A Novel Evaluation Methodology for Assessing Off-Policy Learning Methods in Contextual Bandits

Negar Hassanpour and Russell Greiner *

[1] University of Alberta, hassanpo@ualberta.ca
[2] University of Alberta, rgreiner@ualberta.ca

Abstract. We propose a novel evaluation methodology for assessing off-policy learning methods in contextual bandits. In particular, we provide a way to use data from any given Randomized Control Trial (RCT) to generate a range of observational studies with synthesized "outcome functions" that can match the user's specified degrees of sample selection bias, which can then be used to comprehensively assess a given learning method. This is especially important in evaluating methods developed for precision medicine, where deploying a bad policy can have devastating effects. As the outcome function specifies the real-valued quality of *any* treatment for any instance, we can accurately compute the quality of any proposed treatment policy. This paper uses this evaluation methodology to establish a common ground for comparing the robustness and performance of the available off-policy learning methods in the literature.

Keywords: Contextual Bandits, Off-policy Learning, Evaluation Method

1 Introduction

Precision medicine is a rapidly emerging field that tries to determine which specific treatment (*e.g.*, surgery, drug therapy, etc.) leads to the best outcome for each patient. This not only improves the patients' quality of life, but also typically reduces the health-care costs, especially in situations where the first course of treatment does not provide desirable outcomes, which means the patient has to receive a second (hopefully better) treatment. Many attempt to learn models for precision medicine from observed data; here, it is often necessary to answer counterfactual questions such as: *"Would this patient have lived longer, had she received an alternative treatment?"*. To answer such questions, it is required to know the underlying causal relationships between the patient's attributes and the outcomes associated with each potential treatment. Such causal relationships can only be learned from experimental studies that involve making interventions and collecting data on-line [1]. As such studies are often not available, we have to approximate the causal effects from off-line datasets [2].

In a Randomized Control Trial (RCT), the treatment assignment T to a patient is independent of the patient attributes X (see Fig. 1a). This makes it straightforward to infer the causal effects on a *population* level [2]. However,

* The authors were supported by NSERC and Amii. Data access was provided through a collaborative agreement between University of Alberta and National Institute of Mental Health, setup by James Benoit under the supervision of Dr. Serdar Dursun.

(a) Randomized Control Trial (b) Observational Study

Fig. 1: Types of data collection

it is not possible to run an RCT for every causal query; since, they are at best expensive, or in most cases infeasible. By contrast, in observational studies, the health-care provider suggests a treatment based on the patient's attributes, according to her training and/or the established clinical pathway. This is *sample selection bias*, as the assignment of treatment T depends on the patient's health history/attributes X (see Fig. 1b). However, while RCTs are rare, observational studies are abundant and are therefore our main source of data for developing algorithms that produce the personalized models needed for precision medicine.

The whole setup can be viewed in a *contextual bandit* setting [3] where, given a vector of attributes $x_i \in X$ describing patient i and her received treatment $t_i \in T$, we observe an outcome value $y(\, x_i,\, t_i\,) \in \mathbb{R}$. Treatment is selected according to an established clinical pathway represented by a conditional probability distribution $\pi_0(\, t \,|\, x\,)$. Following the literature, we refer to this as the *logging policy* (a.k.a. "behaviour policy" in reinforcement learning [4]). Also note that, for each patient, we only get to observe the outcome $y(\, x_i,\, t_i\,)$ associated with the received treatment t_i and not the outcomes associated with the alternative treatment(s) $(t \neq t_i)$. Since such *counterfactual* outcome(s) are inherently "unobservable" (and not just "unobserved"), there is no way to truly determine the best personalized treatment for each patient. The case of $y(\, x_i,\, t_i\,) \in \mathbb{R}^{\geq 0}$ might represent life expectancy after treatment, while $y(\, x_i,\, t_i\,) \in \{0, 1\}$ might indicate whether a patient would die or survive. In general, such contextual bandit datasets have the following information: $\mathcal{D} = \{\, [x_i,\, t_i,\, y(\, x_i,\, t_i\,),\, \pi_0(\, t_i \,|\, x_i\,)]\,\}_{i=1..n}$ The goal here is to find the best policy, $\pi^*(\, t \,|\, x\,)$, whose most likely selected treatment $t^* = \arg\max_t \pi^*(\, t \,|\, x\,)$ for patient x, indeed matches the best one. In other words, had we known all outcomes (observed *and* counterfactuals(s)), we want:

