Doubly robust off-policy evaluation with shrinkage

Yi Su [1] Maria Dimakopoulou [2] Akshay Krishnamurthy [3] Miroslav Dudík [3]

Abstract

We propose a new framework for designing estimators for off-policy evaluation in contextual bandits. Our approach is based on the asymptotically optimal doubly robust estimator, but we shrink the importance weights to minimize a bound on the mean squared error, which results in a better bias-variance tradeoff in finite samples. We use this optimization-based framework to obtain three estimators: (a) a weight-clipping estimator, (b) a new weight-shrinkage estimator, and (c) the first shrinkage-based estimator for combinatorial action sets. Extensive experiments in both standard and combinatorial 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 (Li et al., 2011), advertising (Bottou et al., 2013), and search engines (Li et al., 2015) to personalized healthcare (Zhou et al., 2017), are naturally modeled by the *contextual bandit* protocol (Langford & Zhang, 2008), 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 the user's click on the article. The goal is to maximize the reward, but the learner can only observe the reward for chosen actions, and not for the others.

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 (Dudík et al., 2011).

The most accurate approaches to off-policy evaluation are variants of *doubly robust* (DR) estimators (Robins & Rotnitzky, 1995; Bang & Robins, 2005; Dudík et al., 2011). DR estimation begins by fitting a regression model to predict rewards as a function of context and action. The fitted model can be used to impute unobserved rewards of the target policy on the training data, but such a direct estimate is typically biased. Instead, DR adds a correction term obtained by importance weighting the difference between observed rewards and predicted rewards. The resulting approach is unbiased, and it is asymptotically optimal under weaker assumptions than other methods (Rothe, 2016). However, its finite-sample variance can still be quite high when importance weights (also known as inverse propensity scores) are large. Therefore, several works have developed variants of DR that clip or remove large importance weights. Although weight clipping incurs some bias, it substantially decreases the variance and can yield a lower mean squared error (Bembom & van der Laan, 2008; Bottou et al., 2013; Wang et al., 2017; Su et al., 2018). These works motivate weight shrinkage as a heuristic for trading off bias and variance, but they do not provide insight into when and how these different methods should be used.

In this paper, we ask: *What are the systematic strategies for shrinking importance weights?* We seek to answer this question without making strong assumptions about the quality of the reward predictor, but we would like to adapt to its quality. We make the following contributions:

- We derive a general framework for shrinking the importance weights by optimizing a sharp bound on the mean squared error (MSE). We use two bounding techniques. The first is agnostic to the quality of the reward estimator and yields *pessimistic shrinkage* estimators. The second incorporates the quality of the reward predictor and yields *optimistic shrinkage* estimators.
- We provide theoretical justification for the standard practice of weight clipping by showing that it corresponds to pessimistic shrinkage.
- Using optimistic shrinkage, we derive new estimators, which are also applicable to *combinatorial actions*, arising, for example, when a news portal is recommending

[1] Cornell University, Ithaca, NY [2] Netflix, Los Gatos, CA [3] Microsoft Research, New York, NY. Correspondence to: Yi Su <ys756@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

not just a single article, but a list of articles (Cesa-Bianchi & Lugosi, 2012; Swaminathan et al., 2017).

Apart from the conceptual and theoretical contributions above, we also carry out an extensive empirical evaluation. For atomic (i.e., non-combinatorial) actions, we consider 108 experimental conditions derived from 9 real-world data sets and covering a range of data set sizes, feature dimensions, policy overlap (i.e., the magnitude of importance weights), and quality of reward estimators. For combinatorial actions, we consider a standard learning-to-rank data set and vary the quality of reward estimators. In all instances, we demonstrate the efficacy of our shrinkage approach. Via extensive ablation studies, we also identify a robust configuration of our shrinkage approach that we recommend as a practical choice.

Comparison with 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 (Hahn, 1998; Hirano et al., 2003; Imbens et al., 2007; Rothe, 2016), but only few that optimize MSE in finite samples. Most notably, Kallus (2017; 2018) develops the *kernel optimal matching* (KOM) approach that adjusts importance weights by optimizing MSE under smoothness (or parametric) assumptions on the reward function. This method is reminiscent of direct modeling, whose bias can be bounded under smoothness assumptions, but whose performance deteriorates if these assumptions are violated. In contrast, we optimize importance weights with essentially no modeling assumptions. Another difference is that KOM runs in time that is super-linear in the data set size, which prevents its use with large data sets, whereas our approach requires a single pass through the data and readily applies to large-scale scenarios.

