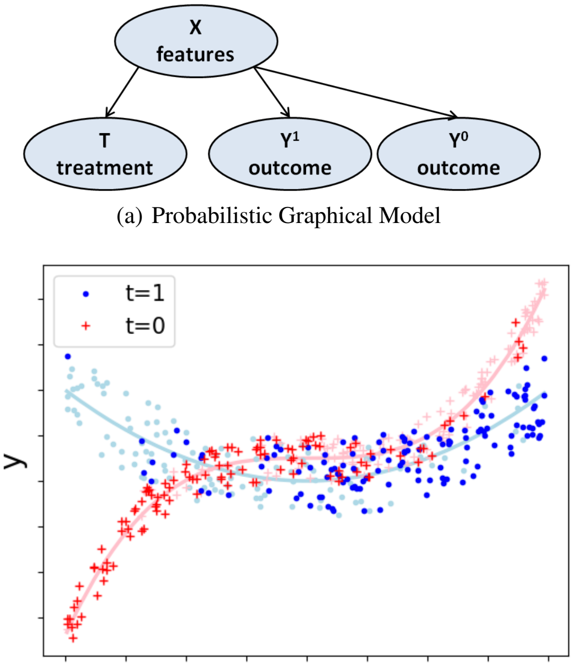
[1] Department of Computing Science, University of Alberta, Edmonton, Canada. Correspondence to: Negar Hassanpour <hassanpo@ualberta.ca>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

[1] www.cmu.edu/acic2018/data-challenge/

[2] This work assumes the "ignorability assumption": there are no covariates that contribute to both treatment selection and outcome (a.k.a., confounders), but are not recorded in x.

[3] Perhaps because recovery after surgery is much faster for younger patients; and that it is just not cost effective to perform such an invasive procedure on older patients.



(a) Probabilistic Graphical Model

(b) Best viewed in color – Blue dots and red pluses illustrate the assigned treatment t with their respective observed outcome y. Light-blue dots and pink pluses illustrate the true counterfactual outcomes; these are never observed. Note that samples with higher (lower) x values have been assigned to $t = 1$ ($t = 0$) more frequently; hence we have sample selection bias.

Figure 1. Observational Dataset

tion by adding *sample-specific* importance sampling weights (see Sec. 2.2), and (ii) the procedure for finding the best set of model hyperparameters (see Sec. 3.2).

2.1. (Shalit et al., 2017)'s Model Overview

Shalit *et al.* (2017) attempt to reduce sample selection bias by learning a common representation space $\Phi(x)$ (Bengio et al., 2013) in which $\Pr(\, x \mid t = 0\,)$ and $\Pr(\, x \mid t = 1\,)$ are as close[4] as possible to each other, provided $\Phi(\,\cdot\,)$ retains enough information that a learned regression model $h^t(\,\Phi\,)$ can generalize well on factual outcomes. See Figure 2 for the model architecture; $h^t(\,\Phi\,)$ and $\Phi(\,x\,)$ are parameterized by deep neural networks trained jointly in an end-to-end fashion.

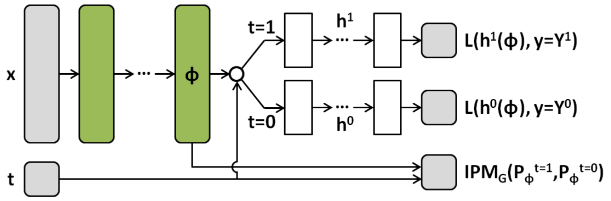


Figure 2. The neural network architecture proposed by (Shalit et al., 2017) for estimating the Individualized Treatment Effect (ITE).

