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@cs.stanford.edu

Abstract

We consider the problem of off-policy policy selection in reinforcement learning: using historical data generated from running one policy to compare two or more policies. We show that approaches based on importance sampling can be *unfair*—they can select the worse of two policies more often than not. We give two examples where the unfairness of importance sampling could be practically concerning. We then present sufficient conditions to theoretically guarantee fairness and a related notion of safety. Finally, we provide a practical importance sampling-based estimator to help mitigate one of the systematic sources of unfairness resulting from using importance sampling for policy selection.

1 INTRODUCTION

In this paper, we consider the problem of *off-policy policy selection*: using historical data generated from running one policy to compare two or more policies. Off-policy policy selection methods can be used, for example, to decide which policy should be deployed when two or more batch reinforcement learning (RL) algorithms suggest different policies or when a data-driven policy is compared to a policy designed by a human expert. The primary contribution of this paper is that we show that the *importance sampling* (IS) estimator (Precup et al., 2000), which lies at the foundation of many policy selection and policy search algorithms (Mandel et al., 2014; Levine and Koltun, 2013; Thomas et al., 2015b), is often *unfair* when used for policy selection: when comparing two policies, the worse of the two policies may be returned more than half the time.

After formalizing our notion of fairness, we show that unfairness can occur in both the on-policy and off-policy settings. We further show that in the off-policy setting, using importance sampling for policy selection can be unfair in practically relevant settings. In particular, we show that IS can favor myopic policies—policies that obtain less total reward, but which obtain it early in an episode—as well as favoring policies that produce shorter trajectories. Although IS is an unbiased estimator, this unfairness arises because policy selection involves taking a maximum over estimated quantities. Depending on the distribution of estimates, importance sampling may systematically favor policies that are worse in expectation.

We then present two new approaches for avoiding unfairness when using importance sampling for policy selection. First, we give sufficient conditions under which using importance sampling for policy selection is fair, and provide algorithms that guarantee a related notion of safety. We also describe how our approach to guaranteeing fairness and safety is related to the notions of power analysis and statistical hypothesis testing. Although of theoretical interest, the reliance of this approach on conservative concentration inequalities limits its practicality. Thus, in our second approach, we introduce a new practical IS-based estimator that lacks the theoretical properties of our first approach, but which can help mitigate unfairness due to differing trajectory lengths.

Although there is significant literature surrounding reducing the variance and mean squared error of off-policy policy evaluation methods that use IS-based estimators (Powell and Swann, 1966; Dudík et al., 2011; Jiang and Li, 2016; Thomas and Brunskill, 2017), other challenges associated with off-policy policy selection, such as fairness, have not been explored in the literature. Similar notions of fairness have recently been proposed in the online RL setting, where an algorithm is fair if it never takes a worse action with higher probability than a better action Jabbari et al. (2017). This is similar to our notion of fairness in the offline RL setting, where an al-

gorithm is fair if it does not choose a worse policy with higher probability than the best candidate policy. However, the issues surrounding fairness in the two settings are different due to the different nature of each setting. By introducing a notion of fairness for policy selection and highlighting some limitations of existing IS-based approaches, we hope to motivate further work on developing practical and fair policy selection algorithms.

2 BACKGROUND

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—such as a Markov decision process (MDP) or partially observable Markov decision process (POMDP)—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 $\tau_{1:t-1} \triangleq (O_0, A_1, R_1, O_1, A_2, R_2, \ldots, O_{t-1})$ 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: $V^\pi \triangleq \mathbf{E}\left[\sum_{t=1}^{T} R_t \middle| \tau \sim \pi\right]$, where $\tau \sim \pi$ means that the actions of τ are sampled according to π. The agent's goal is to find and execute a policy with a large value.

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

2.1 ON-POLICY POLICY EVALUATION

Although we are primarily interested in the off-policy setting, i.e., the setting where $\pi_b \neq \pi_e$, we will also discuss the problem of on-policy policy evaluation. This problem arises, for example, when running a randomized control trial or A/B test to compare two policies. In this case, the value of each policy is directly estimated by running it to generate n trajectories, $\tau_1, \tau_2, \ldots, \tau_n$, and then estimating the policy's performance using the *Monte Carlo* estimator: $\hat{V}^{\pi_e}_{\text{MC},n} \triangleq \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} R_{i,t}$, where $R_{i,t}$ and T_i denote the reward of τ_i at time t and the length of τ_i respectively.

2.2 IMPORTANCE SAMPLING

In this paper, we primarily focus on off-policy policy evaluation and selection. Specifically, we focus on estimators that use importance sampling. Model-based off-policy estimators tend to have lower variance than IS-based estimators, but at the cost of being biased and asymptotically incorrect (not consistent estimators of V^π) (Mandel et al., 2014). In contrast, IS-based estimators can provide unbiased estimates of the value. There has been significant interest in using IS-based techniques in RL for policy evaluation (Precup et al., 2000; Jiang and Li, 2016; Thomas and Brunskill, 2016), as well as growing recent interest in using it for policy selection (Mandel et al., 2014) and the related problems of policy search and policy gradient optimization (Jie and Abbeel, 2010; Levine and Koltun, 2013; Thomas et al., 2015b; Wang et al., 2016).

The IS estimator (Precup et al., 2000) is given by: $\hat{V}^{\pi_e}_{\text{IS}} \triangleq \frac{1}{n}\sum_{i=1}^{n} w_i \sum_{t=1}^{T_i} R_{i,t}$, where $w_i \triangleq \prod_{t=1}^{T_i} \frac{\pi_e(a_{i,t}|\tau_{i,1:t-1})}{\pi_b(a_{i,t}|\tau_{i,1:t-1})}$. 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, $\hat{V}^{\pi_e}_{\text{IS}}$ often has high variance. The *weighted importance sampling* (WIS) estimator, $\hat{V}^{\pi_e}_{\text{WIS}} \triangleq \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \sum_{t=1}^{T_i} R_{i,t}$, is a variant of the IS estimator that often has much lower variance, but which is *not* an unbiased estimator of V^{π_e}.

3 FAIR POLICY SELECTION

A policy selection algorithm is given a set of candidate policies and must choose one of the policies for use in the future. Any policy evaluation algorithm (i.e., estimator) can be converted to a policy selection algorithm by simply evaluating each policy using the estimator, and then selecting the policy that has the largest estimated value. Thus we can use the Monte Carlo estimator for on-policy policy selection, and we can use the IS or WIS estimators for off-policy policy selection.

There are two natural properties we would like in a batch policy selection algorithm:

- **Consistency**: In the limit as the amount of historical data goes to infinity, the algorithm should always select the policy that has the largest value.
- **Fairness**: With *any* amount of historical 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.[1]

Exploring and ensuring the fairness of (IS-based) policy selection algorithms is the focus of this paper. There has been recent interest on combining model-based estimators and IS-based estimators (Jiang and Li, 2016; Thomas and Brunskill, 2016), however as both model-based estimators and (as we will show) IS-based estimators are unfair, it is easy to show that these new estimators must also be unfair. Therefore, we restrict our attention to the standard IS and WIS estimators.

Before proceeding, we formally define a way to rank policies. Let the better-than operator $\succ_B$, be such that $\pi_1 \succ_B \pi_2$ is `True` if π_1 is better than π_2 and `False` otherwise, for some notion of "better". For example, we can define $\succ_V$ to order policies based on their values: $\pi_1 \succ_V \pi_2$ is `True` if $V^{\pi_1} > V^{\pi_2}$ and `False` otherwise. Define the optimal policy π^* in a policy class Π to be the policy where $\pi^* \succ_B \pi'$ for all $\pi' \neq \pi^* \in \Pi$.[2] We will now formally define fairness.

Definition 3.1. A policy selection algorithm that chooses policies from a policy class Π is *fair with respect to a better-than operator* $\succ_B$ if whenever the algorithm outputs a policy, the probability that it outputs π^* is at least as large as the probability that it will output any other policy. The algorithm is strictly fair if the probability of outputting policy π^* is strictly greater than the probability of outputting any other policy.

Notice that the probabilistic guarantee in this definition conditions on when the algorithm outputs a policy. This allows for a policy selection algorithm that does not output any policy in cases when it cannot determine which policy is better. Also, notice that the trivial policy selection algorithm that never outputs a policy is fair. However, ideally we want a policy selection algorithm that outputs a policy as often as possible while maintaining fairness. This is an important distinction: although we *want* an algorithm that often outputs a policy, we *require* the algorithm to at least be fair. We now see that this seemingly straightforward property is not satisfied by even the most natural policy selection algorithms.

