Doubly Robust Policy Evaluation and Learning

Miroslav Dudík MDUDIK@YAHOO-INC.COM
John Langford JL@YAHOO-INC.COM
Yahoo! Research, New York, NY, USA 10018

Lihong Li LIHONG@YAHOO-INC.COM
Yahoo! Research, Santa Clara, CA, USA 95054

Abstract

We study decision making in environments where the reward is only partially observed, but can be modeled as a function of an action and an observed context. This setting, known as contextual bandits, encompasses a wide variety of applications including health-care policy and Internet advertising. A central task is evaluation of a new policy given historic data consisting of contexts, actions and received rewards. The key challenge is that the past data typically does not faithfully represent proportions of actions taken by a new policy. Previous approaches rely either on models of rewards or models of the past policy. The former are plagued by a large bias whereas the latter have a large variance.

In this work, we leverage the strength and overcome the weaknesses of the two approaches by applying the *doubly robust* technique to the problems of policy evaluation and optimization. We prove that this approach yields accurate value estimates when we have *either* a good (but not necessarily consistent) model of rewards *or* a good (but not necessarily consistent) model of past policy. Extensive empirical comparison demonstrates that the doubly robust approach uniformly improves over existing techniques, achieving both lower variance in value estimation and better policies. As such, we expect the doubly robust approach to become common practice.

1. Introduction

We study decision making in environments where we receive feedback only for chosen actions. For example, in Internet advertising, we find only whether a user clicked on some of the presented ads, but receive no information about the ads that were not presented. In health care, we only find out success rates for patients who received the treatments, but not for the alternatives. Both of these problems are instances of *contextual bandits* (Auer et al., 2002; Langford & Zhang, 2008). The context refers to additional information about the user or patient. Here, we focus on the offline version: we assume access to historic data, but no ability to gather new data (Langford et al., 2008; Strehl et al., 2011).

Two kinds of approaches address offline learning in contextual bandits. The first, which we call the *direct method* (DM), estimates the reward function from given data and uses this estimate in place of actual reward to evaluate the policy value on a set of contexts. The second kind, called *inverse propensity score* (IPS) (Horvitz & Thompson, 1952), uses importance weighting to correct for the incorrect proportions of actions in the historic data. The first approach requires an accurate model of rewards, whereas the second approach requires an accurate model of the past policy. In general, it might be difficult to accurately model rewards, so the first assumption can be too restrictive. On the other hand, it is usually possible to model the past policy quite well. However, the second kind of approach often suffers from large variance especially when the past policy differs significantly from the policy being evaluated.

In this paper, we propose to use the technique of *doubly robust* (DR) estimation to overcome problems with the two existing approaches. Doubly robust (or doubly protected) estimation (Cassel et al., 1976; Robins et al., 1994; Robins & Rotnitzky, 1995; Lunceford & Davidian, 2004; Kang & Schafer, 2007) is a statistical approach for estimation from incomplete data with an important property: if *either one* of the two estimators (in DM and IPS) is correct, then the estimation is unbiased. This method thus increases the chances of drawing reliable inference.

For example, when conducting a survey, seemingly ancillary questions such as age, sex, and family income may be

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

asked. Since not everyone contacted responds to the survey, these values along with census statistics may be used to form an estimator of the probability of a response conditioned on age, sex, and family income. Using importance weighting inverse to these estimated probabilities, one estimator of overall opinions can be formed. An alternative estimator can be formed by directly regressing to predict the survey outcome given any available sources of information. Doubly robust estimation unifies these two techniques, so that unbiasedness is guaranteed if *either* the probability estimate is accurate *or* the regressed predictor is accurate.

We apply the doubly robust technique to policy value estimation in a contextual bandit setting. The core technique is analyzed in terms of bias in Section 3 and variance in Section 4. Unlike previous theoretical analyses, we do not assume that either the reward model or the past policy model are correct. Instead, we show how the deviations of the two models from the truth impact bias and variance of the doubly robust estimator. To our knowledge, this style of analysis is novel and may provide insights into doubly robust estimation beyond the specific setting studied here. In Section 5, we apply this method to both policy evaluation and optimization, finding that this approach substantially sharpens existing techniques.

1.1. Prior Work

Doubly robust estimation is widely used in statistical inference (see, e.g., Kang & Schafer (2007) and the references therein). More recently, it has been used in Internet advertising to estimate the effects of new features for online advertisers (Lambert & Pregibon, 2007; Chan et al., 2010). Previous work focuses on parameter estimation rather than policy evaluation/optimization, as addressed here. Furthermore, most of previous analysis of doubly robust estimation studies asymptotic behavior or relies on various modeling assumptions (e.g., Robins et al. (1994), Lunceford & Davidian (2004), and Kang & Schafer (2007)). Our analysis is non-asymptotic and makes no such assumptions.

