Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

High Confidence Off-Policy Evaluation

Philip S. Thomas[1,2] **Georgios Theocharous**[1] **Mohammad Ghavamzadeh**[1,3]
[1]Adobe Research, [2]University of Massachusetts Amherst, [3]INRIA Lille
{phithoma,theochar,ghavamza}@adobe.com

Abstract

Many reinforcement learning algorithms use trajectories collected from the execution of one or more policies to propose a new policy. Because execution of a bad policy can be costly or dangerous, techniques for evaluating the performance of the new policy without requiring its execution have been of recent interest in industry. Such off-policy evaluation methods, which estimate the performance of a policy using trajectories collected from the execution of other policies, heretofore have not provided confidences regarding the accuracy of their estimates. In this paper we propose an off-policy method for computing a lower confidence bound on the expected return of a policy.

Introduction

In this paper we show how trajectories generated by some policies, called *behavior policies*, can be used to compute the confidence that the expected return of a different policy, called the *evaluation policy*, will exceed some lower bound. This *high confidence off-policy evaluation* mechanism has immense implications in industry, where execution of a new policy can be costly or dangerous if it performs worse than the policy that is currently being used. There are many applications that are hindered by such safety concerns, including news recommendation systems (Li et al. 2010), patient diagnosis systems (Hauskrecht and Fraser 2000), neuroprosthetic control (Thomas et al. 2009; Pilarski et al. 2011), automatic Propofol administration (Moore et al. 2010), and lifetime value optimization in marketing systems (Silver et al. 2013). In our experiments we show how our algorithm can be applied to a digital marketing problem where confidence bounds are necessary to motivate the potentially risky gamble of executing a new policy.

Although the off-policy evaluation problem has been solved efficiently in the multi-arm bandit case (Li et al. 2011), it is still an open question for sequential decision problems. Existing methods for estimating the performance of the evaluation policy using trajectories from behavior policies do not provide confidence bounds (Maei and Sutton 2010; Liu, Mahadevan, and Liu 2012; Mandel et al. 2014).

Our approach is straightforward—for each trajectory, we use importance sampling to generate an *importance weighted return*, which is an unbiased estimate of the expected return of the evaluation policy (Precup, Sutton, and Singh 2000). These importance weighted returns can be used by a *concentration inequality* (Massart 2007) to get a confidence bound on the expected return. The primary challenge of this approach is that the importance weighted returns have high variance and a large possible range, both of which loosen the bounds produced by existing concentration inequalities. We therefore derive a novel modification of an existing concentration inequality that makes it particularly well suited to this setting.

In the rest of this paper we explain how to generate the unbiased estimates, describe existing concentration inequalities, derive our new concentration inequality, and provide supporting empirical studies that show both the necessity of high confidence off-policy evaluation methods and the viability of our approach.

Preliminaries

Let $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $r_t \in [r_{\min}, r_{\max}]$ be the bounded reward at time t, and $\gamma \in [0,1]$ be a discount factor.[1] We denote by $\pi(a|s,\theta)$ the probability (density) of taking action a in state s when using *policy parameters* $\theta \in \mathbb{R}^{n_\theta}$, where n_θ is a positive integer—the dimension of the policy parameter space. A *trajectory* of length T is an ordered set of states (observations), actions, and rewards: $\tau = \{s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T\}$. We define the (normalized and discounted) *return* of a trajectory to be

$$R(\tau) := \frac{\left(\sum_{t=1}^{T} \gamma^{t-1} r_t\right) - R_-}{R_+ - R_-} \in [0,1],$$

where R_- and R_+ are upper and lower bounds on $\sum_{t=1}^{T} \gamma^{t-1} r_t$. In the absence of domain-specific knowledge, the loose bounds $R_- = r_{\min}(1-\gamma^T)/(1-\gamma)$ and $R_+ = r_{\max}(1-\gamma^T)/(1-\gamma)$ can be used. Let

$$\rho(\theta) := \mathbb{E}\left[R(\tau)|\theta\right]$$

[1]Although we use MDP notation, by replacing states with observations, our results carry over to POMDPs with reactive policies.

denote the performance of policy parameters θ, i.e., the expected discounted return when using policy parameters θ. We assume that all trajectories are of length at most T.

