On the Design of Estimators for Bandit Off-Policy Evaluation

Nikos Vlassis [1] **Aurelien Bibaut** [2] **Maria Dimakopoulou** [1] **Tony Jebara** [1]

Abstract

Off-policy evaluation is the problem of estimating the value of a target policy using data collected under a different policy. We describe a framework for designing estimators for bandit off-policy evaluation. Given a base estimator and a parametrized class of control variates, we seek a control variate in that class that reduces the risk of the base estimator. We derive the population risk as a function of the class parameters and we discuss some approaches for optimizing this function. We present our main results in the context of multi-armed bandits, and we decribe a simple design for contextual bandits that gives rise to an estimator that is shown to perform well in multi-class cost-sensitive classification datasets.

1. Introduction

We address the problem of *off-policy evaluation* in a bandit setting. The problem involves estimating the value (expected reward) of a target policy from logged data collected under an alternative logging policy. This is an important problem, with roots in causality and estimation of treatment effects (Robins & Rotnitzky, 1995; Hahn, 1998; Hirano et al., 2003), and with an active research literature (Dudík et al., 2014; Li et al., 2015; Thomas & Brunskill, 2016; Swaminathan et al., 2017; Wang et al., 2017; Athey & Wager, 2017; Kallus & Zhou, 2018; Farajtabar et al., 2018; Joachims et al., 2018). The setting is of particular interest in recommendation systems (e.g., movie recommendations at Netflix), in which the logged actions are recommended items (e.g., movies) and the logged rewards capture a metric of interest (e.g., watch time). Off-policy evaluation allows testing a much larger number of candidate policies than would be possible by online A/B testing.

In the context of bandit problems, the following question is of great interest: Given a bandit model, what is a low-risk estimator of the counterfactual target value? (Throughout, by 'risk' we will assume mean squared error (MSE).) The question is primarily relevant in the non-asymptotic (finite sample) and/or agnostic (inaccurate model) setting, since asymptotically and under model realizability the Doubly Robust (DR) estimator is known to be risk-optimal (Hahn, 1998). Even for the simple case of a Bernoulli K-armed bandit the question of what constitutes a risk-optimal estimator is still open. Li et al. (2015) have derived a minimax lower bound of risk in the case of K-armed bandits with Gaussian rewards, and have demonstrated that REG, the classical regression (aka direct method) estimator for K-armed bandits, matches the minimax lower bound up to a constant. However, it is not known whether REG is improvable, and whether there exist instance-dependent estimators that dominate REG. Lattimore & Szepesvári (2018, Chapter 16) discuss the problem of computing instance-dependent lower bounds for bandit problems.

In this work we address the problem of *designing* a low-risk estimator for a given bandit instance. We develop our main results in the context of the standard Bernoulli K-armed bandit, but our ideas are generalizable to a wider class of problems, and we demonstrate an application to contextual bandits. We approach the problem in the following way: We assume we possess a base estimator b, for instance, REG. Given knowledge of b, we design a *parametric* class of control variates (CV), and seek a member t from that class such that $\text{MSE}(b-t) < \text{MSE}(b)$. The latter involves an optimization problem over population-dependent quantities. We derive the population risk difference $\text{MSE}(b-t) - \text{MSE}(b)$ and discuss a few ways to optimize this quantity.

Our main contributions can be summarized as follows. First, we show that the population risk is closed-form for a parametrized class of estimators involving polynomial functions of sufficient statistics (Section 3.3). Second, we demonstrate experimentally (Section 4 and Fig.1) that there *do* exist instance-dependent estimators that improve REG for K-armed bandits in the finite-sample regime, and those estimators can be learned from the data and built off totally different estimator classes (e.g., IPS). Finally, we show how a simple application of the idea can give rise to an estimator for contextual bandits that attains state-of-the-art performance in a benchmark dataset.

[1]Netflix, Los Gatos CA, USA [2]Department of Statistics, University of California Berkeley, Berkeley, USA. Correspondence to: Nikos Vlassis <nvlassis@netflix.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. K-armed Bandits

We assume K actions and binary rewards. The logged data consist of n i.i.d. pairs (a_i, r_i) that are assumed to be generated as follows: K Bernoulli rewards r_a are first drawn as $r_a \sim P_a$, for $a = 1, \dots, K$, where P_a are the (unknown) Bernoulli parameters, then actions are drawn as $a_i \sim \mu$ where μ is the logging policy, and finally we observe the rewards $r_i = r_{a_i}$. Using the n logged samples, we want to estimate the value of a target policy π:

$$v = \mathbb{E}_\pi[r] = \sum_a \pi_a P_a \,.$$

We will throughout assume absolute continuity of π with respect to μ, that is $\mu_a > 0$ when $\pi_a > 0$. A well-known unbiased estimator is the inverse propensity scoring (IPS) estimator (Horvitz & Thompson, 1952):

