Toward Minimax Off-policy Value Estimation

Lihong Li
Microsoft Research

Rémi Munos
Google DeepMind

Csaba Szepesvári
University of Alberta

Abstract

This paper studies the off-policy evaluation problem, where one aims to estimate the value of a target policy based on a sample of observations collected by another policy. We first consider the single-state, or multi-armed bandit case, establish a finite-time minimax risk lower bound, and analyze the risk of three standard estimators. For the so-called regression estimator, we show that while it is asymptotically optimal, for small sample sizes it may perform suboptimally compared to an ideal oracle up to a multiplicative factor that depends on the number of actions. We also show that the other two popular estimators can be arbitrarily worse than the optimal, even in the limit of infinitely many data points. The performance of the estimators are studied in synthetic and real problems; illustrating the methods strengths and weaknesses. We also discuss the implications of these results for off-policy evaluation problems in contextual bandits and fixed-horizon Markov decision processes.

1 Introduction

In reinforcement learning including multi-armed bandits, one of the most fundamental problems is *policy evaluation* — estimating the average reward obtained by running a given policy to select actions in an unknown system. A straightforward solution is to simply run the policy and measure the average reward collected. In many applications, however, running a new policy in the actual system can be expensive or even impossible. For example, flying a helicopter with a new policy can be risky as it may lead to crashes; deploying a new ad display policy on a website may be catastrophic to user experience; testing a new treatment on patients may simply be impossible for legal and ethical reasons; *etc.*

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

It is the purpose of *off-policy* evaluation (Precup et al., 2000, Sutton et al., 2010), sometimes referred to as *offline evaluation* in the bandit literature (Li et al., 2011) or *counterfactual reasoning* (Bottou et al., 2013) to overcome this problem. Here, we still aim to estimate the average reward of a target policy, but instead of running it directly, we only have access to a sample of observations made about the unknown system, which may be collected in the past using a *different* policy. Off-policy evaluation has been found useful in several important applications (Langford et al., 2008, Li et al., 2011, Bottou et al., 2013) and can also be regarded as a key building block for policy *optimization* which, as in supervised learning, can often be reduced to evaluation, as long as the complexity of the policy class is well-controlled (Ng and Jordan, 2000). Accordingly, off-policy evaluation was found to be useful in many optimization algorithms for Markov decision processes (e.g., Heidrich-Meisner and Igel 2009) and bandit problems (Auer et al., 2002, Langford and Zhang, 2008, Strehl et al., 2011).

In the context of supervised learning, off-policy learning is known as the *covariate shift* problem, where one estimates losses under changing distributions for model selection (Quiñonero-Candela et al., 2008, Yu and Szepesvári, 2012) and is also related to active learning (Dasgupta, 2011). In statistics, the problem appears in the context of causal effect estimation from controlled experiments (e.g., Holland (1986)), where one is to estimate an intervention's effect on outcomes based on *observational* data that are collected by a different intervention. Thus, results established here may have useful implications in these related problems.

The topic of the present paper is off-policy evaluation in finite settings, under a mean squared error (MSE) criterion. As opposed to the statistics literature (e.g., Hirano et al. (2003)), in addition to the asymptotics, we are also interested in results for *finite sample sizes*. In particular, we are interested in limits of performance (minimax MSE) given fixed policies, but unknown stochastic rewards with bounded mean reward, as well as the performance of estimation procedures compared to the minimax MSE. We are not aware of prior work that would have studied the above problem (i.e., relating the MSE of algorithms to the best possible MSE). The pros and cons of minimax estimation are discussed at length in well-known textbooks (Kiefer,

1987, Lehmann and Casella, 1998). We view achieving the minimax optimal MSE as a modest goal: if an estimator fails to achieve it (up to a constant factor), its use is not recommended unless additional knowledge is available. However, as we will see in our problem setting, even achieving this modest goal is nontrivial. The picture is further complicated by the fact that our estimators are not given a bound on the mean reward function.

Our main results are as follows: We start with a lower bound on the minimax MSE, as well as an asymptotic lower MSE to set a target for the estimation procedures. Next, we derive the exact MSE of the importance sampling estimator (IS), which is shown to have an extra (uncontrollable) factor as compared to the lower bounds. We then consider the weighted version of the IS estimator (WIS) and argue that it shares the same limitation as the IS estimator. Next, we consider the estimator which estimates the mean rewards by sample means, which we call the regression estimator (REG). The motivation of studying this estimator is both its simplicity and also because it is known that a related estimator is asymptotically efficient (Hirano et al., 2003). The main question is whether the asymptotic efficiency transfers into finite-time efficiency. Our answer to this is mixed: We show that for a large class of settings the MSE of REG is within a constant factor of the minimax MSE lower bound; however, the "constant" depends on the number of actions (K), or a lower bound on the variance. We also show that the dependence of the MSE of REG on the number actions is unavoidable. Therefore, while REG is asymptotically optimal, in finite-sample settings it may be less than ideal except for "small" action sets or high noise setting, when it can be thought of as a minimax near-optimal estimator. We also show that for sample sizes up to $\sqrt{K}$ all estimators must suffer a constant MSE.

2 Multi-armed Bandits

We first introduce the problem studied. Let $\mathcal{A} = \{1, 2, \ldots, K\}$ be a finite set of K actions. Data $D^n = ((A_1, R_1), \ldots, (A_n, R_n)) \in (\mathcal{A} \times \mathbb{R})^n$ is generated by the following process: Given a distribution $\pi_D \in \mathbf{\Delta}_{\mathcal{A}}$ over $\mathcal{A}$ (i.e., $\pi_D : \mathcal{A} \to [0, 1]$ such that $\sum_a \pi_D(a) = 1$),

$$A_i \sim \pi_D(\cdot), \quad R_i \sim \Phi(\cdot|A_i), \quad i = 1, \ldots, n\,,$$

for some collection $\Phi = (\Phi(\cdot|a))_{a\in\mathcal{A}}$ of distributions over the reals, indexed by actions. It is assumed that each pair (A_i, R_i) is independent of the others. We think of R_i as the random reward for action A_i, and π_D as a *policy* generating the actions. The problem is to estimate the *value*

$$v_\Phi^\pi := \mathbb{E}_{A\sim\pi, R\sim\Phi(\cdot|A)}[\,R\,] \;\; (= \textstyle\sum_a \pi(a) r_\Phi(a)\,)$$

of some *target policy* $\pi \in \mathbf{\Delta}_{\mathcal{A}}$, possibly different from the data generating policy π_D. Here, in the second expression shown in the parenthesis, $r_\Phi(a) = \mathbb{E}_{R\sim\Phi(\cdot|a)}[R]$ is the *mean reward* of action a. The estimate $\widehat{v}$ must be constructed based on π, π_D, and the data D^n only and we view an estimator $\mathbf{A}$ as a function that maps triplets (π, π_D, D^n) to some estimate $\widehat{v} \in \mathbb{R}$. The quality of an estimate $\widehat{v}$ produced by an estimator is measured by its mean squared error, $\mathrm{MSE}\,(\widehat{v}) := \mathbb{E}\left[(\widehat{v} - v_\Phi^\pi)^2\right]$. The *off-policy value estimation problem in multi-armed bandits* is the problem of constructing estimators of the above form with low mean squared error (MSE). This can be viewed as the simplified version of the full-blown off-policy value estimation problem in Markovian Decision Problems, which is more prevalent in the literature (see the references in the introduction and Section 3).

