Offline Policy Evaluation with New Arms

Ben London
blondon@amazon.com
Amazon

Thorsten Joachims
thorstj@amazon.com
Amazon

Abstract

We study offline policy evaluation in a setting where the target policy can take actions that were not available when the data was logged. We analyze the bias of two popular regression-based estimators in this setting, and upper-bound their biases by a quantity we refer to as the *reward regression risk*. We show that the estimators can be asymptotically unbiased and uniformly convergent if the reward regression risk asymptotically goes to zero. We then upper-bound the reward regression risk using tools from domain adaptation. This analysis motivates using domain adaptation algorithms to train reward predictors for offline policy evaluation. It also suggests future directions for developing improved offline policy optimization algorithms.

1 Introduction

Offline evaluation is a cornerstone of industrial machine learning systems, such as recommendation engines and online advertising platforms, as it enables rapid experimentation without impacting user experiences. Typically, it involves using logged user interactions with an existing policy to estimate the performance of a new policy. This *off-policy* setting is fundamentally a counterfactual inference problem, since it requires us to reason about how the system would have performed had we used the new policy. Much work [26, 13, 28, 29, 49, 19, 40, 47, 1, 51, 14, 17, 25, 24, 30, 44, 50] has been devoted to this problem, drawing on techniques from causal inference, contextual bandits and reinforcement learning. Importantly, most of this work assumes that the actions (sometimes referred to as *arms*) available to the new policy have not changed since the data was logged. However, it is often the case that new arms become available (such as new music, movies or advertisements), and we would like to be able to estimate how a policy using those arms will perform. This setting is the focus of our work.

Off-policy evaluation strategies broadly fall into two categories: *Monte Carlo (MC) methods*, which use observed rewards to directly approximate the expected reward; and *reward regressors*, which estimate a function to predict the expected reward. Both methods face challenges and uncertainties when dealing with new arms. For MC methods, estimating the reward of a new arm is fundamentally impossible (without additional assumptions), since feedback for the new arm has never been observed. The situation is slightly better for reward regressors, since they can impute missing observations. However, predicting the reward of a new arm relies on an assumption that a predictor can generalize to unseen examples—and, in this case, the examples may come from a distribution that is potentially very different from the one it was trained on. Consequently, reward estimates for new arms will almost certainly be biased.

In this work, we analyze two estimators—*direct method* (DM) and *doubly robust* (DR) [13]—in the presence of new arms. We start by characterizing the inherent bias that this setting introduces, and relate this characterization to uniform convergence (which is important for offline policy *optimization*). We show that both asymptotic unbiasedness and uniform convergence depend on a quantity which we term the *reward regression risk*, which is the reward predictor's expected absolute error with respect to the true mean reward—crucially, under the distribution induced by the target policy.

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2020.

Since this distribution is different from the distribution of logged actions, a reward predictor that is trained on the log data might not generalize to the new distribution. We view this problem through the lens of *domain adaptation* [12], a setting in which abundant labeled data is available for a *source* distribution, while only unlabeled data is available for the *target* distribution. Via this connection, we are able to upper-bound the reward regression risk using tools from domain adaptation theory [38]. Our analysis suggests that domain adaptation algorithms might yield reward predictors that generalize better.

One challenging aspect of our setting—especially for policy optimization—is that the target policy might be undetermined at the time of training the reward predictor. This complicates domain adaptation algorithms, which typically assume that the target distribution is fixed. We discuss potential solutions to this problem, but leave it as an area for future work.

2 Preliminaries

Let $\mathcal{X}$ denote a set of *contexts*, such as user requests to a content streaming service. Let $\mathcal{A}(x)$ denote a set of *actions* (a.k.a. *arms*)—such as recommending a certain piece of content—that can be taken in response to a given context, $x \in \mathcal{X}$. For notational convenience, let $\mathcal{A} \triangleq \bigcup_{x \in \mathcal{X}} \mathcal{A}(x)$. Let $\rho : \mathcal{X} \times \mathcal{A} \to [0, 1]$ denote a *reward function*, which measures the goodness of a given action in a given context; e.g., whether the user clicked on or played the recommended content.

A *policy*, π, defines a mapping from contexts to actions. As this mapping may be stochastic, we can think of a policy as inducing a distribution over actions from which we can sample. We write $\pi(x)$ to denote the conditional distribution on $\mathcal{A}(x)$ given x; and similarly, we write $\pi(a \,|\, x)$ to denote the conditional probability of $a \in \mathcal{A}(x)$ given x. Note that when a policy is deterministic, $\pi(a \,|\, x) = 1$ for exactly one action and zero for others.

A policy interacts with the environment through the following process. At each round of interaction, the environment samples a context, x, and reward function, ρ, from a stationary distribution, $\mathbb{D}$; we denote this draw by $(x, \rho) \sim \mathbb{D}$. We also write $x \sim \mathbb{D}_x$ to denote the marginal distribution of contexts, and $\rho \sim \mathbb{D}_{\rho|x}$ to denote the conditional distribution of rewards given a context. Given x, the policy samples an action, $a \sim \pi(x)$, and receives a corresponding reward, $\rho(x, a)$. Since the reward for a given context and action is random, we are interested in an action's mean reward, $\bar{\rho}(x, a) \triangleq \mathbb{E}_{\rho \sim \mathbb{D}_{\rho|x}}[\rho(x, a)]$.

Suppose we have previously launched a policy, π_0, for the purpose of data collection. We will refer to π_0 as the *logging policy*. We denote the set of actions available to the logging policy by $\mathcal{A}_0$. The logging policy may be of any parametric or nonparametric form; all we require from it is that it is randomized and has *full support* on the action set—meaning, for any $x \in \mathcal{X}$ and $a \in \mathcal{A}_0(x)$, it has $\pi_0(a \,|\, x) > 0$. We let the logging policy run for n rounds of interaction and collect a dataset, $S \triangleq (x_i, a_i, p_i, r_i)_{i=1}^n$, where $x_i, \rho_i \sim \mathbb{D}$, $a_i \sim \pi_0(x_i)$, $p_i \triangleq \pi_0(a_i \,|\, x_i)$ and $r_i \triangleq \rho_i(x_i, a_i)$. For ease of exposition, we may write the distribution of S as $(\mathbb{D} \times \pi_0)^n$.

Now, suppose we add some new actions to $\mathcal{A}_0$; let $\mathcal{A}_1 \supset \mathcal{A}_0$ denote the new action set.[1] Given a new policy to be evaluated (sometimes referred to as the *target policy*), π_1, that induces a distribution on $\mathcal{A}_1$, our goal is to estimate its expected reward,

$$R(\pi_1) \triangleq \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi_1(x)} [\rho(x, a)] = \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x} \mathop{\mathbb{E}}_{a\sim\pi_1(x)} [\bar{\rho}(x, a)].$$

We consider two estimators, both of which rely on having a function that, given a context and action, *predicts* the mean reward. We write $h : \mathcal{X} \times \mathcal{A} \to [0, 1]$ to denote such a function. The predictor could come from anywhere, but it is typically learned. For example, h could be the minimizer of a (regularized) least squares objective, trained on a separate sample of log data.[2] We refer to estimators in this family as *reward regressors*.

The first—and arguably simplest—estimator we consider is the so-called *direct method* (DM). The DM estimator works by predicting the mean reward for each action, then weighting by the target

[1] One way to look at this setting is that there was only ever action set $\mathcal{A}_1$ and the logging policy simply had *deficient support* [39], but it will be instructive to explicitly consider separate action sets.

