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

Negar Hassanpour
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8
hassanpo@ualberta.ca

Russell Greiner
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8
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 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 developing methods 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.

1 Introduction

Precision medicine is a rapidly emerging field that tries to determine which 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 such that 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 sufficient 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 [12]. As such studies are not accessible, we have to approximate the causal effects from off-line datasets [7].

In a Randomized Control Trial (RCT), treatment assignment T is independent of the patient attributes X (see Fig. 1a). This makes it straightforward to infer the causal effects on a *population* level [7]. However, it is not possible to run an RCT for every causal query that we might have; since, they are at best expensive, or in most cases, infeasible. By contrast, in observational studies, the health-care provider –according to her training and/or the established clinical pathway– suggests a treatment based on the patient's attributes. This is referred to as *sample selection bias*, that is, the assignment of treatment T depends on the patient's health history/attributes X (see Fig. 1b). However, unlike RCTs, 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 [23] where, given a vector of attributes $x_i \in X$ describing patient i and her received treatment $t_i \in T$ selected according to the established clinical pathway, we observe an outcome value $y(\, x_i,\, t_i\,) \in \Re$. The established clinical pathway can

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

(a) Randomized Control Trial (b) Observational Study

Figure 1: Types of data collection

be represented as 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 [17]). 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 which treatment would have been best for each patient. The case of $y(\, x_i,\, t_i \,) \in \Re^{\geq 0}$ might for example represent life expectancy, while $y(\, x_i,\, t_i \,) \in \{0, 1\}$ might indicate whether a patient would have died or survived as a result of the received treatment. 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 as well as [an estimate of] counterfactuals(s)), we would 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 [17]) is not limited to medical applications. It is also present in applications such as A/B testing, ad-placement systems [2, 8], recommender systems [16], and intelligent tutoring systems [11], to name a few. It is challenging to evaluate a proposed policy, given an observational dataset, due to (i) the inherent unobservability of the counterfactual outcomes, and also (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) Therefore, it is required that any new proposed method for learning a policy from off-line data must first be comprehensively evaluated in terms of its robustness against various degrees of selection bias, prior to deployment. 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 details.

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 an 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 a 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 touch on 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 best treatment that maximizes the outcome. This can be done following Eq. 1 by substituting $y(\, x,\, t \,)$ with the predicted (counter)factual outcomes $\widehat{y}(\, x,\, t \,)$ via the regression fit.

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, modeling bias is a lot more challenging to address, as

the outcome function is often complicated, meaning it is not easy to identify the right model and/or features to produce an accurate estimator. In summary, the outcome prediction method tries to answer the harder question of accurately predicting the counterfactual outcomes, while all we need for an optimal policy is the ability 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 [6, 15] as 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 [17], off-policy learning for contextual bandits [9, 10], and counterfactual learning with causal graphs [2]. First, we formulate a utility function 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, $\pi_0(\, t_i \,|\, x_i\,)$ is the policy used to sample treatments in the training data (*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 [15]. Below we sketch the proof that IPS is unbiased (note $y_i = y(x_i, t_i)$ is a short form notation for the observed outcome):

$$\mathbb{E}[\widehat{U}(\pi)] = \sum_{x_1,t_1} \cdots \sum_{x_n,t_n} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)} y_i\right] \times \pi_0(t_1|x_1) \,\cdots\, \pi_0(t_n|x_n) \times P(x_1) \,\cdots\, P(x_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \sum_{x_i,t_i} \pi_0(t_i|x_i)\, P(x_i) \left[\frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)}\, y_i\right] = \frac{1}{n}\sum_{i=1}^{n} \sum_{x_i,t_i} \pi(t_i|x_i)\, P(x_i)\, y_i \, \frac{1}{n}\sum_{i=1}^{n} U(\pi) = U(\pi)$$

However, the IPS estimator has a high variance due to the $\pi_0(\, t_i \,|\, x_i\,)$ terms in the denominator. That is, some importance sampling weights 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 use the clipped weights [2] as follows:

$$w_i \quad = \quad \min\left\{M, \frac{\pi(t_i|x_i)}{\pi_0(t_i|x_i)}\right\} \tag{4}$$

where M is a hyperparameter that represents an upper bound allowed for the importance sampling weights. Ultimately, with any utility function, the optimal policy π^* is obtained by:

$$\pi^* \quad = \quad \arg\max_{\pi\in\Pi} \widehat{U}(\pi) \quad = \quad \arg\min_{\pi\in\Pi} \widehat{R}(\pi) \tag{5}$$

where Π is the hypothesis space containing all possible policies and $\widehat{R}(\pi) = -\widehat{U}(\pi)$ is the empirical risk. When we use the clipped weights w (see Eq. 4), we denote the risk as $\widehat{R}^M(\pi)$.