$$t^* \quad = \quad \arg\max_t y(\, x,\, t\,) \tag{1}$$

The problem of learning an optimal policy from observational studies (a.k.a. "off-policy learning" [4]) is not limited to medical applications. It also appears in A/B testing, ad-placement systems [5,6], recommender systems [7], intelligent tutoring systems [8], etc. It is challenging to evaluate a proposed policy given an observational dataset, due to (i) the inherent unobservability of the counterfactual outcomes, and (ii) the unknown degree of existing selection bias – *i.e.*, the strength of the link between X and T. To overcome these two issues, we have to somehow create synthetic datasets with full information on outcome (*i.e.*, both factual and counterfactual). However, to the best of our knowledge, all the existing approaches only evaluate their proposed method on an observational study generated under a fixed level of selection bias – see Sec. 3.1 for more de-

tails. This paper addresses this gap by proposing an evaluation methodology to establish a common ground for comparing the efficiency and effectiveness of the proposed methods for off-policy learning from observational datasets. Our contribution is an algorithm that, given an RCT dataset, can synthesize a spectrum of observational datasets that exhibit various degrees of selection bias.

2 Background

All the existing methods for off-line policy learning must address the sample selection bias to effectively learn an optimal policy from bandit feedback data. Many, including [9,10], use rejection sampling to extract a RCT-like subset of data. That is, they find and discard a subset of instances, such that the probability of assigning any treatment becomes uniform for all the remaining samples. This approach is extremely data inefficient, which might be acceptable for applications that have access to enormous amounts of data, such as ad-placement in search engines. However, this inefficiency means it cannot be used for small medical datasets. Besides the methods mentioned above, there are two other main approaches for handling the sample selection bias: (i) outcome prediction and (ii) utility maximization. Below, we briefly summarize such methods.

2.1 Outcome Prediction

Outcome prediction (OP) methods (a.k.a. "direct methods") try to learn a regression fit that predicts the outcome for patient i given feature vector x_i for any treatment $t \in T$ (*i.e.*, $\widehat{y}(\, x_i,\, t\,)$). Assuming the learned regression fit is accurate, we can rank different treatments in terms of their estimated outcome and choose the treatment that has the best outcome. This can be done following Eq. 1 by replacing $y(\, x,\, t\,)$ with the predicted (counter)factual outcomes $\widehat{y}(\, x,\, t\,)$.

Although this method seems straightforward, obtaining an accurate regression fit is often not easy as this requires addressing (i) selection bias and (ii) modeling bias. While there are ways to deal with selection bias, such as importance sampling (*cf.* [5]) modeling bias is more challenging to address, because it is not easy to identify the right model and/or features to produce an accurate estimator. Moreover, the outcome prediction method tries to answer the harder question of accurately predicting the counterfactual outcomes, while all we need to find an optimal policy is to *rank* the potential treatments. Therefore, other methods have been proposed that try to maximize a utility function instead.

2.2 Inverse Propensity Scoring

Inverse Propensity Scoring (IPS), an *importance sampling* technique that adjusts the weights of different instances, is one of the main ways to address the selection bias problem. Here, we use a variant of the early works of Horvitz and Thompson [11] and Rosenbaum and Rubin [12], since we are interested in evaluating a stochastic policy $\pi(\, t \,|\, x\,)$ as opposed to a deterministic one. This variant has been used in many closely-related applications, such as off-policy reinforcement learning [4], off-policy learning for contextual bandits [9,10], and counterfactual learning with causal graphs [5]. First, we formulate a utility function with importance sampling weights $\frac{\pi(\cdot)}{\pi_0(\cdot)}$ as:

$$\widehat{U}_{IPS}(\pi) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\, t_i \,|\, x_i\,)}{\pi_0(\, t_i \,|\, x_i\,)}\; y(\, x_i,\, t_i\,) \tag{2}$$

where n is the number of instances in the given contextual bandit dataset, $\pi_0(\, t_i \,|\, x_i \,)$ is the policy used to sample treatments in the training set (*i.e.*, the "logging policy"), $\pi(\, t_i \,|\, x_i \,)$ is the probability of selecting t_i given x_i by the proposed policy $\pi(\cdot)$, and $y(\, x_i,\, t_i \,)$ is the observed outcome. Note that IPS is an unbiased estimator of the true [unknown] utility $U(\pi)$, meaning:

$$\mathbb{E}_{\mathcal{D}}[\, \widehat{U}_{IPS}(\pi) \,] \quad = \quad U(\pi) \tag{3}$$

for any $\pi(\cdot)$ provided that $\pi_0(\cdot)$ has a non-zero value everywhere in its support [12]. Ultimately, the optimal policy π^* is obtained by:

$$\pi^* \quad = \quad \underset{\pi \in \Pi}{\arg\max}\, \widehat{U}(\pi) \quad = \quad \underset{\pi \in \Pi}{\arg\min}\, \widehat{R}(\pi) \tag{4}$$

where Π is the hypothesis space containing all possible policies and $\widehat{R}(\pi) = -\widehat{U}(\pi)$ is the empirical risk. Although unbiased, the IPS estimator has a high variance due to the $\pi_0(\, t_i \,|\, x_i \,)$ term in its denominator. That is, some importance sampling weights (*i.e.*, $\frac{\pi(\cdot)}{\pi_0(\cdot)}$) will be very large for any instance with small π_0 – *i.e.*, treatments that had a small chance of being selected, but were selected anyway. The most common approach is to clip the weights [5]:

$$w_i \quad = \quad \min \left\{ M, \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)} \right\} \tag{5}$$

where the M hyperparameter is an upper bound for the importance sampling weights. The risk calculated with the clipped weights is denoted as $\widehat{R}^M(\pi)$.

