

---

# Balanced Off-Policy Evaluation in General Action Spaces

---

**Arjun Sondhi**  
Flatiron Health

**David Arbour**  
Adobe Research

**Drew Dimmery**  
Facebook Core Data Science

## Abstract

Estimation of importance sampling weights for off-policy evaluation of contextual bandits often results in *imbalance*—a mismatch between the desired and the actual distribution of state-action pairs after weighting. In this work we present balanced off-policy evaluation (B-OPE), a generic method for estimating weights which minimize this imbalance. Estimation of these weights reduces to a binary classification problem regardless of action type. We show that minimizing the risk of the classifier implies minimization of imbalance to the desired counterfactual distribution. In turn, this is tied to the error of the off-policy estimate, allowing for easy tuning of hyperparameters. We provide experimental evidence that B-OPE improves weighting-based approaches for offline policy evaluation in both discrete and continuous action spaces.

Contextual bandits provide an elegant mechanism for choosing actions to optimize a given reward in the presence of uncertainty. This is done through implementing a *policy*, which defines actions performed given an observed state (Langford and Zhang, 2007). Applications of contextual bandits abound in medicine, where personalized treatments are designed based on known patient history (Tewari and Murphy, 2017), and internet marketing, where advertisements can be tailored to user interests (Li *et al.*, 2010). Unfortunately, in many applied settings, learning an optimal policy may be prohibitively expensive, and experimenting with an untested policy could result in unacceptably negative results, such as patient death. Given this difficulty, an important problem area is *counterfactual* or *off-policy* evaluation (OPE), where the ex-

pected reward of a proposed policy is estimated using logged historical data (states, actions, and rewards). This problem is even more important when attempting to safely deploy a policy for an application that previously used ad-hoc or difficult-to-enumerate rules as a de facto policy.

Our work focuses specifically on improving the estimation of weights commonly used in OPE. Because regression models often give biased results in off-policy settings, modern methods typically incorporate importance sampling to reweight the observed reward data through inverse propensity score (IPS) weighting (Dudík *et al.*, 2014; Thomas and Brunskill, 2016; Wang *et al.*, 2017). These methods have shown strong performance, but they typically assume that the importance weights (and therefore, propensity scores) are known exactly. In practice, this is typically not the case for two reasons: (i) Many policies are high-dimensional or continuous, making it much easier to sample from the policy given a state than to know the propensity score for a given state-action pair, and (ii) Logged data may not be from a probabilistic policy, but from ad-hoc rules set by engineers, perhaps with some randomization from an A/B test. In the absence of an oracle propensity score estimator, importance sampling weights applied to the observed state-action pairs will not necessarily result in the desired distribution. *Balance* measures the quality of the approximation to the counterfactual distribution; when weighted state-actions are imbalanced, off-policy evaluation can be arbitrarily biased.

In this paper, we develop a new estimator of balancing weights for off-policy evaluation in contextual bandit problems with arbitrary action spaces. Our proposed estimator, which we call *balanced off-policy evaluation* (B-OPE), is motivated by optimizing the *balance* between the two policies. B-OPE trains a probabilistic classifier on state-action data from both policies, which is then used to directly estimate density ratios. These can then be plugged into any existing OPE method that involves importance sampling (Kallus and Zhou, 2018; Wang *et al.*, 2017; Dudík *et al.*, 2014; Farajtabar *et al.*, 2018). B-OPE only requires logged data on

states, and the actions which would be taken by both the observed and target policies at those states. It does not require true knowledge or an estimator of either policy density. B-OPE is defined generally with respect to proper scoring rules which admit a wide variety of popular probabilistic classifiers, each of which correspond to a different underlying balance condition.

The main contributions provided in this work are:

1. We introduce B-OPE, a balancing weight estimator for off-policy evaluation in contextual bandits that applies to arbitrary action spaces without knowledge of either the observed or proposed policy density.
2. We show theoretically that B-OPE optimizes the balance condition, i.e. it minimizes a divergence between the observed and proposed state-action distributions.
3. We show that the loss of the classification problem bounds the bias and variance of the off-policy estimate, which allows practitioners to discriminate amongst losses and perform hyperparameter tuning by using cross-validation.

The rest of the paper is structured as follows. We provide an overview of OPE and the concept of balancing weights in Section 1. We then describe B-OPE in Section 2, and explain how classifier probabilities can be used to directly obtain importance sampling weights. In Section 3, we provide a theoretical analysis of our estimator and prove consistency for the counterfactual policy value. We summarize and discuss related work in Section 4. In Section 5, we evaluate B-OPE in numerical experiments, considering both discrete and continuous action spaces. The latter experiments provide an extension of the “classifier trick” of Dudík *et al.* (2011) to continuous action spaces.

## 1 Background and Problem Description

We assume the standard contextual bandit setup. Our data consists of  $n$  independent observations  $(s_i, a_i, r_i)$ . For each unit, a *state*  $s_i$  is observed, an *action*  $a_i$  is taken in accordance with some *policy*  $\pi$  (the distribution of  $a | s$ ), and a *reward*  $r_i$  is observed in response. With slight abuse of notation, we use the notation  $\pi$  to refer to both a policy and its density, and use  $\pi(s)$  to denote the action that would be taken under policy  $\pi$  for a state  $s$ .