Several other papers in machine learning have used ideas related to the basic technique discussed here, although not with the same language. For *benign bandits*, Hazan & Kale (2009) construct algorithms which use reward estimators in order to achieve a worst-case regret that depends on the variance of the bandit rather than time. Similarly, the Offset Tree algorithm (Beygelzimer & Langford, 2009) can be thought of as using a crude reward estimate for the "offset". In both cases, the algorithms and estimators described here are substantially more sophisticated.

2. Problem Definition and Approach

Let $\mathcal{X}$ be an input space and $\mathcal{A} = \{1, \ldots, k\}$ a finite action space. A contextual bandit problem is specified by a distribution D over pairs $(x, \vec{r})$ where $x \in \mathcal{X}$ is the context and $\vec{r} \in [0,1]^{\mathcal{A}}$ is a vector of rewards. The input data has been generated using some unknown policy (possibly adaptive and randomized) as follows:

- The world draws a new example $(x, \vec{r}) \sim D$. Only x is revealed.
- The policy chooses an action $a \sim p(a \mid x, h)$, where h is the history of previous observations (that is, the concatenation of all preceding contexts, actions and observed rewards).
- Reward r_a is revealed. It should be emphasized that other rewards $r_{a'}$ with $a' \neq a$ are not observed.

Note that neither the distribution D nor the policy p is known. Given a data set $S = \{(x, h, a, r_a)\}$ collected as above, we are interested in two tasks: policy evaluation and policy optimization. In policy evaluation, we are interested in estimating the *value* of a stationary policy π, defined as:

$$V^\pi = \mathbf{E}_{(x,\vec{r})\sim D}[r_{\pi(x)}] \ .$$

On the other hand, the goal of policy optimization is to find an optimal policy with maximum value: $\pi^* = \operatorname{argmax}_\pi V^\pi$. In the theoretical sections of the paper, we treat the problem of policy evaluation. It is expected that better evaluation generally leads to better optimization (Strehl et al., 2011). In the experimental section, we study how our policy evaluation approach can be used for policy optimization in a classification setting.

2.1. Existing Approaches

The key challenge in estimating policy value, given the data as described in the previous section, is the fact that we only have partial information about the reward, hence we cannot directly simulate our proposed policy on the data set S. There are two common solutions for overcoming this limitation. The first, called *direct method* (DM), forms an estimate $\hat{\varrho}_a(x)$ of the expected reward conditioned on the context *and* action. The policy value is then estimated by

$$\hat{V}^\pi_{\text{DM}} = \frac{1}{|S|} \sum_{x \in S} \hat{\varrho}_{\pi(x)}(x) \ .$$

Clearly, if $\hat{\varrho}_a(x)$ is a good approximation of the true expected reward, defined as $\varrho_a(x) = \mathbf{E}_{(x,\vec{r})\sim D}[r_a \mid x]$, then the DM estimate is close to V^π. Also, if $\hat{\varrho}$ is unbiased, $\hat{V}^\pi_{\text{DM}}$ is an unbiased estimate of V^π. A problem with this method is that the estimate $\hat{\varrho}$ is formed without the knowledge of π and hence might focus on approximating ϱ mainly in the areas that are irrelevant for V^π and not sufficiently in the areas that are important for V^π; see Beygelzimer & Langford (2009) for a more refined analysis.

The second approach, called *inverse propensity score* (IPS), is typically less prone to problems with bias. Instead of approximating the reward, IPS forms an approximation $\hat{p}(a \mid x, h)$ of $p(a \mid x, h)$, and uses this estimate to correct for the shift in action proportions between the old, data-collection policy and the new policy:

$$\hat{V}^{\pi}_{\text{IPS}} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \frac{r_a \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)}$$

where $\mathbf{I}(\cdot)$ is an indicator function evaluating to one if its argument is true and zero otherwise. If $\hat{p}(a \mid x, h) \approx p(a \mid x, h)$ then the IPS estimate above will be, approximately, an unbiased estimate of V^{π}. Since we typically have a good (or even accurate) understanding of the data-collection policy, it is often easier to obtain a good estimate $\hat{p}$, and thus IPS estimator is in practice less susceptible to problems with bias compared with the direct method. However, IPS typically has a much larger variance, due to the range of the random variable increasing. The issue becomes more severe when $p(a \mid x, h)$ gets smaller. Our approach alleviates the large variance problem of IPS by taking advantage of the estimate $\hat{\varrho}$ used by the direct method.

2.2. Doubly Robust Estimator

Doubly robust estimators take advantage of both the estimate of the expected reward $\hat{\varrho}_a(x)$ and the estimate of action probabilities $\hat{p}(a \mid x, h)$. Here, we use a DR estimator of the form first suggested by Cassel et al. (1976) for regression, but previously not studied for policy learning:

$$\hat{V}^{\pi}_{\text{DR}} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \Bigg[\frac{(r_a - \hat{\varrho}_a(x)) \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)} + \hat{\varrho}_{\pi(x)}(x) \Bigg]. \quad (1)$$

Informally, the estimator uses $\hat{\varrho}$ as a baseline and if there is data available, a correction is applied. We will see that our estimator is accurate if *at least one* of the estimators, $\hat{\varrho}$ and $\hat{p}$, is accurate, hence the name *doubly robust*.

