Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning

Philip S. Thomas PHILIPT@CS.CMU.EDU
Emma Brunskill EBRUN@CS.CMU.EDU

Abstract

In this paper we present a new way of predicting the performance of a reinforcement learning policy given historical data that may have been generated by a different policy. The ability to evaluate a policy from historical data is important for applications where the deployment of a bad policy can be dangerous or costly. We show empirically that our algorithm produces estimates that often have orders of magnitude lower mean squared error than existing methods—it makes more efficient use of the available data. Our new estimator is based on two advances: an extension of the doubly robust estimator (Jiang & Li, 2015), and a new way to mix between model based and importance sampling based estimates.

1. Introduction

The ability to predict the performance of a policy without actually having to use it is crucial to the responsible use of reinforcement learning algorithms. Consider the setting where the user of a reinforcement learning algorithm has already deployed some policy, e.g., for determining which advertisement to show a user visiting a website (Theocharous et al., 2015), for determining which medical treatment to suggest for a patient (Thapa et al., 2005), or for suggesting a personalized curriculum for a student (Mandel et al., 2014). In these examples, using a bad policy can be costly or dangerous, so it is important that the user of a reinforcement learning algorithm be able to predict how well a new policy will perform without having to deploy it.

In this paper we propose a new algorithm for tackling this performance prediction problem, which is called the *off-policy policy evaluation* (OPE) problem. The primary objective in OPE problems is to produce estimates that minimize some notion of error. We select mean squared error, a popular notion of error for estimators, as our loss function. This is in line with previous works that all use (root) mean squared error when empirically validating their methods (Precup et al., 2000; Dudík et al., 2011; Mahmood et al., 2014; Thomas, 2015b; Jiang & Li, 2015).

Given this goal, an estimator should be strongly consistent—its mean squared error should converge almost surely to zero as the amount of available data increases.[1] In this paper we introduce a new strongly consistent estimator, MAGIC, that directly optimizes mean squared error. Our empirical results show that MAGIC can produce estimates with orders of magnitude lower mean squared error than the estimates produced by existing algorithms.

Our new algorithm comes from the synthesis of two new ideas. The first is an extension of the recently proposed *doubly robust* (DR) OPE algorithm (Jiang & Li, 2015). We present a novel derivation of the DR algorithm that removes the assumption that the horizon is finite and known. We also give conditions under which the DR estimator is strongly consistent. We then show how we can reduce the variance of the DR estimator by introducing a small amount of bias—an effective trade-off when minimizing the mean squared error of the estimates. We call our extension of the DR estimator the *weighted doubly robust* (WDR) estimator.

Our second major contribution is a new estimator, which we call the *blending IS and model* (BIM) estimator, that combines two different OPE estimators not just by selecting between them, but by blending them together in a way that minimizes the mean squared error. The combination of these two contributions results in a particularly powerful new OPE algorithm that we call the *model and guided importance sampling combined* (MAGIC) estimator, which uses BIM to combine a purely model-based estimator with WDR. In our simulations, MAGIC has the best general performance, often exhibiting orders of magnitude lower mean squared error than prior state-of-the-art estimators.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

The research reported here was supported by a NSF CAREER grant 1350984 and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130215 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

[1] In Appendix A we define strong consistency and present Lemma 3, which elucidates its connection to mean squared error.

2. Notation

We assume that the reader is familiar with reinforcement learning (Sutton & Barto, 1998) and adopt notational standard MDPNv1 for *Markov decision processes* (Thomas, 2015a, MDPs). Although our results carry over to the setting where the states, actions, and rewards are continuous random variables with density functions, for simplicity, our notation assumes that the state, action, and reward sets are finite. Let $H \coloneqq (S_0, A_0, R_0, S_1, \dots)$ be a *trajectory*,and $g(H) \coloneqq \sum_{t=0}^{\infty} \gamma^t R_t$ denote the *return* of a trajectory. We assume that $R_t \in [r_{\min}, r_{\max}]$ for (possibly unknown) finite constants $r_{\min}$ and $r_{\max}$. Let $\gamma \in [0, 1]$ for the finite-horizon setting and $\gamma \in [0, 1)$ for the indefinite and infinite horizon settings so that $g(H)$ is bounded. We use the discounted objective function, $v(\pi) \coloneqq \mathbf{E}[g(H)|H \sim \pi]$, where $H \sim \pi$ denotes that H was generated using the policy π. We use superscripts to denote which trajectory a term comes from, e.g., S_t^H. Let v^π and q^π be the *state value function* and *state-action value function* for policy π—for all $(\pi, s, a) \in \Pi \times \mathcal{S} \times \mathcal{A}$, let $v^\pi(s) \coloneqq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s, \pi\right]$ and $q^\pi(s, a) \coloneqq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s, A_0 = a, \pi\right]$. Notice that v without a superscript denotes the objective function, while v^π denotes a value function, and that the two are related: $v(\pi) = \sum_{s \in \mathcal{S}} \Pr(S_0 = s) v^\pi(s)$.

Let *historical data*, D, be a set of $n \in \mathbb{N}_{>0}$ trajectories and the known policies, called *behavior policies*, that were used to generate them: $D \coloneqq \{(H_i, \pi_i)\}_{i=1}^n$, where $H_i \sim \pi_i$. When we write H_i, we *always* mean that $H_i \sim \pi_i$. Let $\rho_t(H, \pi_e, \pi_b) \coloneqq \prod_{i=0}^{t} \pi_e\left(A_i^H \middle| S_i^H\right) / \pi_b\left(A_i^H \middle| S_i^H\right)$, be an *importance weight*, which is the probability of the first t steps of H under the *evaluation policy*, π_e, divided by its probability under the behavior policy, π_b (Precup et al., 2000, Section 2). We write ρ_t^i and ρ_t as shorthand for $\rho_t(H_i, \pi_e, \pi_i)$ and $\rho_t(H, \pi_e, \pi_b)$. Let $\rho_{-1}^i \coloneqq 1$ for all i. One of the primary challenges will be to combat the high variance and large range of the importance weights, ρ_t.