The task addressed in this work is as follows: given a proposed policy  $\pi_1$  and logged data  $(s, a, r)$  collected following a policy  $\pi_0$  (the “factual” data), estimate the

expected reward of following  $\pi_1$  on the observed states (the “counterfactual” data). We denote the reward function as  $r(a, s)$ , and an estimated reward function as  $\hat{r}(a, s)$ .

We assume the following throughout:

**A 1.**  $\pi_0(a, s) = 0 \implies \pi_1(a, s) = 0 \forall a \in A, s \in S$

**A 2.**  $0 < \frac{\pi_1(a, s)}{\pi_0(a, s)} \leq \alpha_1 < \infty$

**A 3.**  $0 \leq r(a, s) \leq \alpha_2 < \infty, \forall s, a \in S \times A$

**A 4.** *The distribution of rewards across potential actions is independent of policy, conditional on state.*

### 1.1 Off-Policy Estimation

We now briefly review the different classes of off-policy estimation. Throughout this section we assume that  $(s_i, a_i, r_i)$  are data collected under the observed policy  $\pi_0$ , and  $a'_i$  is an action that would be taken under the proposed policy  $\pi_1$ .

The *direct method* approach to this problem fits a regression model  $\hat{r}(a, s)$  to approximate the reward function  $r(a, s)$  under the observed policy  $\pi_0$ . The counterfactual policy value,  $V_{\pi_1} := \mathbb{E}_{\pi_1}[r]$ , is estimated as an average over the predicted value for actions from the new policy:

$$\hat{V}^{DM} = \frac{1}{n} \sum_{i=1}^n \hat{r}(s_i, a'_i)$$

In order for the resulting estimate to be consistent, the reward model  $\hat{r}$  needs to generalize well to the reward distribution that would be observed under policy  $\pi_1$ . In practice, this method can be badly biased if the observed state-action data is not close to the counterfactual distribution (Dudík *et al.*, 2011).

*Importance sampling* reweights the observed rewards by an inverse propensity score (IPS), and a rejection sampling term, i.e.,

$$\hat{V}^{IPS} = \frac{1}{n} \sum_{i=1}^n r_i \frac{\mathbb{1}_{a_i}(a'_i)}{\hat{\pi}_0(a_i | s_i)}$$

Importance sampling is unbiased when the IPS is estimated well, but it often suffers from high variance. The *self-normalized importance sampling estimator* (also called the “weighted” or Hájek estimator) has been used to reduce variance, at the cost of small finite-sample bias, while maintaining consistency (Swaminathan and Joachims, 2015; Cochran, 1977),

$$\hat{V}^{SNIS} = \frac{\sum_{i=1}^n r_i \frac{\mathbb{1}_{a_i}(a'_i)}{\hat{\pi}_0(a_i | s_i)}}{\sum_{i=1}^n \frac{\mathbb{1}_{a_i}(a'_i)}{\hat{\pi}_0(a_i | s_i)}}$$

In continuous action spaces, Kallus and Zhou (2018) recently proposed an IPS-based method that replaces

the indicator function  $\mathbb{1}_{a_i}(\cdot)$  with a kernel smoothing term  $K$  having bandwidth  $h$ , i.e.,

$$V^{KIS} = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{a'_i - a_i}{h}\right) \frac{r_i}{\hat{\pi}_0(a_i|s_i)}$$

The corresponding self-normalized importance sampling estimator is defined analogously.

Finally, *doubly robust* estimators combine the direct method and importance sampling approaches. These methods weigh the residuals from the direct method regression with IPS. This reduces the variance of the resulting estimator and maintains consistency if either the direct method regression model or the importance sampling weights are correctly specified (Dudík *et al.*, 2014; Thomas and Brunskill, 2016). For discrete or continuous action spaces, the reward is estimated as

$$\hat{V}^{DR} = \frac{1}{n} \sum_{i=1}^n (r_i - \hat{r}(s_i, a_i)) \frac{J(a_i, a'_i)}{\hat{\pi}_0(a_i|s_i)} + \hat{r}(s_i, a'_i)$$

where  $J(a_i, a'_i)$  is a suitable rejection sampling term. In a similar manner, the SWITCH estimator of Wang *et al.* (2017) combines these two estimators by using IPS unless the weight is too large, in which case it uses the direct method.

## 1.2 Balance

Estimated IPS weights for OPE do not typically ensure balanced counterfactual state-action pairs. The true importance weights,  $\rho$ , imply the *balance condition*

$$\mathbb{E}_{\pi_0} [\phi(a) \otimes \psi(s) \rho(a, s)] = \mathbb{E}_{\pi_1} [\phi(a) \otimes \psi(s)], \quad (1)$$

where  $\phi$  and  $\psi$  are any real-valued (possibly vector) functions of  $a$  and  $s$ , respectively. In other words, the observed state-action distribution is weighted to exactly match the target state-action distribution, resulting in consistent off-policy estimates. However, proper propensity score model specification is difficult to test, and can be particularly difficult to obtain with continuous actions.