Several recent works study how to improve DR estimators under similar assumptions as we make here (Wang et al., 2017; Farajtabar et al., 2018; Su et al., 2018), focusing either on weight shrinkage or on training of the reward predictor. However, to our knowledge, we are the first to provide a detailed theoretical and empirical investigation of the interplay between these two design components. For example, in Table 2, we show that the *more robust doubly robust* (MRDR) approach for training of the reward predictor (Farajtabar et al., 2018) performs poorly in combination with weight shrinkage. More generally, different estimators may require different reward predictors. This specific finding has practical implications that are missing in prior work.

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 $\eta(x, a) := \mathbb{E}\,[\,r \mid x, a\,]$ for the expected reward, conditioned on a given context and action.

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 $\pi(x, a, r) := D(x)\pi(a \mid x)D(r \mid x, a)$ 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 $V(\pi) := \mathbb{E}_{(x,a,r)\sim\pi}[r]$.

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:

$$\begin{aligned}\mathrm{MSE}\big(\hat{V}(\pi)\big) &= \mathrm{Bias}\big(\hat{V}(\pi)\big)^2 + \mathrm{Var}\big[\hat{V}(\pi)\big],\\ \mathrm{Bias}\big(\hat{V}(\pi)\big) &:= \Big|\mathbb{E}\big[\hat{V}(\pi) - V(\pi)\big]\Big|.\end{aligned}$$

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

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

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 ensures that the importance weights are well defined and $\hat{V}_{\mathrm{IPS}}(\pi)$ is an unbiased estimator of $V(\pi)$. If there is a 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, but it can be biased due to approximation errors in fitting $\hat{\eta}$.

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

$$\begin{aligned}&\hat{V}_{\text{DR}}(\pi;\hat{\eta})\\ &\quad:= \hat{V}_{\text{DM}}(\pi;\hat{\eta}) + \frac{1}{n}\sum_{i=1}^{n} w(x_i,a_i)\big(r_i - \hat{\eta}(x_i,a_i)\big). \quad (1)\end{aligned}$$

The DR estimator applies IPS to a shifted reward, using $\hat{\eta}$ as a control variate to decrease the variance of IPS, while preserving its unbiasedness. DR is asymptotically optimal, as long as it is possible to derive sufficiently good reward predictors $\hat{\eta}$ given enough data (Rothe, 2016).

However, even when the reward predictor $\hat{\eta}$ is perfect, 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. As mentioned in Section 1, several approaches seek a more favorable bias–variance trade-off by shrinking the importance weights. 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 $\hat{V}_{\text{DR}}$. 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

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

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 $\hat{w}$:

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

We assume that $0 \le \hat{w} \le w$, 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 $\hat{w} = w$:

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

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* $0 \le \hat{w} \le w$ *then*

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

See appendix for the proof of Proposition 1 (as well as other mathematical statements from this paper).

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 Section 5 we derive 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

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

The loss $L(\hat{\eta})$ quantifies the quality of $\hat{\eta}$. We use it to bound the bias by applying the Cauchy–Schwarz inequality to (4):

$$\begin{aligned}\text{Bias}(\hat{w}) \le &\sqrt{\mathbb{E}_{\mu}\Big[\tfrac{1}{z(x,a)}\big(\hat{w}(x,a) - w(x,a)\big)^2\Big]}\\ &\cdot\sqrt{L(\hat{\eta})}. \quad (5)\end{aligned}$$

To bound the variance, we invoke Proposition 1 and focus on bounding the quantity $\mathbb{E}_\mu\left[\hat{w}^2(r-\hat{\eta})^2\right]$:

$$\begin{aligned}
&\mathbb{E}_\mu\left[\hat{w}^2(x,a)\bigl(r-\hat{\eta}(x,a)\bigr)^2\right] \\
&\quad\le \sqrt{\mathbb{E}_\mu\left[\tfrac{1}{z(x,a)}\hat{w}^4(x,a)\right]}\sqrt{\mathbb{E}_\mu\left[z(x,a)\bigl(r-\hat{\eta}(x,a)\bigr)^4\right]} \\
&\quad\le \sqrt{\mathbb{E}_\mu\left[\tfrac{w^2(x,a)}{z(x,a)}\hat{w}^2(x,a)\right]}\sqrt{L(\hat{\eta})}, \qquad (6)
\end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality, and the second from the fact that $\hat{w}^2(x,a)\le w^2(x,a)$ and $|r-\hat{\eta}(x,a)|\le 1$.