The following is (Shalit et al., 2017)'s objective function:

$$J(h,\Phi) = \min_{h,\Phi} \frac{1}{n}\sum_{i=1}^{n} \omega_i \cdot L[\, h^{t_i}\,(\Phi(x_i)),\, y_i\,] + \lambda\cdot\mathcal{R}(h) + \alpha\cdot \text{IPM}_G(\,\{\Phi(x_i)\}_{i:\,t_i=0},\; \{\Phi(x_i)\}_{i:\,t_i=1}\,) \tag{1}$$

where $L[\, h^{t_i}\,(\Phi(x_i)),\, y_i\,] = [\, h^{t_i}\,(\Phi(x_i)) - y_i\,]^2$ is the squared loss of predicting the factual outcome for sample i,

$$\omega_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}, \quad \text{with} \quad u = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{2}$$

$\mathcal{R}(h)$ is the regularization term for penalizing model complexity, and $\text{IPM}_G\,(\{\Phi(x_i)\}_{i:\,t_i=0}, \{\Phi(x_i)\}_{i:\,t_i=1})$ is the measure of distance between the two distributions $\Pr(\, x \mid t = 0\,)$ and $\Pr(\, x \mid t = 1\,)$.

The first term of $J(h,\Phi)$ minimizes a weighted sum of the factual loss, which is a standard supervised machine learning objective. More specifically, we can see from Eq. (2) that $2\omega_i = \frac{1}{\Pr(t_i)}$, where $\Pr(\, t_i\,)$ is simply the frequentist probability of assigning $t_i \in \{0, 1\}$ to the whole population. Notice that $2\omega_i$ can be rewritten as follows:

$$\frac{1}{\Pr(\,t_i\,)} = \frac{\Pr(\,t_i\,)}{\Pr(\,t_i\,)} + \frac{1-\Pr(\,t_i\,)}{\Pr(\,t_i\,)} = 1 + \frac{\Pr(\,\neg t_i\,)}{\Pr(\,t_i\,)} \tag{3}$$

where $\neg t_i$ represents the alternative treatment that was **not** administered to patient i.

This weighting scheme makes sense because the first addend, "1" (corresponding to the "factual part" of the weight function) minimizes the factual loss (on the observed outcomes); this is the simplest mission of any regression model. The second weight term (*i.e.*, $\frac{\Pr(\,\neg t_i\,)}{\Pr(\,t_i\,)}$; corresponding to the "counterfactual part" of the weight function), on the other hand, attempts to re-weight the factual loss to emphasize the loss for the instances in the treatment arm that has fewer instances.

2.2. Proposed Weighting Scheme

The weights in Eq. (3) attempt to balance the factual loss in accordance with the number of instances within each

[4] This *"closeness"* is measured based on the Integral Probability Metric (IPM) measure of distance. Specifically, they use the following two IPMs: (i) Maximum Mean Discrepancy (Gretton et al., 2012) and (ii) Wasserstein distance (Attouch et al., 2014; Cuturi & Doucet, 2014).

treatment arm. However, it might be beneficial to have weights that incorporate the context information of each instance x, in addition to its respective treatment t. For example, if an instance x_i (with assigned treatment t_i) is "far" from other instances with the same assigned treatment [5], then we want to make sure that our outcome prediction network learns this instance well. To do so, we replace the counterfactual part of the weight function in Eq. (2) with the following importance sampling weights:

$$\begin{aligned}\frac{\Pr(\,\Phi(x_i)\,|\,\neg t_i\,)}{\Pr(\,\Phi(x_i)\,|\,t_i\,)} &= \frac{\frac{\pi(\neg t_i|\Phi(x_i))\,\cdot\,\Pr(\Phi(x_i))}{\Pr(\neg t_i)}}{\frac{\pi(t_i|\Phi(x_i))\,\cdot\,\Pr(\Phi(x_i))}{\Pr(t_i)}}\\ &= \frac{\Pr(t_i)}{\Pr(\neg t_i)}\cdot\frac{\pi(\neg t_i|\Phi(x_i))}{\pi(t_i|\Phi(x_i))} \qquad (4)\\ &= \frac{\Pr(\,t_i\,)}{1-\Pr(\,t_i\,)}\cdot\frac{1-\pi(t_i|\Phi(x_i))}{\pi(t_i|\Phi(x_i))}\end{aligned}$$