We begin by showing that even Monte Carlo estimation is unfair when used for on-policy policy selection. Suppose we want to select the better of two policies, π_1 and π_2, in a multi-armed bandit (MAB) domain with three actions a_1, a_2, and a_3 with rewards and probabilities as described in Table 1. Notice that $V^{\pi_1} = r$ and $V^{\pi_2} = p$. So, if $r < p$, then $V^{\pi_1} < V^{\pi_2}$. However, notice that using one trajectory ($n = 1$), the Monte Carlo estimator is unfair with respect to $\succ_V$ if $r < p < 0.5$ since $\Pr(\hat{V}^{\pi_1}_{\text{MC},1} \leq \hat{V}^{\pi_2}_{\text{MC},1}) = p$. We can similarly show that Monte Carlo policy selection is unfair using n trajectories for $n > 1$, as long as r and p are sufficiently small.

Table 1: The probability of each action under π_1 and π_2 for the example domain where Monte Carlo estimation is unfair. Rewards, R, are deterministic.

	$a_1(R=0)$	$a_2(R=r)$	$a_3(R=1)$
π_1	0	1	0
π_2	$1-p$	0	p

4 UNFAIRNESS OF IMPORTANCE SAMPLING POLICY SELECTION

Unsurprisingly, importance sampling is also unfair with respect to $\succ_V$. However, the unfairness of importance sampling can be arbitrarily worse than the unfairness of the Monte Carlo estimator, in that for any n, we can construct a domain such that IS policy selection is unfair even though the Monte Carlo estimator will always pick the correct policy with even a single sample! We provide one such example in Supplementary Material A. In this section, we present two examples that highlight how the unfairness of importance sampling can arise in counter-intuitive ways in practically interesting settings, motivating the importance of caring about satisfying fairness.

4.1 IS FAVORS MYOPIC POLICIES

In the following example we show that even when comparing two policies that are equally close to the behavior policy, importance sampling can still be unfair. In particular, we show that using IS for policy selection could be biased in favor of myopic policies, which could be of significant 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 POMDP but we are modeling it as a MDP). Thus, we may be interested in comparing full-horizon planning (or an approximation thereof) to myopic planning, and the following exam-

[1] For simplicity, hereafter we assume that there are no two candidate policies that are equally good.

[2] We assume such a best policy exists; however, for some reasonable better-than operators, this may not be true, as we explore in Section 5.1.

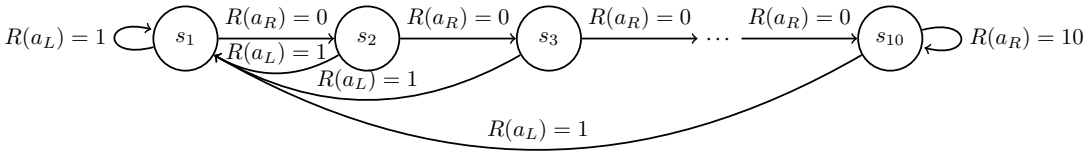

Figure 1: Domain in Section 4.1. The agent is in a chain of length 10. In each state, the agent can either go right (a_R) which progresses the agent along the chain and gives a reward of 0 unless the agent is in s_{10}, in which case it gives a reward of 10 (and keeps the agent in the s_{10}), or go left (a_L), which takes the agent back to state s_1 and gives a reward of 1.

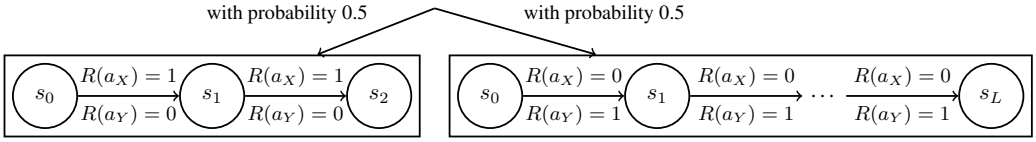

Figure 2: Domain in Section 4.2. The agent is placed uniformly at random in either a chain MDP of length 2 or a chain of length L. At each time step, action a_X deterministically gives a reward of 1 to the agent if the agent is in the chain of length 2 and 0 otherwise, and action a_Y deterministically 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.

ple shows that IS can sometimes favor policies resulting from myopic planning even when full horizon planning is optimal.

Consider the MDP 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 probability distributions 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. For example, 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}, 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.

4.2 IS FAVORS SHORTER TRAJECTORIES

Importance sampling can also 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 solving problems in an educational game or a user chatting with a dialogue system. In such systems, a bad policy may cause a user to leave the system sooner resulting in a short trajectory, which makes it particularly problematic that importance sampling can favor policies that assign higher probability to shorter trajectories. The following example shows that importance sampling can favor policies that generate shorter trajectories even when they are clearly worse.

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. Consider the case where $L = 80$. Clearly π_Y is the better policy, because it incurs a lot of reward when we encounter trajectories of length 80, while only los-

Table 2: Median estimates, out of 100 simulations, of different estimators using 100 samples of π_X and π_Y in the domain in Section 4.2.

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

ing out on a small reward when encountering the short trajectories. Table 2 shows the median estimate, out of 100 simulations, of the Monte Carlo estimator, 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 longer trajectories than a shorter one (because there are exponentially many longer trajectories and so each individual trajectory has an exponentially smaller weight than an individual short trajectory).

5 A NEW KIND OF FAIRNESS

The examples above illustrate that even the most straightforward way of evaluating policies could misguide someone about which policy is actually better under the objective of maximizing the expected sum of reward. Although we showed that off-policy policy selection using IS can be much less fair than doing on-policy Monte Carlo estimation, the fact that on-policy estimation is also unfair suggests that perhaps we are using the wrong measure of performance to construct our better-than operators. The problem is that, even if π_1 usually produces trajectories with more reward than those produced by π_2, it could still be that $V^{\pi_1} < V^{\pi_2}$ if there is a rare trajectory with very large reward that is more likely under π_2. It is therefore worth considering a different notion of "better-than" that better captures which policy is likely to perform better given a fixed and finite amount of historical data. This can be captured using the following better-than operator, which we refer to as better-than with respect to Monte Carlo estimation: $\pi_1 \succ_{\text{MC},n} \pi_2$ is `True` if $\Pr(\hat{V}^{\pi_1}_{\text{MC},n} > \hat{V}^{\pi_2}_{\text{MC},n}) \geq \Pr(\hat{V}^{\pi_1}_{\text{MC},n} < \hat{V}^{\pi_2}_{\text{MC},n})$, and `False` otherwise.

Notice that this notion of better-than is also exactly what we would want to use if we want to find a policy where we optimize our chance of beating a baseline in an experiment or A/B test with n samples (i.e., obtaining a statistically significant result). Also, notice that we have a different better-than operator for each n, which is not necessarily related to the number of samples that we have from our behavior policy π_b. (For example, we may have collected data from interactions with 1,000 users, but want to deploy a policy and expect one million users to use it; in that case we may want to be fair with respect to $\succ_{\text{MC},10^6}$.) We will use $\succ_{\text{MC}}$ to refer to this notion of better-than in general (even though it is not technically a better-than operator without specifying the number of samples). Of course, for this new better-than operator, the Monte Carlo estimator is fair by definition. It is interesting to note that if the distribution over the sum of rewards under the two policies are symmetric distributions (e.g., normal distributions), then the two better-than operators are equivalent. It may seem as though it is easier to satisfy fairness for importance sampling with respect to $\succ_{\text{MC},n}$ than with respect to $\succ_V$; however, we will show below that, at least in some sense, this is not the case. However, we can present conditions where both are satisfied simultaneously. We will henceforth consider and present results with respect to both notions of fairness (i.e., fairness with respect to $\succ_V$ and fairness with respect to $\succ_{\text{MC},n}$). But we will first consider some counterintuitive properties of this new better-than operator.

5.1 FAIRNESS AND NON-TRANSITIVITY

It is a well-known result in probability theory that there exists random variables X, Y, and Z such that $\Pr(X > Y) > 0.5$ and $\Pr(Y > Z) > 0.5$, but $\Pr(Z > X) > 0.5$ as well, indicating that comparing between random variables in such a way is non-transitive (Trybuła, 1961; Gardner, 1970). We claim that this non-transitivity holds for the ordering induced by the $\succ_{\text{MC},n}$ operator as well. This is shown in Supplementary Material B. In fact, it is possible that for any policy π, there is another policy π' where $\succ_{\text{MC},n}(\pi', \pi) = \texttt{True}$, meaning a best policy might not exist with respect to $\succ_{\text{MC}}$.