Let $\hat{r}^\pi(s, a, t) \in [r_{\min}^{\text{model}}, r_{\max}^{\text{model}}]$ denote an approximate model's prediction of R_t if $S_0 = s$, $A_0 = a$, and the policy π is used to generate actions, $A_1, A_2, \dots$, where $r_{\min}^{\text{model}}$ and $r_{\max}^{\text{model}}$ are finite constants. Let $\hat{r}^\pi(s, t) \coloneqq \sum_{a \in \mathcal{A}} \pi(a|s) \hat{r}^\pi(s, a, t)$, be a prediction of R_t if $S_0 = s$ and the policy π is used to generate actions $A_0, A_1, \dots$. Let $\hat{v}^\pi(s) \coloneqq \sum_{t=0}^{\infty} \gamma^t \hat{r}^\pi(s, t)$ and $\hat{q}^\pi(s, a) \coloneqq \sum_{t=0}^{\infty} \gamma^t \hat{r}^\pi(s, a, t)$ be the model's estimates of $v^\pi(s)$ and $q^\pi(s, a)$. We assume that $\hat{r}^\pi(\overset{\infty}{s}, a, t) = 0$ for all $(\pi, a, t) \in \Pi \times \mathcal{A} \times \mathbb{N}_{\geq 0}$, where $\overset{\infty}{s}$ is the terminal absorbing state. Although better models will tend to improve our estimates, we make no assumptions about the veracity of the approximate model's predictions.

3. Off-Policy Policy Evaluation (OPE)

The problem of *off-policy policy evaluation* (OPE) is defined as follows. We are given an evaluation policy, π_e, *historical data*, D, and an approximate model. Our goal is to produce an estimator, $\hat{v}(D)$, of $v(\pi_e)$ that has low *mean squared error* (MSE): $\text{MSE}(\hat{v}(D), v(\pi_e)) \coloneqq \mathbf{E}\left[\left(\hat{v}(D) - v(\pi_e)\right)^2\right]$. We use capital letters to denote random variables, and so the random terms in expected values are always the capitalized letters (e.g. D is a random variable). We assume that the process producing states, actions, and rewards is an MDP with unknown initial state distribution, transition function, and reward function. We assume that the evaluation policy, π_e, the behavior policies, π_i, $i \in \{1, \dots, n\}$, and the discount parameter, γ, are known. For a review of OPE methods, see the works of Precup et al. (2000) or Thomas (2015b, Chapter 3). More recent methods can be found in the works of Jiang & Li (2015) and Mandel et al. (2016).

4. Doubly Robust (DR) Estimator

The *doubly robust* (DR) estimator (Jiang & Li, 2015) is a new unbiased estimator of $v(\pi_e)$ that achieves promising empirical and theoretical results by leveraging an approximate model of an MDP to decrease the variance of the unbiased estimates produced by ordinary importance sampling (Precup et al., 2000). It is doubly robust in that it will provide "good" estimates if either **1)** the model is accurate or **2)** the behavior policies are known. By "good" it is meant that if the former does not hold then the estimator will remain unbiased (although it might have high variance and thus high mean squared error), and if the latter does not hold then if the model has low error the doubly robust estimator will also tend to have low error. Doubly robust estimators were introduced and remain popular in the statistics community (Rotnitzky & Robins, 1995).

The work that introduced the DR estimator for MDPs (Jiang & Li, 2015) derived it as a generalization of a doubly robust estimator for bandits (Dudík et al., 2011). This may be why the DR estimator was derived only for the finite horizon setting where the horizon is known (every trajectory must terminate within $L < \infty$ time steps, and L must be known). It also resulted in a recursive definition of the DR estimator that can be difficult to interpret. In Appendix B we instead derive the DR estimator for MDPs as an application of control variates. Our new derivation holds without assumptions on the horizon and gives the intuitive non-recursive definition, where $w_t^i = \rho_t^i / n$:

$$\mathrm{DR}(D) \coloneqq \sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i R_t^{H_i} \qquad (1)$$
$$-\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t\left(w_t^i\hat{q}^{\pi_e}\left(S_t^{H_i},A_t^{H_i}\right)-w_{t-1}^i\hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right).$$

In Appendix B we show that this definition is equivalent to that of Jiang & Li (2015) when the horizon is finite and known, and we provide several new theoretical results pertaining to the DR estimator. Specifically, we give conditions for DR to be an unbiased estimator without assumptions on the horizon, and we give the first proofs that it is a strongly consistent estimator. Although these are important properties to establish, we relegate them to an appendix due to space limitations.

The non-recursive definition of the DR estimator presented in (1) also reveals the close relationship of the DR estimator to *advantage sum* estimators. Advantage sum estimators were introduced as a way to lower the variance of on-policy Monte Carlo performance estimates for a setting that is a generalization of the (partially observable) MDP setting (Zinkevich et al., 2006; White & Bowling, 2009). The DR estimator for the on-policy setting can be found in the work of Zinkevich et al. (2006, Equation 8). One may therefore view the DR estimator (Jiang & Li, 2015) as the extension of the advantage sum estimator (Zinkevich et al., 2006) to the off-policy setting or as the extension of the doubly robust estimator for bandits (Dudík et al., 2011) to the sequential setting. We are therefore not the first to show that the DR estimator can be viewed as an application of control variates, since White (2009) and Veness et al. (2011, Section 3.1) point out that the advantage sum estimator is an application of control variates. Still, our derivation in Appendix B of the DR estimator is novel.

The DR estimator is not purely model based, since it uses importance weights. However, it is also not a model-free importance sampling method, since it uses an approximate model to decrease the variance of its estimates. We therefore refer to it as a *guided importance sampling* method, since the approximate model is used to guide, but not completely replace, the importance sampling estimates.

5. Weighted Doubly Robust (WDR) Estimator

Empirical and theoretical results show that the DR estimator developed by Jiang & Li (2015) can significantly reduce the variance of ordinary importance sampling without introducing bias. The fact that it does not introduce bias is important when the estimator is used to produce confidence bounds on $v(\pi_e)$ (Thomas, 2015b). However, in practice these confidence bounds often require an impractical amount of data before they are tight enough to be useful, and so approximate confidence bounds (e.g., bootstrap confidence bounds) are used instead (Theocharous et al., 2015). When using these approximate confidence bounds, the strict requirement that an OPE estimator be an unbiased estimator of $v(\pi_e)$ is not necessary. Furthermore, sometimes the goal of OPE is not to produce confidence bounds, but to produce the best possible estimate of $v(\pi_e)$, in order to determine whether π_e should be used instead of the current behavior policy or as an internal mechanism in a policy search algorithm (Levine & Koltun, 2013). In these cases, the "best" estimator is typically defined as the one that has the lowest *mean squared error* (MSE), even if it is not well suited to creating confidence bounds. For example, in their experiments, Precup et al. (2000), Dudík et al. (2011), Mahmood et al. (2014), Thomas (2015b), and Jiang & Li (2015) all use the MSE when evaluating methods.