2.3 Doubly Robust Estimator

As discussed earlier, methods based on OP enjoy a low variance estimate at the cost of a high bias. On the other hand, although IPS is an unbiased estimator, it suffers from a high variance. This motivated Dudík *et al.* [13] to propose the Doubly Robust (DR) estimator, which leverages the strengths and mitigates the weaknesses of the above mentioned methods. Prior to [13], Robins *et al.* [14] had proposed a variant of this method for evaluating deterministic policies that have binary choice of treatment. DR's utility function is formulated as:

$$\widehat{U}_{DR}(\pi) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)} \Big(y(x_i, t_i) - \widehat{y}(x_i, t_i) \Big) \quad + \quad \mathbb{E}_{t \sim \pi|x_i} \big[\widehat{y}(x_i, t) \big] \right] \tag{6}$$

where $\widehat{y}(x, t)$ is the regression fit (obtained from an OP method) that predicts the (counter)factual outcome for any given patient x and treatment $t \in T$.

2.4 Self-Normalized Estimator

In order to alleviate the high variance problem, the doubly robust estimator employs a regression fit to predict counterfactual outcomes to be used as *additive* terms in the utility function (see Eq. 6). Alternatively, Swaminathan and Joachims [15] take a *multiplicative* approach by proposing the Self-Normalized (SN) estimator, which is a stochastic variant of the method proposed by Hirano *et al.* [16] for evaluating deterministic policies with a binary choice of treatment. Self-Normalized (SN) estimator uses the fact that $\mathbb{E}\left[\sum_{i=1}^{n} \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)} \right] = n$, which motivates replacing n with $\sum_{i=1}^{n} \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)}$ in Eq. 2, leading to:

$$\widehat{U}_{SN}(\pi) \;=\; \sum_{i=1}^{n} \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)}\, y(x_i, t_i) \Bigg/ \sum_{i=1}^{n} \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)} \tag{7}$$

The intuition is: since the main source of variance is the importance sampling weights appearing in the numerator of the utility function, having a similar factor in the denominator may cancel out some of the variability.

2.5 Counterfactual Risk Minimization

Swaminathan and Joachims [17] studied the variance of the IPS estimator with clipped weights under the Empirical Risk Minimization (ERM) principle (see Eq. 4) to prove the following generalization bound:

$$P\left[\exists\pi\in\Pi : R(\pi) \;>\; \widehat{R}^M(\pi) + \lambda\sqrt{\frac{\widehat{Var}\left(\,u(\cdot)\,\right)}{n}} + C\right] \quad\leq\quad \gamma \tag{8}$$

where $R(\pi)$ is the true risk, $u(\cdot) = y(x,t)\min\left\{M, \frac{\pi(t|x)}{\pi_0(t|x)}\right\}$, $\widehat{Var}(\cdot)$ is the estimated variance of $u(\cdot)$, $0 < \gamma < 1$ bounds the probability that this generalization bound fails, and λ and C are constants. This suggests adding the square-root term to the ERM objective function:

$$\pi^* \;=\; \arg\min_{\pi\in\Pi}\left\{\widehat{R}^M(\pi) \;+\; \lambda\sqrt{\frac{\widehat{Var}\left(\,u(\pi)\,\right)}{n}}\right\} \tag{9}$$

This addition to the objective function yields the Counterfactual Risk Minimization (CRM) principle [17], which is designed to penalize high empirical variance in weighted observed outcomes. The CRM principle can be used along with any utility function described in Secs. 2.2 to 2.4.

3 Evaluation Methodology

As described earlier, observational studies inherently contain partial information, as we only observe the outcome for the one treatment that was administered to each patient. When evaluating a new learned policy, however, due to the variance of the utility estimators, it is challenging to know if the new proposed policy is indeed better than the in-place policy. The best way to determine the effectiveness of a new policy is to actually deploy it on-site, record the factual outcomes for a reasonable period of time, and then analyze the results. However, it is neither ethical, nor allowed by the health-care community, to deploy a policy that has a chance of producing results that may reduce the patients' quality of life. Therefore, we need to synthesize bandit datasets in such a way that their counterfactual outcomes are also known, merely for the purpose of evaluation. In the following two sections, we first review the existing evaluation methodology [18] and point out its shortcomings, and then explain our proposed evaluation methodology, which allows for a more comprehensive assessment of a proposed algorithm in terms of robustness to various degrees of selection bias.