Balancing weights, like B-OPE, seek to address this difficulty by explicitly optimizing for balance (Liu *et al.*, 2018; Kallus, 2018). If only weights which attain balance are acceptable, then directly optimizing the balance criteria is desirable. Such estimators have been shown to provide strong results in their respective applications even under misspecification. However, there are three main limitations to these approaches: (i) they focus on discrete action spaces, (ii) they involve hyperparameters that must be set by heuristics, or (iii) they are computationally intractable. Our proposed estimator, B-OPE, is computationally simple

and applies to arbitrary action spaces. We further show that B-OPE minimizes the following measure of imbalance between the two policies:

**Definition 1.** Let  $\phi$  and  $\psi$  be real-valued functions of  $a$  and  $s$ , respectively. The  $L_1$  functional discrepancy between the observed policy  $\pi_0$  and the proposed policy  $\pi_1$ , with importance weights  $\hat{\rho}$  is given by  $\|\mathbb{E}_{\pi_0} [\phi(a) \otimes \psi(s) \hat{\rho}(a, s)] - \mathbb{E}_{\pi_1} [\phi(a) \otimes \psi(s)]\|_1$

Note that a consistent estimator  $\hat{\rho}$  will result in this distance going to zero by definition.

## 2 Balanced Importance Sampling

*Balanced off-policy evaluation* (B-OPE) is a simple method for estimating balancing importance sampling weights, and the central contribution of this work. B-OPE leverages classifier-based density ratio estimation (Sugiyama *et al.*, 2012; Menon and Ong, 2016) to learn importance sampling ratios. Specifically, off-policy evaluation using B-OPE consists of four steps:

1. Create a supervised learning problem using the concatenated proposed policy instances  $(s, a')$  and observed policy instances  $(s, a)$ , as covariates and giving a label ( $C$ ) of 0 to the observed policy and 1 to the proposed policy.
2. Learn a classifier to distinguish between the observed and proposed policy.
3. Take the importance sampling ratio as  $\hat{\rho}(a_i, s_i) = \frac{\hat{p}(C=1|a_i, s_i)}{\hat{p}(C=0|a_i, s_i)}$ .
4. Replace IPS weights with the B-OPE estimates in any OPE method which uses them.

Step three arrives at the importance sampler through an application of Bayes rule (Bickel *et al.*, 2009),

$$\frac{P(C=1|a, s)}{P(C=0|a, s)} = \frac{\pi(a, s|C=1)P(C=1)}{\pi(a, s|C=0)P(C=0)} = \frac{\pi_1(a, s)}{\pi_0(a, s)}$$

where  $\frac{P(C=1)}{P(C=0)} = 1$  by design. As an example of step four, replacing IPS in the self-normalized weighting estimator would provide:

$$\hat{V}^{B-OPE} = \frac{\sum_{i=1}^n J(a_i, a'_i) \hat{\rho}(a_i, s_i) r_i}{\sum_{i=1}^n J(a_i, a'_i) \hat{\rho}(a_i, s_i)} \quad (2)$$

where  $J$  defines a rejection sampler term between the observed action  $a_i$  and the proposed action  $a'_i$ . For discrete action spaces, this is simply  $\mathbb{1}_{a_i}(a'_i)$ . For continuous actions, we use the kernel term of Kallus and Zhou, that is  $J(a_i, a'_i) = \frac{1}{h} K\left(\frac{a'_i - a_i}{h}\right)$ . We analyze the theoretical properties of this estimator in section 3, although B-OPE may be combined with a model of rewards. Our experiments in Section 5 also include

the performance of the SWITCH estimator of Wang *et al.* (2017) when using B-OPE weights.

B-OPE works with a wide variety of common classification models, constrained by the following assumption:

**A 5.** *The classifier is trained using a strictly proper composite loss<sup>1</sup>,  $\ell$ , with a twice differentiable Bayes risk,  $f$ .*

This assumption includes a large number of widely used loss functions, such as logistic, exponential, and mean squared error, as well as models commonly used for distribution comparison, such as the kernel based density ratio estimators of Sugiyama *et al.* (2012), and maximum mean discrepancy (Kallus, 2018).

Given that B-OPE targets the policy density ratio, it minimizes imbalance, as given in Definition 1.

**Proposition 1.** *The  $L_1$  functional discrepancy between the observed policy  $\pi_0$  and the proposed policy  $\pi_1$  under B-OPE is bounded by*

$$\begin{aligned} & \|\mathbb{E}_{\pi_0} [\phi(a) \otimes \psi(s) \hat{\rho}(a, s)] - \mathbb{E}_{\pi_1} [\phi(a) \otimes \psi(s)]\|_1 \\ & \leq \|\mathbb{E}_{\pi_0} [\phi(a) \otimes \psi(s) B(\hat{\rho}, \rho)]\|_1 \end{aligned}$$

where  $B$  is a Bregman divergence.

