Optimal and Adaptive Off-policy Evaluation in Contextual Bandits

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the problem of off-policy evaluation—estimating the value of a target policy using data collected by another policy—under the contextual bandit model. We establish a minimax lower bound on the mean squared error (MSE), and show that it is matched up to constant factors by the inverse propensity scoring (IPS) estimator. Since in the multi-armed bandit problem the IPS is suboptimal [8], our result highlights the difficulty of the contextual setting with non-degenerate context distributions. We further consider improvements on this minimax MSE bound, given access to a reward model. We show that the existing doubly robust approach, which utilizes such a reward model, may continue to suffer from high variance even when the reward model is perfect. We propose a new estimator called SWITCH which more effectively uses the reward model and achieves a superior bias-variance tradeoff compared with prior work. We prove an upper bound on its MSE and demonstrate its benefits empirically on a diverse collection of datasets, often seeing orders of magnitude improvements over a number of baselines.

1 Introduction

Contextual bandits refer to a learning setting where the learner repeatedly observes a context, takes an action and observes a reward signal for the quality of the chosen action in the observed context. Crucially, there is no information on the quality of all the remaining actions that were not chosen for the context. As an example, consider online movie recommendation where context describes the information about a user, actions are possible movies to recommend and a reward can be whether the user enjoys the recommended movie. The framework applies equally well to several other applications such as online advertising, web search, personalized medical treatment, etc. The goal of the learner is to come up with a policy, that is a scheme for mapping contexts into actions. A common question which arises in such settings is, given a candidate *target policy*, what is the expected reward it obtains? A simple way of answering the question is by letting the policy to choose actions (such as make movie recommendations to users), and compute the reward it obtains. Such *online evaluation*, is typically costly and time consuming since it involves exposing users to an untested experimental policy, and does not easily scale to evaluating the performance of many different policies.

Off-policy evaluation refers to an alternative paradigm for answering the same question. Suppose we have existing logs from the existing system (which might be choosing actions from a very different *logging policy* than the one we seek to evaluate). Can we estimate the expected reward of the *target policy*? This question has been extensively researched in the contextual bandit model (see, e.g., [7, 3, 10, 9] and references therein). In particular, there are several estimators which are unbiased under mild assumptions, such as inverse propensity scoring (IPS) [6], and sharp estimates on their mean squared error (MSE) for policy evaluation are well-known [5].

Submitted to 30th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

While the IPS-style methods make no attempt at all to model the underlying dependence of rewards on contexts and actions, such information is often available. The simplest approach to off-policy evaluation, given such a model, is to simply use the model to predict the reward for the target policy's action on each context. We call this estimator the model-based approach or the direct method (DM). The key drawback of DM is that it can be arbitrarily biased when the model is misspecified. Some approaches, such as the doubly-robust method (DR) [5] (also see the references therein for its origin in statistics and application in causal inference, e.g., [11, 1]), combine the model with an IPS-style unbiased estimation and remain consistent, with sharp estimates of the MSE.

All these works focus on developing specific methods alongside upper bounds on their MSE. Little work, on the other hand, exists on the question of the fundamental statistical hardness of off-policy evaluation and the optimality (or the lack of) of the existing methods. A notable exception is the recent work of Li et al. [8], who study off-policy evaluation in multi-armed bandits—a special case of our setting, without any contexts—and provide a minimax lower bound on the MSE. Their result shows the suboptimality of IPS (and DR) due to an excessive variance of the importance weights. This result is rather intriguing as it hints at one of two possibilities: (i) IPS and variants are also suboptimal for contextual bandit setting and we should develop better estimators, or (ii) the contextual bandit setting has qualitatively different upper and lower bounds that match. In this quest, our paper makes the following key contributions:

1. We provide the first rate-optimal lower bound on the MSE for off-policy evaluation in contextual bandits. In contrast with context-free multi-armed bandits [8], our lower bound matches the MSE upper bound for IPS up to constants, so long as the contexts have a non-degenerate distribution. This highlights the challenges of the contextual setting; even if the reward as a function of contexts and actions has no variance, the lower bound stays non-trivial in contrast with context-free multi-armed bandits.
2. We propose a new class of estimators called the SWITCH estimators, that adaptively interpolate between an available reward model and IPS. We show that SWITCH has MSE no worse than IPS in the worst case, but is robust to large importance weights. We also show that SWITCH can have a drastically smaller variance than alternatives for combining IPS with a reward model, such as DR.
3. We conduct experiments showing that the new estimator performs significantly better than existing approaches on simulated contextual bandit problems using real-life multiclass classification data sets.