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

$$\begin{aligned}
\mathrm{MSE}(\hat{w}) \le \mathbb{E}_\mu&\left[\tfrac{1}{z(x,a)}\bigl(\hat{w}(x,a)-w(x,a)\bigr)^2\right]L(\hat{\eta}) \\
&+\sqrt{\mathbb{E}_\mu\left[\tfrac{w^2(x,a)}{z(x,a)}\hat{w}^2(x,a)\right]}\sqrt{L(\hat{\eta})}+\frac{1}{n}.
\end{aligned}$$

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 $\lambda\in[0,\infty]$, it can be obtained by minimizing

$$\lambda\mathbb{E}_\mu\left[\tfrac{1}{z(x,a)}\bigl(\hat{w}(x,a)-w(x,a)\bigr)^2\right]+\mathbb{E}_\mu\left[\tfrac{w^2(x,a)}{z(x,a)}\hat{w}^2(x,a)\right]$$

with respect to $\hat{w}$. This objective decomposes across contexts and actions. Taking the derivative with respect to $\hat{w}(x,a)$ and setting it 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 $\hat{V}_{\mathrm{DRos}}(\pi;\hat{\eta},\lambda)$. Note that this estimator does not depend on z, although it was included in the optimization objective. When $\lambda=0$, we have $\hat{w}(x,a)=0$ corresponding to DM. As $\lambda\to\infty$, the weights increase and in the limit become equal to $w(x,a)$, corresponding to standard DR.

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 bounds

$$\mathrm{Bias}(\hat{w})\le\mathbb{E}_\mu\left[|\hat{w}(x,a)-w(x,a)|\right], \qquad (7)$$

$$\mathbb{E}_\mu\left[\hat{w}(x,a)^2(r-\hat{\eta}(x,a))^2\right]\le\mathbb{E}_\mu\left[\hat{w}(x,a)^2\right]. \qquad (8)$$

As before, we do not optimize the resulting MSE bound directly and instead solve for the Pareto front points parameterized by $\lambda\in[0,\infty]$ (we scale λ by a factor of two to obtain the solution that more cleanly matches the clipping estimator):

$$\operatorname*{Minimize}_{\hat{w}}\ 2\lambda\mathbb{E}_\mu\left[|\hat{w}(x,a)-w(x,a)|\right]+\mathbb{E}_\mu\left[\hat{w}(x,a)^2\right].$$

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 (Kang et al., 2007; Strehl et al., 2010; Su et al., 2018) (see Appendix A for detailed calculations). We refer to the resulting estimator as $\hat{V}_{\mathrm{DRps}}(\pi;\hat{\eta},\lambda)$, for *doubly robust with pessimistic shrinkage*. Similarly to optimistic shrinkage, we recover DM for $\lambda=0$, and DR as $\lambda\to\infty$.

4. Shrinkage for Combinatorial Actions

We showcase the generality of our optimization-based approach by deriving a shrinkage estimator for *combinatorial actions* (also called *slates*), which arise, for example, when recommending a ranked list of items.

In contextual combinatorial bandits, the actions are represented as vectors $\mathbf{a}\in\mathbb{R}^d$ for some dimension d and the action space $\mathcal{A}\subseteq\mathbb{R}^d$ is typically exponentially large in d.

Example 1 (Ranking and NDCG)**.** Consider the task of recommending a ranked list of items such as images or web pages. The context x is the query submitted by a user together with a user profile. The action $\mathbf{a}$ represents a ranked list of ℓ items out of m. The list $(i_1,\dots,i_\ell)$, where $i_j\in\{1,\dots,m\}$, is encoded into an action vector $\mathbf{a}\in\{0,1\}^{\ell m}$ via *ℓ-hot encoding*, i.e., we split $\mathbf{a}$ into ℓ blocks of size m, and in the block j we set the i_j-th coordinate to 1 and all others to 0. As a reward we use a standard information-retrieval metric called the *normalized discounted cumulative gain*, defined as $\mathrm{NDCG}(x,\mathbf{a}):=\mathrm{DCG}(x,\mathbf{a})/\mathrm{DCG}^\star(x)$, where $\mathrm{DCG}(x,\mathbf{a}):=\sum_{j=1}^{\ell}\frac{2^{\mathrm{rel}(x,i_j)}-1}{\log_2(j+1)}$, $\mathrm{DCG}^\star(x):=\max_{\mathbf{a}'}\mathrm{DCG}(x,\mathbf{a}')$, and $\mathrm{rel}(x,i)$ is some intrinsic measure of item relevance. (See, e.g., Swaminathan et al., 2017.)

