Appendix

A Concentration inequalities and other technical lemmas

Lemma A.1 (Bernsteins Inequality (Sridharan, 2002)). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \leq \xi$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[x_i]$, then for any $\epsilon > 0$ we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\sigma^2 + 2\xi\epsilon/3}}.$$

Lemma A.2 (Multiplicative Chernoff bound (Chernoff et al., 1952)). *Let X be a Binomial random variable with parameter p, n. For any $\delta > 0$, we have that*

$$\mathbb{P}[X < (1-\delta)pn] < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np}.$$

A slightly weaker bound that suffices for our propose is the following:

$$\mathbb{P}[X < (1-\delta)pn] < e^{-\frac{\delta^2 pn}{2}}.$$

B Related settings

Markov Decision Processes have a long history of associated research (Puterman, 1994; Sutton and Barto, 1998), but many theoretical problems in the basic tabular setting remain an active area of research as of today. We briefly review the other settings and connect them to our results.

Regret bound and sample complexity in the online setting. The bulk of existing work focuses on online learning, where the agent interacts with the MDP with the interests of identifying the optimal policy or minimizing the regret against the optimal policy. The optimal regret is obtained by (Azar et al., 2017) using a model-based approach which translates into a sample complexity bound of $O(H^3SA/\epsilon^2)$, which matches the lower bound of $\Omega(H^3SA/\epsilon^2)$(Azar et al., 2013). The method is however not "uniform PAC" where the state of the art sample complexity remains $O(H^4SA/\epsilon^2)$ (Dann et al., 2017). Model-free approaches that require a space constraint of $O(HSA)$ were studied by Jin et al. (2018) which implies a sample complexity bound of $O(H^4SA/\epsilon^2)$.

Sample complexity with a generative model. Another sequence of work assumes access to a generative model where one can sample from s_{t+1} and r_t given any s_t, a_t in time $O(1)$ (Kearns and Singh, 1999). Sidford et al. (2018) is the first that establishes the optimal sample complexity of $\tilde{\Theta}(H^3SA/\epsilon^2)$ under this setting (counting H generative model calls as one episode). Agarwal et al. (2019) establishes a similar results by estimating the parameters of the MDP model using maximum-likelihood estimation.

C Proof of the main theorem

To analyze the MSE upper bound $\mathbb{E}_\mu[(\widehat{v}^\pi_{\mathrm{TMIS}} - v^\pi)^2]$, we create a fictitious surrogate $\widetilde{v}^\pi_{\mathrm{TMIS}}$, which is an unbiased version of $\widehat{v}^\pi_{\mathrm{TMIS}}$. A few auxiliary lemmas are first presented and Bellman equations are used for deriving variance decomposition in a recursive way. Second order moment of marginalized state distribution $\widetilde{d}^\pi_t$ can then be bounded by analyzing its variance.

C.1 Fictitious tabular MIS estimator.

The fictitious estimator[9] $\widetilde{v}^\pi$ fills in the gap of state-action location (s_t, a_t) of the true estimator $\widehat{v}^\pi$ where $n_{s_t,a_t} = 0$. Specifically, it replaces every component in $\widehat{v}^\pi$ with a fictitious counterpart, *i.e.* $\widetilde{v}^\pi := \sum_{t=1}^{H} \langle \widetilde{d}^\pi_t, \widetilde{r}^\pi_t \rangle$,

[9] We replcace the notation of $\widetilde{v}^\pi_{\mathrm{TMIS}}$ with just $\widetilde{v}^\pi$ throughout the proof. $\widetilde{v}^\pi$ always denotes fictitious tabular MIS estimator.

with $\widetilde{d}_t^\pi = \widetilde{P}_t^\pi \widetilde{d}_{t-1}^\pi$ and $\widetilde{P}_t^\pi(s_t|s_{t-1}) = \sum_{a_{t-1}} \widetilde{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1})$, $\widetilde{r}_t^\pi(s_t) = \sum_{a_t} \widetilde{r}_t(s_t, a_t)\pi(a_t|s_t)$. In particular, let E_t denotes the event $\{n_{s_t,a_t} \geq n d_t^\mu(s_t, a_t)(1-\theta)\}$[10], then

$$\widetilde{r}_t(s_t, a_t) = \widehat{r}_t(s_t, a_t)\mathbf{1}(E_t) + r_t(s_t, a_t)\mathbf{1}(E_t^c)$$
$$\widetilde{P}_{t+1}(\cdot|s_t, a_t) = \widehat{P}_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t) + P_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t^c).$$

where $0 < \theta < 1$ is a parameter that we will choose later.

The name "fictitious" comes from the fact that $\widetilde{v}^\pi$ is *not implementable* using the data[11], but it creates a bridge between $\widehat{v}^\pi$ and v^π because of its unbiasedness, see Lemma C.5. Also, for simplicity of the proof, throughout the rest of the paper we denote: $\mathcal{D}_t := \left\{s_{1:t}^{(i)}, a_{1:t}^{(i)}, r_{1:t-1}^{(i)}\right\}_{i=1}^n$. Also, in the base case, we denote $\mathcal{D}_1 := \left\{s_1^{(i)}, a_1^{(i)}\right\}_{i=1}^n$ and that $r_t^\pi(s_t) := \mathbb{E}_\pi[r_t^{(1)}|s_t^{(1)} = s_t] = \sum_{a_t} \mathbb{E}[r_t^{(1)}|s_t^{(1)} = s_t, a_t^{(1)} = a_t]\pi(a_t|s_t) := \sum_{a_t} r_t(s_t, a_t)\pi(a_t|s_t)$. Then we have the following preliminary auxiliary lemmas.

Lemma C.1. *$\widetilde{d}_t^\pi$ and $\widetilde{r}_{t-1}^\pi$ are deterministic given $\mathcal{D}_t$. Moreover, given $\mathcal{D}_t$, $\widetilde{P}_{t+1,t}^\pi$ is unbiased of $P_{t+1,t}^\pi$ and $\widetilde{r}_t^\pi$ is unbiased of r_t^π.*

