Doubly robust off-policy evaluation with shrinkage

Yi Su
Cornell University
ys756@cornell.edu

Maria Dimakopoulou
Netflix
madima@stanford.edu

Akshay Krishnamurthy
Microsoft Research
akshay@cs.umass.edu

Miroslav Dudík
Microsoft Research
mdudik@microsoft.edu

Abstract

We design a new family of estimators for off-policy evaluation in contextual bandits. Our estimators are based on the asymptotically optimal approach of doubly robust estimation, but they shrink importance weights to obtain a better bias-variance tradeoff in finite samples. Our approach adapts importance weights to the quality of a reward predictor, interpolating between doubly robust estimation and direct modeling. When the reward predictor is poor, we recover previously studied weight clipping, but when the reward predictor is good, we obtain a new form of shrinkage. To navigate between these regimes and tune the shrinkage coefficient, we design a model selection procedure, which we prove is never worse than the doubly robust estimator. Extensive experiments on bandit benchmark problems show that our estimators are highly adaptive and typically outperform state-of-the-art methods.

1 Introduction

Many real-world applications, ranging from online news recommendation [18], advertising [4], and search engines [19] to personalized healthcare [28], are naturally modeled by the *contextual bandit* protocol [17], where a learner repeatedly observes a context, takes an action, and accrues reward. In news recommendation, the context is any information about the user, such as history of past visits, the action is the recommended article, and the reward could indicate, for instance, the user's click on the article. The goal is to maximize the reward, but the learner can only observe the reward for the chosen actions, and not for other actions.

In this paper we study a fundamental problem in contextual bandits known as *off-policy evaluation*, where the goal is to use the data gathered by a past algorithm, known as the *logging policy*, to estimate the average reward of a new algorithm, known as the *target policy*. High-quality off-policy estimates help avoid costly A/B testing and can also be used as subroutines for optimizing a policy [7, 1].

The key challenge in off-policy evaluation is distribution mismatch: the decisions made by the target policy differ from those made by the logging policy that collected the data. Three standard approaches tackle this problem. The first approach, known as *inverse propensity scoring* (IPS) [12], corrects for this mismatch by reweighting the data. The resulting estimate is unbiased, but may exhibit intolerably high variance when the importance weights (also known as inverse propensity scores) are large. The second approach, *direct modeling* (DM) or imputation, sidesteps the problem of large importance weights and directly fits a regression model to predict rewards. Non-parametric variants of direct modeling are asymptotically optimal [13], but in finite samples, direct modeling often suffers from a large bias [7]. The third approach, called the *doubly robust* (DR) estimator [20, 2, 7], combines IPS and DM: it first estimates the reward by DM, and then estimates a correction term by IPS. The approach is unbiased, its variance is smaller than that of IPS, and it is asymptotically optimal under

Preprint. Under review.

weaker assumptions than DM [21]. However, since DR uses the same importance weights as IPS, its variance can still be quite high, unless the reward predictor is highly accurate. Therefore, several works have developed variants of DR that clip or remove large importance weights. Weight clipping incurs a small bias, but substantially decreases the variance, yielding a lower mean squared error than standard DR [3, 4, 27, 23].

In this paper, we continue this line of work, by developing a systematic approach for designing estimators with favorable finite-sample performance. Our approach involves shrinking the importance weights to directly optimize a sharp bound on the mean squared error (MSE). We use two styles of upper bounds to obtain two classes of estimators. The first is based on an upper bound that is agnostic to the quality of the reward estimator and yields the previously studied weight clipping [16, 22, 23], which can be interpreted as *pessimistic shrinkage*. The second is based on an upper bound that incorporates the quality of the reward predictor, yielding a new estimator, which we call DR with *optimistic shrinkage*. Both classes of estimators involve a hyperparameter and specific choices produce the unbiased doubly robust estimator and the low-variance direct-modeling estimator; the optimal hyperparameter improves on both of these.

To tune the hyperparameter and navigate between the two estimator classes, we design a simple model-selection procedure. Model selection is crucial to our approach, but is important even for classical estimators, as their performance is highly dataset-dependent. In contrast with supervised learning, where cross-validation is a simple and effective strategy, distribution mismatch makes model selection quite challenging in contextual bandits. Our model selection approach again involves optimizing a bound on the MSE. Combined with our shrinkage estimators, we prove that our final estimator is never worse than DR. Thus, our estimators retain the asymptotic optimality of DR, but with improved finite-sample performance.

We evaluate our approach on benchmark datasets and compare its performance with a number of existing estimators across a comprehensive range of conditions. While our focus is on tuning importance weights, we also vary how we train a reward predictor, including the recently proposed *more robust doubly robust* (MRDR) approach [9], and apply our model selection to pick the best predictor. Our experiments show that our approach typically outperforms state-of-the-art methods in both off-policy evaluation and off-policy learning settings. We also find that the choice of the reward predictor changes, depending on whether weight shrinkage is used or not. Via extensive ablation studies, we identify a robust configuration of our shrinkage approach that we recommend as a practical choice.

Other related work. Off-policy estimation is studied in observational settings under the name *average treatment effect* (ATE) estimation, with many results on asymptotically optimal estimators [10, 11, 13, 21], but only few that optimize MSE in finite samples. Most notably, Kallus [14, 15] adjusts importance weights by optimizing MSE under smoothness (or parametric) assumptions on the reward function. This can be highly effective when the assumptions hold, but the assumptions are difficult to verify in practice. In contrast, we optimize importance weights without making any modeling assumptions other than boundedness of rewards.

2 Setup

We consider the *contextual bandits* protocol, where a decision maker interacts with the environment by repeatedly observing a *context* $x \in \mathcal{X}$, choosing an *action* $a \in \mathcal{A}$, and observing a *reward* $r \in [0, 1]$. The context space $\mathcal{X}$ can be uncountably large, but we assume that the action space $\mathcal{A}$ is finite. In the news recommendation example, x describes the history of past visits of a given user, a is a recommended article, and r equals one if the user clicks on the article and zero otherwise. We assume that contexts are sampled *i.i.d.* from some distribution $D(x)$ and rewards are sampled from some conditional distribution $D(r \mid x, a)$. We write $\eta(x, a) \coloneqq \mathbb{E}\left[r \mid x, a \right]$ for the expected reward, conditioned on a given context and action, and refer to η as the *regression function*.

The behavior of a decision maker is formalized as a conditional distribution $\pi(a \mid x)$ over actions given contexts, referred to as a *policy*. We also write $\pi(x, a, r) \coloneqq D(x)\pi(a \mid x)D(r \mid x, a)$ for the joint distribution over context-action-reward triples when actions are selected by the policy π. The expected reward of a policy π, called the *value* of π, is denoted as $V(\pi) \coloneqq \mathbb{E}_{(x,a,r)\sim\pi}[r]$.

In the off-policy evaluation problem, we are given a dataset $\{(x_i, a_i, r_i)\}_{i=1}^{n} \sim \mu$ consisting of context-action-reward triples collected by some *logging policy* μ, and we would like to estimate the value of a *target policy* π. The quality of an estimator $\hat{V}(\pi)$ is measured by the *mean squared error*

$$\mathrm{MSE}\big(\hat{V}(\pi)\big) := \mathbb{E}\Big[\big(\hat{V}(\pi) - V(\pi)\big)^2\Big],$$

where the expectation is with respect to the data generation process. In analyzing the error of an estimator, we rely on the decomposition of MSE into the bias and variance terms:

$$\mathrm{MSE}\big(\hat{V}(\pi)\big) = \mathrm{Bias}\big(\hat{V}(\pi)\big)^2 + \mathrm{Var}\big[\hat{V}(\pi)\big], \qquad \mathrm{Bias}\big(\hat{V}(\pi)\big) := \mathbb{E}\big[\hat{V}(\pi) - V(\pi)\big].$$

We consider three standard approaches for off-policy evaluation. The first two are *direct modeling* (DM) and *inverse propensity scoring* (IPS). In the former, we train a reward predictor $\hat{\eta} : \mathcal{X} \times \mathcal{A} \to [0,1]$ and use it to impute rewards. In the latter, we simply reweight the data. The two estimators are:

$$\hat{V}_{\mathrm{DM}}(\pi;\hat{\eta}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi(a \mid x_i)\hat{\eta}(x_i,a), \qquad \hat{V}_{\mathrm{IPS}}(\pi) := \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i \mid x_i)}{\mu(a_i \mid x_i)} r_i.$$

For a concise notation, let $w(x,a) := \pi(a \mid x)/\mu(a \mid x)$, denote the *importance weight*. We make a standard assumption that π is absolutely continuous with respect to μ, meaning that $\mu(a \mid x) > 0$ whenever $\pi(a \mid x) > 0$. This condition ensures that the importance weights are well defined and that $\hat{V}_{\mathrm{IPS}}(\pi)$ is an unbiased estimator for $V(\pi)$. However, if there is substantial mismatch between π and μ, then the importance weights will be large and $\hat{V}_{\mathrm{IPS}}(\pi)$ will have large variance. On the other hand, given any fixed reward predictor $\hat{\eta}$ (fit on a separate dataset), $\hat{V}_{\mathrm{DM}}(\pi)$ has low variance, independent of the distribution mismatch, but it is typically biased due to approximation errors in fitting $\hat{\eta}$.