2.3 Doubly Robust Estimator

As discussed earlier, methods based on outcome prediction enjoy a low variance estimate at the cost of a high bias. On the other hand, although IPS method is an unbiased estimator, it suffers from a high variance. This motivated [3] to propose the Doubly Robust (DR) estimator, which leverages the strengths and mitigate the weaknesses of the above mentioned methods. Prior to [3], [14] had proposed a variant of this method for evaluating deterministic policies that have binary choice of treatment. The DR 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 a method based on outcome prediction that predicts the (counter)factual outcome for a given patient represented by x with respect to any 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 *et al.* [20] take a *multiplicative* approach by proposing the Self-Normalized (SN) estimator, which is a stochastic variant of the method proposed by [5] 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 *et al.* [19] studied the variance of IPS estimator with clipped weights under the Empirical Risk Minimization (ERM) principle (Eq. 5) 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, $u(\cdot) = y(x,t) \min\left\{M, \frac{\pi(t|x)}{\pi_0(t|x)}\right\}$, $0 < \gamma < 1$ is the probability that this generalization bound fails, and λ and C are constants. This suggests addition of the square-root term to the ERM objective function as follows:

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

This addition to the objective function yields the Counterfactual Risk Minimization (CRM) principle [19]. The intuition here is 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 get to 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 clinical pathway. 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 [1] 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.* [1] proposed an approach that converts the training partition of a full-information multi-label supervised dataset $\mathcal{D}^* = \{\, [\,(x_i,\, t_i^*)]\,\}_{i=1..n}$ with $t_i^* \in \{0,1\}^k$ [1] into a partial-information bandit dataset for training off-policy learning methods.[2] Here, we view each label t_i^* as the best possible treatment for the patient i – that is, the outcome $y(\, x_i,\, t_i\,)$ 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^*. 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 $h_0(\cdot)$ represents the underlying mechanism that creates this observational study. This allows a single subject to appear many (r) times in the dataset, each time with a different treatment (of course $r < 2^k$). In many applications such as

[1] This is not one-hot encoding, as there are cases where more than one label might apply to a single instance – *e.g.*, a News article concerning political initiatives on climate change.

[2] Note that the test set remains intact for evaluating the learned policy in terms of classification accuracy.

ad-placement in a search engine, this underlying mechanism (*i.e.*, the deployed algorithm) is known and logged as the propensity scores $\pi_0(\cdot)$ during data collection [2]. To mimic the same situation (*i.e.*, knowing the underlying assignment mechanism), in [18], $h_0(t|x)$ is set to be the logistic regression, whose parameters were 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 developed 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 appropriately consider such interactions, a single class label seems more appropriate.
2. Using the Jaccard index to define outcomes y implies assigning equal importance to various treatment options. However, in medicine, this assumption does not hold since the consequences of being wrong about one treatment might be catastrophic, but for another, might be minor. Here, continuous measures such as *survival time* ($y \in \Re^{\geq 0}$) that directly correspond to the consequences of the assigned treatment on the patient's health status are more appropriate.
3. Unlike applications such as ad-placement where the underlying mechanism of action selection is known, this mechanism in medical observational studies (*i.e.*, the clinical pathway) cannot be fully understood. In reality, we never have access to [even a small] subset of data with ground truth labels. Therefore, the propensities $\pi_0(t_i|x_i)$ have to be calculated directly from the bandit dataset, as opposed to readily deriving them from $h_0(t|x)$ (*i.e.*, estimated from 5% of supervised dataset).

Next, we discuss our proposed evaluation methodology and its advantages over the existing approach.