where $\pi(\,t\,|\,\Phi(x)\,)$ is the probability of assigning treatment t given the context in the Φ representation space (a.k.a., propensity score). We use a Logistic Regression (LR) model to fit the propensity score function:

$$\pi(t|\Phi(x)) = \frac{1}{1+e^{-(2t-1)(\Phi(x)\cdot W+b)}} \tag{5}$$

To learn the LR parameters $[W, b]$, we attempt to minimize the following cost function:

$$\min_{W,b}\ \frac{1}{n}\sum_{i=1}^{n} C[\,W, b, \Phi(x), t\,] \tag{6}$$

$$\text{where} \quad C[\,W, b, \Phi(x), t\,] = -\log[\,\pi(t_i|\Phi(x_i))\,] \tag{7}$$

Since π depends on Φ, we update $[W, b]$ with every update of the parameters of Φ and h.

Hence, this is a multi-objective[6] optimization problem that we try to solve interactively. That is, each training iteration consists of two steps:

(i) minimize Eq. (1) using stochastic gradient descent and 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 Eq. (4), with parameters W and b held fixed during optimization.

(ii) minimize Eq. (6) with U and V held fixed and update the parameters of the propensity score function (*i.e.*, W and b).

This procedure is described in more details in Algorithm 1. Also note that both objective functions are computed for one mini-batch at a time.

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

```
Input: Factual samples $(x_1, t_1, y_1), ..., (x_n, t_n, y_n)$, scaling parameter $\alpha > 0$, regularization parameter $\lambda > 0$, loss function $L(\cdot,\cdot)$, representation network $\Phi_U$ with initial weights $[U]$, outcome network $h_V$ with initial weights $[V]$, function family $G$ for IPM, propensity network $\pi_W$ 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 $\{i_1, i_2, ..., i_m\} \in \{1, 2, ..., n\}$
    Calculate the gradient of the IPM term:
        $g_1 = \nabla_U \mathrm{IPM}_G(\{\Phi_U(x_{i_j})\}_{t_{i_j}=0}, \{\Phi_U(x_{i_j})\}_{t_{i_j}=1})$
    Calculate the proposed importance sampling weights $\omega_{i_j}$ from $W$ and Pr( $t$ ) following Eq. (4)
    Calculate the gradients of the empirical loss:
        $g_2 = \nabla_U \frac{1}{m}\sum_j \omega_{i_j} \cdot L[\,h_V^{t_{i_j}}(\Phi_U(x_{i_j})), y_{i_j}\,]$
        $g_3 = \nabla_V \frac{1}{m}\sum_j \omega_{i_j} \cdot L[\,h_V^{t_{i_j}}(\Phi_U(x_{i_j})), y_{i_j}\,]$
    Obtain step size scalar or matrix $\eta_1$ with standard neural net methods (e.g., Adam (Kingma & Ba, 2015))
    Update weights of the representation and hypothesis networks:
        $[U,\ V] \leftarrow [U - \eta_1(\alpha g_1 + g_2),\ V - \eta_1(g_3 + 2\lambda V)]$
    Calculate gradients of the propensity network's cost function:
        $g_4 = \nabla_W \frac{1}{m}\sum_j C(W, b, \Phi_U(x_{i_j}), t_{i_j})$
        $g_5 = \nabla_b\ \frac{1}{m}\sum_j C(W, b, \Phi_U(x_{i_j}), t_{i_j})$
    Obtain $\eta_2$
    Update the propensity network's weights:
        $[W,\ b] \leftarrow [W,\ b] - [\eta_2 g_4,\ \eta_2 g_5]$
end for
```

3. Experiments

3.1. Evaluation Criteria

We use Effect-Normalized Root Mean Squared Error

$$\text{ENoRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\hat{e}_i}{e_i}\right)^2} \quad \text{with} \quad \begin{cases}\hat{e}_i = \hat{y}_i^1 - \hat{y}_i^0\\ e_i = y_i^1 - y_i^0\end{cases} \tag{8}$$

to measure the accuracy for Individualized Treatment Effect Estimation (ITE), where $\hat{y}_i^1$ and $\hat{y}_i^0$ are predicted outcomes for treatments 1 and 0 respectively. Note that only one of y_i^1 or y_i^0 is observed during training (including hyperparameter selection). Therefore, ENoRMSE can be calculated only if the dataset is synthetic and the true counterfactual outcomes are available.