The proof for this proposition is in the supplement. When  $\hat{\rho} = \rho$  this discrepancy is trivially equal to 0. The degree to which balance is attained is implied by the quality of the approximation of  $\hat{\rho}$  to  $\rho$ . The divergence used in this bound is determined by the classifier used in the estimation of the weights. For example, when the B-OPE classifier is trained to minimize the log-loss,  $B$  is the Jensen-Shannon divergence. Bregman divergences define a wide variety of divergences including KL divergence and maximum mean discrepancy (Huszár, 2013) that are often considered in the analysis of off-policy evaluation and covariate shift (Kallus, 2018; Bickel *et al.*, 2009; Gretton *et al.*, 2009). Proposition 3 of Menon and Ong (2016) shows that minimizing the scoring rule in the classifier is equivalent to minimization of the divergence. This demonstrates that minimization of the B-OPE classifier loss is tied to minimization of imbalance.

### 3 Theoretical Properties

B-OPE lets us formally tie classifier performance to the quality of our off-policy evaluation. In this section, we make this explicit by:

<sup>1</sup>A loss is strictly composite if the Bayes-optimal score is given by  $\bar{s}^* = \Psi \circ \hat{p}(C = 1|s, a)$  where  $\Psi$  is a link function  $\Psi: [0, 1] \rightarrow \mathbb{R}$ . Readers should see Buja *et al.* (2005) and Reid and Williamson (2010) for complete treatments of strictly proper composite losses.

1. Describing bounds for the bias and variance of the off-policy estimate in terms of the error for the density ratio.
2. Using results from prior work to show that minimizing the risk of the binary classifier used in B-OPE is equivalent to minimizing the error of the density ratio.
3. Combining these results to show consistency of B-OPE (described in Section 2) for off-policy evaluation.

An immediate consequence of these properties is that hyper-parameter tuning and model selection to minimize the risk of the binary classifier used in B-OPE *directly translates* to minimizing the error in off-policy evaluation via imbalance minimization. Importantly, this property is *not* shared by weights based on propensity score estimation (Kang *et al.*, 2007).

Let  $p(a, s) := \frac{\pi_1(a, s)}{\pi_1(a, s) + \pi_0(a, s)}$  denote the true class probability of observing data  $(a, s)$  under the target policy  $\pi_1$  instead of the observed policy  $\pi_0$ . This is estimated with a probabilistic classifier  $\hat{p}(a, s)$  on labelled state-action data. Additionally, let

$$\rho(a, s) := \frac{\pi_1(a, s)}{\pi_0(a, s)} = \frac{p(a, s)}{1 - p(a, s)}$$

denote the true policy density ratio, with estimator  $\hat{\rho}$ . We assume the classifier has regret that decays with increasing  $n$ .

**A 6.** *Let  $\hat{p}(a, s)$  be a probabilistic classifier such that  $\text{regret}(\hat{p}; \mathcal{D}, \ell) = O(n^{-\epsilon})$  for some constant  $\epsilon \in (0, 1)$ .*

Next, we require that our importance sampling weight estimator,  $\hat{\rho}$ , is independent of the observed rewards  $r$ . This can be easily achieved through sample splitting, training the classifier  $\hat{p}$  and applying B-OPE on independent datasets.

**A 7.** *Given observed state-action data, the density ratio estimator  $\hat{\rho}$  is independent of the observed rewards  $r(\pi_0(s), s)$ .*

Finally, we require certain regularity conditions and rates to use in our theoretical results.

**A 8.** *(i) The functions  $\pi_0(a, s)$ ,  $\pi_1(a, s)$ ,  $\rho(a, s)$ , and  $\hat{\rho}(a, s)$  have bounded second derivatives with respect to  $a$ , and*

*(ii) In the continuous action domain, the bandwidth parameter  $h = O(n^{-1/5})$ .*

We now show that the importance sampling estimator using B-OPE in equation (2) is asymptotically unbiased, and derive a bound for its variance. We accomplish this by characterizing the asymptotic quantities in terms of the Bregman divergence between the estimated and true density ratios. In the propositions

below, we use  $r_{\pi_1}$  to denote  $r(\pi_1(s), s)$  and  $\rho_{\pi_1}$  to denote  $\rho(\pi_1(s), s)$ .

**Proposition 2.** *In discrete action spaces, the expected bias of  $\hat{V}^{B-OPE}$  obeys the following bound:*

$$|\mathbb{E}_{\pi_1}[r] - \mathbb{E}_{\pi_0}[\mathbb{1}_a(\pi_1(s))\hat{\rho}(a, s)r(a, s)]| \leq \mathbb{E}_{\pi_0}[B(\rho, \hat{\rho})r_{\pi_1}]$$

*In continuous action spaces, the expected bias of  $\hat{V}^{B-OPE}$  obeys the following bound*

$$\left| \mathbb{E}_{\pi_1}[r] - \mathbb{E}_{\pi_0} \left[ \frac{1}{h} K \left( \frac{a - \pi_1(s)}{h} \right) \hat{\rho}(a, s) r(a, s) \right] \right| \leq \mathbb{E}_{\pi_0}[B(\rho, \hat{\rho})r_{\pi_1}] + o(h^2)$$

**Proposition 3.** *In discrete action spaces, the variance of  $\hat{V}^{B-OPE}$  obeys the following bound*

$$\text{Var}_{\pi_0} [\hat{V}^{B-OPE}] \leq \frac{1}{n} \left( \mathbb{E}_{\pi_1}[\rho_{\pi_1} r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[r_{\pi_1}^2(\delta^2 + 2\delta\rho_{\pi_1})] \right)$$

*In continuous action spaces, the variance of  $\hat{V}^{B-OPE}$  obeys the following bound*

$$\text{Var}_{\pi_0} [\hat{V}^{B-OPE}] \leq \frac{R(K)}{nh} \left( \mathbb{E}_{\pi_1}[\rho_{\pi_1} r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[r_{\pi_1}^2(\delta^2 + 2\delta\rho_{\pi_1})] \right) + o\left(\frac{1}{nh}\right)$$

where  $R(K) = \int K(u)^2 du$  and  $\delta = B(\rho, \hat{\rho})$ .

