Stitched Trajectories for Off-Policy Learning

Scott Sussex [1] **Omer Gottesman** [1] **Yao Liu** [2] **Susan Murphy** [1] **Emma Brunskill** [2] **Finale Doshi-Velez** [1]

Abstract

We study the problem of off-policy evaluation in RL settings. Importance sampling methods provide unbiased estimators for the value of a policy, without the need to learn an explicit model of the environment, but suffer from high variance, in part due to small effective sample sizes when evaluating deterministic policies. We introduce a method of stitching together sequences from different trajectories to increase the effective sample size for importance sampling estimators, thus reducing their variance while retaining their unbiasedness. We demonstrate that our method reduces the policy value estimation error on several synthetic toy examples.

1. Introduction

Off-policy evaluation (OPE) methods aim to evaluate the performance of an evaluation policy using data collected under a different behavior policy. OPE is desirable in situations where deploying the evaluation policy in practice may be costly, dangerous, or unethical. In the healthcare domain, for example, one might want to estimate the value of a new treatment policy using observational data collected by clinicians (Gottesman et al., 2018) rather than put patients at risk by subjecting them to an untested treatment regime.

Existing approaches to the OPE problem can be classified into two main categories: model-based and importance sampling methods. Model-based methods involve learning an approximate model of the environment and using that model to estimate the value of the evaluation policy (Sutton & Barto, 2017). The main drawback of model-based methods is that their estimates can be highly biased if the modeling assumptions do not hold.

Importance sampling (IS) methods estimate the value of a policy by computing a weighted average of the rewards collected in trajectories observed under the behavior policy. The IS weight of a trajectory is given by the ratio of the probability of the trajectory being observed under the evaluation policy compared to the behavior policy. IS methods are appealing as they can provide unbiased estimates for the value of the evaluation policy, without requiring any knowledge of the environment's dynamics. The downside of IS methods is that their estimators often have very large variance (Thomas & Brunskill, 2016).

Two main reasons lead to the high variance of IS methods. First, if policies that are likely under the evaluation policy are unlikely under the behavior policy, many IS weights will be either very large or very small. Second, for real world applications, we are often interested in estimating the value of a deterministic evaluation policy but collect data using a stochastic behavior policy. This results in the IS weight of a trajectory being equal to zero for all trajectories in which the action taken does not match the action which would have been taken under the evaluation policy, leading to very small effective sample sizes even for large datasets.

We introduce a method of stitching partial trajectories for the setting where we have a deterministic evaluation policy and a stochastic behavior policy. Our method takes trajectories with IS weight 0—that is, trajectories that would have been thrown out—and splits and stitches them into new trajectories. These new trajectories can be combined with the original trajectories with non-zero IS weight to create a new data set; existing IS-based OPE methods can then be applied to generate a value estimate with a greater effective sample size. We prove that, under some assumptions about the underlying Markov decision process, our method of stitched trajectories retains the unbiasedness of ordinary importance sampling and decreases its variance. We also incorporate the method into ordinary, weighted (Precup et al., 2000) and weighted doubly robust (Jiang & Li, 2016) IS-based OPE methods, and show empirically that it reduces mean squared error (MSE) in estimating the value of the evaluation policy in three synthetic domains.

[1]Harvard University, Cambridge, MA, USA [2]Stanford University, Stanford, CA, USA. Correspondence to: Scott Sussex <scottsussex@college.harvard.edu>, Omer Gottesman <gottesman@fas.harvard.edu >, Yao Liu <yaoliu@stanford.edu>, Susan Murphy <samurphy@fas.harvard.edu>, Emma Brunskill <ebrun@cs.stanford.edu>, Finale Doshi-Velez <finale@seas.harvard.edu>.