In practice, it is rare to have an accurate estimation of either ϱ or p. Thus, a basic question is: How does this estimator perform as the estimates $\hat{\varrho}$ and $\hat{p}$ deviate from the truth? The following two sections are dedicated to bias and variance analysis, respectively, of the DR estimator.

3. Bias Analysis

Let Δ denote the additive deviation of $\hat{\varrho}$ from ϱ, and δ a multiplicative deviation of $\hat{p}$ from p:

$$\Delta(a, x) = \hat{\varrho}_a(x) - \varrho_a(x),$$
$$\delta(a, x, h) = 1 - p(a \mid x, h)/\hat{p}(a \mid x, h) \ .$$

We express the expected value of $\hat{V}^{\pi}_{\text{DR}}$ using $\delta(\cdot, \cdot, \cdot)$ and $\Delta(\cdot, \cdot)$. To remove clutter, we introduce shorthands ϱ_a for $\varrho_a(x)$, $\hat{\varrho}_a$ for $\hat{\varrho}_a(x)$, $\mathbf{I}$ for $\mathbf{I}(\pi(x) = a)$, p for $p(\pi(x) \mid x, h)$, $\hat{p}$ for $\hat{p}(\pi(x) \mid x, h)$, Δ for $\Delta(\pi(x), x))$, and δ for $\delta(\pi(x), x, h)$. In our analysis, we assume that the estimates $\hat{p}$ and $\hat{\varrho}$ are fixed independently of S (e.g., by splitting the original data set into S and a separate portion for estimating $\hat{p}$ and $\hat{\varrho}$). To evaluate $\mathbf{E}[\hat{V}^{\pi}_{\text{DR}}]$, it suffices to focus on a single term in Eq. (1), conditioning on h:

$$\begin{aligned}
&\mathbf{E}_{(x,\vec{r}) \sim D, a \sim p(\cdot \mid x, h)} \left[\frac{(r_a - \hat{\varrho}_a)\mathbf{I}}{\hat{p}} + \hat{\varrho}_{\pi(x)} \right] \\
&= \mathbf{E}_{x,\vec{r},a|h} \left[\frac{(r_a - \varrho_a - \Delta)\mathbf{I}}{\hat{p}} + \varrho_{\pi(x)} + \Delta \right] \\
&= \mathbf{E}_{x,a|h} \left[\frac{(\varrho_a - \varrho_a)\mathbf{I}}{\hat{p}} + \Delta\big(1 - \mathbf{I}/\hat{p}\big) \right] + \mathbf{E}_x[\varrho_{\pi(x)}] \\
&= \mathbf{E}_{x|h}\big[\Delta\big(1 - p/\hat{p}\big)\big] + V^{\pi} = \mathbf{E}_{x|h}[\Delta\delta] + V^{\pi} \ .
\end{aligned} \quad (2)$$

Even though x is independent of h, the conditioning on h remains in the last line, because δ, p and $\hat{p}$ are functions of h. Summing across all terms in Eq. (1), we obtain the following theorem:

Theorem 1 *Let Δ and δ be defined as above. Then, the bias of the doubly robust estimator is*

$$|\mathbf{E}_S[\hat{V}^{\pi}_{\text{DR}}] - V^{\pi}| = \frac{1}{|S|} \Big| \mathbf{E}_S \Big[\sum_{(x,h) \in S} \Delta\delta \Big] \Big| \ .$$

If the past policy and the past policy estimate are stationary (i.e., independent of h), the expression simplifies to

$$|\mathbf{E}[\hat{V}^{\pi}_{\text{DR}}] - V^{\pi}| = |\mathbf{E}_x[\Delta\delta]| \ .$$

In contrast (for simplicity we assume stationarity):

$$|\mathbf{E}[\hat{V}^{\pi}_{\text{DM}}] - V^{\pi}| = |\mathbf{E}_x[\Delta]|$$
$$|\mathbf{E}[\hat{V}^{\pi}_{\text{IPS}}] - V^{\pi}| = |\mathbf{E}_x[\varrho_{\pi(x)}\delta]| \ ,$$

where the second equality is based on the observation that IPS is a special case of DR for $\hat{\varrho}_a(x) \equiv 0$.

In general, neither of the estimators dominates the others. However, if *either* $\Delta \approx 0$, *or* $\delta \approx 0$, the expected value of the doubly robust estimator will be close to the true value, whereas DM requires $\Delta \approx 0$ and IPS requires $\delta \approx 0$. Also, if $\Delta \approx 0$ and $\delta \ll 1$, DR will still outperform DM, and similarly for IPS with roles of Δ and δ reversed. Thus, DR can effectively take advantage of both sources of information for better estimation.

