# Appendix

## A Technical lemmas

**Lemma A.1** (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 looser bound that suffices for our propose:

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

**Lemma A.2** (Hoeffdings Inequality Sridharan (2002)). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \le \xi_i$ with probability 1. Then for any $\epsilon > 0$ we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \ge \epsilon\right) \le e^{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}\xi_i^2}}.$$

**Lemma A.3** (Bernsteins Inequality). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \le \xi$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[x_i]$, then with probability $1-\delta$ we have*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \le \sqrt{\frac{2\sigma^2 \cdot \log(1/\delta)}{n}} + \frac{2\xi}{3n}\log(1/\delta)$$

**Lemma A.4** (Mcdiarmids Inequality (Sridharan, 2002)). *Let $x_1, ..., x_n$ be independent random variables and $S : X^n \to \mathbb{R}$ be a measurable function which is invariant under permutation and let the random variable $Z$ be given by $Z = S(x_1, x_2, ..., x_n)$. Assume $S$ has bounded difference: i.e.*

$$\sup_{x_1,...,x_n,x_i'} |S(x_1, ..., x_i, ..., x_n) - S(x_1, ..., x_i', ..., x_n)| \le \xi_i,$$

*then for any $\epsilon > 0$ we have*

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \ge \epsilon) \le 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}\xi_i^2}}.$$

**Lemma A.5** (Azuma-Hoeffding inequality). *Suppose $X_k$, $k = 1, 2, 3, ...$ is a martingale and $|X_k - X_{k-1}| \le c_k$ almost surely. Then for all positive integers $N$ and any $\epsilon > 0$,*

$$\mathbb{P}(|X_N - X_0| \ge \epsilon) \le 2e^{-\frac{\epsilon^2}{2\sum_{i=1}^{N} c_i^2}}.$$

**Lemma A.6** (Freedman's inequality Tropp et al. (2011)). *Let $X$ be the martingale associated with a filter $\mathcal{F}$ (i.e. $X_i = \mathbb{E}[X|\mathcal{F}_i]$) satisfying $|X_i - X_{i-1}| \le M$ for $i = 1, ..., n$. Denote $W := \sum_{i=1}^{n}\mathrm{Var}(X_i|\mathcal{F}_{i-1})$ then we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \ge \epsilon, W \le \sigma^2) \le 2e^{-\frac{\epsilon^2}{2(\sigma^2 + M\epsilon/3)}}.$$

*Or in other words, with probability $1-\delta$,*

$$|X - \mathbb{E}[X]| \le \sqrt{8\sigma^2 \cdot \log(1/\delta)} + \frac{2M}{3}\cdot\log(1/\delta), \quad Or \quad W \ge \sigma^2.$$

**Lemma A.7** (Best arm identification lower bound Krishnamurthy et al. (2016)). *For any $A \ge 2$ and $\tau \le \sqrt{1/8}$ and any best arm identification algorithm that produces an estimate $\hat{a}$, there exists a multi-arm bandit problem for which the best arm $a^\star$ is $\tau$ better than all others, but $\mathbb{P}[\hat{a} \ne a^\star] \ge 1/3$ unless the number of samples $T$ is at least $\frac{A}{72\tau^2}$.*

## B On error metric for OPE

In this section, we discuss the metric considered in this work. Traditionally, most works directly use *Mean Square Error* (MSE) $\mathbb{E}[(\widehat{v}^\pi - v^\pi)^2]$ as the criterion for measuring OPE methods *e.g.* Thomas and Brunskill (2016); Thomas (2015); Thomas et al. (2017); Farajtabar et al. (2018), or equivalently, by proposing unbiased estimators and discussing its variance *e.g.* Jiang and Li (2016). Alternately, one can consider bounding the absolute difference between $v^\pi$ and $\widehat{v}^\pi$ with high probability (*e.g.* Duan et al. (2020)), *i.e.* $|\widehat{v}^\pi - v^\pi| \le \epsilon_{\text{prob}}$ *w.h.p.* Generally speaking, high probability bound can be seen as a stricter criterion compared to MSE since