[5] Consider the one ● instance on the top left of Figure 1(b), as there are no other $t = 1$ instances close by.

[6] Indeed, there are two objectives to optimize: Eqs. (1) & (6).

3.2. Hyperparameter Selection

As counterfactual outcomes are inherently unobservable, it is not possible to use standard cross-validation to select hyperparameters. We therefore need to provide some estimation of the true effect as a surrogate for the e term in denominator of Eq. (8).

Shalit *et al.* (2017) used the observed outcome $y_{j(i)}$ of the nearest neighbor (nn[7]) $j(i)$ to instance i in the alternative treatment group $t_{j(i)} = \neg t_i = 1 - t_i$ as a surrogate for the counterfactual outcome for instance i. They then calculate $e_{nn} = (2t_i - 1)(y_i^{t_i} - y_{j(i)}^{\neg t_i})$. Substituting e_{nn} with the true e in denominator of Eq. (8) gives ENoRMSE$_{nn}$ which is a proxy for the true ENoRMSE. Finally, Shalit *et al.* (2017) rank models trained with different sets of hyperparameters based on ENoRMSE$_{nn}$ and select the best one.

We explore this approach, considering ENoRMSE$_{nn}$ calculated using the one nearest neighbor method based on either the original x space (*i.e.*, ENoRMSE$_{nn}$) or the common representation space $\Phi(x)$ (*i.e.*, ENoRMSE$_{nn_\Phi}$). However, our empirical results on synthetic datasets showed that neither ENoRMSE$_{nn}$ nor ENoRMSE$_{nn_\Phi}$ gave a decent ranking of the models trained with different sets of hyperparameters in terms of the true ENoRMSE. Therefore, in this paper, we decided to use a stronger method for finding a surrogate for the true individualized treatment effect e: the Bayesian Additive Regression Trees (BART) method (Chipman et al., 2010). Our empirical results show that ENoRMSE$_{\text{BART}}$ provides rankings that are much closer to those of the true ENoRMSE and therefore, the hyperparameters selected based on its rankings yield models with better performance.

3.3. Datasets

We use synthetic datasets provided at the 2018 Atlantic Causal Inference Data Challenge. The x matrix for each of these datasets are sampled from a covariates file of a real-world medical measurements taken from the Linked Birth and Infant Death Data (LBIDD)[8]. LBIDD's covariate file is comprised of 100,000 instances, over 177 features.

There are 24 synthetic datasets (number of instances $n_m \in \{1, 2.5, 5, 10, 25, 50\} \times 10^3$ for $m \in \{1, ..., 24\}$) for which we have access to both the factual as well as the counterfactual tables. Factual tables contain treatment bits as well as the observed outcomes. Counterfactual tables – which are only to be used for evaluation purposes – contain the true outcomes for treatments 0 and 1 (*i.e.*, y^0 and y^1 respectively). For each synthetic dataset, a data generating process determines t, y^0, and y^1 given the respective sampled x matrix. This data generating processes have not been revealed by the challenge organizers.

Table 1. The aggregated ENoRMSE (lower is better). The model hyperparameters for both CFR and CFR-ISW methods are selected based on the ranking provided by ENoRMSE$_{\text{BART}}$ (see Sec. 3.2). The entry in **bold** is the best for each row.

	Datasets	1-nn	BART	CFR	CFR-ISW
	All	75.32	20.03	8.92	**1.07**
# Instances	1 k	86.44	141.58	10.51	**1.72**
	2.5 k	47.45	23.35	15.27	**0.73**
	5 k	38.04	9.51	2.81	**0.93**
	10 k	40.25	2.96	1.22	**0.81**
	25 k	25.69	1.52	**0.89**	1.03
	50 k	94.45	16.13	11.12	**1.14**

Table 2. Comparison of the two hyperparameter selection methods; *i.e.*, ranking based on either ENoRMSE$_{\text{1-NN}}$ or ENoRMSE$_{\text{BART}}$ (see Sec. 3.2) in terms of the true aggregated ENoRMSE (lower is better). The entry in **bold** is the best for each ITE method.

