## A. Details of Couterexamples

In this section we provide details of computing the variance in Figure 1. For each MDP, there are totally four possible trajectories (product of two actions and two steps), and the probabilities of them under behavior policy are all $1/4$. We list the return of different estimators for those four trajectories, then compute the variance of the estimators.

| | Probabilities of path | Example 1a IS | PDIS | MIS | Example 1b IS | PDIS | MIS | Example 1c IS | PDIS | MIS |
|---|---|---|---|---|---|---|---|---|---|---|
| $a_1, a_1$ | 0.25 | 1.44 | 1.2 | 1.2 | 0 | 0 | 0 | 0 | 0 | 0 |
| $a_1, a_2$ | 0.25 | 1.92 | 2.16 | 2.0 | 1.44 | 1.44 | 1.2 | 0.96 | 0.96 | 0.8 |
| $a_2, a_1$ | 0.25 | 0.96 | 0.8 | 0.8 | 0.64 | 0.8 | 0.8 | 0.96 | 0.8 | 0.8 |
| $a_2, a_2$ | 0.25 | 1.28 | 1.44 | 1.6 | 1.92 | 1.76 | 2.0 | 1.28 | 1.44 | 1.6 |
| Expectation | | 1.4 | 1.4 | 1.4 | 1 | 1 | 1 | 0.8 | 0.8 | 0.8 |
| Variance | | 0.12 | 0.2448 | 0.2 | 0.5424 | 0.4528 | 0.52 | 0.2304 | 0.2688 | 0.32 |

Table 1: Importance sampling returns and the variance. See figure 1 for the problem structure.

## B. Proof of Lemma 1

*Proof.* In this proof, we use $\tau$ to denote the trajectory without reward: $\tau_{1:t} = \{s_k, a_k\}_{k=1}^{t}$. Since $\mathbb{E}(\rho_{1:t}|s_t, a_t) = \mathbb{E}(\rho_{1:t-1}|s_t, a_t)\rho_t$, we only need to prove that $\mathbb{E}(\rho_{1:t-1}|s_t, a_t) = \frac{d^\pi(s_t)}{d^\mu(s_t)}$.

$$\begin{aligned}
\mathbb{E}(\rho_{1:t-1}|s_t, a_t) &= \int \prod_{k=1}^{t-1} \frac{\pi(s_k, a_k)}{\mu(s_k, a_k)} p_\mu(\tau_{1:t-1}|s_t, a_t) \mathrm{d}\tau_{1:t-1} && (1)\\
&= \int \frac{p_\pi(\tau_{1:t-1})}{p_\mu(\tau_{1:t-1})} p_\mu(\tau_{1:t-1}|s_t, a_t) \mathrm{d}\tau_{1:t-1} && (2)\\
&= \int \frac{p_\pi(\tau_{1:t-1})}{p_\mu(\tau_{1:t-1})} \frac{p_\mu(\tau_{1:t-1}) p(s_t|\tau_{1:t-1}) \mu(a_t|s_t)}{p_\mu(s_t, a_t)} \mathrm{d}\tau_{1:t-1} && (3)\\
&= \int \frac{p_\pi(\tau_{1:t-1})}{p_\mu(\tau_{1:t-1})} \frac{p_\mu(\tau_{1:t-1}) p(s_t|s_{t-1}, a_{t-1}) \mu(a_t|s_t)}{d_t^\mu(s_t)\mu(a_t|s_t)} \mathrm{d}\tau_{1:t-1} && (4)\\
&= \frac{1}{d_t^\mu(s_t)} \int p(s_t|s_{t-1}, a_{t-1}) p_\pi(\tau_{1:t-1}) \mathrm{d}\tau_{1:t-1} && (5)\\
&= \frac{1}{d_t^\mu(s_t)} \int p(s_t|\tau_{1:t-1}) p_\pi(\tau_{1:t-1}) \mathrm{d}\tau_{1:t-1} && (6)\\
&= \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} && (7)
\end{aligned}$$

□

## C. Proofs for Finite Horizon Case

### C.1. Proof of Lemma 2

*Proof.* Since $\mathbb{E}\left(\sum_t \mathbb{E}(Y_t|X_t)\right) = \mathbb{E}\left(\sum_t Y_t\right)$, we just need to compute the difference between the second moment of $\sum_t Y_t$ and $\sum_t \mathbb{E}(Y_t|X_t)$:

$$\begin{aligned}
\mathbb{E}\left(\sum_t \mathbb{E}(Y_t|X_t)\right)^2 =& \mathbb{E}\left(\sum_t \left(\mathbb{E}(Y_t|X_t)\right)^2 + 2\sum_{t<k} \mathbb{E}(Y_t|X_t)\mathbb{E}(Y_k|X_k)\right) && (8)\\
=& \sum_t \mathbb{E}\left(\mathbb{E}(Y_t|X_t)\right)^2 + 2\sum_{t<k} \mathbb{E}(\mathbb{E}(Y_t|X_t)\mathbb{E}(Y_k|X_k)) && (9)\\
\leq& \sum_t \mathbb{E}\left(\mathbb{E}(Y_t^2|X_t)\right) + 2\sum_{t<k} \mathbb{E}(\mathbb{E}(Y_t|X_t)\mathbb{E}(Y_k|X_k)) && (10)\\
=& \sum_t \mathbb{E}(Y_t^2) + 2\sum_{t<k} \mathbb{E}(\mathbb{E}(Y_t|X_t)\mathbb{E}(Y_k|X_k)) && (11)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(\sum_t Y_t\right)^2 =& \mathbb{E}\left(\sum_t Y_t^2 + 2\sum_{t<k} Y_t Y_k\right) && (12)\\
=& \sum_t \mathbb{E}(Y_t^2) + 2\sum_{t<k} \mathbb{E}(Y_t Y_k) && (13)
\end{aligned}$$

Thus we finished the proof by taking the difference between $\mathbb{E}\left(\sum_t Y_t\right)^2$ and $\mathbb{E}\left(\sum_t \mathbb{E}(Y_t|X_t)\right)^2$. $\square$

