Importance Sampling for Fair Policy Selection

Shayan Doroudi
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
shayand@cs.cmu.edu

Philip S. Thomas
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
philipt@cs.cmu.edu

Emma Brunskill
Computer Science Department
Stanford University
Stanford, CA 94305
ebrun@stanford.edu

Abstract

We consider the problem of *off-policy policy selection* in reinforcement learning settings: using historical data generated from running some policy to compare a set of two or more new policies. Policy selection methods can be used, for example, to decide which policy should be deployed next when two or more batch reinforcement learning algorithms suggest different policies or when we want to compare a policy derived from data to a policy constructed by an expert. We show that existing approaches to policy selection based on importance sampling can be *unfair*: they can select the worse of two policies more often than not. We present two illustrative examples to show that this unfairness can adversely impact policy selection scenarios that may arise in practical settings. We then give sufficient conditions for when we can apply existing techniques to do policy selection fairly. Our hope is that this work will lead to more researchers thinking about the problems that arise in off-policy policy selection and how we may mitigate these problems, which we believe has been largely ignored in the literature.

Keywords: policy selection, policy evaluation, importance sampling

Acknowledgements

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130215 and R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Dept. of Education.

1 Introduction

In this paper, we consider the problem of *off-policy policy selection* in reinforcement learning settings: using historical data generated from running some policy to compare a set of two or more new policies. Policy selection methods can be used, for example, to decide which policy should be deployed next when two or more batch reinforcement learning algorithms suggest different policies or when we want to compare a policy derived from data to a policy constructed by an expert. Importance sampling, a technique for predicting the performance of one policy using data generated from running a different policy [4], is at the foundation of many policy selection and policy search algorithms [3, 1, 2, 5, 6]. In this paper, we introduce the notion of fairness for policy selection algorithms, which we believe has not been considered in prior work. In the case of comparing two policies, we say that a policy selection algorithm is *fair* if it selects the better of the two policies more often than it selects the worse of the two policies. The primary contribution of this paper is that we show that standard policy selection algorithms based on importance sampling are often unfair. We illustrate this with two concrete examples of settings that may arise in practice. We then present sufficient conditions for when we can use importance sampling to make fair comparisons, which is a first step towards fair policy selection.

2 Background

2.1 Reinforcement Learning

We consider sequential decision making settings in stochastic domains. In such domains, an agent interacts with the environment, and in doing so, it generates a trajectory, $\tau \triangleq (O_0, A_1, R_1, O_1, A_2, R_2, \ldots, A_T, R_T, O_T)$, which is a sequence of observations, actions, and rewards, with trajectory length T. The observations and rewards are generated by the environment according to a stochastic process that is unknown. The agent chooses actions according to a stochastic policy π, which is a conditional probability distribution over actions A_t given the partial trajectory $\tau_{1:t-1} \triangleq (O_1, A_1, R_1, O_2, A_2, R_2, \ldots, O_{t-1})$ of prior observations, actions, and rewards. The value of a policy π, V^π, is the expected sum of rewards when the policy is used:

$$V^\pi \triangleq \mathbb{E}\left[\sum_{t=1}^{T} R_t \middle| \tau \sim \pi\right]$$

The agent's goal is to find and execute a policy with a large value.

In this paper, we consider offline (batch) reinforcement learning where we have a batch of data, called historical data, that was generated from some known behavior policy π_b. We are interested in doing **batch off-policy policy selection**: identifying a good policy for use in the future based on estimating its performance using the data from π_b. This typically involves policy estimation or evaluation of a policy π_e. If $\pi_e = \pi_b$ this is known as **on-policy policy evaluation**. Otherwise it is known as **off-policy policy evaluation**.

2.2 Importance Sampling

In this paper, we focus on estimators that use importance sampling for off-policy policy selection. Model-based off-policy estimators tend to have lower variance than importance-sampling-based estimators, but at the cost of being biased and asymptotically incorrect (not consistent estimators of V^π) [3]. In contrast, importance sampling-based estimators can provide unbiased estimates of the value of a policy.