3.2 The Proposed Approach

In addition to overcoming the shortcomings of the existing evaluation methodology, we want to address the following requirements: (i) designing a bandit dataset that is as realistic as possible. in terms of its similarity to an actual medical observational study (see Sec. 3.2.1); and (ii) includes a way for the user to generate many different observational studies from a single RCT dataset, in order to run comprehensive evaluation of learning methods for contextual bandits (see Sec. 3.2.2).

3.2.1 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 bias. This makes sense because there is no sample 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 $y \in \Re^{\geq 0}$), 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 $t_i \in \{t_0, t_1, \dots\}$ (calculated on each treatment arm separately), which measures the amount of variance in the response variable that can be explained by the observed explanatory variables.[3]

Given a RCT dataset with two treatment arms (*i.e.*, $t \in \{0, 1\}$), we first fit two Gaussian Process (GP) [13] 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, with which we can 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

[3] A low R^2 measure means there is unobserved confounder(s) that [significantly] contribute to the outcome.

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:

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

Note this applies to both $t = 0$ and $t = 1$. This procedure ensures that any synthetic RCT data generated by random re-assignment of treatments will have an $\widehat{ATE}$ close to the original ATE. Figs. 5a and 5b in Appendix show the original and a sample synthetic scatter plots respectively.

We also want our synthetic datasets to match the R_t^2 measure on every treatment arm t. To do so, we first calculate $\widehat{R}^2$ for each treatment arm, 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 as follows:

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

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.

Now, with this complete set of outcomes (both observed as well as conterfactual), 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 a $h_0(\cdot)$ function. The next section elaborates on how our proposed evaluation methodology is able to synthesize various observational studies, covering a wide range of sample selection bias.

3.2.2 Various Generating Policies $h_0(\cdot)$

Unlike [18, 19, 20], the 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 the positive class (*i.e.*, $t^* = 1$) and negative for the negative class (*i.e.*, $t^* = 0$). More concretely, to sample from a positive class, we sample according to $\sigma(z)$ and to sample from a negative class, we sample according to $1 - \sigma(z)$.

The α parameter controls the degree of selection bias. With $\alpha = 0$, $\sigma(z)$ would be a uniform distribution, which results in sampling a RCT bandit dataset. Larger α creates a more biased bandit dataset. That is, there is a higher chance for samples to be assigned with their respective ground truth treatment. At the limit of $\alpha = \infty$, we would have a step function at β. Clearly, if $\beta = 0$, then all the sampled treatments would be the ground truth and therefore, the bandit data would be equivalent to the supervised data. Fig. 3 in Appendix shows the effect of changing α on the respective generated bandit dataset. We can also simulate the tendency towards prescribing a certain treatment more that the alternative(s) (*i.e.*, conservatism) by modifying the β parameter. As such, a $\beta > 0$ would assign treatment 0 to more patients and treatment 1 to fewer ones, and vice versa for $\beta < 0$. Fig. 4 in Appendix shows the effect of changing β with constant $\alpha = 0.5$.

4 Experiments, Results, and Discussions

To demonstrate the benefits of the proposed evaluation methodology, we use two RCT datasets:
Acupuncture RCT [22, 21] studies 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 value in their measurements. 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 of the results can be found in the Appendix (see

Figs. 6, 7, and 8). [4] For this outcome, $ATE = -6.15$, $R_0^2 = 0.68$, and $R_1^2 = 0.33$. 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 [4] 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). 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. 3.2.1. **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.[5] **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. 2 summarizes the performance results; showing the effect of changing α at $\beta = 0$ in Figs. 2a and 2c, and changing β at $\alpha = 0.05$ in Figs. 2b and 2d. Each point on the plots represents the mean classification error rate across 25 runs and its respective error bar indicates a fraction 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", and for some, "no acupuncture" or "St. John's Wort" indeed 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 because a higher α produces a bandit dataset with a higher tendency towards sampling ground truth treatments more often (see Fig. 3b). Observe in Figs. 2a and 2c that OP's prediction of (counter)factual outcomes is not accurate as α moves away from 0, resulting in a bad performance for DR also. IPS's performance is also correlated with α and it tends to do worse as α increases, as opposed to Logger. 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 (*cf.*, Sec. 2.4). Therefore, 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 "1" labeled (*i.e.*, its z in Eq. 12 is positive), 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 [mostly not statistically significant though]. But, it does not do so for IPS and DR. In general, our results indicate that, if a reliable OP method is available , then DR, amongst other methods, is the most effective and robust 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.