The task is clearly infeasible if $\pi_D(a) = 0$ for some $a \in \mathcal{A}$, hence in what follows we always assume that $\pi_D(a) > 0$ for all actions $a \in \mathcal{A}$. The question then is how sensitive an estimator will be or must be to small values of π_D. The fact that π_D is small alone will not necessarily lead to high MSE. For example, if π agrees with π_D, then the fact that some values of $\pi_D(a)$ are small will not matter. Similarly, if the *reward variance* $\sigma_\Phi^2(a) := \mathbb{V}_{R\sim\Phi(\cdot|a)}(R)$ is very small, even a few reward observations at the a are sufficient to estimate the *mean reward* $r_\Phi(a)$ with a small error, mitigating the negative effect a small probability $\pi_D(a)$. A "reasonable" estimator exploits these effects. Indeed, from a reasonable estimator we expect that if a problem instance is "easier," the estimator will have a smaller MSE, i.e., the estimator should *adapt* to the difficulty of problem instances. A rigorous study of this problem is the main topic of the present paper.

The rest of the paper is organized as follows: To define what can be reasonably expected from an estimator, in Section 2.2 we will first establish several lower bounds on the MSE of unrestricted estimators, both for finite n and when $n \to \infty$. In the next sections (Section 2.3 and Section 2.4) we will investigate several popular estimation methods, comparing upper bounds on their MSE to the previously obtained lower bounds, thus highlighting their strengths and weaknesses. These findings are complemented in Section 2.5 with simulation results both on synthetic and real-world data (illuminating the strengths and weaknesses of the theoretical results), while in Section 3 we will look at the implications of our results beyond multi-armed bandits.

2.1 Notation

We shall denote by $\pi_D \otimes \Phi$ the common distribution underlying the random pairs (A_i, R_i) (i.e., $\pi_D \otimes \Phi$ is a distribution on $\mathcal{A} \times \mathbb{R}$). We let $\pi_D^* := \min_a \pi_D(a)$ and $\pi(B) := \sum_{a\in B} \pi(a)$ for $B \subseteq \mathcal{A}$. For convenience, we identify a function $f : \mathcal{A} \to \mathbb{R}$ with the K-dimensional vector whose kth component is $f(k)$. Thus, r_Φ, σ_Φ^2, etc. are considered vectors. For $x, y \in \mathbb{R}^K$, $x \le y$ means $x_i \le y_i$ for all $1 \le i \le K$. The set of all distribution families

$(\Phi(\cdot|a))_{a\in\mathcal{A}}$ indexed by $\mathcal{A}$ is denoted by Ψ. We let $\mathbb{R}_+$ denote the set of nonnegative reals and for $\sigma^2 \in \mathbb{R}_+^K$, we denote by Ψ_{σ^2} the set of $\Phi \in \Psi$ such that $\sigma_\Phi^2 \le \Phi$, while we denote by $\Psi_{\sigma^2,R_{\max}}$ the subset of Ψ_{σ^2} such that for any $\Phi \in \Psi_{\sigma^2,R_{\max}}$, $0 \le r_\Phi(a) \le R_{\max}$ holds for all $a \in \mathcal{A}$.

The following quantities will facilitate discussions:

$$V_1 := \mathbb{E}\left[\mathbb{V}\left(\frac{\pi(A)}{\pi_D(A)}R|A\right)\right] = \sum_a \frac{\pi^2(a)}{\pi_D(a)}\sigma_\Phi^2(a)\,, \quad (1a)$$

$$V_2 := \mathbb{V}\left(\mathbb{E}\left[\frac{\pi(A)}{\pi_D(A)}R|A\right]\right) = \mathbb{V}\left(\frac{\pi(A)}{\pi_D(A)}r_\Phi(A)\right)$$
$$= \sum_a \frac{\pi^2(a)}{\pi_D(a)}r_\Phi(a)^2 - (v_\Phi^\pi)^2\,. \quad (1b)$$

Note that V_1 and V_2 are functions of Φ, π_D and π, but this dependence is suppressed. Also, V_1 and V_2 are independent in that there are no constants $c, C > 0$ such that $cV_1 \le V_2 \le CV_1$ for any π, π_D, Φ. For subsets $B \subseteq \mathcal{A}$, we denote by $p_{B,n}$ the probability that none of the actions in D^n falls into B; that is, $p_{B,n} = \mathbb{P}(A_1, \ldots, A_n \notin B)$. Hence, $p_{B,n} = (1 - \pi_D(B))^n$. For singletons $B = \{a\}$, the shorthand $p_{a,n}$ is used instead of $p_{\{a\},n}$.

2.2 Lower Bounds

We start with establishing a minimax lower bound that characterizes the inherent hardness of the off-policy value estimation problem. As noted before, an estimator $\mathbf{A}$ is considered as a function that maps (π, π_D, D^n) to an estimate of v_Φ^π, denoted $\widehat{v}_{\mathbf{A}}(\pi, \pi_D, D^n)$. Fix $\sigma^2 := (\sigma^2(a))_{a\in\mathcal{A}}$. We consider the minimax optimal MSE over the class of problems where $\Phi \in \Psi_{\sigma^2,R_{\max}}$:

$$R_n^*(\pi, \pi_D, R_{\max}, \sigma^2) :=$$
$$\inf_{\mathbf{A}} \sup_{\Phi\in\Psi_{\sigma^2,R_{\max}}} \mathbb{E}_{\pi_D\otimes\Phi}\left[(\widehat{v}_{\mathbf{A}}(\pi, \pi_D, D^n) - v_\Phi^\pi)^2\right]\,,$$

where by $\mathbb{E}_{\pi_D\otimes\Phi}$ we denote the expectation operator underlying the probability measure $\mathbb{P}_{\pi_D\otimes\Phi}$ under which the joint distribution of the data $D^n = ((A_1, R_1), \ldots, (A_n, R_n))$ is $(\pi_D \otimes \Phi)^n$. The restriction on the magnitude of the mean reward function through $R_{\max}$ is necessary because $\lim_{R_{\max}\to\infty} R_n^*(\pi, \pi_D, R_{\max}, \sigma^2) = \infty$. The intuitive explanation of this is that for any $n > 0$, the probability that for some action $a \in \mathcal{A}$ there is no reward observed for a is positive. Under this event no estimator can guess a correct value of the underlying mean reward. Of course, one may object that an estimator may not need to estimate the mean reward of the actions, but a rigorous formal argument shows that this does not allow any estimator to escape from having an unbounded MSE when the range of r_Φ is unbounded.