The proofs are deferred to the supplement. The implication of Proposition 2 is that the expected bias of B-OPE is bounded from above by the Bregman divergence between the true density ratio between the observed and proposed policy and the model estimate of the density ratio. The specific Bregman divergence depends on the choice of classifier  $\hat{p}$ . We can then appeal to Proposition 3 of Menon and Ong (2016) to provide an explicit link between the risk of the classifier and the Bregman divergence between  $\rho(a, s)$  and  $\hat{\rho}(a, s)$ .

We now prove the consistency of the B-OPE estimator given in (2):

**Proposition 4.** *Under Assumptions 1-8, and with bounded variance of the Bregman divergence, the B-OPE estimator is consistent for the counterfactual policy value, that is, as  $n \rightarrow \infty$ ,  $\hat{V}^{B-OPE} \rightarrow \mathbb{E}_{\pi_1}[r]$ .*

*Proof.* Based on Propositions 2 and 3, by selecting a Bregman divergence of the form in Proposition 3 of Menon and Ong (2016), we can bound the bias and variance in terms of the classifier  $\hat{p}$  regret. Recall from Assumption 6, this regret scales as  $O(n^{-\epsilon})$  for  $\epsilon \in$

$(0, 1)$ . Then, since rewards  $r$  are bounded, and  $h = O(n^{-1/5})$  we have that the bias tends to 0 as  $n \rightarrow \infty$ .

We can apply a similar argument for the variance, by decomposing

$$\mathbb{E}_{\pi_0}[B(\rho, \hat{\rho})^2] = \text{Var}_{\pi_0}[B(\rho, \hat{\rho})] + \mathbb{E}_{\pi_0}[B(\rho, \hat{\rho})]^2.$$

Then, given that  $\text{Var}_{\pi_0}[B(\rho, \hat{\rho})]$ ,  $\rho$ , and  $r$  are bounded, we have that the variance bound in Proposition 3 also goes to 0 as  $n \rightarrow \infty$ .  $\square$

The full proof and technical details for these results can be found in the supplement. It is worth briefly discussing the implications of Propositions 1-3 combined with Proposition 3 of Menon and Ong (2016) which ties classifier risk to the quality of the density ratio estimate. Proposition 1 implies that optimizing classifier performance directly translates into optimizing the quality of the importance sampler. In short, B-OPE allows for principled tradeoffs between imbalance (and the bias that comes with it) against variance in finite samples. The bias and variance of the estimated policy evaluation can be minimized by optimizing for classifier performance. Because the classifier risk is directly tied to the quality of the off-policy estimate, the problem is essentially reduced to model selection for supervised learning. As sample sizes increase, however, B-OPE maintains consistency and reduces imbalance. Even under misspecification, B-OPE seeks to minimize imbalance. In this case, bias will not vanish asymptotically, although imbalance will.

## 4 Related Work

Related work can roughly be divided into three categories: off-policy evaluation of contextual bandits, balancing estimators, and density ratio estimation. The most closely related work is prior work on off-policy evaluation for contextual bandits. Li *et al.* (2011) introduced the use of rejection sampling for offline evaluation of contextual bandit problems. Within the causal inference community there is a long literature on the use of doubly robust estimators (c.f. Bang and Robins, 2005; Kang *et al.*, 2007; Tan, 2010; Cao *et al.*, 2009). Dudík *et al.* (2011) later proposed the use of doubly robust estimation for off-policy evaluation of contextual bandits, combining the doubly robust estimator of causal effects with a rejection sampler. Since then, several works have sought to minimize the variance and improve robustness of the doubly robust estimator. Farajtabar *et al.* (2018) and Wang *et al.* (2017) present work to minimize the variance of the estimators by reducing the dependence on the inverse propensity score in high variance settings. Swaminathan and Joachims (2015) use a Hájek style estimator (Hájek

and others, 1964). Later work from Thomas (2015) and Swaminathan and Joachims (2015) build on this work to improve estimation.

A second related line of work is balancing estimators. Under correct specification of the conditional model Rosenbaum and Rubin (1983) show balance of the propensity score. More recently, a growing literature seeks to develop balancing estimators which are robust to mis-specification. Hainmueller (2012) and Zubizarreta (2015) provide optimization-based procedures which define weights that are balancing but are not necessarily valid propensity scores. Imai and Ratkovic (2014) later defined an estimator which strives to find a valid propensity score subject to balancing constraints. This was extended to general treatment regimes by Fong *et al.* (2018).