### C.2. Proof of Theorem 1

*Proof.* Let $\tau_{1:t}$ be the first $t$steps in a trajectory: $(s_1, a_1, r_1, \ldots, s_t, a_t, r_t)$, then $\rho_{1:t} r_t = \mathbb{E}(\rho_{1:T} r_t|\tau_{1:t})$. To prove the inequality between the variance of importance sampling and per decision importance sampling, we apply Lemma 2 to the variance, letting $Y_t = r_t\rho_{1:T}$ and $X_t = \tau_{1:t}$. Then it is sufficient to show that for any $1 \leq t < k \leq T$,

$$\mathbb{E}(r_t r_k \rho_{1:T}\rho_{1:T}) = \mathbb{E}(Y_t Y_k) \geq \mathbb{E}(\mathbb{E}(Y_t|X_t)\mathbb{E}(Y_k|X_k)) = \mathbb{E}(r_t r_k \rho_{1:t}\rho_{1:k}) \tag{14}$$

To prove that, it is sufficient to show $\mathbb{E}(r_t r_k \rho_{1:T}\rho_{1:T}|\tau_{1:t}) \geq \mathbb{E}(r_t r_k \rho_{1:t}\rho_{1:k}|\tau_{1:t})$. Since

$$\begin{aligned}
\mathbb{E}\left(r_t r_k \rho_{1:t}\rho_{1:k}|\tau_{1:t}\right) &= r_t\rho_{1:t}^2 \mathbb{E}\left(r_k\rho_{t+1:k}|\tau_{1:t}\right) && (15)\\
&= r_t\rho_{1:t}^2 \mathbb{E}\left(r_k\rho_{t+1:T}|\tau_{1:t}\right) && (16)\\
&= r_t\rho_{1:t}^2 \mathbb{E}\left(r_k\rho_{t+1:T}|\tau_{1:t}\right)\mathbb{E}\left(\rho_{t+1:T}|\tau_{1:t}\right) && (17)\\
& && (18)
\end{aligned}$$

Given $\tau_{1:t}$, $r_k$ and $\rho_{t+1:T}$ can be viewed as $r_{k-t+1}$ and $\rho_{1:T-t+1}$ on a new trajectory. Then according to the statement of theorem, $r_{k-t+1}\rho_{1:T-t+1}$ and $\rho_{1:T-t+1}$ are are positively correlated. Now we can upper bound $\mathbb{E}\left(r_t r_k \rho_{1:t}\rho_{1:k}|\tau_{1:t}\right)$ by:

$$\begin{aligned}
r_t\rho_{1:t}^2 \mathbb{E}\left(r_k\rho_{t+1:T}|\tau_{1:t}\right)\mathbb{E}\left(\rho_{t+1:T}|\tau_{1:t}\right) &\leq r_t\rho_{1:t}^2 \mathbb{E}\left(r_k\rho_{t+1:T}\rho_{t+1:T}|\tau_{1:t}\right) && (19)\\
&= \mathbb{E}\left(r_k r_t\rho_{1:T}\rho_{1:T}|\tau_{1:t}\right) && (20)
\end{aligned}$$

This implies $\mathbb{E}(r_t r_k \rho_{1:T}\rho_{1:T}) \geq \mathbb{E}(r_t r_k \rho_{1:t}\rho_{1:k})$ by taking expectation over $\tau_{1:t}$, and finish the proof. $\square$

### C.3. Proof of Theorem 2

*Proof.* Using lemma 2 by $Y_t = \rho_{1:t} r_t$ and $X_t = s_t, a_t, r_t$ , we have that the variance of $\hat{v}_{\text{SIS}}$ is smaller than the variance of $\hat{v}_{\text{PDIS}}$ if for any $t < k$:

$$\begin{aligned}
\mathbb{E}\left[\rho_{1:t}\rho_{0:k} r_t r_k\right] \geq& \mathbb{E}\left[\mathbb{E}(\rho_{1:t}|s_t, a_t)\mathbb{E}(\rho_{0:k}|s_k, a_k) r_t r_k\right] && (21)\\
=& \mathbb{E}\left[\frac{d_t^\pi(s,a)}{d_t^\mu(s,a)}\frac{d_k^\pi(s,a)}{d_k^\mu(s,a)} r_t r_k\right] && (22)
\end{aligned}$$

The second line follows from Lemma 1 to simplify $\mathbb{E}(\rho_{0:t}|s_t, a_t)$. To show that, we will transform the above equation into a an expression about two covariances. To proceed we subtracting $\mathbb{E}(\rho_{1:t}r_t)\mathbb{E}(\rho_{1:k}r_k)$ from both sides, and note that the resulting left hand side is simply the covariance:

$$\begin{aligned}\mathrm{Cov}\left[\rho_{1:t}r_t, \rho_{0:k}r_k\right] =&\mathbb{E}\left[\rho_{1:t}\rho_{1:k}r_t r_k\right] - \mathbb{E}(\rho_{1:t}r_t)\mathbb{E}(\rho_{1:k}r_k)\\ \geq&\mathbb{E}\left[\frac{d_t^\pi(s,a)}{d_t^\mu(s,a)}\frac{d_k^\pi(s,a)}{d_k^\mu(s,a)}r_t r_k\right] - \mathbb{E}(\rho_{1:t}r_t)\mathbb{E}(\rho_{1:k}r_k)\end{aligned} \tag{23}$$

We now expand the second term in the right hand side

$$\mathbb{E}(\rho_{1:t}r_t)\mathbb{E}(\rho_{1:k}r_k) =\mathbb{E}(r_t\mathbb{E}(\rho_{1:t}|s_t,a_t))\mathbb{E}(r_k\mathbb{E}(\rho_{1:k}|s_k,a_k)) \tag{24}$$

$$=\mathbb{E}\left[\frac{d_t^\pi(s,a)}{d_t^\mu(s,a)}r_t\right]\mathbb{E}\left[\frac{d_k^\pi(s,a)}{d_k^\mu(s,a)}r_k\right] \tag{25}$$

This shows that both sides of 23 are covariances. The result then follows under the assumption of the proof. $\square$

## D. Proofs for infinite horizon case

### D.1. Proof of Theorem 3

*Proof.* We can write the log of likelihood ratio as sum of random variables on a Markov chain,

$$\log\rho_{1:T} = \sum_{t=1}^{T}\log\rho_t = \sum_{t=1}^{T}\log\left(\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\right) \tag{26}$$