The first part of the theorem below shows that the minimax MSE scales quadratically with $R_{\max}$, while the second part shows that when $R_{\max}$ or n is large, the minimax MSE scales with V_1/n where V_1 is defined like in (1a), with σ_Φ^2 replaced by σ^2—the largest possible variance within the class $\Psi_{\sigma^2,R_{\max}}$. Thus, as expected, larger variances make the problem harder, though V_1 captures more finely the relationship between π, π_D and the variances. The final part shows that the constant multiplying V_1/n can be increased to 1 asymptotically, as n becomes large.

Theorem 1. *For any $n > 0$, π_D, π, $R_{\max}$ and σ^2, one has*

$$R_n^*(\pi, \pi_D, R_{\max}, \sigma^2) \ge \frac{1}{4}R_{\max}^2 \max_{B\subset\mathcal{A}} \pi^2(B)p_{B,n}\,.$$

Furthermore, provided that

$$\max_a \frac{\pi(a)\sigma^2(a)}{\pi_D(a)}\sqrt{\frac{0.6}{V_1}} \le \sqrt{n}R_{\max}\,, \quad (2)$$

we also have that

$$R_n^*(\pi, \pi_D, R_{\max}, \sigma^2) \ge 0.02\,\frac{V_1}{n}\,,$$

where $V_1 = \sum_a \frac{\pi^2(a)}{\pi_D(a)}\sigma^2(a)$. Finally,

$$\liminf_{n\to\infty} \frac{R_n^*(\pi, \pi_D, R_{\max}, \sigma^2)}{V_1/n} \ge 1. \quad (3)$$

One may wonder about the necessity of condition (2) required by the second lower bound. However, intuitively, a lower bound of the form V_1/n can only hold when $R_{\max}$ is large compared to at least V_1/n since the minimax MSE over $\Psi_{\sigma^2,R_{\max}}$ converges to zero as $R_{\max} \to 0$. Indeed, a minimax estimator for the class $\Psi_{\sigma^2,R_{\max}}$ may well use the knowledge of $R_{\max}$ to limit its loss by exploiting that the value to be estimated lies in the interval $[0, R_{\max}]$, hence, only estimates that belong to this interval make sense. A well known technique to exploit such knowledge is to truncate a preliminary estimate to this interval. However, in this paper we focus on estimators that have no a priori knowledge of an upper bound on the range of rewards, hence we will not consider this problem. On a related note, it is possible to extend the proof to remove condition (2) at the expense of a more complicated lower bound. We also leave this to future work.

Proof sketch. Full proofs for the three parts are given in Appendices A.1–A.3, respectively. The first part's proof follows the intuition already given in the text. The second part is proved by standard lower bounding techniques: We choose two problems with similar reward distributions, Φ_1 and Φ_2, so that achieving ε MSE within $\{\Phi_1, \Phi_2\}$ is equivalent to telling which of Φ_1 and Φ_2 is the true distribution that generated data D^n. Fano's inequality is then applied to yield the desired result. The third part is proved directly by the Cramer-Rao lower bound. □

A simple corollary of the previous theorem is that the minimax risk is constant when the number of samples is "small" and the worst target policy is chosen:

Corollary 1. *For* $K \geq 2$, $n \leq \sqrt{K}$, $\sup_\pi R_n^*(\pi, \pi_D, R_{\max}, \sigma^2) = \Omega(R_{\max}^2)$.

Proof. Choose $B \subset \mathcal{A}$ to minimize $\pi_D(B)$ subject to the constraint $|B| = \lfloor \sqrt{K} \rfloor$. Note that $\mathbb{P}(A_1, \ldots, A_n \notin B) = (1-\pi_D(B))^n \geq (1-\frac{|B|}{K})^n \geq (1-\frac{1}{\sqrt{K}})^{\sqrt{K}} \geq (1-\frac{1}{\sqrt{2}})^{\sqrt{2}}$. Choosing π such that $\pi(B) = 1$ gives the result. □

In particular, the result means that in a worst-case sense, no estimator can achieve a nontrivial MSE for small sample sizes, or alternatively, all estimators are equally poor in this regime, at least in the above worst-case sense. The proof also reveals that the worst-case target policy is supported on the subset of $\Theta(\sqrt{K})$ actions that π_D is the least likely to sample from. We conjecture that the result can be strengthened by increasing the upper limit on n.

2.3 Importance Sampling Estimators

One of the most popular estimators is known as the propensity score estimator in the statistical literature (Rosenbaum and Rubin, 1983, 1985), or the importance weighting estimator (Bottou et al., 2013). We call it the importance sampling (IS) estimator, as it estimates the unknown value using likelihood ratios, or importance weights:

$$\widehat{v}_{\mathrm{IS}}(\pi, \pi_D, D^n) := \frac{1}{n} \sum_{i=1}^n \frac{\pi(A_i)}{\pi_D(A_i)} R_i .$$

This estimator is *unbiased*: $\mathbb{E}[\widehat{v}_{\mathrm{IS}}(\pi, \pi_D, D^n)] = v_\Phi^\pi$, implying that the MSE is purely contributed by the variance of the estimator. The main result in this subsection shows that this estimator does not achieve the minimax lower bound up to *any* constant. The proof (given in Appendix A.4) is based on a direct calculation using the law of total variance.

Proposition 1. $\mathrm{MSE}\left(\widehat{v}_{\mathrm{IS}}(\pi, \pi_D, D^n)\right) = (V_1 + V_2)/n$.