$$\hat{v}_{\text{IPS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_i}{\mu_i} r_i \,,$$

where we write $\pi_i = \pi_{a_i}$ and $\mu_i = \mu_{a_i}$. Note that the IPS estimator can be written

$$\hat{v}_{\text{IPS}} = \frac{1}{n} \sum_{i=1}^{n} \sum_a \mathbb{I}[A_i = a] \frac{\pi_a}{\mu_a} r_i = \sum_a n_a^+ \frac{\pi_a}{n \mu_a}$$

where we defined $n_a^+ = \sum_{i=1}^n \mathbb{I}[A_i = a]\, r_i$ the number of observed positives for arm a. The variance of the IPS estimator is (see derivation in the next section)

$$\text{Var}(\hat{v}_{\text{IPS}}) = \frac{1}{n} \Big(\sum_a P_a \frac{\pi_a^2}{\mu_a} - v^2 \Big),$$

which is known to be suboptimal for K-armed bandits, as it does not match, up to any constant, the minimax lower bound of risk for this problem (Li et al., 2015). An alternative estimator is REG, that is order optimal for this problem:

$$\hat{v}_{\text{REG}} = \sum_a n_a^+ \frac{\pi_a \, \mathbb{I}[n_a > 0]}{n_a}$$

where $n_a = \sum_{i=1}^n \mathbb{I}[A_i = a]$ is the number of times (out of n) that arm a was chosen.

We note that IPS and REG are both of the same form $\hat{v} = \sum_a n_a^+ f(n_a)$, for some choice of function f. A natural question is whether there exists some other member in the same family (for some function f) that exhibits improved risk. This question has prompted our approach, which aims at designing a new estimator by adding to a base estimator a control variate that is optimized to reduce overall risk.

3. The Family of Estimators

We will present the main ideas by assuming that the base estimator b and the control variate t are both of the following form:

$$b = \sum_a n_a^+ g(n_a), \qquad t = \sum_a n_a^+ f(n_a),$$

for some choice of functions g and f, and we will write $b = \sum_a n_a^+ g_a$ and $t = \sum_a n_a^+ f_a$ for brevity.

The question of interest is under what conditions on b, t, and the problem instance, a new estimator defined by $b - t$ could dominate the base estimator, that is, $\text{MSE}(b-t) < \text{MSE}(b)$. We collect all t-dependent terms of the risk difference function $\text{MSE}(b-t) - \text{MSE}(b)$ into a function R:

$$R = \big(\mathbb{E}[b] - \mathbb{E}[t] - v\big)^2 + \text{Var}[t] - 2\text{Cov}(b, t), \quad (1)$$

where $v \equiv v_\pi$ is the true target value of π. Next we show how to compute the various terms in (1).

3.1. Variance

The following holds for any function f:

$$\begin{aligned}
\text{Var}(t) &= \text{Var}(\mathbb{E}[t|(n_a)]) + \mathbb{E}(\text{Var}[t|(n_a)]) \\
&= \text{Var}\Big(\sum_a P_a n_a f_a\Big) + \sum_a P_a(1-P_a)\mathbb{E}[n_a f_a^2] \\
&= \sum_a P_a^2 \text{Var}(n_a f_a) + \sum_{a \neq a'} P_a P_{a'} \text{Cov}(n_a f_a, n_{a'} f_{a'}) \\
&\qquad + \sum_a P_a(1-P_a)\mathbb{E}[n_a f_a^2], \quad (2)
\end{aligned}$$

and the individual variance/covariance terms can be computed as usual via expectations:

$$\text{Cov}(n_a f_a, n_{a'} f_{a'}) = \mathbb{E}[n_a f_a n_{a'} f_{a'}] - \mathbb{E}[n_a f_a]\,\mathbb{E}[n_{a'} f_{a'}].$$

When t is in the IPS family, that is, f is constant w.r.t. (n_a), then $\mathbb{E}[t] = n \sum_a P_a f_a \mu_a$ and it is easy to verify that (2) simplifies to

$$\text{Var}\big(t | f = \text{const}\big) = n \sum_a P_a f_a^2 \mu_a - \frac{1}{n}\big(\mathbb{E}[t]\big)^2. \quad (3)$$

The variance of IPS is obtained by using $f_a = \frac{\pi_a}{n\mu_a}$ in (3):

$$\text{Var}\big(t | t = \text{IPS}\big) = \frac{1}{n}\Big(\sum_a P_a \frac{\pi_a^2}{\mu_a} - v^2 \Big).$$

where $\mathbb{E}[t] = v$ since IPS is unbiased.

