SUPPLEMENTARY MATERIAL

A UNFAIRNESS OF IMPORTANCE SAMPLING

Suppose we want to use importance sampling to select the better of two policies, π_a 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^{π_b} = p + (1 - p)r and V^{π_a} = 1, so a fair policy selection algorithm should choose π_a 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_a. 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 π_2 is never sampled, IS will choose π_b, since in that case \( V^{π_b} = 0 \). π_b will never sample π_2 with probability \( (1 - p)^n \). Thus IS is unfair as long as \( (1 - p)^n ≤ 0.5 \), or as long as \( p ≤ 1 - 0.5^{1/n} ≈ \ln(2)/n \) 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

<table>
<thead>
<tr>
<th></th>
<th>( a_1 )</th>
<th>( a_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R = r &lt; 1 )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( R = 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \pi_e \)

\( \pi_b \)

\( 1 - p \)

\( p \)

B NON-TRANSITIVITY

Theorem B.1 (Non-Transitivity of \( >_{MC,n} \)). The relation induced by \( >_{MC,n} \) is non-transitive. Specifically, there exists policies \( π_1, π_2, \) and \( π_3 \) where

\[
\Pr(\hat{V}^{π_1}_{MC,n} > \hat{V}^{π_2}_{MC,n}) = \Pr(\hat{V}^{π_2}_{MC,n} > \hat{V}^{π_3}_{MC,n}) = \Pr(\hat{V}^{π_3}_{MC,n} > \hat{V}^{π_1}_{MC,n}) = 1 - φ \approx 0.618 \text{ where } φ = \frac{\sqrt{5} + 1}{2} \text{ is the golden ratio. Moreover, it is possible that for any policy } π, \text{ there is another policy } π' \text{ where } >_{MC,n}(π', π) = \text{ True.}
\]

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

\[
\Pr(\hat{V}^{π_1}_{MC,n} > \hat{V}^{π_2}_{MC,n}) = 1 - (1 - p)^n
\]

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

\[
\Pr(\hat{V}^{π_3}_{MC,n} > \hat{V}^{π_1}_{MC,n}) = (1 - q^n) + q^n(1 - p)^n
\]

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

\[
\Pr(\hat{V}^{π_1}_{MC,n} > \hat{V}^{π_2}_{MC,n}) = φ - 1
\]

\[
\Pr(\hat{V}^{π_2}_{MC,n} > \hat{V}^{π_3}_{MC,n}) = φ - 1
\]

\[
\Pr(\hat{V}^{π_3}_{MC,n} > \hat{V}^{π_1}_{MC,n}) = (2 - φ) + (φ - 1)(2 - φ) = φ - 1
\]

We now show that for this multi-armed bandit, there is no optimal policy with respect to \( >_{MC,n} \). 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 \( \Pr(\hat{V}^{π_1}_{MC,n} > \hat{V}^{π_2}_{MC,n}) = p(φ - 1) ≥ q(φ - 1) = \Pr(\hat{V}^{π_3}_{MC,n} > \hat{V}^{π_1}_{MC,n}) \). 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
Proof of Theorem 6.1. Suppose without loss of generality that $V_{\pi_1} > V_{\pi_2}$. Let $\hat{V}_\pi = \sum_{t=1}^{T_\pi} R_{i,t}$ (i.e., the estimate of the value of policy $\pi$ using only $\tau_i^\pi$). Now let

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

Note that the range of $X_i$ is $[-V_{\pi_1}^{Max}, V_{\pi_1}^{Max}]$. Let $\omega$ be the difference between the upper and lower bounds of $X_i$, that is, $\omega = V_{\pi_1}^{Max} + V_{\pi_2}^{Max}$. 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:

$$\Pr(\bar{X} \leq 0) = \Pr(\bar{X} - \mathbb{E}[\bar{X}] \leq -\mathbb{E}[\bar{X}])$$

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

Note that

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

and $\mathbb{E}[\bar{X}] = V_{\pi_1} - V_{\pi_2}$. Thus, if we want to guarantee

$$\Pr(\bar{V}_{MC,n}^{\pi_1} - \bar{V}_{MC,n}^{\pi_2} \leq 0) \leq \delta$$

we can simply guarantee

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

Solving for $\omega$, 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

$$\Pr\left(\bar{V}_{MC,n}^{\pi_1} - \bar{V}_{MC,n}^{\pi_2} > 0\right) \geq 0.5$$

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:

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

\[\square\]

Theorem 6.2. Using importance sampling for policy selection when we have $n$ samples from the behavior policy is fair with respect to $\pi$, 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 $\hat{V}_i^{\pi} = \sum_{t=1}^{T_i} w_i^{\pi} R_{i,t}$ (i.e., the estimate of the value of policy $\pi$ using only $\tau_i^\pi$). Now let

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

Note that the range of $X_i$ is $[-w_{\pi_1}^{\max} V_{\max}^{\pi_1} w_{\pi_2}^{\max} V_{\max}^{\pi_2}].$ Let $\omega$ be the difference between the upper and lower bounds of $X_i$, that is, $\omega = w_{\pi_1}^{\max} V_{\max}^{\pi_1} + w_{\pi_2}^{\max} V_{\max}^{\pi_2}$. Because $\tau_i$ and $\tau_j$ 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. \[\square\]

Theorem 6.3. For any two policies $\pi_1$ and $\pi_2$, behavior policy $\pi_b$, 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 $\pi_{MC,kn}$ provided that there exists $\epsilon > 0$ and $\delta < 0.5$ such that $\Pr\left(V_{MC,kn}^{\pi_1} - V_{MC,kn}^{\pi_2} \geq \epsilon\right) \geq 1 - \delta$ and

$$(w_{\pi_1}^{\max} + 1) V_{\max}^{\pi_1} + (w_{\pi_2}^{\max} + 1) V_{\max}^{\pi_2} \leq \epsilon \sqrt{\frac{2n}{\ln \frac{1}{\delta}}}$$

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 $\Pr\left(V_{MC,kn}^{\pi_1} - V_{MC,kn}^{\pi_2} \geq \epsilon\right) \geq 1 - \delta$. Recall that IS uses trajectories $\tau_1, \ldots, \tau_n \sim \pi_b$. Consider additional random samples $\tau_1^{\pi_1}, \ldots, \tau_n^{\pi_1} \sim \pi_1$ and $\tau_1^{\pi_2}, \ldots, \tau_n^{\pi_2} \sim \pi_2$. Note that these samples are all independent from each other. For $i \in \{1, 2, \ldots, n\}$, let

$$\hat{V}_{i}^{\pi} = 1/k \sum_{j=1}^{k} R_{i,j}$$

(i.e., the estimate of the value of policy $\pi$ using only samples $\tau_i^{\pi_1}, \tau_i^{\pi_2}, \ldots, \tau_i^{\pi_k}$). Furthermore let $\hat{V}_{IS,i}^{\pi} = \sum_{t=1}^{T_i} w_i^{\pi} R_{i,t}$. Now let

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

Notice that the range of $X_i$ is $[-w_{\pi_1}^{\max} V_{\max}^{\pi_1} w_{\pi_2}^{\max} V_{\max}^{\pi_2}].$ Let $\omega$ be the difference between the upper and lower bounds of $X_i$, that is, $\omega = (w_{\pi_1}^{\max} + 1) V_{\max}^{\pi_1} + (w_{\pi_2}^{\max} + 1) V_{\max}^{\pi_2}$. Let

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

and

$$\mathbb{E}[\hat{X}] = (V_{\pi_1} - V_{\pi_2}) - (V_{\pi_1} - V_{\pi_2}) = 0$$

Thus we have that

$$\Pr(\hat{V}_{IS}^{\pi_1} - \hat{V}_{IS}^{\pi_2} \leq 0) = \Pr(\hat{X} \leq -(\hat{V}_{IS}^{\pi_1} - \hat{V}_{IS}^{\pi_2}))$$

\[\square\]
Thus, we can use Hoeffding’s inequality to find that:

\[
\Pr \left( V_{\pi_1} - V_{\pi_2} \leq 0 \Big| \hat{V}_{MC,kn}^\pi \leq \epsilon \right) \leq \exp \left( -\frac{2\epsilon^2}{\omega^2} \right)
\]

So if we want to guarantee

\[
\Pr \left( V_{\pi_1} - V_{\pi_2} \leq 0 \right) \leq \gamma
\]

we can simply guarantee

\[
\exp \left( -\frac{2\epsilon^2 (\hat{V}_{MC,kn}^\pi - \hat{V}_{MC,kn}^{\pi_2})^2}{\omega^2} \right) \leq \gamma
\]