The good news is that over fifty years ago, Trybuła (1965) showed that for any m independent random variables $X_1, \ldots, X_m$, $\min\{\Pr(X_1 > X_2), \ldots \Pr(X_{m-1} > X_m), \Pr(X_m > X_1)\} < 0.75$. Thus, in order to avoid such non-transitivity, we motivate a new kind of fairness:

Definition 5.1. An algorithm is *transitively fair* with respect to a better-than operator $\succ_B$ if it is fair with respect to $\succ_B$ and if the algorithm does not output a policy when comparing two policies from a set of policies π_1, $\ldots$, π_k such that $\pi_i \succ_B \pi_{i+1}$ for $i \in \{1, \ldots, k-1\}$ and

$\pi_k \succ_B \pi_1$.

Clearly any algorithm that is fair with respect to $\succ_V$ is also transitively fair, as comparing real numbers is a transitive operation. This is not the case for any algorithm that is fair with respect to $\succ_{\text{MC},n}$, but one way to ensure transitive fairness in this case is to not only be fair with respect to $\succ_{\text{MC},n}$, but to also not output a policy unless $\Pr(\hat{V}^{\pi_1}_{\text{MC},n} > \hat{V}^{\pi_2}_{\text{MC},n}) \geq 0.75$.

6 GUARANTEEING FAIRNESS

Given that importance sampling is not fair in general, we would like to understand under what conditions we can guarantee importance sampling can be used to do fair policy selection. Recall that even Monte Carlo estimation is not fair, so we would also like to give conditions for which we can guarantee on-policy policy selection can be done fairly. Thus, we are interested in guaranteeing the following four notions of fairness: **(1)** fairness with respect to $\succ_V$ when we have samples from each policy, **(2)** fairness with respect to $\succ_{\text{MC},n}$ when we have samples from each policy, **(3)** fairness with respect to $\succ_V$ when we have samples from a behavior policy, and **(4)** fairness with respect to $\succ_{\text{MC},n}$ when we have samples from a behavior policy.

Notice that case (2) is satisfied whenever we use Monte Carlo estimation for policy selection by definition of $\succ_{\text{MC}}$. We now give theorems describing the conditions under which we can satisfy the remaining three cases. Let V^{π}_{Max} be the largest value that could result from policy π and w^{π}_{Max} be the largest importance weight possible for policy π with samples drawn from behavior policy π_b. In what follows, for simplicity, we assume that the minimum possible value of a trajectory for all policies is 0.[3] Our results can be extended in the case that this is not true by considering the minimum possible value for each policy. Theorem 6.1 gives the conditions for which using the on-policy Monte Carlo estimator is fair for policy selection, when comparing between two policies.

Theorem 6.1. *Using the on-policy Monte Carlo estimator for policy selection when we have n samples from each of policies π_1 and π_2 is fair provided that $V^{\pi_1}_{Max} + V^{\pi_2}_{Max} \leq |V^{\pi_1} - V^{\pi_2}|\sqrt{\frac{2n}{\ln 2}}$ We can guarantee strict fairness if the inequality above is strict.*

Similarly, Theorem 6.2 gives conditions for which using the importance sampling estimator for policy selection is guaranteed to be fair with respect to $\succ_V$. Algorithm 1 is a fair policy selection algorithm that guarantees fairness whenever the condition in Theorem 6.2 is met, and otherwise returns `No Fair Comparison`, provided that $\epsilon \geq |V^{\pi_1} - V^{\pi_2}|$ and $\delta \geq 0.5$. Setting $\delta = 0.5$ is sufficient to guarantee fairness, but we can guarantee stronger notions of fairness by choosing some $\delta > 0.5$ (i.e., whenever the algorithm outputs a policy, it outputs the better policy with probability at least $\delta > 0.5$). While we only consider comparing between two policies in this section, we can extend our results to the case where we select a policy from a class of $n \geq 2$ policies, as we show in Supplementary Material D.

Theorem 6.2. *Using importance sampling for policy selection when we have n samples from the behavior policy is fair with respect to $\succ_V$, provided that*

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

We can guarantee strict fairness if the inequality above is strict.

Theorems 6.1 and 6.2 can both be shown with a simple application of Hoeffding's inequality; the proofs are given in Supplementary Material C. Alternatively, we can use other concentration inequalities to obtain fairness conditions/algorithms of a similar form. Notice that Theorem 6.2 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 selecting a policy whose value is ϵ worse than that of the better policy. Thus ϵ can be thought of as a hypothetical effect size as would be encountered in hypothesis testing. To make the analogue with hypothesis testing more clear, notice that if we fix the policies that we want to compare, we can instead convert Theorems 6.1 and 6.2 to give lower bounds on n that guarantee fairness; that is, we can ask how many samples do we need before we can fairly compare between two specific policies. This is analogous to doing a power analysis in the hypothesis testing literature. A critical difference is that in hypothesis testing we are typically interested in minimizing the probability of a bad event, whereas here we are ensuring that the better of the two policies is chosen more often. Furthermore, in the off-policy case, we are testing counterfactual hypotheses—hypotheses that we never run in the real-world.

Notice that Theorem 6.2 is satisfied for a much smaller subset of policies than Theorem 6.1 as $w^{\pi_1}_{\text{Max}}$ and $w^{\pi_2}_{\text{Max}}$

[3] In the off-policy case, all our results also hold under the very mild assumption that for each policy we evaluate there is some trajectory that has non-zero probability under π_b but has 0 probability under the evaluation policy.

can be huge (exponential in the trajectory length). This is not surprising and matches our intuition (as seen in the examples above) that IS can be very unfair even in cases where on-policy selection is fair.

Algorithm 1 Off-Policy FPS-$\succ_V$

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

In order to satisfy fairness with respect to $\succ_{MC,n}$, the condition on the difference between the two policies will be in terms of the difference between typical Monte Carlo estimates of the value, as shown in Theorem 6.3. We can also satisfy transitive fairness with a stricter assumption on the difference between the two policies.

Theorem 6.3. *For any two policies π_1 and π_2, behavior policy π_b, and for all $k \in \{1, 2, 3, \ldots\}$, using importance sampling for policy selection when we have n samples from the behavior policy is fair with respect to $\succ_{MC,kn}$ provided that there exists $\epsilon > 0$ and $\delta < 0.5$ such that $\Pr\left(|\hat{V}_{MC,kn}^{\pi_1} - \hat{V}_{MC,kn}^{\pi_2}| \geq \epsilon\right) \geq 1 - \delta$ and $(w_{Max}^{\pi_1}+1)V_{Max}^{\pi_1} + (w_{Max}^{\pi_2}+1)V_{Max}^{\pi_2} \leq \epsilon \sqrt{\frac{2n}{\ln \frac{1-\delta}{0.5-\delta}}}$. Importance sampling in this setting is transitively fair, provided that $\delta \leq 0.25$.*

The theorem essentially says that as long as the conditions hold, we can guarantee fairness with respect to $\succ_{\text{MC},m}$ for any m that is a multiple of n (which is slightly weaker than being able to satisfy fairness for all $m \geq n$). Notice that if we compare Theorem 6.2 with Theorem 6.3 for the same choice of ϵ and for any choice of $\delta < 0.5$ (in Theorem 6.3), then provided that the conditions on the difference between the two policies for both theorems hold, the latter is satisfied for a strict subset of policies. Thus, we can use Theorem 6.3 to simultaneously guarantee fairness with respect to $\succ_V$ and with respect to $\succ_{\text{MC},m}$ (provided that the policy effect size conditions of both theorems hold). It might seem strange that satisfying $\succ_{\text{MC},n}$ requires a stricter test than for satisfying $\succ_V$, as it might seem as though importance sampling would be more likely to choose the same policy as the Monte Carlo estimator than the policy that has a higher value (when they are not the same). However, this is not necessarily the case, as the policy that importance sampling chooses will depend on the behavior policy.

6.1 SAFETY

As mentioned above, our conditions on fairness require us to have a reasonable estimate of an effect size between the policies we want to compare. It would be worthwhile to have a guarantee that does not require us to speculate about this unknown quantity. In this section, we give another type of guarantee which we refer to as *safety*. Safety has been considered as a property for policy evaluation and policy improvement algorithms (Thomas et al., 2015a,b). Here we extend the property to apply to policy selection algorithms.