3.1 The Existing Approach

Beygelzimer *et al.* [18] proposed an approach that converts the training partition of a full-information binary multi-label supervised dataset[3] $\mathcal{D}^* = \{\,[\,(x_i,\, t_i^*)]\,\}_{i=1..n}$ with $t_i^* \in \{0,1\}^k$into a partial-information bandit dataset for training off-policy

[3] This is not one-hot encoding as there may be instances with multiple associated labels – *e.g.*, a news article concerning political initiatives on climate change.

learning methods.[4] They view each label t_i^* as the best possible treatment for patient i – that is, the outcome value-function $y(\, x_i,\, t_i\,)$ is defined such that $y(\, x_i,\, t_i^*\,) > y(\, x_i,\, \neg t_i\,)$, where $\neg t_i$ is any of the treatments other than t_i^*, as this ensures that Eq. 1 holds, and so the optimal policy $\pi^*(t|x_i)$ will prefer t_i^*. One can convert this supervised dataset into a bandit dataset by sampling a set of new labels $t_i \sim h_0(\, t \,|\, x_i\,)$ for each x_i, where $h_0(\cdot)$ is the underlying mechanism that creates this observational study. This allows a single subject to appear many $(r < 2^k)$ times in the dataset, each time associated with a different treatment.

In many applications, such as ad-placement, the underlying treatment assignment mechanism (*i.e.*, the deployed algorithm) is known [5]. To mimic the same situation, Swaminathan and Joachims [19] set $h_0(t|x)$ to be a logistic regression function, whose parameters are learned from a small portion (*e.g.*, 5%) of the supervised training set. This $h_0(\cdot)$ is then used to guide the sampling process of new labels t_i for each x_i, and log the propensities $\pi_0(t_i|x_i) = h_0(t_i|x_i)$. Finally, the outcome $y(x_i, t_i)$ is calculated as the Jaccard index between the supervised label t_i^* and the bandit label(s) t_i(s). This completes the procedure of generating a bandit dataset $\mathcal{D} = \{[x_i, t_i, y(x_i, t_i), \pi_0(t_i|x_i)]\}_{i=1..n}$.

There are several reasons why this evaluation framework is not appropriate for assessing off-policy learning methods for medical observational studies:

1. It is not clear how to map the concept of binary multi-label $\in \{0,1\}^k$ to treatment, *e.g.*, letting each bit in a label vector to be 1 (resp. 0) refer to taking (resp. not taking) a certain medication, a multi-label target would mean a combination of several drugs. However, due to drug interactions, such combinations might neutralize the effect of the treatment or worse, be detrimental to the patient's health. Therefore, unless there is a principled way to consider such interactions, a single class label seems more appropriate.
2. Using the Jaccard index to define outcomes implies assigning equal importance to various treatment options. However, this assumption does not hold in medicine since receiving the wrong treatment might be catastrophic for some cases while minor for others. Here, continuous measures such as *survival time* ($y \in \mathbb{R}^{\geq 0}$) that directly correspond to the consequences of the assigned treatment on the patient's health status seem more appropriate.
3. Unlike applications such as ad-placement, where the underlying mechanism of action selection is known, it may not be fully understood in medical observational studies (*e.g.*, clinical pathways). In reality, we never have access to [even a small] subset of data with ground truth labels. Hence, the propensities $\pi_0(\cdot)$ have to be calculated directly from the bandit dataset, as opposed to readily deriving them from $h_0(\cdot)$ (*i.e.*, estimated from 5% of supervised data).

3.2 The Proposed Approach

This section discusses our proposed evaluation methodology and its advantages over the existing approach. In addition to overcoming the shortcomings of the existing approach, we want to address the following requirements: (i) design a bandit dataset that is as realistic as possible in terms of similarity to an actual medical observational study (Sec. 3.2); and (ii) include a procedure to generate

[4] Note that the test set remains intact for evaluating the learned policy.

many different observational studies from a single RCT dataset to allow for comprehensive evaluation of learning methods for contextual bandits (Sec. 3.2).