$$\begin{aligned}\mathbb{E}[(\widehat{v}^\pi - v^\pi)^2] &= \mathbb{E}[(\widehat{v}^\pi - v^\pi)^2 \mathbb{1}_E] + \mathbb{E}[(\widehat{v}^\pi - v^\pi)^2 \mathbb{1}_{E^c}] \\ &\le \epsilon_{\text{prob}}(\delta)^2 \cdot (1-\delta) + H^2 \cdot \delta,\end{aligned}$$

where $E$ is the event that $\epsilon_{\text{prob}}$ error holds and $\delta$ is the failure probability. As a result, if both $\delta$ and $\epsilon_{\text{prob}}(\delta)$ can be controlled small, then the high probability bound implies a result for MSE bound. This is realistic, since $\delta$ mostly appears inside the logarithmic term of $\epsilon_{\text{prob}}(\delta)$ so the second term can be scaled to sufficiently small without affecting the polynomial dependence for the first term.

Table 2: Summary of Uniform OPE results for $H$-horizon non-stationary setting

| Method/Analysis | Policy class | Guarantee | Sample complexity |
|---|---|---|---|
| Simulation Lemma | All policies | $\epsilon$-uniform convergence | $O(H^4S^2/d_m\epsilon^2)$ |
| Theorem 3.3 | All policies | $\epsilon$-uniform convergence | $O(H^4S/d_m\epsilon^2)$ |
| Theorem 3.5 | All deterministic policies | $\epsilon$-uniform convergence | $O(H^3S/d_m\epsilon^2)$ |
| Theorem 3.7 | local policies | $\epsilon$-uniform convergence | $O(H^3/d_m\epsilon^2)$ |
| Minimax lower bound (Theorem 3.8) | ———— | over class $\mathcal{M}_{d_m}$ | $\Omega(H^3/d_m\epsilon^2)$ |

## C Some preparations

In this section we present some results that are critical for proving the main theorems.

**Lemma C.1.** *For any $0 < \delta < 1$, there exists an absolute constant $c_1$ such that when total episode $n > c_1 \cdot 1/d_m \cdot \log(HSA/\delta)$, then with probability $1-\delta$,*

$$n_{s_t,a_t} \ge n \cdot d_t^\mu(s_t,a_t)/2, \quad \forall\, s_t, a_t.$$

If state $s_t$ is not accessible, then $n_{s_t,a_t} = d_t^\mu(s_t,a_t) = 0$ so the lemma holds trivially.[6]

*Proof of Lemma C.1.* Define $E := \{\exists t, s_t, a_t \text{ s.t. } n_{s_t,a_t} < nd_t^\mu(s_t,a_t)/2\}$. Then combining the multiplicative Chernoff bound (Lemma A.1 in the Appendix) and a union bound over each $t$,$s_t$ and $a_t$, we obtain

$$\begin{aligned}\mathbb{P}[E] &\le \sum_t \sum_{s_t} \sum_{a_t} \mathbb{P}[n_{s_t,a_t} < nd_t^\mu(s_t,a_t)/2] \\ &\le HSA \cdot e^{-\frac{n \cdot \min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}{8}} = HSA \cdot e^{-\frac{n \cdot d_m}{8}} := \delta\end{aligned}$$

solving this for $n$ then provides the stated result. □

Now we define: $N := \min_{t,s_t,a_t} n_{s_t,a_t}$, then above implies $N \ge nd_m/2$ (recall $d_m$ in Assumption 2.2). Now we aggregate only the first $N$ pieces of data in each state-action $(s_t,a_t)$[7] of off-policy data $\mathcal{D}$ and they consist of a new dataset $\mathcal{D}' = \{(s_t, a_t, s_{t+1}^{(i)}, r_t^{(i)}) : i = 1, ..., N; t \in [H]; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$, and is a subset of $\mathcal{D}$. For the rest of paper, we will use either $\mathcal{D}'$ or the original $\mathcal{D}$ to create OPEMA $\widehat{v}^\pi$ (only for theoretical analysis purpose). Whether $\mathcal{D}$ or $\mathcal{D}'$ is used will be stated clearly in each context.