In the next section, we will see that

$$\lim_{n\to\infty} \frac{R_n^*(\pi, \pi_D, R_{\max}, \sigma^2)}{V_1/n} = 1 , \tag{4}$$

showing that $\frac{\widehat{v}_{\mathrm{IS}}(\pi, \pi_D, D^n)}{R_n^*(\pi, \pi_D, R_{\max}, \sigma^2)} = 1 + \frac{V_2}{V_1} + \omega(1)$, i.e., the risk of IS is (asymptotically) $1 + \frac{V_2}{V_1}$ times the optimal risk; the larger V_2 and the smaller V_1 are, the worse is the risk of IS in the limit compared to the optimum.

A modification of the IS estimator, known as the *weighted importance sampling estimator*, is meant to overcome this weakness. This estimator is given by

$$\widehat{v}_{\mathrm{WIS}} = \sum_{i=1}^n \frac{\frac{\pi(A_i)}{\pi_D(A_i)}}{\sum_{j=1}^n \frac{\pi(A_j)}{\pi_D(A_j)}} R_i .$$

By the law of large numbers, $\frac{1}{n}\sum_j \frac{\pi(A_j)}{\pi_D(A_j)} \to \mathbb{E}\left[\frac{\pi(A_j)}{\pi_D(A_j)}\right] = 1$ as $n \to \infty$, showing that the WIS estimator is consistent. Using the delta method, its asymptotic MSE is given by: (Liu, 2001)

$$\mathrm{MSE}\left(\widehat{v}_{\mathrm{IS}}\right) + \frac{(v_\Phi^\pi)^2}{n} \mathbb{V}\left(\frac{\pi(A)}{\pi_D(A)}\right) - \frac{2v_\Phi^\pi}{n} \mathbf{Cov}\left(\frac{\pi(A)}{\pi_D(A)}, \frac{\pi(A)R}{\pi_D(A)}\right),$$

where $(A, R) \sim \pi_D \otimes \Phi$. Therefore, when $\frac{\pi(A)}{\pi_D(A)}$ and $\frac{\pi(A)R}{\pi_D(A)}$ are highly correlated, as often seen in practice, WIS is a more efficient estimator than IS. Unfortunately, WIS still fails short of being asymptotically minimax optimal. Appendix A.5 gives a full proof of the following theorem:

Theorem 2. *Assume Gaussian reward distributions. Then, for some constants* $(c_a)_{a\in A}$ *that depend on* π, π_D *only but not on the reward variances or means, it holds that* $\mathrm{MSE}\left(\widehat{v}_{\mathrm{WIS}}(\pi, \pi_D, D^n)\right) = \frac{V_1 + \sum_b c_b \pi^2(b)}{n} + \omega\left(\frac{1}{n}\right)$.

Based on (4), we see that $\frac{\mathrm{MSE}(\widehat{v}_{\mathrm{WIS}}(\pi, \pi_D, D^n))}{R_n^*(\pi, \pi_D, R_{\max}, \sigma^2)} = 1 + \frac{\sum_b c_b \pi^2(b)}{V_1} + \omega(\frac{1}{V_1})$. Since V_1 can be made arbitrarily small while keeping $\sum_b c_b \pi^2(b)$ (which only depends on π and π_D) constant, we indeed see that even WIS fails to be asymptotically minimax optimal.

2.4 Regression Estimator

For convenience, define $n(a) := \sum_{i=1}^n \mathbb{I}(A_i = a)$ to be the number of samples for action a in D^n, and $R(a) := \sum_{i=1}^n \mathbb{I}(A_i = a) R_i$ the total rewards of a. The regression estimator (REG) is given by

$$\widehat{v}_{\mathrm{Reg}}(\pi, D^n) := \sum_a \pi(a) \widehat{r}(a),$$

$$\text{where} \qquad \widehat{r}(a) := \begin{cases} 0, & \text{if } n(a) = 0; \\ \frac{R(a)}{n(a)}, & \text{otherwise} . \end{cases}$$

For brevity, we will also write $\widehat{r}(a) = \mathbb{I}\{n(a) > 0\} \frac{R(a)}{n(a)}$, where we take $\frac{0}{0}$ to be zero. The name of the estimator comes from the fact that it estimates the reward function, and the problem of estimating the reward function can be thought of as a regression problem.

Interestingly, as can be verified by direct calculation, the REG estimator can also be written as

$$\widehat{v}_{\mathrm{Reg}}(\pi, D^n) = \frac{1}{n} \sum_{i=1}^n \frac{\pi(A_i)}{\widehat{\pi}_D(A_i)} R_i , \tag{5}$$

where $\widehat{\pi}_D(a) = \frac{n(a)}{n}$ is the empirical estimate of $\pi_D(a)$. Hence, the only difference between the IS estimator and REG is that the former uses π_D to reweight the data, while the latter uses the *empirical estimates* $\widehat{\pi}_D$. It may appear that IS is superior since it uses the "right" quantity. Surprisingly, REG turns out to be much more robust, as will

be shown shortly; see Appendix D for further discussions of a related problem. The robustness of regression estimators is also independently suggested by Nicol (2015).

For the next statement, the counterpart of Proposition 1, the following quantities will be useful:

$$b_n := \sum_a \pi(a) r_\Phi(a) p_{a,n}\,,$$

$$V_{0,n} := b_n^2 + \sum_a \pi^2(a) r_\Phi^2(a)\, p_{a,n}(1-p_{a,n}) \quad \text{and}$$

$$V_{3,n} := \sum_a \mathbb{E}\left[\frac{\mathbb{I}\{n(a)>0\}}{\widehat{\pi}_D(a)} - \frac{1}{\pi_D(a)}\right] \pi^2(a)\sigma^2(a)\,.$$

Note that $b_n = v_\Phi^\pi - \mathbb{E}\left[\widehat{v}_{\mathrm{Reg}}\right]$ is the negative bias of $\widehat{v}_{\mathrm{Reg}}$.

Proposition 2. *Fix π, π_D and assume $r_\Phi \geq 0$. Then it holds that $\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}(\pi, D^n)\right) \leq V_{0,n} + \frac{V_1+V_{3,n}}{n}$. Furthermore, if Φ consist of normal distributions, $\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}\right) \geq \frac{V_1}{n} + 4b_n^2\left(1+\frac{V_1}{n}\right) + \frac{2}{n}\sum_a \frac{\pi^2(a)}{\pi_D(a)}\sigma_\Phi^2(a) p_{a,n}$.*

A full proof is given in Appendix A.6.

Here comes the main result of this section that characterizes the MSE of REG in terms of the minimax optimal MSE.