Solving for \( \omega \), we must have that:

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

Notice that

\[
\Pr \left( \hat{V}_{IS}^{\pi_1} - \hat{V}_{IS}^{\pi_2} \geq 0 \right)
\]

\[
\geq \Pr \left( \hat{V}_{IS}^{\pi_1} - \hat{V}_{IS}^{\pi_2} \leq 0 | \hat{V}_{MC,kn}^{\pi_1} - \hat{V}_{MC,kn}^{\pi_2} \geq \epsilon \right)
\]

\[
\times \Pr \left( \hat{V}_{MC,kn}^{\pi_1} - \hat{V}_{MC,kn}^{\pi_2} \geq \epsilon \right)
\]

\[
\geq (1 - \gamma)(1 - \delta)
\]

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

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

which is what we want.

As long as \( \delta \leq 0.25 \), we have that the IS is transitively fair since any fair algorithm with respect to \( \succ_{MC,kn} \) is transitively fair whenever \( \Pr(|\hat{V}_{MC,kn}^{\pi_1} - \hat{V}_{MC,kn}^{\pi_2}| > 0) \geq 0.75 \)

### 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 \( \Pi \), behavior policy \( \pi_b, p = 0.5 \), and fair off-policy policy selection algorithm \( \text{FPS} \), Algorithm 3 is a strictly fair policy selection algorithm with when we have \( n \) samples drawn from \( \pi_b \).

---

**Algorithm 3 Off-Policy FPS for \( k \) policies**

**Input** \( \Pi, V_{\Pi,\text{Max}}, \epsilon, p \)

\( T = \{\tau_1, \tau_2, \ldots, \tau_n \sim \pi_b\}, \text{FPS} \)

\( \delta \leftarrow (1 - p)/(2k + 3) \)

\( \pi^* \leftarrow \Pi, \text{next} \)

\( \text{Eliminated} \leftarrow \emptyset \)

\( \text{CurrBeat} \leftarrow \emptyset \)

**repeat**

\( \pi' \leftarrow (\Pi, \text{CurrBeat}).\text{next} \)

\( \text{winner} \leftarrow \text{FPS}(\pi^*, \pi', V_{\Pi,\text{Max}}^\pi, V_{\Pi,\text{Max}}', \epsilon, \delta, T) \)

**if** \( \text{winner} == \pi^* \) **then**

\( \text{Eliminated} \leftarrow \text{Eliminated} \cup \{\pi'\} \)

\( \text{CurrBeat} \leftarrow \text{CurrBeat} \cup \{\pi'\} \)

**else if** \( \text{winner} == \pi' \) **then**

\( \pi^* \leftarrow \pi' \)

\( \text{Eliminated} \leftarrow \text{Eliminated} \cup \{\pi^*\} \)

\( \text{CurrBeat} \leftarrow \text{CurrBeat} \cup \{\pi^*\} \)

**else**

\( \pi^* \leftarrow (\Pi, \text{Eliminated}).\text{next} \)

\( \text{Eliminated} \leftarrow \text{Eliminated} \cup \{\pi^*, \pi'\} \)

\( \text{CurrBeat} \leftarrow \emptyset \)

**end if**

**until** \( \text{len(Eliminated)} == k - 1 \) or \( \text{len(CurrBeat)} == k \)

**if** \( \text{len(Eliminated)} == k - 1 \) **then**

**return** \( \pi^* \)

**else**

**return** No Fair Comparison

**end if**
Proof of Theorem 6.4. 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 only one policy is 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 \( 2k - 3 \) 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.

\[ \text{Proof of Theorem 6.4.} \]

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 \( \pi_1 \) and \( \pi_2 \), behavior policy \( \pi_b \), \( \omega = w_{\pi_1}^{\pi_2} V_{\pi_1}^{\pi_2} + w_{\pi_2}^{\pi_1} V_{\pi_2}^{\pi_1} \) and \( \delta \leq 0.5 \), Algorithm 2 is a safe policy selection algorithm with respect to \( \succ V \) with probability \( 1 - \delta \).

\[ \text{Proof of Theorem 6.4.} \]

