Optimal and Adaptive Off-policy Evaluation in Contextual Bandits

Yu-Xiang Wang [1] **Alekh Agarwal** [2] **Miroslav Dudík** [2]

Abstract

We study the off-policy evaluation problem—estimating the value of a target policy using data collected by another policy—under the contextual bandit model. We consider the general (agnostic) setting without access to a consistent model of rewards and establish a minimax lower bound on the mean squared error (MSE). The bound is matched up to constants by the inverse propensity scoring (IPS) and doubly robust (DR) estimators. This highlights the difficulty of the agnostic contextual setting, in contrast with multi-armed bandits and contextual bandits with access to a consistent reward model, where IPS is suboptimal. We then propose the SWITCH estimator, which can use an existing reward model (not necessarily consistent) to achieve a better bias-variance tradeoff than IPS and DR. We prove an upper bound on its MSE and demonstrate its benefits empirically on a diverse collection of data sets, often outperforming prior work by orders of magnitude.

1. Introduction

Contextual bandits refer to a learning setting where the learner repeatedly observes a context, takes an action and observes a reward for the chosen action in the observed context, *but no feedback on any other action*. An example is movie recommendation, where the context describes a user, actions are candidate movies and the reward measures if the user enjoys the recommended movie. The learner produces a policy, meaning a mapping from contexts to actions. A common question in such settings is, given a *target policy*, what is its expected reward? By letting the policy choose actions (e.g., recommend movies to users), we can compute its reward. Such *online evaluation* is typically costly since it exposes users to an untested experimental policy, and does not scale to evaluating many different target policies.

Off-policy evaluation is an alternative paradigm for the same question. Given logs from the existing system, which might be choosing actions according to a very different *logging policy* than the one we seek to evaluate, can we estimate the expected reward of the *target policy*? There are three classes of approaches to address this question: the *direct method* (DM), also known as regression adjustment, *inverse propensity scoring* (IPS) (Horvitz & Thompson, 1952) and *doubly robust* (DR) estimators (Robins & Rotnitzky, 1995; Bang & Robins, 2005; Dudík et al., 2011; 2014).

Our first goal in this paper is to study the optimality of these three classes of approaches (or lack thereof), and more fundamentally, to quantify the statistical hardness of off-policy evaluation. This problem was previously studied for multi-armed bandits (Li et al., 2015) and is related to a large body of work on asymptotically optimal estimators of average treatment effects (ATE) (Hahn, 1998; Hirano et al., 2003; Imbens et al., 2007; Rothe, 2016), which can be viewed as a special case of off-policy evaluation. In both settings, a major underlying assumption is that rewards can be consistently estimated from the features (i.e., covariates) describing contexts and actions, either via a parametric model or non-parametrically. Under such consistency assumptions, it has been shown that DM and/or DR are optimal (Imbens et al., 2007; Li et al., 2015; Rothe, 2016),[1] whereas standard IPS is not (Hahn, 1998; Li et al., 2015), but it becomes (asymptotically) optimal when the true propensity scores are replaced by suitable estimates (Hirano et al., 2003).

Unfortunately, consistency of a reward model can be difficult to achieve in practice. Parametric models tend to suffer from a large bias (see, e.g., the empirical evaluation of Dudík et al., 2011) and non-parametric models are limited to small dimensions, otherwise non-asymptotic terms become too large (see, e.g., the analysis of non-parametric regression by Bertin et al., 2004). Therefore, here we ask: *What can be said about hardness of policy evaluation in the absence of reward-model consistency?*

In this pursuit, we provide the first rate-optimal lower bound on the mean-squared error (MSE) for off-policy evaluation

[1]Carnegie Mellon University, Pittsburgh, PA [2]Microsoft Research, New York, NY. Correspondence to: Yu-Xiang Wang <yuxiangw@cs.cmu.edu>, Alekh Agarwal <alekha@microsoft.com>, Miroslav Dudík <mdudik@microsoft.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]The precise assumptions vary for each estimator, and are somewhat weaker for DR than for DM.

in contextual bandits without consistency assumptions. Our lower bound matches the upper bounds of IPS and DR up to constants, when given a non-degenerate context distributions. This result is in contrast with the suboptimality of IPS under previously studied consistency assumptions, which implies that the two settings are qualitatively different.

Whereas IPS and DR are both minimax optimal, our experiments (similar to prior work) show that IPS is readily outperformed by DR, even when using a simple parametric regression model that is not asymptotically consistent. We attribute this to a lower variance of the DR estimator. We also empirically observe that while DR is generally highly competitive, it is sometimes substantially outperformed by DM. We therefore ask whether it is possible to achieve an even better bias-variance tradeoff than DR. We answer affirmatively and propose a new class of estimators, called the SWITCH estimators, that *adaptively interpolate* between DM and DR (or IPS). We show that SWITCH has MSE no worse than DR (or IPS) in the worst case, but is robust to large importance weights and can achieve a substantially smaller variance than DR or IPS.

We empirically evaluate the SWITCH estimators against a number of strong baselines from prior work, using a previously used experimental setup to simulate contextual bandit problems on real-world multiclass classification data. The results affirm the superior bias-variance tradeoff of SWITCH estimators, with substantial improvements across a number of problems.

In summary, the first part of our paper initiates the study of optimal estimators in a finite-sample setting and without making strong modeling assumptions, while the second part shows how to practically exploit domain knowledge by building better estimators.