3.2. Covariance

The following holds for any functions g, f:

$$\begin{aligned}
\text{Cov}(b, t) = \mathbb{E}[\text{Cov}(b, t|(n_a))] + \\
\text{Cov}(\mathbb{E}[b|(n_a)], \mathbb{E}[t|(n_a)]), \quad (4)
\end{aligned}$$

and using the fact that $\mathrm{Cov}(n_a^+ g_a, n_{a'}^+ f_{a'}|(n_a)) = 0$ for $a \neq a'$, the first term in (4) simplifies to

$$\begin{aligned}\mathbb{E}[\mathrm{Cov}(b,t|(n_a))] &= \mathbb{E}\Big[\sum_a \mathrm{Cov}(n_a^+ g_a, n_a^+ f_a|(n_a))\Big] \\ &= \sum_a \mathbb{E}[(n_a^+)^2 g_a f_a|(n_a)] - \sum_a P_a^2\, \mathbb{E}[n_a g_a]\, \mathbb{E}[n_a f_a] \\ &= \sum_a P_a^2 \mathbb{E}[n_a^2 g_a f_a] + \sum_a P_a(1-P_a)\mathbb{E}[n_a g_a f_a] \\ &\quad - \sum_a P_a^2\, \mathbb{E}[n_a g_a]\, \mathbb{E}[n_a f_a]. \quad (5)\end{aligned}$$

The second term $\mathrm{Cov}(\mathbb{E}[b|(n_a)], \mathbb{E}[t|(n_a)])$ in (4) can be similarly computed by writing

$$\mathbb{E}[b|(n_a)] = \sum_a P_a n_a g_a, \quad \mathbb{E}[t|(n_a)] = \sum_a P_a n_a f_a,$$

and expanding the covariance $\mathrm{Cov}(\mathbb{E}[b|(n_a)], \mathbb{E}[t|(n_a)])$ as a sum of covariances as in (2). (An interesting question, which we do not pursue here, is to identify conditions on g, f under which $\mathrm{Cov}(\mathbb{E}[b|(n_a)], \mathbb{E}[t|(n_a)]) \geq 0$.)

3.3. Computing the Expectations

From the above we see that the function R requires evaluating various expectations of functions of the counts n_a. For instance, when the base estimator b is REG and the control variate t is in the IPS family (i.e., f is constant), then the expectations in (5) read

$$\begin{aligned}\mathbb{E}[n_a^2 g_a f_a] &= \pi_a f_a \mathbb{E}\big[n_a \mathbb{I}[n_a > 0]\big] = \pi_a f_a \mathbb{E}[n_a], \\ \mathbb{E}[n_a g_a f_a] &= \pi_a f_a \mathbf{P}[n_a > 0] = \pi_a f_a (1 - (1-\mu_a)^n).\end{aligned}$$

A key result is that the required expectations can be evaluated in closed-form when the functions g and f are polynomial functions of n_a. This is possible by resorting to the following formula for the expectations of certain functions of multinomial counts (Mosiman, 1962):

$$\mathbb{E}[(n_a)_m] = (n)_m\, \mu_a^m,$$

where $(x)_m = x(x-1)\cdots(x-m+1)$ is the m'th order *falling factorial* of x. The set of falling factorials form a basis for the polynomial ring, so we can express any polynomial of n_a as a linear combination of such functions. The following identities are also known to hold:

$$\begin{aligned}x(x)_m &= (x)_{m+1} + m(x)_m \\ x^2(x)_m &= (x)_{m+2} + (2m+1)(x)_{m+1} + m^2(x)_m \\ (x)_m(x)_n &= \sum_{k=0}^{m} \binom{m}{k}\binom{n}{k} k!\,(x)_{m+n-k}\,.\end{aligned}$$

4. An Example

Let us consider a simple example in which b is IPS and t is of the simpler form

$$t = \sum_a n_a f_a,$$

where f_a are constant w.r.t. n_a. Note that in that case the estimator $b - t$ can be viewed as doing *reward shaping*, as it can be written as:

$$\hat{v} \equiv b - t = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\pi_i}{\mu_i} r_i - f_i\Big), \quad (6)$$

where $f_i \equiv f_{a_i}$. Let us assume that we are interested in obtaining an unbiased estimator $\hat{v}$. In that case, the requirement that t is zero-mean amounts to imposing a linear constraint on f_a:

$$\mathbb{E}[t] = 0 \;\Leftrightarrow\; \sum_a \mu_a f_a = 0.$$

Since the base estimator b is unbiased, we can verify (by recapitulating the analysis of the previous section or by performing a direct calculation) that the function R in (1) reads

$$R = \frac{1}{n}\Big(\sum_a \mu_a f_a^2 - 2\sum_a \pi_a P_a f_a\Big) + \text{const.}$$