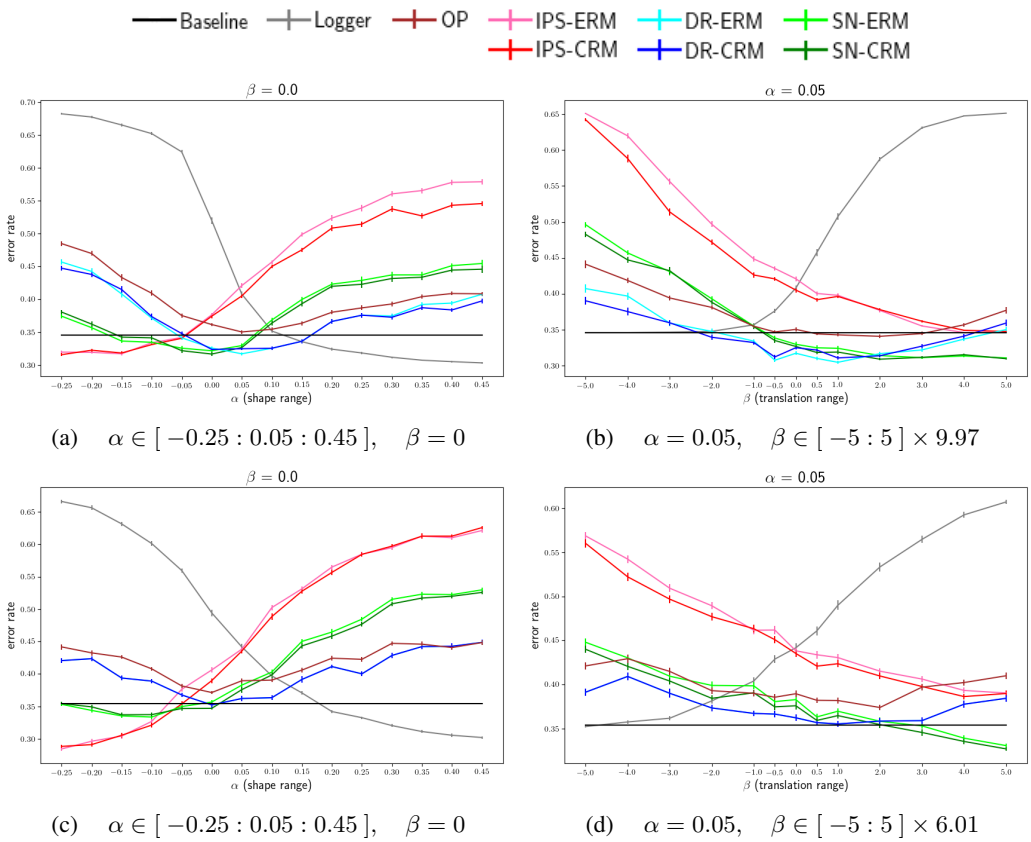
5 Future directions and Conclusions

The proposed evaluation methodology can be extended in the following three directions:

1. Increase the pool size of possible treatments from $|T| = 2$ (*i.e.*, $t \in \{0, 1\}$) to $|T| > 2$. This can cover cases such as combining several medications or other medical interventions.
2. Explore ways to extend this evaluation methodology for sequential observational studies (*i.e.*, following a course of treatment). This is not trivial since the space of all possible decisions grows exponentially as we progress through the course of treatment.

[4] These are similar, showing that the ranking of the methods in terms of performance is fairly robust.