Theorem 3 (Minimax Optimality of the Regression Estimator)**.** *Let $D_n = \{(A_i, R_i)\}_{i=1,\ldots,n}$ be an i.i.d. sample from (π_D, Φ). Then, the following hold:*

(i) For any $\pi, \pi_D \in \mathbf{\Delta}_K$, $\sigma^2 \in \mathbb{R}_+^K$, $R_{\max} > 0$, $\Phi \in \Psi_{\sigma^2, R_{\max}}$, $n > 0$ such that (2) holds,

$$\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}(\pi, D_n)\right) \leq \{C + 250\}\; R_n^*\,, \qquad (6)$$

where $R_n^* = R_n^*(\pi, \pi_D, R_{\max}, \sigma^2)$ and $C = \min(4K^2, 50K \max_a \frac{r_\Phi^2(a)}{\sigma_\Phi^2(a)})$.

(ii) A suboptimality factor of $\Omega(K)$ in the above result is unavoidable: For $K > 2$, there exists (π, π_D) such that for any $n \geq 1$,

$$\frac{\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}(\pi, D_n)\right)}{R_n^*(\pi, \pi_D, R_{\max}, 0)} \geq n e^{-2n/(K-1)}\,.$$

In particular, for $n = \frac{K-1}{2}$, this ratio is at least $\frac{K-1}{2e}$.

(iii) $\widehat{v}_{\mathrm{Reg}}$ is asymptotically minimax optimal:

$$\limsup_{n\to\infty} \frac{\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}(\pi, D_n)\right)}{R_n^*(\pi, \pi_D, R_{\max}, \sigma^2)} \leq 1\,.$$

While in the proof we will upper bound $V_{3,n}$ in terms of $O(V_1)$, there remains a gap between the lower and upper bounds in this proposition as the second term in the definition of $V_{0,n}$ cannot be matched by any of the terms in the lower bound. Nevertheless, the result shows that for $R_{\max}$ large (or n large), the MSE of REG is upper bounded by a constant multiple of the minimax MSE over $\Psi_{\sigma^2, R_{\max}}$.

However, there are two limitations with the first result in the theorem. First, it only holds for restricted values of $R_{\max}$ (or n) when (2) holds. This is because the lower bound on the minimax MSE expressed in terms of V_1/n only holds for a restricted range of values, which as explained is due to that when $R_{\max}$ is small, V_1/n cannot be a lower bound. As a result, for such small values of $R_{\max}$, REG cannot be "near-minimax" over the class $\Psi_{\sigma^2, R_{\max}}$. As mentioned earlier, if one is given the prior information that the problem instance belongs to $\Psi_{\sigma^2, R_{\max}}$, this can be exploited by introducing a truncation. In the case of REG, this could be done by truncating the estimates of the mean reward to lie in $[0, R_{\max}]$. Although it would be interesting to check whether with this modification REG becomes near-minimax optimal, since here we are more interested in the case when no upper bound on the range of rewards is known, we do not pursue this direction. The second issue with the first bound is that the constant multiplier of the minimax optimal MSE that allows us to bound the MSE of REG in terms of the minimax optimal MSE scales with the number of actions K. In fact, the multiplier scales quadratically with K. In the second part of the theorem we show that a linear scaling of the multiplier as a function of K is inevitable: For $n = \Theta(K)$, the MSE of REG will be at least $\Omega(K)$ times larger than the minimax optimal MSE.

Finally, the last part of the result shows that although for small values of n, the MSE of REG can be significantly larger than the minimax optimal MSE, asymptotically, as $n \to \infty$, the MSE of REG is optimal. This result also shows that the minimax optimal MSE is asymptotically equal to V_1/n.

Proof sketch of Theorem 3. Full proofs for the three parts are given in Appendix A.7. For the first part, we use Proposition 2: $\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}(\pi, D_n)\right) \leq V_{0,n} + \frac{V_1+V_{3,n}}{n}$. We then prove that $V_{3,n} \leq 4V_1$ and that $V_{0,n}$ is upper bounded by $\min\left(K^2 \max_a \pi^2(a) r_\Phi^2(a) p_{a,n}, K \max_{a\in\mathcal{A}}\left(\frac{r_\Phi^2(a)}{\sigma^2(a)}\right)\frac{V_1}{n}\right)$. We conclude by using Theorem 1 to upper bound each term in the previous min by $O(R_n^*)$ provided that (2) holds.

For the second part, we choose $\pi(a) = \pi_D(a) = 1/K$, $r_\Phi(a) = 1$. For $K \geq 2$, $p_{a,n} = (1-1/K)^n = e^{-n\log(1/(1-1/K))} = e^{-n\log(1+1/(K-1))} \geq e^{-n/(K-1)}$. Hence, $\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}\right) \geq \left(\mathbb{E}\left[\widehat{v}_{\mathrm{Reg}} - v_\Phi^\pi\right]\right)^2 = \left(\sum_a \pi(a) r_\Phi(a) p_{a,n}\right)^2 \geq e^{-2n/(K-1)}$. Now consider IS. Choosing $\sigma^2 = 0$, we have $V_1 = 0$ and so by Proposition 1,

$$\sup_{\Phi: 0\leq r_\Phi\leq 1, \sigma_\Phi^2 = 0} \mathrm{MSE}\left(\widehat{v}_{\mathrm{IS}}\right) = \sup_{\Phi: 0\leq r_\Phi\leq 1, \sigma_\Phi^2 = 0} \frac{V_2}{n} \leq \frac{1}{n}\,.$$

Hence, $\frac{\mathrm{MSE}(\widehat{v}_{\mathrm{Reg}})}{R_n^*(\pi,\pi_D,1,0)} \geq \frac{e^{-2n/(K-1)}}{\sup_{\Phi: 0\leq r_\Phi\leq 1, \sigma_\Phi^2=0} \mathrm{MSE}(\widehat{v}_{\mathrm{IS}})} \geq n e^{-2n/(K-1)}$.

Finally, for the last part, we derive a refined bound $V_{3,n} = O(V_1\sqrt{\log(n)/n})$ to derive that for any $\pi, \pi_D, \sigma^2, \Phi$ such

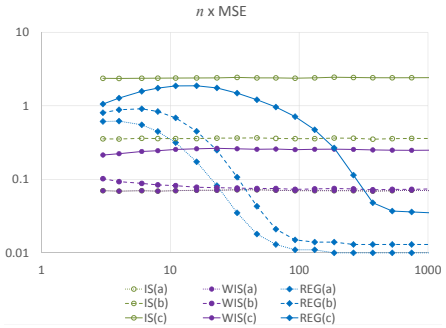

Figure 1: nMSE of IS, WIS and REG in the first synthetic experiment. IS(a) and WIS(a) are almost identical.