Hence, we can characterize the set of minimum-variance unbiased estimators within the above class of control variates t by the optima of the following linearly-constrained quadratic program (QP):

$$\min_{f_a} \quad \sum_a \mu_a f_a^2 - 2\sum_a \pi_a P_a f_a \quad (7)$$

$$\text{s.t.} \quad \sum_a \mu_a f_a = 0\,. \quad (8)$$

Alternatively, when the number K of actions is very large, as in the case of linear bandits (Ghosh et al., 2017) or slate bandits (Swaminathan et al., 2017), we can consider a parametrized family of offsets of the form $f_a = w^\intercal \phi_a$, where $w \in \mathbb{R}^d$ is a d-dimensional vector and ϕ_a are d-dimensional feature vectors describing each action. In that case, the set of minimum-variance unbiased estimators are characterized by a QP in the vector w:

$$\min_{w} \quad w^\intercal \Phi\, w - 2 w^\intercal \sum_a \pi_a P_a \phi_a \quad (9)$$

$$\text{s.t.} \quad w^\intercal \sum_a \mu_a \phi_a = 0\,, \quad (10)$$

where $\Phi = \sum_a \mu_a \phi_a \phi_a^\intercal$ is a weighted covariance matrix.

4.1. Optimization

Solving the QP in (7)-(8), or the analogous one in (9)-(10), presents a difficulty, as the second term of the objective function contains unknown quantities (the Bernoulli parameters P_a, for $a = 1, \ldots, K$). Next we discuss a few approaches to dealing with this.

Eliminating the unknown quantities. One approach is to approximate the original QP with a new QP in which the unknown quantities are imputed by some constant values (for instance, when a good prior is known for the Bernoulli parameters P_a). As an example, in the case of K-armed bandits with Bernoulli rewards we have $P_a \in [0, 1]$, and hence we can consider imputing $P_a = 1$ which approximates the second term in (7) with $2\pi_a f_a$. That would give $f_a^* = 1 - \frac{\pi_a}{\mu_a}$, in which case the corresponding estimator in (6) boils down to IPS, showing that the above approximation is too loose. For a richer parametric family, a question of interest is whether the minimax solution over P_a could give an estimator that has lower risk than the base estimator. We are not aware of any work that directly addresses this question. We raise as an open problem the design of instance-dependent control variates that guarantee risk improvement for any value of the true population parameters.

Approximating the unknown quantities by sample surrogates. A more practical approach to dealing with the unknown quantities in (7) is to approximate the population expectations with sample averages (for instance, approximate the P_a with empirical success rates) and then solve the approximate QP. A similar approach was used by Farajtabar et al. (2018) in a contextual bandits setting involving a parametrized class of doubly-robust estimators. In our example above, an idea would be to rewrite the second sum in (7) as $\sum_a \mu_a P_a \frac{\pi_a}{\mu_a} f_a$, and note that this is a population average over the logging policy μ and the true reward models P_a. Hence, one can approximate this expectation using the logged data, giving rise to the following QP:

$$\begin{aligned} \min_{f_a} \quad & \sum_a \mu_a f_a^2 - \frac{2}{n} \sum_{i=1}^n \frac{\pi_i}{\mu_i} r_i f_i \\ \text{s.t.} \quad & \sum_a \mu_a f_a = 0 \,. \end{aligned} \tag{11}$$

Rewriting the second sum in (11) using indicator variables

$$\sum_{i=1}^n \frac{\pi_i}{\mu_i} r_i f_i = \sum_{i=1}^n \sum_a \mathbb{I}[A_i = a] \frac{\pi_a}{\mu_a} r_i f_a = \sum_a \frac{\pi_a}{\mu_a} f_a n_a^+$$

and solving the resulting QP, gives

$$f_a^* = -\frac{\pi_a n_a^+}{n \mu_a^2} + \sum_{a'} \frac{\pi_{a'} n_{a'}^+}{n \mu_{a'}} \,. \tag{12}$$

This corresponds to using a control variate

$$t = \sum_a \frac{\pi_a n_a^+}{n \mu_a} \left(1 - \frac{n_a}{n \mu_a}\right) \tag{13}$$

$$= \frac{1}{n} \sum_{i=1}^n \frac{\pi_i}{\mu_i} \left(2 r_i - \frac{n_{a_i}^+}{n \mu_i}\right) . \tag{14}$$

Note from (13) that this t belongs to the family of CVs from Section 3. Also note that the terms in the parenthesis of (14) converge in the limit of large n to $2P_{a_i} - P_{a_i}$, and hence the estimator converges to the true v. It is easy to verify that the expectation of t is

$$\mathbb{E}[t] = \frac{1}{n} \sum_a \frac{\pi_a}{\mu_a} P_a (1 - \mu_a) \,,$$

and hence the derived estimator $b - t$ is asymptotically unbiased.

However, the above brings up a subtle issue (that is also present in the approach of Farajtabar et al. (2018)). When optimizing an empirical surrogate of population risk, whatever comes out of the optimization is a function of the sample, and hence it is random. This randomness would have induced correlations that are essentially ignored in the optimization problem. For example, note that the optimal f_a^* in (12) are random since they depend on n_a^+, but the parametric CV class in (6) had assumed non-random f_a. One should therefore view the plug-in approach above primarily as a 'guide' that allows identifying a candidate class of control variates, and which should be followed by an analysis of the risk of the corresponding estimators from that class (using for instance the tools presented in Section 3). Clearly, every iteration of this procedure will require dealing with the unknown population parameters. Eventually, the latter will need to be estimated empirically (by imputation, sample surrogates, or sample-splitting (Dudík et al., 2014; Athey & Wager, 2017)), but the hope is that a 'good-enough' parametric class can already be identified by optimization at the population level.