By the strong law of large number on Markov chain (Breiman, 1960):

$$\frac{1}{T}\log\rho_{1:T} = \frac{1}{T}\sum_{i=1}^{T}\log\left(\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}\right) \to_{a.s.} \mathbb{E}_{d^\mu}\log\left(\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}\right) = -c \tag{27}$$

If $\pi \neq \mu$, the strict concavity of $\log$ function implies that:

$$c = \mathbb{E}_{d^\mu}\log\left(\frac{\pi(a|s)}{\mu(a|s)}\right) < \log\mathbb{E}_{d^\mu}\left(\frac{\pi(a|s)}{\mu(a|s)}\right) = 0 \tag{28}$$

Thus $\frac{1}{T}\log\rho_{1:T} \to_{a.s.} c$ and $\rho_{1:T}^{1/T} \to_{a.s.} e^{-c}$. Since $r_t \leq 1$, $|\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t|^{1/T} \leq \rho_{1:T}^{1/T}T^{1/T}$. Since $T^{1/T} \to 1$, $\overline{\lim}_{T\to\infty}|\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t|^{1/T} < e^{-c}$. $\square$

### D.2. Proof of Corollary 1

*Proof.* $\rho_{1:T} \to_{a.s.} 0$ directly follows from $\rho_{1:T}^{1/T} \to_{a.s.} e^{-c}$ in Theorem 3. For $\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t$, if there exist $\epsilon > 0$ such that $\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t > \epsilon$ for any T, then:

$$\overline{\lim}_{T\to\infty}\left|\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t\right|^{1/T} \geq \overline{\lim}_{T\to\infty}\epsilon^{1/T} = 1$$

This contradicts $e^{-c} > \overline{\lim}_T|\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t|^{1/T}$ So $\overline{\lim}_T\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t \leq 0$, which implies that $\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t \to_{a.s.} 0$ $\square$

### D.3. Proof of Lemma 3

*Proof.* Let $f(s,a) = \log\frac{\pi(s,a)}{\mu(s,a)}$. According to Assumption 3, $|f(s,a)| < \infty$. Since $B(s,a) \geq 1$, $\frac{|f(s,a)|}{\sqrt{B(s,a)}} < \infty$. Since $f^2$ and $B$ are both finite, $\mathbb{E}_{d^\mu}f^2 < \infty$ and $\mathbb{E}_{d^\mu}B < \infty$. Now we satisfy the condition of Lemma 3 in (Glynn and Olvera-Cravioto, 2019): in the proof of Lemma 3 in in (Glynn and Olvera-Cravioto, 2019) they used their Assumption

i) Harris Chain, which is our Assumption 1, their Assumption vii) $||f||_s qrtV$ bounded (whic is satisfied by our bound on B in Assumption 2), which is explained by f is bounded and $\sqrt{B} >= 1$, and finally their assumption iv), which is our assumption 2. The only difference is we assume a "petite" K which is a slight generalization of the "small" set K (See discussion in (Meyn and Tweedie, 2012, Section 5)). The proof in Meyn and Glynn 1996 also used petite (which is the part where Glynn and Olvera-Cravioto need assumption iv)). This assumption (drift condition) is often necessary for quantitative analysis of general state Markov Chains. The geometric ergodicity for general state MC is also defined with a petite/small set. By Thm 15.0.1 in Meyn and Tweedie the drift property is equivalent to geometric ergodicity. According to Lemma 3 in (Glynn and Olvera-Cravioto, 2019), whose proof is similar with Theorem 2.3 in (Glynn et al., 1996), we have that there exist a solution $\hat{f}$ to the following Poisson's equation:

$$\hat{f}(s,a) - \mathbb{E}_{\cdot|s,a}\hat{f}(s',a') = f(s,a) - \mathbb{E}_{d^\mu} f(s,a) \tag{29}$$

satisfying $|\hat{f}(s,a)| < c_1\sqrt{B(s,a)}$ for some constant $c_1$. Following from the Poisson's equation we have:

$$\begin{aligned}
\log\rho_{1:T} + Tc =& \sum_{t=1}^{T}\left(f(s_t,a_t) - \mathbb{E}_{d^\mu} f(s,a)\right) && (30)\\
=& \sum_{t=1}^{T}\left(\hat{f}(s_t,a_t) - \mathbb{E}_{s',a'|s_t,a_t}\hat{f}(s',a')\right) && (31)\\
=& \hat{f}(s_1,a_1) - \hat{f}(s_{T+1},a_{T+1}) + \sum_{t=2}^{T+1}\left(\hat{f}(s_t,a_t) - \mathbb{E}_{s',a'|s_{t-1},a_{t-1}}\hat{f}(s',a')\right) && (32)
\end{aligned}$$

$\left(\hat{f}(s_t,a_t) - \mathbb{E}_{s',a'|s_{t-1},a_{t-1}}\hat{f}(s',a')\right)$ are martingale differences. The absolute value of difference is upper bounded by $2\|\hat{f}\|_\infty \le 2c_1\sqrt{\|B\|_\infty}$. □

### D.4. Proof of Theorem 4

**Lemma 1.** *If $\mathbb{E}_\mu[\rho^2|s] \le M_\rho^2$ for any s, $\mathbb{E}\left[\rho^2_{0:k}\right] \le M_\rho^{2k}$*

*Proof.*

$$\begin{aligned}
\mathbb{E}\left[\rho^2_{0:t}\right] &= \mathbb{E}\left[\prod_{i=1}^{k}\rho_i^2\right] && (33)\\
&= \mathbb{E}\left[\left(\prod_{i=1}^{k-1}\rho_i^2\right)\mathbb{E}_{s_k,a_k}\left[\rho_k^2|s_1,a_1,s_2,\dots,s_{k-1},a_{k-1}\right]\right] && (34)\\
&\le \mathbb{E}\left[\left(\prod_{i=1}^{k-1}\rho_i^2\right)M_\rho^2\right] && (35)\\
&= M_\rho^2\mathbb{E}\left[\prod_{i=1}^{k-1}\rho_i^2\right] && (36)\\
&\dots && (37)\\
&= M_\rho^{2k} && (38)
\end{aligned}$$