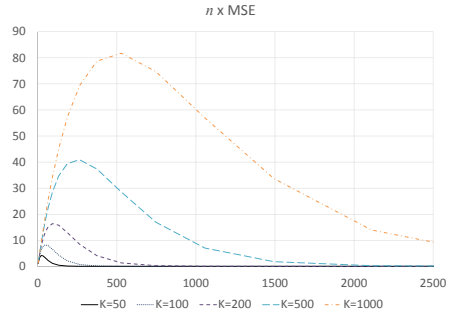

Figure 2: nMSE of REG in the second synthetic experiment. The curves correspond to different numbers of arms.

that $\sigma_\Phi^2 \le \sigma^2$, we have, for n large enough, $\mathrm{MSE}\left(\widehat{v}_{\mathrm{Reg}}\right) \le V_{0,n} + \frac{V_1+V_3}{n} \le ce^{-n/c} + \frac{V_1}{n}\left(1 + c\sqrt{\frac{\ln n}{n}}\right)$, where $c > 0$ is a problem dependent constant. Combining this with (3) of Theorem 1 gives the desired result. □

Summary. The results so far can be summarized as follows: The asymptotic MSE of REG is V_1/n, which is optimal in an asymptotic sense for any instance $\Phi \in \Psi$. The REG estimator is minimax optimal up to a constant multiplier of $O(K^2)$ starting from a well defined range of values for $R_{\max}$ (or n). In this bound, the constant cannot be reduced below $\Omega(K)$, thus for an intermediate range of sample sizes, REG will work worse as the number of actions K becomes large. No algorithm can achieve nontrivial MSE in a worst-case sense for small value of n, i.e., when $n = O(\sqrt{K})$. The IS/WIS estimators are suboptimal, even in an asymptotic sense. Both IS and WIS will have a positive MSE even when the variance of the reward distribution for each action is zero. The MSE of IS is negatively impacted by the variability of the scaled mean reward. While WIS improves most of the time on IS, this relies on the correlation between the importance weights and the importance weights multiplied by the random reward.

2.5 Simulation Results

This subsection corroborates our analysis with simulation results that empirically demonstrate the impact of key quantities on the MSE of the three estimators. We will first use a synthetic setup to demonstrate the behavior of various estimators that is predicted by our analysis above. Then, we use a real dataset to show such phenomena can indeed happen in realistic problems. In all experiments, we repeat the data-generation process (with π_D) $10,000$ times, and compute the MSE of each estimator.

2.5.1 Synthetic Data

Two sets of synthetic experiments are done. All reward distributions are normal distributions with $\sigma^2 = 0.01$ and different means. We then plot normalized MSE (MSE multiplied by sample size n), or nMSE, against n.

The first experiment is to compare the finite-time as well as asymptotic accuracy of $\widehat{v}_{\mathrm{IS}}$, $\widehat{v}_{\mathrm{WIS}}$ and $\widehat{v}_{\mathrm{Reg}}$. We choose $K = 10$, $r_\Phi(a) = a/K$, $\pi(a) \propto a$. Three choices of π_D are used: (a) $\pi_D(a) \propto a$, (b) $\pi_D(a) = 1/K$, and (c) $\pi_D(a) \propto (K - a)$. These choices lead to increasing values of V_2 (with V_1 approximately fixed).

As seen in Figure 1, the nMSE of $\widehat{v}_{\mathrm{IS}}$ remains constant as n increases, equal to $V_1 + V_2$, as predicted in Proposition 1. The nMSE of $\widehat{v}_{\mathrm{WIS}}$ is much smaller and remains roughly unchanged as well. In contrast, the nMSE of $\widehat{v}_{\mathrm{Reg}}$ is large when n is small, because of the high bias, and then quickly converges to the asymptotic minimax rate V_1 (Theorem 3, part iii). As V_2 can be arbitrarily larger than V_1, it follows that $\widehat{v}_{\mathrm{Reg}}$ is preferred over $\widehat{v}_{\mathrm{IS}}$, as least for sufficiently large n that is needed to drive the bias down.[1] Furthermore, although $\widehat{v}_{\mathrm{WIS}}$ can be most efficient when sample size is small, it is inferior to $\widehat{v}_{\mathrm{Reg}}$ asymptotically.

The second experiment is to study how K affects the nMSE of $\widehat{v}_{\mathrm{Reg}}$. Here, we choose $\pi_D = 1/K$, $r_\Phi(a) = a/K$, $\pi(a) \propto a$, and vary $K \in \{50, 100, 200, 500, 1000\}$. As Figure 2 shows, a larger K presents a greater challenge to $\widehat{v}_{\mathrm{Reg}}$, which is consistent with Theorem 3 (part i). Not only does the maximum nMSE grow approximately linearly with K, the number of samples needed for nMSE to start decreasing also scales roughly as $K/2$, perfectly consistent with part ii of Theorem 3.

2.5.2 Real-world Data

We now examine the performance of these popular estimators in a more realistic scenario, using actual data collected on a major commercial search engine. When a query is submitted, the search engine returns a SERP (Search Engine Result Page) that contains an ordered list of URLs. If

[1] It should be noted that in practice, after D^n is generated, it is easy to quantify the bias of $\widehat{v}_{\mathrm{Reg}}$ simply by identifying the set of actions a with $n(a) = 0$.

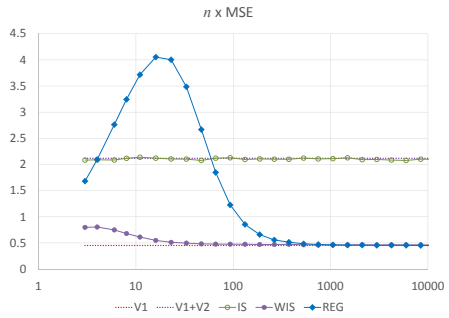

Figure 3: nMSE for query "facebook" ($K = 2178$). The asymptotic rates V_1 and $V_1 + V_2$ are provided for reference.