2. Background and Notation

We give the notation we use for Markov decision processes (MDPs). s_t, a_t, r_t refer to the state, action and reward during a trajectory at time t. In this paper we assume the sets of possible states, actions and rewards are finite. Let $z = (s_0, a_0, r_0, s_1, ...)$ be a trajectory and let $g(z) = \sum_{t=0}^{\infty} \gamma^t r_t$ denote the return of a trajectory. Set $\gamma \in [0, 1)$ to be the rate we discount the future. The evaluation policy is denoted π_e and the behavior policy π_b. We let $V(\pi) := E[g(z)|z \sim \pi]$ where $z \sim \pi$ means that z is sampled using policy π. We are interested in estimating $V(\pi_e)$. The transition probabilities of the MDP are denoted $p(s_{t+1}|a_t, s_t)$. The reward function $r(r_t|s_{t+1}, a_t, s_t)$ gives the probability of receiving reward r_t. We are interested in the setting where both r and the transition probabilities are unknown.

There exist many IS-based methods to estimate the value of a policy given a dataset of trajectories D. Ordinary importance sampling provides an unbiased estimator for the value function, but with large variance that can make the method impractical (Precup et al., 2000). Weighted importance sampling provides a biased estimator that is still consistent and has lower variance (Precup et al., 2000). More recently, doubly robust off-policy evaluation methods provide an unbiased estimator if either the value function or importance ratios are accurately estimated (Jiang & Li, 2016). These estimators have been shown to have lower variance that makes them more practical than alternatives. Doubly robust methods do require an initial estimate for the value function, however this can be obtained from our off-policy data. Weighted doubly robust estimators can lower the variance further at the cost of introducing a small amount of bias (Thomas & Brunskill, 2016).

Define $\rho(z, \pi_e, \pi_b) = p(z|\pi_e)/p(z|\pi_b) = \Pi_{i=0}^{\infty} \pi_e(a_i|s_i)/\pi_b(a_i|s_i)$. Let D be the set of trajectories generated under the behavior policy. Ordinary importance sampling gives an estimate for the value of an evaluation policy $V(\pi_e)$ as $\frac{\sum_{z \in D} \rho(z, \pi_e, \pi_b) g(z)}{|D|}$. For a given trajectory z, the notation $\rho_{i:j}$ refers to $\rho(z_{i:j}, \pi_e, \pi_b)$ where $z_{i:j} = (s_i, a_i, r_i, ...s_j)$.

For the setting where π_e is deterministic and π_b is stochastic, $\rho(z, \pi_e, \pi_b)$ will be 0 if there is an action in the trajectory that π_e would not take. This will result in some number of trajectories having 0 importance weight, especially so if π_e and π_b differ a large amount; these 0-weighted trajectories in turn reduce our effective sample size and increase the variance of our value estimate.

Finally, closer to this work, Fonteneau et al. (2013) also constructs new trajectories from existing data. Their work was focused on value estimation in the batch off-policy case (Fonteneau et al., 2013), but aims to simulate a Monte Carlo estimator rather than use importance sampling.

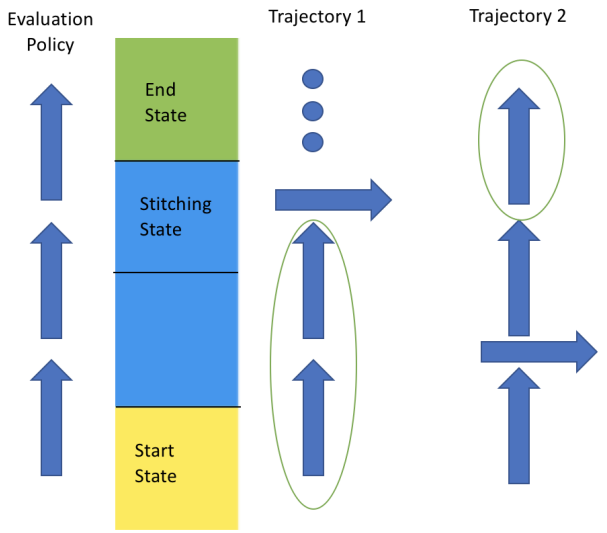

Figure 1. Demonstration of a stitching policy on a deterministic four state corridor gridworld. Trajectory 1 does not follow the evaluation policy after the stitching state. We sample any trajectory that does not follow the evaluation policy before the stitching state- trajectory 2. We then assemble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