Although unbiasedness might seem like a desirable property of an estimator, when the goal is to minimize MSE, it often is not. In general, the MSE of an estimator, $\hat{\theta}$, of a statistic, θ, can be decomposed into its variance and its squared bias: $\mathrm{MSE}(\hat{\theta},\theta) = \mathbf{E}[(\theta-\hat{\theta})^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2$, where $\mathrm{Bias}(\hat{\theta}) \coloneqq \mathbf{E}[\hat{\theta}] - \theta$. The optimal estimator in terms of MSE is typically one that balances this bias-variance trade-off, not one with zero bias. Therefore, in the context of minimizing MSE, strong asymptotic consistency, which requires the MSE of an estimator to almost surely converge to zero as the amount of available data increases, is a more desirable property than unbiasedness.

In this section we propose a new OPE estimator that we call the *weighted doubly robust* (WDR) estimator. The WDR estimator comes from applying a simple well-known extension to importance sampling estimators to the DR estimator to produce a new guided importance sampling method. This extension does not directly optimize the bias-variance trade-off, but it does tend to significantly better balance it while maintaining asymptotic consistency. More specifically, WDR is based on *weighted importance sampling* (Powell & Swann, 1966) as opposed to ordinary importance sampling (Hammersley & Handscomb, 1964). For further discussion of the benefits of weighted importance sampling over ordinary importance sampling, see the work of Thomas (2015b, Section 3.8). Weighted importance sampling has been used before for OPE (Precup et al., 2000), but not with the DR estimator.

We define the WDR estimator as the DR estimator in (1), except where $w_t^i \coloneqq \rho_t^i / \sum_{j=1}^{n}\rho_t^j$.[2] Intuitively it is

[2]Just as DR-v2 extends the DR estimator (Jiang & Li, 2015, Section 4.4), one can create the WDR-v2 estimator by replacing $\hat{q}^{\pi_e}(S_t,A_t)$ with $\hat{r}^{\pi_e}(S_t,A_t,0)+\gamma\hat{v}^{\pi_e}(S_{t+1})$ in (1). For the domains presented here, these variants did not outperform the original DR and WDR estimators.

clear that this estimator is asymptotically correct because $\mathbf{E}[\rho_t^j] = 1$, and so by the law of large numbers the denominator of w_t^i will converge to n. Although WDR is not an unbiased estimator of $v(\pi_e)$, its bias follows a pattern that is both predictable and also sometimes desirable. When there is only a single trajectory, i.e., $n = 1$, $\mathrm{WDR}(D)$ is an unbiased estimator of the performance of the behavior policy, since $w_t^1 = 1$ for all t. If there is a single behavior policy, π_b, as the number of trajectories increases, the expected value of $\mathrm{WDR}(D)$ shifts from $v(\pi_b)$ towards $v(\pi_e)$.

In Appendix C we establish two different sets of assumptions that are sufficient to show that WDR is a strongly consistent estimator of $v(\pi_e)$.

6. Empirical Studies (WDR)

In order to both show the empirical benefits of WDR over existing importance sampling estimators and better motivate our second major contribution, we present an empirical comparison of different OPE methods.[3] We compare to a broad sampling of model-free importance sampling estimators, definitions of which can be found in the work of Thomas (2015b, Chapter 3): *importance sampling* (IS), *per-decision importance sampling* (PDIS), *weighted importance sampling* (WIS), and *consistent weighted per-decision importance sampling* (CWPDIS). We also compare to the guided importance sampling *doubly robust* (DR) estimator (Jiang & Li, 2015).

Lastly, we compare to the *approximate model* (AM) estimator, which uses all of the available data to construct an approximate model of the MDP.[4] The performance of the evaluation policy on the approximate model is typically easy to compute and can be used as an estimate of $v(\pi_e)$. For example, in our experiments the approximate model maintains an estimate, $\widehat{d}_0$, of the initial state distribution, and so we define $\mathrm{AM} := \sum_{s \in \mathcal{S}} \widehat{d}_0(s) \hat{v}^{\pi_e}(s)$. Notice that unlike the importance sampling based methods, AM does not include any importance weights (ρ_t terms).

Here we provide an overview the results detailed in Appendix D. We used three domains: **1)** a 4×4 gridworld previously constructed specifically for evaluating OPE methods (Thomas, 2015b, Section 2.5); **2)** *ModelFail*, a partially observable, deterministic, 4-state domain with horizon $L = 2$ and in which 3 of the states are aliased (appear identical to the agent), which means that the agent's observations are not Markovian and thus that the approximate (MDP) model is incorrect, even asymptotically; and **3)** *ModelWin*, a stochastic 4-state MDP with $L = 20$, where the model that we use can perfectly represent the true MDP.

In our simulations, WDR dominated the other importance sampling and guided importance sampling estimators (but not AM). Not only did WDR always achieve the lowest mean squared error of these estimators, but no other single (guided) importance sampling estimator was able to always achieve mean squared errors within an order of magnitude of WDR's (e.g., Figure 1a). Note that, as expected, WDR significantly outperforms AM on the ModelFail domain. However, AM significantly outperforms WDR on the ModelWin domain, which was designed so that the model quickly converges to the true MDP.

One might wonder why DR and WDR can do worse than AM even though they incorporate the approximate model. Although this question has been discussed before by Jiang & Li (2015, Section 4.2), we review it here. Notice that we can write the DR and WDR estimators as:

$$\mathrm{WDR}(D) := \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{v}^{\pi_e}(S_0^{H_i})}_{\text{(a)}} \tag{2}$$
$$+ \sum_{i=1}^{n}\sum_{t=0}^{\infty} \gamma^t w_t^i \left[\underbrace{R_t^{H_i} - \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) + \gamma \hat{v}^{\pi_e}\left(S_{t+1}^{H_i}\right)}_{\text{(b)}} \right].$$

If the approximate model is perfect, then **(a)** is both a low variance and unbiased estimator of $v(\pi_e)$. If the approximate model is perfect and R_t and S_{t+1} are deterministic functions of S_t and A_t, then **(b)** is zero, and so the second term is always zero and WDR is an excellent estimator. However, if R_t or S_{t+1} is *not* a deterministic function of S_t and A_t—if the state transitions or rewards are stochastic—then **(b)** is not necessarily zero. If the importance weights, w_t^i, have high variance, then even slightly non-zero values of **(b)** can result in high mean squared error.