Proof of Lemma C.1. By construction of the estimator, $\widetilde{d}_t^\pi$ and $\widetilde{r}_{t-1}^\pi$ only depend on $\mathcal{D}_t$, therefore $\widetilde{d}_t^\pi$ and $\widetilde{r}_{t-1}^\pi$ given $\mathcal{D}_t$ are constants. For the second argument, we have $\forall s_t, s_{t+1}$,

$$\begin{aligned}
\mathbb{E}[\widetilde{P}_{t+1,t}^\pi(s_{t+1}|s_t)|\mathcal{D}_t] &= \sum_{a_t} \mathbb{E}[\widetilde{P}_{t+1,t}(s_{t+1}|s_t, a_t)|\mathcal{D}_t]\pi(a_t|s_t) \\
&= \sum_{a_t} \Big(\mathbf{1}(E_t)\mathbb{E}[\widehat{P}_{t+1,t}(s_{t+1}|s_t, a_t)|\mathcal{D}_t] + \mathbf{1}(E_t^c)P_{t+1,t}(s_{t+1}|s_t, a_t)\Big)\pi(a_t|s_t) \\
&= \sum_{a_t} \Big(\mathbf{1}(E_t)P_{t+1,t}(s_{t+1}|s_t, a_t) + \mathbf{1}(E_t^c)P_{t+1,t}(s_{t+1}|s_t, a_t)\Big)\pi(a_t|s_t) \\
&= \sum_{a_t} P_{t+1,t}(s_{t+1}|s_t, a_t)\pi(a_t|s_t) = P_{t+1,t}^\pi(s_{t+1}|s_t),
\end{aligned}$$

where the third equal sign comes from the fact that conditional on E_t, $\hat{P}(s_{t+1}|s_t, a_t)$ — the empirical mean — is unbiased. The result about $\tilde{r}_t^\pi$ can be derived using a similar fashion. ∎

Using Lemma C.1, we can derive the following recursions for expectation and variance:

Lemma C.2. *For $h = 1, ..., H$, we have*

$$\mathbb{E}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \sum_{t=1}^{h-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle \middle| \mathcal{D}_{h-1}\right] = \langle \widetilde{d}_{h-1}^\pi, V_{h-1}^\pi\rangle + \sum_{t=1}^{h-2}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle, \tag{14}$$

$$\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \sum_{t=1}^{h}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] = \mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle \widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle \middle| \mathcal{D}_h\right]\right] + \mathrm{Var}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \sum_{t=1}^{h-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] \tag{15}$$

Proof. The proof of Lemma C.2 can be found in Lemma B.2 and Lemma 4.1 in Xie et al. (2019) by coupling the standard Bellman equation:

$$V_h^\pi = r_h^\pi + [P_{h+1,h}^\pi]^T V_{h+1}^\pi \tag{16}$$

with the total law of expectations and the total law of variances. ∎

Lemma C.3 (Boundedness of Tabular MIS estimators)**.** $0 \leq \widehat{v}^\pi \leq HR_{\max}$, $0 \leq \widetilde{v}^\pi \leq HR_{\max}$.

[10] More rigorously, E_t depends on the specific pair s_t, a_t and should be written as $E_t(s_t, a_t)$. However, for brevity we just use E_t and this notation should be clear in each context.

[11] It depends on unknown information such as d_t^μ, $P_{t,t-1}^\pi$, exact conditional expectation of the reward r_t^π and so on.

Proof. we show $\widehat{P}_t^\pi(\cdot|s_{t-1})$ is a (degenerated) probability distribution for all t, s_{t-1}.

$$\begin{aligned}\sum_{s_t}\widehat{P}_t^\pi(s_t|s_{t-1}) &= \sum_{s_t}\sum_{a_{t-1}}\widehat{P}_t(s_t|s_{t-1},a_{t-1})\pi(a_{t-1}|s_{t-1}) \\ &= \sum_{a_{t-1}}\sum_{s_t}\widehat{P}_t(s_t|s_{t-1},a_{t-1})\pi(a_{t-1}|s_{t-1}) \qquad \text{This is since } |\mathcal{A}|,|\mathcal{S}|<\infty \\ &= \sum_{a_{t-1}}\sum_{s_t}\frac{n_{s_t,s_{t-1},a_{t-1}}}{n_{s_{t-1},a_{t-1}}}\pi(a_{t-1}|s_{t-1}) \\ &\le \sum_{a_{t-1}}\pi(a_{t-1}|s_{t-1}) = 1\end{aligned} \tag{17}$$

The last line is inequality since $\widehat{P}_t(s_t|s_{t-1},a_{t-1}) = 0$ when $n_{s_{t-1},a_{t-1}} = 0$. Following the same logic, it is easy to show $\widetilde{P}_t^\pi(\cdot|s_{t-1})$ is a non-degenerated probability distribution.

Next note $\sum_{s_1}\widehat{d}_1^\pi(s_1) = \sum_{s_1}\widehat{d}_1^\mu(s_1) = \sum_{s_1}\frac{n_{s_1}}{n} = 1$. Suppose $\widehat{d}_{t-1}^\pi(\cdot)$ is a (degenerated) probability distribution, then from $\widehat{d}_t^\pi = \widehat{P}_t^\pi\widehat{d}_{t-1}^\pi$ and (17), by induction we know $\widehat{d}_t^\pi(\cdot)$ is a (degenerated) probability distribution for all t.

Using Assumption 2.1, it is easy to show $\hat{r}_t^\pi(s_t) \le R_{\max}$ for all s_t, then combining all results above we have $\widehat{v}^\pi := \sum_{t=1}^H\langle\widehat{d}_t^\pi, \hat{r}_t^\pi\rangle \le HR_{\max}$. Similarly, $\widetilde{v}^\pi \le HR_{\max}$.

■

The boundedness of Tabular-MIS estimator cannot be inherited by the State-MIS estimator since $\widehat{v}_{\text{SMIS}}^\pi$ explicitly uses importance weights and there is no reason for it to be less than $HR_{\max}$. As a result, we do not need an extra projection step for our estimation to be valid (see Xie et al. (2019) Lemma B.1). Thanks to the following lemma, throughout the rest of the analysis we only need to consider $\widetilde{v}^\pi$.

Lemma C.4. *Let $\widehat{v}^\pi$ be the Tabular-MIS estimator and $\widetilde{v}^\pi$ be the fictitious version of TMIS we described above with parameter θ. Then the MSE of the TMIS and fictitious TMIS satisfies*

$$\mathbb{E}[(\widehat{v}^\pi - v^\pi)^2] \le \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2] + 3H^3SAR_{\max}^2 e^{-\frac{\theta^2 n\min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}{2}}$$