[5] Our implementation of IPS (and SN below) is obtained from Policy Optimizer for Exponential Models (POEM [18]). 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) $\alpha \in [\,-0.25 : 0.05 : 0.45\,], \quad \beta = 0$ (b) $\alpha = 0.05, \quad \beta \in [\,-5 : 5\,] \times 9.97$

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

Figure 2: Mean and $\frac{1}{10}\times$standard deviation of the classification error rates obtained on the "Acupuncture" (top) and "Hypericum" (bottom) datasets; best viewed in color

3. We can create observational datasets that contain censored samples, *i.e.*, only a *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 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.*, [1]), 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 ($y \in \Re^{\geq 0}$). Using the proposed evaluation methodology, we assessed several prominent off-policy learning methods in contextual bandits – namely, outcome prediction, Inverse Propensity Scoring, Doubly Robust [3], Self-Normalized [20], and Counterfactual Risk Minimization principle [19] – 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 [1]'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.

Acknowledgement

Data access was provided through a collaborative agreement between University of Alberta and National Institute of Mental Health (NIMH), setup by James Benoit under the supervision of Dr. Serdar Dursun.

References

[1] A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM, 2009.

[2] L. Bottou, J. Peters, J. Q. Candela, D. X. Charles, M. Chickering, E. Portugaly, D. Ray, P. Y. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[3] M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *International Conference on Machine Learning*, 2011.

[4] H. D. T. S. Group et al. Effect of hypericum perforatum (st john's wort) in major depressive disorder: a randomized controlled trial. *Jama*, 287(14):1807–1814, 2002.

[5] K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[6] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

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

[8] L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934. ACM, 2015.

[9] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[10] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 297–306, New York, NY, USA, 2011.

[11] Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *Educational Data Mining (EDM)*, 2014.

[12] J. Pearl. *Causality*. Cambridge university press, 2009.

[13] C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[14] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[15] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55, 1983.

[16] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, pages 1670–1679, 2016.

[17] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[18] A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

[19] A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

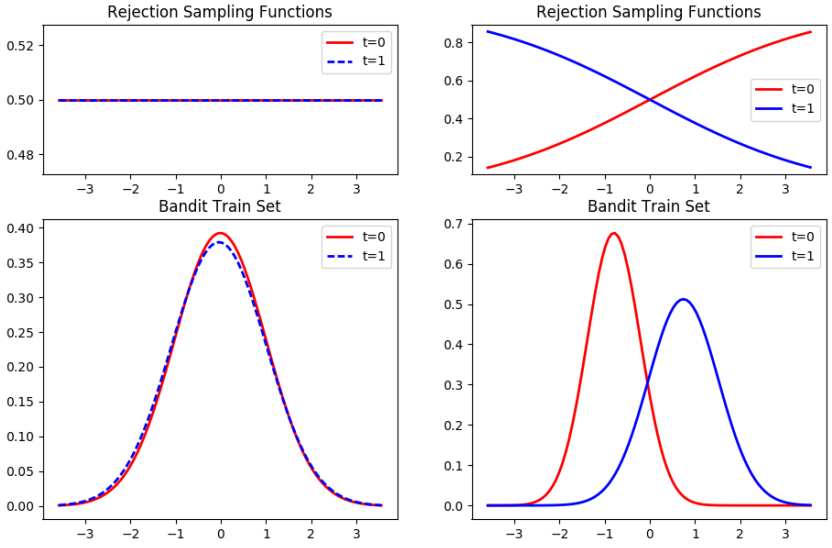
[20] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, pages 3231–3239, 2015.

[21] A. J. Vickers. Whose data set is it anyway? sharing raw data from randomized trials. *Trials*, 7(1):15, 2006.

[22] A. J. Vickers, R. W. Rees, C. E. Zollman, R. McCarney, C. M. Smith, N. Ellis, P. Fisher, and R. Van Haselen. Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *Bmj*, 328(7442):744, 2004.