Designing a Bandit Dataset: We require that the designed bandit dataset be as similar as possible to a real medical observational dataset. Therefore, instead of converting a supervised dataset to a bandit dataset, we directly work with a real-world RCT dataset as the source and from it synthesize various observational studies with different degrees of sample selection bias. [5] This makes sense because there is no selection bias in RCT datasets and therefore, one can often estimate the counterfactual outcomes reliably. In addition to the primary constraints (*e.g.*, $t \in \{0, 1\}$ and $y \in \mathbb{R}^{\geq 0}$), we want to preserve the statistical characteristics of the original (source) RCT dataset; characteristics such as:

(i) Average Treatment Effect: $ATE = \frac{1}{N_1} \sum y(x_i, 1) - \frac{1}{N_0} \sum y(x_i, 0)$, where N_1 (resp., N_0) is the number of subjects assigned to $t = 1$ (resp. $t = 0$); and

(ii) Coefficient of Determination: $R_t^2 = 1 - \frac{\sum [y(x_i,t_i) - f_t(x_i)]^2}{\sum [y(x_i,t_i) - \bar{y}]^2}$ for $t_i \in \{t_0, t_1, \dots\}$, where $\bar{y}$ is the mean of y. Coefficient of Determination is calculated on each treatment arm separately and measures the amount of variance in the response variable that can be explained by the observed explanatory variables.[6]

Given a RCT dataset with two treatment arms (*i.e.*, $t \in \{0, 1\}$), we first fit two Gaussian Process (GP) [20] models $f_t(\cdot)$ to calculate an initial estimate of the counterfactual outcomes; one for each treatment arm. More concretely, $f_t(x)$ provides a mean $\mu_t(x)$ along with a standard deviation $\sigma_t(x)$ that indicates the confidence of estimation at any point x in the function domain. We can now calculate the counterfactual outcome for each subject. For example, the counterfactual outcome for a subject whose received treatment was $t = 1$ (with observed outcome $y(x_i, 1)$) is defined as: $\widehat{y}(x_i, 0) = \mu_0(x_i) + k_0 \times \sigma_0(x_i)$, where k_0 is determined such that the average *personalized* treatment effect calculated on the N_1 subjects whose received treatment was $t = 1$ (*i.e.*, $\widehat{ATE}_1 = \frac{1}{N_1} \sum_i \text{ s.t. } {}_{t_i=1} \big(y(x_i, 1) - \widehat{y}(x_i, 0)\big)$) matches the ATE calculated on the original RCT dataset. Solving for $\widehat{ATE}_1 = ATE$ and $\widehat{ATE}_0 = ATE$ yields:

$$k_t = \Big(ATE - (2t-1)\frac{1}{N_{\neg t}} \sum \big(\mu_t(x_i) - y(x_i, \neg t_i)\big)\Big) \Big/ \frac{1}{N_t} \sum \sigma_t(x_i) \qquad (10)$$

This procedure ensures that any synthetic RCT data generated by random reassignment of treatments will have an $\widehat{ATE}$ close to the original ATE.

We also want our synthetic datasets to match the R_t^2 measure on every treatment arm t. To do so, we first calculate $\widehat{R}_t^2$ for each treatment arm t, on all subjects, using either observed, or counterfactual outcomes as derived in the previous step. Then, if the $\widehat{R}_t^2$ was higher than the original R_t^2 value, we modify the counterfactual outcomes by adding noise to them as follows:

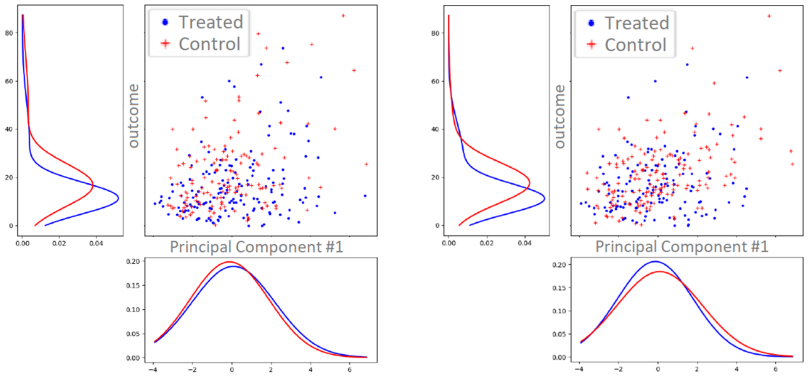
$$\hat{y}(x_i, t) \quad += \quad e_t \times \epsilon_i \quad , \quad t \neq t_i \qquad (11)$$

[5] This means the X values are realistic. By contrast, we do not know whether the X values from a supervised dataset look like realistic [medical] observational studies.

[6] A low R^2 measure suggests that there must exist [some] unobserved confounder(s) that [significantly] contribute to the outcome.

where e_t is the amplitude of the noise (tuned such that $\widehat{R}_t^2$ matches R_t^2) and $\epsilon \sim U(-0.5, 0.5)$. As $\mathbb{E}[\epsilon] = 0$, $\hat{y}(x_i, t)$'s expected increase is 0, and therefore we expect that $\widehat{ATE}$ would not change.