[2] Conceivably, one could train the reward predictor using the same data that is used to evaluate the target policy, though this would introduce a source of bias.

policy's distribution. For a target policy, π_1, log dataset, S, and reward predictor, h, we denote the DM estimator by

$$\hat{R}_{\text{DM}}(\pi_1, S; h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{a \sim \pi_1(x_i)} [h(x_i, a)].$$

A similar, yet more sophisticated, method is the *doubly robust* (DR) estimator. DR combines reward regression with Monte Carlo estimation to get the "best of both worlds." Essentially, DR corrects the residual error of the reward prediction whenever the true reward is observed. Typically, DR does not assume access to the true propensities, instead using an estimate of the logging policy to approximate the propensities. Here, we consider a variant of the DR estimator that uses exact propensities:

$$\hat{R}_{\text{DR}}(\pi_1, S; h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{a \sim \pi_1(x_i)} \left[\frac{\mathbb{1}\{a = a_i\}}{p_i} (r_i - h(x_i, a)) + h(x_i, a) \right].$$

We want to guarantee that these estimators possess certain statistical properties, which we now formalize. In the following definitions, let Π_0 denote a class of logging policies with full support on $\mathcal{A}_0$, and let Π_1 denote a class of target policies that use $\mathcal{A}_1$. Given a target policy, $\pi_1 \in \Pi_1$, and a log dataset, S, collected under $\pi_0 \in \Pi_0$, let $\hat{R}(\pi_1, S)$ denote a generic estimator for the target policy's expected reward, $R(\pi_1)$.

Definition 1. A reward estimator is *unbiased* if, for all $\mathbb{D}$, $n \geq 1$, $\pi_0 \in \Pi_0$ and $\pi_1 \in \Pi_1$,

$$\mathop{\mathbb{E}}_{S \sim (\mathbb{D} \times \pi_0)^n} [\hat{R}(\pi_1, S)] = R(\pi_1).$$

Definition 2. A reward estimator is *asymptotically unbiased* if, for all $\mathbb{D}$, $\pi_0 \in \Pi_0$ and $\pi_1 \in \Pi_1$,

$$\lim_{n \to \infty} \mathop{\mathbb{E}}_{S \sim (\mathbb{D} \times \pi_0)^n} [\hat{R}(\pi_1, S)] = R(\pi_1).$$

Definition 3. A reward estimator *uniformly converges in probability* if, for any $\mathbb{D}$, $\pi_0 \in \Pi_0$, $\epsilon > 0$ and $\delta \in (0, 1)$, there exists an integer, $n_0(\epsilon, \delta)$, for which all $n > n_0(\epsilon, \delta)$ satisfy

$$\Pr_{S \sim (\mathbb{D} \times \pi_0)^n} \left\{ \sup_{\pi_1 \in \Pi_1} \left| \hat{R}(\pi_1, S) - R(\pi_1) \right| > \epsilon \right\} \leq \delta.$$

Unbiasedness guarantees that, on average (over realizations of the log data), the estimator is an accurate prediction of the true reward. Asymptotic unbiasedness only guarantees that the bias vanishes as the dataset grows—yet this can be very useful if the decay rate is fast (e.g., $\mathrm{O}(n^{-1})$). Note that unbiasedness implies asymptotic unbiasedness.

Uniform convergence in probability is a useful property for both offline evaluation and *optimization*. If the tail bound holds simultaneously for all policies in Π_1, then we can optimize the estimator to find the best target policy.

Our goal is to identify conditions under which these properties hold for the DM and DR estimators.

3 Reward Regression Risk

In this section, we characterize bias and uniform convergence using a quantity we refer to as the *reward regression risk*. We will show that this quantity is critical to offline evaluation in the presence of new arms.

For a given target policy, π_1, a reward predictor, h, and a set of *excluded* arms, $\mathcal{E}$ (to be defined in the sequel), we define the reward regression risk as the predictor's expected absolute error on a new example, (x, a), drawn from $(\mathbb{D}_x \times \pi_1)$, when a is not in $\mathcal{E}(x)$ (denoting the arms excluded in the given context):

$$\mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{E}) \triangleq \mathop{\mathbb{E}}_{x \sim \mathbb{D}_x} \mathop{\mathbb{E}}_{a \sim \pi_1(x)} [\mathbb{1}\{a \notin \mathcal{E}(x)\} \, |h(x, a) - \bar{\rho}(x, a)|]. \tag{1}$$

When $\mathcal{E} = \emptyset$ (i.e., no arms are excluded), we will simply write $\mathcal{L}_{\pi_1}(h, \bar{\rho}) \triangleq \mathcal{L}_{\pi_1}(h, \bar{\rho}; \emptyset)$. Since the indicator function and absolute error are both nonnegative, we have that $\mathcal{L}_{\pi_1}(h, \bar{\rho}) \geq \mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{E})$.

Recall that h is typically *trained* to predict mean reward, $\bar{\rho}$, using supervised learning—albeit under the logging distribution, π_0, not the target distribution, π_1. From this perspective, the reward regression risk measures the predictor's ability to generalize from the training data to the target distribution. The remainder of this section will show that asymptotic unbiasedness and uniform convergence both depend on $\mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{E})$ being $\mathrm{o}(1)$. While this is rather straightforward in standard supervised learning, the challenge here is that the target distribution is determined by a target policy, which may be very different from the policy that generated the training data. Indeed, the fact that the target policy has access to actions that were not available to the logging policy only exacerbates the problem. We formalize this issue in Section 4 using concepts from domain adaptation.

3.1 Characterizing Bias

We now analyze the biases of the DM and DR estimators for target policies that use new actions. Irrespective of the action set, the DM estimator is unbiased if and only if the reward predictor is an unbiased estimate of the mean reward (conditioned on the context and action) [13, 50]. In contrast, because we have assumed that exact propensities are available, the DR estimator can be unbiased even if the reward predictor is biased [13]—provided the target policy uses the same action set as the logging policy. Yet, when the target policy uses a new, expanded action set, $\mathcal{A}_1 \supset \mathcal{A}_0$, both DM and DR are biased. Moreover, we can upper-bound their biases by the reward regression risk.

Proposition 1. *If π_0 has full support on $\mathcal{A}_0$, but π_1 uses actions $\mathcal{A}_1 \supset \mathcal{A}_0$, then*

$$\underbrace{\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n}\left[\hat{R}_{\mathrm{DM}}(\pi_1, S; h)\right] - R(\pi_1)}_{\textit{DM bias}} = \mathop{\mathbb{E}}_{\substack{x\sim\mathbb{D}_x \\ a\sim\pi_1(x)}}\Big[\underbrace{h(x,a) - \bar{\rho}(x,a)}_{\textit{reward predictor bias}}\Big] \le \mathcal{L}_{\pi_1}(h, \bar{\rho}), \tag{2}$$

and

$$\underbrace{\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n}\left[\hat{R}_{\mathrm{DR}}(\pi_1, S; h)\right] - R(\pi_1)}_{\textit{DR bias}} = \mathop{\mathbb{E}}_{\substack{x\sim\mathbb{D}_x \\ a\sim\pi_1(x)}}\Big[\underbrace{\mathbb{1}\{a \notin \mathcal{A}_0(x)\}\left(h(x,a) - \bar{\rho}(x,a)\right)}_{\textit{reward predictor bias w.r.t. new actions}}\Big] \le \mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{A}_0). \tag{3}$$

A consequence of the above is that, if the reward regression risk is $\mathrm{o}(1)$, then the DM and DR estimators are asymptotically unbiased. It is also of note that the reward regression risk for DR is less than that of DM, since $\mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{A}_0) \le \mathcal{L}_{\pi_1}(h, \bar{\rho})$. We will use this fact in Section 4, where we upper-bound $\mathcal{L}_{\pi_1}(h, \bar{\rho})$, which thereby upper-bounds $\mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{A}_0)$.

3.2 Characterizing Uniform Convergence

We can use the above characterizations of estimator bias to characterize uniform convergence. Uniform convergence in probability requires a bound on $|\hat{R}(\pi_1, S; h) - R(\pi_1)|$ that holds for *all* $\pi_1 \in \Pi_1$ with high probability. To do so will require some additional machinery; in this work, we use a generalization of the *Rademacher complexity* for vector-valued functions.