Standard importance weighting techniques, such as DR and IPS, can fail dramatically in the combinatorial setting, because their variance scales linearly with the size of $\mathcal{A}$, which is typically exponential in d. However, if the expected reward is linear in $\mathbf{a}$, i.e., $\eta(x,\mathbf{a})=\boldsymbol{\eta}(x)^\top\mathbf{a}$ for some (unknown) function $\boldsymbol{\eta}:\mathcal{X}\to\mathbb{R}^d$, then it is possible to achieve variance polynomial in d using the *pseudo-inverse* estimator of Swaminathan et al. (2017). Given a reward predictor $\hat{\boldsymbol{\eta}}:\mathcal{X}\to\mathbb{R}^d$, it is also possible to obtain the DR variant of this estimator (DR-PI):

$$\hat{V}_{\text{DR-PI}}(\pi;\hat{\boldsymbol{\eta}}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\eta}}_i^\top \mathbf{q}_{\pi,x_i} + \mathbf{w}_i^\top \mathbf{a}_i (r_i - \hat{\boldsymbol{\eta}}_i^\top \mathbf{a}_i), \quad (9)$$

where $\hat{\boldsymbol{\eta}}_i \coloneqq \hat{\boldsymbol{\eta}}(x_i)$, $\mathbf{q}_{\pi,x_i} \coloneqq \mathbb{E}_\pi[\mathbf{a} \mid x_i]$, $\mathbf{w}_i \coloneqq \Gamma^\dagger_{\mu,x_i}\mathbf{q}_{\pi,x_i}$, $\Gamma_{\mu,x_i} \coloneqq \mathbb{E}_\mu[\mathbf{a}\mathbf{a}^\top \mid x_i]$, and † denotes the matrix pseudo-inverse. The vector $\mathbf{w}_i$ plays the role of the importance weight, while the first term corresponds to the direct modeling approach. Swaminathan et al. (2017) establish that this estimator is unbiased when $\eta(x,\mathbf{a})$ is linear in $\mathbf{a}$ and $\operatorname{span}\big(\operatorname{supp}\pi(\cdot \mid x)\big) \subseteq \operatorname{span}\big(\operatorname{supp}\mu(\cdot \mid x)\big)$, which is a linear relaxation of absolute continuity. Note that NDCG in Example 1 satisfies the linearity assumption.

We next derive the shrunk variant of DR-PI, following the optimistic bounding technique from Section 3.1. A formal difference is that we seek a vector-valued map $\hat{\mathbf{w}}: \mathcal{X} \to \mathbb{R}^d$. Since $\mathbf{w}(x)^\top\mathbf{a}$ can be negative, we formalize the shrinkage property as $\hat{\mathbf{w}}(x)^\top\mathbf{a} = c(x,\mathbf{a})\mathbf{w}(x)^\top\mathbf{a}$ for some $c(x,\mathbf{a}) \in [0,1]$. Also, analogously to non-combinatorial setup, we assume that $\hat{\boldsymbol{\eta}}(x)^\top\mathbf{a} \in [0,1]$ for all $\mathbf{a}$. Now all the steps from Section 3.1, except for Proposition 1 (to which we return below), go through under substitution $w(x,\mathbf{a}) = \mathbf{w}(x)^\top\mathbf{a}$, $\hat{w}(x,\mathbf{a}) = \hat{\mathbf{w}}(x)^\top\mathbf{a}$, and $\hat{\eta}(x,\mathbf{a}) = \hat{\boldsymbol{\eta}}(x)^\top\mathbf{a}$. The resulting (optimistic) shrinkage estimator takes form

$$\hat{V}_{\text{DRos-PI}}(\pi;\hat{\boldsymbol{\eta}},\lambda)$$
$$\coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\eta}}_i^\top \mathbf{q}_{\pi,x_i} + \frac{\lambda \mathbf{w}_i^\top\mathbf{a}_i}{\lambda + (\mathbf{w}_i^\top\mathbf{a}_i)^2}(r_i - \hat{\boldsymbol{\eta}}_i^\top\mathbf{a}_i). \quad (10)$$

The detailed derivation is in Appendix B. To our knowledge this is the first weight-shrinkage estimator for contextual combinatorial bandits.

To finish the section, we derive a combinatorial variant of Proposition 1, establishing a tight, but simple-to-optimize proxy for the variance of a DR-PI. This requires an additional assumption that for each x, the logging policy is supported on a linearly independent set of actions $\mathcal{B}_x \subseteq \mathcal{A}$; this requirement is typically easy to satisfy in practice (see, e.g., Section 6.2). We write $B_x \in \mathbb{R}^{d\times|\mathcal{B}_x|}$ for the matrix with columns $\mathbf{a} \in \mathcal{B}_x$, and $\mathbf{v}_{\pi,x}$ for the unique vector such that $B_x\mathbf{v}_{\pi,x} = \mathbf{q}_{\pi,x}$. Finally, let $\operatorname{Var}(\hat{\mathbf{w}})$ denote the variance of a DR-PI estimator with the shrunk weight map $\hat{\mathbf{w}}$.