[6] In general, non-accessible state will not affect our results so to make our presentation succinct we will not mention non-accessible state for the rest of paper unless necessary.

[7] Note we can do this since by definition $N \le n_{s_t,a_t}$ for all $s_t, a_t$.

**Remark C.2.** *It is worth mentioning that when use $\mathcal{D}'$ to construct $\widehat{v}^\pi$, $n^{\mathcal{D}'}_{s_t,a_t} = N$ for all $s_t, a_t$. Also, $N := \min n^{\mathcal{D}}_{s_t,a_t}$ (note $n^{\mathcal{D}}_{s_t,a_t}$ is the count from $\mathcal{D}$) itself is a random variable and in the extreme case we could have $N = 0$ and if that happens $\widehat{v}^\pi = 0$ (since in that case $\widehat{P}_t \equiv 0$ and $\widehat{d}^\pi_t$ is degenerated). However, there is only tiny probability $N$ will be small, as guaranteed by Lemma C.1.*

We wanted to point out that this technique of dropping certain amount of data, is not uncommon for analyzing model-based method in RL: e.g. Rmax exploration (Brafman and Tennenholtz, 2002) for online episodic setting (see [Jiang (2018), Notes on Rmax exploration] Section 2 Algorithm for tabular MDP. The data they use is the known set $K$ with parameter $m$, in step3 data pairs observed more than $m$ times are not recorded).

### C.1 Fictitious OPEMA estimator.

Similar to Xie et al. (2019); Yin and Wang (2020), we introduce an unbiased version of $\widehat{v}^\pi$ to fill in the gap at $(s_t, a_t)$ where $n_{s_t,a_t}$ is small. Concretely, every component in $\widehat{v}^\pi$ is substituted by the fictitious counterpart, *i.e.* $\widetilde{v}^\pi := \sum_{t=1}^H \langle \widetilde{d}^\pi_t, \widetilde{r}^\pi_t \rangle$, with $\widetilde{d}^\pi_t = \widetilde{P}^\pi_t \widetilde{d}^\pi_{t-1}$ and $\widetilde{P}^\pi_t(s_t|s_{t-1}) = \sum_{a_{t-1}} \widetilde{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1})$. In particular, consider the high probability event in Lemma C.1, *i.e.* let $E_t$ denotes the event $\{n_{s_t,a_t} \geq n d^\mu_t(s_t, a_t)/2\}$[8], then we define

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

Similarly, for the OPEMA estimator uses data $\mathcal{D}'$, the fictitious estimator is set to be

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

where $E$ denote the event $\{N \geq n d_m/2\}$.

$\widetilde{v}^\pi$ creates a bridge between $\widehat{v}^\pi$ and $v^\pi$ because of its unbiasedness and it is also bounded by $H$ (see Lemma B.3 and Lemma B.5 in Yin and Wang (2020) for those preliminary results). Also, $\widetilde{v}^\pi$ is identical to $\widehat{v}^\pi$ with high probability, as stated by the following lemma.

**Lemma C.3.** *For any $0 < \delta < 1$, there exists an absolute constant $c_1$ such that when total episode $n > c_1 d_m \cdot \log(HSA/\delta)$, then with probability $1 - \delta$,*

$$\sup_{\pi\in\Pi} |\widehat{v}^\pi - \widetilde{v}^\pi| = 0.$$

*Proof.* This Lemma is a direct corollary of Lemma C.1 by considering the event $E_1 := \{\exists t, s_t, a_t \text{ s.t. } n_{s_t,a_t} < n d^\mu_t(s_t, a_t)/2\}$ or $\{N < n d_m/2\}$ since $\widehat{v}^\pi$ and $\widetilde{v}^\pi$ are identical on $E^c_1$. □

Note $\widehat{v}^\pi$ and $\widetilde{v}^\pi$ even equal to each other uniformly over all $\pi$ in $\Pi$. This is not surprising since only logging policy $\mu$ will decide if they are equal or not. This lemma shows how close $\widehat{v}^\pi$ and $\widetilde{v}^\pi$ are. Therefore in the following it suffices to consider the uniform convergence of $\sup_{\pi\in\Pi} |\widetilde{v}^\pi - v^\pi|$.