Definition 4. For an arbitrary class of vector-valued functions, $\mathcal{F} \subseteq \{\mathcal{Z} \to \mathbb{R}^k\}$, and a dataset, $S \in \mathcal{Z}^n$, let

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \sigma_{i,j} f_j(z_i)\right]$$

denote the *empirical Rademacher complexity*, where each $\sigma_{i,j}$ is an independent random variable, uniformly distributed over $\{\pm 1\}$. Let $\mathfrak{R}_n \triangleq \mathbb{E}_S[\hat{\mathfrak{R}}_S(\mathcal{F})]$ denote the *Rademacher complexity*. Note that we obtain the traditional Rademacher complexity when $k = 1$.

In the following, we assume that the true and predicted rewards are bounded, and that the propensities under the logging policy are uniformly lower-bounded by some value. We argue that both assumptions are reasonable because the reward function and logging policy are typically user-defined.

Proposition 2. *Assume that $\tau \triangleq \inf_{x\in\mathcal{X},\, a\in\mathcal{A}_0(x)} \pi_0(a \,|\, x) > 0$, and let $K \triangleq \max_{x\in\mathcal{X}} |\mathcal{A}_1(x)|$. Then, for any $n \ge 1$ and $\delta \in (0, 1)$:*

1. *with probability at least* $1-\delta$ *over draws of* $S \sim (\mathbb{D} \times \pi_0)^n$, *all* $\pi_1 \in \Pi_1$ *satisfy*

$$\left|\hat{R}_{\text{DM}}(\pi_1, S; h) - R(\pi_1)\right| \leq \mathcal{L}_{\pi_1}(h, \bar{\rho}) + 2\sqrt{K}\mathfrak{R}_n(\Pi_1) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}};$$

2. *with probability at least* $1-\delta$ *over draws of* $S \sim (\mathbb{D} \times \pi_0)^n$, *all* $\pi_1 \in \Pi_1$ *satisfy*

$$\left|\hat{R}_{\text{DR}}(\pi_1, S; h) - R(\pi_1)\right| \leq \mathcal{L}_{\pi_1}(h, \bar{\rho}; \mathcal{A}_0) + \frac{2\sqrt{K}}{\tau}\mathfrak{R}_n(\Pi_1) + \frac{2-\tau}{\tau}\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

Proposition 2 tells us that the estimators uniformly converge in probability if the reward regression risk and the Rademacher complexity decrease as a function of n. There are many analyses of Rademacher complexity [2, 23, 37] that can be applied to control this term. Thus, the remaining question is whether the reward regression risk is $\mathrm{o}(1)$.

4 Interpretation as Domain Adaptation

As shown in the previous section, asymptotic unbiasedness and uniform convergence depend on the reward predictor's ability to generalize. Unfortunately, it must generalize to actions that it has never seen rewards for, distributed according to a policy that is likely different from the one that generated the training data. One way to view this problem is through the lens of *domain adaptation* [12]. In domain adaptation, abundant training data is available from a *source distribution*, $\mathbb{P}$, while little or no training data is available from the *target distribution* of interest, $\mathbb{Q}$. In our setting, the source distribution is determined by the logging policy, while the target distribution is determined by the target policy.

Many theoretical studies have analyzed the risk of domain adaptation (e.g., [3, 5, 34, 4, 16, 8, 11, 52]; see [38] for a survey). In these studies, it is customary to define risk with respect to a deterministic labeling function, $f : \mathcal{Z} \to \mathcal{Y}$, that provides the true label for any example. The labeling function may be domain-specific, but for our purposes it is unnecessary to make this distinction. Importantly, the labeling function may not be in the class of hypotheses, $\mathcal{H} \subseteq \{\mathcal{Z} \to \mathcal{Y}\}$, being considered.

For a loss function, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and a given hypothesis, $h \in \mathcal{H}$, the target risk,

$$\mathcal{L}_{\mathbb{Q}}(h, f) \triangleq \mathop{\mathbb{E}}_{z \sim \mathbb{Q}}[\ell(h(z), f(z))],$$

is the expected loss under $\mathbb{Q}$. This modular notation can also denote risk under an alternative distribution, such as the source risk, $\mathcal{L}_{\mathbb{P}}(h, f)$; or, risk with respect to a different labeling function, such as another hypothesis, $\mathcal{L}_{\mathbb{Q}}(h, h')$, for $h' \in \mathcal{H}$. Notably, we can recreate the reward regression risk (for $\mathcal{E} = \emptyset$) with $\mathbb{Q} = (\mathbb{D}_x \times \pi_1)$, $f = \bar{\rho}$ and $\ell(y, y') = |y - y'|$.

A typical risk bound for domain adaptation involves a term that captures how much the source distribution differs from the target distribution. In this paper, we use a formalism known as *discrepancy* [34, 8, 11]. Unlike common *divergence* measures (such as χ^2 or Kullback-Liebler), a discrepancy accounts for properties of the learning problem, such as the hypothesis class and loss function. Moreover, discrepancy is a *pseudometric*, since it obeys the axioms of identity, symmetry and subadditivity (i.e., the triangle inequality), but not necessarily distinguishability. Our analysis will use the following definition of discrepancy.

Definition 5 ([34])**.** For a hypothesis class, $\mathcal{H}$, and a loss function, ℓ, the *discrepancy* between distributions $\mathbb{P}$ and $\mathbb{Q}$ is

$$\operatorname{disc}(\mathbb{P}, \mathbb{Q}) \triangleq \sup_{h, h' \in \mathcal{H}} \left|\mathcal{L}_{\mathbb{P}}(h, h') - \mathcal{L}_{\mathbb{Q}}(h, h')\right|. \tag{4}$$

Using the discrepancy, with $\ell(y, y') = |y - y'|$, we can upper-bound the reward regression risk. To improve readability, we omit $\mathbb{D}_x$ from our notation of the source and target distributions.

Proposition 3. *Let* $\mathcal{H} \subseteq \{\mathcal{X} \times \mathcal{A} \mapsto [0, 1]\}$ *denote a class of reward predictors. For any* $n \geq 1$ *and* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ *over draws of* $S \triangleq (x_i, a_i, r_i)_{i=1}^n \sim (\mathbb{D} \times \pi_0)^n$, *the* *following holds for all $h \in \mathcal{H}$ and $\pi_1 \in \Pi_1$:*

$$\mathcal{L}_{\pi_1}(h, \bar{\rho}) \leq \frac{1}{n} \sum_{i=1}^{n} |h(x_i, a_i) - r_i| + \lambda_{\mathcal{H}}(\pi_0, \pi_1) + \operatorname{disc}(\pi_0, \pi_1) + 2\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}, \quad (5)$$

$$\textit{where} \quad \lambda_{\mathcal{H}}(\pi_0, \pi_1) \triangleq \min_{h \in \mathcal{H}} \mathcal{L}_{\pi_0}(h, \bar{\rho}) + \mathcal{L}_{\pi_1}(h, \bar{\rho}). \quad (6)$$

Our analysis is inspired by [11, Proposition 5], but we use a different discrepancy and change the order of quantifiers to hold uniformly over all target distributions. Another useful feature of our bound is that, because the labeling function happens to be an expectation of random labelings, we are able to state the empirical risk in terms of the (noisy) observed rewards. This stands in contrast to some other domain adaptation bounds [3, 5, 34, 4, 8, 11], which use the empirical risk with respect to the labeling function. In our setting, the labeling function (i.e., mean reward) is unobservable.

The term $\lambda_{\mathcal{H}}(\pi_0, \pi_1)$—which has appeared in prior work [3, 5, 4, 52]—quantifies the *approximation error*; that is, how well the hypothesis class fits the mean reward function, under both distributions. Importantly, when $\bar{\rho} \in \mathcal{H}$ (i.e., the model is correctly specified), the approximation error is zero.

The Rademacher complexity, $\mathfrak{R}_n(\mathcal{H})$, can be upper-bounded analytically, using properties of the hypothesis class and loss function [2, 23, 37], which typically motivates some form of structural risk minimization (e.g., regularization). If $\mathfrak{R}_n(\mathcal{H})$ is upper-bounded by a decreasing function of n, then the discrepancy can be estimated from data (for a given target policy).