Proposition 2. *Assume that $\mu(\cdot \mid x)$ is supported on a linearly independent set of actions for every x. If $\hat{\mathbf{w}}(x)^\top\mathbf{a} = c(x,\mathbf{a})\mathbf{w}(x)^\top\mathbf{a}$ for some $c(x,\mathbf{a}) \in [0,1]$, then*

$$\left|\operatorname{Var}(\hat{\mathbf{w}}) - \frac{1}{n}\mathbb{E}_\mu\Big[(\hat{\mathbf{w}}^\top\mathbf{a})^2\big(r - \hat{\boldsymbol{\eta}}^\top\mathbf{a}\big)^2\Big]\right| \le \frac{1}{n}\mathbb{E}_x[\|\mathbf{v}_{\pi,x}\|_1^2].$$

Note that the quantity $\|\mathbf{v}_{\pi,x}\|_1$ on the right-hand side only depends on the set $\mathcal{B}_x$, but not on the probabilities with which μ chooses $\mathbf{a} \in \mathcal{B}_x$. Non-combinatorial setting of Section 3 is a special case of the linearly independent setting, where $d = |\mathcal{A}|$ and actions are represented by standard basis vectors. In this case, $\|\mathbf{v}_{\pi,x}\|_1 = 1$ and we recover Proposition 1. We can always select $\mathcal{B}_x$ to be an (approximate) *barycentric spanner* and achieve $\|\mathbf{v}_{\pi,x}\|_1 = O(d)$ (Awerbuch & Kleinberg, 2008; Dani et al., 2008).

5. Model Selection

All of our shrinkage estimators have hyperparameters which we condense into a tuple θ. For example $\theta = (\hat{\eta}, \mathrm{o}, \lambda)$ denotes that we are using a reward predictor $\hat{\eta}$ and optimistic shrinkage with the parameter λ. To select among these hyperparameters, we propose and analyze a simple model selection procedure.

Let $\hat{V}_\theta$ denote the estimator parameterized by θ. We consider the procedure that estimates the variance of $\hat{V}_\theta$ by sample variance $\widehat{\operatorname{Var}}(\theta)$, and bounds the bias of $\hat{V}_\theta$ by a data-dependent upper bound $\operatorname{BiasUB}(\theta)$. The only requirement is that for all θ, $\operatorname{Bias}(\theta) \le \operatorname{BiasUB}(\theta)$ (with high probability), and that $\operatorname{BiasUB}(\theta) = 0$ whenever $\operatorname{Bias}(\theta) = 0$; this holds for both bias bounds from Section 3, as they become zero when $\hat{w} = w$. Now, to choose θ from a set of hyperparameters Θ, we optimize the estimate of the MSE:

$$\hat{\theta} \leftarrow \operatorname*{Minimize}_{\theta\in\Theta}\ \operatorname{BiasUB}(\theta)^2 + \widehat{\operatorname{Var}}(\theta).$$

The next theorem shows that this procedure always compares favorably with all the unbiased estimators included in Θ, up to an asymptotically negligible term $O(n^{-3/2})$. In particular, the procedure is asymptotically optimal whenever Θ includes a standard (non-shrunk) DR.

Theorem 3. *Let Θ be a finite set of hyperparameter values and let $\Theta_0 \coloneqq \{\theta \in \Theta : \operatorname{Bias}(\theta) = 0\}$ denote the subset of unbiased estimators. Assume that with probability $1 - \delta/2$ we have $\operatorname{Bias}(\theta) \le \operatorname{BiasUB}(\theta)$ for all $\theta \in \Theta$. Then there exists a universal constant C such that with probability at least $1-\delta$ we have*

$$\operatorname{MSE}(\hat{\theta}) \le \min_{\theta_0\in\Theta_0} \operatorname{MSE}(\theta_0) + C\log(|\Theta|/\delta)/n^{3/2}.$$

There are many strategies to construct data-dependent bias bounds with the required properties. The three bounds in our experiments take form of sample averages that 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 adjusted variants where we add twice the standard error.

Our model selection procedure is related to MAGIC (Thomas & Brunskill, 2016) as well as the procedure for the

SWITCH estimator (Wang et al., 2017). Unlike MAGIC, we pick a single hyperparameter value θ rather than aggregating several, and we use different bias and variance estimates. SWITCH uses our pessimistic bias bound (7), but with no theoretical justification. We use two additional bounding strategies, which are empirically shown to help, and provide theoretical justification in the form of an oracle inequality.

6. 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.

6.1. Non-combinatorial Setting

Datasets. Following prior work (Dudík et al., 2014; Wang et al., 2017; Farajtabar et al., 2018; Su et al., 2018), 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 4 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 $y \sim \mu(\cdot \mid x)$ and then observing a deterministic or stochastic reward r. The value of n varies across experimental conditions.

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. (2018). 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.