However, none of these directly address the problem of off-policy evaluation for contextual bandits. Kallus (2018) introduces a method for balanced policy evaluation that relies on a regularized estimator that seeks to minimize the maximum mean discrepancy (Gretton *et al.*, 2012). Calculation of weights is achieved through a quadratic program, which presents computational challenges as sample size grows large. It is interesting to note that the proposed evaluation procedure of Kallus (2018) fits within the assumptions of B-OPE where the scoring rule is maximum mean discrepancy (a strictly proper scoring rule) and the model is learned with variance regularization. The accompanying classifier can be defined via a modification of support vector machine classification (Bickel *et al.*, 2009). Dimakopoulou *et al.* (2018) propose balancing in the context of online learning linear contextual bandits by reweighting based on the propensity score. This differs from this work in the focus on online learning rather than policy evaluation and the use of a linear model-based propensity score which provides mean balance only in the case of correct specification. Wu and Wang (2018) propose a method which seeks to minimize an  $f$ -divergence to minimize regret, similar to the target in this work. However in the setting of Wu and Wang (2018) access to the true propensities are assumed, whereas B-OPE estimates the density ratio directly from observed and proposed state action pairs.

The final line of related work is density ratio estimation. The use of classification for density ratio estimation dates back to at least Qin (1998). Later work leverages classification for covariate shift adaptation (Bickel *et al.*, 2007, 2009) and two-sample testing (Friedman, 2004; Lopez-Paz and Oquab, 2017). However, this work is the first time classifier-based density estimation has been adapted for off-policy evaluation. There is also a growing literature on

density ratio estimation that does not rely on classification models. These methods largely rely on kernels to perform estimation (Huang *et al.*, 2007; Sugiyama *et al.*, 2012). KL importance estimation (KLIEP) (Sugiyama *et al.*, 2008), and least squares importance fitting (LSIF) (Kanamori *et al.*, 2009) are the most directly relevant, given their ability to optimize hyper-parameters via cross validation. Interestingly, Menon and Ong (2016) provides a loss for classification based density ratio estimation that reproduces KLIEP and LSIF. Thus, these estimators can be included inside of B-OPE by considering the corresponding loss functions for the classifier.

## 5 Experiments

In the experiments that follow, we evaluate direct method, importance sampling, and SWITCH estimators for off-policy evaluation and show that estimators which use B-OPE typically outperform those which use standard IPS. For the latter two methods, we compare inverse propensity score and B-OPE weights, and use the self-normalized versions of the estimators given in Section 1. In the SWITCH estimator, the threshold parameter  $\tau$  is selected using the tuning method suggested by Wang *et al.* (2017). We defer our results for doubly robust estimators to the supplement, but found the same trends in those evaluations. The direct method, propensity score, and B-OPE estimators are all trained as gradient boosted tree classifiers (or regressors for the continuous evaluations).

### 5.1 Discrete Action Spaces

We begin by evaluating the accuracy of B-OPE for the value of an unobserved policy in the discrete reward setting. We employ the method of Dudík *et al.* (2011) to turn a  $k$ -class classification problem into a  $k$ -armed contextual bandit problem. We split the data, training a classifier on one half of the data (**train**). This classifier defines our target policy, wherein the action taken is the label predicted. The reward is defined as an indicator of whether the predicted label is the true label. The optimal policy, then, is to take an action equal to the true label in the original data. Evaluating a policy corresponds to estimating the actor’s accuracy at identifying the true label.

In the second half of the dataset (**test**) we retain only a ‘partially labeled’ dataset where we uniformly sample actions (labels) and observe the resulting rewards. The **train** half of the data is also used to train direct method, propensity score, and B-OPE models. OPE estimators based on these models are then applied to the **test** data to estimate the relevant quantities for off-policy evaluation methods. We compare the ex-

