Bootstrapping with Models:
Confidence Intervals for Off-Policy Evaluation

Josiah P. Hanna, Peter Stone, Scott Niekum
{jphanna,pstone,sniekum}@cs.utexas.edu
The University of Texas at Austin
2317 Speedway, Austin, TX 78712 USA

Abstract

In many reinforcement learning applications, it is desirable to determine confidence interval lower bounds on the performance of any given policy *without executing said policy*. In this context, we propose two bootstrapping off-policy evaluation methods which use learned MDP transition models in order to estimate lower confidence bounds on policy performance with limited data. We empirically evaluate the proposed methods in a standard policy evaluation tasks.[1]

Introduction

As *reinforcement learning* (RL) methods find application in the real world, it will be critical to establish the performance of policies with high confidence before they are executed. This problem is known as the *high confidence off-policy evaluation problem* (HCOPE). We propose data-efficient approximate solutions to this problem.

We propose two new bootstrap methods which use models to lower the variance of off-policy value estimates. We empirically evaluate both methods on two policy evaluation tasks and show these methods are far more data-efficient than existing importance sampling based approaches. Finally, we combine theoretical and empirical results to make specific recommendations about when to use different off-policy confidence bound methods in practice.

Problem Statement

We formalize our problem as a Markov decision process (MDP) defined as $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma, d_0 \rangle$ where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ is a distribution over next states given the current state and action, $r : \mathcal{S} \times \mathcal{A} \to [-r_{\texttt{min}}, r_{\texttt{max}}]$ is a reward function, $\gamma \in [0, 1]$ is a discount factor, and d_0 is an initial state distribution. An agent samples actions from a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ which is a probability mass function for actions conditioned on a given state.

A trajectory, H of length L is defined as a state-action history, $S_1, A_1, S_2, ...S_L, A_L$ where $S_1 \sim d_0$, $A_t \sim \pi(\cdot|s_t)$, and $S_{t+1} \sim P(\cdot|S_t, A_t)$. The return of a trajectory is $g(H) = \sum_{t=1}^{L} \gamma^{t-1} r(S_t, A_t)$. The policy, π, and transition dynamics, P, induce a distribution over trajectories, p_π. We also write $H \sim \pi$ to denote a trajectory sampled by executing π. The expected discounted return of a policy, π, is defined as $V(\pi) := E_{H \sim \pi}[g(H)]$.

Algorithm 1 Bootstrap Confidence Interval
Input is π_e, $\mathcal{D}$, δ, and a number of bootstrap estimates, B.

```
input $\pi_e, \mathcal{D}, \delta, B$
output $1-\delta$ confidence interval lower bound on $V(\pi_e)$.
for all $i \in [1, B]$ do
    $\tilde{\mathcal{D}}_i = \{H_1^i, ..., H_n^i\}$ where $H_j^i \sim \mathcal{D}$
    $\hat{V}_i$ = Off-PolicyEstimate($\pi_e, \tilde{\mathcal{D}}_i$)
end for
sort($\{\hat{V}_i | i \in [1, B]\}$) // Sort ascending
$l \leftarrow \lfloor \delta B \rfloor$
Return $\hat{V}_l$
```

Given a set of n trajectories, $\mathcal{D} = \{H_1, .., H_n\}$, where $H_i \sim \pi_b$ for some π_b, an evaluation policy, π_e, and a confidence level, δ, we propose two methods to approximate a confidence interval lower bound, $V_{-,\delta}(\pi_e)$, on $V(\pi_e)$ such that $V_{-,\delta}(\pi_e) < V(\pi_e)$ with probability $1 - \delta$.

Bootstrapping Policy Lower Bounds

Bootstrapping is a statistical technique that samples with replacement from a given sample to approximate confidence intervals. See Efron (1979) for further reading on bootstrap confidence intervals. Define **Off-PolicyEstimate** to be any method that takes a data set of trajectories $\mathcal{D}$ and a policy π_e and returns a policy value estimate $\hat{V}(\pi_e)$. Algorithm 1 gives a Bootstrap Confidence Interval procedure for computing a confidence interval on $\hat{V}(\pi_e)$, as computed by **Off-PolicyEstimate**. Since we desire lower bounds on $V(\pi_e)$ we give pseudocode for a bootstrap lower bound. The method is equally applicable to upper bounds and two sided intervals. A similar method using *weighted importance sampling* was proposed by Thomas et al. (2015).

Direct Model-Based Bootstrapping

The model-based off-policy estimator, MB, computes $\hat{V}(\pi_e)$ by first using all trajectories in $\mathcal{D}$ to build a model $\hat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \hat{P}, r, \gamma, \hat{d}_0 \rangle$ where $\hat{P}$ and $\hat{d}_0$ are estimated from data

[1] An extended version of this work can be found at https://arxiv.org/abs/1606.06126

generated by the behavior policy, π_b. Then we estimate $\hat{V}(\pi_e)$ as the value of π_e acting in $\hat{\mathcal{M}}$. Algorithm 1 with MB as **Off-PolicyEstimate** defines MB-BOOTSTRAP.