We assume that we are given a data set, $\mathcal{D}$, that consists of n trajectories, $\{\tau_i\}_{i=1}^n$, each labeled by the policy parameters that generated them, $\{\theta_i\}_{i=1}^n$, i.e.,[2]

$$\mathcal{D} = \big\{ (\tau_i, \theta_i) : i \in \{1, \ldots, n\}, \tau_i \text{ generated using } \theta_i \big\}.$$

Note that $\{\theta_i\}_{i=1}^n$ are *behavior* polices—those that generated the batch of data (trajectories). Finally, we denote by θ the *evaluation* policy—the one that should be evaluated using the data set $\mathcal{D}$. Although some trajectories in $\mathcal{D}$ may have been generated using the evaluation policy, we are particularly interested in the setting where some or all of the behavior policies are different from the evaluation policy. As described in the introduction, our goal is to present a mechanism that takes a confidence, $(1-\delta) \in [0,1]$, as input and returns a corresponding lower bound, $\rho_- \in [0,1]$, for the performance of the evaluation policy, $\rho(\theta)$. The mechanism should also be able to take a lower bound, $\rho_- \in [0,1]$, as input and return the confidence, $1-\delta$, that ρ_- is a lower bound on $\rho(\theta)$.

Generating Unbiased Estimates of $\rho(\theta)$

Our approach relies on our ability to take an element $(\tau, \theta_i) \in \mathcal{D}$, i.e., a trajectory τ generated by a behavior policy θ_i, and compute an unbiased estimate, $\hat{\rho}(\theta, \tau, \theta_i)$, of the performance of the evaluation policy $\rho(\theta)$. We use *importance sampling* (Precup, Sutton, and Singh 2000) to generate these unbiased estimates:[3]

$$\hat{\rho}(\theta, \tau, \theta_i) = R(\tau)\frac{\Pr(\tau|\theta)}{\Pr(\tau|\theta_i)} := \underbrace{R(\tau)}_{\text{return}} \underbrace{\prod_{t=1}^{T} \frac{\pi(a_t|s_t,\theta)}{\pi(a_t|s_t,\theta_i)}}_{\text{importance weight}}, \quad (1)$$

where $\Pr(\tau|\theta)$ is the probability that trajectory τ is generated by following policy θ. Note that we do *not* need to require $\pi(a|s,\theta_i) > 0$ for all s and a in (1), since division by zero can never occur in this equation. This is because a_t would have never been chosen in trajectory τ_i if $\pi(a_t|s_t,\theta_i) = 0$.

For each θ_i, $\hat{\rho}(\theta, \tau, \theta_i)$ is a random variable that can be sampled by generating a trajectory, τ, using policy parameters θ_i, and then using (1). If $\pi(a|s,\theta) = 0$ for all s and a where $\pi(a|s,\theta_i) = 0$, then importance sampling is unbiased, i.e., $\mathbb{E}[\hat{\rho}(\theta, \tau, \theta_i)] = \rho(\theta)$. However, it is important to consider what happens when this is not the case—when there is one or more state-action pair, s, a, where $\pi(a|s,\theta) > 0$ but $\pi(a|s,\theta_i) = 0$. In this case $\hat{\rho}(\theta, \tau, \theta_i)$ is a biased estimator because it does not have access to data that can be used to evaluate the outcome of taking action a in state s. For simplicity, we avoid this by assuming that if $\pi(a|s,\theta_i) = 0$, then $\pi(a|s,\theta) = 0$. In the appendix we explain this in more detail and show that *the lower bound that we propose holds even without this assumption*.

[2]Note that θ_i denotes the parameter vector of the ith trajectory and not the ith element of θ.

[3]*Per-decision importance sampling* (Precup, Sutton, and Singh 2000) is another unbiased estimator that could be used in place of ordinary importance sampling. Here we use ordinary importance sampling due to its simplicity.

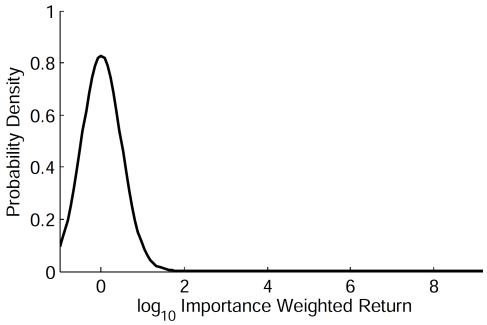

Figure 1: Empirical estimate of the *probability density function* (PDF) of $\hat{\rho}(\theta, \tau, \theta_i)$ on a simplified version of the mountain-car problem (Sutton and Barto 1998) with $T = 20$. The behavior policy, θ_i, corresponds to a suboptimal policy and the evaluation policy, θ, is selected along the natural policy gradient from θ_i. The PDF is estimated from 100,000 trajectories. It is important to note that while the tightest upper bound on $\hat{\rho}(\theta, \tau, \theta_i)$ is approximately $10^{9.4}$, the largest observed importance weighted return is only around 316. The sample mean is about $0.191 \approx 10^{-0.72}$. Note that the horizontal axis is scaled logarithmically. Also, the upper bound on $\hat{\rho}(\theta, \tau, \theta_i)$ was computed using a brute force search for the s, a pair that maximizes $\pi(a|s,\theta)/\pi(a|s,\theta_i)$. Without domain-specific knowledge, this state-action pair could occur at every time step and could result in a return of 1, making the largest possible importance weighted return $(\pi(a|s,\theta)/\pi(a|s,\theta_i))^T$.