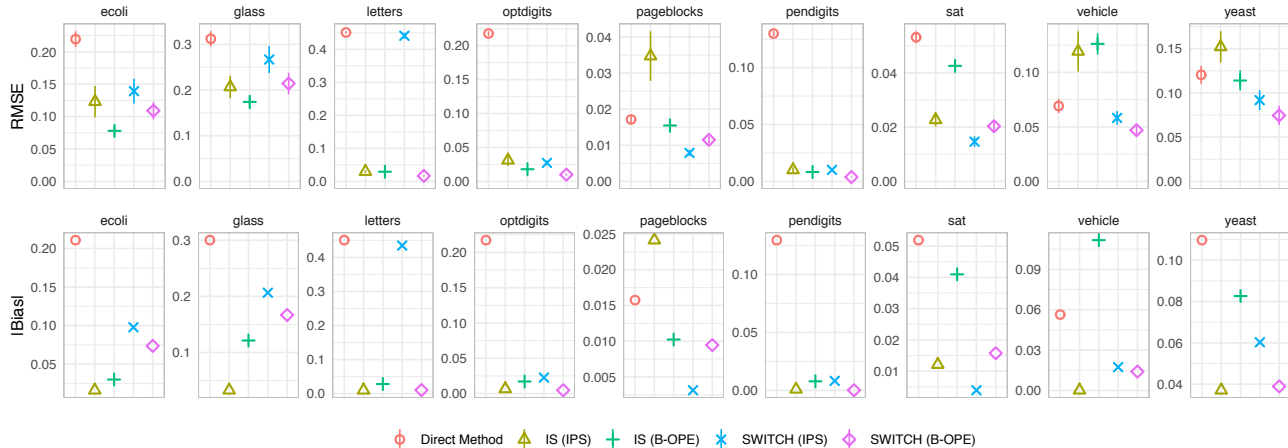


Figure 1: Root mean-squared error (RMSE) and bias for discrete action spaces using the classifier trick of Dudík *et al.* (2014). Points labeled “IS” use only weighting plus rejection sampling with either IPS or B-OPE. Points labeled “SWITCH” use the method of Wang *et al.* (2017) to adaptively combine importance sampling and direct estimators. RMSE is lower when using B-OPE compared to IPS; bias is typically slightly higher from B-OPE.

pected reward estimates to the true mean reward of the target policy applied to the `test` data. For each dataset, this process is repeated over 100 iterations, where we vary the actions under the observed uniform policy.

Our target policy model is trained as a multi-class random forest classifier. These models use the default hyperparameter values from scikit-learn with the exception of the number of trees. In order to provide increasingly complex policies to evaluate, we increase the number of trees as a function of sample size:  $\lceil 10 \times n^{\frac{1}{4}} \rceil$ . The propensity score, B-OPE and direct method (one-vs-rest) models are gradient boosted decision trees with default XGBoost hyperparameters with the exception of the number of boosting iterations. In order to adapt the estimator to the size of the dataset, the number of iterations is set as a function of sample size:  $\lceil 20 \times \sqrt{n} \rceil$ . We use the same datasets from the UCI repository (Dua and Graff, 2017) used by Dudík *et al.* (2011), and summarize their characteristics in the supplement. For some datasets, we removed classes with low frequencies to avoid issues when data splitting.

The results of the OPE estimators are summarized in Figure 1, where we plot the root mean squared error and bias averaged over 100 iterations. We see that the direct method estimator tends to be heavily biased for the true policy value, compared to B-OPE and IPS. The direct method generally performs quite poorly in terms of overall accuracy. The standard B-OPE estimator performs at least as well as and typically better than the IPS estimator. This also holds for the corresponding SWITCH estimators. While B-OPE often

has slightly higher bias than IPS, it strikes a better balance between bias and variance, leading to substantially improved accuracy in most cases.

## 5.2 Continuous Action Spaces

For the continuous action case, we provide a novel extension of the same transformation employed in the previous section for evaluation of discrete actions. We take a selection of datasets with continuous outcomes, and train a regression model on the `train` half of the data, which constitutes our target policy. The reward of a prediction (defined to be an action in our evaluation) is the negative of the Euclidean distance to the true label. Thus, it is optimal to choose actions equal to the true outcome as in the discrete evaluation. Evaluating the behavior policy is equivalent to estimating the mean squared error of the predictive model.

As before, we retain the `test` data for evaluation, while using the `train` data to train direct method, propensity score, and B-OPE models. For our observed policy, we sample actions from the empirical distribution of `train` labels, and compute the corresponding rewards. We then estimate the target policy value, repeating this over 300 iterations. We retain the same basic models from the previous section for this evaluation, swapping out classifiers for regressors as appropriate.

We use datasets from the UCI repository (Dua and Graff, 2017) and Kaggle, and summarize their characteristics in the supplement. The policy we evaluate is given by training a random forest regression to predict the continuous outcome. We also use gradient boosted regression trees for training direct method, propensity

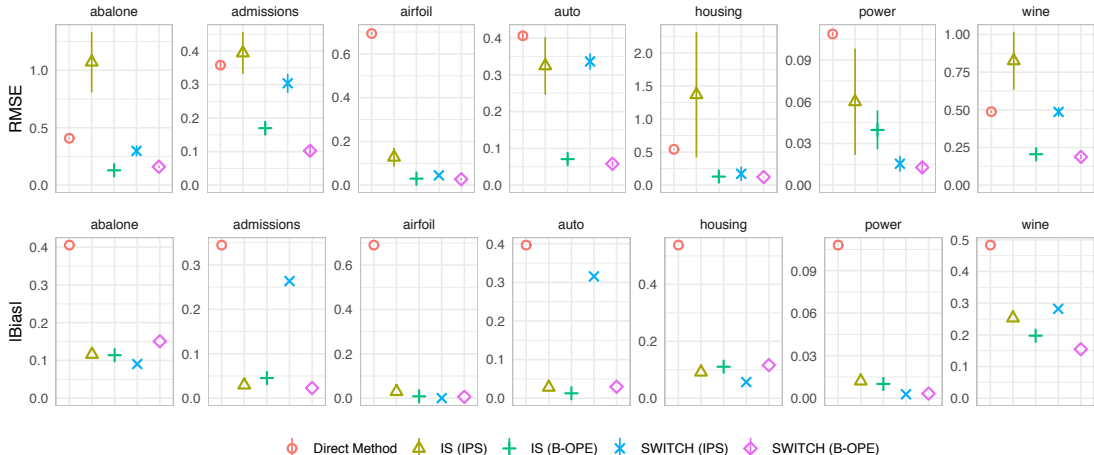


Figure 2: RMSE and bias for continuous action spaces using a modification of the classifier trick of Dudík *et al.* (2014) for regression detailed in Section 5.2. Points labeled “IS” use only weighting plus rejection sampling with either IPS or B-OPE. Points labeled “SWITCH” use the method of Wang *et al.* (2017) to adaptively combine importance sampling and direct estimators. RMSE and bias are typically lower using B-OPE than IPS.