Proposition 4. *For any $\pi_0 \in \Pi_0$, $\pi_1 \in \Pi_1$, $n_0 \geq 1$, $n_1 \geq 1$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $S_0 \sim (\mathbb{D}_x \times \pi_0)^{n_0}$ and $S_1 \sim (\mathbb{D}_x \times \pi_1)^{n_1}$,*

$$|\operatorname{disc}(\pi_0, \pi_1) - \operatorname{disc}(\mathbb{S}_0, \mathbb{S}_1)| \leq 4\mathfrak{R}_{n_0}(\mathcal{H}) + 4\mathfrak{R}_{n_1}(\mathcal{H}) + \sqrt{\frac{\ln \frac{4}{\delta}}{2n_0}} + \sqrt{\frac{\ln \frac{4}{\delta}}{2n_1}},$$

where $\mathbb{S}_0$ and $\mathbb{S}_1$ are the empirical distributions of S_0 and S_1.

Proposition 4 is similar to [34, Corollary 7], but uses the Rademacher complexity instead of the empirical Rademacher complexity.

4.1 Optimizing Discrepancy

Equation 5 tells us that the reward regression risk is reduced when the discrepancy between the source and target distributions is small. This leads us to ask whether it is possible to *optimize* the source distribution to minimize discrepancy. To be clear, the source distribution is fixed and cannot be modified, but we can train a reward predictor against an alternative empirical distribution using the *weighted* empirical risk, $\frac{1}{n} \sum_{i=1}^{n} w_i \, |h(x_i, a_i) - r_i|$, where $w_i \geq 0$ is an *importance weight*. This motivates a two-step procedure for training a reward predictor that adapts to the target policy: (1) optimize an alternative source distribution (i.e., policy), $\pi_0' \in \Pi_0$, to minimize the discrepancy, $\operatorname{disc}(\pi_0', \pi_1)$; (2) use the weighted empirical risk, with $w_i = \pi_0'(a_i \,|\, x_i)/p_i$, to train a reward predictor. To account for optimizing π_0', we adapt Proposition 3 to hold uniformly for all $\pi_0' \in \Pi_0$.

Proposition 5. *Fix $\pi_0 \in \Pi_0$, and assume that $\tau \triangleq \inf_{x \in \mathcal{X},\, a \in \mathcal{A}_0(x)} \pi_0(a \,|\, x) > 0$. Then, for any $n \geq 1$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $S \triangleq (x_i, a_i, p_i, r_i)_{i=1}^{n} \sim (\mathbb{D} \times \pi_0)^n$, the following holds for all $h \in \mathcal{H}$, $\pi_0' \in \Pi_0$ and $\pi_1 \in \Pi_1$:*

$$\mathcal{L}_{\pi_1}(h, \bar{\rho}) \leq \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_0'(a_i \,|\, x_i)}{p_i} \, |h(x_i, a_i) - r_i| + \lambda_{\mathcal{H}}(\pi_0', \pi_1) + \operatorname{disc}(\pi_0', \pi_1) + \frac{2}{\tau} \mathfrak{R}_n(\mathcal{H}) + \frac{1}{\tau} \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

By Proposition 4, for a given π_0' and π_1, we can estimate $\operatorname{disc}(\pi_0', \pi_1)$ using unlabeled datasets, $S_0' \sim (\mathbb{D}_x \times \pi_0')^{n_0}$ and $S_1 \sim (\mathbb{D}_x \times \pi_1)^{n_1}$. We can thus replace $\operatorname{disc}(\pi_0', \pi_1)$ in step (1) of the above procedure with the empirical discrepancy, $\operatorname{disc}(\mathbb{S}_0', \mathbb{S}_1)$. This approach was originally proposed by Mansour et al. [34]; we refer the reader to their work for methods to perform the optimization.

An issue with the above approach is that it assumes that the target distribution (i.e., target policy) is known when training the reward predictor. One aspect that distinguishes our setting from that of traditional domain adaptation is that the target distribution may be undetermined at this point, since the target policy may be derived from the reward predictor. Such is the case in offline policy optimization, where the target policy is trained to maximize a reward estimator. If the estimator requires a reward predictor, then optimally training the reward predictor depends on the target policy. We are thus faced with a circular dependency—a so-called "chicken-and-egg" problem—in which each optimization depends on the other.

We propose to break this dependency by introducing an intermediate distribution, $\mathbb{S}_1'$, which acts as a surrogate for $\mathbb{S}_1$ when training the reward predictor. It is reasonable to assume that the new action set, $\mathcal{A}_1$, is known at that point; thus, given any sample of contexts from the environment, we can construct unlabeled examples from the target domain—albeit not distributed by the target distribution. For an intermediate policy, $\pi_1' \in \Pi_1$, and a sample of contexts, $(x_i)_{i=1}^{n_1} \sim \mathbb{D}_x$, we can sample an action for each context, $a_i \sim \pi_1'(x_i)$, to create an unlabeled dataset, S_1', with empirical distribution $\mathbb{S}_1'$. We can then optimize an empirical source distribution, $\mathbb{S}_0'$, to minimize $\operatorname{disc}(\mathbb{S}_0', \mathbb{S}_1')$, and use $\mathbb{S}_0'$ to train a reward predictor (via weighted empirical risk minimization). This predictor may not be optimal for π_1, but it may be "good enough" if π_1' is "close enough" to π_1. We motivate this idea by noting that $\operatorname{disc}(\mathbb{S}_0', \mathbb{S}_1) \leq \operatorname{disc}(\mathbb{S}_0', \mathbb{S}_1') + \operatorname{disc}(\mathbb{S}_1', \mathbb{S}_1)$, via the triangle inequality. If π_1 is close to π_1', then $\operatorname{disc}(\mathbb{S}_1', \mathbb{S}_1)$ will be small.

The question then becomes how to select π_1' such that it will be close to an as yet undetermined π_1. This problem is conceptually similar to selecting a *prior* in Bayesian learning. If we have prior knowledge of what π_1 might be, then we can make an informed guess. Lacking any such prior knowledge, we could make π_1' uniformly random on $\mathcal{A}_1$. Our choice of π_1' can also inform how we select π_1. If π_1 is optimized with respect to an estimator based on h, which depends on π_1', then constraining π_1 to be close to π_1' ensures that the optimization objective is accurate. This can be accomplished via regularization [45, 32] or trust-region methods [41].

5 Related Work

Offline policy evaluation (and optimization) has been studied extensively in the literature on contextual bandits [26, 13, 28, 29, 1, 51, 17, 30, 50], reinforcement learning [33, 49, 19, 47, 14, 7] and counterfactual learning [43, 6, 45, 46, 40, 20, 21, 25, 24, 32, 44]. Many of these works analyzed the bias, variance and uniform convergence of reward estimators. Our work differs in that we explicitly consider a scenario in which the action set is expanded after data collection.

This setting has not received attention until very recently. Notably, Sachdeva et al. [39] explored policy optimization in this setting, but view it as the logging policy having deficient support. They proposed several learning strategies, one of which involves an estimator that is reminiscent of DR, and independently arrived at a characterization of the bias that mirrors our own. However, their characterization does not provide guidance on how to assess or minimize the bias.

Support deficiency has multiple manifestations. In our work, the logging policy's support is a subset of the target policy's support, but the reverse could be true. Thomas and Brunskill [48] studied importance weighting in a setting in which the target distribution's support is a subset of the source (i.e., logging) distribution's support. Support deficiency also arises in off-policy (not necessarily offline) reinforcement learning. With rich action spaces, the target policy will almost certainly select actions not taken by the behavioral policy. To prevent this from happening, one can constrain the policy space to be close to the behavioral policy [41], or constrain the action space to be similar to the actions selected by the behavioral policy [31]. Alternatively, one can simply impute zero reward for actions not taken by the behavioral policy [15].