2. Setup

In contextual bandit problems, the learning agent observes a context x, takes an action a and observes a scalar reward r for the action chosen in the context. Here the context x is a feature vector from some domain $\mathcal{X} \subseteq \mathbb{R}^d$, drawn according to a distribution λ. Actions a are drawn from a finite set $\mathcal{A}$. Rewards r have a distribution conditioned on x and a denoted by $D(r \mid x, a)$. The decision rule of the agent is called a policy, which maps contexts to distributions over actions, allowing for randomization in the action choice. We write $\mu(a \mid x)$ and $\pi(a \mid x)$ to denote the *logging* and *target* policies respectively. Given a policy π, we extend it to a joint distribution over (x, a, r), where $x \sim \lambda$, action $a \sim \pi(a \mid x)$, and $r \sim D(r \mid x, a)$. With this notation, given n i.i.d. samples $(x_i, a_i, r_i) \sim \mu$, we wish to compute the *value of* π:

$$v^\pi = \mathbb{E}_\pi[r] = \mathbb{E}_{x\sim\lambda}\mathbb{E}_{a\sim\pi(\cdot|x)}\mathbb{E}_{r\sim D(\cdot|a,x)}[r]. \quad (1)$$

In order to correct for the mismatch in the action distributions under μ and π, it is typical to use *importance weights*, defined as $\rho(x, a) \coloneqq \pi(a|x)/\mu(a|x)$. For consistent estimation, it is standard to assume that $\rho(x, a) \neq \infty$, corresponding to *absolute continuity* of π with respect to μ, meaning that whenever $\pi(a|x) > 0$, then also $\mu(a|x) > 0$. We make this assumption throughout the paper. In the remainder of the setup we present three common estimators of v^π.

The first is the inverse propensity scoring (IPS) estimator (Horvitz & Thompson, 1952), defined as

$$\hat{v}^\pi_{\text{IPS}} = \sum_{i=1}^n \rho(x_i, a_i) r_i. \quad (2)$$

IPS is unbiased and makes no assumptions about how rewards might depend on contexts and actions. When such information is available, it is natural to posit a parametric or non-parametric model of $\mathbb{E}[r \mid x, a]$ and fit it on the logged data to obtain a reward estimator $\hat{r}(x, a)$. Policy evaluation can now simply be performed by scoring π according to $\hat{r}$ as

$$\hat{v}^\pi_{\text{DM}} = \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}} \pi(a \mid x_i)\hat{r}(x_i, a), \quad (3)$$

where the DM stands for *direct method* (Dudík et al., 2011), also known as *regression adjustment* or *imputation* (Rothe, 2016). IPS can have a large variance when the target and logging policies differ substantially, and parametric variants of DM can be inconsistent, leading to a large bias. Therefore, both in theory and practice, it is beneficial to combine the approaches into a *doubly robust* estimator (Cassel et al., 1976; Robins & Rotnitzky, 1995; Dudík et al., 2011), such as the following variant,

$$\hat{v}^\pi_{\text{DR}} = \frac{1}{n}\sum_{i=1}^n \Big[\rho(x_i, a_i)\big(r_i - \hat{r}(x_i, a_i)\big) + \sum_{a\in\mathcal{A}} \pi(a \mid x_i)\hat{r}(x_i, a)\Big]. \quad (4)$$

Note that IPS is a special case of DR with $\hat{r} \equiv 0$. In the sequel, we mostly focus on IPS and DR, and then suggest how to improve them by further interpolating with DM.

3. Limits of Off-policy Evaluation

In this section, we study the off-policy evaluation problem in a minimax setup. After setting up the framework, we present our lower bound and the matching upper bounds for IPS and DR under appropriate conditions.

While minimax optimality is standard in statistical estimations, it is not the only notion of optimality. An alternative framework is that of asymptotic optimality, which establishes Cramer-Rao style bounds on the asymptotic variance of estimators. We use the minimax framework, because it

is the most amenable to finite-sample lower bounds, and is complementary to previous asymptotic results, as we discuss after presenting our main results.

3.1. Minimax Framework

Off-policy evaluation is a statistical estimation problem, where the goal is to estimate v^π given n i.i.d. samples generated according to a policy μ. We study this problem in a standard minimax framework and seek to answer the following question. What is the smallest MSE that *any* estimator can achieve in the worst case over a large class of contextual bandit problems? As is usual in the minimax setting, we want the class of problems to be rich enough so that the estimation problem is not trivial, and to be small enough so that the lower bounds are not driven by complete pathologies. In our problem, we fix λ, μ and π, and only take worst case over a class of reward distributions. This allows the upper and lower bounds to depend on λ, μ and π, highlighting how these ground-truth parameters influence the problem difficulty. The family of reward distributions $D(r \,|\, x, a)$ that we study is a natural generalization of the class studied by Li et al. (2015) for multi-armed bandits. We assume we are given maps $R_{\max} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ and $\sigma : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$, and define the class of reward distributions $\mathcal{R}(\sigma, R_{\max})$ as[2]

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

Note that σ and $R_{\max}$ are allowed to change over contexts and actions. Formally, an estimator is 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 $\mathcal{R}(\sigma, R_{\max})$, denoted by $R_n(\pi; \lambda, \mu, \sigma, R_{\max})$, is defined as

$$\inf_{\hat{v}} \sup_{D(r|x,a) \in \mathcal{R}(\sigma, R_{\max})} \mathbb{E}\left[(\hat{v} - v^\pi)^2 \right]. \tag{5}$$

Recall that the expectation is taken over the n samples drawn from μ, along with any randomness in the estimator. The main goal of this section is to obtain a lower bound on the minimax risk. To state our bound, recall that $\rho(x,a) = \pi(a \mid x)/\mu(a \mid x) < \infty$ is an importance weight at (x,a). We make the following technical assumption on our problem instances, described by tuples of the form $(\pi, \lambda, \mu, \sigma, R_{\max})$:

Assumption 1. *There exists $\epsilon > 0$ such that $\mathbb{E}_\mu\left[(\rho\sigma)^{2+\epsilon}\right]$ and $\mathbb{E}_\mu\left[(\rho R_{\max})^{2+\epsilon}\right]$ are finite.*

This assumption is only a slight strengthening of the assumption that $\mathbb{E}_\mu[(\rho\sigma)^2]$ and $\mathbb{E}_\mu[(\rho R_{\max})^2]$ be finite, which is required for consistency of IPS (see, e.g., Dudík et al., 2014). Our assumption holds for instance when the context space is finite, because then both ρ and $R_{\max}$ are bounded.