The third approach, called the *doubly robust* (DR) estimator [8], combines DM and IPS:

$$\hat{V}_{\mathrm{DR}}(\pi;\hat{\eta}) := \hat{V}_{\mathrm{DM}}(\pi;\hat{\eta}) + \frac{1}{n}\sum_{i=1}^{n} w(x_i,a_i)\big(r_i - \hat{\eta}(x_i,a_i)\big). \tag{1}$$

The DR estimator applies IPS to a shifted versions of rewards, using $\hat{\eta}$ as a control variate to decrease the variance of IPS. DR preserves unbiasedness of IPS and achieves asymptotic optimality, as long as it is possible to derive sufficiently good reward predictors $\hat{\eta}$ given enough data [21].

When $\hat{\eta} = 0$, DR recovers IPS and is plagued by the same large variance. However, even when the reward predictor $\hat{\eta}$ is perfect, any intrinsic stochasticity in the rewards may cause the terms $r_i - \hat{\eta}(x_i,a_i)$, appearing in the DR estimator, to be far from zero. Multiplied by large importance weights $w(x_i,a_i)$, these terms yield large variance for DR in comparison with DM. Several approaches seek a more favorable bias-variance trade-off by clipping, removing, or re-scaling the importance weights [27, 23, 14, 15]. Our work also seeks to systematically replace the weights $w(x_i,a_i)$ with new weights $\hat{w}(x_i,a_i)$ to bring the variance of DR closer to that of DM.

In practice, $\hat{\eta}$ is biased due to approximation errors, so in this paper we make no assumptions about its quality. At the same time, we would like to make sure that our estimators can adapt to high-quality $\hat{\eta}$ if it is available. To motivate our adaptive estimator, we assume that $\hat{\eta}$ is trained via weighted least squares regression on a separate dataset than used in $\hat{V}_{\mathrm{DR}}$. That is, for a dataset $\{(x_j,a_j,r_j)\}_{j=1}^{m} \sim \mu$, we consider a weighting function $z : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ and solve

$$\hat{\eta} := \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{m}\sum_{j=1}^{m} z(x_j,a_j)\big(f(x_j,a_j) - r_j\big)^2, \tag{2}$$

where $\mathcal{F}$ is some function class of reward predictors. Natural choices of the weighting function z, explored in our experiments, include $z(x,a) = 1$, $z(x,a) = w(x,a)$ and $z(x,a) = w^2(x,a)$. We stress that the assumption on how we fit $\hat{\eta}$ only serves to guide our derivations, but we make no specific assumptions about its quality. In particular, we do not assume that $\mathcal{F}$ contains a good approximation of η.

3 Our Approach: DR with Shrinkage

Our approach replaces the importance weight mapping $w : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ in the DR estimator (1) with a new weight mapping $\hat{w} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ found by directly optimizing sharp bounds on the

MSE. The resulting estimator, which we call the *doubly robust estimator with shrinkage* (DRs) thus depends on both the reward predictor $\hat{\eta}$ and the weight mapping $\hat{w}$:

$$\hat{V}_{\text{DRs}}(\pi; \hat{\eta}, \hat{w}) \coloneqq \hat{V}_{\text{DM}}(\pi; \hat{\eta}) + \frac{1}{n} \sum_{i=1}^{n} \hat{w}(x_i, a_i)\big(r_i - \hat{\eta}(x_i, a_i)\big). \tag{3}$$

We assume that $0 \leq \hat{w} \leq w$, justifying the terminology "shrinkage". For a fixed choice of π and $\hat{\eta}$, we will seek the mapping $\hat{w}$ that minimizes the MSE of $\hat{V}_{\text{DRs}}(\pi; \hat{\eta}, \hat{w})$, which we simply denote as $\text{MSE}(\hat{w})$. We similarly write $\text{Bias}(\hat{w})$ and $\text{Var}(\hat{w})$ for the bias and variance of this estimator.

We treat $\hat{w}$ as the optimization variable and consider two upper bounds on MSE: an optimistic one and a pessimistic one. In both cases, we separately bound $\text{Bias}(\hat{w})$ and $\text{Var}(\hat{w})$. To bound the bias, we use the following expression, derived from the fact that $\hat{V}_{\text{DRs}}$ is unbiased when $\hat{w} = w$:

$$\begin{aligned} \text{Bias}(\hat{w}) &= \mathbb{E}\big[\hat{V}_{\text{DRs}}(\pi; \hat{\eta}, \hat{w})\big] - \mathbb{E}\big[\hat{V}_{\text{DRs}}(\pi; \hat{\eta}, w)\big] \\ &= \mathbb{E}_{\mu}\big[\big(\hat{w}(x, a) - w(x, a)\big)\big(r - \hat{\eta}(x, a)\big)\big]. \end{aligned} \tag{4}$$

To bound the variance, we rely on the following proposition, which states that it suffices to focus on the second moment of the terms $\hat{w}(x_i, a_i)\big(r_i - \hat{\eta}(x_i, a_i)\big)$:

Proposition 1. *If* $0 \leq \hat{w}(x, a) \leq w(x, a)$ *and* $|\,\hat{\eta}(x, a) - \eta(x, a)\,| \leq 1$, *then*

$$\left| \text{Var}(\hat{w}) - \frac{1}{n}\mathbb{E}_{\mu}\Big[\hat{w}^2(x, a)\big(r - \hat{\eta}(x, a)\big)^2\Big] \right| \leq \frac{5}{n}.$$

The proof of Proposition 1 and other mathematical statements from this paper are in the appendix.

We derive estimators for two different regimes depending on the quality of the reward predictor $\hat{\eta}$. Since we do not know the quality of $\hat{\eta}$ a priori, in the next section we obtain a model selection procedure to select between these two estimators.

3.1 DR with Optimistic Shrinkage

Our first family of estimators is based on an optimistic MSE bound, which adapts to the quality of $\hat{\eta}$, and which we expect to be tighter when $\hat{\eta}$ is more accurate. Recall that $\hat{\eta}$ is trained to minimize weighted square loss with respect to some weighting function z, which we denote as

$$L(\hat{\eta}) \coloneqq \mathbb{E}_{\mu}\Big[z(x, a)\big(r - \hat{\eta}(x, a)\big)^2\Big].$$

We next bound the bias and variance in terms of the weighted square loss $L(\hat{\eta})$.

To bound the bias we apply the Cauchy-Schwarz inequality to (4):

$$\text{Bias}(\hat{w}) \leq \sqrt{\mathbb{E}_{\mu}\left[\tfrac{1}{z(x,a)}\big(\hat{w}(x, a) - w(x, a)\big)^2\right]}\sqrt{L(\hat{\eta})}. \tag{5}$$

To bound the variance, we invoke Proposition 1 and focus on bounding the quantity $\mathbb{E}_{\mu}\big[\hat{w}^2(r - \hat{\eta})^2\big]$. Using the Cauchy-Schwarz inequality, the fact that $0 \leq \hat{w} \leq w$, and $|r - \hat{\eta}(x, a)| \leq 1$, we obtain

$$\begin{aligned} \mathbb{E}_{\mu}\Big[\hat{w}^2(x, a)\big(r - \hat{\eta}(x, a)\big)^2\Big] &\leq \sqrt{\mathbb{E}_{\mu}\left[\tfrac{1}{z(x,a)}\hat{w}^4(x, a)\right]}\sqrt{\mathbb{E}_{\mu}\left[z(x, a)\big(r - \hat{\eta}(x, a)\big)^4\right]} \\ &\leq \sqrt{\mathbb{E}_{\mu}\left[\tfrac{w^2(x,a)}{z(x,a)}\hat{w}^2(x, a)\right]}\sqrt{L(\hat{\eta})}. \end{aligned} \tag{6}$$