A related problem in on-policy reinforcement learning is being able to adapt when presented with novel actions [18]. However, the offline (off-policy) setting presents different challenges; since interaction with the environment is impossible, bias is a factor.

We are not the first to recognize a connection between off-policy evaluation and domain adaptation (specifically, the subproblem of *covariate shift* [42]); for example, it is noted in [29, 46, 22, 30]. The closest prior work to our own is from Johansson et al. [22], who applied Cortes and Mohri's [8] domain adaptation theory to motivate a representation learning approach to counterfactual inference. It is also worth highlighting Liu et al.'s [30] improved DR estimator (dubbed *triply robust*), which addresses the covariate shift problem using robust reward regression.

6 Discussion and Future Work

While a substantial body of work has previously focused on reducing the variance of policy evaluation, bias remains a key source of error on problems where some arms were not observed during data collection. To address this problem, we characterized the bias that DM and DR can incur on support-deficient policy evaluation problems. In particular, we identified the reward regression risk as the key quantity for characterizing bias, and leveraged work in domain adaptation to derive bounds on the reward regression risk. This opens up new avenues of research for developing improved policy evaluation and optimization algorithms.

In particular, the problem of how to train an optimal reward predictor for downstream policy optimization is ripe for future work. Further, we plan to derive new bounds that account for the variance caused by importance weighting (both in reward regression and policy evaluation), such as empirical Bernstein bounds [35] or those based on the Rényi divergence: [9]; this will provide a full characterization of estimator error that can be optimized to control both the bias and variance.

References

[1] A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. Effective evaluation using logged bandit feedback from multiple loggers. *Knowledge Discovery and Data Mining*, 2017.

[2] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3, 2003.

[3] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Neural Information Processing Systems*, 2007.

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79, 2010.

[5] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Neural Information Processing Systems*, 2008.

[6] L. Bottou, J. Peters, J. Qui nonero Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14, 2013.

[7] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. Chi. Top-K off-policy correction for a REINFORCE recommender system. *Web Search and Data Mining*, 2019.

[8] C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519, 2014.

[9] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Neural Information Processing Systems*, 2010.

[10] C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang. Structured prediction theory based on factor graph complexity. In *Neural Information Processing Systems*, 2016.

[11] C. Cortes, M. Mohri, and A. Medina. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20, 2019.

[12] H. Daumè and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26, 2006.

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

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

[15] S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. In *ICML*, 2019.

[16] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning*, 2013.

[17] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. Offline A/B testing for recommender systems. *Web Search and Data Mining*, 2018.

[18] A. Jain, A. Szot, and J. Lim. Generalization to new actions in reinforcement learning. In *International Conference on Machine Learning*, 2020.

[19] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

[20] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *ACM Conference on Web Search and Data Mining*, 2017.

[21] T. Joachims, A. Swaminathan, and M. de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

[22] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, 2016.

[23] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Neural Information Processing Systems*, 2008.

[24] N. Kallus. Balanced policy evaluation and learning. In *Neural Information Processing Systems*, 2018.

[25] N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *Artificial Intelligence and Statistics*, 2018.

[26] J. Langford, A. Strehl, and J. Vaughan. Exploration scavenging. In *International Conference on Machine Learning*, 2008.

[27] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, 1991.

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

[29] L. Li, R. Munos, and C. Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, 2015.

[30] A. Liu, H. Liu, A. Anandkumar, and Y. Yue. Triply robust off-policy evaluation. *CoRR*, abs/1911.05811, 2019.

[31] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Off-policy policy gradient with state distribution correction. In *Uncertainty in Artificial Intelligence*, 2019.

[32] B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, 2019.

[33] A. Mahmood, H. Hasselt, and R. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Neural Information Processing Systems*, 2014.

[34] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, 2009.

[35] A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Computational Learning Theory*, 2009.

[36] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 1989.

[37] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012. ISBN 978-0-262-01825-8.

[38] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *CoRR*, abs/2004.11829, 2020.

[39] N. Sachdeva, Y. Su, and T. Joachims. Off-policy bandits with deficient support. In *Knowledge Discovery and Data Mining*, 2020.

[40] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, 2016.

[41] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

[42] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 2000.

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

[44] Y. Su, L. Wang, M. Santacatterina, and T. Joachims. CAB: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, 2019.

[45] A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16, 2015.

[46] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems*, 2015.

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

[48] P. Thomas and E. Brunskill. Importance sampling with unequal support. In *AAAI*, 2017.

[49] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, 2015.

[50] N. Vlassis, A. Bibaut, M. Dimakopoulou, and T. Jebara. On the design of estimators for bandit off-policy evaluation. In *International Conference on Machine Learning*, 2019.

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

[52] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, 2019.

A Deferred Proofs

This appendix contains proofs deferred from the main manuscript.

A.1 Proof of Proposition 1

Using linearity of expectation, we can express the bias of the DM estimator as the bias of the reward predictor:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n}\left[\hat{R}_{\mathrm{DM}}(\pi_1,S;h)\right] - R(\pi_1) &= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n}\mathop{\mathbb{E}}_{a\sim\pi_1(x_i)}[h(x_i,a)] - \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[\rho(x,a)] \\
&= \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[h(x,a)] - \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[\rho(x,a)] \\
&= \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[h(x,a) - \mathop{\mathbb{E}}_{\rho\sim\mathbb{D}_{\rho|x}}[\rho(x,a)]\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[h(x,a) - \bar{\rho}(x,a)].
\end{aligned}$$

We then apply Jensen's inequality to get the reward regression risk:

$$\mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[h(x,a) - \bar{\rho}(x,a)] \leq \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[|h(x,a) - \bar{\rho}(x,a)|] = \mathcal{L}_{\pi_1}(h,\bar{\rho};\emptyset) = \mathcal{L}_{\pi_1}(h,\bar{\rho}).$$

For the DR estimator, the bias is characterized by the bias of the reward predictor on the new actions:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n}\left[\hat{R}_{\mathrm{DR}}(\pi_1,S;h)\right] - R(\pi_1) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n}\mathop{\mathbb{E}}_{a\sim\pi_1(x_i)}\left[\frac{\mathbb{1}\{a=a_i\}}{p_i}\left(r_i - h(x_i,a)\right) + h(x_i,a)\right] - \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}[\rho(x,a)] \\
&= \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\mathop{\mathbb{E}}_{a'\sim\pi_0(x)}\left[\frac{\mathbb{1}\{a=a'\}}{\pi_0(a'\,|\,x)}\left(\rho(x,a) - h(x,a)\right)\right] + h(x,a) - \rho(x,a)\right] \\
&= \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\left(\sum_{a'\in\mathcal{A}_0(x)}\pi_0(a'\,|\,x)\,\frac{\mathbb{1}\{a=a'\}}{\pi_0(a'\,|\,x)}\left(\rho(x,a) - h(x,a)\right)\right) + h(x,a) - \rho(x,a)\right] \\
&= \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\mathbb{1}\{a\in\mathcal{A}_0(x)\}\left(\rho(x,a) - h(x,a)\right) + h(x,a) - \rho(x,a)\right] \\
&= \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\mathbb{1}\{a\notin\mathcal{A}_0(x)\}\left(h(x,a) - \rho(x,a)\right)\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\mathbb{1}\{a\notin\mathcal{A}_0(x)\}\left(h(x,a) - \bar{\rho}(x,a)\right)\right].
\end{aligned}$$

As before, we use Jensen's inequality:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\mathbb{1}\{a\notin\mathcal{A}_0(x)\}\left(h(x,a) - \bar{\rho}(x,a)\right)\right] &\leq \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_1(x)}\left[\mathbb{1}\{a\notin\mathcal{A}_0(x)\}\left|\left(h(x,a) - \bar{\rho}(x,a)\right)\right|\right] \\
&= \mathcal{L}_{\pi_1}(h,\bar{\rho};\mathcal{A}_0).
\end{aligned}$$

Thus completes the proof.