Proof of Lemma C.4. Define $E := \{\exists t, s_t, a_t \text{ s.t. } n_{s_t,a_t} < nd_t^\mu(s_t,a_t)(1-\theta)\}$. Similarly to Lemma B.1 in the appendix of Xie et al. (2019), we have

$$\begin{aligned}&\mathbb{E}[(\widehat{v}^\pi - v^\pi)^2] \le \mathbb{E}[(\widehat{v}^\pi - \widetilde{v}^\pi)^2] = \mathbb{E}[(\widehat{v}^\pi - \widetilde{v}^\pi)^2] + 2\mathbb{E}[(\widehat{v}^\pi - \widetilde{v}^\pi)(\widetilde{v}^\pi - v^\pi)] + \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2] \\ =&\mathbb{P}[E]\mathbb{E}\big[(\widehat{v}^\pi - \widetilde{v}^\pi)^2 + 2(\widehat{v}^\pi - \widetilde{v}^\pi)(\widetilde{v}^\pi - v^\pi)\big|E\big] + \mathbb{P}[E^c]\cdot 0 + \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2] \\ \le&3\mathbb{P}[E]H^2R_{\max}^2 + \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2],\end{aligned}$$

where the last inequality uses Lemma C.3. Then combining the multiplicative Chernoff bound (Lemma A.2 in the Appendix) and a union bound over each t,s_t and a_t, we get that

$$\mathbb{P}[E] \le \sum_t\sum_{s_t}\sum_{a_t}\mathbb{P}[n_{s_t,a_t} < nd_t^\mu(s_t,a_t)(1-\theta)] \le HSAe^{-\frac{\theta^2 n\min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}{2}},$$

which provides the stated result. ■

Lemma C.4 tells that MSE of two TMISs differs by a quantity $3H^3SAR_{\max}^2 e^{-\frac{\theta^2 n\min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}{2}}$ and this illustrates that the gap between two MSE's can be sufficiently small as long as $n \ge \frac{\text{polylog}(S,A,H,n)}{\min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}$.

C.2 Variance and Bias of Fictitious tabular MIS estimator.

Lemma C.5 (Xie et al. (2019) Lemma B.2)**.** *Tabular-MIS estimator is unbiased:* $\mathbb{E}[\widetilde{v}^\pi] = v^\pi$ *for all* $\theta < 1$.

Lemma C.6 (Variance decomposition)**.**

$$\begin{aligned}\mathrm{Var}[\widetilde{v}^\pi] =& \frac{\mathrm{Var}[V_1^\pi(s_1^{(1)})]}{n} \\ &+ \sum_{h=1}^{H}\sum_{s_h}\sum_{a_h}\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h,a_h}}\mathbf{1}(E_h)\right]\pi(a_h|s_h)^2\mathrm{Var}\left[(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\Big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right].\end{aligned} \tag{18}$$

where $V_t^\pi(s_t)$ *denotes the value function under* π *which satisfies the Bellman equation*

$$V_t^\pi(s_t) = r_t^\pi(s_t) + \sum_{s_{t+1}} P_t^\pi(s_{t+1}|s_t)V_{t+1}^\pi(s_{t+1}).$$

Remark C.7. *Note even though the construction of TMIS and SMIS are different, both fictitious estimators are unbiased for* v^π*. Therefore the MSE of MIS estimators are dominated by the variance of the fictitious estimators. Comparing Lemma C.6 with Lemma 4.1 in Xie et al. (2019) we can see our Tabular-MIS estimator achieves a lower bound, and it is essentially asymptotic optimal, as explained by Remark 3.2.*

Proof of Lemma C.6. The proof relies on applying Lemma C.2 in a recursive way. One key observation is

To begin with the following variance decomposition, which applies (15) recursively.

$$\begin{aligned}\mathrm{Var}[\widetilde{v}^\pi] =& \mathbb{E}\mathrm{Var}[\widetilde{v}^\pi|\mathcal{D}_H] + \mathrm{Var}[\mathbb{E}[\widetilde{v}^\pi|\mathcal{D}_H]] \\ =& \mathbb{E}\left[\mathrm{Var}[\langle\widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathrm{Var}[\mathbb{E}[\langle\widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H] + \sum_{t=1}^{H-1}\langle\widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\ =& \mathbb{E}\left[\mathrm{Var}[\langle\widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathrm{Var}[\langle\widetilde{d}_H^\pi, r_H^\pi\rangle + \sum_{t=1}^{H-1}\langle\widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\ =& \mathbb{E}\left[\mathrm{Var}[\langle\widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathrm{Var}[\langle\widetilde{d}_H^\pi, V_H^\pi\rangle + \sum_{t=1}^{H-1}\langle\widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\ =& \mathbb{E}\left[\mathrm{Var}[\langle\widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_H^\pi, V_H^\pi\rangle + \langle\widetilde{d}_{H-1}^\pi, \widetilde{r}_{H-1}^\pi\rangle\Big|\mathcal{D}_{H-1}\right]\right] \\ &+ \mathrm{Var}\left[\langle\widetilde{d}_{H-1}^\pi, V_{H-1}^\pi\rangle + \sum_{t=1}^{H-2}\langle\widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] = \ldots \\ =& \mathbb{E}\left[\mathrm{Var}[\langle\widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \sum_{h=1}^{H-1}\mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle\widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle\Big|\mathcal{D}_h\right]\right] + \mathrm{Var}\left[\langle\widetilde{d}_1^\pi, V_1^\pi\rangle\right]\end{aligned}$$

Now let us analyze $\mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle\widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle\Big|\mathcal{D}_h\right]\right]$. Note $\widetilde{P}_{h+1,h}^\pi(\cdot, s_h)$ and $\widetilde{r}_h^\pi(s_h)$ for each s_h are conditionally independent given $\mathcal{D}_h$, since $\mathcal{D}_h$ partitions the n episodes into S disjoint sets according to the states $s_h^{(i)}$ at time h. Similarly, $\tilde{P}_{h+1}(\cdot|s_h, a_h)$ and $\tilde{r}_h^\pi(s_h, a_h)$ for each (s_h, a_h) are also conditionally independent given $\mathcal{D}_h$. These observations imply:

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^{\pi}, V_{h+1}^{\pi}\rangle + \langle \widetilde{d}_{h}^{\pi}, \widetilde{r}_{h}^{\pi}\rangle \middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h}\mathrm{Var}\left[\widetilde{d}_{h}^{\pi}(s_h)\langle \widetilde{P}_{h+1,h}^{\pi}(\cdot,s_h), V_{h+1}^{\pi}\rangle + \widetilde{d}_{h}^{\pi}(s_h)\cdot\widetilde{r}_{h}^{\pi}(s_h)\middle|\mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h}\widetilde{d}_{h}^{\pi 2}(s_h)\mathrm{Var}\left[\sum_{a_h}\langle \widetilde{P}_{h+1}(\cdot|s_h,a_h)\cdot\pi(a_h|s_h), V_{h+1}^{\pi}\rangle + \sum_{a_h}\widetilde{r}_h(s_h,a_h)\cdot\pi(a_h|s_h)\middle|\mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h}\widetilde{d}_{h}^{\pi}(s_h)^2\sum_{a_h}\pi(a_h|s_h)^2\mathrm{Var}\left[\langle \widetilde{P}_{h+1}(\cdot|s_h,a_h), V_{h+1}^{\pi}\rangle + \widetilde{r}_h(s_h,a_h)\middle|\mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h}\widetilde{d}_{h}^{\pi}(s_h)^2\sum_{a_h}\pi(a_h|s_h)^2\mathbf{1}(E_t)\mathrm{Var}\left[\frac{1}{n_{s_h,a_h}}\sum_{i|s_h^{(i)}=s_h,a_h^{(i)}=a_h}(V_{h+1}^{\pi}(s_{h+1}^{(i)})+r_h^{(i)})\middle|\mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h}\widetilde{d}_{h}^{\pi}(s_h)^2\sum_{a_h}\pi(a_h|s_h)^2\cdot\frac{\mathbf{1}(E_t)}{n_{s_h,a_h}}\cdot\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(i)})+r_h^{(i)})\middle|s_h^{(i)}=s_h,a_h^{(i)}=a_h\right]\right] \\
=&\sum_{s_h}\sum_{a_h}\pi(a_h|s_h)^2\cdot\mathbb{E}\left[\frac{\widetilde{d}_{h}^{\pi}(s_h)^2}{n_{s_h,a_h}}\cdot\mathbf{1}(E_t)\right]\cdot\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(i)})+r_h^{(i)})\middle|s_h^{(i)}=s_h,a_h^{(i)}=a_h\right].
\end{aligned}
\tag{19}
$$

The second line and the fourth line use the conditional independence for s_t and (s_t, a_t) respectively. The fifth line uses that when $n_{s_h,a_h} < nd_h^{\mu}(s_h,a_h)(1-\theta)$, the conditional variance is 0. The sixth line uses the fact that episodes are iid.

Plug (19) into the above variance decomposition and uses $V_{H+1}=0$, we finally get

$$
\begin{aligned}
\mathrm{Var}[\widetilde{v}^{\pi}] =&\frac{\mathrm{Var}[V_1^{\pi}(s_1^{(1)})]}{n} \\
&+\sum_{h=1}^{H}\sum_{s_h}\sum_{a_h}\mathbb{E}\left[\frac{\widetilde{d}_{h}^{\pi}(s_h)^2}{n_{s_h,a_h}}\mathbf{1}(E_h)\right]\pi(a_h|s_h)^2\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)})+r_h^{(1)})\middle|s_h^{(1)}=s_h,a_h^{(1)}=a_h\right].
\end{aligned}
$$

■

C.3 Bounding the variance of $\widetilde{d}_h^{\pi}(s_h)$.

Applying the definition of variance, we directly have

$$
\mathbb{E}\left[\frac{\widetilde{d}_{h}^{\pi}(s_h)^2}{n_{s_h,a_h}}\mathbf{1}(E_h)\right] \le \frac{(1-\theta)^{-1}}{nd_h^{\mu}(s_h,a_h)}\mathbb{E}\left[\widetilde{d}_{h}^{\pi}(s_h)^2\right] = \frac{(1-\theta)^{-1}}{nd_h^{\mu}(s_h,a_h)}(d_h^{\pi}(s_h)^2 + \mathrm{Var}[\widetilde{d}_{h}^{\pi}(s_h)]), \tag{20}
$$

where we use the fact that $\widetilde{d}_h^{\pi}(s_h)$ is unbiased (which can be proved by induction through applying total law of expectations and the recursive relationship $\widetilde{d}_t^{\pi} = \widetilde{P}_t^{\pi}\widetilde{d}_{t-1}^{\pi}$). Therefore the only thing left is to bound the the variance of $\widetilde{d}_h^{\pi}(s_h)$. To tackle it, we consider bounding the covariance matrix of $\widetilde{d}_h^{\pi}(s_h)$. As we shall see in Lemma C.8, fortunately, we are able to derive an identical result of Lemma B.4 in Xie et al. (2019) for our Tabular-MIS estimator, which helps greatly in bounding the the variance of $\widetilde{d}_h^{\pi}(s_h)$.

Lemma C.8 (Covariance of $\widetilde{d}_h^{\pi}$ with TMIS)**.**

$$
\begin{aligned}
\mathrm{Cov}(\widetilde{d}_h^{\pi}) \preceq&\frac{(1-\theta)^{-1}}{n}\sum_{t=1}^{h-1}\mathbb{P}_{h+1,t+1}^{\pi}\mathrm{diag}\left[\sum_{s_t,a_t}\frac{d_t^{\pi}(s_t)^2+\mathrm{Var}(\widetilde{d}_t^{\pi}(s_t))}{d_t^{\mu}(s_t)}\frac{\pi(a_t|s_t)^2}{\mu(a_h|s_t)}\mathbb{P}_{t+1,t}(\cdot|s_t,a_t)\right]\left[\mathbb{P}_{h+1,t+1}^{\pi}\right]^T \\
&+\frac{1}{n}\mathbb{P}_{h,1}^{\pi}\mathrm{diag}\left[d_1^{\pi}\right]\left[\mathbb{P}_{h,1}^{\pi}\right]^T.
\end{aligned}
$$

where $\mathbb{P}_{h,t}^{\pi} = \mathbb{P}_{h,h-1}^{\pi}\cdot\mathbb{P}_{h-1,h-2}^{\pi}\cdot\ldots\cdot\mathbb{P}_{t+1,t}^{\pi}$ *— the transition matrices under policy* π *from time* t *to* h *(define* $\mathbb{P}_{h,h}^{\pi} := I$*).*

Proof of Lemma C.8. We start by applying the law of total variance to obtain the following recursive equation

$$\mathrm{Cov}[\widetilde{d}_h^\pi] = \mathbb{E}\left[\mathrm{Cov}\left[\widetilde{\mathbb{P}}_{h,h-1}^\pi \widetilde{d}_{h-1}^\pi \middle| \mathcal{D}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}_{h,h-1}^\pi \widetilde{d}_{h-1}^\pi \middle| \mathcal{D}_{h-1}\right]\right] \tag{21}$$

$$= \mathbb{E}\left[\mathrm{Cov}\left[\sum_{s_{h-1}} \widetilde{\mathbb{P}}_{h,h-1}^\pi(\cdot|s_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1}) \middle| \mathcal{D}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}_{h,h-1}^\pi \widetilde{d}_{h-1}^\pi \middle| \mathcal{D}_{h-1}\right]\right] \tag{22}$$

$$= \underbrace{\mathbb{E}\left[\sum_{s_{h-1}} \mathrm{Cov}\left[\widetilde{\mathbb{P}}_{h,h-1}^\pi(\cdot|s_{h-1}) \middle| \mathcal{D}_{h-1}\right] \widetilde{d}_{h-1}^\pi(s_{h-1})^2\right]}_{(*)} + \mathbb{P}_{h,h-1}^\pi \mathrm{Cov}[\widetilde{d}_{h-1}^\pi][\mathbb{P}_{h,h-1}^\pi]^T. \tag{23}$$

