Artificial Trajectories for Off-Policy Learning

Scott Sussex, Omer Gottesman

scottsussex@college.harvard.edu, gottesman@fas.harvard.edu

Abstract

Improving estimators for off-policy learning is an area of ongoing interest. The use of importance weights when evaluating a deterministic target policy leads to a small effective sample size. We explore the use of artificial trajectories to increase effective sample size in off policy learning. We evaluate our method empirically and show a reduction in mean squared error using artificial trajectories compared to existing methods.

Method

Artificial trajectories are constructed to increase the number of trajectories with nonzero importance weight. Stitching policies must produce importance weights where the MDP transition policies do not need to be known. We find a stitching policy that achieves this.

For a trajectory that reaches stitching points $s'_1 ... s'_k$, the importance weight for the trajectory is

$$\frac{\rho}{1 + \Sigma(1 - b(a'_i|s'_i))}$$

where ρ is the importance weight that would be calculated for the trajectory if there were no stitching policy, and a'_i is the action taken by the evaluation policy in state s'_i.

Since a stitching policy only modifies importance weights, the method can be incorporated into existing importance sampling techniques without having to trade off bias for a decrease in variance. Artificial trajectories methods can form an estimate when other methods might sample no trajectories with nonzero importance weights.

Our stitching policy places assumptions on the stitching points selected.

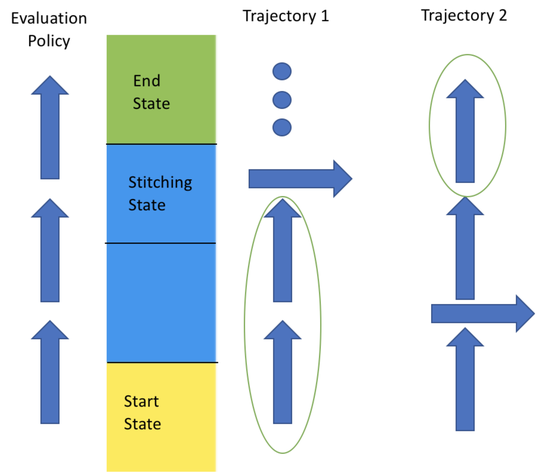

Figure 1: Demonstration of a stitching policy on a deterministic four state corridor gridworld.Trajectory 1 does not follow the evaluation policy after the stitching state. We sample any trajectory that does not follow the evaluation policy before the stitching state- trajectory 2. We then assemble the circled parts of each to form an artificial trajectory.

Experiments

We evaluate our method on stochastic gridworlds. The evaluation policy is the optimal deterministic policy. The behavior policy is ϵ-greedy with $\epsilon = 0.5$.

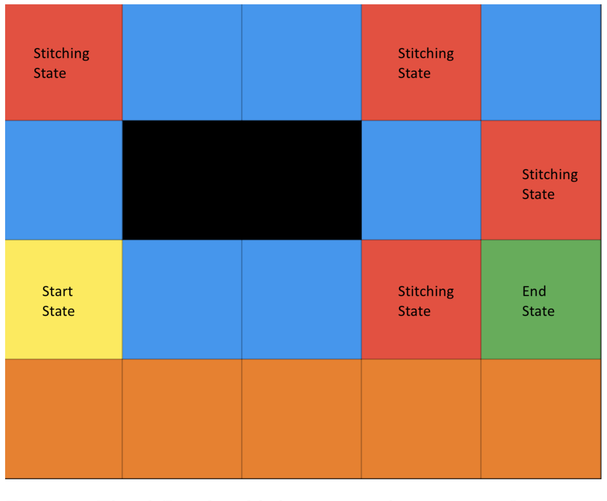

Figure 2: The cliff gridworld showing stitching points. Orange tiles indicate the pit states.