Combining the bounds (5) and (6) with Proposition 1 yields the following bound on $\text{MSE}(\hat{w})$:

$$\text{MSE}(\hat{w}) \leq \mathbb{E}_{\mu}\left[\tfrac{1}{z(x,a)}\big(\hat{w}(x, a) - w(x, a)\big)^2\right] L(\hat{\eta}) + \sqrt{\mathbb{E}_{\mu}\left[\tfrac{w^2(x,a)}{z(x,a)}\hat{w}^2(x, a)\right]}\sqrt{L(\hat{\eta})} + \frac{5}{n}.$$

A direct minimization of this bound appears to be a high dimensional optimization problem. Instead of minimizing the bound directly, we note that it is a strictly increasing function of the two expectations that appear in it. Thus, its minimizer must be on the Pareto front with respect to the two expectations, meaning that for some choice of $\lambda \in [0, \infty]$, the minimizer can be obtained by solving

$$\underset{\hat{w}}{\text{Minimize}} \;\; \lambda \mathbb{E}_\mu \left[\tfrac{1}{z(x,a)} \big(\hat{w}(x,a) - w(x,a)\big)^2 \right] + \mathbb{E}_\mu \left[\tfrac{w^2(x,a)}{z(x,a)} \hat{w}^2(x,a) \right].$$

The objective decomposes across contexts and actions. Taking the derivative with respect to $\hat{w}(x,a)$ and setting to zero yields the solution

$$\hat{w}_{\mathrm{o},\lambda}(x,a) = \frac{\lambda}{w^2(x,a) + \lambda} w(x,a),$$

where "o" above is a mnemonic for optimistic shrinkage. We refer to the DRs estimator with $\hat{w} = \hat{w}_{\mathrm{o},\lambda}$ as the *doubly robust estimator with optimistic shrinkage* (DRos) and denote it by $\hat{V}_{\text{DRos}}(\pi; \hat{\eta}, \lambda)$. Note that this estimator does not depend on z, although it was included in the optimization objective.

3.2 DR with Pessimistic Shrinkage

Our second estimator family makes no assumptions on the quality of $\hat{\eta}$ beyond the range bound $\hat{\eta}(x,a) \in [0,1]$, which implies $|\hat{\eta}(x,a) - r| \le 1$ and yields the bias and second-moment bounds

$$\text{Bias}(\hat{w}) \le \mathbb{E}_\mu\big[|\hat{w}(x,a) - w(x,a)|\big], \qquad \mathbb{E}_\mu\big[\hat{w}(x,a)^2 (r - \hat{\eta}(x,a))^2\big] \le \mathbb{E}_\mu\big[\hat{w}(x,a)^2\big]. \tag{7}$$

As before, we do not optimize the resulting MSE bound directly and instead solve for the Pareto front points parameterized by $\lambda \in [0, \infty]$:

$$\underset{\hat{w}}{\text{Minimize}} \;\; \lambda \mathbb{E}_\mu\big[|\hat{w}(x,a) - w(x,a)|\big] + \mathbb{E}_\mu\big[\hat{w}(x,a)^2\big].$$

The objective again decomposes across context-action pairs, yielding the solution

$$\hat{w}_{\mathrm{p},\lambda}(x,a) = \min\{\lambda,\, w(x,a)\},$$

which recovers (and justifies) existing weight-clipping approaches [16, 22, 23]. We refer to the resulting estimator as $\hat{V}_{\text{DRps}}(\pi; \hat{\eta}, \lambda)$, for *doubly robust with pessimistic shrinkage*, since we have used the worst-case bounds in the derivation. See Appendix A for detailed calculations.

3.3 Basic Properties

The two shrinkage estimators, for a suitable choice of λ, are never worse than DR or DM. This is an immediate consequence of their form and serves as a basic sanity check.

Proposition 2. *Let $\hat{V}$ denote either $\hat{V}_{\text{DRos}}$ or $\hat{V}_{\text{DRps}}$. Then for any $\hat{\eta}$ there exists $\lambda^\star \ge 0$ such that*

$$\text{MSE}\big(\hat{V}(\pi; \hat{\eta}, \lambda^\star)\big) \le \min\Big\{ \text{MSE}\big(\hat{V}_{\text{DM}}(\pi; \hat{\eta})\big),\, \text{MSE}\big(\hat{V}_{\text{DR}}(\pi; \hat{\eta})\big) \Big\}.$$

Both estimators actually *interpolate* between $\hat{V}_{\text{DM}}$ and $\hat{V}_{\text{DR}}$. As λ varies, we obtain $\hat{V}_{\text{DM}}$ for $\lambda = 0$ and $\hat{V}_{\text{DR}}$ as $\lambda \to \infty$. The optimal choice of λ is therefore always competitive with both. While we do not know $\lambda^\star$, in the next section, we derive a model selection procedure that finds a good λ.

4 Model Selection

All of our estimators can be written as finite sums of the form $\hat{V}(\theta) = \frac{1}{n} \sum_{i=1}^n Z_i(\theta)$, where θ are some fixed hyperparameters and $Z_i(\theta)$ are *i.i.d.* random variables. For example $\theta = (\hat{\eta}, \mathrm{o}, \lambda)$ denotes that we are using a reward predictor $\hat{\eta}$ and the optimistic shrinkage with the parameter λ. To choose hyperparameters, we first estimate the variance of $\hat{V}(\theta)$ by the sample variance

$$\widehat{\text{Var}}(\theta) := \widehat{\text{Var}}\big(\hat{V}(\theta)\big) = \frac{1}{n(n-1)} \sum_{i=1}^n \Big(Z_i(\theta) - \bar{Z}(\theta) \Big)^2 \qquad \text{where } \bar{Z}(\theta) := \frac{1}{n} \sum_{i=1}^n Z_i(\theta).$$

We also form a data-dependent upper bound on the bias, which we call $\text{BiasUB}(\theta)$. The only requirement is that for all θ, $\text{Bias}(\theta) \le \text{BiasUB}(\theta)$ (with high probability), and that $\text{BiasUB}(\theta) = 0$ whenever $\text{Bias}(\theta) = 0$; this holds for both bias bounds from the previous section, as they become zero when $\hat{w} = w$. Now, we simply choose θ from some class of hyperparameters Θ to optimize the estimate of the MSE:

$$\hat{\theta} \leftarrow \underset{\theta \in \Theta}{\text{Minimize}} \;\; \text{BiasUB}(\theta)^2 + \widehat{\text{Var}}(\theta).$$

This model selection procedure is related to MAGIC [26] as well as the model selection procedure for the SWITCH estimator [27]. In comparison with MAGIC, we pick a single parameter value θ rather than aggregating several, and we use different bias and variance estimates. SWITCH uses the pessimistic bias bound (7), whereas we will also use two additional bounding strategies.

Theorem 3. *Let Θ be a finite set of hyperparameter values and let $\Theta_0 := \{\theta \in \Theta : \text{Bias}(\theta) = 0\}$ denote the subset corresponding to unbiased procedures. Then*

$$\text{MSE}(\hat{\theta}) \leq \min_{\theta_0 \in \Theta_0} \text{MSE}(\theta_0) + o(n^{-1}).$$

The theorem shows that the MSE of our final estimator after model selection is no worse than the best unbiased estimator in consideration. The $o(n^{-1})$ term is asymptotically negligible, as the MSE itself is $\Theta(1/n)$. In particular, since the doubly robust estimator is unbiased and a special case of our estimators with $\lambda = \infty$, we retain asymptotic optimality as long as we include $\lambda = \infty$ in the model selection. In fact, since we may perform model selection over many different choices for $\hat{\eta}$, our estimator is competitive with the doubly robust approach using the best reward predictor.

There are many natural ways to construct data-dependent upper bounds on the bias with the required properties. The three we use in our experiments involve using samples to approximate expectations in: (i) the expression for the bias given in (4), (ii) the optimistic bias bound in (5), and (iii) the pessimistic bias bound in (7). In our theory, these estimates need to be adjusted to obtain high-probability confidence bounds. In our experiments we evaluate both the basic estimates and a adjusted variant where we add twice the standard error.

5 Experiments

We evaluate our new estimators on the tasks of off-policy evaluation and off-policy learning, and compare their performance with previous estimators. Our secondary goal is to identify the configuration of the shrinkage estimator that is most robust for use in practice.