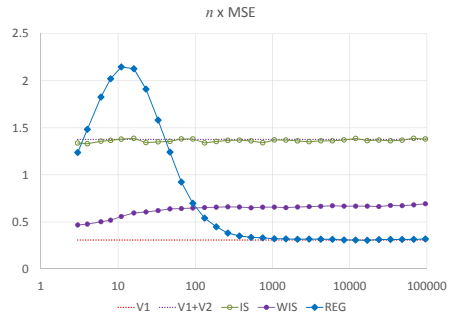

Figure 4: nMSE for query "gmail" ($K = 648$). The asymptotic rates V_1 and $V_1 + V_2$ are provided for reference.

the page contains useful results, the user clicks on the page more likely. For the purpose here, each query defines a multi-armed bandit, where actions are the possible SERPs, and the reward is 1 if the page is clicked on and 0 otherwise.

Over a long period of time, due to constant engineering efforts to improve it, the search engine inevitably displays diversified results — for the same query it may return different SERPs in different time. We first choose a few popular queries such as "facebook" and "gmail". Then, for each of them, we collect all SERPs that have been returned by the search engine in a six-month period, together with their frequencies (i.e., how many times they were returned) and average click probabilities. To avoid unreliable click probabilities, SERPs with low frequencies are removed.

For a fixed query, the data above can be used to build a bandit model (the actions and each action's Bernoulli reward distribution). The sampling probability, $\pi_D(a)$, is the relative frequency of a in the data. Finally, the target policy π is one that chooses uniformly at random the 10 arms with highest frequencies. The off-policy evaluation problem is to estimate the click rate of π, using data collected by π_D. The setup above is intended to mimic realistic scenarios where (good) target policies tend to choose similar arms, and we are interested in estimating click rates from search log, without running expensive online experiments.

Results for query "facebook" is given in Figure 3, where nMSE of the three estimators are compared as sample size increases. Similar to the synthetic experiments, IS is asymptotically non-optimal. The nMSE of REG is relative large with intermediate sample size, but decreases very fast to the asymptotic minimax optimum as n grows. The accuracy of WIS is particularly strong in this case, enjoying a very small nMSE with small sample size and is competitive with REG in the limit. However, for another popular query, "gmail", as shown in Figure 4, nMSE of WIS fails to converge to the asymptotic minimax optimum. Therefore, despite the popularity of WIS in empirical studies, it is not necessarily the most accurate estimator.

Results for the other popular queries we tried are qualitatively similar to one of the two shown here. They suggest our theoretical findings do provide useful insights and predictions for both finite-time and asymptotic accuracy of these popular estimators in real-world applications.

Finally, it is worth mentioning that we also ran preliminary experiments with an estimator that simply combines IS/WIS and REG using $\widehat{v}_{\mathrm{Reg}} + \widehat{v}_{\mathrm{WIS}} \sum_{a:n(a)=0} \pi(a)$ (or replace $\widehat{v}_{\mathrm{WIS}}$ with $\widehat{v}_{\mathrm{IS}}$) to obtain the best of both worlds. The results were encouraging but due to space limitation they are not reported here. It also remains for future work to study the theoretical properties of such estimators.

3 Extensions

In this section, we consider extensions of our previous results to contextual bandits and Markovian Decision Processes, while implications to semi-supervised learning are discussed in the supplementary material.

3.1 Contextual Bandits

The problem setup is as follows: In addition to the finite action set $\mathcal{A} = \{1, 2, \ldots, K\}$, we are also given a context set $\mathcal{X} = \{1, 2, \ldots, M\}$. A policy now is a map $\pi : \mathcal{X} \to [0,1]^{\mathcal{A}}$ such that for any $x \in \mathcal{X}$, $\pi(x)$ is a probability distribution over the action space $\mathcal{A}$. For notational convenience, we will use $\pi(a|x)$ instead of $\pi(x)(a)$. The set of policies over $\mathcal{X}$ and $\mathcal{A}$ will be denoted by $\Pi(\mathcal{X}, \mathcal{A})$. The process generating the data $D^n = \{(X_i, A_i, R_i)\}_{1 \le i \le n}$ is described by the following: (X_i, A_i, R_i) are independent copies of (X, A, R), where $X \sim \mu(\cdot)$, $A \sim \pi_D(\cdot|X)$ and $R \sim \Phi(\cdot|A, X)$ for some unknown family of distributions $\{\Phi(\cdot|a, x)\}_{a \in \mathcal{A}, x \in \mathcal{X}}$ and *known* policy $\pi_D \in \Pi(\mathcal{X}, \mathcal{A})$ and context distribution μ. For simplicity, we fix $R_{\max} = 1$.

We are also given a *known* target policy $\pi \in \Pi(\mathcal{X}, \mathcal{A})$ and want to estimate its value, $v_\Phi^{\pi,\mu} := \mathbb{E}_{X \sim \mu, A \sim \pi(\cdot|X), R \sim \Phi(\cdot|A,X)}[R]$ based on the knowledge of D^n, π_D, μ and π, where the quality of an estimate $\widehat{v}$ constructed based on D^n (and π, π_D, μ) is measured by

its mean squared error, $\mathrm{MSE}\left(\widehat{v}\right) := \mathbb{E}\left[(\widehat{v} - v_\Phi^{\pi,\mu})^2\right]$, just like in the case of contextless bandits. Let $\sigma_\Phi^2(x,a) = \mathbb{V}(R)$ for $R \sim \Phi(\cdot|x,a)$, $x \in \mathcal{X}, a \in \mathcal{A}$. An estimator $\mathbf{A}$ can be considered as a function that maps (μ, π, π_D, D^n) to an estimate of $v_\Phi^{\pi,\mu}$, denoted $\widehat{v}_{\mathbf{A}}(\mu, \pi, \pi_D, D^n)$. Fix $\sigma^2 := (\sigma^2(x,a))_{x\in\mathcal{X}, a\in\mathcal{A}}$. The minimax optimal risk subject to $\sigma_\Phi^2(x,a) \le \sigma^2(x,a)$ for all $x \in \mathcal{X}, a \in \mathcal{A}$ is defined by $R_n^*(\mu, \pi, \pi_D, \sigma^2) := \inf_{\mathbf{A}} \sup_{\Phi:\sigma_\Phi^2 \le \sigma^2} \mathbb{E}\left[(\widehat{v}_{\mathbf{A}}(\mu, \pi, \pi_D, D^n) - v_\Phi^{\pi,\mu})^2\right]$.