Since the smallest possible return is zero and the importance weights are nonnegative, the *importance weighted returns*, $\hat{\rho}(\theta, \tau, \theta_i)$, are bounded from below by zero. However, when the action selected in a state by the behavior policy has low probability under the behavior policy but high probability under the evaluation policy, i.e., $\pi(a_t|s_t,\theta_i)$ is small and $\pi(a_t|s_t,\theta)$ is large, then the corresponding importance weighted return, $\hat{\rho}(\theta, \tau, \theta_i)$, might be large. So, the random variables $\hat{\rho}(\theta, \tau, \theta_i)$ are bounded from below by zero, have expected value in $[0,1]$, and may have a large upper bound. This means that $\hat{\rho}(\theta, \tau, \theta_i)$ often has a very long tail, as shown in Fig. 1. Thus, the primary challenge of our endeavor is to measure and account for this large range and high variance to produce a tight bound on $\rho(\theta)$.

Concentration Inequality

We consider three concentration inequalities that provide probability bounds on how a random variable deviates from its expectation. Let $X_1, \ldots, X_n$ be n independent real-valued bounded random variables such that for each $i \in \{1, \ldots, n\}$, we have $\Pr(X_i \in [0, b_i]) = 1$ and $\mathbb{E}[X_i] = \mu$. In the context or our problem, the X_i correspond to the importance weighted returns, $\hat{\rho}(\theta, \tau, \theta_i)$, the uniform mean, μ, is $\rho(\theta)$, and b_i is an upper bound on $\hat{\rho}(\theta, \tau, \theta_i)$. Recall from

the last section that b_i can be exceedingly large—about $10^{9.4}$ in the example of Fig. 1. In the following, for simplicity, we assume that all the random variables have the same upper bound, b.

Chernoff-Hoeffding (CH) inequality: This bound indicates that with probability at least $1-\delta$, we have

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}.$$

Maurer & Pontil's empirical Bernstein (MPeB) inequality (Maurer and Pontil 2009, Theorem 11)**:** This bound replaces the true (unknown in our setting) variance in Bernstein's inequality with the sample variance. The MPeB inequality states that with probability at least $1-\delta$, we have[4]

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{7b\ln(2/\delta)}{3(n-1)} - \frac{1}{n}\sqrt{\frac{\ln\left(2/\delta\right)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2}.$$

Anderson (AM) inequality: The Anderson inequality (Anderson 1969) is based on the Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky, Kiefer, and Wolfowitz 1956), and the variant we use here is with the optimal constants found by Massart (1990). The AM inequality states that with probability at least $1-\delta$, we have

$$\mu \geq z_n - \sum_{i=0}^{n-1}(z_{i+1} - z_i)\min\left\{1, \frac{i}{n} + \sqrt{\frac{\ln\left(2/\delta\right)}{2n}}\right\}.$$

where $z_1, \ldots, z_n$ are the samples of the random variables $X_1, X_2, \ldots, X_n$, sorted such that $z_1 \leq z_2 \leq \ldots \leq z_n$, and $z_0 = 0$. Unlike the CH and MPeB bounds, which hold for independent random variables, the AM inequality only holds for independent and identically distributed (i.i.d.) random variables, i.e., $X_1, \ldots, X_n$ should also be identically distributed. In the context of our problem, this means that the AM bound can only be used when all of the trajectories in $\mathcal{D}$ were generated by a single behavior policy.

Notice that the effect of the range, b, is decreased in MPeB relative to CH, since in MPeB the range is divided by n, whereas in CH it is divided by $\sqrt{n}$. While CH is based on the sample mean and MPeB is based on the sample mean and variance, AM takes into account the entire sample cumulative distribution function. This allows AM to only depend on the largest observed sample and not b. This can be a significant improvement in situations like the example of Fig. 1, where the largest observed sample is about 316, while b is approximately $10^{9.4}$. However, despite AM's desirable property that it does not depend on the range of the random variable, it is not suitable for our problem for two reasons: **1)** It tends to be looser than MPeB for random variables without long tails, due to its inherent reliance on the Kolmogorov-Smirnov statistic (Diouf and Dufour 2005, Section 4). **2)** As discussed earlier, unlike CH and MPeB, AM can be applied only if the random variables are i.i.d., and to the best of our knowledge, it is not obvious how to extend it to the setting in which the random variables are only independent and not identically distributed (in the context of our problem, this is when $\mathcal{D}$ is generated by multiple behavior policies).

[4]To obtain this from Maurer and Pontil's Thm. 11, we first normalize X and then apply Thm. 11 with $1 - X$ instead of X.