A.2 Proof of Proposition 2

Most of the proof is the same for the DM and DR estimators, so we can consider a generic estimator, $\hat{R}$. Since the reward predictor, h, is given, we omit it from our notation for the remainder of the proof. To reduce notation further, we will omit the subscript when taking expectations over $S\sim(\mathbb{D}\times\pi_0)^n$.

Via the triangle inequality, we have that

$$\left|\hat{R}(\pi,S) - R(\pi)\right| \leq \left|\hat{R}(\pi,S) - \mathbb{E}[\hat{R}(\pi,S)]\right| + \left|\mathbb{E}[\hat{R}(\pi,S)] - R(\pi)\right|.$$

The righthand term is the estimator's bias, which we characterized in Section 3.1 for DM and DR. The lefthand term is the difference of the estimator and its expectation; and since the estimator is essentially just an average of bounded, i.i.d. random variables, we can upper-bound this difference with high probability using standard concentration inequalities. However, to make such a bound hold for *all* target policies will require the Rademacher complexity.

For notational convenience, let $z \triangleq (x, a, p, r)$ denote an example. We define function classes

$$\mathcal{F}_{\Pi}^{\text{DM}} \triangleq \left\{ (x, a, p, r) \mapsto \mathop{\mathbb{E}}_{a' \sim \pi(x)} [h(x, a')] : \pi \in \Pi \right\}$$

and

$$\mathcal{F}_{\Pi}^{\text{DR}} \triangleq \left\{ (x, a, p, r) \mapsto \mathop{\mathbb{E}}_{a' \sim \pi(x)} \left[\frac{\mathbb{1}\{a' = a\}}{p} (r - h(x, a')) + h(x, a') \right] : \pi \in \Pi \right\}.$$

For each $\pi \in \Pi$ and its corresponding $f_{\pi}^{\text{DM}} \in \mathcal{F}_{\Pi}^{\text{DM}}$ and $f_{\pi}^{\text{DR}} \in \mathcal{F}_{\Pi}^{\text{DR}}$, we have

$$\hat{R}_{\text{DM}}(\pi, S) = \frac{1}{n} \sum_{i=1}^{n} f_{\pi}^{\text{DM}}(z_i) \quad \text{and} \quad \hat{R}_{\text{DR}}(\pi, S) = \frac{1}{n} \sum_{i=1}^{n} f_{\pi}^{\text{DR}}(z_i).$$

Further, since the rewards are bounded by $[0, 1]$, we have that the range of f_{π}^{DM} is bounded by $[0, 1]$; and since the propensities are lower-bounded by τ, the range of f_{π}^{DR} is bounded by $[1 - \tau^{-1}, \tau^{-1}]$.

We then define a function

$$\Phi_{\mathcal{F}_{\Pi}}(S) \triangleq \sup_{f_{\pi} \in \mathcal{F}_{\Pi}} \frac{1}{n} \sum_{i=1}^{n} f_{\pi}(z_i) - \mathbb{E}[f_{\pi}(z)] = \sup_{\pi \in \Pi} \hat{R}(\pi, S) - \mathbb{E}[\hat{R}(\pi, S)],$$

which is the maximum difference of a generic estimator and its expectation. We can upper-bound the expectation of $\Phi_{\mathcal{F}_{\Pi}}$ using the Rademacher complexity:

$$\mathbb{E}[\Phi_{\mathcal{F}_{\Pi}}(S)] \leq 2\mathfrak{R}_n(\mathcal{F}_{\Pi}).$$

(See [37] for details.) Further, for any two datasets, $S, S' \in \mathcal{Z}^n$, that differ by a single example, we have that

$$\left| \Phi_{\mathcal{F}_{\Pi}^{\text{DM}}}(S) - \Phi_{\mathcal{F}_{\Pi}^{\text{DM}}}(S') \right| \leq \frac{1}{n} \quad \text{and} \quad \left| \Phi_{\mathcal{F}_{\Pi}^{\text{DR}}}(S) - \Phi_{\mathcal{F}_{\Pi}^{\text{DR}}}(S') \right| \leq \frac{2 - \tau}{\tau n},$$

since f_{π}^{DM} and f_{π}^{DR} are uniformly bounded. Thus, by McDiarmid's inequality [36], we have with probability at least $1 - \delta/2$ over draws of $S \sim (\mathbb{D} \times \pi_0)^n$ that

$$\Phi_{\mathcal{F}_{\Pi}^{\text{DM}}}(S) \leq \mathbb{E}[\Phi_{\mathcal{F}_{\Pi}^{\text{DM}}}(S)] + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} \leq 2\mathfrak{R}_n(\mathcal{F}_{\Pi}^{\text{DM}}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}};$$

or, alternatively,

$$\Phi_{\mathcal{F}_{\Pi}^{\text{DR}}}(S) \leq \mathbb{E}[\Phi_{\mathcal{F}_{\Pi}^{\text{DR}}}(S)] + \frac{2 - \tau}{\tau} \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} \leq 2\mathfrak{R}_n(\mathcal{F}_{\Pi}^{\text{DR}}) + \frac{2 - \tau}{\tau} \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

These are one-sided bounds, but we can derive equivalent bounds for

$$\Phi_{\mathcal{F}_{\Pi}}^{-}(S) \triangleq \sup_{f_{\pi} \in \mathcal{F}_{\Pi}} \mathbb{E}[f_{\pi}(z)] - \frac{1}{n} \sum_{i=1}^{n} f_{\pi}(z_i) = \sup_{\pi \in \Pi} \mathbb{E}[\hat{R}(\pi, S)] - \hat{R}(\pi, S).$$

Since

$$\sup_{\pi \in \Pi} \left| \hat{R}(\pi, S) - \mathbb{E}[\hat{R}(\pi, S)] \right| = \max \left\{ \Phi_{\mathcal{F}_{\Pi}}(S), \Phi_{\mathcal{F}_{\Pi}}^{-}(S) \right\},$$

we have with probability at least $1 - \delta$ that

$$\sup_{\pi \in \Pi} \left| \hat{R}_{\text{DM}}(\pi, S) - \mathbb{E}[\hat{R}_{\text{DM}}(\pi, S)] \right| \leq 2\mathfrak{R}_n(\mathcal{F}_{\Pi}^{\text{DM}}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}};$$

or, alternatively,

$$\sup_{\pi \in \Pi} \left| \hat{R}_{\text{DR}}(\pi, S) - \mathbb{E}[\hat{R}_{\text{DR}}(\pi, S)] \right| \leq 2\mathfrak{R}_n(\mathcal{F}_{\Pi}^{\text{DR}}) + \frac{2 - \tau}{\tau} \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

Putting it all together, for any $\delta \in (0, 1)$, with probability at least $1-\delta$ over draws of $S \sim (\mathbb{D} \times \pi_0)^n$, the following holds for any $\pi_1 \in \Pi_1$:

$$\begin{aligned}\left|\hat{R}_{\text{DM}}(\pi_1, S) - R(\pi_1)\right| &\leq \left|\mathbb{E}[\hat{R}_{\text{DM}}(\pi_1, S)] - R(\pi_1)\right| + \sup_{\pi \in \Pi_1} \left|\hat{R}_{\text{DM}}(\pi, S) - \mathbb{E}[\hat{R}_{\text{DM}}(\pi, S)]\right| \\ &\leq \left|\mathbb{E}[\hat{R}_{\text{DM}}(\pi_1, S)] - R(\pi_1)\right| + 2\mathfrak{R}_n(\mathcal{F}^{\text{DM}}_{\Pi_1}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} \\ &\leq \mathcal{L}_{\pi_1}(h, \bar{\rho}) + 2\mathfrak{R}_n(\mathcal{F}^{\text{DM}}_{\Pi_1}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.\end{aligned}$$

In the last line, we used the bias characterization from Equation 2. A similar derivation applies to DR using the bias characterization from Equation 3.