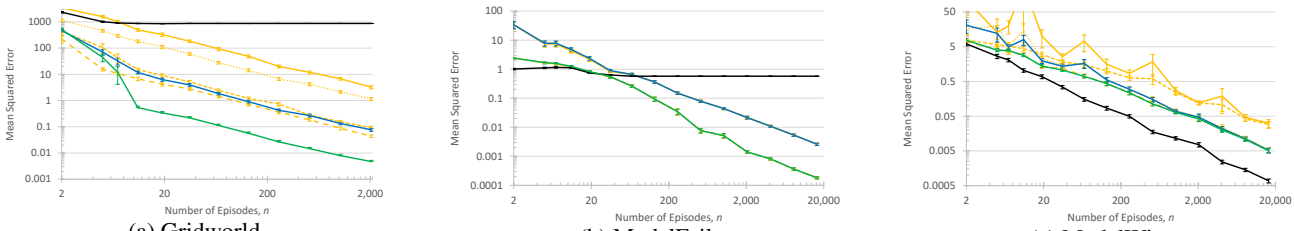
In summary, while WDR tends to outperform the other importance sampling estimators, sometimes AM can produce estimates with much lower MSE. This trend is also visible in the results of Jiang & Li (2015), where AM performs better than DR. Ideally we would like an estimator that combines WDR and AM or switches between them to always achieve the performance of the better estimator. In the following sections we show how this can be done.

7. Blending IS and Model (BIM) Estimator

In this section we show how two OPE estimators can be merged into a single estimator that exhibits the desirable properties of both. Before doing so, we establish some ter-

[3] The raw data for all experiments in this paper is provided in the supplemental spreadsheet.

[4] This model-based estimator has been called the *direct method* in previous work (Dudík et al., 2011), however, in other previous work *direct methods* are model-free while *indirect methods* are model-based (Sutton & Barto, 1998, Section 9.2).



(a) Gridworld (b) ModelFail (c) ModelWin

Figure 1: Empirical results for three different experimental setups. All plots in this paper have the same format: they show the mean squared error of different estimators as n, the number of episodes in D, increases. Both axes always use a logarithmic scale and standard error bars are included from 128 trials. All plots use the following legend:

IS PDIS WIS CWPDIS DR AM WDR

minology. We divide OPE estimators into three classes. The first class we call *importance sampling estimators*. We define this class to include all estimators that, when L is finite, are defined using all of the importance weights $\rho_0, \rho_1, \ldots, \rho_{L-1}$. Notice that this includes IS, PDIS, WIS, CWPDIS, DR, and WDR. The second class we call *purely model-based estimators*. We define this class to include all estimators that do not contain any ρ_t terms for $t \geq 0$. The only purely model-based estimator in this paper is AM. Finally, we call the third class *partial importance sampling estimators*. These estimators are those that do not fall into either of the other two classes—estimators that use importance weights, ρ_t, but only for $t < L-1$. We will introduce one such estimator later in this section.

We contend that importance sampling estimators and purely model-based estimators are two extremes on a spectrum of estimators. Importance sampling estimators tend to be strongly consistent. That is, as more historical data becomes available, their estimates become increasingly accurate. However, their use of importance weights means that they *all* (including DR and WDR) also can have high variance relative to purely model-based estimators. This is evident in the results on the ModelWin domain.

On the other end of the spectrum, purely model-based estimators like AM are often *not* strongly consistent. If the approximate model uses function approximation or if there is some partial observability, then the approximate model may not converge to the true MDP. So, as more historical data becomes available, the estimates of AM may converge to a value other than $v(\pi_e)$. Thus, purely model-based estimators tend to have high bias, even asymptotically, as evidenced by the AM curve in Figure 1b. However, purely model-based methods also tend to have low variance because they do not contain any ρ_t terms.

Between these two extremes lies a range of partial importance sampling estimators. Estimators that are close to the purely model-based estimators use ρ_t terms only for small t, while estimators that are close to importance sampling estimators use ρ_t terms with large t approaching $L-1$. Before formally defining one such partial importance sampling estimator, we present a few additional definitions. First, let $\text{IS}^{[0:j]}(D)$ denote an estimate of $\mathbf{E}[\sum_{t=0}^{j} \gamma^t R_t | H \sim \pi_e]$, constructed from D using an importance sampling method like PDIS or WDR, which uses importance weights up to and including ρ_j. Similarly, let $\text{AM}^{[j:\infty]}(D)$ denote a primarily model-based prediction from D of $\mathbf{E}[\sum_{t=j}^{\infty} \gamma^t R_t | H \sim \pi_e]$ that may not use ρ_t terms with $t \geq j$.

We can now define a partial importance sampling estimator that we call the *off-policy j-step return*, $g^{(j)}(D)$, which uses an importance sampling based method to predict the outcome of using π_e up until R_j is generated, and the approximate model estimator to predict the outcomes thereafter. That is, let j denote the *length* of the j-step return and for all $j \in \mathbb{N}_{\geq -1}$, let[5]

$$g^{(j)}(D) \coloneqq \text{IS}^{[0:j]}(D) + \text{AM}^{[j+1:\infty]}(D)$$
$$g^{(\infty)}(D) \coloneqq \lim_{j \to \infty} g^{(j)}(D). \tag{3}$$

Notice that $g^{(-1)}(D)$ is a purely model-based estimator, $g^{(\infty)}(D)$ is an importance sampling estimator, and the other off-policy j-step returns are partial importance sampling estimators that blend between these two extremes. When j is small, the off-policy j-step return is similar to AM, using importance sampling to predict only a few early rewards. When j is large, it uses importance sampling to predict most of the rewards and the model only for rewards at the end of a trajectory. So, as j increases we expect the variance of the return to increase, but the bias to decrease.