Datasets. Following prior work [8, 25, 27, 9, 23], we simulate bandit feedback on 9 UCI multi-class classification datasets. This lets us evaluate estimators in a broad range of conditions and gives us ground-truth policy values (see Table 3 in the appendix for the dataset statistics). Each multi-class dataset with k classes corresponds to a contextual bandit problem with k possible actions coinciding with classes. We consider either *deterministic rewards*, where on multiclass example (x, y^*), the action y yields the reward $r = \mathbf{1}\{y = y^*\}$, or *stochastic rewards* where $r = \mathbf{1}\{y = y^*\}$ with probability 0.75 and $r = 1 - \mathbf{1}\{y = y^*\}$ otherwise. For every dataset, we hold out 25% of the examples to measure ground-truth. On the remaining 75% of the dataset, we use logging policy μ to simulate n bandit examples by sampling a context x from the dataset, sampling an action $y \sim \mu(\cdot|x)$ and then observing a deterministic or stochastic reward r. The value of n varies across experimental conditions. [1]

Policies. We use the 25% held-out data to obtain logging and target policies as follows. We first obtain two deterministic policies $\pi_{1,\text{det}}$ and $\pi_{2,\text{det}}$ by training two logistic models on the same data, but using either the first or second half of the features. We obtain stochastic policies parameterized by (α, β), following the *softening* technique of Farajtabar et al. [9]. Specifically, $\pi_{1,(\alpha,\beta)}(a \mid x) = (\alpha + \beta u)$ if $a = \pi_{1,\text{det}}(x)$ and $\pi_{1,(\alpha,\beta)}(a \mid x) = \frac{1-\alpha-\beta u}{k-1}$ otherwise, where $u \sim \text{Unif}([-0.5, 0.5])$. In off-policy evaluation experiments, we consider a fixed target and several choices of logging policy (see Table 1). In off-policy learning we use $\pi_{1,(0.9,0)}$ as the logging policy.

Table 1: Policies.

	base	α	β
target	$\pi_{1,\text{det}}$	0.9	0
logging	$\pi_{1,\text{det}}$	0.7	0.2
	$\pi_{1,\text{det}}$	0.5	0.2
	—	$1/k$	0
	$\pi_{2,\text{det}}$	0.3	0.2
	$\pi_{2,\text{det}}$	0.5	0.2
	$\pi_{2,\text{det}}$	0.95	0.1

Reward predictors. We obtain reward predictors $\hat{\eta}$ by training linear models via weighted least squares with ℓ_2 regularization, using 2-fold cross-validation to tune the regularization parameter. We

[1] If n is the size of the remaining 75%, we use each example exactly once, but there is still variation in the ordering of examples, actions taken, and rewards.

	$\hat{\eta} \equiv 0$	$z \equiv 1$	$z = w$	$z = w^2$	MRDR
DM	0 (0)	47 (23)	45 (22)	41 (31)	11 (5)
DR	27 (2)	86 (9)	90 (4)	85 (5)	65 (0)
snDR	63 (7)	80 (2)	85 (8)	69 (4)	54 (0)
DRs	23 (19)	44 (16)	35 (4)	62 (35)	18 (2)

	DRps	DRos
$\hat{\eta} \equiv 0$	21	51
$z \equiv 1$	58	28
$z = w$	55	30
$z = w^2$	55	29
MRDR	49	29

Table 2: Ablation analysis. Left: we compare reward predictors using a fixed estimator (with oracle tuning if applicable). We report the number of conditions where a regressor is statistically indistinguishable from the best and, in parenthesis, the number of conditions where it statistically dominates all others. Right: we compare different shrinkage types using a fixed reward predictor (with oracle tuning) reporting the number of conditions where one statistically dominates the other.

experiment with weights $z(x,a) \in \{1, w(x,a), w^2(x,a)\}$ and we also consider the special weight design of Farajtabar et al. [9], which we call MRDR (see Appendix C for details). In evaluation experiments, we use $1/2$ of the bandit data to train $\hat{\eta}$; in learning experiments, we use $1/3$ of the bandit data to train $\hat{\eta}$. In addition to the four trained reward predictors, we also consider $\hat{\eta} \equiv 0$.

Baselines. We include a number of estimators in our evaluation: the direct modeling approach (DM), doubly-robust (DR) and its self-normalized variant (snDR), our approach (DRs), and the doubly-robust version of the SWITCH estimator of Wang et al. [27], which also performs a form of weight clipping.[2] Note that DR with $\hat{\eta} \equiv 0$ is identical to inverse propensity scoring (IPS); we refer to its self-normalized variant as snIPS. Our estimator and SWITCH have hyperparameters, which are selected by their respective model selection procedures (see Appendix C for details about the hyperparameter grid).

5.1 Off-policy Evaluation

We begin by evaluating different configurations of DRs via an ablation analysis. Then we compare DRs with baseline estimators. We have a total of 108 experimental conditions: for each of the 9 datasets we use 6 logging policies and consider stochastic or deterministic rewards. Except for the learning curves below, we always take n to be all available bandit data (75% of the overall dataset).

We measure performance with clipped MSE, $\mathbb{E}\big[(\hat{V} - V(\pi))^2 \wedge 1\big]$, where $\hat{V}$ is the estimator and $V(\pi)$ is the ground-truth value (computed on the held-out 25% of the data). We use 500 replicates of bandit-data generation to estimate the MSE; statistical comparisons are based on paired t-tests at significance level 0.05. In some of our ablation experiments, we pick the best hyperparameters against the test set on a per-replicate basis, which we call *oracle tuning* and always call out explicitly.

Ablation analysis. Table 2 displays the results of two ablation studies, one evaluating different reward predictors and the other evaluating the optimistic and pessimistic shrinkage types.

On the left, for each fixed estimator type (e.g., DR) we compare different reward predictors by reporting the number of conditions where it is statistically indistinguishable from the best and the number of conditions where it statistically dominates all other estimators using that predictor. For DRs we use oracle tuning for the shrinkage type and coefficient. The table shows that weight shrinkage strongly influences the choice of regressor. For example, $z \equiv 1$ and $z = w$ are top choices for DR, but with the inclusion of shrinkage in DRs, $z = w^2$ emerges as the best single choice.[3] In our comparison experiments below, we therefore restrict our attention to $z = w^2$ and additionally also consider $\hat{\eta} \equiv 0$, because it allows including IPS as a special case of DRs. Somewhat surprisingly, MRDR is in our experiments dominated by other reward predictors (except for $\hat{\eta} \equiv 0$), and this remains true even with a deterministic target policy (see Table 4 in the appendix).

On the right of Table 2, we compare optimistic and pessimistic shrinkage when paired with a fixed reward predictor (using oracle tuning for the shrinkage coefficient). We report the number of times that one estimator statistically dominates the other. The results suggest that both shrinkage types are important for robust performance across conditions, so we consider both choices going forward.

[2] For simplicity we call this estimator SWITCH, although Wang et al. call it SWITCH-DR.

[3] The oracle can always select the shrinkage parameter in DRs to recover DM or DR, but, according to the table, the oracle choices for $z \equiv 1$ and $z = w$ lead to inferior performance compared with $z = w^2$.

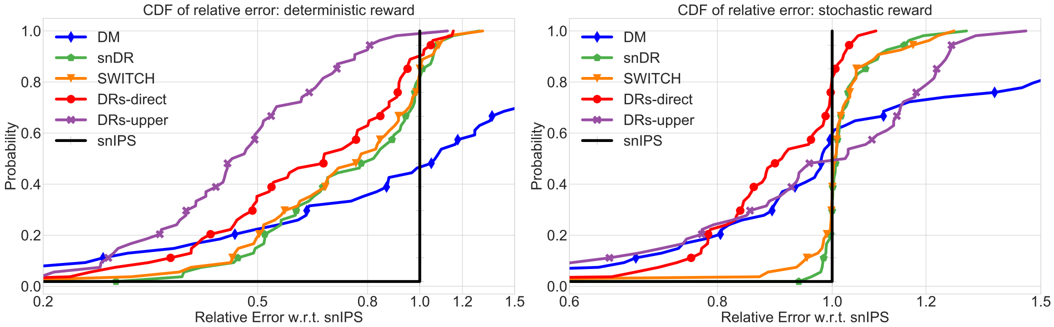

Figure 1: CDF plots of normalized MSE aggregated across all conditions with deterministic rewards (left) and stochastic rewards (right). See Table 6 in the appendix for statistical significance.