□

*Proof.* Define $Y = \rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t$ and $Z = \mathbb{1}(Y > v^\pi/2)$, then $v^\pi = \mathbb{E}(Y)$. By the law of total variance,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\mathbb{E}(Y|Z)) + \mathbb{E}(\mathrm{Var}(Y|Z)) && (39)\\
&\ge \mathrm{Var}(\mathbb{E}(Y|Z)) && (40)\\
&= \mathbb{E}(\mathbb{E}(Y|Z))^2 - (v^\pi)^2 && (41)\\
&\ge \Pr(Y > v^\pi/2)(\mathbb{E}(Y|Y > v^\pi/2))^2 - (v^\pi)^2 && (42)
\end{aligned}$$

Now we are going to lower bound $\mathbb{E}(Y|Y > v^\pi/2)$. We can rewrite $\mathbb{E}(Y) = v^\pi$ as:

$$
\begin{aligned}
v^\pi =& \mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|Z)) && (43)\\
=& \Pr(Y > v^\pi/2)\mathbb{E}(Y|Y > v^\pi/2) + \Pr(Y \le v^\pi/2)\mathbb{E}(Y|Y \le v^\pi/2) && (44)\\
\le& \Pr(Y > v^\pi/2)\mathbb{E}(Y|Y > v^\pi/2) + 1 \times v^\pi/2 && (45)
\end{aligned}
$$

So $\mathbb{E}(Y|Y > v^\pi/2) \ge \frac{v^\pi}{2\Pr(Y > v^\pi/2)}$. Substitute this into the RHS of Equation 42:

$$
\mathrm{Var}(Y) \ge \frac{(v^\pi)^2}{4\Pr(Y > v^\pi/2)} - (v^\pi)^2 \tag{46}
$$

Now we are going to upper bound $\Pr(Y > v^\pi/2)$. Recall that we define $c = \mathbb{E}_{d^\mu} D_{\mathrm{KL}}(\mu||\pi) = -\mathbb{E}_{d^\mu} \log\left(\frac{\pi(a|s)}{\mu(a|s)}\right)$. Now we define $c(T) = -\mathbb{E}_{d^\mu_{1:T}} \log\left(\frac{\pi(a|s)}{\mu(a|s)}\right) = -\frac{1}{T}\mathbb{E}_\mu[\log \rho_{1:T}]$.

$$
\begin{aligned}
&\Pr(Y > v^\pi/2) && (47)\\
=& \Pr(\rho_{1:T}\sum_{t=1}^{T}\gamma^{t-1}r_t > v^\pi/2) \le \Pr(\rho_{1:T}T > v^\pi/2) && (48)\\
=& \Pr\left(\rho_{1:T} > \frac{v^\pi}{2T}\right) && (49)\\
=& \Pr\left(\log\rho_{1:T} > \log v^\pi - \log(2T)\right) && (50)\\
=& \Pr\left(\frac{\log\rho_{1:T}}{T} > \frac{\log v^\pi - \log(2T)}{T}\right) && (51)\\
=& \Pr\left(\frac{\log\rho_{1:T}}{T} + c + \frac{\hat{f}(s_{T+1}, a_{T+1}) - \hat{f}(s_1, a_1)}{T} > c + \frac{\log v^\pi - \log(2T) + \hat{f}(s_{T+1}, a_{T+1}) - \hat{f}(s_1, a_1)}{T}\right) && (52)
\end{aligned}
$$

Since $\log v^\pi$ is a constant, $\hat{f}(s_{T+1}, a_{T+1}) - \hat{f}(s_1, a_1)$ could be upper bounded by constant $2c_1\sqrt{\|B\|_\infty}$, and $\lim_{T\to\infty}\frac{\log(2T)}{T} = 0$, we know that $\lim_{T\to\infty}\frac{\log v^\pi - \log(2T) + \hat{f}(s_{T+1}, a_{T+1}) - \hat{f}(s_1, a_1)}{T} = 0$. So there exists a constant $T_0 > 0$ such that for all $T > T_0$,

$$
\frac{\log v^\pi - \log(2T) + \hat{f}(s_{T+1}, a_{T+1}) - \hat{f}(s_1, a_1)}{T} > -\frac{c}{2}
$$

Therefore for all $T > T_0$:

$$
\Pr(Y > v^\pi/2) \le \Pr\left(\frac{\log\rho_{1:T}}{T} + c + \frac{\hat{f}(s_{T+1}, a_{T+1}) - \hat{f}(s_1, a_1)}{T} > c/2\right)
$$

According to Lemma 3, and Azuma's inequality(Azuma, 1967), we have:

$$
\Pr(Y > v^\pi/2) \le \exp\left(\frac{-Tc^2}{8c_1^2\|B\|_\infty}\right)
$$

Thus we can lower bound the variance of importance sampling estimator $Y$:

$$
\mathrm{Var}(Y) \ge \frac{(v^\pi)^2}{4}\exp\left(\frac{Tc^2}{8c_1^2\|B\|_\infty}\right) - (v^\pi)^2 \tag{53}
$$

If the one step likelihood ratio is upper bounded by $U_\rho$, then the variance of importance sampling estimator can be upper bounded by:

$$
\begin{aligned}
\mathrm{Var}(\hat{v}_{\mathrm{IS}}) = \mathbb{E}[Y^2] - (v^\pi)^2 =& \mathbb{E}\left[\rho_{0:T}^2\left(\sum_{t=1}^{T}\gamma^{t-1}r_t\right)^2\right] - (v^\pi)^2 && (54)\\
\le& T^2\mathbb{E}\left[\rho_{0:T}^2\right] - (v^\pi)^2 && (55)\\
\le& T^2U_\rho^{2T} - (v^\pi)^2 && (56)
\end{aligned}
$$

Following from lemma 1, the variance term can also be upper bounded by:

$$\begin{aligned}\mathrm{Var}(\hat{v}_{\text{IS}}) = \mathbb{E}[Y^2] - (v^\pi)^2 = \mathbb{E}\left[\rho_{0:T}^2\left(\sum_{t=1}^{T}\gamma^{t-1}r_t\right)^2\right] - (v^\pi)^2 &\quad (57)\\ \leq T^2\mathbb{E}\left[\rho_{0:T}^2\right] - (v^\pi)^2 &\quad (58)\\ \leq T^2 M_\rho^{2T} - (v^\pi)^2 &\quad (59)\end{aligned}$$