		CFR		CFR-ISW	
	Datasets	1-nn	BART	1-nn	BART
	All	15.04	**8.92**	5.23	**1.07**
# Instances	1 k	**8.08**	10.51	2.21	**1.72**
	2.5 k	36.33	**15.27**	0.82	**0.73**
	5 k	5.79	**2.81**	1.05	**0.93**
	10 k	1.45	**1.22**	15.11	**0.81**
	25 k	1.01	**0.89**	1.28	**1.03**
	50 k	17.98	**11.12**	**1.11**	1.14

3.4. Results and Discussions

Table 1 summarizes the performance of four different methods in terms of ENoRMSE. The four methods are:

- 1-NN: The one nearest neighbor method for finding the counterfactual outcomes as described in the first paragraph of Sec. 3.2.
- BART: Bayesian Additive Regression Trees method (Chipman et al., 2010) for finding the ITE.
- CFR: CounterFactual Regression method proposed in (Shalit et al., 2017) whose best set of hyperparameters is determined based on ENoRMSE$_{\text{BART}}$.
- CFR-ISW: CounterFactual Regression with Importance Sampling Weights; the proposed method.

The first row of Table 1 reports the aggregated ENoRMSE of all the 24 datasets (*i.e.*, A). The next rows report the aggregated ENoRMSE for datasets with the same number

[7] In (Shalit et al., 2017), the nearest neighbor is identified based on a distance metric defined on the x space.

[8] Source: National Center for Health Statistics

of instances – *i.e.*, A_n for $n \in \{1, 2.5, 5, 10, 25, 50\} \times 10^3$. A_n and A respectively are calculated as follows:

$$\begin{aligned} A_n &= \sqrt{\frac{1}{|D_n|} \sum_{i \in D_n} \text{ENoRMSE}(D_n)} \\ A &= \sqrt{\frac{1}{\sum_{n \in S} n} \cdot \sum_{n \in S} n \cdot A_n^2} \end{aligned} \tag{9}$$

where D_n is set of all the datasets that have n instances and $S = \{ 1, 2.5, 5, 10, 25, 50 \} \times 10^3$ is set of the different dataset sizes.

Results in Table 1 show that incorporating the proposed importance sampling weights into the factual loss improves the ENoRMSE measure on almost all datasets by a large margin. Also, results in Table 1 show that, although neither 1-NN nor BART ITE methods perform well in terms of ENoRMSE, results in Table 2 show that hyperparameter selection for both CFR and CFR-ISW ITE methods based on $\text{ENoRMSE}_{\text{BART}}$ results achieves far better results in terms of ENoRMSE compared to that of $\text{ENoRMSE}_{\text{1-NN}}$.

References

Attouch, Hedy, Buttazzo, Giuseppe, and Michaille, Gârard. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*, volume 17. Siam, 2014.

Bengio, Yoshua, Courville, Aaron, and Vincent, Pascal. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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

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

Cuturi, Marco and Doucet, Arnaud. Fast computation of wasserstein barycenters. In Xing, Eric P. and Jebara, Tony (eds.), *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, 13(Mar):723–773, 2012.

Kingma, Diederik P and Ba, Jimmy Lei. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

Li, Lihong, Chu, Wei, Langford, John, and Wang, Xuanhui. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the 4th International Conference on Web Search and Data Mining*, Hong Kong, 2011.

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

Shalit, Uri, Johansson, Fredrik D., and Sontag, David. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3076–3085, 2017.

Strehl, Alex, Langford, John, Li, Lihong, and Kakade, Sham M. Learning from logged implicit exploration data. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, pp. 2217–2225. 2010.

Swaminathan, Adith and Joachims, Thorsten. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research (JMLR)*, 16, 2015.