Reward predictors. We obtain reward predictors $\hat{\eta}$ by training linear models via weighted least squares with ℓ_2 regularization. We consider weights $z(x, a) \in \{1, w(x, a), w^2(x, a)\}$ as well as the *more robust doubly robust*, or MRDR, weight design of Farajtabar et al. (2018) (see Appendix D). 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 $\hat{\eta} \equiv 0$. The remaining bandit data is used to calculate the value of each estimator.

Baselines. We include a number of estimators in our evaluation: the direct modeling approach (DM), doubly-robust approach (DR) and its self-normalized variant (snDR), our approach (DRs), and the doubly-robust version of the SWITCH estimator of Wang et al. (2017), which also performs a form of weight clipping.[1] 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 D for details about the hyperparameter grid).

6.1.1. 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 (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. We conduct two ablation studies: one evaluating different reward predictors and the other evaluating the optimistic and pessimistic shrinkage types.

In Table 2, for each fixed estimator type (e.g., DR) we evaluate each reward predictor 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 predictors. For DRs we use oracle tuning for the shrinkage type and coefficient λ. 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 choice. In our comparison experiments below,

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

Table 1. Policy parameters used in the experiments.

	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

Table 2. Comparison of reward predictors using a fixed estimator (with oracle tuning if applicable); reporting the number of conditions where a regressor is statistically as good as the best and, in parenthesis, the number of conditions where it statistically dominates all others.

	$\hat{\eta} \equiv 0$	$z \equiv 1$	$z = w$	$z = w^2$	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)

Table 3. Comparison of shrinkage types using a fixed reward predictor (with oracle tuning); reporting the number of conditions where one statistically dominates the other.

	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

we run each method with its best reward predictor: DM with $z \equiv 1$, snDR with $z = w$, and DRs and SWITCH with $z = w^2$. For DRs and SWITCH, we additionally also consider $\hat{\eta} \equiv 0$, because it allows including IPS as their special case. Somewhat surprisingly, in our experiments, MRDR is dominated by other reward predictors (except for $\hat{\eta} \equiv 0$), and this remains true even with a deterministic target policy (see Table 5 in the appendix).

In Table 3, we compare optimistic and pessimistic shrinkage when paired with a fixed reward predictor (using oracle tuning for λ). We report how many times each 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.

Comparisons. In Figure 1 (left two plots), we compare our new estimator with the baselines. We visualize the results by plotting the cumulative distribution function (CDF) of the normalized MSE of each method (normalized by the MSE of snIPS) across the experimental conditions. Better performance corresponds to CDF curves towards the top-left corner, meaning the method achieves a lower MSE more frequently. The first plot summarizes 54 conditions where the reward is deterministic, while the second plot considers the 54 stochastic reward conditions. For DRs we consider two model selection procedures outlined in Section 5 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 these estimates twice their standard error, before taking minimum, more closely matching our theory. For DRs, we 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 (the first 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 (the second plot), 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 the right two plots of Figure 1, plotting the MSE as a function of the number of samples for one choice of a logging policy and dataset, first with deterministic rewards and then with stochastic rewards. Because of a more robust performance, we therefore advocate for DRs-direct as our final method.

6.1.2. Off-policy Learning

Following prior work (Swaminathan & Joachims, 2015a;b; Su et al., 2018), 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 $\hat{\mathbf{u}} = \operatorname{argmin}_{\mathbf{u}} [-\hat{V}(\pi_{\mathbf{u}}) + \gamma \|\mathbf{u}\|^2]$ 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 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 $z = 1/\mu(a \mid x)$ and $z = 1/\mu(a \mid x)^2$.

In Figure 3, we show 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 report this value normalized by the average value for IPS. For DM and DR, we select the hyperparameter γ and reward predictor optimally in hindsight, while for DRs we use our model selection. Note that we do not compare with SWITCH here as it is not amenable to gradient-based optimization (Su et al., 2018). We find that off-policy learning using DRs-direct always outperforms the baselines, with the exception of the *optdigits* dataset, where all the methods perform similarly.

6.2. Combinatorial Setting

We empirically evaluate the performance of shrinkage-based estimator in the ranking problem introduced in Example 1. Following Swaminathan et al. (2017), we generate contextual bandit data from the fully labeled MSLR-WEB10K