4. Variance Analysis

In the previous section, we argued that the expected value of $\hat{V}^{\pi}_{\text{DR}}$ compares favorably with IPS and DM. In this section, we look at the variance of DR. Since large deviation

bounds have a primary dependence on variance, a lower variance implies a faster convergence rate. We treat only the case with stationary past policy, and hence drop the dependence on h throughout.

As in the previous section, it suffices to analyze the second moment (and then variance) of a single term of Eq. (1). We use a similar decomposition as in Eq. (2). To simplify derivation we use the notation $\varepsilon = (r_a - \varrho_a)\mathbf{I}/\hat{p}$. Note that, conditioned on x and a, the expectation of ε is zero. Hence, we can write the second moment as

$$\begin{aligned}
&\mathbf{E}_{x,\vec{r},a}\left[\left(\frac{(r_a - \hat{\varrho}_a)\mathbf{I}}{\hat{p}} + \hat{\varrho}_{\pi(x)}\right)^2\right] \\
&\quad= \mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{E}_x[\varrho^2_{\pi(x)}] + 2\mathbf{E}_{x,a}\left[\varrho_{\pi(x)}\Delta\left(1 - \mathbf{I}/\hat{p}\right)\right] \\
&\qquad+ \mathbf{E}_{x,a}\left[\Delta^2\left(1 - \mathbf{I}/\hat{p}\right)^2\right] \\
&\quad= \mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{E}_x[\varrho^2_{\pi(x)}] + 2\mathbf{E}_x\left[\varrho_{\pi(x)}\Delta\delta\right] \\
&\qquad+ \mathbf{E}_x\left[\Delta^2\left(1 - 2p/\hat{p} + p/\hat{p}^2\right)\right] \\
&\quad= \mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{E}_x[\varrho^2_{\pi(x)}] + 2\mathbf{E}_x\left[\varrho_{\pi(x)}\Delta\delta\right] \\
&\qquad+ \mathbf{E}_x\left[\Delta^2\left(1 - 2p/\hat{p} + p^2/\hat{p}^2 + p(1-p)/\hat{p}^2\right)\right] \\
&\quad= \mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{E}_x\left[\left(\varrho_{\pi(x)} + \Delta\delta\right)^2\right] \\
&\qquad+ \mathbf{E}_x\left[\Delta^2 \cdot p(1-p)/\hat{p}^2\right] \\
&\quad= \mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{E}_x\left[\left(\varrho_{\pi(x)} + \Delta\delta\right)^2\right] \\
&\qquad+ \mathbf{E}_x\left[\frac{1-p}{p} \cdot \Delta^2(1-\delta)^2\right] .
\end{aligned}$$

Summing across all terms in Eq. (1) and combining with Theorem 1, we obtain the variance:

Theorem 2 *Let Δ, δ and ε be defined as above. If the past policy and the policy estimate are stationary, then the variance of the doubly robust estimator is*

$$\mathbf{Var}\left[\hat{V}^{\pi}_{\mathrm{DR}}\right] = \frac{1}{|S|}\left(\mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{Var}_x\left[\varrho_{\pi(x)} + \Delta\delta\right] + \mathbf{E}_x\left[\frac{1-p}{p} \cdot \Delta^2(1-\delta)^2\right]\right) .$$

Thus, the variance can be decomposed into three terms. The first accounts for randomness in rewards. The second term is the variance of the estimator due to the randomness in x. And the last term can be viewed as the importance weighting penalty. A similar expression can be derived for the IPS estimator:

$$\mathbf{Var}\left[\hat{V}^{\pi}_{\mathrm{IPS}}\right] = \frac{1}{|S|}\left(\mathbf{E}_{x,\vec{r},a}[\varepsilon^2] + \mathbf{Var}_x\left[\varrho_{\pi(x)} - \varrho_{\pi(x)}\delta\right] + \mathbf{E}_x\left[\frac{1-p}{p} \cdot \varrho^2_{\pi(x)}(1-\delta)^2\right]\right) .$$

The first term is identical, the second term will be of similar magnitude as the corresponding term of the DR estimator, provided that $\delta \approx 0$. However, the third term can be much larger for IPS if $p(\pi(x) \mid x) \ll 1$ and $|\Delta|$ is smaller than $\varrho_{\pi(x)}$. In contrast, for the direct method, we obtain

$$\mathbf{Var}\left[\hat{V}^{\pi}_{\mathrm{DM}}\right] = \frac{1}{|S|}\mathbf{Var}_x\left[\varrho_{\pi(x)} + \Delta\right] .$$

Thus, the variance of the direct method does not have terms depending either on the past policy or the randomness in the rewards. This fact usually suffices to ensure that it is significantly lower than the variance of DR or IPS. However, as we mention in the previous section, the bias of the direct method is typically much larger, leading to larger errors in estimating policy value.

5. Experiments

This section provides empirical evidence for the effectiveness of the DR estimator compared to IPS and DM. We consider two classes of problems: multiclass classification with bandit feedback in public benchmark datasets and estimation of average user visits to an Internet portal.