Finally, to make the bounds more interpretable, we upper-bound the Rademacher terms by the Rademacher complexity of Π_1. To do so, we use a generalization of Talagrand's contraction lemma [27] for vector-valued function classes, due to Cortes et al. [10]. For all $z \in \mathcal{Z}$ and $\pi, \pi' \in \Pi$, we have

$$|f^{\text{DM}}_{\pi}(z) - f^{\text{DM}}_{\pi'}(z)| = |\langle h(x, \cdot), \pi(\cdot \mid x) - \pi'(\cdot \mid x)\rangle| \leq \sqrt{K} \left\|\pi(\cdot \mid x) - \pi'(\cdot \mid x)\right\|_2 .$$

(Here, we treated $h(x, \cdot)$ and $\pi(\cdot \mid x)$ as vectors. The inequality follows from Cauchy-Schwarz and the assumption that there can be at most K actions, each of which has predicted reward at most 1.) Thus, each $f^{\text{DM}}_{\pi} \in \mathcal{F}^{\text{DM}}_{\Pi_1}$ is a $\sqrt{K}$-Lipschitz (with respect to the 2-norm) function of π. By a similar reasoning,

$$|f^{\text{DR}}_{\pi}(z) - f^{\text{DR}}_{\pi'}(z)| \leq \frac{\sqrt{K}}{\tau} \left\|\pi(\cdot \mid x) - \pi'(\cdot \mid x)\right\|_2 ;$$

meaning, each $f^{\text{DR}}_{\pi} \in \mathcal{F}^{\text{DR}}_{\Pi_1}$ is $(\sqrt{K}/\tau)$-Lipschitz. (Here, we used the fact that the range of f^{DR}_{π} is bounded by $[1-\tau^{-1}, \tau^{-1}]$, and $\left|1 - \tau^{-1}\right| \leq \left|\tau^{-1}\right|$ for $\tau \in (0, 1)$.) Having established Lipschitzness with respect to the 2-norm, we apply [10, Lemma 5] and have

$$\mathfrak{R}_n(\mathcal{F}^{\text{DM}}_{\Pi_1}) \leq \sqrt{K}\mathfrak{R}_n(\Pi_1) \quad \text{and} \quad \mathfrak{R}_n(\mathcal{F}^{\text{DR}}_{\Pi_1}) \leq \frac{\sqrt{K}}{\tau}\mathfrak{R}_n(\Pi_1).$$

A.3 Proof of Proposition 3

We begin with some notation. Recall that $\mathcal{L}_{\pi_1}(h, \bar{\rho})$ is the reward regression risk (Equation 1) under a target policy, π_1. We can think of this quantity as the target risk. We similarly let

$$\mathcal{L}_{\pi_0}(h, \bar{\rho}) \triangleq \mathop{\mathbb{E}}_{x \sim \mathbb{D}_x} \mathop{\mathbb{E}}_{a \sim \pi_0(x)} [|h(x, a) - \bar{\rho}(x, a)|]$$

denote the source risk under the logging policy, π_0. Note that both risks use the *marginal* distribution of contexts, $\mathbb{D}_x$, since the labels are given by the mean reward, $\bar{\rho}$. We will also use the source risk under the joint distribution (i.e., without marginalizing out ρ), which we denote by

$$\mathcal{L}_{\pi_0}(h) \triangleq \mathop{\mathbb{E}}_{(x,\rho) \sim \mathbb{D}} \mathop{\mathbb{E}}_{a \sim \pi_0(x)} [|h(x, a) - \rho(x, a)|].$$

Finally, using $\mathbb{S}$ to denote the empirical distribution of the dataset, S, we let

$$\mathcal{L}_{\mathbb{S}}(h) \triangleq \mathop{\mathbb{E}}_{(x,a,r) \sim \mathbb{S}} [|h(x, a) - r|] = \frac{1}{n} \sum_{i=1}^{n} |h(x_i, a_i) - r_i|$$

denote the empirical source risk with respect to S. Our goal will be to upper-bound $\mathcal{L}_{\pi_1}(h, \bar{\rho})$ by a function of $\mathcal{L}_{\mathbb{S}}(h)$, such that the bound holds for all $h \in \mathcal{H}$ and $\pi_1 \in \Pi_1$ simultaneously.

First, we use the discrepancy (with $\ell(y, y') = |y - y'|$) to convert the target risk to the source risk. With

$$h^\star \in \mathop{\arg\min}_{h \in \mathcal{H}} \mathcal{L}_{\pi_0}(h, \bar{\rho}) + \mathcal{L}_{\pi_1}(h, \bar{\rho}), \tag{7}$$

we have that

$$\begin{aligned}\mathcal{L}_{\pi_1}(h,\bar{\rho}) &\le \mathcal{L}_{\pi_1}(h^\star,\bar{\rho}) + \mathcal{L}_{\pi_1}(h,h^\star) \\ &= \mathcal{L}_{\pi_1}(h^\star,\bar{\rho}) + \mathcal{L}_{\pi_1}(h,h^\star) - \mathcal{L}_{\pi_0}(h,h^\star) + \mathcal{L}_{\pi_0}(h,h^\star) \\ &\le \mathcal{L}_{\pi_1}(h^\star,\bar{\rho}) + \operatorname{disc}(\pi_1,\pi_0) + \mathcal{L}_{\pi_0}(h,h^\star) \\ &\le \mathcal{L}_{\pi_1}(h^\star,\bar{\rho}) + \operatorname{disc}(\pi_1,\pi_0) + \mathcal{L}_{\pi_0}(h^\star,\bar{\rho}) + \mathcal{L}_{\pi_0}(h,\bar{\rho}) \\ &= \operatorname{disc}(\pi_0,\pi_1) + \lambda_{\mathcal{H}}(\pi_0,\pi_1) + \mathcal{L}_{\pi_0}(h,\bar{\rho}). \end{aligned} \tag{8}$$

The first and last inequalities follow from the subadditivity of the loss function. The last line follows from the definitions of $\lambda_{\mathcal{H}}(\pi_0,\pi_1)$ (Equation 6) and $h^\star$ (Equation 7), as well as the symmetry of the discrepancy (Equation 4).

Next, we apply Jensen's inequality (since the loss is a convex function of $\bar{\rho}$, and $\bar{\rho}$ is an expectation) to upper-bound the source risk under $(\mathbb{D}_x \times \pi_0)$ by the source risk under $(\mathbb{D} \times \pi_0)$:

$$\begin{aligned}\mathcal{L}_{\pi_0}(h,\bar{\rho}) &= \mathop{\mathbb{E}}_{x\sim\mathbb{D}_x}\mathop{\mathbb{E}}_{a\sim\pi_0(x)}\left[\left|h(x,a) - \mathop{\mathbb{E}}_{\rho\sim\mathbb{D}_{\rho|x}}[\rho(x,a)]\right|\right] \\ &\le \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}}\mathop{\mathbb{E}}_{a\sim\pi_0(x)}[|h(x,a)-\rho(x,a)|] = \mathcal{L}_{\pi_0}(h). \end{aligned} \tag{9}$$

Combining Equations 8 and 9, we have that

$$\mathcal{L}_{\pi_1}(h,\bar{\rho}) \le \operatorname{disc}(\pi_0,\pi_1) + \lambda_{\mathcal{H}}(\pi_0,\pi_1) + \mathcal{L}_{\pi_0}(h) \tag{10}$$

almost surely (i.e., with probability 1) for all $h \in \mathcal{H}$ and $\pi_1 \in \Pi_1$.

We therefore only need to upper-bound $\mathcal{L}_{\pi_0}(h)$ for all $h \in \mathcal{H}$. To accomplish this, we can use any (one-sided) risk bound for supervised learning, such as [37, Theorem 11.3]. Since the loss function is 1-Lipschitz and (in this case) bounded by $[0,1]$, we thus have, with probability at least $1-\delta$ over $S \sim (\mathbb{D} \times \pi_0)^n$, that all $h \in \mathcal{H}$ satisfy

$$\mathcal{L}_{\pi_0}(h) \le \mathcal{L}_{\mathbb{S}}(h) + 2\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}. \tag{11}$$

Crucially, this bound holds irrespective of π_1. Combining Equations 10 and 11 completes the proof.