Symbols and notations. A context x is a feature vector in $\mathcal{X}$, possibly $\mathbb{R}^d$ or $\{0, 1\}^d$ for some large d. The stationary distribution of contexts is denoted by $\mathcal{D}_x$. Actions, denoted as a, are drawn from a set $\mathcal{A}$. A policy is a function from contexts to distributions over actions, which allows for modeling randomized action choice. We will use $\mu(a|x)$ and $\pi(a|x)$ to denote the *logging* and *target* policies respectively. We use $\rho(x, a)$ to denote the *importance weights* $\pi(a|x)/\mu(a|x)$. Rewards r have a distribution conditioned on x and a denoted by $\mathcal{D}(r|x, a)$. Given a policy π which is a distribution over actions given contexts, we extend it to a joint distribution over triples (x, a, r), where x is drawn according to $\mathcal{D}_x$, action a according to $\pi(a|x)$, and r according to $\mathcal{D}(r|x, a)$. For a policy π, we refer to its expected reward as its value, formally defined as $v^\pi := \mathbb{E}_\pi[r]$.

2 Main results

In this section, we present our main results but leave technical details to the full paper.

2.1 The limit of model-free off-policy evaluation

Off-policy evaluation is intrinsically a statistical estimation problem, where the goal is to estimate v^π. We study this problem in a standard minimax framework: given n iid samples according to a policy μ, what is the smallest mean square error (MSE) *any* estimator can achieve for evaluating a fixed policy π, in the worst case over a particular class of data-generating distributions? Specifically, we generalize the results of Li et al. [8] for multi-armed bandits. We analyze the off-policy evaluation problems given a fixed $\mathcal{D}_x$, μ and π and consider the worst case over a class of reward-generating distributions. Our worst-case bounds are thus allowed to depend on properties of $\mathcal{D}_x$, μ and π.

To formulate our class of reward-generating functions, assume we are given maps $R_{\max} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ and $\sigma : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$. The class of conditional distributions $\mathcal{R}(\sigma, R_{\max})$ is defined as

$$\mathcal{R}(\sigma, R_{\max}) := \Big\{ \mathcal{D}(r|x,a) : 0 \le \mathbb{E}_{\mathcal{D}}[r|x,a] \le R_{\max}(x,a) \text{ and } \operatorname{Var}_{\mathcal{D}}[r|x,a] \le \sigma^2(x,a) \text{ for all } x,a \Big\}.$$

Note that σ^2 and $R_{\max}$ are allowed to change over contexts and actions. Formally, let an estimator be any function $\hat{v} : (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^n \to \mathbb{R}$ that takes n data points collected by μ and outputs an estimate of v^π. The minimax risk of off-policy evaluation over the class $\mathcal{R}(\sigma^2, R_{\max})$ is defined as

$$R_n(\mathcal{D}_x, \pi, \mu, \sigma, R_{\max}) := \inf_{\hat{v}} \sup_{\mathcal{D}(r|x,a) \in \mathcal{R}(\sigma, R_{\max})} \mathbb{E}\left[(\hat{v} - v^\pi)^2 \right]. \tag{1}$$

We prove the following lower bound on the minimax risk:

Theorem 1 (Minimax rate). *For sufficiently large n and $|\mathcal{X}|$*[1]*, we have*

$$R_n(\mathcal{D}_x, \pi, \mu, \sigma, R_{\max}) = \Theta\left(\frac{1}{n} \Big(\mathbb{E}_\mu\Big[\rho^2(x,a)\sigma^2(x,a)\Big] + \mathbb{E}_\mu\Big[\rho^2(x,a) R_{\max}^2(x,a)\Big] \Big) \right),$$

where $\rho(x,a) = \pi(a|x)/\mu(a|x)$.

This result matches the MSE upper bound for the IPS estimator [6], meaning that the estimator is unimprovable beyond constant factors in the worst-case. This is somewhat surprising because IPS was shown to be strictly suboptimal in multi-arm bandits. Specifically, the minimax rate for multi-arm bandits is just $\frac{1}{n}\mathbb{E}_\mu[\rho^2\sigma^2]$, meaning that the second term which depends on $\rho^2 R_{\max}^2$ is the sub-optimality of IPS in that setting, which can be arbitarily large when the rewards are deterministic so that $\sigma \equiv 0$. On the other hand, a non-degenerate context distribution (meaning that there is a large number of unique contexts) leads to a significant variance in policy evaluation even when rewards are deterministic, due to the randomness in the draw of contexts. This randomness is responsible for the gap between contextual and non-contextual lower bounds.