Let \( \hat{V}_i = \sum_{t=1}^{T_i} w_i^\tau_t R_{i,t} \) (i.e., the estimate of the value of policy \( \pi \) using only \( \tau_i \)). Now let

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

Note that the range of \( X_i \) is \( [-w_{\pi_1}^{\pi_2} V_{\pi_1}^{\pi_2}, w_{\pi_2}^{\pi_1} V_{\pi_2}^{\pi_1}] \). Because \( \tau_i \) and \( \tau_j \) 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:

\[ \Pr \left( X_i - E[X_i] \geq -\omega \frac{\ln(1/\gamma)}{2n} \right) \geq 1 - \gamma \]

and

\[ \Pr \left( X_i - E[X_i] \leq \omega \frac{\ln(1/\gamma)}{2n} \right) \geq 1 - \gamma \]

where \( \omega = w_{\pi_1}^{\pi_2} V_{\pi_1}^{\pi_2} + w_{\pi_2}^{\pi_1} V_{\pi_2}^{\pi_1} \). Note that

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

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

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

and

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

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

\[ \text{Theorem 6.5.} \]

For any two policies \( \pi_1 \) and \( \pi_2 \), where \( \Pr \left( \left| \hat{V}_{MC,\pi_1}^{\pi_2} - \hat{V}_{MC,\pi_2}^{\pi_1} \right| \geq \delta \right) \geq 1 - \delta_{MC} \) any behavior policy \( \pi_b \), \( \omega = (w_{\pi_1}^{\pi_2} + 1)V_{\pi_1}^{\pi_2} + (w_{\pi_2}^{\pi_1} + 1)V_{\pi_2}^{\pi_1} \), and \( \delta \leq 0.5 \) for some \( \delta \leq 0.5 \) and for all \( k \in \{1, 2, 3, \ldots\} \), Algorithm 2 is a safe policy selection algorithm with respect to \( \succ_{MC,\pi_b} \) with probability \( 1 - \delta_{MC} \). Recall that Algorithm 2 receives as input \( \tau_1, \ldots, \tau_n \sim \pi_b \). Consider additional random samples \( \tau_1^{\pi_1}, \ldots, \tau_n^{\pi_1} \sim \pi_1 \) and \( \tau_1^{\pi_2}, \ldots, \tau_n^{\pi_2} \sim \pi_2 \). Note that these samples are all independent from each other. For \( i \in \{1, 2, \ldots, n\} \), let

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

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

**Theorem 6.6.** There exists policies $\pi_1$, $\pi_2$, and behavior policy $\pi_b$ for which Algorithm 2 with inputs as described in Theorem 6.4 is not a safe policy selection algorithm with respect to $\succ_{MC}$ with $p = 0.5$ when we have a single sample drawn from $\pi_b$.

**Proof of Theorem 6.6.** Consider a world where there are three trajectories: $\tau_1$ with reward 0.0001, $\tau_2$ with reward 0.0002, and $\tau_3$ with reward 1. We want to select between two policies: $\pi_1$, which places probability 1 on $\tau_2$ and $\pi_2$ which places probability 0.51 on $\tau_1$ and probability 0.49 on $\tau_3$. When we only have one sample from each policy, $\Pr(\hat{V}_{\pi_1} > \hat{V}_{\pi_2}) = 0.51 > 0.5$, but clearly $V_{\pi_1} \ll V_{\pi_2}$. Now consider using IS with behavior policy $\pi_b$ which places probability 0.48 on $\tau_1$ and probability 0.01 on $\tau_2$ and probability 0.51 on $\tau_3$. If we apply Algorithm 2 with the inputs to guarantee that the algorithm is safe with respect $\succ_{\nu}$ (as given in Theorem 6.4), we find that whenever $\pi_b$ samples from $\tau_3$,

\[
V_{\pi_1}^{IS} - V_{\pi_2}^{IS} + (u_{\pi_1} V_{\pi_1}^{IS} + u_{\pi_2} V_{\pi_2}^{IS}) \sqrt{\frac{\ln 4}{2n}} = 0(1) - 0.49(0.01) + 0.51(0.0002) \approx -0.060 < 0
\]

Since this event occurs with probability 0.51, we find that Algorithm 2 returns $\pi_2$ more than half the time, indicating that Algorithm 2 is not a safe policy with respect to $\succ_{MC}$. 

\[\square\]