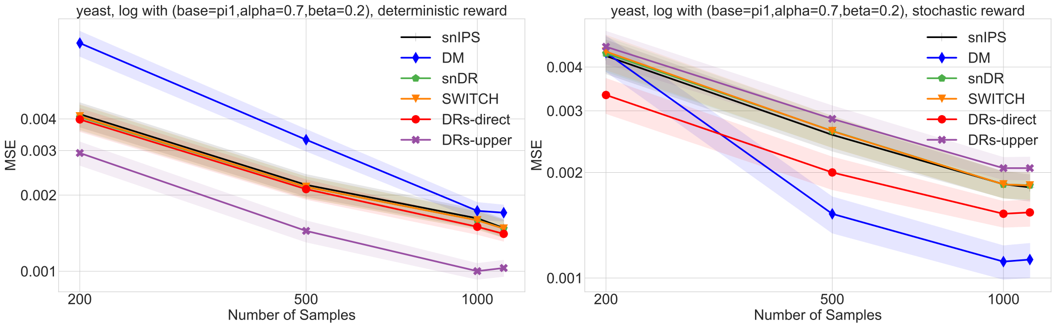

Figure 2: MSE for a varying number of samples n, for the dataset yeast and logging policy $\pi_{1,(0.7,0.2)}$, with deterministic rewards (left) and stochastic rewards (right).

Comparisons. In Figure 1, we compare our new estimator with the baselines. We visualize the results by plotting the cumulative distribution function (CDF) of the normalized MSE (normalized by the MSE of snIPS) across conditions for each method. Better performance corresponds to CDF curves towards top-left corner, meaning the method achieves a lower MSE more frequently. The left plot summarizes 54 conditions where the reward is deterministic, while the right plot considers the 54 stochastic reward conditions. For DRs we consider two model selection procedures outlined in Section 4 that differ in their choice of BiasUB. DRs-direct estimates the expectations in the expressions in Eqs. (4), (5), and (7) (corresponding to the bias and bias bounds) by empirical averages and takes their pointwise minimum. DRs-upper adds to theses estimates twice their standard error, before taking minimum, more closely matching our theory. For DRs, we only use the zero reward predictor and the one trained with $z = w^2$, and we always select between both shrinkage types. Since SWITCH also comes with a model selection procedure, we use it to select between the same two reward predictors as DRs.

In the deterministic case (left plot) we see that DRs-upper has the best aggregate performance, by a large margin. DRs-direct also has better aggregate performance than the baselines on most of the conditions. In the stochastic case, DRs-direct has similarly strong performance, but DRs-upper degrades considerably, suggesting this model selection scheme is less robust to stochastic rewards. We illustrate this phenomenon in Figure 2, plotting the MSE as a function of the number of samples for one choice of a logging policy and dataset, with deterministic rewards on the left and stochastic on the right. Because of a more robust performance, we therefore advocate for DRs-direct as our final method.

5.2 Off-policy Learning

Following prior work [24, 25, 23], we learn a stochastic linear policy $\pi_{\mathbf{u}}$ where $\pi_{\mathbf{u}}(a \mid x) \propto \exp\{\mathbf{u}^\top \mathbf{f}(x, a)\}$ and $\mathbf{f}(x, a)$ is a featurization of context-action pairs. We solve ℓ_2-regularized empirical risk minimization $\hat{\mathbf{u}} = \mathrm{argmin}_{\mathbf{u}}\big[-\hat{V}(\pi_{\mathbf{u}}) + \gamma\|\mathbf{u}\|^2\big]$ via gradient descent, where $\hat{V}$ is a policy-value estimator and $\gamma > 0$ is a hyperparameter. For these experiments, we partition the data into four quarters: one full information segment for

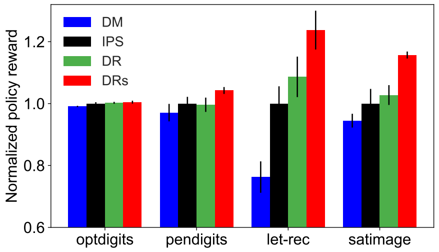

Figure 3: Learning experiments.

training the logging policy and as a test set, and three bandit segments for (1) training reward predictors, (2) learning the policy, and (3) hyperparameter tuning and model selection. The logging policy is $\pi_{1,(0.9,0)}$ and since there is no fixed target policy, we consider three reward predictors: $\hat{\eta} \equiv 0$, and $\hat{\eta}$ trained with $z = \frac{1}{\mu(a|x)}$ and $z = \frac{1}{\mu(a|x)^2}$.

In Figure 3 we display the performance of four methods (DM, DR, IPS, and DRs-direct) on four of the UCI datasets. For each method, we compute the average value of the learned policy on the test set (averaged over 10 replicates) and we report this value normalized by that for IPS. For DM and DR, we select the hyperparamater γ and reward predictor optimally in hindsight, while for DRs we use our model selection method. Note that we do not compare with SWITCH here as it is not amenable to gradient-based optimization [23]. Except for the opt-digits dataset, where all the methods are comparable, we find that off-policy learning using DRs-direct always outperforms the baselines.

Acknowledgments

We thank Alekh Agarwal for valuable input in the early discussions about this project. Part of this work was completed while Yi Su was visiting Microsoft Research.

References

[1] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv:1702.02896*, 2017.

[2] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 2005.

[3] Oliver Bembom and Mark J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical report, UC Berkeley, 2008.

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

[5] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[6] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[7] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

[8] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

[9] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, 2018.

[10] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 1998.

[11] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 2003.

[12] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952.

[13] Guido Imbens, Whitney Newey, and Geert Ridder. Mean-squared-error calculations for average treatment effects. *ssrn.954748*, 2007.

[14] Nathan Kallus. A Framework for Optimal Matching for Causal Inference. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[15] Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, 2018.

[16] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 2007.

[17] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, 2008.

[18] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *International Conference on Web Search and Data Mining*, 2011.

[19] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *International Conference on World Wide Web*, 2015.

[20] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 1995.

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

[22] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, 2010.

[23] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. Cab: Continuous adaptive blending estimator for policy evaluation and learning. In *International Conference on Machine Learning*, 2018.

[24] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, 2015.

[25] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015.

[26] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

[27] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, 2017.

[28] Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 2017.

A Derivation of Shrinkage Estimators

In this section we provide detailed derivations for the two estimators.

We first derive the pessimistic version. Recall that the optimization problem decouples across (x, a), so we focus on a single (x, a) pair, and we omit explicit dependence on these. Fixing $\lambda \geq 0$, we must solve

$$\underset{0 \leq \hat{w} \leq w}{\text{Minimize}} \;\; \mu \hat{w}^2 + \lambda \,|\, \mu(\hat{w} - w) \,| \,.$$

The optimality conditions are that

$$2\mu\hat{w} + \lambda\mu v = 0 \qquad \text{and} \qquad v \in \partial\,|\hat{w} - w| = \begin{cases} 1 \text{ if } & \hat{w} > w \\ [-1, 1] \text{ if } & \hat{w} = w \\ -1 \text{ if } & \hat{w} < w \end{cases}$$

The first case for v cannot occur, since setting $v = 1$ would make $\hat{w} = -\lambda/2 \leq 0$ (according to the first equation), but we know that $w \geq 0$. If $\hat{w} < w$ then we must have $\hat{w} = \lambda/2$. And so, we get

$$\hat{w}_{\text{p},\lambda}(x, a) = \min\{\lambda/2, w(x, a)\},$$

which is the clipped estimator.

For the optimistic version, the optimization problem is

$$\underset{0 \leq \hat{w} \leq w}{\text{Minimize}} \;\; \mu \hat{w}^2 w^2/z + \lambda\mu(\hat{w} - w)^2/z$$

The optimality conditions are

$$2\mu w^2 \hat{w}/z + 2\lambda\mu(\hat{w} - w)/z = 0.$$

This gives the optimistic estimator

$$\hat{w}_{\text{o},\lambda}(x, a) = \frac{\lambda}{w(x, a)^2 + \lambda} w(x, a).$$

Notice that this estimator does not depend on the weighting function z, so it does not depend on how we train the regression model.