[2] Technically, the inequalities in the definition of $\mathcal{R}(\sigma, R_{\max})$ need to hold almost surely with $x \sim \lambda$ and $a \sim \mu(\cdot \mid x)$.

3.2. Minimax Lower Bound for Off-policy Evaluation

With the minimax setup in place, we now give our main lower bound on the minimax risk for off-policy evaluation and discuss its consequences. Our bound depends on a parameter $\gamma \in [0,1]$ and a derived indicator random variable $\xi_\gamma(x,a) := \mathbf{1}(\mu(x,a) \le \gamma)$, which separates out the pairs (x,a) that appear "frequently" under μ.[3] As we will see, the "frequent" pairs (x,a) (where $\xi_\gamma = 0$) correspond to the intrinsically realizable part of the problem, where consistent reward models can be constructed. The "infrequent" pairs (where $\xi_\gamma = 1$) constitute the part that is non-realizable in the worst-case. When $\mathcal{X} \subseteq \mathbb{R}^d$ and λ is continuous with respect to the Lebesgue measure, then $\xi_\gamma(x,a) = 1$ for all $\gamma \in [0,1]$, so the problem is non-realizable everywhere in the worst-case. Our result uses the following problem-dependent constant (defined with the convention $0/0 = 0$):

$$C_\gamma := 2^{2+\epsilon} \max\left\{ \frac{\mathbb{E}_\mu[(\rho\sigma)^{2+\epsilon}]^2}{\mathbb{E}_\mu[(\rho\sigma)^2]^{2+\epsilon}}, \frac{\mathbb{E}_\mu[\xi_\gamma(\rho R_{\max})^{2+\epsilon}]^2}{\mathbb{E}_\mu[\xi_\gamma(\rho R_{\max})^2]^{2+\epsilon}} \right\}.$$

Theorem 1. *Assume that a problem instance satisfies Assumption 1 with some $\epsilon > 0$. Then for any $\gamma \in [0,1]$ and any $n \ge \max\{16 C_\gamma^{1/\epsilon}, 2 C_\gamma^{2/\epsilon} \mathbb{E}_\mu[\sigma^2/R_{\max}^2]\}$, the minimax risk $R_n(\pi; \lambda, \mu, \sigma, R_{\max})$ satisfies the lower bound*

$$\frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\right] + \mathbb{E}_\mu\left[\xi_\gamma \rho^2 R_{\max}^2\right]\Big(1 - 350 n \gamma \log(5/\gamma)\Big)}{700 n}.$$

The bound holds for every $\gamma \in [0,1]$, and we can take the maximum over γ. In particular, we get the following simple corollary under continuous context distributions.

Corollary 1. *Under conditions of Theorem 1, assume further that λ has a density relative to Lebesgue measure. Then*

$$R_n(\pi; \lambda, \mu, \sigma, R_{\max}) \ge \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\right] + \mathbb{E}_\mu\left[\rho^2 R_{\max}^2\right]}{700 n}.$$

If λ is a mixture of a density and point masses, then $\gamma = 0$ will exclude the point masses from the second term of the lower bound. In general, choosing $\gamma = \mathcal{O}\big(1/(n \log n)\big)$ excludes the contexts likely to appear multiple times, and ensures that the second term in Theorem 1 remains non-trivial (when $\mu(x,a) \le \gamma$ with positive probability).

Before sketching the proof of Theorem 1, we discuss its preconditions and implications.

[3] Formally, $\mu(x,a)$ corresponds to $\mu(\{(x,a)\})$, i.e., the measure under μ of the set $\{(x,a)\}$. For example, when λ is a continuous distribution then $\mu(x,a) = 0$ everywhere.

Preconditions of the theorem: The theorem assumes the existence of a (problem-dependent) constant C_γ which depends on the constant γ and various moments of the importance-weighted rewards. When $R_{\max}$ and σ are bounded (a common situation), C_γ measures how heavy-tailed the importance weights are. Note that $C_\gamma < \infty$ for all $\gamma \in [0, 1]$ whenever Assumption 1 holds, and so the condition on n in Theorem 1 is eventually satisfied as long as the random variable $\sigma / R_{\max}$ has a bounded second moment. This is quite reasonable since in typical applications the *a priori* bound on expected rewards is on the same order or larger than the *a priori* bound on the reward noise. For the remainder of the discussion, we assume that n is appropriately large so the preconditions of the theorem hold.