2.2 Adaptive estimation with an auxiliary direct estimator

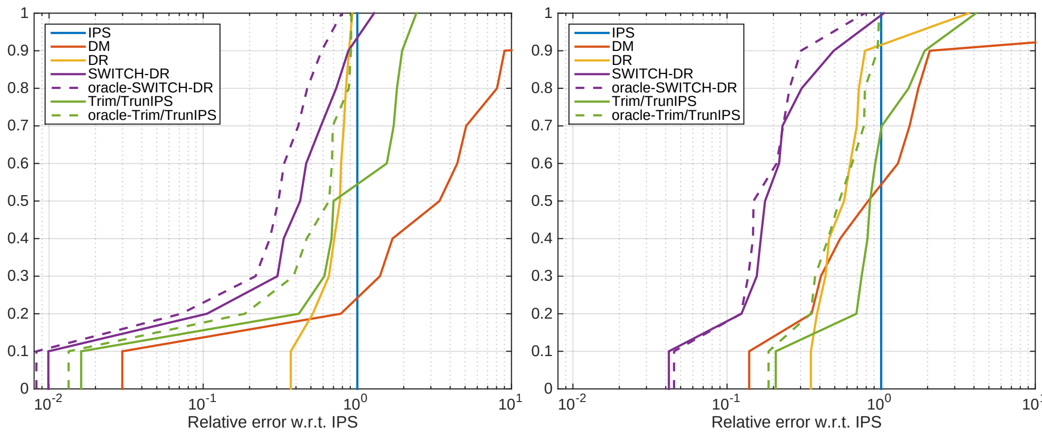
Clearly we cannot do any better than IPS in the worst-case, and yet its upper bound has a dependence on ρ^2, which results in a severe degradation of performance when the importance weights are large. Prior works [4, 5] attempt to address this issue by the development of a doubly robust (DR) estimator which combines IPS with a reward model, when the latter is available. The combination is done in a way that the overall estimator remains unbiased, albeit with a smaller variance when the reward model is good. However, the DR can pay a steep price for being unbiased. Even if we have access to reward model $\hat{r}(x,a)$ such that $\hat{r}(x,a) \equiv \mathbb{E}[r|x,a]$, that is the true conditional expectation, DR suffers from a large variance depending on importance weights whenever the rewards have non-trivial conditional variance. On the other hand, the direct method, which simply estimates v^π using $\sum_{i=1}^n \hat{r}(x_i, \pi(x_i))/n$, has no dependence on the importance weights in this extreme case.

Indeed, this drawback of DR leads to it being sub-optimal, similar to IPS, in the multi-armed bandit setting of Li et al. [8], and naturally leads to the question: *is there a better way to combine a reward model with IPS that achieves a better MSE?* Stated differently, DR is on one extreme end of bias-variance tradeoff by requiring no bias. Could we do better by allowing a small bias and obtaining a significant variance reduction in the process?

Since we are seeking to avoid variance due to excessive importance weights, it is natural to handle the context-action pairs with large importance weights (i.e., those that result in a large variance) separately. To this end, we decompose the value of a policy into two components, based on how large the importance weights are relative to a threshold τ. Under expectation operators, we write ρ instead of a more verbose $\rho(x,a)$:

$$\begin{aligned} v^\pi &= \mathbb{E}_\pi[r] = \mathbb{E}_\pi[r\mathbf{1}(\rho \le \tau)] + \mathbb{E}_\pi[r\mathbf{1}(\rho > \tau)] \\ &= \mathbb{E}_\mu[\rho r \mathbf{1}(\rho \le \tau)] + \mathbb{E}_{\mathcal{D}_x}\Big[\sum_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{D}}[r|x,a]\pi(a|x)\mathbf{1}(\rho(x,a) > \tau) \Big]. \end{aligned}$$

[1] n needs to be larger than a constant that depends only on μ and π, and $|\mathcal{X}| > Cn\log|\mathcal{X}|$, for a C that measures how uniform $\mathcal{D}_x$ is. If $\mathcal{X}$ is a continuous domain, then we only need that $\mathcal{D}_x$ is a probability density.



(a) Deterministic reward (b) Noisy reward

Figure 1: Cumulative distribution of the mean-squared error (MSE) of various estimators over 10 UCI data sets. The results are normalized by dividing MSE by the MSE of IPS. In the left-hand panel, the UCI labels are used as is; in the right-hand panel, additional label noise is added. Curves towards top-left part of the plot are better, as they achieve smaller values of MSE. Methods in dashed lines are "cheating" by choosing the threshold parameter τ to optimize test MSE. As we see, SWITCH-DR dominates all other methods and our empirical tuning of τ is not too far from the optimal possible.