The decomposition of the covariance in the third line uses that $\mathrm{Cov}(X+Y) = \mathrm{Cov}(X) + \mathrm{Cov}(Y)$ when X and Y are statistically independent and the columns of $\widetilde{\mathbb{P}}_{h,h-1}$ are independent when conditioning on $\mathcal{D}_{h-1}$.

$$(*) = \mathbb{E}\left[\sum_{s_{h-1}}\sum_{a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \mathrm{Cov}\left[\widetilde{\mathbb{P}}_h(\cdot|s_{h-1},a_{h-1}) \middle| \mathrm{Data}_{h-1}\right] \widetilde{d}_{h-1}^\pi(s_{h-1})^2\right] \tag{24}$$

$$= \mathbb{E}\left[\sum_{s_{h-1}}\sum_{a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \mathbf{1}(E_{h-1}) \mathrm{Cov}\left[\widehat{\mathbb{P}}_h(\cdot|s_{h-1},a_{h-1}) \middle| \mathrm{Data}_{h-1}\right] \widetilde{d}_{h-1}^\pi(s_{h-1})^2\right] \tag{25}$$

$$= \mathbb{E}\left[\sum_{s_{h-1}}\sum_{a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \frac{\mathbf{1}(E_{h-1})}{n_{s_{h-1},a_{h-1}}} \mathrm{Cov}\left[\mathbf{e}_{s_h^{(1)}} \middle| s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}\right] \widetilde{d}_{h-1}^\pi(s_{h-1})^2\right] \tag{26}$$

$$= \sum_{s_{h-1},a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \mathbb{E}\left[\frac{\widetilde{d}_{h-1}^\pi(s_{h-1})^2}{n_{s_{h-1},a_{h-1}}} \mathbf{1}(E_{h-1})\right] \Big[\mathrm{diag}[\mathbb{P}_h(\cdot|s_{h-1},a_{h-1})] \tag{27}$$

$$- \mathbb{P}_h(\cdot|s_{h-1},a_{h-1}) \cdot \mathbb{P}_h(\cdot|s_{h-1},a_{h-1})^T\Big] \tag{28}$$

$$\prec \sum_{s_{h-1}}\sum_{a_{h-1}} \left\{\frac{d_{h-1}^\pi(s_{h-1})^2 + \mathrm{Var}[\widetilde{d}_{h-1}^\pi(s_{h-1})]}{n d_{h-1}^\mu(s_{h-1})(1-\theta)} \frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})} \mathrm{diag}[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]\right\} \tag{29}$$

The second line uses the fact that conditional on E_{h-1}^c, the variance of $\widetilde{\mathbb{P}}(\cdot|s_{h-1},a_{h-1})$ is zero given Data_h. The third line uses the basic property of empirical average, and the fourth line comes from the fact

$$\begin{aligned}
&\mathrm{Cov}\left[\mathbf{e}_{s_h^{(1)}} \middle| s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}\right] \\
=&\mathbb{E}\left[\mathbf{e}_{s_h^{(1)}} \cdot \mathbf{e}_{s_h^{(1)}}^T \middle| s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}\right] \\
&- \mathbb{E}\left[\mathbf{e}_{s_h^{(1)}} \middle| s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}\right] \cdot \mathbb{E}\left[\mathbf{e}_{s_h^{(1)}} \middle| s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}\right]^T \\
=&\mathrm{diag}(\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})) - \mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]^T
\end{aligned}$$

The last line (29) uses the fact that $\mathbb{P}_{h,h-1}^\pi(\cdot|s_{h-1})[\mathbb{P}_{h,h-1}^\pi(\cdot|s_{h-1})]^T$ is positive semidefinite, $n_{s_{h-1},a_{h-1}} \geq n d_{h-1}^\mu(s_{h-1},a_{h-1})(1-\theta)$ and the definition of variance for $\widetilde{d}_{h-1}^\pi(s_{h-1})$. Combining (23) and (29) and by recursively apply them, we get the stated results. ■

Benefitting from the identical semidefinite ordering bound on $\mathrm{Cov}(\widetilde{d}_h^\pi)$ for TMIS and SMIS, we can borrow the following results from Xie et al. (2019) for our Tabular-MIS estimator.

Lemma C.9 (Corollary 2 of Xie et al. (2019)). *For $h = 1$, we have $\mathrm{Var}[\widetilde{d}_1^\pi(s_1)] = \frac{1}{n}(d_h^\pi(s_1) - d_h^\pi(s_1)^2)$, and for $h = 2, 3, ..., H$, we have:*

$$\mathrm{Var}[\widetilde{d}_h^\pi(s_h)] \leq \frac{(1-\theta)^{-1}}{n} \sum_{t=2}^{h} \sum_{s_t} \mathbb{P}_{h,t}^\pi(s_h|s_t)^2 \varrho(s_t) + \frac{1}{n} \sum_{s_1} \mathbb{P}_{h,1}^\pi(s_h|s_1)^2 d_1(s_1)$$

where $\varrho(s_t) := \sum_{s_{t-1}} \left(\frac{d^\pi_{t-1}(s_{t-1})^2 + \mathrm{Var}(\widetilde{d}^\pi_{t-1}(s_{t-1}))}{d^\mu_{t-1}(s_{t-1})} \sum_{a_{t-1}} \frac{\pi(a_{t-1}|s_{t-1})^2}{\mu(a_{t-1}|s_{t-1})} \mathbb{P}_{t,t-1}(s_t|s_{t-1}, a_{t-1}) \right)$.

Lemma C.10 (Error propagation: Theorem B.1 of Xie et al. (2019))**.** *Let* $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ *and* $\tau_s := \max_{t,s_t} \frac{d^\pi_t(s_t)}{d^\mu_t(s_t)}$. *If* $n \geq \frac{2(1-\theta)^{-1} t \tau_a \tau_s}{\max\{d^\pi_t(s_t), d^\mu_t(s_t)\}}$ *for all* $t = 2, ..., H$, *then for all* $h = 1, 2, ..., H$ *and* s_h, *we have that:*

$$\mathrm{Var}[\widetilde{d}^\pi_h(s_h)] \leq \frac{2(1-\theta)^{-1} h \tau_a \tau_s}{n} d^\pi_h(s_h).$$

Before giving the proof of Theorem 3.1, we first prove Lemma 3.4.