With this complete set of outcomes (observed as well as counterfactual), we can determine the best treatment for each patient (*i.e.*, ground truth labels), following Eq. 1. It is also possible to synthesize any observational study (including RCT) by simply designing an appropriate $h_0(\cdot)$ function. Figs. 2a and 2b respectively show the scatter plots of an original RCT dataset and a sample synthetic RCT generated from it,following the procedure described above. The next section explains how our proposed evaluation methodology can synthesize various observational studies, covering a wide range of selection bias.



(a) Original RCT (b) Synthetic RCT

Fig. 2: The proposed method generates a synthetic RCT dataset (right) that is very similar to a real RCT dataset (left)

Various Generating Policies $h_0(\cdot)$: Unlike [19,17,15], our proposed evaluation methodology decouples the generating policy $h_0(\cdot)$ from the supervised dataset. This means we can easily design different $h_0(\cdot)$ policies with various degrees of selection bias and/or conservatism, which in turn enables us to study the behaviours/robustness of different learning algorithms under such various circumstances. In order to create a bandit dataset, our basis function for sampling labels (*i.e.*, treatments) is a sigmoid function:

$$\sigma(z) \quad = \quad \sigma_{\alpha,\beta}(z) \quad = \quad \frac{1}{1 + e^{-\alpha(z-\beta)}} \tag{12}$$

where $|z|$ is the amount of improvement in outcome for a patient in case she receives the best treatment; z is positive for $t^* = 1$ class and negative for $t^* = 0$ class. For instances in $t^* = 1$ class, the treatments t are then drawn according to $\sigma(z)$ via rejection sampling and for $t^* = 0$ class according to $1 - \sigma(z)$.

Parameter α in Eq. 12 controls the degree of selection bias. With $\alpha = 0$, $\sigma(z)$ is a uniform distribution, which results in synthesizing a RCT dataset. Increased α creates a more biased dataset such that at the limit of $\alpha = \infty$, $\sigma(z)$ becomes a step function at β. At $\beta = 0$, a larger α increases the chance that a sampled treatment t is equal to its respective ground truth label t^*. We can simulate the

tendency towards prescribing a certain treatment more than the alternative(s) (*i.e.*, conservatism) by modifying β. As such, a $\beta > 0$ would assign treatment 0 to more patients and treatment 1 to fewer ones, and vice versa for $\beta < 0$.