Suppose we have a batch of trajectories $\tau_1, \tau_2, \ldots, \tau_n$ sampled independently from executing a behavior policy π_b, but we want to estimate the value of another policy π_e. We can use the **importance sampling (IS) estimator** [4], which is given by:

$$\hat{V}_{\text{IS}}^{\pi_e} \triangleq \frac{1}{n}\sum_{i=1}^{n} w_i \sum_{t=1}^{T_i} R_{i,t}$$

where

$$w_i = \frac{\prod_{t=1}^{T_i} \pi_e(a_{i,t}|\tau_{i,1:t-1})}{\prod_{t=1}^{T_i} \pi_b(a_{i,t}|\tau_{i,1:t-1})}$$

The IS estimator is an unbiased and strongly consistent estimator of V^{π_e} if $\pi_e(a|\tau_{1:t-1}) = 0$ for all actions, a, and partial trajectories, $\tau_{1:t-1}$, where $\pi_b(a_t|\tau_{1:t-1}) = 0$. However, the IS estimator often has very large variance (which is at the root of why it can be unfair for policy selection, as we will show below). The **weighted importance sampling (WIS) estimator** is another estimator where instead of dividing the sum of the IS estimates for each trajectory by the number of trajectories, we divide by the sum of the importance weights as follows:

$$\hat{V}_{\text{WIS}}^{\pi_e} \triangleq \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \sum_{t=1}^{T_i} R_{i,t}$$

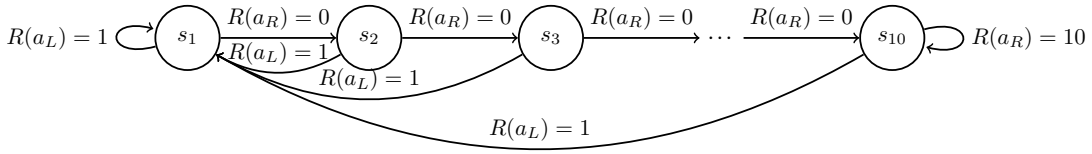

Figure 1: Domain in Section 3.1. The agent is in a chain of length 10. In each state, the agent can either go right (a_R) which progresses the agent along the chain and gives a reward of 0 unless the agent is in s_10, in which case it gives a reward of 10 (and keeps the agent in the s_{10}), or go left (a_L), which takes the agent back to state s_1 and gives a reward of 1.

This estimator has less variance than the importance sampling estimator, but at the expense of adding some bias.

2.3 Policy Selection

A **policy selection** algorithm is any algorithm that takes as input an arbitrary number of policies, and outputs one of those policies. Any estimator used for policy evaluation can be transformed into a policy selection algorithm by simply evaluating each input policy and selecting the one that performs best under the estimator. Because policy evaluation is often used to do policy selection, the problem of policy selection has not been adequately studied independent of policy evaluation to our knowledge, even though it is perhaps the more important of the two problems since policy selection underlies the decision of what policy to use in practice. There are at least two properties that are desirable to have in a policy selection algorithm:

- Consistency: In the limit as the number of trajectories of historical data goes to infinity, the algorithm should always select the policy that has the largest value.
- Fairness: With *any* amount of data, the probability that the algorithm selects a policy with the largest value should be greater than the probability that it selects a policy that does not have the largest value. When choosing between two policies, this implies that the algorithm should choose the better policy at least half the time.

Since model-based approaches to policy evaluation are biased when the model class is inaccurate, they also do not satisfy these properties in general. For example, comparing the estimated value of the optimal policy from a set of models is both inconsistent and unfair (as even in the limit of infinite data, it may *always* pick the wrong policy) [3]. Importance sampling on the other hand, is consistent when used for policy selection as it is an unbiased and consistent estimator of the value function (so in the limit of infinite data, using IS will always lead to choosing the better policy); however we now show that it is not a fair policy selection algorithm.

3 Unfairness of Importance Sampling

We will give two examples that show the unfairness of importance sampling and how they can arise in counterintuitive ways in practically interesting settings, motivating why we should care about satisfying fairness.

3.1 Example 1: Bias Towards Myopic Policies

In this example, we show that using IS for policy selection could be biased in favor of myopic policies, which could be of great practical concern. This may come up in practical settings where we are interested in comparing more heuristic methods of planning (e.g., short look-ahead) to full-horizon planning methods. If we have the correct model class, full horizon planning is expected to be optimal, however it is both computationally expensive (so possibly not even tractable) and potentially sub-optimal if our model class is incorrect (e.g., our state representation is inaccurate or the world is a partially observable Markov decision process but we are modeling it as a fully observable Markov decision process). Thus, we may be interested in comparing full-horizon planning (or an approximation thereof) to myopic planning, and the following example shows that IS can sometimes favor policies resulting from myopic planning.