Comparison with upper bounds: The setting of Corollary 1 is typical of many contextual bandit applications. In this setting both IPS and DR achieve the minimax risk up to a multiplicative constant. Let $r^*(x, a) := \mathbb{E}[r \mid x, a]$. Recall that DR is using an estimator $\hat{r}(x, a)$ of $r^*(x, a)$, and IPS can be viewed as a special case of DR with $\hat{r} \equiv 0$. By Lemma 3.3(i) of Dudík et al. (2014), the MSE of DR is

$$\begin{aligned}\mathbb{E}[(\hat{v}^\pi_{\mathrm{DR}} - v^\pi)^2] \\ = \frac{1}{n}\Big(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathrm{Var}_{x\sim D}\mathbb{E}_{a\sim\mu(\cdot|x)}[\rho r^*] \\ + \mathbb{E}_{x\sim D}\mathrm{Var}_{a\sim\mu(\cdot|x)}[\rho(\hat{r} - r^*)]\Big). \qquad (6)\end{aligned}$$

Note that $0 \le r^* \le R_{\max}$, so if the estimator $\hat{r}$ also satisfies $0 \le \hat{r} \le R_{\max}$, we obtain that the risk of DR (with IPS as a special case) is at most $\mathcal{O}\big(\frac{1}{n}(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\rho^2 R^2_{\max}])\big)$. This means that IPS and DR are unimprovable, in the worst case, beyond constant factors. Another implication is that the lower bound of Corollary 1 is sharp, and the minimax risk is precisely $\Theta\big(\frac{1}{n}(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\rho^2 R^2_{\max}])\big)$. While IPS and DR exhibit the same minimax rates, Eq. (6) also immediately shows that DR will be better than IPS whenever $\hat{r}$ is even moderately good (better than $\hat{r} \equiv 0$).

Comparison with asymptotic optimality results: As discussed in Section 1, previous work on optimal off-policy evaluation, specifically the average treatment estimation, assumes that it is possible to consistently estimate $r^*(x, a) = \mathbb{E}[r \mid x, a]$. Under such an assumption it is possible to (asymptotically) match the risk of DR with the perfect reward estimator $\hat{r} \equiv r^\star$, and this is the best possible asymptotic risk (Hahn, 1998). This optimal risk is $\frac{1}{n}\big(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathrm{Var}_{x\sim D}\mathbb{E}_\pi[r^* \mid x]\big)$, corresponding to the first two terms of Eq. (6), with no dependence on $R_{\max}$. Several estimators achieve this risk, *including the multiplicative constant*, under various consistency assumptions (Hahn, 1998; Hirano et al., 2003; Imbens et al., 2007; Rothe, 2016). Note that this is strictly below our lower bound for continuous λ. That is, consistency assumptions yield a better asymptotic risk than possible in the agnostic setting. The gap in constants between our upper and lower bounds is due to the finite-sample setting, where lower-order terms cannot be ignored, but have to be explicitly bounded. Indeed, apart from the result of Li et al. (2015), discussed below, ours is the first finite-sample lower bound for off-policy evaluation.

Comparison with multi-armed bandits: For multi-armed bandits, equivalent to contextual bandits with a single context, Li et al. (2015) show that the minimax risk equals $\Theta(\mathbb{E}_\mu[\rho^2\sigma^2]/n)$ and is achieved, e.g., by DM, whereas IPS is suboptimal. They also obtain a similar result for contextual bandits, assuming that each context appears with a large-enough probability to estimate its associated rewards by empirical averages (amounting to realizability). While we obtain a larger lower bound, this is not a contradiction, because we allow arbitrarily small probabilities of individual contexts and even continuous distributions, where the probability of any single context is zero.

On a closer inspection, the first term of our bound in Theorem 1 coincides with the lower bound of Li et al. (2015) (up to constants). The second term (optimized over γ) is non-zero only if there are contexts with small probabilities relative to the number of samples. In multi-armed bandits, we recover the bound of Li et al. (2015). When the context distribution is continuous, or the probability of seeing repeated contexts in a data set of size n is small, we get the minimax optimality of IPS.

One of our key contributions is to highlight this *agnostic contextual* regime where IPS is optimal. In the *non-contextual* regime, where each context appears frequently, the rewards for each context-action pair can be consistently estimated by empirical averages. Similarly, the asymptotic results discussed earlier focus on a setting where rewards can be consistently estimated thanks to parametric assumptions or smoothness (for non-parametric estimation), with the goal of asymptotic efficiency. Our work complements that line of research. In many practical situations, we wish to evaluate policies on high-dimensional context spaces, where the consistent estimation of rewards is not a feasible option. In other words, the agnostic contextual regime dominates.

The distinction between the contextual and non-contextual regime is also present in our proof, which combines a non-contextual lower bound due to the reward noise, similar to the analysis of Li et al. (2015), and an additional bound arising for non-degenerate context distributions. This latter result is a key technical novelty of our paper.

Proof sketch: We only sketch some of the main ideas here and defer the full proof to Appendix A. For simplicity, we discuss the case where λ is a continuous distribution. We consider two separate problem instances corresponding to the two terms in Theorem 1. The first part is relatively straightforward and reduces the problem to Gaussian mean

estimation. We focus on the second part which depends on $R_{\max}$. Our construction defines a prior over the reward distributions, $D(r \mid x, a)$. Given any (x, a), a problem instance is given by

$$\mathbb{E}[r \mid x, a] = \eta(x, a) = \begin{cases} R_{\max}(x, a) & \text{w.p. } \theta(x, a), \\ 0 & \text{w.p. } 1 - \theta(x, a), \end{cases}$$

for $\theta(x, a)$ to be appropriately chosen. Once η is drawn, we consider a problem instance defined by η where the rewards are deterministic and the only randomness is in the contexts. In order to lower bound the MSE across all problems, it suffices to lower bound $\mathbb{E}_\theta[\mathrm{MSE}_\eta(\hat{v})]$. That is, we can compute the MSE of an estimator for each individual η, and take expectation of the MSEs under the prior prescribed by θ. If the expectation is large, we know that there is a problem instance where the estimator incurs a large MSE.