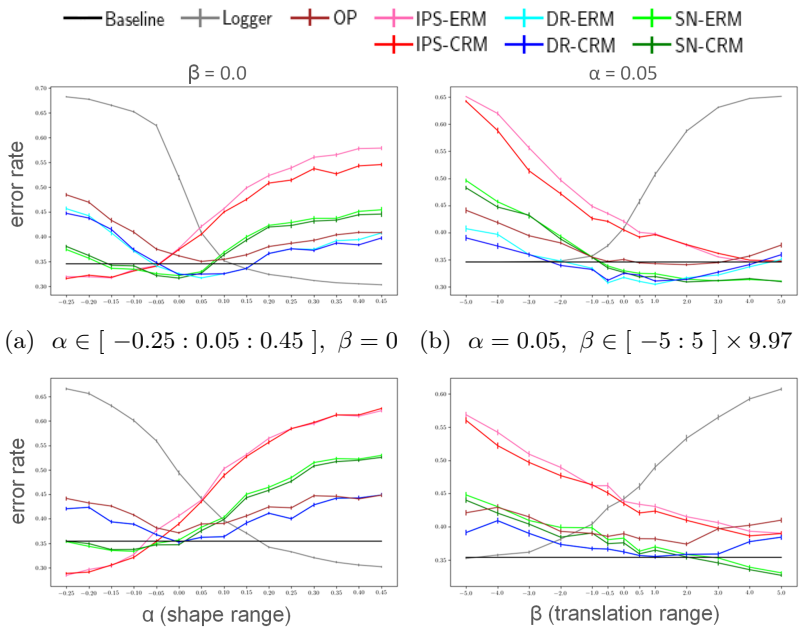
4 Experiments, Results, and Discussions

We experimented the proposed framework with the following two RCT datasets:
Acupuncture RCT [21,22] was designed to study the potential benefit of acupuncture (in addition to the standard care) for treatment of chronic headache disorders. It has 18 features, all measured prior to applying any treatment. There were two main outcomes: "severity score" and "headache frequency", each measured at two points in time: "immediately after the treatment is completed" and "at one year follow-up". Out of 401 participants, we use the 295 subjects with no missing values. Due to space limitation, we only report the performance results on one of the main outcomes (*i.e.*, severity score at one year follow-up) in the main text (the rest are closely similar).For this outcome, $ATE = -6.15$, $R_0^2 = 0.68$, and $R_1^2 = 0.33$, while our synthesized RCTs has $\widehat{ATE} = -6.17(0.76)$, $\hat{R}_0^2 = 0.60(0.05)$, and $\hat{R}_1^2 = 0.36(0.07)$.
Hypericum RCT [23] was designed to assess the acute efficacy of a standardized extract of the herb St. John's Wort in treatment of patients with major depression disorder. This study has three arms (placebo, hypericum, and an SSRI medication). The primary outcome measure is Hamilton Depression scale at the end of week 8. We compiled 278 features from assessment forms. In our experiments, we use the "hypericum" and "SSRI" as the binary treatment options (82 and 79 patients in each arm respectively). The original RCT has $ATE = -2.25$, $R_0^2 = 0.24$, and $R_1^2 = 0.00$. Our synthesized RCTs has $\widehat{ATE} = -2.65(0.97)$, $\hat{R}_0^2 = 0.15(0.10)$, and $\hat{R}_1^2 = 0.04(0.09)$.

The following methods are compared in terms of classification accuracy on the ground truth labels derived following the procedure described in Sec. 3.2. **Baseline:** predict the majority class. **Logger:** use the logging policy $\pi_0(\cdot)$ as the classifier. **Outcome Prediction:** find a regression fit with a simple linear least squares method with L2 regularization (OP), then use Eq. 1 to predict the best label (*i.e.*, treatment). **Inverse Propensity Scoring:** use the ERM objective function (IPS-ERM), as well as the CRM objective function (IPS-CRM) to learn a new policy $\pi(t|x)$ that acts as our classifier.[7] **Doubly Robust:** use the same regression function as OP for the regression component with either ERM (DR-ERM) or CRM (DR-CRM) objective functions to obtain $\pi(t|x)$. **Self-Normalized:** use either ERM or CRM objective functions (SN-ERM and SN-CRM respectively) to obtain $\pi(t|x)$.

Fig. 3 summarizes the performance results; showing the effect of changing α at $\beta = 0$ in Figs. 3a and 3c, and changing β at $\alpha = 0.05$ in Figs. 3b and 3d. Each point on the plots represents the mean classification error rate across 25 runs and

[7] Our implementation of IPS (and SN below) is obtained from Policy Optimizer for Exponential Models (POEM [19]). We extended POEM substantially to include a way to deal with the missing components (*i.e.*, OP and DR), as well as implementation of the proposed evaluation methodology.



(a) $\alpha \in [\ -0.25 : 0.05 : 0.45\ ],\ \beta = 0$ (b) $\alpha = 0.05,\ \beta \in [\ -5 : 5\ ] \times 9.97$

(c) $\alpha \in [\ -0.25 : 0.05 : 0.45\ ],\ \beta = 0$ (d) $\alpha = 0.05,\ \beta \in [\ -5 : 5\ ] \times 6.01$

Fig. 3: Mean and $\frac{1}{10}\times$standard deviation of the classification error rates on the "Acupuncture" (top) and "Hypericum" (bottom) datasets; best viewed in color

its respective error bar indicates a fraction (10%) of the standard deviation of the error rates (in order to maintain the plots' clarity). Also note that the Baseline accuracy for neither of the datasets is 1.0, meaning that not all patients benefit from receiving "acupuncture" or "SSRI"; indeed, for some, "no acupuncture" or "St. John's Wort" achieves a better outcome, *i.e.*, personalized medicine.

Effects of changing α. As α increases, it is trivial that the Logger's accuracy would improve since a higher α produces a bandit dataset with a higher tendency towards sampling the ground truth treatments more often. Moreover, OP's prediction of (counter)factual outcomes is not accurate as α moves away from 0, resulting in a bad performance for DR as well. IPS's performance is also correlated with α and it tends to do worse as α increases, as opposed to that of Logger's. SN tends to perform worse as $|\alpha|$ increases since this imposes π_0 to be small in parts of its domain which, due to weight clipping, results in $\mathbb{E}\left[\sum_{i=1}^{n} \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)}\right] \neq n$. This breaks the fundamental idea of SN (see Sec. 2.4). This suggests that SN is only useful for datasets that are close to RCT.

Effects of changing β. We know that the effect of changing β varies relative to the degree of class imbalance in a dataset. In ours, since the majority class is labeled as "1", a $\beta > 0$ would result in assigning fewer samples with label $t = 1$. This, in turn, would generate a bandit dataset that is more exploratory (which

SN seems to prefer); but, on the other hand, is far from the true underlying label distribution (which is not desirable for OP and DR).

ERM versus CRM principle. We found that the CRM principle often improves SN's performance (not statistically significant though). However, it does not do so for IPS and DR. In general, our results indicate that if a reliable OP method is available, then DR is the most effective and robust method for various α and β values. However, we should bear in mind that most OP (and as a result DR) methods require more processing power than IPS and SN. Therefore, we face a trade-off between a quick response versus a more accurate one.