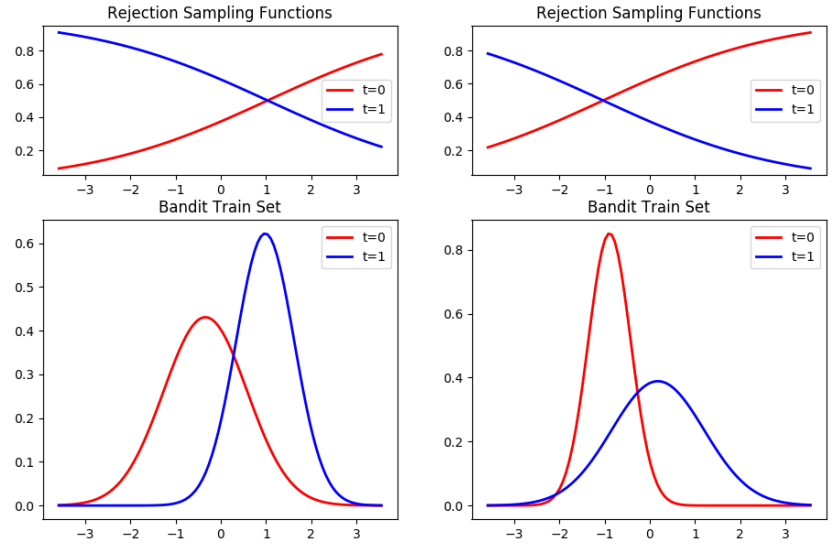
[23] C.-C. Wang, S. R. Kulkarni, and H. V. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

Appendix



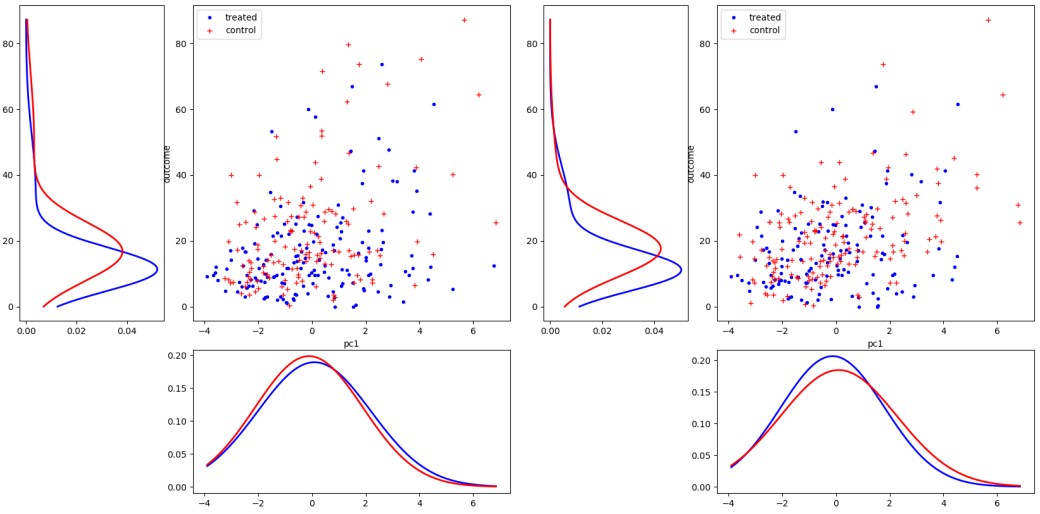
(a) $\alpha = 0$ (b) $\alpha = 0.5, \beta = 0$

Figure 3: Effect of changing α on selection bias. The bandit dataset (bottom) is sampled as guided by the respective sampling function (top) for each treatment. Fig. 3a shows that with $\alpha = 0$, since the sampling distributions for both treatments are uniform, $\forall\beta$ the resulting bandit dataset is RCT.



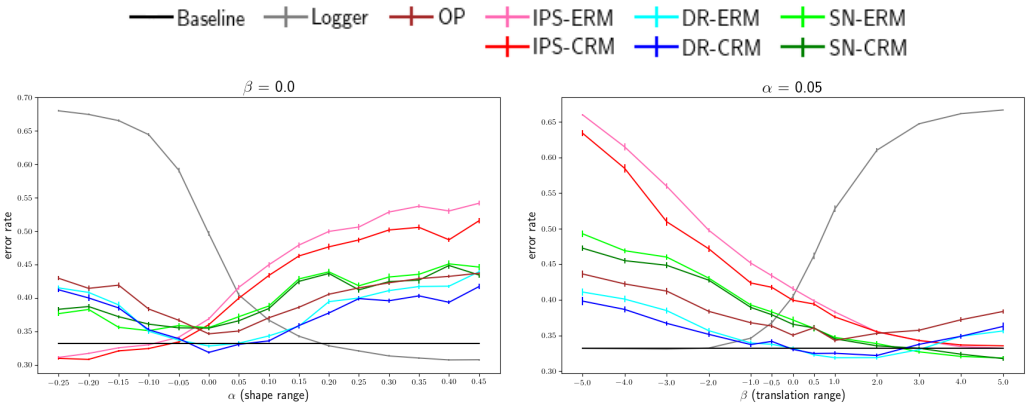
(a) $\alpha = 0.5, \beta = 1$ (b) $\alpha = 0.5, \beta = -1$