A key insight in our proof is that this expectation can be lower bounded by $\mathrm{MSE}_{\mathbb{E}_\theta[\eta(x,a)]}(\hat{v})$, corresponding to the MSE of a single problem instance with the actual *rewards*, rather than $\eta(x, a)$, drawn according to θ and with the mean reward function $\mathbb{E}_\theta[\eta(x, a)]$. This is powerful, since this new problem instance has stochastic rewards, just like Gaussian mean estimation, and is amenable to standard techniques. The lower bound by $\mathrm{MSE}_{\mathbb{E}_\theta[\eta(x,a)]}(\hat{v})$ is only valid when the context distribution λ is rich enough (e.g., continuous). In that case, our reasoning shows that with enough randomness in the context distribution, a problem with even a deterministic reward function is extremely challenging.

4. Incorporating Reward Models

As discussed in the previous section, it is generally possible to beat our minimax bound when consistent reward models exist. We also argued that even in the absence of a consistent model, when DR and IPS both achieve optimal risk rates, the performance of DR on finite samples will be better than IPS as long as the reward model is even moderately good (see Eq. 6). However, under a large reward noise σ, DR may still suffer from high variance when the importance weights are large, even when given a perfect reward model. In this section, we derive a class of estimators that leverage reward models to directly address this source of high variance, in a manner very different from the standard DR approach.

4.1. The SWITCH Estimators

Our starting point is the observation that insistence on maintaining unbiasedness puts the DR estimator at one extreme end of the bias-variance tradeoff. Prior works have considered ideas such as truncating the rewards or importance weights when the importance weights are large (see, e.g., Bottou et al. 2013), which can dramatically reduce the variance at the cost of a little bias. We take the intuition a step further and propose to estimate the rewards for actions by two distinct strategies, based on whether they have a large or a small importance weight in a given context. When importance weights are small, we continue to use our favorite unbiased estimators, but switch to directly applying the (potentially biased) reward model on actions with large importance weights. Here, "small" and "large" are defined via a *threshold parameter* τ. Varying this parameter between 0 and ∞ leads to a family of estimators which we call the SWITCH estimators as they switch between an agnostic approach (such as DR or IPS) and the direct method.

We now formalize this intuition, and begin by decomposing v^π according to importance weights:

$$\begin{aligned}\mathbb{E}_\pi[r] &= \mathbb{E}_\pi[r\mathbf{1}(\rho \le \tau)] + \mathbb{E}_\pi[r\mathbf{1}(\rho > \tau)] \\ &= \mathbb{E}_\mu[\rho r\mathbf{1}(\rho \le \tau)] \\ &\quad + \mathbb{E}_{x\sim\lambda}\Big[\sum_{a\in\mathcal{A}} \mathbb{E}_D[r \mid x, a]\, \pi(a \mid x)\, \mathbf{1}(\rho(x, a) > \tau)\Big].\end{aligned}$$

Conceptually, we split our problem into two. The first problem always has small importance weights, so we can use unbiased estimators such as IPS or DR. The second problem, where importance weights are large, is addressed by DM. Writing this out leads to the following estimator:

$$\begin{aligned}\hat{v}_{\text{SWITCH}} &= \frac{1}{n}\sum_{i=1}^{n} [r_i \rho_i \mathbf{1}(\rho_i \le \tau)] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \hat{r}(x_i, a)\pi(a \mid x_i)\mathbf{1}(\rho(x_i, a) > \tau). \quad (7)\end{aligned}$$

Note that the above estimator specifically uses IPS on the first part of the problem. When DR is used instead of IPS, we refer to the resulting estimator as SWITCH-DR. The reward model used within the DR part of the SWITCH-DR estimator can be the same or different from the reward model used to impute rewards in the second part. We next present a bound on the MSE of the SWITCH estimator using IPS. A similar bound holds for SWITCH-DR.

Theorem 2. *Let $\epsilon(a, x) := \hat{r}(a, x) - \mathbb{E}[r|a, x]$ be the bias of $\hat{r}$ and assume $\hat{r}(x, a) \in [0, R_{\max}(x, a)]$ almost surely. Then for $\hat{v}_{\text{SWITCH}}$, with $\tau > 0$, the MSE is at most*

$$\begin{aligned}&\frac{2}{n}\Big\{\mathbb{E}_\mu\big[(\sigma^2 + R_{\max}^2)\,\rho^2\mathbf{1}(\rho \le \tau)\big] + \mathbb{E}_\pi\big[R_{\max}^2\mathbf{1}(\rho > \tau)\big]\Big\} \\ &\qquad + \mathbb{E}_\pi\big[\epsilon\mathbf{1}(\rho > \tau)\big]^2.\end{aligned}$$

The proposed estimator interpolates between DM and IPS. For $\tau = 0$, SWITCH coincides with DM, while $\tau \to \infty$ yields IPS. Consequently, SWITCH estimator is minimax optimal when τ is appropriately chosen. However, unlike IPS and DR, the SWITCH and SWITCH-DR estimators are by design more robust to large (or heavy-tailed) importance weights. Several estimators related to SWITCH have been previously studied:

1. Bottou et al. (2013) consider a special case of SWITCH with $\hat{r} \equiv 0$, meaning that all the actions with large importance weights are eliminated from IPS. We refer to this method as *Trimmed IPS*.
2. Thomas & Brunskill (2016) study an estimator similar to SWITCH in the more general setting of reinforcement learning. Their *MAGIC* estimator can be seen as using several candidate thresholds τ and then evaluating the policy by a weighted sum of the estimators corresponding to each τ. Similar to our approach of automatically determining τ, they determine the weighting of estimators via optimization (as we discuss below).