Next by using a fictitious analogy of state-action expression as in equation (1), we have:

$$\begin{aligned}
\sup_{\pi\in\Pi} |\widetilde{v}^\pi - v^\pi| &= \sup_{\pi\in\Pi} |\sum_{t=1}^H \langle \widetilde{d}^\pi_t, \widetilde{r}_t \rangle - \sum_{t=1}^H \langle d^\pi_t, r_t \rangle| \\
&= \sup_{\pi\in\Pi} |\sum_{t=1}^H \langle \widetilde{d}^\pi_t, \widetilde{r}_t \rangle - \sum_{t=1}^H \langle \widetilde{d}^\pi_t, r_t \rangle + \sum_{t=1}^H \langle \widetilde{d}^\pi_t, r_t \rangle - \sum_{t=1}^H \langle d^\pi_t, r_t \rangle| \\
&\leq \underbrace{\sup_{\pi\in\Pi} |\sum_{t=1}^H \langle \widetilde{d}^\pi_t - d^\pi_t, r_t \rangle|}_{(*)} + \underbrace{\sup_{\pi\in\Pi} |\sum_{t=1}^H \langle \widetilde{d}^\pi_t, \widetilde{r}_t - r_t \rangle|}_{(**)}
\end{aligned} \tag{3}$$

[8] 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.

We first deal with $(**)$ by the following lemma.

**Lemma C.4.** *We have with probability* $1-\delta$*:*

$$\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\le O(\sqrt{\frac{H^2\log(HSA/\delta)}{n\cdot d_m}})$$

*Proof of Lemma C.4.* Since $|\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\le||\widetilde{d}_t^{\pi}||_1\cdot||\widetilde{r}_t-r_t||_\infty$, we obtain

$$|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\le\sum_{t=1}^{H}||\widetilde{d}_t^{\pi}||_1\cdot||\widetilde{r}_t-r_t||_\infty=\sum_{t=1}^{H}||\widetilde{r}_t-r_t||_\infty,$$

where we used $\widetilde{d}_t^{\pi}(\cdot)$ is a probability distribution. Therefore above expression further indicates $\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\le\sum_{t=1}^{H}||\widetilde{r}_t-r_t||_\infty$. Now by a union bound and Hoeffding inequality (Lemma A.2),

$$\begin{aligned}\mathbb{P}(\sup_t||\widetilde{r}_t-r_t||_\infty>\epsilon)&=\mathbb{P}(\sup_{t,s_t,a_t}|\widetilde{r}_t(s_t,a_t)-r_t(s_t,a_t)|>\epsilon)\\&\le HSA\cdot\mathbb{P}(|\widetilde{r}_t(s_t,a_t)-r_t(s_t,a_t)|>\epsilon)\\&=HSA\cdot\mathbb{P}(|\widehat{r}_t(s_t,a_t)-r_t(s_t,a_t)|\mathbb{1}(E_t)>\epsilon)\\&\le 2HSA\cdot\mathbb{E}[\mathbb{E}[e^{-2n_{s_t,a_t}\epsilon^2}|E_t]]\\&\le 2HSA\cdot\mathbb{E}[\mathbb{E}[e^{-nd_m\epsilon^2}|E_t]]=2HSA\cdot e^{-nd_m\epsilon^2}:=\frac{\delta}{2}.\end{aligned}$$

where we use $\mathbb{P}(A)=\mathbb{E}[\mathbb{1}_A]=\mathbb{E}[\mathbb{E}[\mathbb{1}_A|X]]$. Solving for $\epsilon$, then it follows:

$$\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\le\sum_{t=1}^{H}||\widetilde{r}_t-r_t||_\infty\le O(\sqrt{\frac{H^2\log(HSA/\delta)}{n\cdot d_m}})$$

with probability $1-\delta$. The case for $E=\{N\ge nd_m/2\}$ can be proved easily in a similar way.