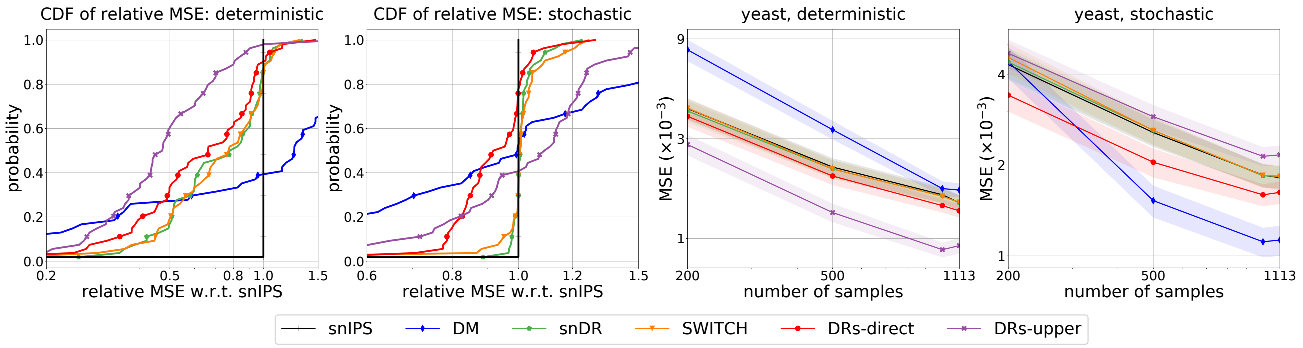

Figure 1. From left to right: (1) CDF of relative MSE w.r.t. snIPS for deterministic rewards, 54 conditions in total; (2) CDF of relative MSE w.r.t. snIPS for stochastic rewards, 54 conditions in total; (3) learning curves on *yeast* dataset, using base policy π_1 with $\alpha = 0.7$ and $\beta = 0.2$, deterministic reward; (4) learning curves on *yeast* dataset, using base policy π_1 with $\alpha = 0.7$ and $\beta = 0.2$, stochastic reward.

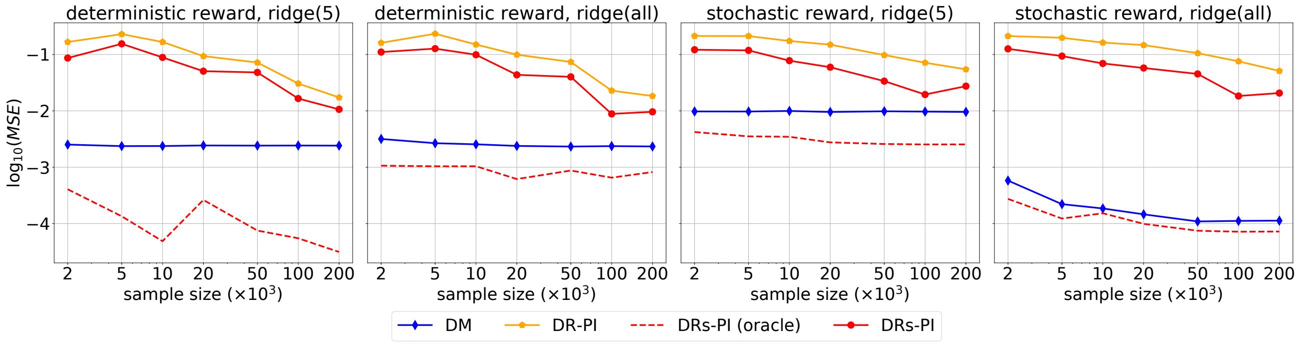

Figure 2. Off-policy evaluation with combinatorial actions. MSE as a function of sample sizes for two reward distributions (deterministic and stochastic), and two reward predictors (ridge regression with five and all features). The MSE of DR-PI is significantly larger than DR-PIs in all cases (with p-value below 0.013 according to a paired t-test).

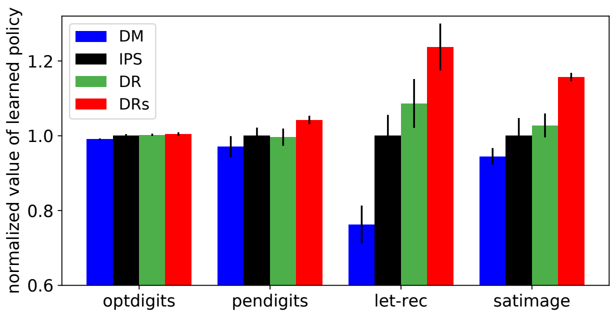

Figure 3. Off-policy learning experiments.

dataset (Qin & Liu, 2013). The dataset has 10K queries, with up to 1251 judged documents for each query. The contexts x are the queries and actions $\mathbf{a}$ represent lists of documents. For each query x and document i, the dataset contains a relevance judgement $\mathrm{rel}(x, i) \in \{0, 1, 2, 3, 4\}$. We consider two types of rewards: *deterministic rewards*, $r = \mathrm{NDCG}(x, \mathbf{a})$ (see definition in Example 1); and *stochastic rewards*, where r is drawn from a Bernoulli distribution with $p = 0.25 + 0.5 \cdot \mathrm{NDCG}(x, \mathbf{a})$. We use data for 10% of the queries to train relevance predictors used to define logging and target policies; the remaining data is used for the bandit protocol. The ground truth is determined using all the data.