3. Stitching trajectories to increase effective sample size

When evaluating a deterministic policy, a trajectory will have non-zero importance sampling weight only if every action in the trajectory is the action which would have been taken under the evaluation policy. For long histories, the probability of finding such a trajectory decreases exponentially with the number of actions. The intuition for our approach is that if we have two trajectories which visited a given state, $s_* \in S_*$, where for the first trajectory all actions up to state s_* agree with the evaluation policy, before not following the evaluation policy at s_*, and for the second trajectory some actions before state s_* differ from the evaluation policy, we can stitch the portion of the first trajectory before s_* to the portion of the second trajectory from s_* onwards to generate a new trajectory which is more likely to have a non-zero IS weight.

For simplicity here we present an algorithm for the case of S_* containing just one state, and provide the full algorithm using multiple stitching states in the appendix.

Algorithm 1 formally details how to stitch together trajectories in our original dataset D to generate a new sample of trajectories, D'. If a trajectory z has nonzero importance weight until reaching a stitching state s_* at time T^{s_*}, but in state s_* takes an action a for which $\pi_e(a|s_*) = 0$, a second

trajectory is found. The second trajectory must have zero importance weight for the section before reaching state s_*. The portion of the first trajectory until reaching state s_* is concatenated to the portion of the second trajectory starting at state s_*, and vice versa, to create two new trajectories—one with possibly non-zero importance weight—that replace the existing two trajectories which both had zero importance weight. Formally, the *concatonate* function in algorithm 1 takes in $(s_0^1, a_0^1, r_0^1, s_1^1, ...s_{t_1}^1, a_{t_1}^1, r_{t_1}^1, ...)$ from some set D_1 and $(s_{t_2}^2, a_{t_2}^2, r_{t_2}^2, ...)$ from some set D_2. It outputs $(s_0^1, a_0^1, r_0^1, s_1^1, ...s_{t_1}^1, a_{t_2}^2, r_{t_2}^2, ...)$.

Input: Stitching state s_*, dataset $D \sim \pi_b$
Output: D'
initialization;
D_1 := [];
D_2 := [];
D' := [];
for $z \in D$ **do**
  if $\rho_{0:T^s_*} > 0$ *and* $\rho_{T^{s_*}:T^{s_*}+1} = 0$ **then**
    add (z, T^{s_*}) to D_1;
  else
    if $\rho_{0:T^{s_*}} = 0$ **then**
      add (z, T^{s_*}) to D_2;
    end
  end
end
for $(z^1,t_1) \in D_1$ **do**
  sample $(z^2,t_2) \in D_2$;
  add $concatenate(z^1_{0:t_1}, z^2_{t_2:})$ to D';
  add $concatenate(z^2_{0:t_2}, z^1_{t_1:})$ to D';
  remove z^2 from D_2;
  remove z^1 and z^2 from D;
end
$D' = D + D'$;

Algorithm 1:

Note that when we sample from D_2, we sample without replacement. Also note that each element of D_1 is sampled at most once. This is to ensure the theoretical variance reduction guarantees proven later. In Algorithm 1, for every trajectory in D_1 we must be able to assign it a trajectory in D_2. Since we sample from D_2 without replacement, we must set the constraint that $|D_1| \leq |D_2|$. This constraint means that under the behavior policy there must be less trajectories that reach the stitching state following the evaluation policy, and then immediately stop following the evaluation policy, than there are trajectories that reach the stitching state by not following the evaluation policy. We must seek to select a stitching state such that this constraint is satisfied.