Our goal in the rest of this section is to extend MPeB so that it is useful for our policy evaluation problem, i.e., so that it is independent of the range of the random variables and able to handle random variables that have different ranges but the same mean. This results in a new concentration inequality that combines the desirable properties of MPeB (general tightness and applicability to random variables that are not identically distributed) with those of AM (no direct dependence on the range of the random variables). In the context of our policy evaluation problem, it also removes the need to determine a tight upper bound on the largest possible importance weighted return, which may require expert consideration of domain-specific properties.

Our new bound is an extension of MPeB that relies on two key insights: **1)** removing the upper tail of a distribution can only lower its expected value, and **2)** MPeB can be generalized to handle random variables with different ranges if it is simultaneously specialized to random variables with the same mean. We prove our new concentration inequality in Thm. 1. To prove this theorem, we collapse the tail of the distribution of the random variables, normalize the random variables so that the MPeB inequality can be applied, and then use MPeB to generate a lower-bound from which we extract a lower-bound on the mean of the original random variables. Our approach for collapsing the tails of the distributions and then bounding the means of the new distributions is similar to bounding the truncated mean and is a form of Winsorization (Wilcox and Keselman 2003). Later we will discuss how the threshold values, c_i, can be selected automatically from the data.

Theorem 1. *Let $X_1, \ldots, X_n$ be n independent real-valued bounded random variables such that for each $i \in \{1, \ldots, n\}$, we have $\Pr(0 \leq X_i) = 1$, $\mathbb{E}[X_i] \leq \mu$, and the fixed real-valued threshold $c_i > 0$. Let $\delta > 0$ and $Y_i := \min\{X_i, c_i\}$. Then with probability at least $1-\delta$, we have*

$$\mu \geq \underbrace{\left(\sum_{i=1}^{n}\frac{1}{c_i}\right)^{-1}\sum_{i=1}^{n}\frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^{n}\frac{1}{c_i}\right)^{-1}\frac{7n\ln(2/\delta)}{3(n-1)}}_{\text{term that goes to zero as } 1/n \text{ as } n\to\infty}$$

$$- \underbrace{\left(\sum_{i=1}^{n}\frac{1}{c_i}\right)^{-1}\sqrt{\frac{\ln(2/\delta)}{n-1}\sum_{i,j=1}^{n}\left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{n} \text{ as } n\to\infty}. \tag{2}$$

Proof. We define n independent random variables, $\mathbf{Z} = \{Z_i\}_{i=1}^n$, as $Z_i := \frac{Y_i}{c_i}$. Thus, we have

$$\bar{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{c_i}. \tag{3}$$

Since $\mathbb{E}[Y_i] \leq \mathbb{E}[X_i] \leq \mu$, we may write

$$\mathbb{E}[\bar{Z}] = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}[Y_i]}{c_i} \leq \frac{\mu}{n}\sum_{i=1}^{n}\frac{1}{c_i}. \tag{4}$$

Notice that the Z_i random variables, and therefore also the $(1 - Z_i)$ random variables, are n independent random variables with values in $[0, 1]$. So, using Thm. 11 of Maurer and Pontil (2009), with probability at least $1 - \delta$, we have

$$\mathbb{E}[1 - \bar{Z}] \leq 1 - \bar{Z} + \sqrt{\frac{2V_n(1 - \mathbf{Z}) \ln(2/\delta)}{n}} + \frac{7 \ln(2/\delta)}{3(n-1)}, \quad (5)$$

where the empirical variance, $V_n(1 - \mathbf{Z})$, is defined as

$$V_n(1 - \mathbf{Z}) := \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} \big((1 - Z_i) - (1 - Z_j)\big)^2 = \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2. \quad (6)$$

The claim follows by replacing $\bar{Z}$, $\mathbb{E}[\bar{Z}]$, and $V_n(1 - \mathbf{Z})$ in (5) with (3), (4), and (6). ■

Remark 1: Notice that if $\Pr(X_i \leq b_i) = 1$ and $c_i = b_i$ for all i, then Thm. 1 degenerates to Thm. 11 of Maurer and Pontil (2009).

Remark 2: Thm. 1 allows us to take evaluation policy parameters θ, a set of trajectories, $\mathcal{D}$, generated by several behavior policies, and a confidence level, $(1 - \delta)$, as input, and return a probabilistic lower bound on the performance of this policy, $\rho(\theta) = \mu$. The lower bound is the *right-hand-side* (RHS) of (2).