Policies. Each query-document pair (x, i) is described by a feature vector $\mathbf{f}(x, i)$, partitioned into title and body features, denoted $\mathbf{f}_t$ and $\mathbf{f}_b$. We train two regression models to predict relevance: a lasso model *lasso*$_b$ based on $\mathbf{f}_b$, and a tree model *tree*$_t$ based on $\mathbf{f}_t$. The model *lasso*$_b$ is used to select the top 20 scoring documents; the action $\mathbf{a}$ is a list of 5 documents out of these 20. In the notation of Example 1, $m = 20$, $\ell = 5$. The target policy is deterministic and chooses $\mathbf{a}$ that lists top 5 documents according to *tree*$_t$. The logging policy is supported on a basis $\mathcal{B}_x \subseteq \mathcal{A}$ for each x. The basis contains the "greedy action" that lists top 5 documents according to *lasso*$_b$ as well as actions obtained by replacing items on the top position and up to two additional positions of the greedy action, resulting in the total of 96 elements in $\mathcal{B}_x$ (see Appendix D.2 for details). The logging policy is ϵ-greedy: on each context, ϵ is drawn uniformly from the set $\{2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}\}$ and is included as part of the context, creating a skew in the importance weights $\mathbf{w}(x)^\top \mathbf{a}$.

Reward predictors. We consider two reward predictors $\hat{\eta}$ trained on logged data. Both are trained via ridge regression, but differ in feature sets they consider: *ridge(all)* is trained on all features, *ridge(5)* is trained on the five features that are most correlated with the reward.

Baselines. We compare our method (DRs-PI) with DM and DR-PI.[2] In DRs-PI we select the hyperparameter λ from a geometrically spaced grid using our model selection procedure with the empirical version of Eq. (4) in place of bias bound and also consider the oracle tuning of λ from the same grid (details in Appendix D.2).

Results and discussion. In Figure 2 we show the MSE of all the methods as a function of sample size, averaged over 20 replicates. Across all conditions, DRs-PI outperforms DR by a factor of 1.5 or more (note that MSE is reported on log scale). A more striking result is the superior quality of the oracle-tuned DRs-PI. It shows that the shrinkage strategy is highly effective in achieving a good bias–variance trade-off, but to unlock its potential in combinatorial settings requires improvements in model selection.

7. Conclusion

In this paper, we have derived shrinkage-based doubly-robust estimators for off-policy evaluation using a principled optimization-based framework. Our approach recovers the weight-clipping estimator from prior work and also yields novel optimistic shrinkage estimators for both atomic and combinatorial settings. Extensive experiments demonstrate the efficacy of these estimators and highlight the role of model selection in achieving good performance. Thus, the next step is to develop model selection procedures for off-policy evaluation that can close the gap with oracle tuning. We look forward to pursuing this direction in future work.

Acknowledgements

This work was partially completed during Yi's and Maria's internships at Microsoft Research. Yi is also supported by the Bloomberg Data Science Fellowship. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers or sponsors.

References

Awerbuch, B. and Kleinberg, R. Online linear optimization and adaptive routing. *J. Comput. Syst. Sci.*, 74(1):97–114, 2008.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 2005.

Bembom, O. and van der Laan, M. J. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical report, UC Berkeley, 2008.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 2013.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 2012.

Dani, V., Hayes, T. P., and Kakade, S. M. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, 2008.

De la Pena, V. and Giné, E. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

Dudík, M., Erhan, D., Langford, J., Li, L., et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, 2018.

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

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 2003.

Imbens, G., Newey, W., and Ridder, G. Mean-squared-error calculations for average treatment effects. *ssrn.954748*, 2007.

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

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, 2018.

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

[2] DR-PI dominates self-normalized version of DR-PI as well as the standard pseudo-inverse estimator (i.e., with $\hat{\eta} \equiv 0$).

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, 2008.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *International Conference on Web Search and Data Mining*, 2011.

Li, L., Chen, S., Kleban, J., and Gupta, A. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *International Conference on World Wide Web*, 2015.

Qin, T. and Liu, T. Introducing LETOR 4.0 datasets. *arXiv:1306.2597*, 2013.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 1995.

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

Strehl, A., Langford, J., Li, L., and Kakade, S. M. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, 2010.

Su, Y., Wang, L., Santacatterina, M., and Joachims, T. Cab: Continuous adaptive blending estimator for policy evaluation and learning. In *International Conference on Machine Learning*, 2018.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, 2015a.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015b.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, 2017.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Wang, Y.-X., Agarwal, A., and Dudik, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, 2017.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 2017.