B Proofs

Proof of Proposition 1. The law of total variance gives

$$\begin{aligned} \mathrm{Var}(\hat{w}) &= \frac{1}{n} \underset{x,a,r\sim\mu}{\mathrm{Var}} \left(\sum_{a' \in \mathcal{A}} \pi(a' \mid x)\hat{\eta}(x, a') + \hat{w}(x, a)(r - \hat{\eta}(x, a)) \right) \\ &= \underbrace{\underset{x}{\mathbb{E}} \underset{a,r\sim\mu}{\mathrm{Var}} \left(\sum_{a' \in \mathcal{A}} \pi(a' \mid x)\hat{\eta}(x, a') + \hat{w}(x, a)(r - \hat{\eta}(x, a)) \right)}_{=:T_1} \\ &\quad \underbrace{+ \underset{x}{\mathrm{Var}} \underset{a,r\sim\mu}{\mathbb{E}} \left(\sum_{a' \in \mathcal{A}} \pi(a' \mid x)\hat{\eta}(x, a') + \hat{w}(x, a)(r - \hat{\eta}(x, a)) \right)}_{=:T_2} \end{aligned}$$

For the first term, since $\sum_{a' \in \mathcal{A}} \pi(a' \mid x)\hat{\eta}(x, a')$ does not depend on a, r, it does not contribute to the conditional variance, and we get

$$\begin{aligned} T_1 &= \underset{x}{\mathbb{E}} \underset{a,r}{\mathrm{Var}} \left(\, \hat{w}(x, a)(r - \hat{\eta}(x, a)) \, \right) = \underset{x,a,r}{\mathbb{E}} \left[\, \hat{w}(x, a)^2 (r - \hat{\eta}(x, a))^2 \, \right] - \underset{x}{\mathbb{E}} \left[\underset{a,r}{\mathbb{E}} \left[\, \hat{w}(x, a)(r - \hat{\eta}(x, a)) \, \right]^2 \right] \\ &= \underset{x,a,r}{\mathbb{E}} \left[\, \hat{w}(x, a)^2 (r - \hat{\eta}(x, a))^2 \, \right] - \underset{x}{\mathbb{E}} \left[\left(\sum_{a \in \mathcal{A}} \hat{w}(x, a)\mu(a \mid x)(\eta(x, a) - \hat{\eta}(x, a)) \right)^2 \right] \end{aligned}$$

For T_2 we have

$$T_2 = \operatorname*{Var}_x \left(\sum_{a \in \mathcal{A}} \pi(a \mid x)\hat{\eta}(x,a) + \hat{w}(x,a)\mu(a \mid x)(\eta(x,a) - \hat{\eta}(x,a)) \right)$$

Now, using our assumption that $0 \leq \hat{w}(x,a) \leq w(x,a) = \pi(a \mid x)/\mu(a \mid x)$, we get

$$|\, \hat{w}(x,a)\mu(a \mid x) \,| \leq \pi(a \mid x),$$

and hence for all $x \in \mathcal{X}$,

$$\sum_{a \in \mathcal{A}} \hat{w}(x,a)\mu(a \mid x)(\eta(x,a) - \hat{\eta}(x,a)) \leq \sum_{a \in \mathcal{A}} \pi(a \mid x) \cdot |\, \eta(x,a) - \hat{\eta}(x,a) \,| \leq 1.$$

Here in the last step we are using boundedness of the regression function η, and the estimated regression function $\hat{\eta}$.

Thus, for T_1 the second term is at most 1, and for T_2, we have

$$T_2 \leq 2\,\mathbb{E}\left[\left(\sum_{a \in \mathcal{A}} \pi(a \mid x)\hat{\eta}(x,a) \right)^2 \right] + 2\,\mathbb{E}\left[\left(\hat{w}(x,a)\mu(a \mid x)(\eta(x,a) - \hat{\eta}(x,a)) \right)^2 \right] \leq 4,$$

using the above calculation. Therefore, the residual terms add up to at most 5. □

Proof of Proposition 2. We analyze the optimistic version. If $\mathrm{MSE}(\hat{V}_{\mathrm{DM}}) \leq \mathrm{MSE}(\hat{V}_{\mathrm{DR}})$, then we simply take $\lambda = 0$ at which point the objective is clearly minimized by $\hat{w} = 0$. Therefore we recover $\hat{V}_{\mathrm{DM}}$. On the other hand if $\mathrm{MSE}(\hat{V}_{\mathrm{DR}}) \leq \mathrm{MSE}(\hat{V}_{\mathrm{DM}})$, then we set $\lambda \to \infty$ so that the minimizer is $\hat{w} = w$. This recovers the doubly-robust estimator. □

Proof of Theorem 3. The main technical part of the proof is a deviation inequality for the sample variance. For this, let us fix θ, which we drop from notation, and focus on estimating the variance

$$\mathrm{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}(z))^2] \text{ with } \widehat{\mathrm{Var}} = \frac{1}{2n(n-1)} \sum_{i \neq j = 1}^{n} (Z_i - Z_j)^2$$

We have the following lemma

Lemma 4 (Variance estimation)**.** *Let $Z_1, \ldots, Z_n$ be iid random variables, and assume that $|Z_i| \leq R$ almost surely. Then there exists a constant $C > 0$ such that for any $\delta \in (0,1)$, with probability at least $1 - \delta$*

$$\left| \mathrm{Var}(Z) - \widehat{\mathrm{Var}} \right| \leq (C+1) \left(\sqrt{\frac{2R^2 \, \mathrm{Var}(Z) \log(4C/\delta)}{(n-1)}} + \frac{4R^2 \log(4C/\delta)}{n(n-1)} \right).$$

Proof. For this lemma only, define $\mu = \mathbb{E}[Z]$. By direct calculation

$$\mathrm{Var}(Z) = \mathbb{E}\left[Z^2 \right] - \mu^2, \qquad \widehat{\mathrm{Var}} = \frac{1}{n} \sum_{i=1}^{n} Z_i^2 - \frac{1}{n(n-1)} \sum_{i \neq j} Z_i Z_j$$

We work with the second term first. Let $Z_1', \ldots, Z_n'$ be an iid sample, independent of $Z_1, \ldots, Z_n$. Now, by Theorem 3.4.1 of De la Pena and Giné [5], we have

$$\mathbb{P}\left[\left| \frac{1}{n(n-1)} \sum_{i \neq j} (Z_i - \mu)(Z_j - \mu) \right| > t \right] \leq C\mathbb{P}\left[\left| \frac{1}{n(n-1)} \sum_{i \neq j} (Z_i - \mu)(Z_j' - \mu) \right| > t/C \right]$$

for a universal constant $C > 0$. Thus, we have decoupled the U-statistic. Now let us condition on $Z_1, \ldots, Z_n$ and write $X_j = \frac{1}{n-1} \sum_{i \neq j} (Z_i - \mu)$, which conditional on $Z_1, \ldots, Z_n$ is non-random.

We will apply Bernstein's inequality on $\frac{1}{n}\sum_{j=1}^n X_j(Z_j' - \mu)$, which is a centered random variable, conditional on $Z_{1:n}$. This gives that with probability at least $1-\delta$

$$\left|\frac{1}{n}\sum_{j=1}^n X_j(Z_j' - \mu)\right| \le \sqrt{\frac{2\frac{1}{n}\sum_{j=1}^n \mathrm{Var}(X_j Z_j')\log(2/\delta)}{n}} + \frac{2\max_j \sup|\,X_j(Z_j-\mu)\,|\log(2/\delta)}{3n}.$$

$$\le \max_j |X_j| \left(\sqrt{\frac{2\,\mathrm{Var}(Z)\log(2/\delta)}{n}} + \frac{2R\log(2/\delta)}{3n}\right).$$

This bound holds with high probability for any $\{X_j\}_{j=1}^n$. In particular, since $|X_j| \le R$ almost surely, we get that with probability $1-\delta$

$$\left|\frac{1}{n(n-1)}\sum_{i\neq j}(Z_i-\mu)(Z_j-\mu)\right| \le C\sqrt{\frac{2R^2\,\mathrm{Var}(Z)\log(2C/\delta)}{n}} + \frac{2CR^2\log(2C/\delta)}{3n}.$$

The factors of C arise from working through the decoupling inequality.

Let us now address the first term, a simple application of Bernstein's inequality gives that with probability at least $1-\delta$

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i^2 - \mathbb{E}[Z^2]\right| \le \sqrt{\frac{2\,\mathrm{Var}(Z^2)\log(2/\delta)}{n}} + \frac{2R^2\log(2/\delta)}{3n}$$

$$\le \sqrt{\frac{2R^2\,\mathrm{Var}(Z)\log(2/\delta)}{n}} + \frac{2R^2\log(2/\delta)}{3n}.$$

Combining the two inequalities, we obtain the result. □

Since we are estimating the variance of the sample average estimator, we divide by another factor of n. Thus the error terms in Lemma 4 are $O(n^{-3/2})$ and $O(n^{-2})$ respectively, which are both $o(n^{-1})$. We will simply write these error terms are $o(n^{-1})$ from now on.

