Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning

Philip S. Thomas PHILIPT@CS.CMU.EDU
Dhruva Tirumala DTIRUMAL@ANDREW.CMU.EDU
Emma Brunskill EBRUN@CS.CMU.EDU
Carnegie Mellon University

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 estimates and importance sampling based estimates.

1. Introduction

Off-policy policy evaluation (OPE)—using existing data obtained from a behavior policy to predict the performance of an alternate policy—can be very useful in both online and batch reinforcement learning. In online settings, it can enable more data-efficient learning during policy search, instead of relying on Monte Carlo estimates for individual policies. In batch settings, its use is often crucial in order to decide which policy to deploy in the future in high stakes settings, such as determining which advertisement to show a user visiting a website (Theocharous et al., 2015), informing patient treatment strategies (Thapa et al., 2005), or creating adaptive instruction (Mandel et al., 2014).

Typically we are interested in OPE estimators that have good theoretical or algorithmic properties, or both. On the theory side, we may desire estimators that are unbiased, strongly consistent (will converge asymptotically to the true policy value) or have finite sample guarantees on their produced values. On the empirical side, almost all prior work evaluates estimators in terms of their mean squared error (MSE) (Precup et al., 2000; Dudík et al., 2011; Mahmood et al., 2014; Thomas, 2015; Jiang & Li, 2015).

Despite this, to our knowledge there are no existing OPE estimators that try to directly minimize the MSE. In our work we introduce a new framework for constructing OPE estimators that directly attempt to minimize MSE by blending together estimators based on models with estimators that leverage importance sampling. We also introduce a particular such estimator, *model and guided importance sampling combined* (MAGIC), and prove that it is also strongly consistent. Our empirical results show that MAGIC can produce estimates with orders of magnitude lower mean squared error than the estimates produced by existing algorithms.

2. Off-Policy Policy Evaluation (OPE)

We define the OPE problem as follows. We are given an evaluation policy, π_e, (or policies) and n episodes of *historical data*, D, generated from (one or more) behavior policies π_i, $i \in \{1, \ldots, n\}$. Our goal is to produce an estimator, $\hat{v}(D)$, of the performance, $v(\pi_e)$, of an evaluation policy, π_e, which has low *mean squared error* (MSE): $\text{MSE}(\hat{v}(D), v(\pi_e)) \coloneqq \mathbf{E}\left[(\hat{v}(D) - v(\pi_e))^2\right]$. We assume that the process producing states, actions, and rewards is an MDP with an unknown initial state distribution, transition function, and reward function. We assume that the evaluation policy, π_e, the behavior policies, and the discount parameter, γ, are known.

There are two common OPE estimators. One uses an MDP model of the domain to estimate the performance of π_e. This model may be provided by an outside source (such as an expert) or estimated from the historical data, D. Model-based estimates are low variance (Mannor et al., 2007) but are biased and not consistent estimators when the model representation chosen does not match the true environment. Unfortunately this is common in real world settings, since most models will only approximate the real environment.

A second option is to use importance sampling (Precup et al., 2000). Importance sampling works by taking Monte

Carlo estimates of the policy performance, where the distribution of the behavior policy is transformed into the distribution over the evaluation policy using importance sampling. Standard importance sampling is unbiased and strongly consistent, but typically creates high variance estimates, particularly in long horizon domains. Though importance sampling has been used successfully in short horizon settings, like online educational games (Mandel et al., 2014), and there has been work to mitigate the variance of IS estimators (Thomas et al., 2015; Thomas, 2015), the produced estimates are typically too high variance to be informative in long horizon settings with limited data.

A promising recent alternative is the *doubly robust* (DR) estimator (Jiang & Li, 2015). DR is a new unbiased estimator of $v(\pi_e)$ that leverages an approximate model of an MDP to decrease the variance of the unbiased estimates produced by ordinary importance sampling. 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; Heejung & Robins, 2005).

3. Blending IS and Model (BIM) Estimator

Despite its promise, DR is limited because it restricts the resulting estimator to be unbiased. Given the common emphasis on creating minimum MSE estimators, an estimator that incorporates a slight amount of bias may offer substantially lower variance, and thus MSE. Towards this, we introduce a framework that, like DR, combines purely model-based estimators (*approximate model* (AM) estimators) with estimators that incorporate importance sampling, but which does so in a way that directly minimizes MSE.

More specifically, we construct a range of estimators that each blend together AM and IS estimators. We can 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 the j^{th} time step, and the approximate model estimator to predict the outcomes thereafter. Note that $g^{(-1)}(D)$ is a purely model-based estimator, $g^{(\infty)}(D)$ is an importance sampling estimator, and other $g^{j}(D)$ are partial importance sampling estimators that blend between these two extremes. As j increases, we expect the variance of the return to increase, but the bias to decrease.

We propose a new estimator framework, 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. The weights are selected in order to (approximately) minimize the estimated mean squared error. We do this by computing a conservative estimate of the bias of the chosen approximate model, and an estimate of the covariance of the returns.

MAGIC is a particular BIM estimator that uses a weighted extension of the DR estimator, combined with an approximate model estimator. We have proven that MAGIC is a strongly consistent estimator given mild assumptions.

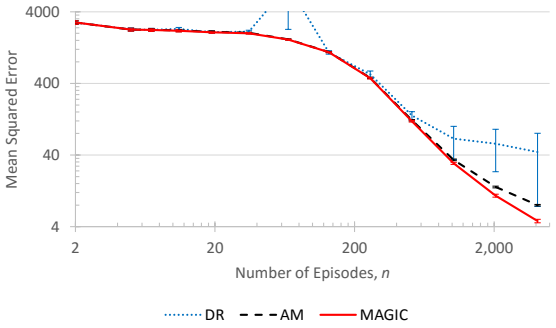

Figure 1: MSE of DR, AM, and MAGIC estimators on the mountain car domain given different amounts of historical data (n episodes) and with standard error bars.

4. Experimental Results

In our simulation results, MAGIC meets or reduces, in some cases by orders of magnitude, the MSE of prior OPE estimators. Whereas our prior experiments considerend only domains with finite states and actions (Thomas & Brunskill, 2016), here in Figure 1 we present results using MAGIC for the canonical mountain car domain, which has continuous states. We use the same behavior and evaluation policies used by Jiang & Li (2015) in the bottom-left plot of their Figure 1. Note that for mountain car, MAGIC always has the lowest MSE for any amount, n, of historical data. Also, MAGIC eventually exceeds the model performance (AM), since the model does not perfectly capture the underlying domain, and substantially outperforms DR by further leveraging the model.

5. Future Work

There exist ample future directions, including improving the underlying bias and variance estimators used to construct the MAGIC estimate, and using MAGIC for policy selection.

References

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.

Heejung, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

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

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.

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.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, February 2007.

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.

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. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015.

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

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, 2015.