□

### D.5. Proof of Theorem 5

*Proof.* Let $Y_t = \rho_{1:t}\gamma^{t-1}r_t$. For the upper bound:

$$\begin{aligned}\mathrm{Var}(\hat{v}_{\text{PDIS}}) =&\mathbb{E}\left(\left(\sum_{t=1}^{T}Y_t\right)^2\right) - (v^\pi)^2 &(60)\\ \leq&\mathbb{E}\left(T\sum_{t=1}^{T}Y_t^2\right) - (v^\pi)^2 &(61)\\ =&T\sum_{t=1}^{T}\mathbb{E}(Y_t^2) - (v^\pi)^2 &(62)\\ =&T\sum_{t=1}^{T}\mathbb{E}(\rho_{0:t}^2\gamma^{2t-2}r_t^2) - (v^\pi)^2 &(63)\\ \leq&T\sum_{t=1}^{T}U_\rho^{2t}\gamma^{2t-2}\mathbb{E}_\mu[(r_t)^2] - (v^\pi)^2 &(64)\end{aligned}$$

Or it can also be bounded as:

$$\begin{aligned}\mathrm{Var}(\hat{v}_{\text{PDIS}}) \leq&T\sum_{t=1}^{T}\mathbb{E}(\rho_{0:t}^2\gamma^{2t-2}r_t^2) - (v^\pi)^2 &(65)\\ =&T\sum_{t=1}^{T}\gamma^{2t-2}\mathbb{E}(\rho_{0:t}^2) - (v^\pi)^2 &(66)\\ \leq&T\sum_{t=1}^{T}\gamma^{2t-2}M_\rho^{2t} - (v^\pi)^2 &(67)\end{aligned}$$

The last step follows from lemma 1. For the lower bound, we notice that $Y_t \geq 0$ for any $t$, then:

$$\mathbb{E}\left(\left(\sum_{t=1}^{T}Y_t\right)^2\right) \geq \mathbb{E}\left(\sum_{t=0}^{T}Y_t^2\right) = \sum_{t=1}^{T}\mathbb{E}(Y_t^2) \quad (68)$$

For each $t$, we will follow a similar proof as how to lower bound part in Theorem 4:

$$\begin{aligned}\mathbb{E}(Y_t^2) =&\mathbb{E}\left(\mathbb{E}(Y_t^2|\mathbb{1}(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2))\right) &(69)\\ \geq&\mathbb{E}\left(\mathbb{E}(Y_t|\mathbb{1}(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2))\right)^2 &(70)\\ \geq&\Pr(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)\left(\mathbb{E}(Y_t|Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)\right)^2 &(71)\end{aligned}$$

Notice that $\mathbb{E}(Y_t) = \gamma^{t-1}\mathbb{E}_\pi(r_t)$,

$$\begin{aligned}&\gamma^{t-1}\mathbb{E}_\pi(r_t) = \mathbb{E}(Y_t) &(72)\\ &= \Pr(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)\mathbb{E}(Y_t|Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2) + \Pr(Y_t \leq \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)\mathbb{E}(Y_t|Y_t \leq \gamma^{t-1}\mathbb{E}_\pi(r_t)/2) &(73)\\ &\leq \Pr(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)\mathbb{E}(Y_t|Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2) + \gamma^{t-1}\mathbb{E}_\pi(r_t)/2 &(74)\end{aligned}$$

So we can lower bound the $\mathbb{E}(Y_t^2)$:

$$\mathbb{E}(Y_t | Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2) \geq \frac{\gamma^{t-1}\mathbb{E}_\pi(r_t)}{2\Pr(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)} \tag{75}$$

$$\mathbb{E}(Y_t^2) \geq \frac{\gamma^{2t-2}\left(\mathbb{E}_\pi(r_t)\right)^2}{4\Pr(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)} \tag{76}$$

Now we are going to upper bound the tail probability $\Pr(Y_t > \gamma^{t-1}\mathbb{E}_\pi(r_t)/2)$:

$$\Pr\left(Y_t | Y_t > \frac{\gamma^{t-1}\mathbb{E}_\pi(r_t)}{2}\right) \tag{77}$$

$$= \Pr\left(\rho_{1:t}\gamma^{t-1}r_t > \frac{\gamma^{t-1}\mathbb{E}_\pi(r_t)}{2}\right) \tag{78}$$

$$\leq \Pr\left(\rho_{1:t} > \frac{\mathbb{E}_\pi(r_t)}{2}\right) \tag{79}$$

$$= \Pr\left(\log\rho_{1:t} > \log\mathbb{E}_\pi(r_t) - \log 2\right) \tag{80}$$

$$= \Pr\left(\frac{1}{t}\log\rho_{1:t} > \frac{\mathbb{E}_\pi(r_t) - \log 2}{t}\right) \tag{81}$$

$$= \Pr\left(\frac{1}{t}\log\rho_{1:t} + c + \frac{\hat{f}(s_{t+1}, a_{t+1}) - \hat{f}(s_1, a_1)}{T} > c + \frac{\mathbb{E}_\pi(r_t) - \log 2 + \hat{f}(s_{t+1}, a_{t+1}) - \hat{f}(s_1, a_1)}{t}\right) \tag{82}$$

Since $|\mathbb{E}_\pi(r_t) - \log 2 + \hat{f}(s_{t+1}, a_{t+1}) - \hat{f}(s_1, a_1)|$ is bounded, there exist some $T_0 > 0$ such that if $t > T_0$, we can lower bound the right hand side in the probability by $c/2$. Then for $t > T_0$, by Azuma's inequality (Azuma, 1967),

$$\Pr\left(Y_t | Y_t > \frac{\gamma^{t-1}\mathbb{E}_\pi(r_t)}{2}\right) \leq \Pr\left(\frac{\log\rho_{1:t}}{t} + c + \frac{\hat{f}(s_{t+1}, a_{t+1}) - \hat{f}(s_1, a_1)}{t} > \frac{c}{2}\right) \tag{83}$$

$$\leq \exp\left(\frac{-tc^2}{8c_1^2\|B\|_\infty}\right) \tag{84}$$

So we have that for $t > T_0$:

$$\mathbb{E}(Y_t^2) \geq \frac{\gamma^{2t-2}\mathbb{E}_\pi(r_t)}{4}\exp\left(\frac{tc^2}{8c_1^2\|B\|_\infty}\right)$$