Figure 4: Effect of changing β on conservatism towards preferring one treatment to the other. The bandit dataset (bottom) is sampled as guided by the respective sampling function (top) for each treatment. For instance, Fig. 4a shows a clinical pathway that has a higher tendency towards selecting treatment 0 for patients.



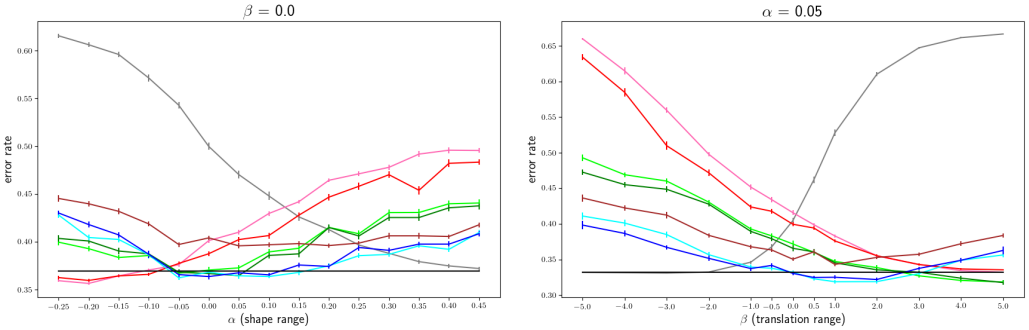
(a) Original RCT (b) Synthetic RCT

Figure 5



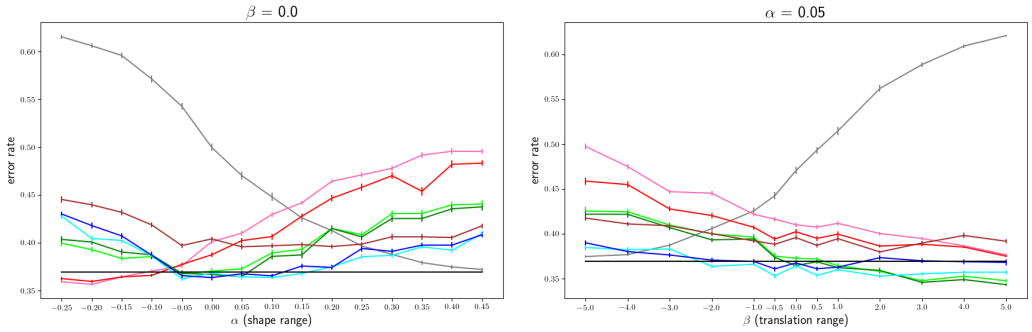
(a) $\alpha \in [\,-0.25 : 0.05 : 0.45\,], \quad \beta = 0$ (b) $\alpha = 0.05, \quad \beta \in [\,-5 : 5\,] \times 9.61$

Figure 6: Mean and $\frac{1}{10}\times$standard deviation of the classification error rates obtained on the "Acupuncture" dataset for outcome "severity score immediately after treatment is completed"



(a) $\alpha \in [\,-0.25 : 0.05 : 0.45\,], \quad \beta = 0$ (b) $\alpha = 0.05, \quad \beta \in [\,-5 : 5\,] \times 4.51$

Figure 7: Mean and $\frac{1}{10}\times$standard deviation of the classification error rates obtained on the "Acupuncture" dataset for outcome "headache frequency at one year follow-up"



(a) $\alpha \in [\, -0.25 : 0.05 : 0.45 \,], \quad \beta = 0$ (b) $\alpha = 0.05, \quad \beta \in [\, -5 : 5 \,] \times 4.04$

Figure 8: Mean and $\frac{1}{10}\times$standard deviation of the classification error rates obtained on the "Acupuncture" dataset for outcome "headache frequency immediately after the treatment is completed