Using our stitching algorithm denoted A, for a trajectory $z \in D'$ that reaches stitching state $s_* \in S_*$ once, the importance weight for the trajectory is

$$\begin{aligned}
\rho'(z) &= \frac{p(z|\pi_e)}{p(z|\pi_b, A)} \\
&= \left[\frac{\Pi_k \pi_e(a_k|s_k) p(s_{k+1}|s_k, a_k)}{\Pi_k \pi_b(a_k|s_k) p(s_{k+1}|s_k, a_k) + \\ \quad \Pi_k \pi_b(a_k|s_k) p(s_{k+1}|s_k, a_k)(1 - \pi_b(a_*|s_*))} \right] \\
&= \frac{\Pi_k \pi(a_k|s_k) p(s_{k+1}|s_k, a_k)}{(2 - \pi_b(a_*|s_*)) \Pi_k \pi_b(a_k|s_k) p(s_{k+1}|s_k, a_k)} \\
&= \frac{\rho(z)}{2 - \pi_b(a_*|s_*)}
\end{aligned} \tag{1}$$

where $\rho(z)$ is the importance weight that would be calculated for the trajectory if there were no stitching algorithm (so using dataset D), and a_* is the action the evaluation policy takes at s_*. Line two comes from the fact the the trajectory could either have been generated under π_b or consist of two stitched trajectories. If a trajectory z does not pass through the stitching state, it has importance weight $\rho(z)$.

For the case of using our stitching algorithm with multiple stitching states and a trajectory that may go through multiple stitching states, $s_i \in S_*$, where a_i is the action the evaluation policy takes at s_i, in the second line we simply sum over all s_i to get $\rho'(z) = \frac{\rho(z)}{2 - b(a_i|s_i)}$.

We maintain the desirable property of importance sampling in having all transition probabilities cancel in the importance weight. Note that our stitching policy results in a weak increase in trajectories with nonzero importance weights and a weak decrease in the importance weight of all trajectories. These two factors suggest that using importance weights for data subject to our stitching policy will result in reduced variance, whilst still allowing for unbiased estimates to be computed. We prove this below:

Theorem 1 Assume that for any stitching state $s_* \in S_*$, $p(s_{t+1}|s_t, a_t)$ and $p(R|s_{t+1}, s_t, a_t)$ are independent of any information in the trajectory before state s_* was reached. Apply our stitching algorithm to a dataset D, producing a dataset D'. $Var(V(\pi_e)|D') \leq Var(V(\pi_e)|D)$, where $V(\pi_e)$ is the estimate of the value of a deterministic policy π_e using ordinary importance sampling. Proof given in appendix.

The importance weights calculated extend naturally to calculating the importance weight for part of a trajectory, allowing stitched trajectories to be easily applied to per decision variants of importance sampling. Finally, existing model based methods often assume that the trajectories are generated following the Markovian property. The modeling assumption the stitched trajectories proof makes is much

weaker, and could be satisfied by using domain knowledge to select stitching states.

4. Experimental Design

We evaluate our method on a corridor gridworld (length 5) and a 3 by 5 gridworld. In both cases, the gridworld is stochastic and possible actions are all four directions. An action to move in a certain direction has probability 0.8 of moving in the specified direction, and probability 0.1 of moving in each of the two perpendicular directions. The reward is deterministic, moving results in loss of reward, and the future is discounted with $\gamma = 0.95$. The evaluation policy is the optimal deterministic policy. The behavior policy is ϵ-greedy with $\epsilon = 0.5$. Reward is -1 per move, -1 for hitting a wall, and 50 for reaching the goal.

Stitching states are chosen by the experimenter and selected in an attempt to maximize the number of stitching states but have approximately probability one of satisfying the constraints of our stitching policy. In the 5 state corridor gridworld, the agent starts on one side and must reach the opposite side. The penultimate state is the stitching state for the stitched trajectories method. In the 5 by 3 gridworld, three stitching states are selected (see figure 2).

We use a number of off-policy techniques to compute an estimate for the value of the optimal policy. For each number of sampled trajectories, we repeat the experiment 100 times and calculate mean squared error. The methods compared are ordinary importance sampling (IS), stitched trajectories ordinary importance sampling (STIS), weighted importance sampling (WIS), stitched trajectories weighted importance sampling (STWIS), per-decision weighted doubly robust (PDWDR), and stitched trajectories per-decision weighted doubly robust (STPDWDR). We repeat 100 times for 50, 200, 500, 1000 and 2500 trajectories each.