Proof of Lemma 3.4. Let value function $V^\pi_h(s_h) = \mathbb{E}_\pi[\sum_{t=h}^H r^{(1)}_t | s^{(1)}_h = s_h]$ and Q-function $Q^\pi_h(s_h, a_h) = \mathbb{E}_\pi[\sum_{t=h}^H r^{(1)}_t | s^{(1)}_h = s_h, a^{(1)}_h = a_h]$, then by total law of variance we obtain (let's suppress the policy π for simplicity):

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{t=1}^h r^{(1)}_t + V_{h+1}(s^{(1)}_{h+1})\right] \\
=&\mathbb{E}\left[\mathrm{Var}\left[\sum_{t=1}^h r^{(1)}_t + V_{h+1}(s^{(1)}_{h+1}) \middle| \mathcal{D}_h\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[\sum_{t=1}^h r^{(1)}_t + V_{h+1}(s^{(1)}_{h+1}) \middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\mathrm{Var}\left[r^{(1)}_h + V_{h+1}(s^{(1)}_{h+1}) \middle| s^{(1)}_h, a^{(1)}_h\right]\right] + \mathrm{Var}\left[\sum_{t=1}^{h-1} r^{(1)}_t + \mathbb{E}\left[V_{h+1}(s^{(1)}_{h+1}) + r^{(1)}_h \middle| s^{(1)}_h, a^{(1)}_h\right]\right] \\
=&\mathbb{E}\left[\mathrm{Var}\left[r^{(1)}_h + V_{h+1}(s^{(1)}_{h+1}) \middle| s^{(1)}_h, a^{(1)}_h\right]\right] + \mathrm{Var}\left[\sum_{t=1}^{h-1} r^{(1)}_t + Q_h(s^{(1)}_h, a^{(1)}_h)\right] \\
=&\mathbb{E}\left[\mathrm{Var}\left[r^{(1)}_h + V_{h+1}(s^{(1)}_{h+1}) \middle| s^{(1)}_h, a^{(1)}_h\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[\sum_{t=1}^{h-1} r^{(1)}_t + Q_h(s^{(1)}_h, a^{(1)}_h) \middle| s^{(1)}_h, r^{(1)}_{1:h-1}\right]\right] \\
&+\mathrm{Var}\left[\mathbb{E}\left[\sum_{t=1}^{h-1} r^{(1)}_t + Q_h(s^{(1)}_h, a^{(1)}_h) \middle| s^{(1)}_h, r^{(1)}_{1:h-1}\right]\right] \\
=&\mathbb{E}\left[\mathrm{Var}\left[r^{(1)}_h + V_{h+1}(s^{(1)}_{h+1}) \middle| s^{(1)}_h, a^{(1)}_h\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[Q_h(s^{(1)}_h, a^{(1)}_h) \middle| s^{(1)}_h, r^{(1)}_{1:h-1}\right]\right] \\
&+\mathrm{Var}\left[\sum_{t=1}^{h-1} r^{(1)}_t + \mathbb{E}\left[Q_h(s^{(1)}_h, a^{(1)}_h) \middle| s^{(1)}_h\right]\right] \\
=&\mathbb{E}\left[\mathrm{Var}\left[r^{(1)}_h + V_{h+1}(s^{(1)}_{h+1}) \middle| s^{(1)}_h, a^{(1)}_h\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[Q_h(s^{(1)}_h, a^{(1)}_h) \middle| s^{(1)}_h\right]\right] + \mathrm{Var}\left[\sum_{t=1}^{h-1} r^{(1)}_t + V_h(s^{(1)}_h)\right],
\end{aligned} \tag{30}$$

where we use Markovian property that $(V_{h+1}(s^{(1)}_{h+1})|\mathcal{D}_h)$ equals $(V_{h+1}(s^{(1)}_{h+1})|s^{(1)}_h, a^{(1)}_h)$ in distribution and $\mathbb{E}\left[V_{h+1}(s^{(1)}_{h+1}) + r^{(1)}_h \middle| s^{(1)}_h, a^{(1)}_h\right] = Q_h(s^{(1)}_h, a^{(1)}_h)$. Then by applying (30) recursively and letting $h = H$, we get the stated result. ∎

Remark C.11. *A straight forward implication of Lemma 3.4 is the following:*

$$\sum_{t=1}^H \mathbb{E}_\pi\left[\mathrm{Var}\left[V^\pi_{t+1}(s^{(1)}_{t+1}) + r^{(1)}_t \middle| s^{(1)}_t, a^{(1)}_t\right]\right] \leq H^2 R^2_{\max}.$$

Combing Lemma C.6 and C.10, we are now ready to prove the main Theorem 3.1.

Proof of Theorem 3.1. Plug the result of Lemma C.10 into Lemma C.6 and uses the unbiasedness of $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ (Lemma C.5) we obtain $\forall\ 0<\theta<1$:

$$\begin{aligned}
&\mathbb{E}[(\widetilde{v}^{\pi}_{\mathrm{TMIS}}-v^{\pi})^2]\\
\leq&\frac{\mathrm{Var}[V_1^{\pi}(s_1^{(1)})]}{n}+\sum_{h=1}^{H}\sum_{s_h,a_h}\frac{(1-\theta)^{-1}}{nd_h^{\mu}(s_h,a_h)}d_h^{\pi}(s_h)^2\pi(a_h|s_h)^2\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)})+r_h^{(1)})\middle|s_h^{(1)}=s_h,a_h^{(1)}=a_h\right].\\
&+\frac{(1-\theta)^{-1}}{n}\sum_{h=1}^{H}\sum_{s_h,a_h}\frac{2(1-\theta)^{-1}h\tau_a\tau_s}{n}\frac{d_h^{\pi}(s_h)}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)})+r_h^{(1)})\middle|s_h^{(1)}=s_h\right]
\end{aligned}\tag{31}$$