A.4 Proof of Proposition 4

We will use the risk definitions given in Appendix A.3, with one modification: for a dataset, S, with empirical distribution $\mathbb{S}$, let

$$\mathcal{L}_{\mathbb{S}}(h,\bar{\rho}) \triangleq \frac{1}{n}\sum_{i=1}^{n}|h(x_i,a_i) - \bar{\rho}(x_i,a_i)|$$

denote the empirical risk using $\bar{\rho}$ as a deterministic labeling function. (Never mind that we cannot actually compute this quantity; it will only be used in the proof.)

Using the triangle inequality,

$$\begin{aligned} &\operatorname{disc}(\pi_0,\pi_1) \le \operatorname{disc}(\pi_0,\mathbb{S}_0) + \operatorname{disc}(\mathbb{S}_0,\mathbb{S}_1) + \operatorname{disc}(\mathbb{S}_1,\pi_1) \\ \text{and}\quad &\operatorname{disc}(\mathbb{S}_0,\mathbb{S}_1) \le \operatorname{disc}(\mathbb{S}_0,\pi_0) + \operatorname{disc}(\pi_0,\pi_1) + \operatorname{disc}(\pi_1,\mathbb{S}_1). \end{aligned}$$

Thus,

$$\begin{aligned}|\operatorname{disc}(\pi_0,\pi_1) - \operatorname{disc}(\mathbb{S}_0,\mathbb{S}_1)| &= \max\{\operatorname{disc}(\pi_0,\pi_1) - \operatorname{disc}(\mathbb{S}_0,\mathbb{S}_1),\ \operatorname{disc}(\mathbb{S}_0,\mathbb{S}_1) - \operatorname{disc}(\pi_0,\pi_1)\} \\ &\le \operatorname{disc}(\pi_0,\mathbb{S}_0) + \operatorname{disc}(\pi_1,\mathbb{S}_1). \end{aligned}$$

We then upper-bound each of the righthand terms using an analysis similar to [34, Corollary 5]. First, we define a class of pairwise differences, $\mathcal{F}_{\text{PD}} \triangleq \{(x,a) \mapsto (h(x,a) - h'(x,a)) : h,h' \in \mathcal{H}\}$, and a class of absolute pairwise differences, $\mathcal{F}_{\text{APD}} \triangleq \{(x,a) \mapsto |h(x,a) - h'(x,a)| : h,h' \in \mathcal{H}\}$. Each member of $\mathcal{F}_{\text{APD}}$ is the absolute value of a member of $\mathcal{F}_{\text{PD}}$. Since the absolute value is 1-Lipschitz, we have via Talagrand's contraction lemma [27] that $\mathfrak{R}_n(\mathcal{F}_{\text{APD}}) \leq \mathfrak{R}_n(\mathcal{F}_{\text{PD}})$. Further,

$$\begin{aligned}\mathfrak{R}_n(\mathcal{F}_{\text{PD}}) &= \mathbb{E}\left[\sup_{h,h' \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(h(x,a) - h'(x,a)\right)\right] \\ &\leq \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x,a)\right] + \mathbb{E}\left[\sup_{h' \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i h'(x,a)\right] = 2\mathfrak{R}_n(\mathcal{H}).\end{aligned}$$

Finally, using an analysis similar to Appendix A.2, we have, with probability $1-\delta$ over draws of both $S_0 \sim (\mathbb{D}_x \times \pi_0)^{n_0}$ and $S_1 \sim (\mathbb{D}_x \times \pi_1)^{n_1}$, that

$$\operatorname{disc}(\pi_0, \mathbb{S}_0) = \sup_{h,h' \in \mathcal{H}} |\mathcal{L}_{\pi_0}(h,h') - \mathcal{L}_{\mathbb{S}_0}(h,h')| \leq 2\mathfrak{R}_{n_0}(\mathcal{F}_{\text{APD}}) + \sqrt{\frac{\ln\frac{4}{\delta}}{2n_0}} \leq 4\mathfrak{R}_{n_0}(\mathcal{H}) + \sqrt{\frac{\ln\frac{4}{\delta}}{2n_0}}$$

and

$$\operatorname{disc}(\pi_1, \mathbb{S}_1) = \sup_{h,h' \in \mathcal{H}} |\mathcal{L}_{\pi_1}(h,h') - \mathcal{L}_{\mathbb{S}_1}(h,h')| \leq 2\mathfrak{R}_{n_1}(\mathcal{F}_{\text{APD}}) + \sqrt{\frac{\ln\frac{4}{\delta}}{2n_1}} \leq 4\mathfrak{R}_{n_1}(\mathcal{H}) + \sqrt{\frac{\ln\frac{4}{\delta}}{2n_1}}.$$

Note that the $\ln(4/\delta)$ comes from bounding each generalization term with probability $\geq 1-\delta/2$. Combining the upper bounds completes the proof.

A.5 Proof of Proposition 5

The proof is similar to that of Proposition 3 (see Appendix A.3), and will reuse the same risk notation. We extend the notation with

$$\mathcal{L}_{\pi_0}(h;\pi_0') \triangleq \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi_0(x)}\left[\frac{\pi_0'(a\,|\,x)}{\pi_0(a\,|\,x)}\,|h(x,a)-\rho(x,a)|\right],$$

and

$$\mathcal{L}_{\mathbb{S}}(h;\pi_0') \triangleq \mathop{\mathbb{E}}_{(x,a,p,r)\sim\mathbb{S}}\left[\frac{\pi_0'(a\,|\,x)}{p}\,|h(x,a)-r|\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_0'(a_i\,|\,x_i)}{p_i}\,|h(x_i,a_i)-r_i|\,.$$

First, we convert $\mathcal{L}_{\pi_1}(h,\bar\rho)$ to $\mathcal{L}_{\pi_0'}(h)$ using the same analysis as Equations 8 and 10:

$$\begin{aligned}\mathcal{L}_{\pi_1}(h,\bar\rho) &\leq \operatorname{disc}(\pi_0',\pi_1) + \lambda_{\mathcal{H}}(\pi_0',\pi_1) + \mathcal{L}_{\pi_0'}(h,\bar\rho) \\ &\leq \operatorname{disc}(\pi_0',\pi_1) + \lambda_{\mathcal{H}}(\pi_0',\pi_1) + \mathcal{L}_{\pi_0'}(h).\end{aligned}$$

Then, noting that $\mathcal{L}_{\pi_0'}(h) = \mathcal{L}_{\pi_0}(h;\pi_0')$ (since all $\pi_0 \in \Pi_0$ have full support), we have

$$\mathcal{L}_{\pi_1}(h,\bar\rho) \leq \operatorname{disc}(\pi_0',\pi_1) + \lambda_{\mathcal{H}}(\pi_0',\pi_1) + \mathcal{L}_{\pi_0}(h;\pi_0'),$$

for any $h \in \mathcal{H}$, $\pi_0' \in \Pi_0$ and $\pi_1 \in \Pi_1$.

The rest of the proof proceeds the same as before, with the exception that the loss function is now $(1/\tau)$-Lipschitz and bounded by $[0, 1/\tau]$, due to the inverse propensity weighting. Thus, with probability at least $1-\delta$ over $S \sim (\mathbb{D}\times\pi_0)^n$, all $h \in \mathcal{H}$ and $\pi_0' \in \Pi_0$ satisfy

$$\mathcal{L}_{\pi_0}(h;\pi_0') \leq \mathcal{L}_{\mathbb{S}}(h;\pi_0') + \frac{2}{\tau}\mathfrak{R}_n(\mathcal{H}) + \frac{1}{\tau}\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}.$$

Since this bound is independent of π_1, it holds for all $h \in \mathcal{H}$, $\pi_0' \in \Pi_0$ and $\pi_1 \in \Pi_1$ simultaneously.