The first part, where the importance weights are small, can still be estimated as before in an unbiased manner. We estimate the second part using our reward model $\hat{r}$. This leads to the following family of SWITCH estimators, parameterized by the threshold τ:

$$\hat{v}_{\text{SWITCH}} = \frac{1}{n}\sum_{i=1}^{n}[r_i \rho_i \mathbf{1}(\rho_i \le \tau)] + \frac{1}{n}\sum_{i=1}^{n}\sum_{a:\rho(x_i,a)>\tau} \hat{r}(x_i,a)\pi(a|x_i).$$

We now analyze the new estimator.

Theorem 2. *Denote $\epsilon(a,x) := \hat{r}(a,x) - \mathbb{E}[r|a,x]$ and assume $\hat{r}$ is pointwise bounded by $R_{\max}$. Then for every $n = 1,2,3,\ldots,$*

$$\mathrm{MSE}(\hat{v}_{\text{SWITCH}}) \le \frac{\mathbb{E}_\pi[R_{\max}^2]}{2n} + \frac{1}{n}\mathbb{E}_\mu\left[\left(2\sigma^2 + \frac{R_{\max}^2}{2}\right)\rho^2\mathbf{1}(\rho \le \tau)\right] + \mathbb{E}_\pi\left[\epsilon \mid \rho > \tau\right]^2 \mathbb{P}_\pi(\rho > \tau)^2,$$

where quantities $R_{\max}$, σ, ρ, and ϵ are functions of random variables x and a.

The first term of the bound is required even when we use DM with a perfect $\hat{r}$. The second term captures the variance of IPS for estimating the part of the problem with importance weights smaller than τ. The third term captures the bias of $\hat{r}$. As τ moves from 0 to ∞, our bound adaptively interpolates between DM and IPS, using DM for the part that causes the high variance for IPS.

The policy value in the region where the importance weights are small can be estimated using any unbiased approach rather than just IPS. For instance, we can use the DR, giving rise to the estimator, which we denote SWITCH-DR.

Automatic parameter tuning: We propose to choose parameter τ by optimizing an empirical estimate of the bias-variance tradeoff. The bias of our estimator is captured by the final term involving ϵ in Theorem 2, where we conservatively bound ϵ by $R_{\max}$. For the variance, we use an empirical estimate arising from the fact that our estimator can be written as a sum of n i.i.d. terms. This conservative procedure ensures that SWITCH (or SWITCH-DR) will perform at least as well as IPS (or DR). It thus remains minimax in the worst case and robust to large ρ. The procedure is related to the MAGIC estimator [12], but uses different estimates of bias and variance. In more detailed experimental results (not included in this abstract), we found that our choice of τ outperforms the MAGIC estimator quite substantially in the contextual bandit setting.

In Figure 1, we compare our estimator with several existing approaches using a similar protocol as in the prior work [4]. The methods are compared by plotting the cumulative distribution of their mean-squared error (MSE) over 10 UCI data sets (converted into a contextual-bandit format). Methods that achieve smaller values of MSE are towards the top-left corner of the plot. Since SWITCH-DR dominated SWITCH in our experiments, we only show SWITCH-DR. In addition to IPS, DM, DR, and SWITCH-DR, we also consider two additional variants of IPS, where importance weights are either capped at τ and renormalized [see, e.g., 2], or the terms with weights larger than τ are removed altogether as described in Bottou et al. [3]. We use a conservative setting of τ based on the error upper bounds of these two methods, and on each data set we select the better of the two, under the name Trim/TrunIPS. Finally, since SWITCH-DR and Trim/TrunIPS depend on the parameter τ, which we tune in a specific manner, we also show their performance for the optimal choice of τ—this serves as a ceiling on their performance, under a possibly smarter tuning of τ.

As we see, SWITCH-DR dominates all other methods and our empirical tuning of τ is not too far from the optimal possible. The advantage of SWITCH-DR is even stronger in the noisy-reward setting, where we add label noise to UCI data.

3 Conclusion

In this paper we carried out minimax analysis of off-policy evaluation in contextual bandits and showed that IPS is optimal in the worst-case. This result highlights the need for using side information, potentially provided by modeling the reward directly, especially when importance weights are too large. Given this observation, we proposed a new class of estimators called SWITCH that can be used to combine any importance sampling estimators, including IPS and DR, with DM. The estimator involves adaptively switching to DM when the importance weights are large and switching to either IPS or DR when the importance weights are small. We showed that the new estimator has favorable theoretical properties and also works well on real-world data.

References

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

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

[3] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[4] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, 2011.

[5] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

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

[7] 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.

[8] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *AISTATS*, 2015.

[9] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

[10] H Lock Oh and Frederick J Scheuren. Weighting adjustment for unit nonresponse. *Incomplete data in sample surveys*, 2:143–184, 1983.

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

[12] Philip S Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *ICML'16*, 2016.