Definition 6.1. A policy selection algorithm that chooses policies from a policy class Π is *safe with probability $1 - \delta$ with respect to a better-than operator $\succ_B$* if whenever it outputs any policy π such that there exists another policy $\tilde{\pi} \in \Pi$ where $\tilde{\pi} \not\succ_B \pi$, it does so with probability at most δ for some $\delta \leq 0.5$.

Thus, a safe policy selection algorithm does not output a sub-optimal policy often; however, it is still possible for a safe policy selection algorithm to output a sub-optimal policy more often than the best policy—but in that case, the algorithm won't output any policy often. This definition is weaker than fairness, but as we will see, we can satisfy it without requiring knowledge about an effect size between the policies. Theorem 6.4 gives the conditions for a safe policy selection algorithm with respect to $\succ_V$ and Theorem 6.5 gives analogous conditions for a safe policy selection algorithm with respect to $\succ_{\text{MC}}$, both using Algorithm 2 as the underlying algorithm. The proofs that these algorithms guarantee safety are given in Supplementary Material E.

Algorithm 2 Off-Policy SPS

input $\pi_1, \pi_2, \omega, p,$
$\tau_1, \tau_2, \ldots, \tau_n \sim \pi_b$
$\beta \leftarrow \omega \sqrt{\frac{\ln(2/(1-p))}{2n}}$
if $\hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} - \beta > 0$ **then**
return π_1
else if $\hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} + \beta < 0$ **then**
return π_2
else
return `No Fair Comparison`
end if

Theorem 6.4. *For any two policies π_1 and π_2, behavior policy π_b, $\omega = w_{Max}^{\pi_1} V_{Max}^{\pi_1} + w_{Max}^{\pi_2} V_{Max}^{\pi_2}$, and $\delta \leq 0.5$, Algorithm 2 is a safe policy selection algorithm with respect to $\succ_V$ with probability $1 - \delta$.*

Theorem 6.5. *For any two policies π_1 and π_2, where $\Pr\left(|\hat{V}_{MC,kn}^{\pi_1} - \hat{V}_{MC,kn}^{\pi_2}| \geq 0\right) \geq 1 - \delta_{MC}$ any behavior policy π_b, $\omega = (w_{Max}^{\pi_1} + 1)V_{Max}^{\pi_1} + (w_{Max}^{\pi_2} + 1)V_{Max}^{\pi_2}$, $p =$*

$1-\delta_{MC}\delta$ *for some* $\delta \leq 0.5$*, and for all* $k \in \{1, 2, 3, \dots\}$*, Algorithm 2 is a safe policy selection algorithm with respect to* $\succ_{MC,kn}$ *with probability* $1-\delta$ *when we have* n *samples drawn from* π_b*. It is a transitively fair policy selection algorithm whenever* $\delta_{MC} \leq 0.25$*.*

Note that these algorithms are analogous to statistical hypothesis testing in that we compare the lower bound of our estimate of the value of one policy with the upper bound of our estimate of the value of another policy. This analogue is similar to how our fair policy selection algorithms shared much in common with doing power analyses for hypothesis testing. Also note that as with fairness, Theorem 6.5 ensures safety with respect to both better-than operators, so if one uses Algorithm 2 with the inputs as described in Theorem 6.5, one does not have to determine which better-than operator one is using. Again, we find that safety with respect to $\succ_{\text{MC}}$ is more difficult to satisfy than safety with respect to $\succ_V$. We can formalize this with the following theorem.

Theorem 6.6. *There exists policies* π_1*,* π_2*, and behavior policy* π_b *for which Algorithm 2 with inputs as described in Theorem 6.4 is not a safe policy selection algorithm with respect to* $\succ_{MC,1}$ *with* $p = 0.5$ *when we have a single sample drawn from* π_b*.*

7 PRACTICAL FAIRNESS: VARYING TRAJECTORY LENGTHS

While the algorithms above provide a way to guarantee fairness, the concentration inequalities we used are naturally quite loose in most cases, and would likely result in returning `No Fair Comparison` in many cases. Practically, it would be desirable to have algorithms that can provide fair comparisons more often. As a first step in this direction, here we discuss a heuristic approach to policy selection for domains where we have varying trajectory lengths, as seen in Section 4.2. The reason for focusing on this particular aspect of unfairness is because it is systematic (potentially arising in any domain where trajectories vary in length), yet it seems like there should be a way to correct for the systematic preference towards shorter trajectories in practically relevant domains. The idea we propose here is to compute an IS-based estimate for each trajectory length individually and then recombine the estimates to get a new estimate. We propose using the following estimator, which we refer to as the *Per-Horizon Weighted Importance Sampling* (PHWIS) estimator:

$$\hat{V}_{\text{PHWIS}} = \sum_{l \in L} W_l \underbrace{\frac{1}{\sum_{\{\tau_i | T_i = l\}} w_i} \sum_{\{\tau_i | T_i = l\}} w_i \sum_{t=1}^{T_i} R_{i,t}}_{\text{WIS estimate on } l\text{-length trajectories}}$$

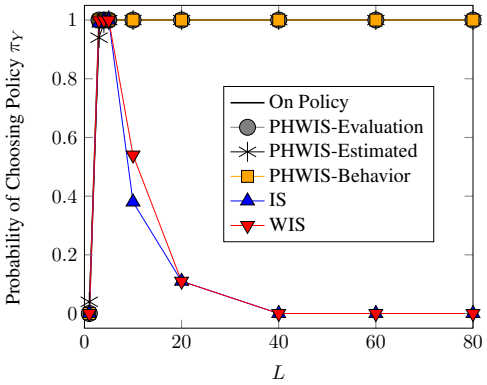

Figure 3: Probability of various estimators choosing π_Y over π_X for different values of L in the domain given in Section 4.2 with 1000 trajectories drawn from the uniform random behavior policy. For each estimator, the probability of outputting π_Y was estimated using 100 independent estimates.

where L is the set of trajectory lengths that appear in the data and W_l is a weight for the relative importance of each trajectory length.

Notice that in the domain in Section 4.2, the length of the trajectories did not depend on the policy that was used to generate them; in such cases, we can use the following weights: W_l (Behavior) $\triangleq \frac{|\{\tau_i | T_i = l\}|}{n}$. The weights simply count the proportion of trajectories in our data (i.e., generated by the behavior policy) that have length l. We will refer to PHWIS with this weighting scheme as PHWIS-Behavior. Now we will see how this estimator performs on the domain in Section 4.2 (Figure 2) where $L \in \{1, 3, 5, 10, 20, 40, 60, 80\}$ given 1000 trajectories from the uniform random behavior policy. Figure 3 shows that while IS and WIS are unfair (choose π_X more often than π_Y) when the long trajectories are of length 20, PHWIS-Behavior always chooses the policy that the on-policy Monte Carlo estimator would choose (i.e., π_X when $L = 1$, and π_Y otherwise).

However, note that in cases where different policies may generate trajectories of different lengths (for example, bad policies causing users to dropout sooner), this simple form of weighting might not work too well. Ideally, we would like to use the following weights: W_l (Evaluation) $\triangleq \Pr(\mathbb{1}(|\tau| = l | \tau \sim \pi_e))$, where π_e is the evaluation policy for which we would like to estimate $\hat{V}^{\pi_e}$. We cannot actually compute these weights because we do not know the probability of any trajectory being generated by the evaluation policy, but when we have ground truth, we can use the PHWIS-Evaluation estimator as a point of comparison. To approximate these weights, we can take the weighted importance sampling estimate of the trajectory lengths generated by the behavior policy: W_l (WIS) $\triangleq \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \mathbb{1}(T_i = l)$

(i.e., rather than reweighing the rewards of all the trajectories generated by the behavior policy, we reweigh the probability that trajectories of length l are generated by the behavior policy). While this is a seemingly reasonable thing to do, this estimate will still suffer from high variance for the same reason the WIS estimator does, which would again lead to assigning small weights for longer trajectories. Instead we propose a heuristic way to make weights for different trajectory lengths have comparable magnitudes: W_l (Estimated) $\triangleq \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i^{1/T_i} \mathbb{1}(T_i = l)$. The idea behind these weights is that they should give us a sense of which weights are preferred by the evaluation policy while maintaining that weights of different trajectory lengths have comparable magnitudes. As we see in Figure 3, PHWIS-Estimated has almost identical performance to PHWIS-Behavior and PHWIS-Evaluation (with just a small probability of choosing the wrong policy when both trajectory lengths are short). However, the true value of using such an estimator comes in domains where the length of a trajectory depends on the policy, and so PHWIS-Behavior may not be sufficient. We now examine one such domain.