score, and B-OPE models. Specifically, to obtain a continuous propensity score, we apply our observed policy to the `train` data, and train a model  $\hat{g}$  to predict actions from state features. Then, conditional on state  $s$ , the action is assumed to come from a normal distribution with mean  $\hat{g}(s)$  and variance  $MSE(\hat{g})$  as is standard practice (Hirano and Imbens, 2004). For each state-action pair  $(s, a)$  in the `test` data, the generalized propensity score is then the density of this distribution at  $a$ .

As in the previous section, we compare B-OPE to IPS (with the Kallus and Zhou (2018) kernel) and the direct method, including the corresponding SWITCH estimators. These results are displayed in Figure 2. We see that B-OPE outperforms the other methods uniformly across all datasets. In contrast to the binary setting, B-OPE does a better job of correcting for bias than IPS. This difference can be accounted for by considering that B-OPE estimates the densities implicitly via binary classification, while IPS must necessarily model the conditional density of action given state. The poor performance reflects the difficulty that many practitioners encounter when modeling continuous conditional distributions. In addition to reducing bias, B-OPE greatly reduces RMSE in most datasets. The B-OPE SWITCH estimator improves on the IPS version in both RMSE and bias in almost all cases. On the `power` dataset, B-OPE incurs half the RMSE of IPS when used within the SWITCH estimator. On `admissions` and `auto`, B-OPE incurs less than one-third of the RMSE than does standard IPS.

## 6 Conclusion

Off-policy evaluation is a critical component for the deployment of contextual bandit solutions in real-world settings. The efficacy of a majority of off-policy evaluation methods relies on the quality of their constituent importance weights. As we have shown, focusing on balance provides an effective means for deriving robust importance weight estimators. In particular, we introduced B-OPE, a simple, flexible, and powerful estimator of balancing weights for off-policy evaluation. B-OPE is easily implemented using off the shelf classifiers and trivially generalizes to arbitrary (e.g. continuous, multi-valued) action types. In Section 3 we tie the bias and variance of our estimator with the risk of the classification task, and show that B-OPE inherently minimizes imbalance. As a consequence of the theoretical results, hyperparameter tuning and model selection can be performed by minimizing classification error using well-known strategies from supervised learning. Experimental evidence indicates that B-OPE provides strong performance for discrete and continuous actions spaces. A natural direction for future work is considering the case of evaluation with sequential decision making and structured action spaces. B-OPE could also be extended to perform policy optimization in all of these settings. It would also be interesting to integrate methods for variance reduction, e.g. Thomas and Brunskill (2016) and Farajtabar *et al.* (2018), to further improve performance.



## References

- Mohan Acharya, Asfia Armaan, and Aneeta Anthony. A comparison of regression models for prediction of graduate admissions. *IEEE International Conference on Computational Intelligence in Data Science*, 2019.
- Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.
- Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.
- Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November, 3*, 2005.
- Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, 2009.
- William G Cochran. *Sampling techniques*. Wiley, 1977.
- Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
- Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. *arXiv preprint arXiv:1812.06227*, 2018.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104, 2011.
- Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.
- Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1446–1455, 2018.
- Christian Fong, Chad Hazlett, Kosuke Imai, et al. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.
- Jerome Friedman. On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA (US), 2004.
- Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.
- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.
- Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.
- Jaroslav Hájek et al. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.
- Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.
- Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.
- Ferenc Huszár. *Scoring rules, divergences and information in Bayesian machine learning*. PhD thesis, University of Cambridge, 2013.
- Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.
- Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251, 2018.
- Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8909–8920, 2018.
- Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.
- Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative

- strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.
- Heysem Kaya, Pinar Tüfekci, and Fikret S Gürgen. Local and global learning methods for predicting power of a combined gas & steam turbine. In *Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE*, pages 13–18, 2012.
- John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 817–824. Citeseer, 2007.
- Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 661–670, New York, NY, USA, 2010. ACM.
- Lihong Li, Wei Chu, John Langford, and Xuanhui 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. ACM, 2011.
- Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2649–2658, 2018.
- David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.
- Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313, 2016.
- Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.
- Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.
- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- J.P. Siebert. *Vehicle Recognition Using Rule Based Methods*. TIRM-87-018. Turing Institute, 1987.
- Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Büna, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.
- Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.
- Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *advances in neural information processing systems*, pages 3231–3239, 2015.
- Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.
- Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.
- Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.
- Philip S Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.
- Pinar Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power and Energy Systems*, 60:126 – 140, 2014.
- Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597, 2017.
- Hang Wu and May Wang. Variance regularized counterfactual risk minimization via variational divergence minimization. In *International Conference on Machine Learning*, pages 5349–5358, 2018.
- José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.