For the model selection result, first apply Lemma 4 for all $\theta \in \Theta$, taking a union bound (This only requires that $|\Theta| = o(\exp(n))$). Further take a union bound over the event that $\mathrm{Bias}(\theta) \le \mathrm{BiasUB}(\theta)$ for all $\theta \in \Theta$, if it is needed. Then, observe that for any $\theta_0 \in \Theta_0$ we have

$$\mathrm{MSE}(\hat\theta) = \mathrm{Bias}(\hat\theta)^2 + \mathrm{Var}(\hat\theta) \le \mathrm{BiasUB}(\theta)^2 + \widehat{\mathrm{Var}}(\hat\theta) + o(n^{-1})$$
$$\le \mathrm{BiasUB}(\theta_0)^2 + \widehat{\mathrm{Var}}(\theta_0) + o(n^{-1}) \le 0 + \mathrm{Var}(\theta_0) + o(n^{-1}) = \mathrm{MSE}(\theta_0) + o(n^{-1}).$$

The first inequality uses Lemma 4 and the fact that Bias $\le$ BiasUB. The second uses that $\hat\theta$ optimizes this quantity, and the third uses the property that $\mathrm{BiasUB}(\theta_0) = 0$ by assumption. □

B.1 Construction of bias upper bounds.

In this section we give detailed construction of bias upper bounds that we use in the model selection procedure. Recall that this is for the analysis only. Empirically we found that using the estimators alone — not the upper bounds — leads to better performance.

Throughout, we fix a set of hyperparameters θ, which we suppress from the notation.

Direct bias estimation. The most straightforward bias estimator is to simply approximate the expectation with a sample average.

$$\widetilde{\mathrm{Bias}} = \frac{1}{n}\sum_{i=1}^n (\,\hat w(x_i,a_i) - w(x_i,a_i)\,)\,(\,r_i - \hat\eta(x_i,a_i)\,)$$

This estimator has finite-sum structure, and naively, the range of each term is $w_\infty = \max_{x,a} w(x,a)$. The variance is at most $\mathbb{E}_\mu[w(x,a)^2]$. Hence Bernstein's inequality gives that with probability at least $1-\delta$

$$\left|\widetilde{\mathrm{Bias}} - \mathrm{Bias}\right| \le \sqrt{\frac{2\,\mathbb{E}_\mu[w(x,a)^2]\log(2/\delta)}{n}} + \frac{2w_\infty\log(2/\delta)}{3n}.$$

Inflating the estimate by the right hand side gives BiasUB, which is a high probability upper bound on Bias.

Pessimistic estimation. The bias bound used in the pessimistic estimator and its natural sample estimator are

$$\mathop{\mathbb{E}}_{\mu}[\,|\,\hat{w}(x,a) - w(x,a)\,|\,], \qquad \widetilde{\text{Bias}} = \frac{1}{n}\sum_{i}\sum_{a}\mu(a \mid x_i)\,|\,\hat{w}(x_i,a) - w(x_i,a)\,|\,.$$

Note that since we have already eliminated the dependence on the reward, we can analytically evaluate the expectation over actions, which will lead to lower variance in the estimate.

Again we perform a fairly naive analysis. Since $\hat{w}(x,a) \le w(x,a)$ and using the fact that $w(x,a) = \pi(a \mid x)/\mu(a \mid x)$ the range of the random variable is simply 1. Therefore, Hoeffding's inequality gives that with probability $1-\delta$

$$\text{Bias} \le \widetilde{\text{Bias}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

and we use the right hand side for our high probability upper bound.

Optimistic estimation. For the optimistic bound, we must estimate two terms, one involving the regressor and one involving the importance weights. We use

$$T_1 := \frac{1}{n}\sum_{i=1}^{n} z(x_i,a_i)(r_i - \hat{\eta}(x_i,a_i))^2, \qquad T_2 := \frac{1}{n}\sum_{i=1}^{n}\sum_{a}\mu(a \mid x_i)(\hat{w}(x_i,a) - w(x_i,a))^2/z(x_i,a).$$

Note here that the former uses sampled actions from μ, but does not involve the importance weight, while the latter involves the importance weight but analytically evaluates the expectation over μ. Thus we can expect that both are fairly low variance.

For both, we use Bernstein's inequality. For T_1 the range is naively $z_\infty = \max_{x,a} z(x,a)$ and the variance is $\mathbb{E}_\mu[z(x,a)^2]$. Thus we get that with probability at least $1-\delta/2$

$$\mathbb{E}\left[\,z(x,a)(r - \hat{\eta}(x,a))^2\,\right] \le T_1 + \sqrt{\frac{2\,\mathbb{E}_\mu[z(x,a)^2]\log(2/\delta)}{n}} + \frac{2z_\infty\log(2/\delta)}{3n}$$

For the second one, the range is $\max_{x,a} w(x,a)/z(x,a)$ and we just use Hoeffding's inequality, so that with probability $1-2\delta$

$$\mathop{\mathbb{E}}_{\mu}\left[\,(\hat{w}(x,a) - w(x,a))^2/z\,\right] \le T_2 + \sqrt{\frac{\max_{x,a}\frac{w(x,a)}{z(x,a)}\log(2/\delta)}{2n}}$$

The high probability upper bound follows by multiplying the two right hand sides together and taking square root.

C Experimental Details and Additional Results

Dataset statistics. We use datasets from the UCI Machine Learning Repository [6]. Dataset statistics are displayed in Table 3.

Hyperparameter grid. For our shrinkage estimators and SWITCH, we choose the shrinkage coefficients from a grid of 30 geometrically spaced values. For the pessimistic estimator and SWITCH, the largest and smallest values in the grid are the 0.05 quantile and 0.95 quantile of the importance weights. For the optimistic estimator, the largest and smallest values are $0.01 \times (w_{0.05})^2$ and $100 \times (w_{0.95})^2$ where $w_{0.05}$ and $w_{0.95}$ are the 0.05 and 0.95 quantile of the importance weights.

For the off-policy learning experiments, we only consider the shrinkage coefficients in $\{0.0, 0.1, 1, 10, 100, 1000, \infty\}$ during training, while for model selection, we use the same grid as in the evaluation experiments.

MRDR. Farajtabar et al. [9] propose training the regression model with a specific choice of weighting z, which we also use in our experiments. When the evaluation policy π is deterministic, they set $z(x,a) = \mathbf{1}\{\pi(x) = a\} \cdot \frac{1-\mu(a|x)}{\mu(a|x)^2}$. For stochastic policies, following the implementation of Farajtabar et al., we sample $a_i \sim \pi(\cdot \mid x_i)$ for each example in the dataset used to train the reward predictor. Then we proceed as if the evaluation policy deterministically chooses a_i on example x_i.

Dataset	Glass	Ecoli	Vehicle	Yeast	PageBlok	OptDigits	SatImage	PenDigits	Letter
Actions	6	8	4	10	5	10	6	10	26
Examples	214	336	846	1484	5473	5620	6435	10992	20000

Table 3: Dataset statistics.

Ablation study for deterministic target policy. Since MRDR is more suited to deterministic policies, we also report the results of our regressor and shrinkage ablations for a deterministic target policy $\pi_{1,\text{det}}$ in Table 4. As with the stochastic policies, the estimator influences the choice of reward predictor, but note that $z = 1$ and $z = w$ are more favorable here. This is likely due to high variance suffered from training with $z = w^2$, because the importance weights are larger with a deterministic policy. Our shrinkage ablation reveals that both estimator types are important also when the target policy is deterministic.

Ablation study for model selection. In Table 5, we show the comparison of different model selection methods under different reward predictor pairs and different shrinkage types. In most cases, Dir-all (DRs-direct where the bias bound is estimated as the pointwise minimum of (1) the bias, (2) the optimistic bound and (3) the pessimistic bound) and Up-all (all bias estimates are adjusted by adding twice standard error before taking pointwise minimum) are most frequently statistically indistinguishable from the best, which suggests that our proposed bias estimate (by taking pointwise minimum of the three) is robust and adaptive.

Comparisons across additional experimental conditions. In Figure 4 and Figure 5, we compare our new estimators, DRs-direct and DRs-upper, with baselines across various conditions (apart from deterministic versus stochastic rewards from the main paper). We first investigate the performance under *friendly logging* (logging and evaluation policies are derived from the same deterministic policy, $\pi_{1,det}$), *adversarial logging* (logging and evaluation policies are derived from different policies $\pi_{1,det}$, $\pi_{2,det}$), and *uniform logging* (logging policy is uniform over all actions). Then we plot the performance in the small sample regime, where we aggregate the 108 conditions (6 logging policies, 9 datasets, deterministic/stochastic reward) at just 200 bandit samples.

