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. $\square$

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

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

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:

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

Note that

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

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

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

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 $k \in \{1, 2, 3, \ldots\}$, 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 $\Pr\left(|\hat{V}^{\pi_1}_{MC,kn} - \hat{V}^{\pi_2}_{MC,kn}| \geq \epsilon\right) \geq 1 - \delta$ 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 $\tau_1, \ldots, \tau_n \sim \pi_b$. Consider additional random samples $\tau_1^{\pi_1}, \ldots, \tau_{kn}^{\pi_1} \sim \pi_1$ and $\tau_1^{\pi_2}, \ldots, \tau_{kn}^{\pi_2} \sim \pi_2$. Note that these samples are all independent from each other. For $i \in \{1, 2, \ldots, n\}$, let

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

(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 $\hat{V}^{\pi}_{\text{IS},i} = \sum_{t=1}^{T_i} w_{i,t}R_{i,t}$. Now let

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

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

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

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

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

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 $\Pr(|\hat{V}^{\pi_1}_{\text{MC},kn} - \hat{V}^{\pi_1}_{\text{MC},kn}| > 0) \geq 0.75$ $\square$

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
    $\pi' \leftarrow$ ($\Pi\backslash$CurrBeat).next
    winner $\leftarrow$ `FPS`(π^*, π', $V^{\pi^*}_{\text{Max}}$, $V^{\pi'}_{\text{Max}}$, ϵ, δ, $\mathcal{T}$)
    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
      $\pi^* \leftarrow$ ($\Pi\backslash$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 $\Pr(|\hat{V}_{\text{MC}}^{\pi_1} - \hat{V}_{\text{MC}}^{\pi_1}| > 0) \geq 0.75$ $\square$

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,

$$V_{\text{IS}}^{\pi_1} - V_{\text{IS}}^{\pi_2} + (w_{\text{Max}}^{\pi_1} V_{\text{Max}}^{\pi_1} + w_{\text{Max}}^{\pi_2} V_{\text{Max}}^{\pi_2})\sqrt{\frac{\ln 4}{2n}}$$

$$= 0(1) - \frac{0.49}{0.51}(1) + \left(\frac{1}{0.01}(0.0002) + \frac{0.51}{0.48}(1)\right)\sqrt{\frac{\ln 4}{2}}$$

$$\approx -0.060 < 0$$

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 $\succ_{\text{MC},1}$ $\square$