Choose $\theta=\sqrt{4\log(n)/(n\min_{t,s_t,a_t}d_t^{\mu}(s_t,a_t))}$. Then by assumption $n>\frac{16\log n}{\min_{t,s_t,a_t}d_t^{\mu}(s_t,a_t)}$ we have $\theta<1/2$, which allows us to write $(1-\theta)^{-1}\leq(1+2\theta)$ in the leading term and $(1-\theta)^{-1}\leq 2$ in the subsequent terms. The condition of Lemma C.10 is satisfied by The second assumption on n. Then, combining (31) with Lemma C.4 we get:

$$\begin{aligned}
\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}}-v^{\pi})^2]\leq&\frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)})+r_h^{(1)})\middle|s_h^{(1)}=s_h,a_h^{(1)}=a_h\right]\\
&\cdot\left(1+\sqrt{\frac{16\log n}{n\min_{t,s_t}d_t^{\mu}(s_t)}}\right)+\frac{3}{n^2}H^3SAR_{\max}^2\\
&+\frac{8\tau_a\tau_s}{n^2}\sum_{h=1}^{H}\sum_{s_h,a_h}\frac{h\cdot d_h^{\pi}(s_h)}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\cdot\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)})+r_h^{(1)})\middle|s_h^{(1)}=s_h,a_h^{(1)}=a_h\right],
\end{aligned}\tag{32}$$

now use Lemma 3.4, we can bound the last term in (32) by

$$\frac{8\tau_a^2\tau_sH}{n^2\cdot d_m}\sum_{t=1}^{H}\mathbb{E}_{\pi}\left[\mathrm{Var}\left[V_{t+1}^{\pi}(s_{t+1}^{(1)})+r_t^{(1)}\middle|s_t^{(1)},a_t^{(1)}\right]\right]\leq\frac{8\tau_a^2\tau_sH^3R_{\max}^2}{n^2\cdot d_m},$$

Combine this term with $\frac{3}{n^2}H^3SAR_{\max}^2$ we obtain the higher order term $O(\frac{\tau_a^2\tau_sH^3R_{\max}^2}{n^2\cdot d_m})$, where we use that pigeonhole principle implies that $S<\tau_s, A<\tau_a$.

This completes the proof.

■

D Proofs of data splitting Tabular-MIS estimator.

We define the fictitious data splitting Tabular-MIS estimator as:

$$\widetilde{v}^{\pi}_{\mathrm{split}}=\frac{1}{N}\sum_{i=1}^{N}\widetilde{v}^{\pi}_{(i)},$$

where each $\widetilde{v}^{\pi}_{(i)}$ is the fictitious Tabular-MIS estimator of $\widehat{v}^{\pi}_{(i)}$. Moreover, we set all $\widetilde{v}^{\pi}_{(1)},\widetilde{v}^{\pi}_{(2)},...,\widetilde{v}^{\pi}_{(N)}$ jointly share the same fictitious parameter θ_M.

Proof of Theorem 3.6. Let $E':=\{\exists\ \widetilde{v}^{\pi}_{(i)}: s.t.\widetilde{v}^{\pi}_{(i)}\neq\widehat{v}^{\pi}_{(i)}\}$, then an argument similar to Lemma C.4 can be derived:

$$\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{split}}-v^{\pi})^2]\leq 3\mathbb{P}[E']H^2R_{\max}^2+\mathbb{E}[(\widetilde{v}^{\pi}_{\mathrm{split}}-v^{\pi})^2],$$

and

$$\mathbb{P}[E']\leq N\sum_t\sum_{s_t}\sum_{a_t}\mathbb{P}[n_{s_t,a_t}<M\cdot d_t^{\mu}(s_t,a_t)(1-\theta_M)]\leq NHSAe^{-\frac{\theta_M^2M\min_{t,s_t,a_t}d_t^{\mu}(s_t,a_t)}{2}},$$

therefore $\mathbb{P}[E']$ will be sufficiently small if $M\geq O(\mathrm{Polylog}(H,S,A,n)/\min_{t,s_t,a_t}d_t^{\mu}(s_t,a_t))$. By near-uniformity we $M\geq O(\mathrm{Polylog}(H,S,A,n)SA)\geq O(\mathrm{Polylog}(H,S,A,n)/\min_{t,s_t,a_t}d_t^{\mu}(s_t,a_t))$.

Moreover, by i.i.d and unbiasedness of $\widetilde{v}^{\pi}_{(i)}$, we have

$$\mathbb{E}[(\widetilde{v}^{\pi}_{\text{split}} - v^{\pi})^2] = \frac{1}{N}\mathbb{E}[(\widetilde{v}^{\pi}_{(1)} - v^{\pi})^2] \leq \frac{1}{N} \cdot O(\frac{H^2SA}{M}) = O(\frac{H^2SA}{n}),$$

by the second assumption on M and Theorem 3.1.

■

We now proof Lemma 3.9, since it will be used to as the intermediate step for proving Theorem 3.8.

Proof of Lemma 3.9. Note that

$$\begin{aligned}
&\mathbb{P}\left[\left\{\exists \pi \in \prod \text{ s.t. } \widetilde{v}^{\pi}_{\text{split}} \neq \widehat{v}^{\pi}_{\text{split}}\right\}\right] \leq N \cdot \mathbb{P}\left[\left\{\exists \pi \in \prod, \text{ s.t. } \widetilde{v}^{\pi}_{(1)} \neq \widehat{v}^{\pi}_{(1)}\right\}\right] \\
&\leq N \cdot \mathbb{P}\left[\{\exists t, s_t, a_t \text{ s.t. } n^{(1)}_{s_t,a_t} < n d^{\mu}_t(s_t,a_t)(1-\theta_M)\}\right] \\
&\leq NHSAe^{-\frac{\theta_M^2 M \min_{t,s_t,a_t} d^{\mu}_t(s_t,a_t)}{2}},
\end{aligned}$$

therefore by near-uniformity $M > \max\left[O(SA \cdot \text{Polylog}(S,H,A,N,1/\delta)), O(H\tau_a\tau_s)\right]$ is sufficient to guarantee the stated result. ■

Now we can prove Theorem 3.8.

Proof of Theorem 3.8. First of all, we have

$$\mathbb{P}\left(|\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| > \epsilon\right) \leq \mathbb{P}\left(|\widehat{v}^{\pi}_{\text{split}} - \widetilde{v}^{\pi}_{\text{split}}| > 0\right) + \mathbb{P}\left(|\widetilde{v}^{\pi}_{\text{split}} - v^{\pi}| > \epsilon\right), \tag{33}$$

Now by Bernstein inequality we have

$$\mathbb{P}\left(|\widetilde{v}^{\pi}_{\text{split}} - v^{\pi}| > \epsilon\right) = \mathbb{P}\left(|\frac{1}{N}\sum_{i=1}^{N}(\widetilde{v}^{\pi}_{(i)} - v^{\pi})| \geq \epsilon\right) \leq \exp\left(-\frac{N\epsilon^2}{2\text{Var}(\widetilde{v}^{\pi}_{(1)}) + 2HR_{\max}\epsilon/3}\right) := \delta/2. \tag{34}$$