7.1 POLICY-DEPENDENT TRAJECTORY LENGTHS

Consider a MDP that has three states—s_0, s_1, and s_2—and two actions—X and Y. The agent starts in s_0 and, on the first time step, if the agent takes action X, they deterministically transition to s_1 and receives a reward of r, and if the agent takes action Y, they transition deterministically to s_2 and receive a reward of 1. Thereafter the agent will always remain in the same state until the trajectory ends and will receive a reward of r whenever it takes action X in state s_1, a reward of 1 whenever it takes action Y in state s_2, and a reward of 0 otherwise. If the first action was X, the trajectory will end with probability 0.05 after each action, and if the first action was Y, the trajectory will end with probability 0.01 after each action. Thus, taking action Y in the beginning will result in a trajectory that is five times as long in expectation.

The behavior policy π_b takes each action with probability 0.5. Again, we want to compare two policies: π_X, which takes action X with probability 0.99, and π_Y, which takes action Y with probability 0.99. Whether π_X or π_Y is better will depend on r. The policies have the same value when $r = 5$, because the difference in rewards offsets the difference in lengths.

Figure 4 shows which policy is chosen when using different estimators for the domain described above with various values for r. We find that the IS, WIS, and even PHWIS-Behavior estimators are unfair regardless of the value of r (except for PHWIS-Behavior when $r = 1$). PHWIS-Evaluation tracks the on-policy estimator reasonably well, only sometimes choosing the wrong policy in the $r = 4$ case. PHWIS-Estimated similarly tracks the on-policy estimator reasonably well, but it sometimes chooses the wrong policy in the $r = 3$ case and is unfair when $r = 4$. Thus, PHWIS-Estimated seems to be a reasonable policy selection estimator to use in such domains, and can be much better than using PHWIS-Behavior in some cases.

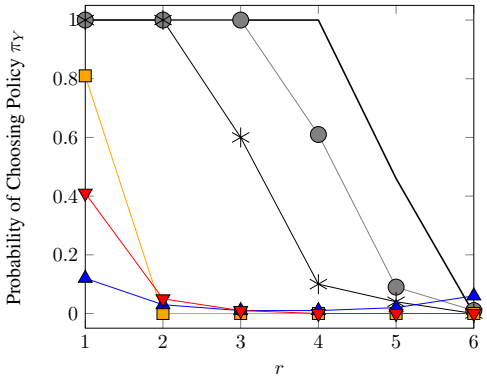

Figure 4: Probability of various estimators choosing π_Y over π_X for different values of r in the domain given in Section 7.1.

8 CONCLUSION

In this paper, we examined the problem of off-policy policy selection and introduced a new property for policy selection algorithms called fairness. We showed that importance sampling is unfair when used for policy selection even though it is an unbiased estimator for policy evaluation. We presented two approaches to deal with this issue, a theoretical solution and a new practical estimator. This is but a first step in tackling the issue of fairness in off-policy policy selection. Our hope is that our introduction of the notion of fairness for policy selection will result in growing interest on the challenges involved in doing off-policy policy selection, including how unfairness propagates to policy search methods that optimize over an infinite class of policies.

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. The research was also supported in part by a NSF CAREER grant.

References

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

M. Gardner. Paradox of nontransitive dice and elusive principle of indifference. *Scientific American*, 223(6): 110, 1970.

S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in reinforcement learning. In *International Conference on Machine Learning*, 2017.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661, 2016.

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

S. Levine and V. Koltun. Guided policy search. In *International Conference on Machine Learning*, pages 1–9, 2013.

T. Mandel, Y.-E. Liu, S. Levine, E. Brunskill, and Z. Popovic. Offline policy evaluation across representations with applications to educational games. In *Internatonal Conference on Autonomous Agents and Multi-Agent Systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

M. Powell and J. Swann. Weighted uniform sampling—a Monte Carlo technique for reducing variance. *IMA Journal of Applied Mathematics*, 2(3):228–236, 1966.

D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*. Citeseer, 2000.

P. S. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

P. S. Thomas and E. Brunskill. Importance sampling with unequal support. In *AAAI*, pages 2646–2652, 2017.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, pages 3000–3006, 2015a.

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

S. Trybuła. On the paradox of three random variables. *Applicationes Mathematicae*, 4(5):321–332, 1961.

S. Trybuła. On the paradox of n random variables. *Applicationes Mathematicae*, 8(2):143–156, 1965.

Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

SUPPLEMENTARY MATERIAL

A UNFAIRNESS OF IMPORTANCE SAMPLING

Suppose we want to use importance sampling to select the better of two policies, π_e and π_b, where we have prior data collected from π_b, in a MAB with two actions a_1 and a_2, with rewards and probabilities as described in Table 3. Notice that $V^{\pi_b} = p + (1-p)r$ and $V^{\pi_e} = 1$, so a fair policy selection algorithm should choose π_e at least half the time since $p + (1-p)r < 1$. If we draw only a single sample from π_b, we get that with probability $1-p$, IS would select π_b over π_e. Thus as long as $p < 0.5$, IS will be unfair. Furthermore, notice that as we decrease p, the gap between the performance of the policies increases, yet the probability that IS chooses the right policy only decreases!

Now suppose we draw n samples from π_b. Notice that as long as τ_2 is never sampled, IS will choose π_b, since in that case $\hat{V}_{\text{IS}}^{\pi_3} = 0$. π_b will never sample τ_2 with probability $(1-p)^n$. Thus IS is unfair as long as $(1-p)^n \geq 0.5$, or as long as $p \leq 1 - 0.5^{1/n} \approx \ln(2)/n$ for large n.

It may appear that this unfairness is not a big problem when we have a reasonable number of samples, but the practical significance of this problem becomes more pronounced in more realistic domains where we have a large number of possible trajectories, or equivalently, a long horizon. For example consider a domain where there are only two actions and the agent must take 50 sequential actions and receives a reward only at the end of a trajectory. Furthermore, consider that the only valuable trajectory is to take a particular action for the entire trajectory (analogous to a_2 above). In this case, we would need over 10^{14} samples just to get IS to be fair!

Table 3: Domain in Supplementary Material A. Rewards are deterministic. The bottom two rows give the probability distributions for π_e and π_b over the two actions.

	a_1 $(R = r < 1)$	a_2 $(R = 1)$
π_e	0	1
π_b	$1-p$	p

B NON-TRANSITIVITY

Theorem B.1 (Non-Transitivity of $\succ_{\text{MC},n}$)**.** *The relation induced by $\succ_{MC,n}$ is non-transitive. Specifically, there exists policies π_1, π_2, and π_3 where* $\Pr(\hat{V}_{MC,n}^{\pi_1} > \hat{V}_{MC,n}^{\pi_2}) = \Pr(\hat{V}_{MC,n}^{\pi_2} > \hat{V}_{MC,n}^{\pi_3}) = \Pr(\hat{V}_{MC,n}^{\pi_3} > \hat{V}_{MC,n}^{\pi_1}) = 1 - \phi \approx 0.618$ *where $\phi = \frac{\sqrt{5}+1}{2}$ is the golden ratio. Moreover, it is possible that for any policy π, there is another policy π' where $\succ_{MC,n}$ $(\pi', \pi) = \texttt{True}$.*

Proof of Theorem B.1. Consider a multi-armed bandit where there are three actions: a_1, which gives a reward of $n+1$ with probability p and a reward of 0 with probability $1-p$, a_2 which always gives a reward of 1, and a_3 which gives a reward of n^2+n+1 with probability $1-q$ and a reward of 0.5 with probability q. Suppose policies π_1, π_2, and π_3 always choose action a_1, a_2, and a_3 respectively. Now suppose we want to estimate the three policies with n on-policy samples from each. We have that π_1 gives a higher reward than π_2 whenever we get the large reward at least once, which happens with probability $1-(1-p)^n$. Thus

$$\Pr(\hat{V}_{\text{MC},n}^{\pi_1} > \hat{V}_{\text{MC},n}^{\pi_2}) = 1 - (1-p)^n$$