□

Note that in order to measure the randomness in reward, sample complexity $n$ only has dependence of order $H^2$, this result implies random reward will only cause error of lower order dependence in $H$. Therefore, in many RL literature deterministic reward is directly assumed. Next we consider $(*)$ in (3) by decomposing $\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle$ into a martingale type representation. This is the key for our proof since with it we can use either uniform concentration inequalities or martingale concentration inequalities to prove efficiency.

### C.2 Decomposition of $\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle$

Let $\widetilde{d}_t^{\pi}\in\mathbb{R}^{S\cdot A}$ denote the marginal state-action probability vector, $\pi_t\in\mathbb{R}^{(S\cdot A)\times S}$ is the policy matrix with $(\pi_t)_{(s_t,a_t),s_t}=\pi_t(a_t|s_t)$ and $(\pi_t)_{(s_t,a_t),s}=0$ for $s\ne s_t$. Moreover, let state-action transition matrix $T_t\in\mathbb{R}^{S\times(S\cdot A)}$ to be $(T_t)_{s_t,(s_{t-1},a_{t-1})}=P_t(s_t|s_{t-1},a_{t-1})$, then we have

$$\widetilde{d}_t^{\pi}=\pi_t\widetilde{T}_t\widetilde{d}_{t-1}^{\pi}\tag{4}$$

$$d_t^{\pi}=\pi_tT_td_{t-1}^{\pi}.\tag{5}$$

Therefore we have

$$\widetilde{d}_t^{\pi}-d_t^{\pi}=\pi_t(\widetilde{T}_t-T_t)\widetilde{d}_{t-1}^{\pi}+\pi_tT_t(\widetilde{d}_{t-1}^{\pi}-d_{t-1}^{\pi})\tag{6}$$

recursively apply this formula, we have

$$\widetilde{d}_t^{\pi}-d_t^{\pi}=\sum_{h=2}^{t}\Gamma_{h+1:t}\pi_h(\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^{\pi}+\Gamma_{1:t}(\widetilde{d}_1^{\pi}-d_1^{\pi})\tag{7}$$

where $\Gamma_{h:t}=\prod_{v=h}^{t}\pi_vT_v$ and $\Gamma_{t+1:t}:=1$. Now let $X=\sum_{t=1}^{H}\langle r_t,\widetilde{d}_t^{\pi}-d_t^{\pi}\rangle$, then we have the following:

**Theorem C.5** (martingale decomposition of $X$: Restate of the fictitious version of Lemma 3.1)**.** *We have:*

$$X = \sum_{h=2}^{H} \langle V_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle + \langle V_1^\pi(s), (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle,$$

*where the inner product is taken w.r.t states.*

*Proof of Theorem C.5.* By applying (7) and the change of summation, we have

$$\begin{aligned}
X &= \sum_{t=1}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle + \langle r_t, \Gamma_{1:t}(\widetilde{d}_1^\pi - d_1^\pi)\rangle\right)\\
&= \sum_{t=1}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \sum_{h=1}^{H}\langle r_t, \Gamma_{1:t}(\widetilde{d}_1^\pi - d_1^\pi)\rangle\\
&= \sum_{t=2}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \sum_{h=1}^{H}\langle r_t, \Gamma_{1:t}(\widetilde{d}_1^\pi - d_1^\pi)\rangle\\
&= \sum_{h=2}^{H}\left(\sum_{t=h}^{H}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \sum_{h=1}^{H}\langle(\pi_1^T\Gamma_{1:t}^T r_t)(s), (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle\\
&= \sum_{h=2}^{H}\left(\underbrace{\langle\sum_{t=h}^{H}\pi_h^T\Gamma_{h+1:t}^T r_t, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle}_{V_h^\pi(s)}\right) + \underbrace{\langle(\sum_{h=1}^{H}\pi_1^T\Gamma_{1:t}^T r_t)(s), (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle}_{V_1^\pi(s)}
\end{aligned}$$