Consider the domain given in Figure 1. Now suppose we have data collected from a behavior policy π_b that takes each action with probability 0.5 and all trajectories have length 200. We want to compare two policies: π_{myopic} which takes a_L with probability 0.99 and a_R with probability 0.01, and π_{opt} which takes a_L with probability 0.01 and a_R with probability 0.99. (Note: the actual optimal policy is to always take a_R, for which π_{opt} is a slightly stochastic version.) Notice that the probability distribution of importance weights is the same for both π_{myopic} and π_{opt}, so both are equally close to the behavior policy in terms of probabilities over trajectories. However, for datasets that are not large enough, the importance sampling estimate will be larger for π_{myopic} than for π_{opt}, even though it is clearly the worse policy. In particular, when we have 1000 samples, **(1)** around 60% of the time, the importance sampling estimate of π_{myopic} is larger than that of π_{opt},

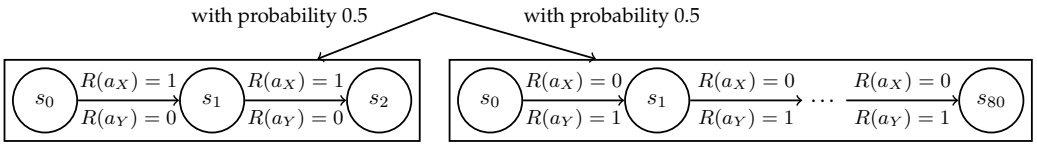

Figure 2: Domain in Section 3.2. The agent is placed uniformly at random in either a chain of length 2 or a chain of length L. At each time step, action a_X deterministically gives a reward of 1 to the agent if the agent is in the chain of length 2 and 0 otherwise, and action a_Y deterministically gives a reward of 1 to the agent if the agent is in the chain of length L and 0 otherwise. Both actions progress the agent along the chain.

	$\hat{V}_{\text{MC}}$	$\hat{V}_{\text{IS}}$	$\hat{V}_{\text{WIS}}$
π_X	1.39	0.98	1.98
π_Y	39.52	0.010	0.020

Table 1: Median estimates, out of 100 simulations, of different estimators using 100 samples of π_X and π_Y in the domain in Section 3.2.

and **(2)** around 95% of the time, the weighted importance sampling estimate of π_{myopic} is larger than that of π_{opt}. Thus both the IS and WIS estimators are unfair for policy selection.

The reason IS is unfair in this case is because one policy only gives high rewards in events that are unlikely under the behavior policy, and hence the behavior policy often does not see the high rewards of this policy as compared to a myopic policy. However, note that these events are still likely enough that we can build a model that would suggest choosing the optimal policy. IS is unable to detect simple patterns that a model-based approach (or even a human briefly looking at the data) would easily infer; this is the cost of having an evaluation technique that places virtually no assumptions on policies.

3.2 Example 2: Systematic Bias Towards Shorter Trajectories

We now show another practically important example where importance sampling can systematically favor policies that assign higher probability to shorter trajectory lengths (in domains where the length of each trajectory may vary). This is a problem that could arise in many practical domains, for example domains where a user is free to leave the system at any time, such as a student doing problems in an educational game or a user chatting with a dialogue system. Moreover, this is especially worrisome when there is some correlation between how long a user stays in the system and the reward that the system obtains. In many cases the reward might be directly proportional to the number of interactions the user has with the system. Even if that is not the case, in many situations, worse policies might bias users to leave the system earlier. For example, in an educational game whose goal is to maximize student learning, we can imagine a policy that gives levels that are too difficult will lead students to leaving the game and hence learning very little, whereas a policy that gives an optimal progression of levels might result in the student to play the game for a longer duration of time and hence learn more. Thus, it is particularly problematic that importance sampling can favor policies that assign higher probability to shorter trajectories even when shorter trajectories are worse than longer ones, which the following example shows to be true.