4.2. Automatic Parameter Tuning

So far we have discussed the properties of the SWITCH estimators assuming that the parameter τ is chosen well. Our goal is to obtain the best of IPS and DM, but a poor choice of τ might easily give us the worst of the two estimators. Therefore, a method for selecting τ plays an essential role. A natural criterion would be to pick τ that minimizes the MSE of the resulting estimator. Since we do not know the precise MSE (as v^π is unknown), an alternative is to minimize its data-dependent estimate. Recalling that the MSE can be written as the sum of variance and squared bias, we estimate and bound the terms individually.

Recall that we are working with a data set (x_i, a_i, r_i) and $\rho_i \coloneqq \pi(a_i \mid x_i)/\mu(a_i \mid x_i)$. Using this data, it is straightforward to estimate the variance of the SWITCH estimator. Let $Y_i(\tau)$ denote the estimated value that π obtains on the data point x_i according to the SWITCH estimator with the threshold τ, that is

$$Y_i(\tau) \coloneqq r_i \rho_i \mathbf{1}(\rho_i \leq \tau) + \sum_{a \in \mathcal{A}} \hat{r}(x_i, a) \pi(a|x_i) \mathbf{1}(\rho(x_i, a) > \tau),$$

and $\bar{Y}(\tau) \coloneqq \frac{1}{n} \sum_{i=1}^n Y_i(\tau)$. Since $\hat{v}_{\text{SWITCH}} = \bar{Y}(\tau)$ and the x_i are i.i.d., the variance can be estimated as

$$\mathrm{Var}(\bar{Y}(\tau)) \approx \frac{1}{n^2} \sum_{i=1}^n (Y_i(\tau) - \bar{Y}(\tau))^2 =: \widehat{\mathrm{Var}}_\tau, \qquad (8)$$

where the approximation above is clearly consistent since the random variables Y_i are appropriately bounded as long as the rewards are bounded, because the importance weights are capped at the threshold τ.

Next we turn to the bias term. For understanding bias, we look at the MSE bound in Theorem 2, and observe that the last term in that theorem is precisely the squared bias. Rather than using a direct bias estimate, which would require knowledge of the error in $\hat{r}$, we will upper bound this term. We assume that the function $R_{\max}(x, a)$ is known. This is not limiting since in most practical applications an *a priori* bound on the rewards is known. Then we can upper bound the squared bias as

$$\mathbb{E}_\pi \big[\epsilon \mathbf{1}(\rho > \tau)\big]^2 \leq \mathbb{E}_\pi \big[R_{\max} \mathbf{1}(\rho > \tau)\big]^2.$$

Replacing the expectation with an average, we obtain

$$\widehat{\mathrm{Bias}}_\tau^2 \coloneqq \left[\frac{1}{n} \sum_{i=1}^n \mathbb{E}_\pi \big[R_{\max} \mathbf{1}(\rho > \tau) \,\big|\, x_i\big]\right]^2.$$

With these estimates, we pick the threshold $\widehat{\tau}$ by optimizing the sum of estimated variance and the upper bound on bias,

$$\widehat{\tau} \coloneqq \operatorname*{argmin}_\tau \widehat{\mathrm{Var}}_\tau + \widehat{\mathrm{Bias}}_\tau^2. \qquad (9)$$

Our upper bound on the bias is rather conservative, as it upper bounds the error of DM at the largest possible value for every data point. This has the effect of favoring the use of the unbiased part in SWITCH whenever possible, unless the variance would overwhelm even an arbitrarily biased DM. This conservative choice, however, immediately implies the minimax optimality of the SWITCH estimator using $\widehat{\tau}$, because the incurred bias is no more than our upper bound, and it is incurred only when the minimax optimal IPS estimator would be suffering an even larger variance.

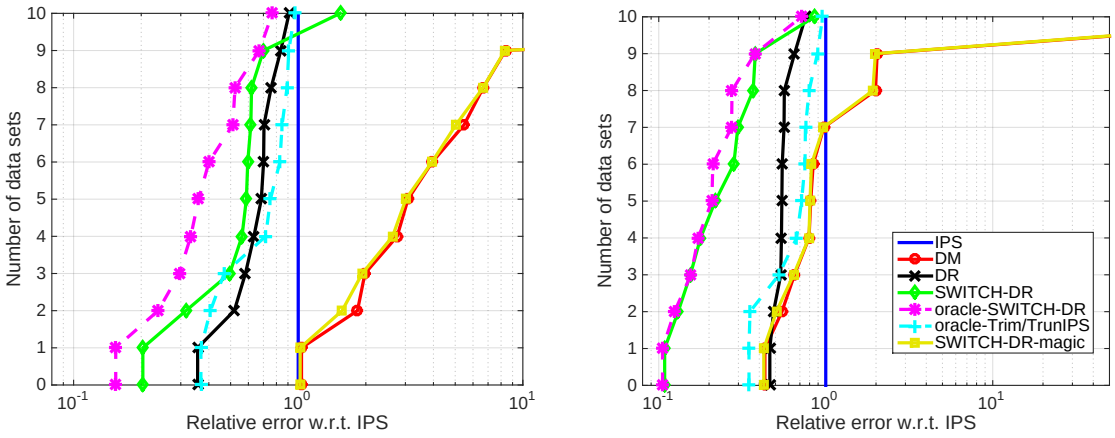
Our automatic tuning is related to the MAGIC estimator of Thomas & Brunskill (2016). The key differences are that we pick only one threshold τ, while they combine the estimates with many different τs using a weighting function. They pick this weighting function by optimizing a bias-variance tradeoff, but with significantly different bias and variance estimators. In our experiments, the automatic tuning using Eq. (9) generally works better than MAGIC.