We repeat the experiment above for a cliff gridworld example (see figure 3). The evaluation and behaviour policy have the same specification as above. Stochasticity of the MDP and rewards are the same as above, with -50 reward for entering a pit state.

5. Results

All method's mean estimate quickly converge on the true value for large numbers of trajectories, with no noticeable bias shown by any method. The methods do show clear differences in their variances. The plot of mean squared error (MSE) over the repeats against number of trajectories demonstrates this. Incorporating stitched trajectories into ordinary importance sampling, weighted importance sampling and per decision weighted doubly robust results in a reduction in MSE in all cases.

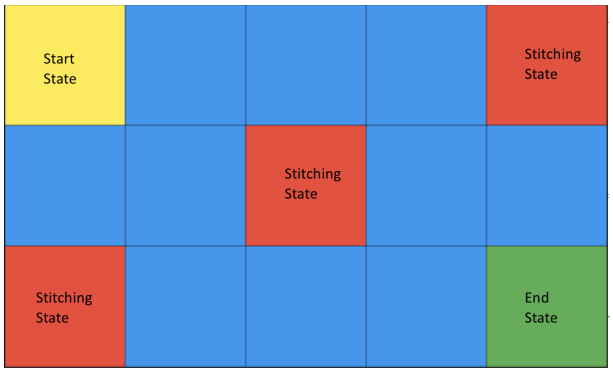

Figure 2. The 3 by 5 gridworld showing stitching states

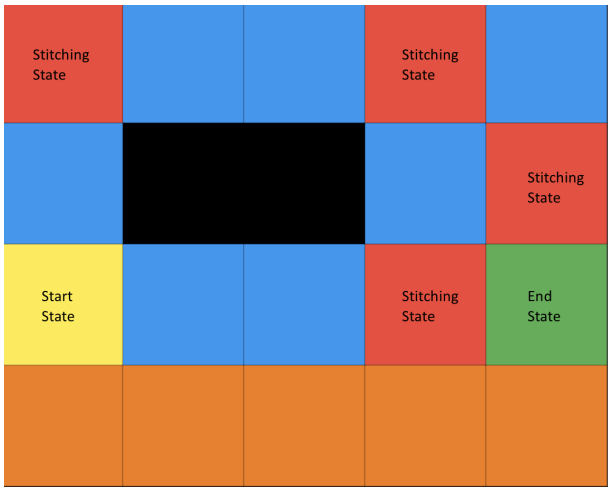

Figure 3. The cliff gridworld showing stitching states. Orange tiles indicate the pit states. Black tiles indicate "walls"- tiles the agent cannot enter.

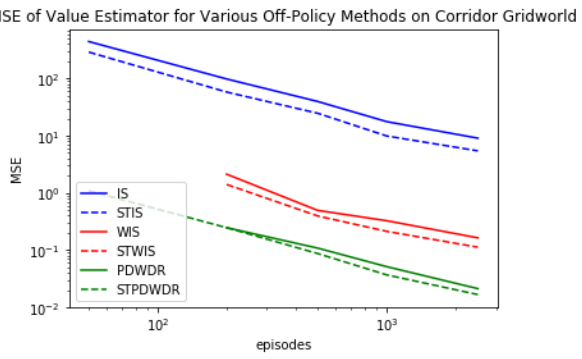

Figure 4. A plot of MSE for various importance sampling methods in a corridor gridworld.

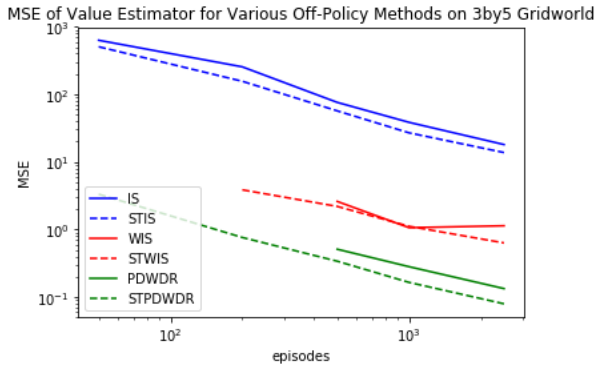

Figure 5. A plot of MSE for various importance sampling methods in a 3 by 5 stochastic gridworld.