Solving (34) and apply Theorem 3.1, we obtain

$$\epsilon \leq \sqrt{\frac{2\text{Var}(\widetilde{v}^{\pi}_{(1)})\log(2/\delta)}{N}} + \frac{2HR_{\max}\log(2/\delta)}{3N} \leq \widetilde{O}(\sqrt{\frac{H^2SA\log(2/\delta)}{M \cdot N}}) + \frac{2HR_{\max}\log(2/\delta)}{3N}. \tag{35}$$

As N goes large, the square root term in (35) will dominate and it seems we only need to consider the square root term in N and treat the second term as the higher order term. However, since $M > \max\left[O(SA \cdot \text{Polylog}(S,H,A,N,1/\delta)), O(H\tau_a\tau_s)\right]$, N cannot be arbitrary large given n. An example is: when $N = n$, then $M = n/N = 1$ does not satisfy the condition. Therefore to make the square root term dominates we need

$$\sqrt{\frac{H^2SA\log(2/\delta)}{M \cdot N}} \geq O(\frac{HR_{\max}\log(2/\delta)}{N}).$$

This translates to

$$M \leq \widetilde{O}(\sqrt{nSA}), \tag{36}$$

where $\widetilde{O}$ absorbs all the Polylog terms.

Therefore under the condition (36), we can really absorb the second term in (35) (as higher order term) and combine it with Lemma 3.9 to get that with probability $1-\delta$,

$$|\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| \leq 0 + \widetilde{O}(\sqrt{\frac{H^2SA}{M \cdot N}}) = \widetilde{O}(\sqrt{\frac{H^2SA}{n}}).$$

■

Proof of Theorem 5.1. The non-uniform result of Theorem 3.8 gives:

$$|\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^2SA}{n}})$$

Note that all nonstationary deterministic polices class have cardinality $|\prod| = A^{HS}$, which implies $\log|\prod| = HS\log A$, therefore combine Lemma 3.9 with a direct union bound and Multiplicative Chernoff bound we obtain

$$\sup_{\pi\in\prod} |\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^3S^2A}{n}})$$

■

E More details about Empirical Results.

Restate Time-varying, non-mixing Tabular MDP in Section 4.

There are two states s_0 and s_1 and two actions a_1 and a_2. State s_0 always has probability 1 going back to itself, regardless of the actions, *i.e.* $P_t(s_0|s_0,a_1)=1$ and $P_t(s_0|s_0,a_2)=1$. For state s_1, at each time step there is one action (we call it a) that has probability $2/H$ going to s_0 and the other action (we call it a') has probability 1 going back to s_1,

$$P_t(s|s_1,a) = \begin{cases} \frac{2}{H} & \text{if } s=s_0; \\ 1-\frac{2}{H} & \text{if } s=s_1. \end{cases} \quad P_t(s|s_1,a') = \begin{cases} 0 & \text{if } s=s_0; \\ 1 & \text{if } s=s_1. \end{cases}$$

and which action will make state s_1 go to state s_0 with probability $2/H$ is decided by a random parameter p_t uniform sampled in $[0,1]$. If $p_t<0.5$, $a=a_1$ and if $p_t\geq 0.5$, $a=a_2$. These $p_1,...,p_H$ are generated by a sequence of pseudo-random numbers. Moreover, one can receive reward 1 at each time step if $t>H/2$ and is in state s_0, and will receive reward 0 otherwise. Lastly, for logging policy, we define it to be uniform:

$$\mu(\cdot|s_0) = \begin{cases} \frac{1}{2} & \text{if } \cdot = a_1; \\ \frac{1}{2} & \text{if } \cdot = a_2. \end{cases} \quad \text{and} \quad \mu(\cdot|s_1) = \begin{cases} \frac{1}{2} & \text{if } \cdot = a_1; \\ \frac{1}{2} & \text{if } \cdot = a_2. \end{cases}$$

For target policy π, we define it as:

$$\pi(\cdot|s_0) = \begin{cases} \frac{1}{2} & \text{if } \cdot = a_1; \\ \frac{1}{2} & \text{if } \cdot = a_2. \end{cases} \quad \text{and} \quad \pi(\cdot|s_1) = \begin{cases} \frac{1}{4} & \text{if } \cdot = a_1; \\ \frac{3}{4} & \text{if } \cdot = a_2. \end{cases}$$

We run this non-stationary MDP model in the `Python` environment and pseudo-random numbers p_t's are generated by keeping `numpy.random.seed(100)`.

We run each methods under $K=100$ macro-replications with data $\mathcal{D}_{(k)} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{(k)}^{i\in[n],t\in[H]}$, and use each $\mathcal{D}_{(k)}$ $(k=1,...,K)$ to construct a estimator $\widehat{v}^{\pi}_{[k]}$, then the (empirical) RMSE is computed as:

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{K}(\widehat{v}^{\pi}_{[k]} - v^{\pi}_{\text{true}})^2}{K}},$$

where v^{π}_{true} is obtained by calculating $P^{\pi}_{t+1,t}(s'|s) = \sum_a P_{t+1,t}(s'|s,a)\pi_t(a|s)$, the marginal state distribution $d^{\pi}_t = P^{\pi}_{t,t-1}d^{\pi}_{t-1}$, $r^{\pi}_t(s_t) = \sum_{a_t} r_t(s_t,a_t)\pi_t(a_t|s_t)$ and $v^{\pi}_{\text{true}} = \sum_{t=1}^{H}\sum_{s_t} d^{\pi}_t(s_t)r^{\pi}_t(s_t)$. Then Relative-RMSE equals to RMSE/v^{π}_{true}.

Other generic IS-based estimators. There are other Importance Sampling based estimators including *weighted importance sampling* (WIS) and *importance sampling with stationary state distribution* (SSD-IS, Liu et al. (2018a)). The empirical comparisons including these methods are well-demonstrated in Xie et al. (2019) and it was empirically shown that they are worse than SMIS. Because of that, we only focus on comparing SMIS and TMIS in our simulation study.

Algorithm 2 Data Splitting Tabular MIS OPE

Input: Logging data $\mathcal{D} = \{\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{t=1}^H\}_{i=1}^n$ from the behavior policy μ. A target policy π which we want to evaluate its cumulative reward. Splitting data size M.

1: Randomly splitting the data $\mathcal{D}$ evenly into N folds, with each fold $|\mathcal{D}^{(i)}| = M$.
2: **for** $i = 1, 2, \ldots, N$ **do**
3: Use Algorithm 1 to estimate $\widehat{v}_{(i)}^{\pi}$ with data $\mathcal{D}^{(i)}$.
4: **end for**
5: Use the mean of $\widehat{v}_{(1)}^{\pi}, \widehat{v}_{(2)}^{\pi}, ..., \widehat{v}_{(N)}^{\pi}$ as the final estimation of v^{π}.