5. Experiments

We next empirically evaluate the proposed SWITCH estimators on the 10 UCI data sets previously used for off-policy evaluation (Dudík et al., 2011). We convert the multi-class classification problem to contextual bandits by treating the labels as actions for a policy μ, and recording the reward of 1 if the correct label is chosen, and 0 otherwise.

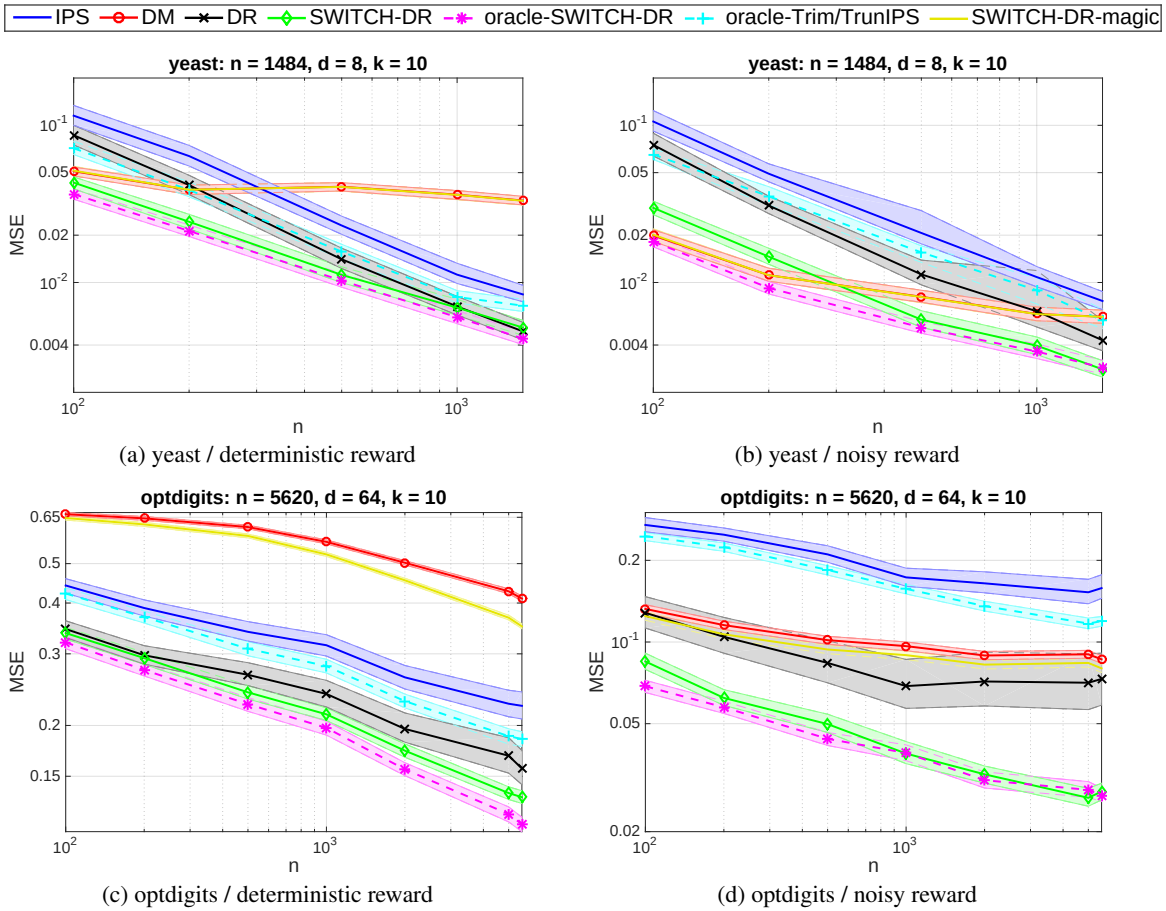
In addition to this *deterministic* reward model, we also consider a *noisy* reward model for each data set, which reveals the correct reward with probability 0.5 and outputs a random coin toss otherwise. Theoretically, this should lead to bigger σ^2 and larger variance in all estimators. In both reward models, $R_{\max} \equiv 1$ is a valid bound.

The target policy π is the deterministic decision of a logistic regression classifier learned on the multi-class data, while the logging policy μ samples according to the probability estimates of a logistic model learned on a covariate-shifted version of the data. The covariate shift is obtained as in prior work (Dudík et al., 2011; Gretton et al., 2009).



(a) Deterministic reward (b) Noisy reward

Figure 1. The number of UCI data sets where each method achieves at least a given Rel. MSE. On the left, the UCI labels are used as is; on the right, label noise is added. Curves towards top-left achieve smaller MSE in more cases. Methods in dashed lines are "cheating" by choosing the threshold τ to optimize test MSE. SWITCH-DR outperforms baselines and our tuning of τ is not too far from the best possible. Each data set uses an n which is the size of the data set, drawn via bootstrap sampling and results are averaged over 500 trials.



(a) yeast / deterministic reward (b) yeast / noisy reward

(c) optdigits / deterministic reward (d) optdigits / noisy reward

Figure 2. MSE of different methods as a function of input data size. *Top:* optdigits data set. *Bottom:* yeast data set.

In each data set with n examples, we treat the uniform distribution over the data set itself as a surrogate of the population distribution so that we know the ground truth of the rewards. Then, in the simulator, we randomly draw i.i.d. data sets of size $100, 200, 500, 1000, 2000, 5000, 10000, \ldots$ until reaching n, with 500 different repetitions of each size. We estimate MSE of each estimator by taking the empirical average of the squared error over the 500 replicates; note that we can calculate the squared error exactly, because we know v^π. For some of the methods, e.g., IPS and DR, the MSE can have a very large variance due to the potentially large importance weights. This leads to very large error bars if we estimate their MSE even with 500 replicates. To circumvent this issue, we report a clipped version of the MSE that truncates the squared error to 1, namely $\mathrm{MSE} = \mathbb{E}[(\hat{v} - v^\pi)^2 \wedge 1]$. This allows us to get valid confidence intervals for our empirical estimates of this quantity. Note that this does not change the MSE estimate of our approach at all, but is significantly more favorable towards IPS and DR. In this section, whenever we refer to "MSE", we are referring to this truncated version.