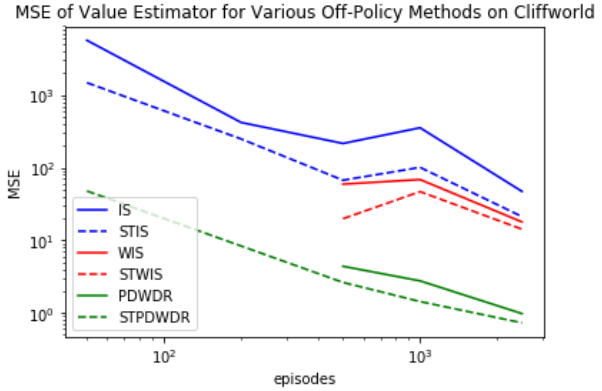

Figure 6. A plot of MSE for various importance sampling methods in a cliff gridworld.

MSE data are not included if a trajectory count number included a repeat where the sampling method could not form an estimate of the policy value, because all importance weights were 0.

6. Conclusion

We have provided two key contributions. First, we gave an algorithm for generating stitched trajectories that results in importance weights which do not depend on MDP transition probabilities. Second, we empirically demonstrated that the use of stitched trajectories for off-policy learning can improve the accuracy of estimators produced from a range of importance sampling estimators. We proved this mathematically for the case of ordinary importance sampling.

We lay the framework for using stitched trajectories in importance sampling, however leave many open questions. More efficient stitching algorithms could be constructed, and theoretical work might compare the variance reductions from each. The work presented here also assumes a discrete state and action space. Future work will include generalizing our framework to the case of continuous states.

Empirical work might evaluate the stitched trajectories method in more complex real world settings, for example on medical records data. A challenge in real-world domains will be the selection of stitching states, and this would be aided by a systematic method for their selection. Finally, the variance reducing properties of stitched trajectories might be combined with the MSE reducing properties of model-based method MAGIC (Thomas & Brunskill, 2016) by including the method of stitched trajectories as the importance sampling method used in MAGIC.

References

Fonteneau, Raphael, Murphy, Susan A., Wehenkel, Louis, and Ernst, Damien. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of Operations Research*, 208(1):383–416, 2013. URL `http://doi.org/10.1007/s10479-012-1248-5`.

Gottesman, Omer, Johansson, Fredrik, Meier, Joshua, Dent, Jack, Lee, Donghun, Srinivasan, Srivatsan, Zhang, Linying, Ding, Yi, Wihl, David, Peng, Xuefeng, Yao, Jiayu, Lage, Isaac, Mosch, Christopher, wei H. Lehman, Li, Komorowski, Matthieu, Komorowski, Matthieu, Faisal, Aldo, Celi, Leo Anthony, Sontag, David, and Doshi-Velez, Finale. Evaluating reinforcement learning algorithms in observational health settings, 2018.

Jiang, Nan and Li, Lihong. Doubly robust off-policy value evaluation for reinforcement learning. In Balcan, Maria Florina and Weinberger, Kilian Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/jiang16.html`.

Precup, Doina, Sutton, Richard S., and Singh, Satinder P. Eligibility traces for off-policy policy evaluation. In *ICML*, pp. 759–766. Morgan Kaufmann, 2000.

Sutton, Richard S. and Barto, Andrew G. *Reinforcement Learning: An Introduction*. Draft Version, Cambridge, MA, USA, 2nd edition, 2017.

Thomas, Philip and Brunskill, Emma. Data-efficient off-policy policy evaluation for reinforcement learning. In Balcan, Maria Florina and Weinberger, Kilian Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148,

New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/thomasa16.html.