5 Future Directions and Conclusions

The proposed evaluation methodology can be extended in three directions:

1. Increase the pool size of possible treatments from $|T| = 2$ (*i.e.*, $t \in \{0, 1\}$).This could cover combining several medications or other medical interventions.
2. Explore ways to extend this evaluation methodology for sequential observational studies. This is not trivial since the space of all possible decisions grows exponentially as we progress through the course of treatment.
3. We can create observational datasets that contain censored samples, *i.e.*, only a *lower bound* of the survival time of some patients is available. Such datasets can then be used to develop and evaluate survival analysis/prediction based methods that can handle sample selection bias.

In this paper, we proposed a novel evaluation methodology for assessing off-policy learning methods in contextual bandits. Unlike the existing methodology (*cf.*, [18]), our approach allows for a comprehensive assessment of the learning methods in terms of performance and robustness with respect to various degrees of sample selection bias. Moreover, it does not require the underlying mechanism for data generation to be known, and it better matches medical applications as it allows the outcomes to be more realistic ($y \in \mathbb{R}^{\geq 0}$). Using the proposed evaluation methodology, we assessed several prominent off-policy learning methods in contextual bandits – namely, outcome prediction, Inverse Propensity Scoring, Doubly Robust [13], Self-Normalized [15], and Counterfactual Risk Minimization principle [17] – on observational datasets synthesized using two RCT datasets. Our analyses identify the conditions under which a certain off-policy learning method performs best (*e.g.*, SN is preferable for a close-to-RCT dataset). Such analysis was not possible with [18]'s evaluation methodology as it has no means to generate such diverse observational datasets in terms of selection bias. Thus, we believe the proposed evaluation methodology should become a standard way for comprehensive assessment of new off-policy learning methods in contextual bandits, especially in costly applications such as precision medicine where deploying a bad policy can have devastating effects.

References

1. Pearl, J.: Causality. Cambridge University Press (2009)
2. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press (2015)

3. Wang, C.C., Kulkarni, S.R., Poor, H.V.: Bandit problems with side observations. IEEE Transactions on Automatic Control **50**(3) (2005)
4. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. Volume 1. MIT Press Cambridge (1998)
5. Bottou, L., Peters, J., Candela, J.Q., Charles, D.X., Chickering, M., Portugaly, E., Ray, D., Simard, P.Y., Snelson, E.: Counterfactual reasoning and learning systems: the example of computational advertising. JMLR **14**(1) (2013)
6. Li, L., Chen, S., Kleban, J., Gupta, A.: 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)
7. Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., Joachims, T.: Recommendations as treatments: Debiasing learning and evaluation. In: Proceedings of the 33rd International Conference on Machine Learning - Volume 48. (2016)
8. Liu, Y.E., Mandel, T., Brunskill, E., Popovic, Z.: Trading off scientific knowledge and user learning with multi-armed bandits. In: Educational Data Mining. (2014)
9. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th international conference on World wide web, ACM (2010)
10. Li, L., Chu, W., Langford, J., Wang, X.: 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)
11. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc **47**(260) (1952)
12. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika (1983)
13. Dudík, M., Langford, J., Li, L.: Doubly robust policy evaluation and learning. International Conference on Machine Learning (2011)
14. Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc **89**(427) (1994)
15. Swaminathan, A., Joachims, T.: The self-normalized estimator for counterfactual learning. In: Advances in Neural Information Processing Systems. (2015)
16. Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**(4) (2003)
17. Swaminathan, A., Joachims, T.: Counterfactual risk minimization: Learning from logged bandit feedback. In: International Conference on Machine Learning. (2015)
18. Beygelzimer, A., Langford, J.: The offset tree for learning with partial labels. In: Proceedings of the 15th ACM SIGKDD, ACM (2009)
19. Swaminathan, A., Joachims, T.: Batch learning from logged bandit feedback through counterfactual risk minimization. JMLR **16** (2015)
20. Rasmussen, C.E., Williams, C.K.: Gaussian processes for machine learning. Volume 1. MIT Press Cambridge (2006)
21. Vickers, A.J., Rees, R.W., Zollman, C.E., McCarney, R., Smith, C.M., Ellis, N., Fisher, P., Van Haselen, R.: Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. BMJ **328**(7442) (2004)
22. Vickers, A.J.: Whose data set is it anyway? sharing raw data from randomized trials. Trials **7**(1) (2006)
23. Hypericum Depression Trial Study Group and others: Effect of Hypericum perforatum (St. John's Wort) in major depressive disorder: a randomized controlled trial. JAMA **287**(14) (2002)