Furthermore, clearly π_2 gives a larger reward than π_3 whenever all samples of π_2 give a reward of 0.5, which happens with probability q^n. Now, finally we see that π_3 gives a larger reward than π_1 whenever it gives at least one sample with a large reward or when both of them give only samples of their small rewards, which happens with probability $(1-q^n) + q^n(1-p)^n$, so

$$\Pr(\hat{V}_{MC,n}^{\pi_3} > \hat{V}_{\text{MC},n}^{\pi_1}) = (1-q^n) + q^n(1-p)^n$$

Now let $p = 1-(2-\phi)^{1/n}$ and $q = (\phi-1)^{1/n}$, where $\phi = \frac{\sqrt{5}+1}{2} \approx 1.618$ is the golden ratio. Thus we have that:

$$\Pr(\hat{V}_{\text{MC},n}^{\pi_1} > \hat{V}_{\text{MC},n}^{\pi_2}) = \phi - 1$$

$$\Pr(\hat{V}_{\text{MC},n}^{\pi_2} > \hat{V}_{\text{MC},n}^{\pi_3}) = \phi - 1$$

$$\Pr(\hat{V}_{\text{MC},n}^{\pi_3} > \hat{V}_{\text{MC},n}^{\pi_1}) = (2-\phi) + (\phi-1)(2-\phi) = \phi - 1$$

We now show that for this multi-armed bandit, there is no optimal policy with respect to $\succ_{\text{MC},n}$. A policy in this setting is simply a distribution over a_1, a_2, and a_3. Equivalently, we can view any policy as a mix of the policies π_1, π_2, and π_3. Suppose a policy π executes π_1 with probability p, π_2 with probability q, and π_3 with probability r. If p is the largest of the probabilities, then notice that $\Pr(\hat{V}_{\text{MC},n}^{\pi_3} > \hat{V}_{\text{MC},n}^{\pi}) = p(\phi-1) \geq q(\phi-1) = \Pr(\hat{V}_{\text{MC},n}^{\pi} > \hat{V}_{\text{MC},n}^{\pi_3})$. We can make a similar argument if q or r are the largest probabilities. Thus, there is no optimal policy. □

C FAIRNESS PROOFS

Theorem 6.1. *Using the on-policy Monte Carlo estimator for policy selection when we have n samples*

from each of policies π_1 and π_2 is fair provided that $V^{\pi_1}_{Max} + V^{\pi_2}_{Max} \leq |V^{\pi_1} - V^{\pi_2}|\sqrt{\frac{2n}{\ln 2}}$ We can guarantee strict fairness if the inequality above is strict.

Proof of Theorem 6.1. Suppose without loss of generality that $V^{\pi_1} > V^{\pi_2}$. Let $\hat{V}_i^{\pi} = \sum_{t=1}^{T_i} R_{i,t}$ (i.e., the estimate of the value of policy π using only τ_i^{π}). Now let

$$X_i = \hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2}$$

Note that the range of X_i is $[-V^{\pi_2}_{\text{Max}}, V^{\pi_1}_{\text{Max}}]$. Let ω be the difference between the upper and lower bounds of X_i, that is, $\omega = V^{\pi_1}_{\text{Max}} + V^{\pi_2}_{\text{Max}}$. Because all $\tau_i^{\pi_1}$ and $\tau_i^{\pi_2}$ are independent of $\tau_j^{\pi_1}$ and $\tau_j^{\pi_2}$ for all $i \neq j$, we know that X_i is independent of X_j for all $i \neq j$. Thus we can use Hoeffding's inequality to find that:

$$\Pr\left(\bar{X} \leq 0\right) = \Pr\left(\bar{X} - \mathbf{E}\left[\bar{X}\right] \leq -\mathbf{E}\left[\bar{X}\right]\right)$$
$$\leq \exp\left(\frac{-2n\mathbf{E}\left[\bar{X}\right]^2}{\omega^2}\right)$$

Note that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2} = \hat{V}^{\pi_1}_{\text{MC},n} - \hat{V}^{\pi_2}_{\text{MC},n}$$

and $\mathbf{E}\left[\bar{X}\right] = V^{\pi_1} - V^{\pi_2}$. Thus, if we want to guarantee

$$\Pr\left(\hat{V}^{\pi_1}_{\text{MC},n} - \hat{V}^{\pi_2}_{\text{MC},n} \leq 0\right) \leq \delta$$

we can simply guarantee

$$\exp\left(\frac{-2n(V^{\pi_1} - V^{\pi_2})^2}{\omega^2}\right) \leq \delta$$

Solving for ω, we must have that:

$$\omega \leq (V^{\pi_1} - V^{\pi_2})\sqrt{\frac{2n}{\ln(1/\delta)}}$$

Substituting $\delta = 0.5$, we can thus guarantee that $\Pr\left(\hat{V}^{\pi_1}_{\text{MC},n} - \hat{V}^{\pi_2}_{\text{MC},n} > 0\right) \geq 0.5$, which guarantees fairness when $V^{\pi_1} > V^{\pi_2}$. Since we do not actually know which policy has a greater value, we can guarantee fairness with the following condition:

$$\omega \leq |V^{\pi_1} - V^{\pi_2}|\sqrt{\frac{2n}{\ln 2}}$$

□

Theorem 6.2. *Using importance sampling for policy selection when we have n samples from the behavior policy is fair with respect to $\succ_V$, provided that*

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

We can guarantee strict fairness if the inequality above is strict.

Proof of Theorem 6.2. Let $\hat{V}_i^{\pi} = \sum_{t=1}^{T_i} w_i^{\pi_e} R_{i,t}$ (i.e., the estimate of the value of policy π using only τ_i). Now let

$$X_i = \hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2}$$

Note that the range of X_i is $[-w^{\pi_2}_{\text{Max}}V^{\pi_2}_{\text{Max}}, w^{\pi_1}_{\text{Max}}V^{\pi_1}_{\text{Max}}]$. Let ω be the difference between the upper and lower bounds of X_i, that is, $\omega = w^{\pi_1}_{\text{Max}}V^{\pi_1}_{\text{Max}} + w^{\pi_2}_{\text{Max}}V^{\pi_2}_{\text{Max}}$. Because τ_i and τ_j are independent for all $i \neq j$, we know that X_i is independent of X_j for all $i \neq j$. The rest of the proof follows exactly as in the proof of Theorem 6.1. □

Theorem 6.3. *For any two policies π_1 and π_2, behavior policy π_b, and for all $k \in \{1, 2, 3, \dots\}$, using importance sampling for policy selection when we have n samples from the behavior policy is fair with respect to $\succ_{MC,kn}$ provided that there exists $\epsilon > 0$ and $\delta < 0.5$ such that $\Pr\left(|\hat{V}^{\pi_1}_{MC,kn} - \hat{V}^{\pi_2}_{MC,kn}| \geq \epsilon\right) \geq 1 - \delta$ and $(w^{\pi_1}_{Max}+1)V^{\pi_1}_{Max} + (w^{\pi_2}_{Max}+1)V^{\pi_2}_{Max} \leq \epsilon\sqrt{\frac{2n}{\ln\frac{1-\delta}{0.5-\delta}}}$. Importance sampling in this setting is transitively fair, provided that $\delta \leq 0.25$.*

Proof of Theorem 6.3. Suppose without loss of generality that $\Pr\left(\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn} \geq \epsilon\right) \geq 1 - \delta$. Recall that IS uses trajectories $\tau_1, \dots, \tau_n \sim \pi_b$. Consider additional random samples $\tau_1^{\pi_1}, \dots, \tau_{kn}^{\pi_1} \sim \pi_1$ and $\tau_1^{\pi_2}, \dots, \tau_{kn}^{\pi_2} \sim \pi_2$. Note that these samples are all independent from each other. For $i \in \{1, 2, \dots, n\}$, let

$$\hat{V}_i^{\pi} = \frac{1}{k}\sum_{j=1}^{k}\sum_{t=1}^{T_{ji}} R_{ji,t}$$

(i.e., the estimate of the value of policy π using only samples $\tau_i^{\pi}, \tau_{2i}^{\pi}, \dots \tau_{ki}^{\pi}$). Furthermore let $\hat{V}^{\pi}_{\text{IS},i} = \sum_{t=1}^{T_i} w_{i,t}R_{i,t}$. Now let

$$X_i = (\hat{V}^{\pi_1}_{\text{IS},i} - \hat{V}^{\pi_2}_{\text{IS},i}) - (\hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2})$$

