Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation

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

Abstract

- Reinforcement learning policies lack **performance guarantees** until they are evaluated in the real world.
- **High Confidence Off-Policy Evaluation** (HCOPE) attempts to place confidence intervals on the value of a policy using existing **off-policy** domain data.
- We introduce two approximate HCOPE methods and demonstrate both **increase data-efficiency** in comparison to the previous state-of-the-art.
- We present a **theoretical bound** on the error in model-based estimates of a policy's value.

Background

Environment modelled as Markov Decision Process:

$$M = (\mathcal{S}, \mathcal{A}, r, P)$$

In state S_t at time step t:

- Agent selects action $A_t \sim \pi(\cdot|S_t)$
- Environment responds with $S_{t+1} \sim P(\cdot|S, A)$
- Reward $r(S_t, A_t)$ received after each action.

The policy and environment determine a distribution over trajectories, $H : S_1, A_1, S_2, A_2, ..., S_L, A_L$

Policy performance measured by its expected sum of rewards:

- $V(\pi) = \mathbb{E}\left[\sum_{t=1}^{L} r(S_t, A_t) \middle| H \sim \pi\right]$ is the expected return of π.

High Confidence Off-Policy Evaluation

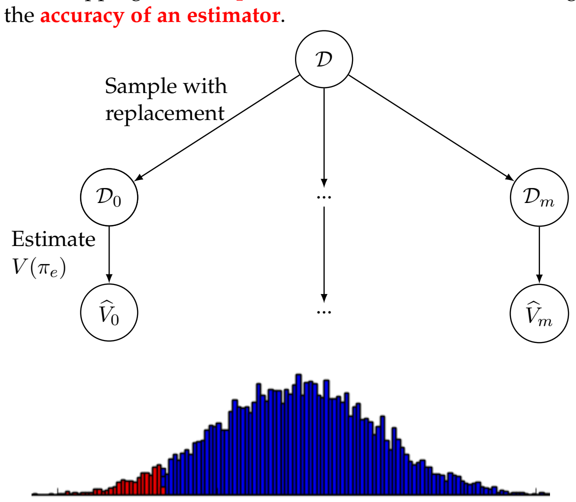
Given:

- An **evaluation policy** π_e.
- A data-set of trajectories, $\mathcal{D}$, generated by a known, **behavior policy** π_b.
- Confidence level $\delta \in [0, 1]$

Determine a **lower bound**, $\hat{V}_{\text{lb}}(\pi_e, \mathcal{D}, \pi_b)$ such that $V(\pi_e) \geq \hat{V}_{\text{lb}}(\pi_e, \mathcal{D}, \pi_b)$ with probability $(1 - \delta)$.

Bootstrap Confidence Intervals

Bootstrapping is a **non-parametric** method of determining the **accuracy of an estimator**.



Acknowledgments

This work has taken place in the Personal Autonomous Robotics Lab (PeARL) and Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. PeARL research is supported in part by the National Science Foundation under grant IIS1208497. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287, IIS-1637736, IIS-1651089), ONR (21C184-01), and AFOSR (FA9550-14-1-0087). Josiah Hanna is supported by an NSF Graduate Research Fellowship. Peter Stone serves on the Board of Directors of, Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

Off-Policy Evaluation

An **off-policy evaluation** (OPE) method predicts $V(\pi_e)$ given trajectories sampled from π_b.

Different OPE methods trade-off **bias** and **variance** differently:

Model-Based OPE

- Use $\mathcal{D}$ to estimate unknown transition probabilities as $\hat{P}$.
- Build a model, $\hat{M} = (\mathcal{S}, \mathcal{A}, r, \hat{P})$
- Estimate $V(\pi_e)$ as the value of π_e in $\hat{M}$.
- MB estimates **reduce variance** at the cost of **high bias** when the model is poor.

Weighted Doubly Robust OPE[2]

- Combines weighted importance-sampling with the state and state-action value functions of an approximate model.
- Approximate model value functions only serve as control variate — lowering variance **without adding model bias**.

Importance Sampling OPE[1]

- Let $\rho_t = \prod_{i=1}^{t} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}$
- $\text{IS}(\pi_e, H, \pi_b) := \rho_L \sum_{t=1}^{L} r(S_t, A_t)$
- **Unbiased** estimator for $V(\pi_e)$; potentially **high variance**.

Contributed Methods

We introduce **two novel bootstrap** off-policy approximate HCOPE methods:

- **MB-BOOTSTRAP** with the model-based estimator.
- **WDR-BOOTSTRAP** with the weighted doubly-robust estimator

Bootstrapping with importance sampling previously proposed by Thomas et al. [3].

Empirical Results

- MB-BOOTSTRAP and WDR-BOOTSTRAP evaluated on Mountain Car and Cliffworld domains.
- For varying n, π_b samples n trajectories and each method computes a **confidence interval lower bound** on $V(\pi_e)$.
- The ideal result is a lower bound that is close to but less than $V(\pi_e)$.
- We compare our proposed methods to bootstrapping with four variants of IS: standard IS, per-decision IS, weighted IS, and per-decision weighted IS.

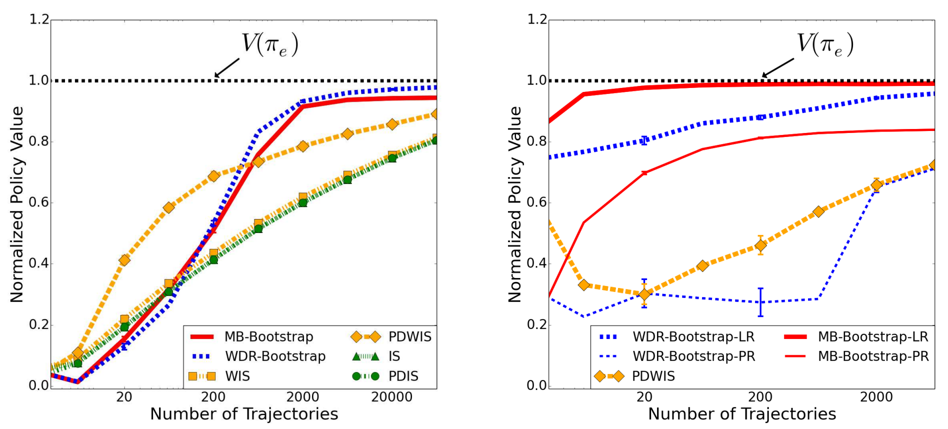


Figure 1: Left: the average empirical lower bound found by each method in the Mountain Car domain. Right: the average empirical lower bound found by each method in the Cliffworld domain. Our proposed methods — MB-BOOTSTRAP and WDR-BOOTSTRAP — **achieve tighter lower bounds** than other evaluated methods.

Method Summary

- Model-Based Bootstrap:
  - Preferable when environment dynamics can be easily estimated.
- Weighted Doubly Robust Bootstrap:
  - Lower bias than MB-BOOTSTRAP in settings where the MB estimator may have **high bias**.
- Cases where only MB-BOOTSTRAP is applicable:
  - **Deterministic** policies
  - **Unknown** behavior policies

Future Work

- Apply theoretical bounds on model bias to **guide model estimation** for MB-BOOSTRAP and WDR-BOOTSTRAP.
- Apply MB-BOOTSTRAP and WDR-BOOTSTRAP to **robotics tasks**.

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

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

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