Remark 3: Despite the nested sum, $\sum_{i,j}$, the RHS of (2) can be evaluated in linear time (a single pass over the samples), since we may write

$$\sum_{i,j=1}^{n} (A_i - A_j)^2 = 2n \sum_{i=1}^{n} A_i^2 - 2 \left(\sum_{i=1}^{n} A_i\right)^2,$$

and so (2) may be rewritten as

$$\mu \geq \left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \left[\sum_{i=1}^{n} \frac{Y_i}{c_i} - \frac{7n \ln(2/\delta)}{3(n-1)} - \sqrt{\frac{2 \ln(2/\delta)}{n-1} \left(n \sum_{i=1}^{n} \left(\frac{Y_i}{c_i}\right)^2 - \left(\sum_{i=1}^{n} \frac{Y_i}{c_i}\right)^2\right)}\right].$$

Remark 4: Whereas in Remark 1 we used a confidence, $1 - \delta$, to compute a lower bound on $\rho(\theta) = \mu$, the bound can also be inverted to produce a confidence from a lower bound, $\mu_- \geq 0$. Let

$$k_1 = \frac{7n}{3(n-1)}, \qquad k_3 = \mu_- \sum_{i=1}^{n} \frac{1}{c_i} - \sum_{i=1}^{n} \frac{Y_i}{c_i},$$

$$k_2 = \sqrt{\frac{2}{(n-1)} \left(n \sum_{i=1}^{n} \left(\frac{Y_i}{c_i}\right)^2 - \left(\sum_{i=1}^{n} \frac{Y_i}{c_i}\right)^2\right)},$$

$$\zeta = \frac{-k_2 + \sqrt{k_2^2 - 4k_1k_3}}{2k_1}.$$

Then our confidence that $\mu \geq \mu_-$ is

$$1 - \delta = \begin{cases} 1 - \min\{1, 2\exp(-\zeta^2)\} & \text{if } \zeta \text{ is real and positive,} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 5: In order to use the result of Thm. 1 in our policy evaluation application, we must select the values of the c_i, i.e., the thresholds beyond which the distributions of the X_i are collapsed. To simplify this procedure, we select a single $c > 0$ and set $c_i = c$ for all i. When c is too large, it loosens the bound just like a large range b does. On the other hand, when c is too small, it decreases the expected values of the Y_i, which also loosens the bound. The optimal c must properly balance this trade-off between the range and mean of the Y_i. Fig. 2 illustrates this trade-off for the mountain car problem described in Fig. 1.

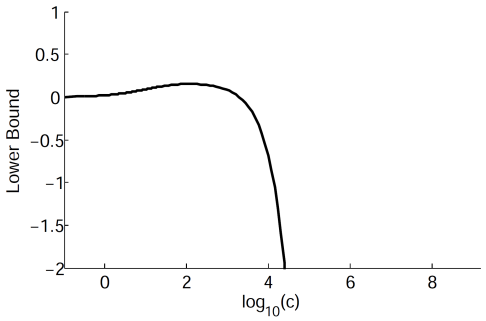

Figure 2: The lower bound μ_- from Thm. 1 when using different values of c on the 100,000 trajectories used to generate Fig. 1. The optimal value of c is around 100, which equates to collapsing the tail of the distribution in Fig. 1 at 100. The curve continues below the horizontal axis down to $-129{,}703$ for $c = 10^{9.4}$, i.e., the upper bound on $\hat{\rho}(\theta, \tau, \theta_i)$.

Thm. 1 requires the thresholds, c_i, to be fixed—i.e., they should not be computed using realizations of any X_i. So, we partition the data set, $\mathcal{D}$, into two sets, $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}$. $\mathcal{D}_{\text{pre}}$ is used to estimate the optimal threshold, c, and $\mathcal{D}_{\text{post}}$ is used to compute the lower bound (the RHS of (2)). The optimal value of c is the one that results in the largest lower bound, i.e., maximizes the RHS of (2). Note that the RHS of (2) depends on the sample mean and sample variance. We will use the sample mean and variance of $\mathcal{D}_{\text{pre}}$ to predict what the sample mean and variance in the RHS of (2) would be if we use $\mathcal{D}_{\text{post}}$ and a specific value of c. We then select a c^* that maximizes this prediction, i.e.,

$$c^* \in \arg\max_c \quad \underbrace{\frac{1}{n_{\text{pre}}} \sum_{i=1}^{n_{\text{pre}}} Y_i}_{\text{prediction of } \mathcal{D}_{\text{post}}\text{'s sample mean}} - \frac{7c \ln(2/\delta)}{3(n_{\text{post}} - 1)} - \sqrt{\frac{\ln(2/\delta)}{n_{\text{post}}} \underbrace{\frac{2}{n_{\text{pre}}(n_{\text{pre}} - 1)} \left(n_{\text{pre}} \sum_{i=1}^{n_{\text{pre}}} Y_i^2 - \left(\sum_{i=1}^{n_{\text{pre}}} Y_i\right)^2\right)}_{\text{prediction of } \mathcal{D}_{\text{post}}\text{'s sample variance}}} \quad (7)$$

Recall that $Y_i := \min\{X_i, c_i\}$, so all the three terms in (7) depend on c. Once an estimate of the optimal threshold, c^*, has been formed from $\mathcal{D}_{\text{pre}}$, Thm. 1 is applied using c^* and the samples in $\mathcal{D}_{\text{post}}$. From our preliminary experiments we found that using $1/20$ of the samples in $\mathcal{D}_{\text{pre}}$ and the remaining $19/20$ in $\mathcal{D}_{\text{post}}$ works well. In our application we know the