Notice that the range of X_i is $[-V^{\pi_2}_{\text{Max}} - w^{\pi_1}_{\text{Max}}V^{\pi_1}_{\text{Max}}, V^{\pi_1}_{\text{Max}} + w^{\pi_2}_{\text{Max}}V^{\pi_2}_{\text{Max}}]$. Let ω be the difference between the upper and lower bounds of X_i, that is, $\omega = (w^{\pi_1}_{\text{Max}} + 1)V^{\pi_1}_{\text{Max}} + (w^{\pi_2}_{\text{Max}} + 1)V^{\pi_2}_{\text{Max}}$.
Notice that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{V}^{\pi_1}_{\text{IS},i} - \hat{V}^{\pi_2}_{\text{IS},i}) - (\hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2})$$
$$= (\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}}) - (\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn})$$

and

$$\mathbf{E}\left[\bar{X}\right] = (V^{\pi_1} - V^{\pi_2}) - (V^{\pi_1} - V^{\pi_2}) = 0$$

Thus we have that

$$\Pr\left(\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}} \leq 0\right) = \Pr\left(\bar{X} \leq -(\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn})\right)$$

$$= \Pr\left(\bar{X} - \mathbf{E}\left[\bar{X}\right] \leq -(\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn})\right)$$

Thus, we can use Hoeffding's inequality to find that:

$$\Pr\left(\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}} \leq 0 | \hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn} \geq \epsilon\right) \leq \exp\left(\frac{-2n\epsilon^2}{\omega^2}\right)$$

So if we want to guarantee

$$\Pr\left(\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}} \leq 0\right) \leq \gamma$$

we can simply guarantee

$$\exp\left(\frac{-2n(\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn})^2}{\omega^2}\right) \leq \gamma$$

Solving for ω, we must have that:

$$\omega \leq (\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn})\sqrt{\frac{2n}{\ln(1/\gamma)}}$$

Notice that

$$\begin{aligned}
&\Pr\left(\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}} \geq 0\right) \\
&\geq \Pr\left(\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}} \leq 0 | \hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn} \geq \epsilon\right) \\
&\quad \times \Pr\left(\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_2}_{\text{MC},kn} \geq \epsilon\right) \\
&\geq (1-\gamma)(1-\delta)
\end{aligned}$$

If we set $\gamma = \frac{0.5-\delta}{1-\delta}$, we have that

$$\Pr\left(\hat{V}^{\pi_1}_{\text{IS}} - \hat{V}^{\pi_2}_{\text{IS}} \geq 0\right) \geq 0.5$$

which is what we want.

As long as $\delta \leq 0.25$, we have that the IS is transitively fairness since any fair algorithm with respect to $\succ_{\text{MC},kn}$ is transitively fair whenever $\Pr(|\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_1}_{\text{MC},kn}| > 0) \geq 0.75$ □

D MULTIPLE COMPARISONS

Here we present an algorithm that uses pairwise comparisons to select amongst $k \geq 2$ policies (Algorithm 3). This algorithm can take as input either of the off-policy fair policy selection algorithms above (or some variant thereof).

Theorem D.1. *For any finite set of k policies Π, behavior policy π_b, $p = 0.5$, and fair off-policy policy selection algorithm `FPS`, Algorithm 3 is a strictly fair policy selection algorithm with when we have n samples drawn from π_b.*

Algorithm 3 Off-Policy FPS for k policies

input Π, V^{Π}_{Max}, ϵ, p
$\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n \sim \pi_b\}$, `FPS`
$\delta \leftarrow (1-p)/(2k+3)$
$\pi^* \leftarrow \Pi$.next
Eliminated $\leftarrow \emptyset$
CurrBeat $\leftarrow \emptyset$
repeat
  $\pi' \leftarrow (\Pi \backslash \text{CurrBeat})$.next
  winner $\leftarrow$ `FPS`$(\pi^*, \pi', V^{\pi^*}_{\text{Max}}, V^{\pi'}_{\text{Max}}, \epsilon, \delta, \mathcal{T})$
  if winner == $\pi*$ **then**
    Eliminated $\leftarrow$ Eliminated $\cup \{\pi'\}$
    CurrBeat $\leftarrow$ CurrBeat $\cup \{\pi'\}$
  else if winner == π' **then**
    $\pi* \leftarrow \pi'$
    Eliminated $\leftarrow$ Eliminated $\cup \{\pi^*\}$
    CurrBeat $\leftarrow$ CurrBeat $\cup \{\pi^*\}$
  else
    $\pi^* \leftarrow (\Pi \backslash \text{Eliminated})$.next
    Eliminated $\leftarrow$ Eliminated $\cup \{\pi^*, \pi'\}$
    CurrBeat $\leftarrow \emptyset$
  end if
until len(Eliminated) == $k-1$ or len(CurrBeat) == k
if len(Eliminated) == $k-1$ **then**
  return π^*
else
  return `No Fair Comparison`
end if

Proof of Theorem D.1. The algorithm essentially applies an algorithm for finding the maximum element of a set, with the exception that whenever it cannot make a fair comparison between two policies, it will eliminate both of those policies from consideration of being better than all other policies with respect to the better-than function. The algorithm must return `No Fair Comparison` if and only if every policy is eliminated. Notice that we only eliminate a policy when it is not returned by `FPS` or when `No Fair Comparison` is returned, which is correct. Notice that until there are $k-1$ policies that are eliminated, at every comparison at least one policy is eliminated and the last of those comparisons must include the only remaining non-eliminated policy. Afterwards it takes at most $k-2$ comparisons with the final policy (comparing it to every other policy other than the one it was already compared to) to determine that no fair comparison is possible, making a total of $2k-3$ comparisons.

On the other hand, the algorithm must return a policy when that policy is outputted by `FPS` from comparisons with every other policy, which is exactly what it does (i.e. when the CurrBeated set includes $k-1$ policies). The maximum number of comparisons it takes to output a policy is the number of comparisons it takes to eliminate $k-1$ other policies plus the number of comparisons it takes to beat $k-2$ policies using the same argument as above, making a total of $2k-3$ comparisons. Thus, if we let $\delta = (1-p)/(2k-3)$ all of the comparisons made by `FPS` will simultaneously hold with probability $1-(2k-3)\delta = p$, and so setting $p = 0.5$ ensures fairness. □

E SAFETY THEOREMS AND PROOFS

In this section, we will prove the theorems for ensuring safety when using importance sampling for policy selection.

Theorem 6.4. *For any two policies π_1 and π_2, behavior policy π_b, $\omega = w_{Max}^{\pi_1} V_{Max}^{\pi_1} + w_{Max}^{\pi_2} V_{Max}^{\pi_2}$, and $\delta \leq 0.5$, Algorithm 2 is a safe policy selection algorithm with respect to $\succ_V$ with probability $1-\delta$.*

Proof of Theorem 6.4. Let $\hat{V}_i^{\pi} = \sum_{t=1}^{T_i} w_i^{\pi_e} R_{i,t}$ (i.e., the estimate of the value of policy π using only τ_i). Now let

$$X_i = \hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2}$$

Note that the range of X_i is $[-w_{\text{Max}}^{\pi_2} V_{\text{Max}}^{\pi_2}, w_{\text{Max}}^{\pi_1} V_{\text{Max}}^{\pi_1}]$. Because τ_i and τ_j are independent for all $i \neq j$, we know that X_i is independent of X_j for all $i \neq j$. Thus we can use Hoeffding's inequality to find that:

$$\Pr\left(\bar{X} - \mathbf{E}\left[\bar{X}\right] \geq -\omega\sqrt{\frac{\ln(1/\gamma)}{2n}}\right) \geq 1-\gamma$$

and

$$\Pr\left(\bar{X} - \mathbf{E}\left[\bar{X}\right] \leq \omega\sqrt{\frac{\ln(1/\gamma)}{2n}}\right) \geq 1-\gamma$$

where $\omega = w_{\text{Max}}^{\pi_1} V_{\text{Max}}^{\pi_1} + w_{\text{Max}}^{\pi_2} V_{\text{Max}}^{\pi_2}$. Note that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2} = \hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2}$$

and $\mathbf{E}\left[\bar{X}\right] = V^{\pi_1} - V^{\pi_2}$. Thus, substituting $(1-p)/2$ for γ, we have that the following two statements hold with probability at least $1-2\gamma = p$,

$$V^{\pi_1} - V^{\pi_2} \geq \hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} - \sqrt{\frac{\ln(2/(1-p))}{2n}}$$

and

$$V^{\pi_1} - V^{\pi_2} \leq \hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} + \sqrt{\frac{\ln(2/(1-p))}{2n}}$$