For $0 < t \leq T_0$, $\mathbb{E}(Y_t^2) \geq 0$ completes the proof. □

### D.6. Proof of Corollary 2

*Proof.* First, $\gamma \geq \exp\left(\frac{-c^2}{16c_1^2\|B\|_\infty}\right)$ indicate $\left(\frac{c^2}{8c_1^2\|B\|_\infty} + 2\log\gamma\right) > 0$. This is necessary for the second condition to hold since $r_t < 1$. The second condition $\mathbb{E}_\pi(r_t) = \Omega\left(\exp\left(\frac{-tc^2}{8c_1^2\|B\|_\infty} - 2t\log\gamma + \epsilon t/2\right)\right)$ implies that there exist a $T_1 > 0$ and a constant $C > 0$ such that $(\mathbb{E}_\pi(r_t))^2 \geq C\left(\exp\left(\frac{-tc^2}{8c_1^2\|B\|_\infty} - 2t\log\gamma + \epsilon t\right)\right)$, for any $t > T_1$. Then let $T > \max\{T_1, T_0\}$, where $T_0$ is the constant in Theorem 5:

$$\mathrm{Var}(\sum_{t=T_0}^{T}\rho_{1_t}\gamma^{t-1}r_t) \geq \sum_{t=1}^{T}\frac{\gamma^{2t-2}(\mathbb{E}_\pi(r_t))^2}{4}\exp\left(\frac{tc^2}{8c_1^2\|B\|_\infty}\right) - (v^\pi)^2 \tag{85}$$

$$\geq \frac{\gamma^{2T-2}(\mathbb{E}_\pi(r_T))^2}{4}\exp\left(\frac{Tc^2}{8c_1^2\|B\|_\infty}\right) - (v^\pi)^2 \tag{86}$$

$$\geq \frac{\gamma^{-2}C}{4}\exp(\epsilon T) - (v^\pi)^2 = \Omega(\exp\epsilon T) \tag{87}$$

□

### D.7. Proof of Corollary 3

*Proof.* If $U_\rho\gamma \leq 1$, $U_\rho^t\gamma^{t-1}\mathbb{E}_\pi(r_t) \leq 1/\gamma$ for any $t$ since $r_t \in [0,1]$. If $U_\rho\gamma\lim\left(\mathbb{E}_\mu(r_T)\right)^{1/T} < 1$, let $\delta = 1 - U_\rho\gamma\lim\left(\mathbb{E}_\mu(r_T)\right)^{1/T} > 0$. There exist a $T_0 > 0$ such that for all $t > T_0$, $U_\rho\gamma(\mathbb{E}_\pi(r_t))^{1/t} \leq U_\rho\gamma(\lim\left(\mathbb{E}_\mu(r_T)\right)^{1/T} + \delta/2(U_\rho\gamma)) = 1 - \delta/2 < 1$. Therefore in both case, for all $T > T_0$, $U_\rho^t\gamma^{t-1}\mathbb{E}_\mu(r_T) \leq 1/\gamma$:

$$\mathrm{Var}(\sum_{t=1}^T \rho_{1:t}\gamma^{t-1}r_t) \leq T\sum_{t=1}^T U_\rho^t\gamma^{t-1}\mathbb{E}_\mu(r_T) \leq T\sum_{t=1}^{T_0} U_\rho^t\gamma^{t-1}\mathbb{E}_\mu(r_T) + T\sum_{t=T_0+1}^{T} U_\rho^t\gamma^{t-1}\mathbb{E}_\mu(r_T) \tag{88}$$

$$\leq TT_0\frac{U_\rho^{T_0}-1}{U_\rho-1} + 2T^2\frac{1}{\gamma} \tag{89}$$

Since $T_0$ is a constant, the variance is $O(T^2)$. □

### D.8. Proof of Theorem 6

*Proof.*

$$\mathrm{Var}\left(\sum_{t=1}^T \frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right) \tag{90}$$

$$=\sum_{t=1}^T \mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right) + 2\sum_{t<k}\mathrm{Cov}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t, \frac{d_k^\pi(s_k,a_k)}{d_k^\mu(s_k,a_k)}\gamma^{k-1}r_k\right) \tag{91}$$

$$\leq\sum_{t=1}^T \mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right) + \sum_{t<k}2\sqrt{\mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right)\mathrm{Var}\left(\frac{d_k^\pi(s_k,a_k)}{d_k^\mu(s_k,a_k)}\gamma^{k-1}r_k\right)} \tag{92}$$

$$\leq\sum_{t=1}^T \mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right) + \sum_{t<k}\left(\mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right) + \mathrm{Var}\left(\frac{d_k^\pi(s_k,a_k)}{d_k^\mu(s_k,a_k)}\gamma^{k-1}r_k\right)\right) \tag{93}$$

$$=T\sum_{t=1}^T \gamma^{2t-2}\mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}r_t\right) \tag{94}$$

$$\leq T\sum_{t=1}^T \gamma^{2t-2}\mathrm{Var}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right) \tag{95}$$

$$=T\sum_{t=1}^T \gamma^{2t-2}\left(\mathbb{E}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2 - 1\right) \tag{96}$$

□

### D.9. Proof of Corollary 4

**Lemma 4.** *If $d_t^\mu(s_t)$ and $d_t^\pi(s_t)$ are asymptotically equi-continuous, $\frac{d^\pi(s)}{d^\mu(s)} \leq U_s$, and $\frac{\pi(a|s)}{\mu(a|s)} \leq U_\rho$, then,*

$$\lim_t \mathbb{E}_{s_t,a_t\sim d_t^\mu}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2 = \mathbb{E}_{s,a\sim d^\mu}\left(\frac{d^\pi(s,a)}{d^\mu(s,a)}\right)^2$$

*Proof.* According to the law of large number on Markov chain (Breiman, 1960), the distribution of $d_t^\mu$ converge to the stationary distribution $d^\mu$ in distribution. By the Lemma 1 in (Boos et al., 1985), $d_t^\mu(s,a)$ converge to $d^\mu(s,a)$ pointwisely, $d_t^\pi(s,a)$ converge to $d^\pi(s,a)$ pointwisely. So $\frac{d_t^\pi(s)}{d_t^\mu(s)}$ converge to $\frac{d^\pi(s)}{d^\mu(s)}$ pointwisely.