We propose a new estimator, which we call the *blending IS and model* (BIM) estimator, that leverages this spectrum of estimators to blend together the IS and AM estimators in a way that minimizes MSE. It does this by computing a weighted average of the different length returns: $\text{BIM}(D) \coloneqq \mathbf{x}^\intercal \mathbf{g}(D)$, where $\mathbf{x} \coloneqq (x_{-1}, x_0, x_1, \ldots)^\intercal$ is an infinite-dimensional weight vector and $\mathbf{g}(D)$ is an infinite-

[5] If prior knowledge about d_0 is available, then one might consider adding $g^{(-2)}(D)$ to denote the model's prediction of $v(\pi_e)$, which might differ from $g^{(-1)}(D)$.

dimensional vector of different length returns, $\mathbf{g}(D) := (g^{(-1)}(D), g^{(0)}(D), \dots,)^\intercal$. The remaining question is then: how should we select the weights, $\mathbf{x}$?

A similar question has been studied before in reinforcement learning research when deciding how to weight j-step returns (not off-policy), as reviewed by Sutton & Barto (1998, Section 7.2). The most common solution, a complex return called the λ-*return*, uses $x_{-1} = 0$ and $x_j = (1-\lambda)\lambda^j$ for all other j. The λ-return is the foundation of the entire TD(λ) family of algorithms, which includes the original linear-time algorithm (Sutton, 1988), least-squares formulations (Bradtke & Barto, 1996; Mahmood et al., 2014), methods for adapting λ (Downey & Sanner, 2010), true-online methods (van Hasselt et al., 2014), and the recent emphatic methods (Mahmood et al., 2015).

Recent work has suggested that the λ-return could be replaced by more statistically principled complex returns like the γ-return (Konidaris et al., 2011) or Ω-return (Thomas et al., 2015). For the finite-horizon setting and for $j \in \{0, \dots, L-1\}$ the γ-return uses $x_j := (\sum_{i=0}^{j} \gamma^{2i})^{-1} / \sum_{\hat{j}=0}^{L-1} (\sum_{i=0}^{\hat{j}} \gamma^{2i})^{-1}$, and the Ω-return uses $x_j = \sum_{i=0}^{L-1} \Omega_n^{-1}(j,i) / \sum_{\hat{j},i=0}^{L-1} \Omega_n^{-1}(\hat{j},i)$, where Ω_n is the $L \times L$ covariance matrix where $\Omega_n(i,j) = \mathrm{Cov}(g^{(i)}(D), g^{(j)}(D))$, and where both the γ and Ω-returns use $x_j = 0$ for $j \notin \{0, \dots, L-1\}$.

The advantage of the γ-return over the λ-return is that it uses a more accurate model of how variance increases with the length of a return, which also eliminates the λ hyperparameter used by the λ-return. The advantages of the Ω-return over the γ-return are that it both uses a yet more-accurate estimate of how variance grows with the length of the return, which is computed from historical data, and that it better accounts for the fact that different length returns are *not* independent, i.e., $g^{(i)}(D)$ and $g^{(j)}(D)$ are not independent even if $i \neq j$.

However, none of these weighting schemes are sufficient for our needs because they do not cause BIM to necessarily be a strongly consistent estimator.[6] This is likely because they were all designed for the setting where only one trajectory is available, i.e., $n = 1$, while strong consistency is a property that deals with performance as $n \to \infty$. Furthermore, they were designed for *on-policy* policy evaluation.

We therefore propose a new weighting scheme (a new complex return for multiple trajectories) that directly optimizes our primary objective: the mean squared error. This new weighting scheme is $\mathbf{x}^\star := \arg\min_{\mathbf{x} \in \mathbb{R}^\infty} \mathrm{MSE}(\mathbf{x}^\intercal \mathbf{g}(D), v(\pi_e))$. Unfortunately, we typically cannot compute $\mathbf{x}^\star$, because we do not know $\mathrm{MSE}(\mathbf{x}^\intercal \mathbf{g}(D), v(\pi_e))$ for any $\mathbf{x}$. Instead, we propose estimating $\mathbf{x}^\star$ by minimizing an approximation of $\mathrm{MSE}(\mathbf{x}^\intercal \mathbf{g}(D), v(\pi_e))$. First, dealing with an infinite number of different return lengths is challenging. To avoid this, we propose only using a subset of the returns, $\{\mathbf{g}^{(j)}(D)\}$, for $j \in \mathcal{J}$, where $|\mathcal{J}| < \infty$. For all $j \notin \mathcal{J}$, we assign $\mathbf{x}_j = 0$. We suggest including -1 and ∞ in $\mathcal{J}$.

To simplify later notation, let $\mathbf{g}_\mathcal{J}(D) \in \mathbb{R}^{|\mathcal{J}|}$ be the elements of $\mathbf{g}(D)$ whose indexes are in $\mathcal{J}$—the returns that will not necessarily be given weights of zero. Also let $\mathcal{J}_j$ denote the j^{th} element in $\mathcal{J}$. We can then estimate $\mathbf{x}^\star$ by:

$$\widehat{\mathbf{x}}^\star \in \arg\min_{\mathbf{x} \in \mathbb{R}^{|\mathcal{J}|}} \mathrm{MSE}(\mathbf{x}^\intercal \mathbf{g}_\mathcal{J}(D), v(\pi_e)),$$

where our estimate of $x^\star_j$ is zero if $j \notin \mathcal{J}$ and our estimate of $x^\star_{\mathcal{J}_j}$ is $\widehat{x}^\star_j$ for $j \in \{1, \dots, |\mathcal{J}|\}$.

Next, to avoid searching all of $\mathbb{R}^{|\mathcal{J}|}$ and also to serve as a form of regularization on $\widehat{\mathbf{x}}^\star$, we limit the set of $\mathbf{x}$ that we consider to the $|\mathcal{J}|$-simplex, i.e., we require $x_j \geq 0$ for all $j \in \{1, \dots, |\mathcal{J}|\}$ and $\sum_{j=1}^{|\mathcal{J}|} x_j = 1$. We write $\Delta^{|\mathcal{J}|}$ to denote this set of weight vectors—the $|\mathcal{J}|$-simplex.