Thus the probability that $V^{\pi_1} - V^{\pi_2} < 0$ but $\hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} - \sqrt{\frac{\ln(2/(1-p))}{2n}} > 0$ or $V^{\pi_1} - V^{\pi_2} > 0$ but $\hat{V}_{\text{IS}}^{\pi_1} + \hat{V}_{\text{IS}}^{\pi_2} - \sqrt{\frac{\ln(2/(1-p))}{2n}} < 0$ is less than p, which means for $p = 1-\delta$, we will output the worse policy according to $\succ_V$ with probability at most δ, which is exactly what we need. □

Theorem 6.5. *For any two policies π_1 and π_2, where $\Pr\left(|\hat{V}_{MC,kn}^{\pi_1} - \hat{V}_{MC,kn}^{\pi_2}| \geq 0\right) \geq 1-\delta_{MC}$ any behavior policy π_b, $\omega = (w_{Max}^{\pi_1}+1)V_{Max}^{\pi_1} + (w_{Max}^{\pi_2}+1)V_{Max}^{\pi_2}$, $p = 1-\delta_{MC}\delta$ for some $\delta \leq 0.5$, and for all $k \in \{1,2,3,\dots\}$, Algorithm 2 is a safe policy selection algorithm with respect to $\succ_{MC,kn}$ with probability $1-\delta$ when we have n samples drawn from π_b. It is a transitively fair policy selection algorithm whenever $\delta_{MC} \leq 0.25$.*

Proof of Theorem 6.5. Recall that Algorithm 2 receives as input $\tau_1, \dots, \tau_n \sim \pi_b$. Consider additional random samples $\tau_1^{\pi_1}, \dots, \tau_{kn}^{\pi_1} \sim \pi_1$ and $\tau_1^{\pi_2}, \dots, \tau_{kn}^{\pi_2} \sim \pi_2$. Note that these samples are all independent from each other. For $i \in \{1,2,\dots,n\}$, let

$$\hat{V}_i^{\pi} = \frac{1}{k}\sum_{j=1}^{k}\sum_{t=1}^{T_{ji}} R_{ji,t}$$

(i.e., the estimate of the value of policy π using only samples $\tau_i^\pi, \tau_{2i}^\pi, \ldots \tau_{ki}^\pi$). Furthermore let $\hat{V}_{\text{IS},i}^\pi = \sum_{t=1}^{T_i} w_{i,t} R_{i,t}$. Now let

$$X_i = (\hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2}) - (\hat{V}_{\text{IS},i}^{\pi_1} - \hat{V}_{\text{IS},i}^{\pi_2})$$

Notice that the range of X_i is $[-V_{\text{Max}}^{\pi_2} - w_{\text{Max}}^{\pi_1} V_{\text{Max}}^{\pi_1}, V_{\text{Max}}^{\pi_1} + w_{\text{Max}}^{\pi_2} V_{\text{Max}}^{\pi_2}]$. Thus we can use Hoeffding's inequality to find that:

$$\Pr\left(\bar{X} - \mathbf{E}\left[\bar{X}\right] \geq -\omega\sqrt{\frac{\ln(1/\delta)}{2n}}\right) \geq 1 - \gamma$$

and

$$\Pr\left(\bar{X} - \mathbf{E}\left[\bar{X}\right] \leq \omega\sqrt{\frac{\ln(1/\delta)}{2n}}\right) \geq 1 - \gamma$$

where $\omega = (w_{\text{Max}}^{\pi_1} + 1)V_{\text{Max}}^{\pi_1} + (w_{\text{Max}}^{\pi_2} + 1)V_{\text{Max}}^{\pi_2}$.
Notice that

$$\begin{aligned}\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i^{\pi_1} - \hat{V}_i^{\pi_2}) - (\hat{V}_{\text{IS},i}^{\pi_1} - \hat{V}_{\text{IS},i}^{\pi_2}) \\ &= (\hat{V}_{\text{MC},kn}^{\pi_1} - \hat{V}_{\text{MC},kn}^{\pi_2}) - (\hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2})\end{aligned}$$

and

$$\mathbf{E}\left[\bar{X}\right] = (V^{\pi_1} - V^{\pi_2}) - (V^{\pi_1} - V^{\pi_2}) = 0$$

Thus, substituting $(1-p)/2$ for γ, we have that the following two statements hold with probability at least $1 - 2\gamma = p$,

$$\hat{V}_{\text{MC},kn}^{\pi_1} - \hat{V}_{\text{MC},kn}^{\pi_2} \geq \hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} - \sqrt{\frac{\ln(2/(1-p))}{2n}}$$

and

$$\hat{V}_{\text{MC},kn}^{\pi_1} - \hat{V}_{\text{MC},kn}^{\pi_2} \leq \hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} + \sqrt{\frac{\ln(2/(1-p))}{2n}}$$

Thus the probability that $\hat{V}_{\text{MC},kn}^{\pi_1} - \hat{V}_{\text{MC},kn}^{\pi_2} < 0$ but $(\hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2}) - \omega\sqrt{\frac{\ln(2/(1-p))}{2n}} > 0$ or $\hat{V}_{\text{MC},kn}^{\pi_1} - \hat{V}_{\text{MC},kn}^{\pi_2} > 0$ but $\hat{V}_{\text{IS}}^{\pi_1} - \hat{V}_{\text{IS}}^{\pi_2} + \omega\sqrt{\frac{\ln(2/(1-p))}{2n}} < 0$ is less than $1 - p$. Now suppose without loss of generality that $P(\hat{V}_{\text{MC},kn}^{\pi_1} > \hat{V}_{\text{MC},kn}^{\pi_2} = \delta_{\text{MC}} > 0.5$. The probability that we output π_2 is at most p/δ_{MC}. So if $p = 1 - \delta_{\text{MC}}\delta$, we output the worse policy with respect to $\succ_{\text{MC},kn}$ with probability at most δ, which is exactly what we need.

As long as $\delta_{\text{MC}} \leq 0.25$, we have that the algorithm is transitively safe since any fair algorithm with respect to $\succ_{\text{MC},kn}$ is transitively fair whenever $\Pr(|\hat{V}_{\text{MC}}^{\pi_1} - \hat{V}_{\text{MC}}^{\pi_1}| > 0) \geq 0.75$ □

Theorem 6.6. *There exists policies π_1, π_2, and behavior policy π_b for which Algorithm 2 with inputs as described in Theorem 6.4 is not a safe policy selection algorithm with respect to $\succ_{MC,1}$ with $p = 0.5$ when we have a single sample drawn from π_b.*

Proof of Theorem 6.6. Consider a world where there are three trajectories: τ_1 with reward 0.0001, τ_2 with reward 0.0002, and τ_3 with reward 1. We want to select between two policies: π_1, which places probability 1 on τ_2 and π_2 which places probability 0.51 on τ_1 and probability 0.49 on τ_3. When we only have one sample from each policy, $\Pr(\hat{V}_{\text{MC},1}^{\pi_1} > \hat{V}_{\text{MC},1}^{\pi_2}) = 0.51 > 0.5$, but clearly $V^{\pi_1} \ll V^{\pi_2}$. Now consider using IS with behavior policy π_b which places probability 0.48 on τ_1 and probability 0.01 on τ_2 and probability 0.51 on τ_3. If we apply Algorithm 2 with the inputs to guarantee that the algorithm is safe with respect $\succ_V$ (as given in Theorem 6.4), we find that whenever π_b samples from τ_3,

$$\begin{aligned}&V_{\text{IS}}^{\pi_1} - V_{\text{IS}}^{\pi_2} + (w_{\text{Max}}^{\pi_1} V_{\text{Max}}^{\pi_1} + w_{\text{Max}}^{\pi_2} V_{\text{Max}}^{\pi_2})\sqrt{\frac{\ln 4}{2n}} \\ &= 0(1) - \frac{0.49}{0.51}(1) + \left(\frac{1}{0.01}(0.0002) + \frac{0.51}{0.48}(1)\right)\sqrt{\frac{\ln 4}{2}} \\ &\approx -0.060 < 0\end{aligned}$$

Since this event occurs with probability 0.51, we find that Algorithm 2 returns π_2 more than half the time, indicating that Algorithm 2 is not a safe policy with respect to $\succ_{\text{MC},1}$ □