5.1. Multiclass Classification with Bandit Feedback

We begin with a description of how to turn a k-class classification task into a k-armed contextual bandit problem. This transformation allows us to compare IPS and DR using *public* datasets for both policy evaluation and learning.

5.1.1. Data Setup

In a classification task, we assume data are drawn IID from a fixed distribution: $(x, c) \sim D$, where $x \in \mathcal{X}$ is the feature vector and $c \in \{1, 2, \ldots, k\}$ is the class label. A typical goal is to find a classifier $\pi : \mathcal{X} \mapsto \{1, 2, \ldots, k\}$ minimizing the classification error:

$$e(\pi) = \mathbf{E}_{(x,c)\sim D}\left[\mathbf{I}(\pi(x) \neq c)\right] .$$

Alternatively, we may turn the data point (x, c) into a cost-sensitive classification example $(x, l_1, l_2, \ldots, l_k)$, where $l_a = \mathbf{I}(a \neq c)$ is the loss for predicting a. Then, a classifier π may be interpreted as an action-selection policy, and its classification error is exactly the policy's expected loss.[1]

To construct a partially labeled dataset, exactly one loss component for each example is observed, following the approach of Beygelzimer & Langford (2009). Specifically, given any $(x, l_1, l_2, \ldots, l_k)$, we randomly select a label

[1] When considering classification problems, it is more natural to talk about minimizing classification errors. This loss minimization problem is symmetric to the reward maximization problem defined in Section 2.

Table 1. Characteristics of benchmark datasets used in Section 5.1.

Dataset	ecoli	glass	letter	optdigits	page-blocks	pendigits	satimage	vehicle	yeast
Classes (k)	8	6	26	10	5	10	6	4	10
Dataset size	336	214	20000	5620	5473	10992	6435	846	1484

$a \sim \text{UNIF}(1, 2, \ldots, k)$, and then only reveal the component l_a. The final data are thus in the form of (x, a, l_a), which is the form of data defined in Section 2. Furthermore, $p(a \mid x) \equiv 1/k$ and is assumed to be known.

Table 1 summarizes the benchmark problems adopted from the UCI repository (Asuncion & Newman, 2007).

5.1.2. Policy Evaluation

Here, we investigate whether the DR technique indeed gives more accurate estimates of the policy value (or classification error in our context). For each dataset:

1. We randomly split data into training and test sets of (roughly) the same size;
2. On the training set with fully revealed losses, we run a direct loss minimization (DLM) algorithm of McAllester et al. (2011) to obtain a classifier (see Appendix A for details). This classifier constitutes the policy π which we evaluate on test data;
3. We compute the classification error on fully observed test data. This error is treated as the ground truth for comparing various estimates;
4. Finally, we apply the transformation in Section 5.1.1 to the test data to obtain a partially labeled set, from which DM, IPS, and DR estimates are computed.

Both DM and DR require estimating the expected conditional loss denoted as $l(x, a)$ for given (x, a). We use a linear loss model: $\hat{l}(x, a) = w_a \cdot x$, parameterized by k weight vectors $\{w_a\}_{a \in \{1,\ldots,k\}}$, and use least-squares ridge regression to fit w_a based on the training set.

Step 4 is repeated 500 times, and the resulting bias and rmse (root mean squared error) are reported in Fig. 1.

As predicted by analysis, both IPS and DR are unbiased, since the probability estimate $1/k$ is accurate. In contrast, the linear loss model fails to capture the classification error accurately, and as a result, DM suffers a much larger bias.

While IPS and DR estimators are unbiased, it is apparent from the rmse plot that the DR estimator enjoys a lower variance, which translates into a smaller rmse. As we shall see next, such an effect is substantial when it comes to policy optimization.

5.1.3. Policy Optimization

We now consider policy optimization (classifier learning). Since DM is significantly worse on all datasets, as indicated in Fig. 1, we focus on the comparison between IPS and DR.

Here, we apply the data transformation in Section 5.1.1 to the *training* data, and then learn a classifier based on the loss estimated by IPS and DR, respectively. Specifically, for each dataset, we repeat the following steps 30 times:

1. We randomly split data into training (70%) and test (30%) sets;
2. We apply the transformation in Section 5.1.1 to the training data to obtain a partially labeled set;
3. We then use the IPS and DR estimators to impute unrevealed losses in the training data;
4. Two cost-sensitive multiclass classification algorithms are used to learn a classifier from the losses completed by either IPS or DR: the first is DLM (McAllester et al., 2011), the other is the Filter Tree reduction of Beygelzimer et al. (2008) applied to a decision tree (see Appendix B for more details);
5. Finally, we evaluate the learned classifiers on the test data to obtain classification error.

Again, we use least-squares ridge regression to build a linear loss estimator: $\hat{l}(x, a) = w_a \cdot x$. However, since the training data is partially labeled, w_a is fitted only using training data $(x, a', l_{a'})$ for which $a = a'$.