Using the bias-variance decomposition of MSE, we have:

$$\begin{aligned}\widehat{\mathbf{x}}^\star &\in \arg\min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathrm{Bias}(\mathbf{x}^\intercal \mathbf{g}_\mathcal{J}(D))^2 + \mathrm{Var}(\mathbf{x}^\intercal \mathbf{g}_\mathcal{J}(D)) \\ &= \arg\min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathbf{x}^\intercal [\Omega_n + \mathbf{b}_n \mathbf{b}_n^\intercal] \mathbf{x},\end{aligned}$$

where n remains the number of trajectories in D, Ω_n is the $|\mathcal{J}| \times |\mathcal{J}|$ covariance matrix where $\Omega_n(i,j) = \mathrm{Cov}(\mathbf{g}^{(\mathcal{J}_i)}(D), \mathbf{g}^{(\mathcal{J}_j)}(D))$ and $\mathbf{b}_n$ is the $|\mathcal{J}|$-dimensional vector with $\mathbf{b}_n(j) = \mathbf{E}[\mathbf{g}^{(\mathcal{J}_j)}(D)] - v(\pi_e)$ for all $j \in \{1, \dots, |\mathcal{J}|\}$.[7] This simplifies the problem of estimating the MSE for all possible $\mathbf{x}$ into estimating two terms: the bias vector, $\mathbf{b}_n$, and the covariance matrix, Ω_n.

Let $\widehat{\mathbf{b}}_n$ and $\widehat{\Omega}_n$ be the estimates of $\mathbf{b}_n$ and Ω_n when there are n trajectories in D. The exact scheme used to estimate $\mathbf{b}_n$ and Ω_n depends on the definitions of $\mathrm{IS}^{[0:j]}(D)$ and $\mathrm{AM}^{[j:\infty]}(D)$. In general, both terms are easier to estimate for unweighted importance sampling estimators like PDIS and DR than for weighted estimators like CWPDIS or WDR.

To make the dependence of BIM on the estimates of Ω_n and $\mathbf{b}_n$ explicit, and to summarize the approximations we have made, we redefine the BIM estimator as:

$$\mathrm{BIM}(D, \widehat{\Omega}_n, \widehat{\mathbf{b}}_n) := (\widehat{\mathbf{x}}^\star)^\intercal \mathbf{g}_\mathcal{J}(D),$$

where $\widehat{\mathbf{x}}^\star \in \arg\min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathbf{x}^\intercal [\widehat{\Omega}_n + \widehat{\mathbf{b}}_n \widehat{\mathbf{b}}_n^\intercal] \mathbf{x}$.

[6] The λ-return with $\lambda = 1$ is defined to be $g^{(\infty)}(D)$ and is consistent, but it does not mix the two OPE methods at all.

[7] Since $\mathbf{b}_n$ (similarly, Ω_n) already has a subscript, we write $\mathbf{b}_n(j)$ to denote the j^{th} element of $\mathbf{b}_n$.

Before continuing, we establish an assumption that will be useful here and later: that the importance weights, ρ_t^i, are bounded above by a finite constant, $\beta \in \mathbb{R}$ (they are always bounded below by zero). This assumption is trivially satisfied in the common setting where the horizon is finite and the state and action sets are finite. Although Assumption 1 requires β to exist, none of our results depend on how large β is. So, in the non-finite state, action, and horizon settings one may ensure that evaluation policies are only considered if they satisfy Assumption 1 for some arbitrarily large β.

Assumption 1 (Bounded importance weight)**.** *There exists a constant $\beta < \infty$ such that for all $(t, i) \in \mathbb{N}_{\geq 0} \times \{1, \ldots, n\}$, $\rho_t^i \leq \beta$ surely.*

We now show that if at least one of the returns included in $\mathcal{J}$ is a strongly consistent estimator of $v(\pi_e)$, Assumption 1 holds, $\beta \in \mathbb{R}$, and if the estimates of $\mathbf{b}_n$ and Ω_n are themselves strongly consistent, then BIM is a strongly consistent estimator of $v(\pi_e)$.

Theorem 1. *If Assumption 1 holds, there exists at least one $j \in \mathcal{J}$ such that $g^{(j)}(D)$ is a strongly consistent estimator of $v(\pi_e)$, $\widehat{\mathbf{b}}_n - \mathbf{b}_n \xrightarrow{a.s.} 0$, and $\widehat{\Omega}_n - \Omega_n \xrightarrow{a.s.} 0$, then* $\mathrm{BIM}(D, \widehat{\Omega}_n, \widehat{\mathbf{b}}_n) \xrightarrow{a.s.} v(\pi_e)$. **Proof** *See Appendix E.*

8. Model and Guided Importance Sampling Combined (MAGIC) Estimator

In this section we propose using the BIM estimator with WDR as the importance sampling estimator, and show how $\mathbf{b}_n$ and Ω_n can be approximated in this setting. The resulting estimator combines purely model based estimates with the estimates of the guided importance sampling algorithm WDR, and so we call it the **m***odel* **a***nd* **g***uided* **i***mportance sampling* **c***ombining* (MAGIC) estimator.

Although the derivation of how to properly define $\mathrm{IS}^{[0:j]}(D)$ and $\mathrm{AM}^{[j:\infty]}(D)$ in order to blend WDR with the approximate model is less obvious than one might expect and therefore an important technical detail, we relegate it to Appendix F due to space restrictions. The resulting definition of an off-policy j-step return is

$$g^{(j)}(D) \coloneqq \sum_{i=1}^{n} g_i^{(j)}(D), \tag{4}$$

where

$$g_i^{(j)}(D) \coloneqq \underbrace{\sum_{t=0}^{j} \gamma^t w_t^i R_t^{H_i}}_{\text{(a)}} + \underbrace{\gamma^{j+1} w_j^i \hat{v}^{\pi_e}(S_{j+1}^{H_i})}_{\text{(b)}} - \underbrace{\sum_{t=0}^{j} \gamma^t \left(w_t^i \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) - w_{t-1}^i \hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right)}_{\text{(c)}}.$$

where **(c)** is the combined control variate for both the importance sampling based term, **(a)**, and the model-based term, **(b)**, and where we use WDR's definition of w_t^i. Another viable definition of $g^{(j)}(D)$ is given in Appendix F.1.