□

## D Proof of uniform convergence in OPE with full policies using standard uniform concentration tools: Theorem 3.3

As a reminder for the reader, the OPEMA estimator used in this section is with data subset $\mathcal{D}'$. Also, by Lemma C.4 we only need to consider $\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle|$.

**Theorem D.1.** *There exists an absolute constant $c$ such that if $n > c\cdot\frac{1}{d_m}\cdot\log(HSA/\delta)$, then with probability $1-\delta$, we have:*

$$\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right| \le O(\sqrt{\frac{H^4\log(HSA/\delta)}{nd_m}}) + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right|\right]$$

*Proof of Theorem D.1.* Note in data $\mathcal{D}' = \{(s_t, a_t, s_{t+1}^{(i)}) : i = 1, ..., N; t = 1, ..., H; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$[9], not only $s_{t+1}^{(i)}$ but also $N$ are random variables.

We first conditional on $N$, then $(s_t, a_t, s_{t+1}^{(i)})$'s are independent samples for all $i, s_t, a_t$ since any sample will not contain information about other samples. Therefore we can regroup $\mathcal{D}'$ into $N$ independent samples with $\mathcal{D}' = \{X^{(i)} : i = 1, ..., N\}$ where $X^{(i)} = \{(s_t, a_t, s_{t+1}^{(i)}), t = 1, ..., H; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$. Now for any $i_0$, change $X^{(i_0)}$ to $X'^{(i_0)} = \{(s_t, a_t, s_{t+1}'^{(i_0)}), t = 1, ..., H; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$ and keep the rest $N-1$ data the same, use this data to

[9]Here we do not include $r_t^{(i)}$ since the quantity $\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle|$ only contains the mean reward function $r_t$.

create new estimator with state-action transition $\widetilde{d}'^{\pi}$, then we have

$$\begin{aligned}
&\left|\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle|-\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t'^{\pi}-d_t^{\pi},r_t\rangle|\right| \\
\leq&\sup_{\pi\in\Pi}\left||\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle|-|\sum_{t=1}^{H}\langle\widetilde{d}_t'^{\pi}-d_t^{\pi},r_t\rangle|\right| \\
\leq&\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle-\sum_{t=1}^{H}\langle\widetilde{d}_t'^{\pi}-d_t^{\pi},r_t\rangle\right| \\
=&\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-\widetilde{d}_t'^{\pi},r_t\rangle\right| \\
=&\sup_{\pi\in\Pi}\left|\sum_{h=2}^{H}\langle\widetilde{V}_h'^{\pi},(\widetilde{T}_h-\widetilde{T}_h')\widetilde{d}_{h-1}^{\pi}\rangle+\langle\widetilde{V}_1'^{\pi},\widetilde{d}_1^{\pi}-\widetilde{d}_1'^{\pi}\rangle\right|,
\end{aligned}$$

where the last equation comes from the trick that substitutes $d_t^{\pi}$ by $\widetilde{d}_t^{\pi}$ in Theorem C.5. By definition, the above equals to