Re-analyzing the obtained CV. We have followed the procedure described in Section 3 to analyze the properties of a CV of the form λt, where $\lambda \in \mathbb{R}$ and t is given by (13). The resulting function R in (1) is quadratic in λ (details omitted) and optimization is straightforward. The next figures show results from a simulation involving 5 and 10 actions, where we see that the resulting estimator improves the base estimator IPS and often it also performs better than REG. In the figures we show the 'oracle' λ_0, which was computed by knowledge of the true P_a, as well as the plug-in $\hat{\lambda}$ which was computed by the plug-in approach outlined above. The graphs show that the largest improvement is obtained in the 'adversarial' setting in which the logging and target policies differ, as predicted by theory (the function R involves terms that contain the ratios $\frac{\pi_a}{\mu_a}$).

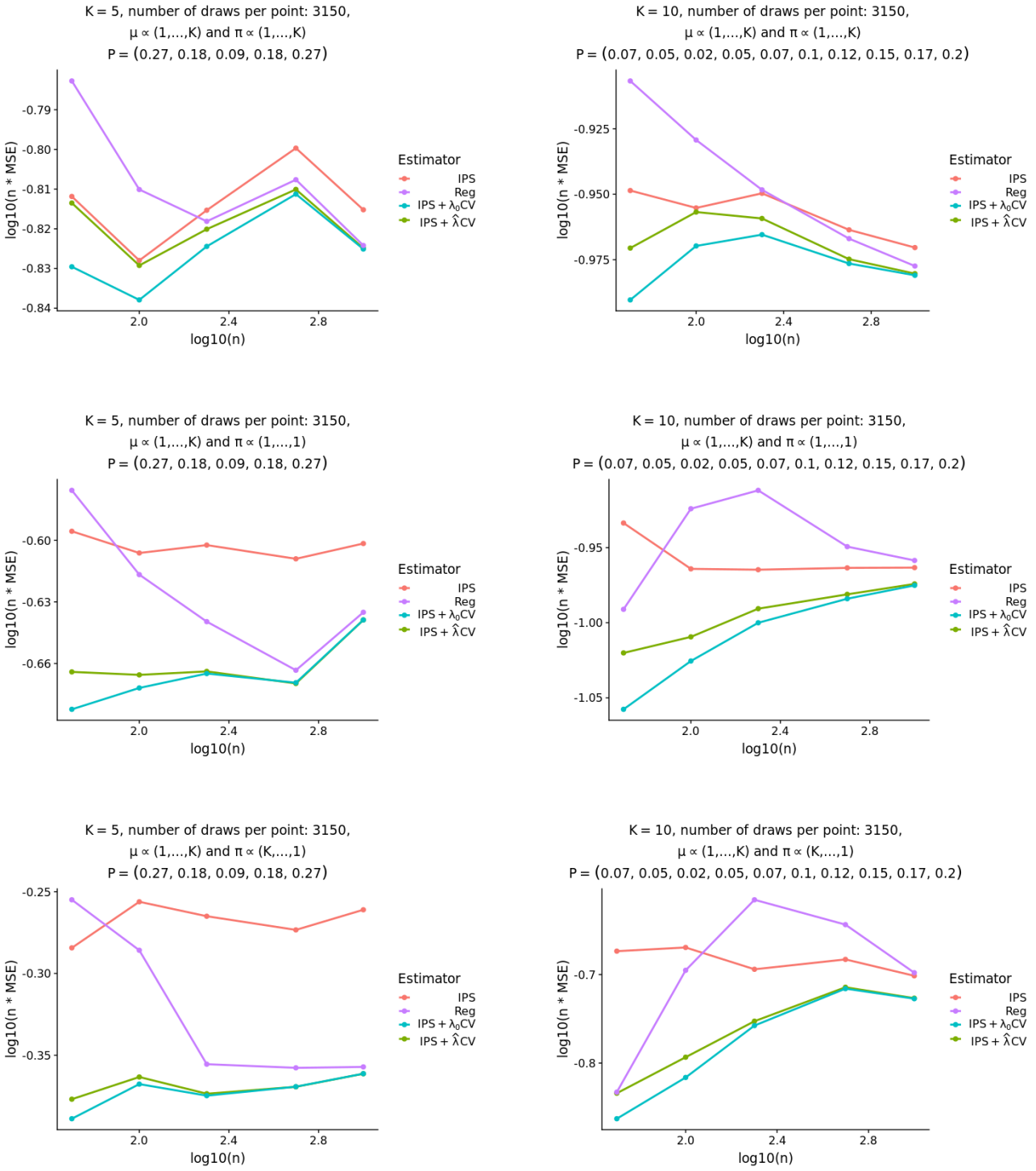

Figure 1. Comparing IPS and REG with an optimized IPS - λt estimator from Section 4.