Consider the domain given in Figure 2. Now suppose we have data collected from a behavior policy π_b that takes each action with probability 0.5. We want to compare two policies: π_X, which takes action a_X with probability 0.99, and π_Y, which takes action a_Y with probability 0.99. Clearly π_Y is the better policy, because it incurs a lot of reward when we encounter trajectories of length 80, while only losing out on a small reward when encountering the short trajectories. Table 1 shows the median estimate, out of 100 simulations, of the Monte Carlo estimator (i.e., the standard on-policy estimator $\hat{V}^{\pi_e}_{\text{MC}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} R_{i,t}$), as well as the median IS and WIS estimates using 1000 samples each. We find that while π_Y is, in actuality, much better, IS essentially only weighs the shorter trajectories, so the estimates only reflect how well the policies do on those trajectories. WIS simply (almost) doubles the estimates because half of the samples have extremely low importance weights. So why does this occur? When using IS in settings where trajectories can have varying lengths, the importance weight of shorter trajectories can be much larger than for longer trajectories, because for longer trajectories, we are multiplying more ratios of probabilities that are more often smaller than one. This happens even if the policy we are evaluating is more likely to produce a longer trajectories than a shorter one (because there are exponentially many longer trajectories and so each individual trajectory has an exponentially smaller weight).

4 Guaranteeing Fairness

We will now show conditions under which we can guarantee fairness when using importance sampling for policy selection.

Theorem 4.1. *Using importance sampling for policy selection when we have n samples from the behavior policy is fair provided that*

$$w^{\pi_1}_{MAX} V^{\pi_1}_{MAX} + w^{\pi_2}_{MAX} V^{\pi_2}_{MAX} \leq |V^{\pi_1} - V^{\pi_2}| \sqrt{\frac{2n}{\ln 2}}$$

where w^{π}_{MAX} is the largest importance weight for policy π and V^{π}_{MAX} is the largest possible value for policy π. In other words, Algorithm 1 is fair provided that $\epsilon \leq |V^{\pi_1} - V^{\pi_2}|$ and $\delta \leq 0.5$.

Theorem 4.1 can be shown with a simple application of Hoeffding's inequality. Alternatively, we can use other concentration inequalities to obtain fair algorithms of a similar form. Additionally, we can extend the algorithms to policy selection with more than two policies by applying a union bound, but that is omitted here for brevity. Notice that Theorem 4.1 tells us that as long as neither policy is too far from the behavior policy in terms of the largest possible importance weight, then we can guarantee fairness, which intuitively makes sense; we can only fairly compare policies that are similar to the behavior policy. However, how far we stray will also depend on how different the values of the policies are from each other. This is a quantity we do not know, so we must pick an ϵ where either we think $\epsilon \geq |V^{\pi_1} - V^{\pi_2}|$ or we are comfortable with the possibility of choosing a policy that is worse than ϵ from the better policy. Theorem 4.1 helps us better understand when we can guarantee fairness and gives hope that importance sampling is still useful for policy selection, but there is still much to do before we can implement fair policy selection to obtain decent policies that are very different from the behavior policy, which is what we would often like in practice. Our hope is that our paper will lead to more researchers thinking about the problems that arise in off-policy policy selection and how we may mitigate these problems, which we believe has been largely ignored in the literature.

Algorithm 1 Fair Policy Selection

Input: $\pi_1, \pi_2, V^{\pi_1}_{MAX}, V^{\pi_2}_{MAX}, \epsilon, \delta$
$\tau_1, \tau_2, \ldots, \tau_n \sim \pi_b$
if $w^{\pi_1}_{MAX} V^{\pi_1}_{MAX} + w^{\pi_2}_{MAX} V^{\pi_2}_{MAX} \leq \epsilon \sqrt{\frac{2n}{\ln 1/\delta}}$ **then**
  return $\max(\hat{V}^{\pi_1}_{IS}, \hat{V}^{\pi_2}_{IS})$
else
  return `No Fair Comparison`
end if

References

[1] Tang Jie and Pieter Abbeel. On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*, pages 1000–1008, 2010.

[2] Sergey Levine and Vladlen Koltun. Guided policy search. In *ICML (3)*, pages 1–9, 2013.

[3] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[4] 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*, pages 759–766, 2000.

[5] P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015.

[6] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.