Average classification errors (obtained in Step 5 above) of the 30 runs are plotted in Fig. 2. Clearly, for policy optimization, the advantage of the DR is even greater than for policy evaluation. In all datasets, DR provides substantially more reliable loss estimates than IPS, and results in significantly improved classifiers.

Fig. 2 also includes classification error of the Offset Tree reduction, which is designed specifically for policy optimization with partially labeled data.[2] While the IPS versions of DLM and Filter Tree are rather weak, the DR versions are competitive with Offset Tree in all datasets, and in some cases significantly outperform Offset Tree.

Finally, we note DR provided similar improvements to two

[2]We used decision trees as the base learner in Offset Trees. The numbers reported here are not identical to those by Beygelzimer & Langford (2009) probably because the filter-tree structures in our implementation were different.

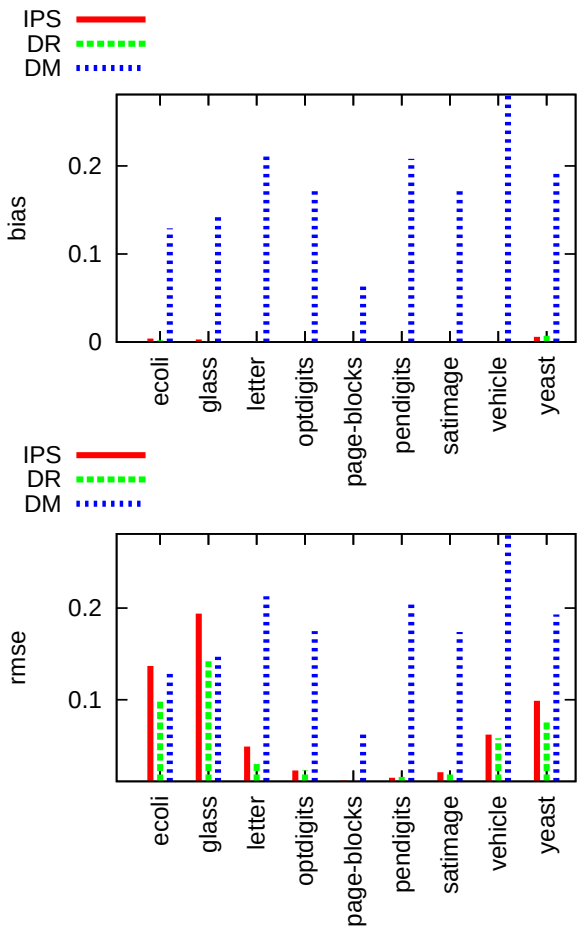

Figure 1. Bias (upper) and rmse (lower) of the three estimators for classification error.

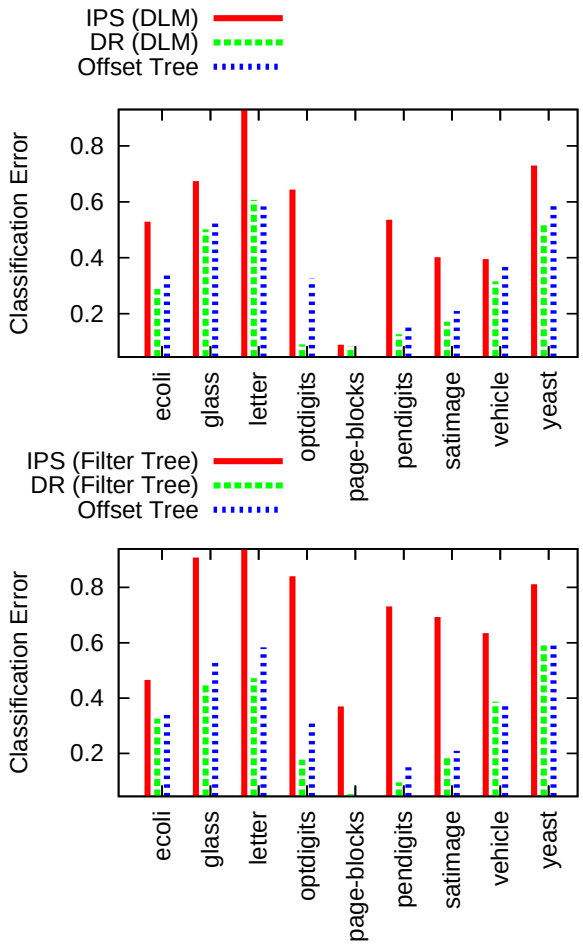

Figure 2. Classification error of direct loss minimization (upper) and filter tree (lower). Note that the representations used by DLM and the trees differ radically, conflating any comparison between the approaches. However, the Offset and Filter Tree approaches share a similar representation, so differences in performance are purely a matter of superior optimization.

very different algorithms, one based on gradient descent, the other based on tree induction. It suggests the DR technique is generally useful when combined with different algorithmic choices.