5. Contextual Bandits

In this section we demonstrate an application of the main ideas to contextual bandits. In particular, using variants of the Doubly Robust estimator as base estimators, we show that by operating outside the manifold of the base estimator we can get estimators with improved risk.

In the stochastic contextual bandit setting, there is a finite set of arms, $a \in \{1, 2, \ldots, K\}$. The environment produces $(x, r) \sim D$, where x is a d-dimensional context vector x and $r = (r(1), \ldots, r(K))$ is the reward associated with each arm in $\{1, 2, \ldots, K\}$. A logging policy μ chooses arm $a \in \{1, 2, \ldots, K\}$ for context x with probability $\mu(a|x)$ and observes the reward only for the chosen arm, $r(a)$. We are interested in estimating the value v of a deterministic target policy π, defined as

$$v = \mathbb{E}_{(x,r)\sim D} r(\pi(x)),$$

from n observations $\{(x_i, a_i, r_i)\}_{i=1}^n$ logged by policy μ.

5.1. Popular Estimators

We provide a brief overview of some popular off-policy evaluation estimators used in the literature:

- Direct Method (DM) forms an estimate $\hat{r}(x, a)$ of the expected reward conditioned on the context and action. The policy value is then estimated by

$$\text{DM} = \frac{1}{n} \sum_{i=1}^{n} \hat{r}(x_i, \pi(x_i)) \,.$$

  DM corresponds to REG in traditional (i.e., non-contextual) multi-armed bandits. If $\hat{r}(x, a)$ is an unbiased estimator of the expected reward conditioned on the context and action, then DM is an unbiased estimator of v. Otherwise DM is biased.

- Inverse Propensity Score (IPS) forms an estimate $\hat{\mu}(a|x)$ of the probability that the logging policy μ chooses action a for context x (propensity). The policy value is then estimated by

$$\text{IPS} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}[\pi(x_i) = a_i]}{\hat{\mu}(a_i|x_i)} r_i \,.$$

  It is often the case that we have access to the logging policy μ and IPS is an unbiased estimator of v. However, IPS may suffer large variance, particularly when the probabilities $\mu(a_i|x_i)$ become small.

- Doubly Robust (DR) (Dudík et al., 2014) takes advantage of both the estimate of the expected reward $\hat{r}(x, a)$ and the estimate of action probabilities $\hat{\mu}(a|x)$ in the estimate DR.

$$\text{DR} = \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{\mathbb{I}[\pi(x_i) = a_i]}{\hat{\mu}(a_i|x_i)} (r_i - \hat{r}(x_i, a_i)) + \hat{r}(x_i, \pi(x_i)) \Big] \,.$$

  If one of the two estimators is unbiased, then DR is unbiased and has much lower variance than IPS. Hence, if the propensities $\mu(a|x)$ are known, DR is unbiased.

- More Robust that Doubly Robust (MRDR) (Farajtabar et al., 2018) is a variation of DR with a DM reward function derived from minimizing the variance of DR. If the model $\hat{r}(x, a)$ of the expected reward conditioned on the context and action is parameterized by β, then MRDR finds the model parameter by solving a weighted least squares problem

$$\beta^* = \operatorname*{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^{n} \Big[\mathbb{I}[\pi(x_i) = a_i] \cdot \frac{1 - \hat{\mu}(a_i|x_i)}{\hat{\mu}(a_i|x_i)^2} (r_i - \hat{r}(x_i, a_i; \beta))^2 \Big] \,.$$

  Similarly to DR, if the propensities $\mu(a|x)$ are known, MRDR is unbiased.

5.2. Control Variates for Doubly Robust Estimators

We propose a new family of estimators for contextual bandits that given an unbiased off-policy evaluation estimator for contextual bandits, it subtract an IPS control variate in order to decrease the risk of the estimator. Particularly, we focus on the doubly robust estimators, DR and MRDR.