$$\mathbb{E}_{s_t,a_t\sim d_t^\mu}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2-\mathbb{E}_{s,a\sim d^\mu}\left(\frac{d^\pi(s,a)}{d^\mu(s,a)}\right)^2 \tag{97}$$

$$=\int_{s,a}\frac{\left(d_t^\pi(s,a)\right)^2}{d_t^\mu(s,a)}\mathrm{d}s\mathrm{d}a-\int_{s,a}\frac{\left(d^\pi(s,a)\right)^2}{d^\mu(s,a)}\mathrm{d}s\mathrm{d}a \tag{98}$$

$$=\int_{s,a}\frac{\left(d_t^\pi(s)\right)^2}{d_t^\mu(s)}\frac{(\pi(a|s))^2}{\mu(a|s)}-\frac{\left(d^\pi(s)\right)^2}{d^\mu(s)}\frac{(\pi(a|s))^2}{\mu(a|s)}\mathrm{d}s\mathrm{d}a \tag{99}$$

$$\leq U_\rho\int_{s,a}\left|\frac{\left(d_t^\pi(s)\right)^2}{d_t^\mu(s)}-\frac{\left(d^\pi(s)\right)^2}{d^\mu(s)}\right|\mathrm{d}s\mathrm{d}a \tag{100}$$

$$\leq U_\rho\int_{s,a}\left|\frac{d^\pi(s)\left(d_t^\pi(s)-d^\pi(s)\right)}{d^\mu(s)}+d_t^\pi(s)\frac{d_t^\pi(s)}{d_t^\mu(s)}-d_t^\pi(s)\frac{d^\pi(s)}{d^\mu(s)}\right|\mathrm{d}s\mathrm{d}a \tag{101}$$

$$\leq U_\rho\int_{s,a}\left|\frac{d^\pi(s)\left(d_t^\pi(s)-d^\pi(s)\right)}{d^\mu(s)}\right|+d_t^\pi(s)\left|\frac{d_t^\pi(s)}{d_t^\mu(s)}-\frac{d^\pi(s)}{d^\mu(s)}\right|\mathrm{d}s\mathrm{d}a \tag{102}$$

$$\leq U_\rho U_s d_{\mathrm{TV}}(d_t^\pi,d^\pi)+U_\rho\int_{s,a}\left|\frac{d_t^\pi(s)}{d_t^\mu(s)}-\frac{d^\pi(s)}{d^\mu(s)}\right|\mathrm{d}s\mathrm{d}a \tag{103}$$

By the law of large number on Markov chain (Breiman, 1960), $d_{\mathrm{TV}}(d_t^\pi,d^\pi)\to 0$. Since $U_\rho$ and $U_s$ are constant, and $\frac{d_t^\pi(s)}{d_t^\mu(s)}\to\frac{d^\pi(s)}{d^\mu(s)}$, the right hand side of equation above converge to zero, which completes the proof. □

Proof of Corollary 4:

*Proof.* Since $\frac{d^\pi(s,a)}{d^\mu(s,a)}$ is bounded by $U_\rho U_s$ and then $\mathbb{E}_{s,a\sim d^\mu}\left(\frac{d^\pi(s,a)}{d^\mu(s,a)}\right)^2$ is bounded by $U_\rho^2 U_s^2$. Following from Lemma 4, there exist $T_0>0$ such that for all $t>T_0$, $\mathbb{E}_{s_t,a_t\sim d_t^\mu}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2\leq 2\mathbb{E}_{s,a\sim d^\mu}\left(\frac{d^\pi(s,a)}{d^\mu(s,a)}\right)^2\leq 2U_\rho^2U_s^2$. Then by Theorem 6, for $T>T_0$

$$\begin{aligned}\mathrm{Var}\left(\sum_{t=1}^T\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\gamma^{t-1}r_t\right)\leq&T\sum_{t=1}^T\gamma^{t-1}\mathbb{E}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2\\ \leq&T\sum_{t=1}^{T_0}\gamma^{t-1}\mathbb{E}\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2+2T(T-T_0)U_\rho^2U_s^2\\ =&O(T^2)\end{aligned}$$

□

### D.10. Proof of Corollary 5

Now we consider an type of approximate MIS estimators, which plug an approximate density ratio into the MIS estimator. More specifially, we consider it use a function $w_t(s_t,a_t)$ to approximate density ratio $\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}$, and construct the estimator as:

$$\hat{v}_{\text{ASIS}}=\sum_{t=1}^T w_t(s,a)\gamma^{t-1}r_t \tag{104}$$

This approximate MIS estimator is often biased based on the choice of $w_t(s,a)$, so we consider the upper bound of their mean square error with respect to $T$ and the error of the ratio estimator.

**Theorem 7.** *$\hat{v}_{ASIS}$ with $w_t$ such that where $\mathbb{E}_\mu\left(w_t(s_t,a_t)-\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2\leq\epsilon_w$*

$$MSE\left(\hat{v}_{ASIS}\right)\leq 2Var(\hat{v}_{SIS})+2T^2\epsilon_w \tag{105}$$

*Proof.*

$$\text{MSE}\left(\sum_{t=1}^{T} w_t(s_t,a_t)\gamma^{t-1}r_t\right) = \mathbb{E}\left(\sum_{t=1}^{T} w_t(s_t,a_t)\gamma^{t-1}r_t - v^{\pi}\right)^2 \tag{106}$$

$$= \mathbb{E}\left(\sum_{t=1}^{T} w_t(s_t,a_t)\gamma^{t-1}r_t - \hat{v}_{\text{SIS}} + \hat{v}_{\text{SIS}} - v^{\pi}\right)^2 \tag{107}$$

$$\leq 2\mathbb{E}\left(\sum_{t=1}^{T} w_t(s_t,a_t)\gamma^{t-1}r_t - \hat{v}_{\text{SIS}}\right)^2 + 2\mathbb{E}\left(\hat{v}_{\text{SIS}} - v^{\pi}\right)^2 \tag{108}$$

$$\leq 2\mathbb{E}\left(\sum_{t=1}^{T} \gamma^{t-1}r_t\left(w_t(s_t,a_t) - \frac{d_t^{\pi}(s_t,a_t)}{d_t^{\mu}(s_t,a_t)}\right)\right)^2 + 2\text{Var}(\hat{v}_{\text{SIS}}) \tag{109}$$

$$\leq 2T\sum_{t=1}^{T} \gamma^{2t-2}\mathbb{E}\left(w_t(s_t,a_t) - \frac{d_t^{\pi}(s_t,a_t)}{d_t^{\mu}(s_t,a_t)}\right)^2 + 2\text{Var}(\hat{v}_{\text{SIS}}) \tag{110}$$

$$\leq 2\text{Var}(\hat{v}_{\text{SIS}}) + 2T\sum_{t=1}^{T}\gamma^{2t-2}\epsilon_w \tag{111}$$

$$\leq 2\text{Var}(\hat{v}_{\text{SIS}}) + 2T^2\epsilon_w \tag{112}$$

□

Proof of Corollary 5:

*Proof.* By Theorem 7 we have that the MSE is bounded by

$$2\text{Var}(\hat{v}_{\text{SIS}}) + 2T^2\epsilon_w \tag{113}$$

According to Corollary 4:

$$2\text{Var}(\hat{v}_{\text{SIS}}) + 2T^2\epsilon_w = O(T^2) + 2T^2\epsilon_w = O(T^2(1+\epsilon_w)) \tag{114}$$

□

## E. Return-Conditional IS estimators

A natural extension of the conditional importance sampling estimators is to condition on the observed returns $G_t$. Precisely we examine the general conditional importance sampling estimator:

$$G_t\mathbb{E}\left[\rho_{1:t}|\phi_t\right] \quad , \tag{115}$$

and consider when $\phi_t = G_t$. An analytic expression for $\mathbb{E}\left[\rho_{1:t}|G_t\right]$ is not available, but we can model this as a regression problem to predict $\mathbb{E}\left[\rho_{1:t}|G_t\right]$ given an input $G_t$. A natural approach is to use ordinary least squares (OLS) estimator to estimate $\mathbb{E}\left[\rho_{1:t}|\phi_t\right]$ viewing $\phi_t$ (or any other statistics $G_t$) as an *input* and $\rho_{1:t}$ as an *output*. While tempting at first glance, we show that this approach produces exactly the same estimates of the expected return as that of the crude importance sampling estimator.

We start by considering the OLS problem associated with the conditional weights in which we want to find a $\widehat{\theta}$ such that $\phi_t^{\top}\widehat{\theta} \approx \mathbb{E}\left[\rho_{1:t}|\phi_t\right]$. Let $\Phi \in \mathbb{R}^{n\times 2}$ be the *design* matrix containing the observed returns $G_t^{(i)}$ after $t$ steps and $Y \in \mathbb{R}^n$ be the vector of importance ratios $\rho_{1:t}^{(i)}$ for each rollout $i$:

$$Y = \begin{bmatrix} \rho_t^{(0)} \\ \vdots \\ \rho_t^{(N)} \end{bmatrix}, \quad \Phi = \begin{bmatrix} G_t^{(0)} & 1 \\ \vdots & \vdots \\ G_t^{(N)} & 1 \end{bmatrix} \quad .$$

The OLS estimator for the return-conditional weights is then $\widehat{Y} = \Phi\widehat{\theta}$ and where $\widehat{\theta} \in \mathbb{R}^2$ is defined as:

$$\widehat{\theta} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y \ .$$

We can now use the approximate return-conditional weights to form a Monte Carlo estimate of the expected return under the target policy:

$$\hat{v}_{\text{RCIS}} \equiv \frac{1}{N} \sum_{i=0}^{N} G_t^{(i)} \widehat{Y}^{(i)} = \frac{1}{N} [1, 0] \Phi^\top \widehat{Y} \ , \tag{116}$$

where $\widehat{Y}^{(i)} = [G_t^{(i)}, 0]^\top \widehat{\theta}$ and the equality follows from the fact that $\Phi^\top Y = [\sum_{i=1}^n \rho_t^{(i)} G_t^{(i)}, \sum_{i=1}^n \rho_t^{(i)}]^\top$. Using this observation, we can also express the crude importance sampling estimator with the linear combination $\Phi^\top Y$, where $Y$ now consists of the *true* weights:

$$\hat{v}_{\text{IS}} \equiv \frac{1}{N} [1, 0] \Phi^\top Y \ . \tag{117}$$

Note that equation (116) differs from (117) only in the term $\widehat{Y} = \Phi\widehat{\theta} = \Phi \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y$ and upon closer inspection, we find that:

$$\Phi^\top \hat{e} = \Phi^\top Y - \Phi^\top \widehat{Y} = \Phi^\top \left(Y - \Phi \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y\right) = \Phi^\top \left(I - H\right) Y = \mathbf{0} \ ,$$

where $\hat{e}$ is residual vector $Y - \widehat{Y}$ and $H = \Phi \left(\Phi^\top \Phi\right)^{-1} \Phi^\top$ is the *hat* matrix. Hence, it follows that the estimate of the expected return made under the crude importance sampling estimator must be identical to the extended estimator which uses approximate return-conditional weights:

$$\hat{v}_{\text{is}} - \hat{v}_{\text{RCIS}} = \frac{1}{n}[1, 0]\Phi^\top Y - \frac{1}{n}[1, 0]\Phi^\top \widehat{Y} = \frac{1}{n}[1, 0] \left(\Phi^\top Y - \Phi^\top \widehat{Y}\right) = [1, 0]\mathbf{0} = 0 \ .$$

This analysis can be generalized to any conditional importance sampling estimator for which $G_t$ can be expressed as a linear combination of $\phi_t$. For example, rather than conditioning on the final return, we could condition on the return so far (the sum of returns to the present) and use $\phi_t = [r_1, r_2, ...r_t]$ with the coefficient vector $[1, 1, ...1, 0]$. Similarly, this negative result carries to reward-conditional weights if the immediate reward $r_t$ can be expressed as linear combination of $\phi_t$, including if $\phi_t$ is simply the immediate reward.

## References


Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19 (3):357–367, 1967.

Dennis D Boos et al. A converse to scheffe's theorem. *The Annals of Statistics*, 13(1):423–427, 1985.

Leo Breiman. The strong law of large numbers for a class of markov chains. *The Annals of Mathematical Statistics*, 31(3): 801–803, 1960.

Peter W. Glynn and Mariana Olvera-Cravioto. Likelihood ratio gradient estimation for steady-state parameters. *Stochastic Systems*, 9(2):83–100, June 2019.

Peter W Glynn, Sean P Meyn, et al. A liapounov bound for solutions of the poisson equation. *The Annals of Probability*, 24(2):916–931, 1996.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.