5.2. Estimating Average User Visits

The next problem we consider is estimating the average number of user visits to a popular Internet portal. Real user visits to the website were recorded for about 4 million *bcookies*[3] randomly selected from all bcookies during March 2010. Each bcookie is associated with a sparse binary feature vector of size around 5000. These features describe browsing behavior as well as other information (such as age, gender, and geographical location) of the bcookie. We chose a fixed time window in March 2010 and calculated the number of visits by each selected bcookie during this window. To summarize, the dataset contains $N = 3854689$ data: $D = \{(b_i, x_i, v_i)\}_{i=1,\ldots,N}$, where b_i is the i-th (unique) bcookie, x_i is the corresponding binary feature vector, and v_i is the number of visits.

If we can sample from D uniformly at random, the sample mean of v_i will be an unbiased estimate of the true average number of user visits, which is 23.8 in this problem. However, in various situations, it may be difficult or impossible to ensure a uniform sampling scheme due to practical constraints, thus the sample mean may not reflect the true quantity of interest. This is known as *covariate shift*, a special case of our problem formulated in Section 2 with $k = 2$ arms. Formally, the partially labeled data consists

[3]A bcookie is unique string that identifies a user. Strictly speaking, one user may correspond to multiple bcookies, but it suffices to equate a bcookie with a user for our purposes here.

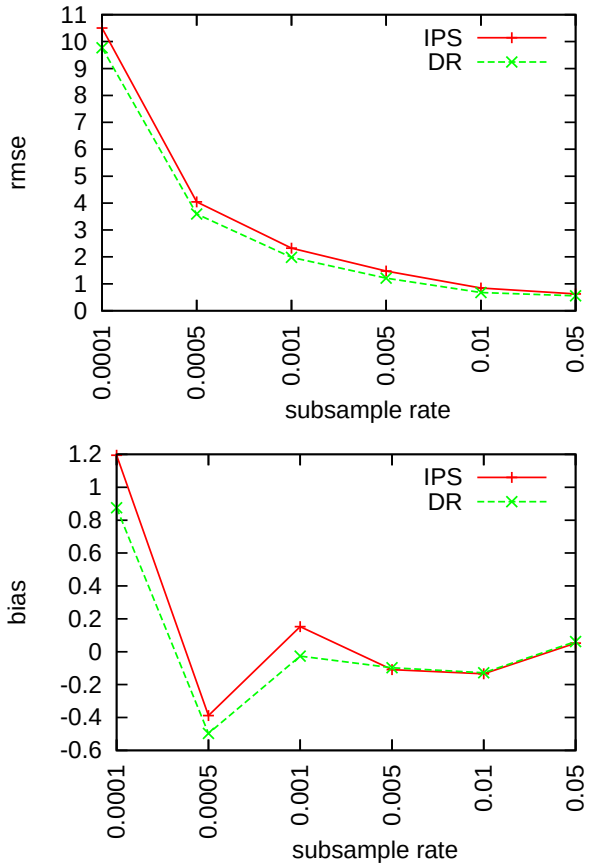

Figure 3. Comparison of IPS and DR: rmse (top), bias (bottom). The ground truth value is 23.8.

of tuples (x_i, a_i, r_i), where $a_i \in \{0, 1\}$ indicates whether bcookie b_i is sampled, $r_i = a_i v_i$ is the observed number of visits, and p_i is the probability that $a_i = 1$. The goal here is to evaluate the value of a constant policy: $\pi(x) \equiv 1$.

To define the sampling probabilities p_i, we adopted a similar approach as in Gretton et al. (2008). In particular, we obtained the first principal component (denoted $\bar{x}$) of all features $\{x_i\}$, and projected all data onto $\bar{x}$. Let $\mathcal{N}$ be a univariate normal distribution with mean $m + (\bar{m} - m)/3$ and standard deviation $(\bar{m} - m)/4$, where m and $\bar{m}$ were the minimum and mean of the projected values. Then, $p_i = \min\{\mathcal{N}(x_i \cdot \bar{x}), 1\}$ was the sampling probability of the i-th bcookie, b_i.

To control data size, we randomly subsampled a fraction $f \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05\}$ from the entire dataset D. For each bcookie b_i in this subsample, set $a_i = 1$ with probability p_i, and $a_i = 0$ otherwise. We then calculated the IPS and DR estimates on this subsample. The whole process was repeated 100 times.

The DR estimator required building a reward model $\hat{\varrho}(x)$, which, given feature x, predicted the average number of visits. Again, least-squares ridge regression was used to fit a linear model $\hat{\varrho}(x) = w \cdot x$ from sampled data.

Fig. 3 summarizes the estimation error of the two methods with increasing data size. For both IPS and DR, the estimation error goes down with more data. In terms of rmse, the DR estimator is consistently better than IPS, especially when dataset size is smaller. The DR estimator often reduces the rmse by a fraction between 10% and 20%, and on average by 13.6%. By comparing to the bias and std metrics, it is clear that DR's gain of accuracy came from a lower variance, which accelerated convergence of the estimator to the true value. These results confirm our analysis that DR tends to reduce variance provided that a reasonable reward estimator is available.