- DR with IPS control variate:

$$\text{DR IPS CV} = \text{DR} - \kappa_{\text{DR}} \cdot \text{IPS} \,.$$

  The coefficient κ_{DR} is chosen so that the risk of the estimator is minimized:

$$\kappa_{\text{DR}} = \frac{\text{Cov}(\text{DR}, \text{IPS})}{\mathbb{E}[IPS]^2 + \text{Var}(\text{IPS})} \,.$$

  Taking sample average equivalents (i.e., replacing the population covariance, variance, and expectation in the above expression with their empirical versions), we obtain a consistent estimator $\hat{\kappa}_{\text{DR}}$ of κ_{DR} given by

$$\hat{\kappa}_{\text{DR}} = \frac{\widehat{\text{Cov}}(\text{DR}, \text{IPS})}{\widehat{\mathbb{E}}[IPS]^2 + \widehat{\text{Var}}(\text{IPS})} \,. \tag{15}$$

- MRDR with IPS control variate:

$$\text{MRDR IPS CV} = \text{MRDR} - \kappa_{\text{MRDR}} \cdot \text{IPS} \,.$$

In analogy with κ_{DR}, the coefficient κ_{MRDR} is chosen so that the risk of the estimator is minimized:

$$\kappa_{\text{MRDR}} = \frac{\text{Cov}(\text{MRDR}, \text{IPS})}{\mathbb{E}[IPS]^2 + \text{Var}(\text{IPS})},$$

and a consistent sample-based estimator $\hat{\kappa}_{\text{MRDR}}$ of κ_{MRDR} is given by

$$\hat{\kappa}_{\text{MRDR}} = \frac{\widehat{\text{Cov}}(\text{MRDR}, \text{IPS})}{\widehat{\mathbb{E}}[IPS]^2 + \widehat{\text{Var}}(\text{IPS})}. \tag{16}$$

5.3. Experiments: Multiclass Classification with Bandit Feedback

As in Dudík et al. (2014), we turn a K-class classification task into a K-armed contextual bandit problem. This transformation allows us to compare different estimators for off-policy evaluation using public datasets. In a classification task, we assume data are drawn IID from a fixed distribution: $(x, c) \sim D$, where $x \in X$ is the context and $c \in \{1, 2, \ldots, K\}$ is the class. Each observation (x, c) is equivalent to observing $(x, r_1, \ldots, r_K)$, where $r_a = \mathbb{I}[a = c]$ is the reward for predicting label a when the true label is c.

We compare the following contextual off-policy estimators DM, IPS, DR, MRDR with our proposed estimators DR & IPS CV and MRDR & IPS CV. We use the same 9 benchmark datasets from the UCI repository (Dua & Graff, 2017; Asuncion & Newman, 2007) as in Dudík et al. (2014).

For the evaluation on a dataset, we follow the methodology of Dudík et al. (2014).

1. We randomly split data into training and test sets of the same size.
2. On the fully labeled training set, we run logistic regression to obtain a classifier $\pi : X \to \{1, 2, \ldots, K\}$. This classifier serves as the deterministic target policy π, which we wish to evaluate.
3. We compute the classification accuracy on fully observed test data. This accuracy is treated as the ground truth for the value v of policy π.
4. We use the test set to construct the logging data with which policy π will be evaluated. We transform the test set into a partially labeled dataset using a stochastic logging policy μ. For any observation $(x, r_1, \ldots, r_K)$ in the test set, we randomly select a label $a \sim \mu(a|x)$ and we only reveal the reward of label a. The final data are of the form (x, a, r_a).

It is common in practice to have mismatches between logging and target policies. Specifically, if for a context x the target policy selects label $\pi(x)$, the logging policy μ selects label $\pi(x)$ with probability $\epsilon = 0.05$ and with probability $1 - \epsilon$ the logging policy μ selects one of the other labels $\{1, 2, \ldots, K\} \backslash \pi(x)$ uniformly at random. Furthermore, the propensities $\mu(a|x)$ are assumed to be known.

DM, DR, MRDR, DR & IPS CV and MRDR & IPS CV require estimating the expected conditional reward denoted as $r(x, a)$ for given (x, a). As in Dudík et al. (2014), we use a linear loss model $\hat{r}(x, a) = w_a \cdot x$ parameterized by K weight vectors $\{w_a\}_{a \in 1, \ldots, K}$ and use least-squares regression to fit w_a based on a partially labeled dataset from the training set.

For each dataset, we repeat step 4, $N = 500$ times. Each repetition j results in a different partially labeled dataset, because the logging policy μ is stochastic. On each dataset evaluate the accuracy of each estimator in terms of RMSE. RMSE is defined as $\sqrt{\sum_{j=1}^{N} (\hat{v}_j - v)^2 / N}$ where v is the true value of the policy and $\hat{v}$ is the value of the policy returned by the estimator on the jth repetition.

The code for all experiments is available by request from the authors.

In Table 1 we report the RMSE of the different estimators for each benchmark. The characteristics of the datasets are reported at the top of the table. In smaller datasets (below 1500 observations), DR & IPS CV always improves over DR, and MRDR & IPS CV always improves over MRDR. In larger datasets (above 5500 observations), the $\hat{\kappa}_{\text{DR}}$ and $\hat{\kappa}_{\text{MRDR}}$ computed in (15) and (16) are very close to zero, making DR & IPS CV coincide with DR, and MRDR & IPS CV coincide with MRDR. This is consistent with the theory, as MRDR and DR are known to by asymptotically efficient. However, in small samples the IPS control variate improves the efficiency of these estimators. Overall, one of DR & IPS CV and MRDR & IPS CV have the lowest RMSE or tie with their respective unbiased counterparts, DR and MRDR. (The confidence intervals overlap for DR/DR-CV, MRDR/MRDR-CV but they also overlap for DR/MRDR. The mean, lower confidence bound and upper confidence bound is consistently below for the CV-variants than the non-CV variants.)