Consider the entries in Ω_n:

$$\mathrm{Cov}\left(g^{(j)}(D), g^{(k)}(D)\right) = \mathrm{Cov}\left(\sum_{i=1}^{n} g_i^{(j)}(D), \sum_{i=1}^{n} g_i^{(k)}(D)\right).$$

Notice that $g_i^{(j)}(D)$ really is a function of all of D, not just H_i, since $w_t^i = \rho_t^i / \sum_{j=1}^{n} \rho_t^j$. This means that, although the terms in the sum, $\sum_{i=1}^{n} g_i^{(j)}(D)$, are identically distributed, they are not independent, due to their shared reliance on D. However, the $g_i^{(j)}(D)$ terms become less dependent as $n \to \infty$ because the only dependence of $g_i^{(j)}(D)$ on trajectories other than H_i comes from the denominator of w_t^i, which converges almost surely to n.

We therefore propose approximating Ω_n using the sample covariance matrix that results from the assumption that $g_i^{(j)}(D)$ and $g_i^{(k)}(D)$ are independent for $j \neq k$. That is, let $\bar{g}_i^{(\mathcal{J}_j)}(D) \coloneqq \frac{1}{n} \sum_{i=1}^{n} g_i^{(\mathcal{J}_j)}(D)$ and

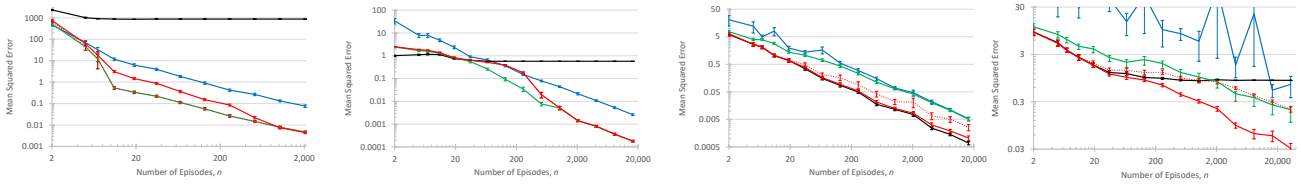
$$\widehat{\Omega}_n(j,k) \coloneqq \frac{n}{n-1} \sum_{i=1}^{n} \left(g_i^{(\mathcal{J}_j)}(D) - \bar{g}_i^{(\mathcal{J}_j)}(D)\right) \times \left(g_i^{(\mathcal{J}_k)}(D) - \bar{g}_i^{(\mathcal{J}_k)}(D)\right). \tag{5}$$

Estimating the bias vector, $\mathbf{b}_n$, is challenging because it has a strong dependence on the value that we wish we knew, $v(\pi_e)$. We cannot use AM's estimate as a stand-in for $v(\pi_e)$ because it would cause us to assume that AM's greatest weakness—its high bias—is negligible. We cannot use WDR's estimate (or any other importance sampling estimator's estimate) because our estimate of $\mathbf{b}_n$ would then conflate the high variance of importance sampling estimates with the bias that we wish to estimate.

When n, the number of trajectories in D, is small, variance tends to be the root cause of high MSE. We therefore propose using an estimate of $\mathbf{b}_n$ that is initially conservative—initially it underestimates the bias—but which becomes correct as n increases. Let $\mathrm{CI}(g^{(\infty)}(D), \delta)$ be a $1-\delta$ confidence interval on the expected value of the random variable $g^{(\infty)}(D) = \mathrm{WDR}(D)$. Intuitively, as n increases we expect that this confidence interval will converge to $g^{(\infty)}(D)$, which in turn converges to $v(\pi_e)$. So, we estimate $\mathbf{b}_n(j)$, the bias of the off-policy j-step return, by its distance from the 10% confidence interval. That is, we estimate $\mathbf{b}_n(j)$ as

$$\widehat{\mathbf{b}}_n(j) \coloneqq \mathrm{dist}\left(g^{(\mathcal{J}_j)}(D), \mathrm{CI}(g^{(\infty)}(D), 0.1)\right),$$

where $\mathrm{dist}(y, \mathcal{Z})$ is the distance between $y \in \mathbb{R}$ and the set $\mathcal{Z} \subseteq \mathbb{R}$, i.e., $\mathrm{dist}(y, \mathcal{Z}) \coloneqq \min_{z \in \mathcal{Z}} |y - z|$. We use both



(a) Gridworld (b) ModelFail (c) ModelWin (d) Hybrid

Figure 2: Empirical comparison of MAGIC to other estimators using the legend from Figure 1. All plots use the following legend (although only Figure 2d includes MAGIC-B):

DR AM WDR MAGIC MAGIC-B

the percentile bootstrap method (Efron & Tibshirani, 1993) and Chernoff-Hoeffding's inequality to construct the confidence interval, and use whichever is tighter. Although in practice the Chernoff-Hoeffding interval is almost always the looser of the two, and so it need not actually be computed, its inclusion simplifies our proofs.

High-level pseudocode suitable for understanding MAGIC is provided in Algorithm 1, while detailed pseudocode suitable for implementations is provided in Appendix G. Recall that $\mathcal{J}$ should include -1 and ∞.

Algorithm 1 MAGIC(D)

1: **Input:** Historical data, $\mathcal{D}$, evaluation policy, π_e, an approximate model, and a set of return-lengths, $\mathcal{J}$.
2: Compute $|\mathcal{J}| \times |\mathcal{J}|$ matrix $\widehat{\Omega}_n$ according to (5).
3: Compute a 90% confidence interval, $[l, u]$, on WDR(D) using the percentile bootstrap method.
4: Compute $|\mathcal{J}| \times 1$ vector $\widehat{\mathbf{b}}_n$, where $\widehat{\mathbf{b}}_n(j) = \text{dist}(g^{(\mathcal{J}_j)}(D), [l, u])$.
5: $\mathbf{x} \leftarrow \arg\min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathbf{x}^\intercal [\widehat{\Omega}_n + \widehat{\mathbf{b}}_n \widehat{\mathbf{b}}_n^\intercal] \mathbf{x}$
6: **return** $\mathbf{x}^\intercal \mathbf{g}_{\mathcal{J}}(D)$

In Theorem 2 we establish conditions under which the MAGIC estimator is a strongly consistent estimator of $v(\pi_e)$. When these conditions are not satisfied, it does *not* mean that the result does not hold or that the MAGIC estimator will perform poorly—it merely means that the theoretical results are not guaranteed by our proofs. Theorem 2 uses a new assumption, Assumption 2, which ensures that all trajectories of interest when evaluating π_e will be produced by all of the behavior policies. This is a standard assumption in OPE and typically precludes the use of deterministic behavior policies.[8]

Assumption 2 (Absolute continuity)**.** *For all* $(s, a, i) \in \mathcal{S} \times \mathcal{A} \times \{1, \ldots, n\}$*, if* $\pi_i(a|s) = 0$ *then* $\pi_e(a|s) = 0$.