	Theorem 1	CH	MPeB	AM
ρ_-	0.154	$-5{,}831{,}000$	$-129{,}703$	0.055

Table 1: A comparison of 95% confidence lower bounds on $\rho(\theta)$. The $100{,}000$ trajectories and evaluation policy are the same as in Fig. 1 and 2. The sample mean (average importance weighted return) is 0.191.

true mean is in $[0, 1]$, so we require $c^* \geq 1$. When some of the random variables are identically distributed, we ensure that they are divided with $1/20$ in $\mathcal{D}_{\text{pre}}$ and $19/20$ in $\mathcal{D}_{\text{post}}$.[5]

Experiments

Mountain Car

We used the mountain car data from Fig. 1 to compare the lower bounds found when using Thm. 1, CH, MPeB, and AM. The results are provided in Table 1. These results reflect our previous discussion—the large possible range of $\hat{\rho}(\theta, \tau, \theta_i)$ causes CH to perform poorly since it scales with $b/\sqrt{n}$. MPeB is the next worst since it scales with $b/(n-1)$. AM performs reasonably well since it depends on the largest observed sample, $z_{100000} \approx 316$, rather than $b \approx 10^{9.4}$. However, as expected, Thm. 1 performs the best because it combines the tightness of MPeB with AM's lack of dependence on b.

Digital Marketing using Real-World Data

Adobe Marketing Cloud is a powerful set of tools that allows companies to fully leverage digital marketing using both automated and manual solutions. It has been deployed widely across the internet, with approximately seven out of every ten dollars transacted on the web passing through one of Adobe's products. Adobe Target, one of the six core components of Adobe Marketing Cloud, allows for automated user-specific targeting of advertisements and campaigns. When a user requests a webpage that contains an advertisement, the decision of which advertisement to show is computed based on a vector containing all of the known features of the user.

This problem tends to be treated as a bandit problem, where an agent treats each advertisement as a possible action and attempts to maximize the probability that the user clicks on the advertisement. Although this greedy approach has been successful, it does not necessarily also maximize the total number of clicks from each user over his or her lifetime. It has been shown that more far-sighted reinforcement learning approaches to this problem can improve significantly upon bandit solutions (Theocharous and Hallak 2013).

In order to avoid the large costs associated with deployment of a bad policy, in this application it is imperative that new policies proposed by RL algorithms are thoroughly evaluated prior to execution. Because off-policy evaluation methods are known to have high variance, estimates of performance without associated confidences are not sufficient. However, our high-confidence off-policy evaluation method *can* provide sufficient evidence supporting the deployment of a new policy to warrant its execution.

For our second case study we used real data, captured with permission from the website of a Fortune 50 company that receives hundreds of thousands of visitors per day and which uses Adobe Target, to train a simulator using a proprietary in-house system identification tool at Adobe. The simulator produces a vector of 31 real-valued features that provide a compressed representation of all of the available information about a user. The advertisements are clustered into two high-level classes that the agent must select between. After the agent selects an advertisement, the user either clicks (reward of $+1$) or does not click (reward of 0) and the feature vector describing the user is updated. We selected $T = 20$ and $\gamma = 1$.

This is a particularly challenging problem because the reward signal is sparse—if each action is selected with probability 0.5 always, only about 0.38% of the transitions are rewarding, since users usually do not click on the advertisements. This means that most trajectories provide no feedback. Also, whether a user clicks or not is close to random, so returns have relatively high variance.

We generated data using an initial baseline policy and then evaluated a new policy proposed by an in-house reinforcement learning algorithm. Fig. 3 shows the 95% confidence lower bound produced using different numbers of trajectories and various concentration inequalities. As in the mountain car example, Thm. 1 significantly outperforms previously existing concentration inequalities.

Fig. 4, gives our confidence for every possible lower bound. This characterizes the risk associated with deployment of the new policy since it gives a confidence bound for every possible outcome—it bounds the probability of different levels of degradation in performance relative to the behavior policy as well as the probability of different amounts of improvement. Fig. 4 is an exceptionally compelling argument for deployment of the new policy in place of the behavior policy.

Conclusion and Future Work

We have presented a technique that can instill the user of an RL algorithm with confidence that a newly proposed policy will perform well, without requiring the new policy to actually be executed. This is accomplished by providing confidence bounds for various levels of performance degradation and improvement. Our ability to compute tight confidence bounds comes from a novel adaptation of an existing concentration inequality to make it particularly well suited to this application. Specifically, it can handle random variables that are not identically distributed and our experiments suggest that it is tight even for distributions with heavy upper tails.