6. Conclusions

We addressed the problem of designing an instance-dependent estimator for a bandit problem. The proposed framework assumes the existence of a base estimator and a parametrized class of control variates, and it involves an optimization problem to locate a control variate in that class that reduces the risk of the base estimator.

Our contributions can be summarized as follows: (1) We proposed the use of parametric control variates for off-

Dataset	ecoli	glass	letter	optdigits	page-blocks	pendigits	satimage	vehicle	yeast
Classes (K)	8	6	26	10	5	10	6	4	10
Dataset size	336	214	20000	5620	5473	10992	6435	846	1484
DM	0.4837	0.3170	0.4063	0.4306	0.0715	0.5281	0.3259	0.4090	0.1914
IPS	0.3074	0.3092	0.0334	0.0815	0.0800	0.0537	0.0724	0.1901	0.1111
DR	0.2136	0.2497	0.0246	0.0411	0.0261	0.0325	0.0402	0.1298	0.0840
MRDR	0.1673	0.3185	**0.0266**	0.0290	**0.0344**	**0.0196**	**0.0302**	0.1194	0.0824
DR IPS CV	0.2099	**0.2271**	0.0246	0.0409	0.0261	0.0325	0.0400	0.1266	0.0827
MRDR IPS CV	**0.1665**	0.3103	**0.0266**	**0.0289**	**0.0344**	**0.0196**	**0.0302**	**0.1182**	**0.0818**

Table 1. RMSE of DM, DR, MRDR, DR & IPS CV and MRDR & IPS CV on UCI benchmark datasets.

policy evaluation, and the possibility of designing instance-dependent control variates via optimization. (2) We showed that the population risk is closed-form for a parametrized class of estimators involving polynomial functions of sufficient statistics (see Section 3.3). (3) We demonstrated experimentally (see Section 4 and Fig.1) that there *do* exist instance-dependent estimators that improve REG in the finite-sample regime, and those estimators can be learned from the data. The existence (and the possibility of automating the design) of instance-dependent estimators that are better than REG is a fascinating empirical finding, which was made possible by the techniques that we developed in this work.

The proposed framework offers many avenues for further research. The most important open question is whether, for any given bandit instance, there exists a family of control variates that guarantees minimax risk improvement for any value of the true population parameters. The existence of such a family, together with the issue of efficiently searching in this family for a near-optimal solution, are interesting open problems.

References

Dua, D. and Graff, C. UCI Machine Learning Repository `http://archive.ics.uci.edu/ml` University of California, Irvine, School of Information and Computer Sciences, 2017.

Asuncion, A, and Newman, D.J. UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science, 2007.

Athey, S, and Wager, S. Efficient policy learning. *arXiv*:1702.02896, 2017.

Bang, Heejung and Robins, James M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Bembom, Oliver and van der Laan, Mark J. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. 2008.

Cassel, Claes M, Särndal, Carl E, and Wretman, Jan H. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Dudík, M, Langford, J, and Li, L. Doubly robust policy evaluation and learning. In *ICML*, 2011.

Dudík, M, Erhan, D, Langford, J, and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Farajtabar, M, Chow, Y, and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *ICML*, 2018.

Ghosh, A, Chowdhury, S R, and Gopalan, A. Misspecified linear bandits. In *AAAI*, 2017.

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pp. 315–331, 1998.

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

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

Imbens, Guido, Newey, Whitney, and Ridder, Geert. Mean-squared-error calculations for average treatment effects. Technical report, 2007.

Greensmith E, Bartlett P L, and Baxter J. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov):1471-1530, 2004.

Joachims, T, Swaminathan, A, and de Rijke, M. Deep learning with logged bandit feedback. In *ICLR*, 2018.

Kallus, N, and Zhou, A. Confounding-robust policy improvement. *NIPS*, 2018.

Lattimore, T, and Szepesvári, C. Bandit Algorithms. *Cambridge University Press*. In press, 2018.

Li, L, Munos, R, and Szepesvári, C. Toward minimax off-policy value estimation. In *AISTATS*, 2015.

Mosimann, J.E. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2), 65-82, 1962.

Robins, J M and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Swaminathan A, Krishnamurthy A, Agarwal A, Dudík M, Langford J, Jose D, Zitouni I. Off-policy evaluation for slate recommendation. In *NIPS*, 2017.

Thomas, P S, and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *ICML*, 2016.

Wang, Y-X, Agarwal, A, and Dudík, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *ICML*, 2017.