Theorem 2 (MAGIC - strongly consistent)**.** *If Assumptions 1 and 2 hold and* $\infty \in \mathcal{J}$*, then* MAGIC$(D) \xrightarrow{a.s.} v(\pi_e)$.
Proof *See Appendix H.*

[8] Assumption 2 could be replaced with a less-restrictive assumption like that used by Thomas (2015b, Section 3.5). We use Assumption 2 because it allows for simplified proofs.

9. Empirical Studies (MAGIC)

Appendix I provides detailed experiments using MAGIC. In this section we provide an overview of these results. The first three plots in Figure 2 correspond to those in Figure 1, but include MAGIC. In general MAGIC does very well, tracking or exceeding the best performance of WDR and AM. However, in Figure 2c MAGIC does not perfectly track AM. The scale is logarithmic, so the difference between MAGIC and AM is small in comparison to the benefit of MAGIC over WDR. We hypothesize that the reason MAGIC does not match AM may be due to error in our estimates of Ω_n and $\mathbf{b}_n$.

Figure 2d is for *Hybrid*, a domain that consists of concatenating ModelFail with ModelWin. This means that early in the trajectories there is partial observability, but later the state is fully observable. This might occur in education domains (initial uncertainty over a student's knowledge) or robotics (positional uncertainty before localizing). In such a setting, MAGIC outperforms all other estimators, even AM and WDR, by automatically leveraging WDR for the parts of trajectories where partial observability causes the model to be inaccurate, and AM for the parts of trajectories where the model is accurate. To emphasize this, we include MAGIC-B (B for *binary*) where $\mathcal{J} = \{-1, \infty\}$, so that BIM can only blend AM and WDR by placing weights on them. The poor performance of MAGIC-B in Figures 2c and 2d supports our use of off-policy j-step returns.

10. Conclusion

We have proposed several new OPE estimators and showed empirically that they outperform existing estimators. While previous OPE estimators that use importance sampling often failed to outperform the approximate model estimator (which does not use importance sampling), our new estimators often do, frequently by orders of magnitude. In cases where the approximate model estimator remains the best estimator, one of our new estimators, MAGIC, performs similarly. In other cases, MAGIC meets or exceeds the performance of state-of-the-art prior estimators. We present some potential avenues of future work in Appendix J.

Acknowledgements

This paper benefited significantly from the feedback and corrections of several people whom we would like to acknowledge and thank: Nan Jiang, Lihong Li, Michael Bowling, Rich Sutton, Scott Niekum, Zhaohan Daniel Guo, Christoph Dann, Shayan Doroudi, and the reviewers.

References

Bartle, Robert G. *The elements of integration and Lebesgue measure*. John Wiley & Sons, 2014.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor-critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Bradtke, S.J. and Barto, A.G. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1–3):33–57, March 1996.

Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, 1997.

Downey, C. and Sanner, S. Temporal difference Bayesian model averaging: A Bayesian perspective on adapting lambda. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 311–318, 2010.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning*, pp. 1097–1104, 2011.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall, London, 1993.

Hammersley, J. M. and Handscomb, D. C. Monte carlo methods, methuen & co. *Ltd., London*, pp. 40, 1964.

Jiang, N. and Li, L. Doubly robust off-policy evaluation for reinforcement learning. *ArXiv*, arXiv:1511.03722v1, 2015.

Konidaris, G. D., Niekum, S., and Thomas, P. S. TD_γ: Re-evaluating complex backups in temporal difference learning. In *Advances in Neural Information Processing Systems 24*, pp. 2402–2410, 2011.

Levine, S. and Koltun, V. Guided policy search. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 1–9, 2013.

Mahmood, A. R., Hasselt, H., and Sutton, R. S. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems 27*, 2014.

Mahmood, A. R., Yu, H., White, M., and Sutton, R. S. Emphatic temporal-difference learning. *ArXiv*, arXiv:1507.01569, 2015.

Mandel, T., Liu, Y., Levine, S., Brunskill, E., and Popović, Z. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, 2014.

Mandel, T., Liu, Y., Brunskill, E., and Popović, Z. Offline evaluation of online reinforcement learning algorithms. In *Proceedings of the Thirtieth Conference on Artificial Intelligence*, 2016.

Mittelhammer, R. C. *Mathematical statistics for economics and business*, volume 78. Springer, 1996.

Powell, M. J. D. and Swann, J. Weighted uniform sampling: a Monte Carlo technique for reducing variance. *Journal of the Institute of Mathematics and its Applications*, 2(3):228–236, 1966.

Precup, D., Sutton, R. S., and Singh, S. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766, 2000.

Rotnitzky, A. and Robins, J. M. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820, 1995.

Sen, P. K. and Singer, J. M. *Large Sample Methods in Statistics An Introduction With Applications*. Chapman & Hall, 1993.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Sutton, R.S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Thapa, D., Jung, I., and Wang, G. Agent based decision support system using reinforcement learning under emergency circumstances. *Advances in Natural Computation*, 3610:888–892, 2005.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2015.

Thomas, P. S. A notation for Markov decision processes. *ArXiv*, arXiv:1512.09075v1, 2015a.

Thomas, P. S. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015b.

Thomas, P. S., Niekum, S., Theocharous, G., and Konidaris, G. D. Policy evaluation using the Ω-return. In *Advances in Neural Information Processing Systems*, 2015.

van Hasselt, H., Mahmood, A. R., and Sutton, R. S. Off-policy TD(λ) with true online equivalence. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

Veness, J., Lanctot, M., and Bowling, M. Variance reduction in monte-carlo tree search. In *Advances in Neural Information Processing Systems*, pp. 1836–1844, 2011.

White, M. A general framework for reducing variance in agent evaluation. Master's thesis, University of Alberta, 2009.

White, M. and Bowling, M. Learning a value analysis tool for agent evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1976–1981, 2009.

Zinkevich, M., Bowling, M., Bard, N., Kan, M., and Billings, D. Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI)*, pp. 573–578, 2006.