The main observation is that the estimation problem for the contextual case can actually be reduced to the contextless bandit case by treating the context-action pairs as "actions" belonging to the product space $\mathcal{X} \times \mathcal{A}$. For any policy π, by slightly abusing notation, let $(\mu \otimes \pi)(x,a) = \mu(x)\pi(a|x)$ be the joint distribution of (X, A) when $X \sim \mu(\cdot)$, $A \sim \pi(\cdot|X)$. (We also let $\mu \otimes \pi(B) = \sum_{(x,a)\in B}(\mu \otimes \pi)(x,a)$ for any $B \subset \mathcal{X} \times \mathcal{A}$.) This way, we can map any contextual policy evaluation problem defined by μ,π_D, π, Φ and a sample size n into a contextless policy evaluation problem defined by $\mu \otimes \pi_D$, $\mu \otimes \pi$, Φ with action set $\mathcal{X} \times \mathcal{A}$.

Thus all results reported in previous sections apply to this contextual bandit setting (see Theorem 5 of Appendix B).

3.2 Markov Decision Processes

The results in Section 2 can also be naturally extended to fixed-horizon, finite Markov decision processes (MDPs). An MDP is described by a tuple $M = \langle \mathcal{X}, \mathcal{A}, P, \Phi, \nu, H \rangle$, where $\mathcal{X} = \{1, \dots, N\}$ is the set of states, $\mathcal{A} = \{1, \dots, K\}$ the set of actions, P the transition kernel, $\Phi : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ the reward function, ν the start-state distribution, and H the horizon. A policy $\pi : \mathcal{X} \mapsto [0,1]^K$ maps states to distributions over actions, and we use $\pi(a|x)$ to denote the probability of choosing action a in state x. Given a policy $\pi \in \Pi(\mathcal{X}, \mathcal{A})$, a trajectory of length H, denoted $T = (X, A, R)$ (for $X \in \mathcal{X}^H$, $A \in \mathcal{A}^H$, and $R \in \mathbb{R}^H$), is generated as follows: $X(1) \in \nu(\cdot)$; for $h \in \{1, \dots, H\}$, $A(h) \sim \pi(\cdot|X(h))$, $R(h) \sim \Phi(\cdot|X_{(h)}, A_{(h)})$, and $X(h+1) \sim P(\cdot|X_{(h)}, A_{(h)})$. The policy value is defined by $v_\Phi^\pi := \mathbb{E}_T[\sum_{h=1}^H R(h)]$. For simplicity, we again assume $R_{\max} = 1$. The off-policy evaluation problem is to estimate v_Φ^π from data $D^n = \{T_t\}_{1\le t\le n}$, where each trajectory T_t is independently generated by an exploration policy $\pi_D \in \Pi(\mathcal{X}, \mathcal{A})$. We assume an unknown reward distribution Φ; other quantities including ν, P, H, π, and π_D are all known. The quality of an estimate $\widehat{v}$ is measured by its MSE: $\mathrm{MSE}\left(\widehat{v}\right) := \left[(\widehat{v} - v_\Phi^\pi)^2\right]$.

By considering a length-H trajectory of state-actions as an "action", one can apply all the results from the previous sections to this setting (see Theorem 6 of Appendix C).

Finally, us note that the exponential dependence of the minimax risk on the horizon H is unavoidable. An example is the "combination lock" MDP with N states $\mathcal{X} = \{1, \dots, N\}$ and $K = 2$ actions $\mathcal{A} = \{L, R\}$; the start state is $x_* = 1$. In any state x, action L takes the learner back to the initial state x_*, while action R takes the learner to state $x + 1$. Assume reward is always 0 except in state N where it can be 0 or $R_{\max}$. It is easy to verify that, if there exists constant p_* such that $p_* \le \pi_D(L|x)$ for all x, then it takes exponentially many steps to reach state N from x_* under policy π_D. Consequently, it requires at least exponentially many trajectories to evaluate a policy π that always takes action R, no matter what evaluation algorithm is used.

4 Conclusions

We have studied the fundamental problem of off-policy evaluation. We focused on the case when there is only one state, also known as the problem of off-policy evaluation problem in multi-armed bandits. Despite the simplicity of this problem, we found that it has a surprisingly rich structure. Our paper is best viewed as making the first steps towards exploring this structure.

In particular, we proved new results that reveal the weaknesses of both the simple "importance sampling" (IS) and its more sophisticated weighted (WIS) version. These are confirmed empirically on both synthetic and real-world data. We have not found such results formally proved in the literature, despite the popularity of these estimators. We also considered another estimator, REG, which estimates the mean reward for each action. Our analysis indicates that REG has different qualities. While it less exposed to the magnitude of importance ratios, it may suffer in the low data regime (as compared to an ideal, optimal estimator), which may happen in practice often when the number of actions is large. This was also confirmed by the experiments. In Section 2.5.2 we also proposed an estimator that combines IS/WIS and REG to merge their strengths, but it remains for future work to explore the properties of this estimator. Another interesting problem is to design near-minimax estimators for the case when a bound on the mean reward function is known (the above methods do not use this knowledge even if available).

Finally, in the paper, we focused on the simplest contextless, finite setting, and showed that our results can be extended to more complex settings like contextual bandits and MDPs. Under additional regularity assumptions, the off-policy value estimation problem can be solved more efficiently. Studying such structures and corresponding minimax estimators is another interesting future direction.

Acknowledgements

We sincerely thank Jin Kim and Imed Zitouni for preparing and sharing the search log data used in our experiments. This work was partially supported by grants from Alberta Innovates Technology Futures and NSERC.

References

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis Xavier Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.

V. Heidrich-Meisner and C. Igel. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *ICML*, pages 401–408, 2009.

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

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(6):945–960, 1986.

Il'dar Abdulovich Ibragimov and Rafail Zalmanovich Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer, 1981.

J. Kiefer. *Introduction to statistical inference*. Springer, 1987.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 20*, pages 1096–1103, 2008.

John Langford, Alexander L. Strehl, and Jennifer Wortman. Exploration scavenging. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning*, pages 528–535, 2008.

E.L. Lehmann and G. Casella. *Theory of Point Estimation*. 2 edition, 1998.

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 International Conference on Web Search and Web Data Mining (WSDM-11)*, pages 297–306, 2011.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

A. Y. Ng and M. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *UAI*, pages 406–415, 2000.

Olivier Nicol. *Data-driven Evaluation of Contextual Bandit Algorithms and Applications to Dynamic Recommendation*. PhD thesis, University of Lille, 2015.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-00)*, pages 759–766, 2000.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, editors. *Covariate Shift and Local Learning by Distribution Matching*. MIT Press, 2008.

P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

P. Rosenbaum and D. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1985.

Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*, pages 2217–2225, 2011.

Richard S. Sutton, Hamid R. Maei, and Csaba Szepesvári. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems 22 (NIPS-99)*, pages 1609–1616, 2010.

Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, 2012.