If a model can capture the true MDP's dynamics or generalize well to unseen parts of the state-action space then MB estimates can have low variance. However, models reduce variance at the cost of adding bias to the estimate. Model bias in MDPs can come from a lack of data or an incorrect choice of function approximator. The second source of bias is more problematic since even as $n \to \infty$ the bootstrap model estimates will converge to a different value from $V(\pi_e)$. However, in settings with low model bias, MB-BOOTSTRAP will have lower variance and thus tighter confidence bounds.

Bootstrapped Weighted Doubly Robust

We also propose *weighted doubly robust bootstrapping* (WDR-bootstrap) with the recently proposed WDR off-policy estimator for settings where the MB estimator may exhibit high bias. The WDR estimator uses per-decision weighted importance sampling (PDWIS) and a model to reduce variance in the estimate (Thomas and Brunskill 2016). Given a model and its state and action value functions for π_e, $\hat{v}_{\pi_e}$ and $\hat{q}_{\pi_e}$, the WDR estimator is defined as:

$$\text{WDR}(\mathcal{D}) := \text{PDWIS}(\mathcal{D}) - \sum_{i=1}^{n}\sum_{t=0}^{L}\gamma^t\left(w_t^i \hat{q}_{\pi_e}(S_t^i, A_t^i) - w_{t-1}^i \hat{v}_{\pi_e}(S_t^i)\right)$$

The second term serves as a control variate with expectation zero and thus WDR is an unbiased estimator of the consistent PDWIS estimator.

Although WDR is biased (since PDWIS is biased), the consistency property of PDWIS ensures that the bootstrap estimates of WDR-BOOTSTRAP will converge to the correct estimate as n increases. Empirical results have shown that WDR can acheive lower MSE than MB in domains where the model converges to an incorrect model (Thomas and Brunskill 2016). However, they also demonstrated situations where the MB evaluation is more efficient at acheiving low MSE than WDR when the variance of the PDWIS weights is high. We empirically analyze the trade-off when using these estimators with bootstrapping off-policy confidence bounds.

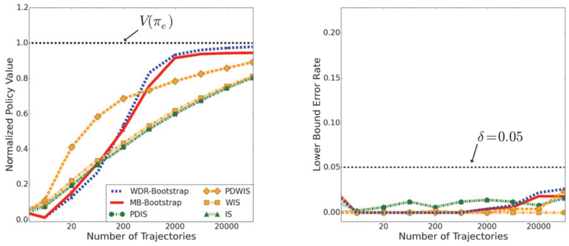
Empirical Results

We empirically evaluate MB-BOOTSTRAP and WDR-BOOTSTRAP by estimating 95 % confidence intervals ($\delta = 0.05$) in the standard mountain car task (Sutton and Barto 1998). In this domain we build tabular models (the mountain car state-action space is discretized) which cannot generalize from observed (s, a) pairs. We compute the model action value function, $\hat{q}_{\pi_e}$, and state value function, $\hat{v}_{\pi_e}$ with value-iteration for WDR. We use Monte Carlo rollouts to estimate $\hat{V}$ with MB. We also show results for importance sampling (IS) BCa-bootstrap methods from Thomas et. al. (2015). To the best of our knowledge, these IS methods are the current state-of-the-art for approximate HCOPE.

Figure 1a displays the average empirical 95 % confidence interval lower bound found by each method. The ideal result is a lower bound, $V_{-,\delta}(\pi_e)$, that is as large as possible subject to $V_{-,\delta}(\pi_e) < V(\pi_e)$. As a general trend we note that our proposed methods—MB-BOOTSTRAP and WDR-BOOTSTRAP—get closer to this ideal result with less data than all other methods. Figure 1b displays the empirical error rate for MB-BOOTSTRAP and WDR-BOOTSTRAP and shows that they approximate the allowable 5% error in each domain.

Figure 1a shows that both of our methods (WDR-BOOTSTRAP and MB-BOOTSTRAP) outperform IS based methods. The notable trend here is that both methods produce approximately the same average lower bound. Therefore, even though MB will eventually converge to $V(\pi_e)$ it does so no faster than WDR which can produce good estimates even when the model is inaccurate.

Figure 1b shows that the MB-BOOTSTRAP and WDR-BOOTSTRAP error rate is much lower than the required error rate yet Figure 1a shows the lower bound is no looser. Since MB-BOOTSTRAP and WDR-BOOTSTRAP are low variance estimators, the average bound can be tight with a low error rate. It is also notable that since bootstrapping only approximates the 5% allowable error rate all methods can do worse then 5% when data is extremely sparse (only two trajectories).



(a) Lower Bound

(b) Error Rate

Figure 1: The 95% lower bound on $V(\pi_e)$ and empirical error rate computed by each method.

Conclusion

We have introduced two novel methods—MB-BOOTSTRAP and WDR-BOOTSTRAP—that approximate confidence intervals for off-policy evaluation. Empirically, our methods yield superior data-efficiency and tighter lower bounds on $V(\pi_e)$ than state-of-the-art importance sampling based methods.

References

Efron, B. e. a. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7(1):1–26.

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

Thomas, P., and Brunskill, E. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1604.00923*.

Thomas, P. S.; Theocharous, G.; and Ghavamzadeh, M. 2015. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*.