Comparisons across all reward predictors. In Table 6–Table 9, we compare the performance of DRs-direct and DRs-upper against baselines across various choices of reward predictors. Specifically, we consider cases when DRs is employing model selection to pick from two predictors and construct a separate comparison table for each of the following sets: $\{\hat{\eta} \equiv 0, \text{MRDR}\}$, $\{\hat{\eta} \equiv 0, z \equiv 1\}$, $\{\hat{\eta} \equiv 0, z = w\}$ as well as $\{\hat{\eta} \equiv 0, z = w^2\}$ (matching the setting of the main paper). We report the number of conditions where each estimator is statistically indistinguishable from the best, and the number of conditions where each estimator statistically dominates all others. DRs-upper is most often in the top group and most often the unique winner. DRs-direct is also better than snIPS, snDR, and SWITCH. These results suggest that our shrinkage estimators are robust to different choices of reward predictors, and not just limited to the recommended set $\{\hat{\eta} \equiv 0, z = w^2\}$.

Robustness of DRs-direct **and** DRs-upper **(w.r.t. inclusion of more reward predictors)** In Figure 6, we test the robustness of our proposed methods as we incorporate more reward predictors. Our practical suggestions is to use $\{\hat{\eta} \equiv 0, z = w^2\}$ (shown as DRs-direct and DRs-upper in the figure). Here we also evaluate these methods when selecting from all reward predictors in the set $\{\hat{\eta} \equiv 0, z \equiv 1, z = w, z = w^2, \text{MRDR}\}$ (shown as DRs-direct (all) and DRs-upper (all) in the figure). For DRs-direct, the curves almost match, suggesting that it is quite robust. However, DRs-upper is less robust to including additional reward predictors.

	$\hat{\eta} \equiv 0$	$z \equiv 1$	$z = w$	$z = w^2$	MRDR
DM	0 (0)	54 (30)	59 (23)	35 (4)	24 (6)
DR	28 (1)	94 (11)	85 (0)	85 (1)	85 (0)
snDR	65 (7)	86 (7)	79 (0)	72 (0)	71 (0)
DRs	14 (9)	51 (17)	65 (14)	54 (6)	47 (4)

	DRps	DRos
$\hat{\eta} \equiv 0$	13	59
$z \equiv 1$	29	55
$z = w$	26	66
$z = w^2$	30	67
MRDR	29	63

Table 4: Ablation analysis for *deterministic* target policy $\pi_{1,\text{det}}$ across experimental conditions.Left: we compare reward predictors using a fixed estimator (with oracle tuning if applicable). We report the number of conditions where a regressor is statistically indistinguishable from the best and, in parenthesis, the number of conditions where it statistically dominates all others. Right: we compare different shrinkage types using a fixed reward predictor (with oracle tuning) reporting the number of conditions where one statistically dominates the other.

	Dir-all	Dir-naive	Dir-opt	Dir-pes	Up-all	Up-naive	Up-opt	Up-pes
0-pes	71	67	74	78	63	60	79	79
0-opt	75	68	71	81	64	62	66	81
0-best	63	59	67	74	56	54	64	76
w^2-pes	47	41	5	3	72	71	5	3
w^2-opt	47	40	3	2	73	70	3	2
w^2-best	47	42	4	3	75	72	4	3
best-pes	51	46	7	5	69	70	6	5
best-opt	49	46	7	3	69	70	5	3
best-best	50	46	8	4	71	70	6	4
all-best	49	46	7	6	65	67	6	6

Table 5: Comparison of model selection methods when paired with different reward predictor sets and shrinkage types. As in other tables, we record the number of conditions in which this model selection method is statistically indistinguishable from the best, for fixed reward predictor set and shrinkage types. Columns are indexed by model selection methods, "Dir" denotes taking sample average and "Up" denotes inflating sample averages with twice the standard error. "Naive" denotes directly estimating bias, "opt" denotes estimating optimistic bias bound, "pes" denotes pessimistic bias bound, and "all" denotes taking the pointwise minimum of all three. Rows are indexed by reward predictors: $\hat{\eta} \equiv 0$, $z = w^2$, "best" denotes selecting over both, and "all" denotes selecting over these and additionally $z = 1$, $z = w$, and MRDR. Rows are also indexed by shrinkage type, optimistic, pessimistic, and best, which denotes model selection over both.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	5	36	5	4	43	63
Unique Best	0	15	0	0	13	46

Table 6: Significance testing for different estimators across all conditions (DM, snDR use reward $z = w^2$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = w^2\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

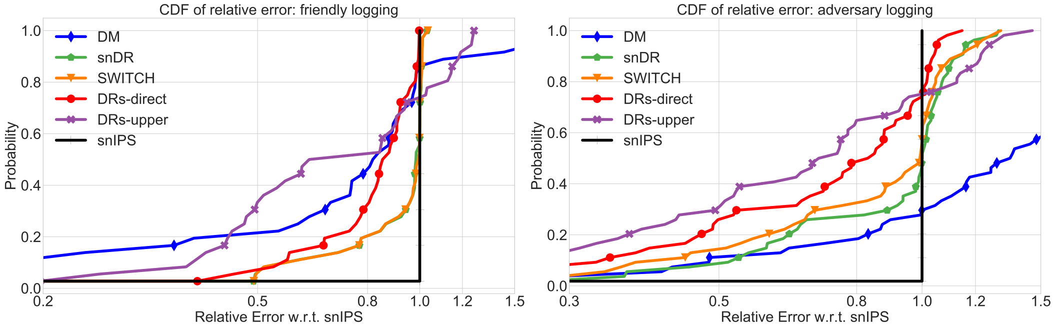

Figure 4: CDF plots of normalized MSE aggregated across all conditions with friendly scenario (left) and adversary scenario (right).

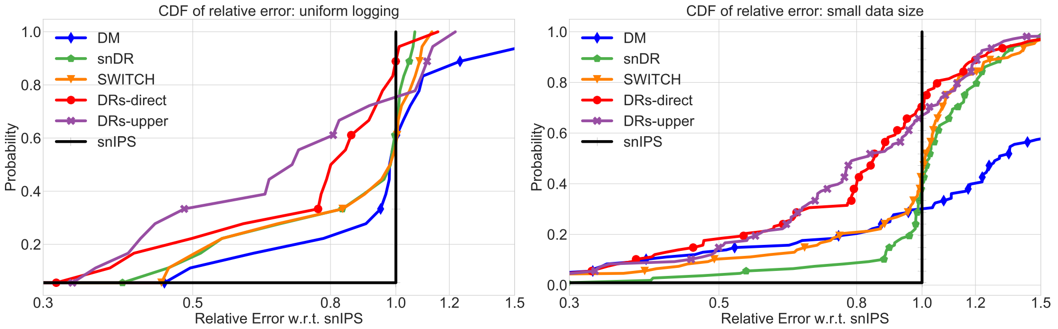

Figure 5: CDF plots of normalized MSE aggregated across all conditions with uniform logging policy scenario (left) and small data regime (right).

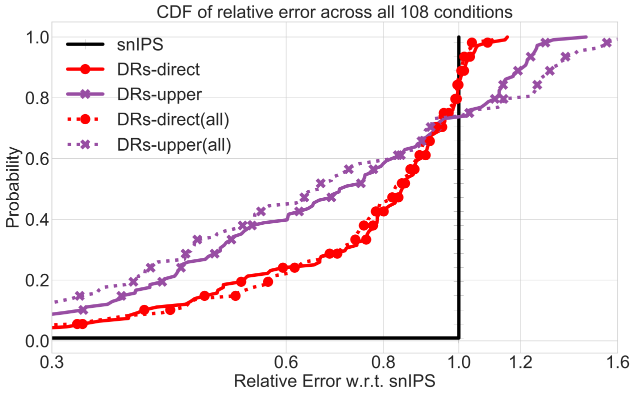

Figure 6: Robustness test for DRs-direct and DRs-upper. DRs-direct (all) and DRs-upper (all) means the corresponding method with reward predictor select from all possible cases $\{\hat{\eta} = 0, z = 1, z = w, z = w^2\}$ and MRDR.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	10	17	8	10	37	73
Unique Best	0	7	0	0	17	57

Table 7: Significance testing for different estimators across all conditions (DM, snDR use reward estimated from MRDR, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, \text{MRDR}\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	10	39	19	15	25	49
Unique Best	0	23	3	1	5	43

Table 8: Significance testing for different estimators across all conditions (DM, snDR use reward $z = 1$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = 1\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	3	44	5	5	30	58
Unique Best	0	23	0	0	6	51

Table 9: Significance testing for different estimators across all conditions (DM, snDR use reward $z = w$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = w\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.