6. Conclusions

Doubly robust policy estimation is an effective technique which virtually always improves on the widely used inverse propensity score method. Our analysis shows that doubly robust methods tend to give more reliable and accurate estimates. The theory is corroborated by experiments on both benchmark data and a large-scale, real-world problem.

In the future, we expect the DR technique to become common practice in improving contextual bandit algorithms. As an example, it is interesting to develop a variant of Offset Tree that can take advantage of better reward models, rather than a crude, constant reward estimate (Beygelzimer & Langford, 2009).

Acknowledgements

We thank Deepak Agarwal for first bringing the doubly robust technique to our attention.

A. Direct Loss Minimization

Given cost-sensitive multiclass classification data $\{(x, l_1, \ldots, l_k)\}$, we perform approximate gradient descent on the policy loss (or classification error). In the experiments of Section 5.1, policy π is specified by k weight vectors $\theta_1, \ldots, \theta_k$. Given $x \in \mathcal{X}$, the policy predicts as follows: $\pi(x) = \arg\max_{a \in \{1,\ldots,k\}} \{x \cdot \theta_a\}$.

To optimize θ_a, we adapt the "towards-better" version of the direct loss minimization method of McAllester et al. (2011) as follows: given any data $(x, l_1, \ldots, l_k)$ and the current weights θ_a, the weights are adjusted by

$$\theta_{a_1} \leftarrow \theta_{a_1} + \eta x, \qquad \theta_{a_2} \leftarrow \theta_{a_2} - \eta x$$

where $a_1 = \arg\max_a \{x \cdot \theta_a - \epsilon l_a\}$, $a_2 =$

$\arg\max_a \{x \cdot \theta_a\}$, $\eta \in (0, 1)$ is a decaying learning rate, and $\epsilon > 0$ is an input parameter.

For computational reasons, we actually performed batched updates rather than incremental updatess. We found that the learning rate $\eta = t^{-0.3}/2$, where t is the batched iteration, worked well across all datasets. The parameter ϵ was fixed to 0.1 for all datasets. Updates continued until the weights converged.

Furthermore, since the policy loss is not convex in the weight vectors, we repeated the algorithm 20 times with randomly perturbed starting weights and then returned the best run's weight according to the learned policy's loss in the training data. We also tried using a holdout validation set for choosing the best weights out of the 20 candidates, but did not observe benefits from doing so.

B. Filter Tree

The Filter Tree (Beygelzimer et al., 2008) is a reduction from cost-sensitive classification to binary classification. Its input is of the same form as for Direct Loss Minimization, but its output is a binary-tree based predictor where each node of the Filter Tree uses a binary classifier—in this case the J48 decision tree implemented in Weka 3.6.4 (Hall et al., 2009). Thus, there are 2-class decision trees in the nodes, with the nodes arranged as per a Filter Tree. Training in a Filter Tree proceeds bottom-up, with each trained node filtering the examples observed by its parent until the entire tree is trained.

Testing proceeds root-to-leaf, implying that the test time computation is logarithmic in the number of classes. We did not test the all-pairs Filter Tree, which has test time computation linear in the class count similar to DLM.

References

Asuncion, A. and Newman, D. J. UCI machine learning repository, 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Computing*, 32(1):48–77, 2002.

Beygelzimer, A. and Langford, J. The offset tree for learning with partial labels. In *KDD*, pp. 129–138, 2009.

Beygelzimer, A., Langford, J., and Ravikumar, P. Multiclass classification with filter-trees. Unpublished technical report: http://www.stat.berkeley.edu/∼pradeepr/paperz/filter-tree.pdf, 2008.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620, 1976.

Chan, D., Ge, R., Gershony, O., Hesterberg, T., and Lambert, D. Evaluating online ad campaigns in a pipeline: causal models at scale. In *KDD*, 2010.

Gretton, A., Smola, A. J., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Dataset shift in machine learning. In *Covariate Shift and Local Learning by Distribution Matching*, pp. 131–160. MIT Press, 2008.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

Hazan, E. and Kale, S. Better algorithms for benign bandits. In *SODA*, pp. 38–47, 2009.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47:663–685, 1952.

Kang, J. D. Y. and Schafer, J. L. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22(4):523–539, 2007. With discussions.

Lambert, D. and Pregibon, D. More bang for their bucks: assessing new features for online advertisers. In *ADKDD*, 2007.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. In *NIPS*, pp. 1096–1103, 2008.

Langford, J., Strehl, A. L., and Wortman, J. Exploration scavenging. In *ICML*, pp. 528–535, 2008.

Lunceford, J. K. and Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

McAllester, D., Hazan, T., and Keshet, J. Direct loss minimization for structured prediction. In *NIPS*, pp. 1594–1602, 2011.

Robins, J. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, 90:122–129, 1995.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89(427):846–866, 1994.

Strehl, A., Langford, J., Li, L., and Kakade, S. Learning from logged implicit exploration data. In *NIPS*, pp. 2217–2225, 2011.