We compare SWITCH and SWITCH-DR against the following baselines: 1. *IPS*; 2. *DM trained via logistic regression*; 3. *DR*; 4. *Truncated and Reweighted IPS (TrunIPS)*; and 5. *Trimmed IPS (TrimIPS)*.

In DM, we train $\hat{r}$ and then evaluate the policy on the same contextual bandit data set. Following Dudík et al. (2011), DR is constructed by randomly splitting the contextual bandit data into two folds, estimating $\hat{r}$ on one fold, and then evaluating π on the other fold and vice versa, obtaining two estimates. The final estimate is the average of the two. TrunIPS is a variant of IPS, where importance weights are capped at a threshold τ and then renormalized to sum to one (see, e.g., Bembom & van der Laan, 2008). TrimIPS is a special case of SWITCH due to Bottou et al. (2013) described earlier, where $\hat{r} \equiv 0$.

For SWITCH and SWITCH-DR as well as TrunIPS and TrimIPS we select the parameter τ by our automatic tuning from Section 4.2. To evaluate our tuning approach, we also include the results for the τ tuned optimally in hindsight, which we refer to as the *oracle* setting, and also show results obtained by the multi-threshold MAGIC approach. In all these approaches we optimize among 21 possible thresholds, from an exponential grid between the smallest and the largest importance weight observed in the data, considering all actions in each observed context.

In order to stay comparable across data sets and data sizes, our performance measure is the relative MSE with respect to the IPS. Thus, for each estimator $\hat{v}$, we calculate $\mathrm{Rel.\ MSE}(\hat{v}) = \frac{\mathrm{MSE}(\hat{v})}{\mathrm{MSE}(\hat{v}_{\mathrm{IPS}})}$.

The results are summarized in Figure 1, plotting the number of data sets where each method achieves at least a given relative MSE.[4] Thus, methods that achieve smaller MSE across more data sets are towards the top-left corner of the plot, and a larger area under the curve indicates better performance. Some of the differences in MSE are several orders of magnitude large since the relative MSE is shown on the logaritmic scale. As we see, SWITCH-DR dominates all baselines and our empirical tuning of τ is not too far from the best possible. The automatic tuning by MAGIC tends to revert to DM, because its bias estimate is too optimistic and so DM is preferred whenever IPS or DR have some significant variance. The gains of SWITCH-DR are even greater in the noisy-reward setting, where we add label noise to UCI data.

In Figure 2, we illustrate the convergence of MSE as n increases. We select two data sets and show how SWITCH-DR performs against baselines in two typical cases: (i) when the direct method works well initially but is outperformed by IPS and DR as n gets large, and (ii) when the direct method works poorly. In the first case, SWITCH-DR outperforms both DM and IPS, while DR improves over IPS only moderately. In the second case, SWITCH-DR performs about as well as IPS and DR despite a poor performance of DM. In all cases, SWITCH-DR is robust to additional noise in the reward, while IPS and DR suffer from higher variance. Results for the remaining data sets are in Appendix D.

6. Conclusion

In this paper we have carried out minimax analysis of off-policy evaluation in contextual bandits and showed that IPS and DR are minimax optimal in the worst-case, when no consistent reward model is available. This result complements existing asymptotic theory with assumptions on reward models, and highlights the differences between agnostic and consistent settings. Practically, the result further motivates the importance of using side information, possibly by modeling rewards directly, especially when importance weights are too large. Given this observation, we propose a new class of estimators called SWITCH that can be used to combine any importance weighting estimators, including IPS and DR, with DM. The estimators adaptively switch between DM when the importance weights are large and either IPS or DR when the importance weights are small. We show that the new estimators have favorable theoretical properties and also work well on real-world data. Many interesting directions remain open for future work, including high-probability upper bounds on the finite-sample MSE of SWITCH estimators, as well as sharper finite-sample lower bounds under realistic assumptions on the reward model.

[4] For clarity, we have excluded SWITCH, which significantly outperforms IPS, but is dominated by SWITCH-DR. Similarly, we only report the better of oracle-TrimIPS and oracle-TrunIPS.

Acknowledgments

The work was partially completed during YW's internship at Microsoft Research NYC from May 2016 to Aug 2016. The authors would like to thank Lihong Li and John Langford for helpful discussions, Edward Kennedy for bringing our attention to related problems and recent developments in causal inference, and an anonymous reviewer for pointing out relevant econometric references and providing valuable feedback that helped connect our work with research on average treatment effects.

References

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.

Bertin, Karine et al. Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli*, 10(5):873–888, 2004.

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

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

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

Gretton, Arthur, Smola, Alex, Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten, and Schölkopf, Bernhard. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3(4):5, 2009.

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

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

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Horvitz, Daniel G and Thompson, Donovan 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.

Lafferty, John, Liu, Han, and Wasserman, Larry. Minimax theory, 2008. URL http://www.stat.cmu.edu/~larry/=sml/Minimax.pdf.

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

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

Rothe, Christoph. The value of knowing the propensity score for estimating average treatment effects. *IZA Discussion Paper Series*, 2016.

Sierpiński, Wacław. Sur les fonctions d'ensemble additives et continues. *Fundamenta Mathematicae*, 3:240–246, 1922.

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