Our scheme for automatically selecting the threshold parameter, c, is *ad hoc*. This could be improved, especially by a method for adaptively determining how many samples should be used to select c. Second, our creation of a practical lower bound on the expected return of a policy might

[5] In our policy evaluation application, the importance weighted returns from two trajectories are identically distributed if the trajectories were generated by the same behavior policy.

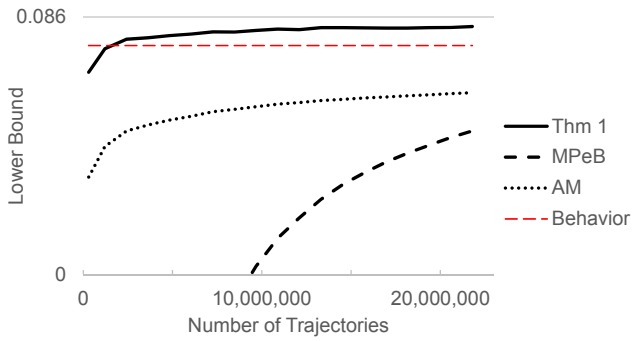

Figure 3: 95% confidence lower bound (unnormalized) on $\rho(\theta)$ using trajectories generated using the simulator described in the text. The behavior policy's true expected return is approximately 0.0765 (per step click probability of $0.0765/T \approx .38\%$) and is plotted as "Behavior". The evaluation policy's true expected return is approximately 0.086 (per step click probability of $\approx .43\%$). These two estimates of the policies' performances were computed by deploying them in the simulator for 1 million trajectories and computing the average return. The curves are averaged over ten trials, and the largest standard deviation is 0.002. CH never achieved a lower bound above -22, and so it is not visible in this plot. Notice that, using Thm. 1, we are able to guarantee that the evaluation policy is an improvement upon the behavior policy with at least 95% confidence using only 2 million trajectories (recall that the website has several hundred thousand visitors *per day*).

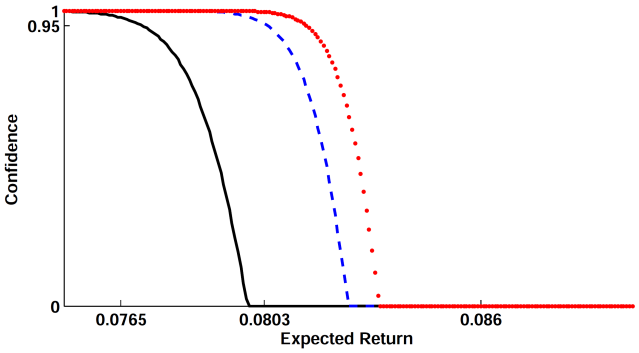

Figure 4: Our confidence that $\rho(\theta)$ is at least the (unnormalized) lower bound specified on the horizontal axis, as computed using Thm. 1. As in Fig. 3, the initial behavior policy's expected return is approximately 0.0765 and the evaluation policy's true (unknown in practice) expected return is approximately 0.086, both of which are marked on the horizontal axis. First we used 2 million trajectories from the behavior policy to compute lower bounds on the performance of the evaluation policy for every possible confidence. The result of this experiment is shown by the solid black line. The 95% confidence lower bound was 0.077. Although this means guaranteed improvement with high confidence, it is only a minor improvement that may not warrant the high overhead costs associated with deploying a new policy. We therefore collected an additional 3 million trajectories using the initial behavior policy and recomputed the lower bounds using all 5 million trajectories. The result of this experiment is shown by the dashed blue line, where the 95% confidence lower bound is 0.0803. We then deployed the evaluation policy and collected 1 million on-policy trajectories and recomputed the lower bounds once again. The result of this experiment using all 6 million trajectories is shown by the dotted red line. Using all 6 million trajectories resulted in a 95% confidence lower bound of 0.81, which supports continued deployment of the evaluation policy.

allow for the transfer of bandit algorithms to the sequential decision making setting.

Appendix

In this appendix we show that, for any $(\tau, \theta_i) \in \mathcal{D}$, $\hat{\rho}(\theta, \tau, \theta_i)$ is a random variable whose expectation is a lower bound on the performance of the evaluation policy, $\rho(\theta)$. Let $\mathcal{Y}$ and $\mathcal{Z}$ be the sets of trajectories, τ, such that $\Pr(\tau|\theta) \neq 0$ and $\Pr(\tau|\theta_i) \neq 0$, respectively, and let $\mathcal{Y}^c$ be the complement of $\mathcal{Y}$. Our claim comes from the following series of (in)equalities:

$$\begin{aligned}
\mathbb{E}_{\tau\sim\theta_i}\left[\hat{\rho}(\theta,\tau,\theta_i)\right] &= \mathbb{E}_{\tau\sim\theta_i}\left[R(\tau)\frac{\Pr(\tau|\theta)}{\Pr(\tau|\theta_i)}\right] \qquad (8)\\
&= \int_{\mathcal{Z}} R(\tau)\Pr(\tau|\theta)\,\mathrm{d}\tau\\
&= \int_{\mathcal{Y}} R(\tau)\Pr(\tau|\theta)\,\mathrm{d}\tau + \underbrace{\int_{\mathcal{Y}^c\cap\mathcal{Z}} R(\tau)\Pr(\tau|\theta)\,\mathrm{d}\tau}_{\overset{(a)}{=}0}\\
&\quad - \int_{\mathcal{Y}\cap\mathcal{Z}^c} R(\tau)\Pr(\tau|\theta)\,\mathrm{d}\tau\\
&\overset{(b)}{\leq} \int_{\mathcal{Y}} R(\tau)\Pr(\tau|\theta)\,\mathrm{d}\tau = \mathbb{E}_{\tau\sim\theta}\left[R(\tau)\right] = \rho(\theta),
\end{aligned}$$

where $\mathbb{E}_{\tau\sim\theta}$ denotes the expected value when the trajectories, τ, are generated using policy parameters θ. In (8),

(a) This integral is zero because, from the definition of $\mathcal{Y}$, we have $\Pr(\tau|\theta) = 0$, for each $\tau \in \mathcal{Y}^c \cap \mathcal{Z}$.
(b) This inequality holds because $R(\tau) \geq 0$.

It is clear from (8) that if the support of $\Pr(\tau|\theta)$ (the evaluation policy) is a subset of the support of $\Pr(\tau|\theta_i)$ (the behavior policy), then $\int_{\mathcal{Y}\cap\mathcal{Z}^c} R(\tau)\Pr(\tau|\theta)d\tau = 0$, and as a result, $\mathbb{E}_{\tau\sim\theta_i}\left[\hat{\rho}(\theta,\tau,\theta_i)\right] = \rho(\theta)$, which means $\hat{\rho}(\theta,\tau,\theta_i)$ is an unbiased estimate of $\rho(\theta)$. However, if this is not the case, i.e., $\mathbb{E}_{\tau\sim\theta_i}\left[\hat{\rho}(\theta,\tau,\theta_i)\right] \leq \rho(\theta)$, our results are still valid, because later in the paper we will find a lower-bound on $\mathbb{E}_{\tau\sim\theta_i}\left[\hat{\rho}(\theta,\tau,\theta_i)\right]$, which would also be a lower bound on $\rho(\theta)$ (our quantity of interest).

References

Anderson, T. W. 1969. Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. *Bulletin of The International and Statistical Institute* 43:249–251.

Diouf, M. A., and Dufour, J. M. 2005. Improved nonparametric inference for the mean of a bounded random variable with applicaiton to poverty measures.

Dvoretzky, A.; Kiefer, J.; and Wolfowitz, J. 1956. Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics* 27:642–669.

Hauskrecht, M., and Fraser, H. 2000. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial Intelligence in Medicine* 18:221–244.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 661–670.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth International Conference on Web Search and Web Data Mining*, 297–306.

Liu, B.; Mahadevan, S.; and Liu, J. 2012. Regularized off-policy TD-learning. In *Advances in Neural Information Processing Systems*.

Maei, H. R., and Sutton, R. S. 2010. GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Third Conference on Artificial General Intelligence*, 91–96.

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

Massart, P. 1990. The tight constraint in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability* 18(3):1269–1283.

Massart, P. 2007. *Concentration Inequalities and Model Selection*. Springer.

Maurer, A., and Pontil, M. 2009. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, 115–124.

Moore, B.; Panousis, P.; Kulkarni, V.; Pyeatt, L.; and Doufas, A. 2010. Reinforcement learning for closed-loop propofol anesthesia: A human volunteer study. In *Innovative Applications of Artificial Intelligence*, 1807–1813.

Pilarski, P. M.; Dawson, M. R.; Degris, T.; Fahimi, F.; Carey, J. P.; and Sutton, R. S. 2011. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *Proceedings of the 2011 IEEE International Conference on Rehabilitation Robotics*, 134–140.

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

Silver, D.; Newnham, L.; Barker, D.; Weller, S.; and McFall, J. 2013. Concurrent reinforcement learning from customer interactions. In *The Thirtieth International Conference on Machine Learning*.

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

Theocharous, G., and Hallak, A. 2013. Lifetime value marketing using reinforcement learning. In *The 1st Multidisciplinary Conference on Reinforcement Learning and Decision Making*.

Thomas, P. S.; Branicky, M. S.; van den Bogert, A. J.; and Jagodnik, K. M. 2009. Application of the actor-critic architecture to functional electrical stimulation control of a human arm. In *Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence*, 165–172.

Wilcox, R. R., and Keselman, H. J. 2003. Modern robust data analysis methods: Measures of central tendency. *Psychological Methods* 8(3):254–274.