- Ordinary importance sampling (IS)
- Weighted importance sampling (WIS)
- Artificial trajectories ordinary importance sampling (ATIS)
- Artificial trajectories weighted importance sampling (ATWIS)
- Per-decision weighted doubly robust (PDWDR)

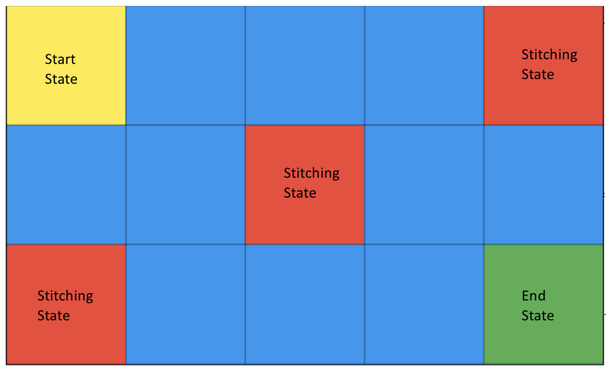

Figure 3: The 5 by 3 gridworld used in experiments.

Results

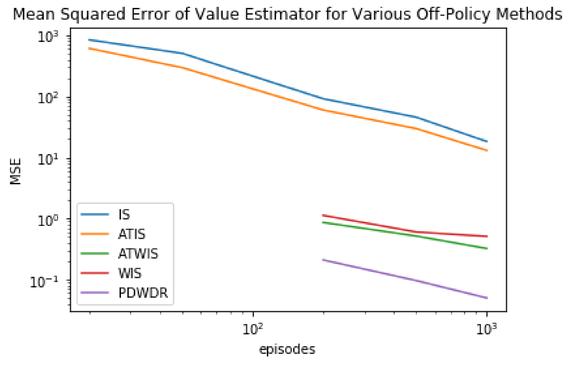

Figure 4: A plot of MSE for various importance sampling methods in a corridor gridworld.

All methods quickly tended towards having an accurate mean estimate, so the MSE graphs represent estimator variance. Artificial trajectories methods appear to lower mean squared error for ordinary and weighted importance sampling in both cases.

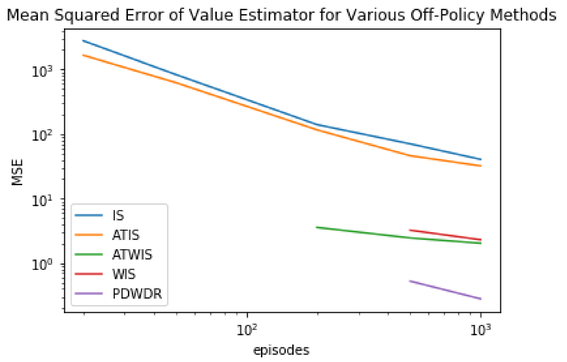

Figure 5: A plot of MSE for various importance sampling methods in a 3 by 5 stochastic gridworld.

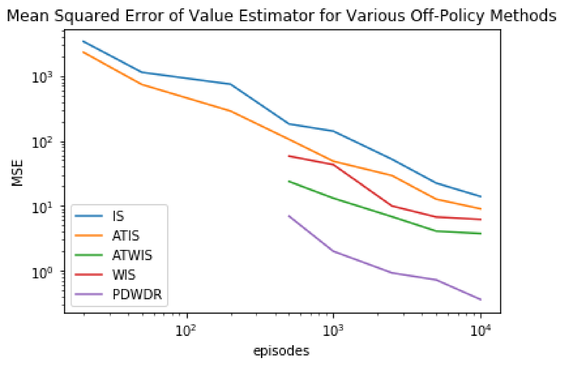

Figure 6: A plot of MSE for various importance sampling methods in a cliff gridworld.

Conclusion

We provide an encouraging first attempt with using artificial trajectories for importance sampling, however leave many open problems:

- Compare variance reductions from different stitching policies.
- Evaluate the method in more complex settings.
- Test the robustness of the method to the breaking of the Markovian property.