$$\begin{aligned}
=&\sup_{\pi\in\Pi}\left|\sum_{h=2}^{H}\langle\widehat{V}_h'^{\pi},(\widehat{T}_h-\widehat{T}_h')\widehat{d}_{h-1}^{\pi}\rangle+\langle\widehat{V}_1'^{\pi},\widehat{d}_1^{\pi}-\widehat{d}_1'^{\pi}\rangle\right|\cdot\mathbb{1}(E) \\
\leq&\sup_{\pi\in\Pi}\left(\sum_{h=2}^{H}||(\widehat{T}_h-\widehat{T}_h')^T\widehat{V}_h'^{\pi}||_{\infty}||\widehat{d}_{h-1}^{\pi}||_1+|\langle\widehat{V}_1'^{\pi},\widehat{d}_1^{\pi}-\widehat{d}_1'^{\pi}\rangle|\right)\cdot\mathbb{1}(E) \\
\leq&\sup_{\pi\in\Pi}\left(\sum_{h=2}^{H}||(\widehat{T}_h-\widehat{T}_h')^T\widehat{V}_h'^{\pi}||_{\infty}+|\langle\widehat{V}_1'^{\pi},\widehat{d}_1^{\pi}-\widehat{d}_1'^{\pi}\rangle|\right)\cdot\mathbb{1}(E)
\end{aligned}$$

Note the change of a single $X^{(i_0)}$ will only change two entries of each row of $(\widehat{T}_h-\widehat{T}_h')^T$ by $1/N$ since with data $\mathcal{D}'$, $n_{s_t,a_t}=N$ for all $s_t,a_t$. Or in other words, given $E$,

$$\widehat{T}_h^T-\widehat{T}_h'^T=\begin{bmatrix}0&...&0&\frac{1}{N}&0&...&-\frac{1}{N}&...&0\\0&\frac{1}{N}&0&...&-\frac{1}{N}&...&...&...&0\\.....&&&&&&&&\\-\frac{1}{N}&0&...&0&...&...&0&...&\frac{1}{N}\end{bmatrix},$$

where the locations of $1/N,-1/N$ in each row are random as it depends on how different is $X'^{(i_0)}$ from $X^{(i_0)}$. However, based on this fact, it is enough for us to guarantee

$$||(\widehat{T}_h-\widehat{T}_h')^T\widehat{V}_h'^{\pi}||_{\infty}\leq\frac{2}{N}||\widehat{V}_h'^{\pi}||_{\infty}\leq\frac{2}{N}(H-h+1)\leq\frac{2}{N}H$$

and same result holds for $|\langle\widehat{V}_1'^{\pi},\widehat{d}_1^{\pi}-\widehat{d}_1'^{\pi}\rangle|\leq 2H/N$ given $N$.

Combine all the results above, for a single change of $X^{(i_0)}$ we have

$$\left|\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle|-\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t'^{\pi}-d_t^{\pi},r_t\rangle|\right|\leq 2\frac{H^2}{N}\mathbb{1}(E)\leq 2\frac{H^2}{N}$$

for any fixed $N$. If we let $Z=S(X^{(1)},...,X^{(N)})=\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle|$, then for a given $N$ by independence and above bounded difference condition we can apply Mcdiarmid inequality Lemma A.4 (where $\xi_i=2H^2/N$) to obtain

$$\mathbb{P}(|Z-\mathbb{E}[Z]|\geq\epsilon|N)\leq 2e^{-\frac{N\epsilon^2}{2H^4}}:=\frac{\delta}{2}\tag{8}$$

Now note when $n > O(\frac{1}{d_m} \cdot \log(HSA/\delta))$, by Lemma C.1 we can obtain $N > nd_m/2$ with probability $1-\delta/2$, combining this result and solving $\epsilon$ in (8), we have

$$\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle\right| \le O(\sqrt{\frac{H^4\log(HSA/\delta)}{n\cdot d_m}}) + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle\right|\right]$$

with probability $1-\delta$.

□

Next before bounding $\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle\right|\right]$, we first give a useful lemma.

Let $\gamma\in(0,1)$ to be any threshold parameter. Then we first have the following lemma:

**Lemma D.2.** *Recall by definition $P_h(s_h,|s_{h-1},a_{h-1}) = T_h(s_h,|s_{h-1},a_{h-1})$. It holds that with probability $1-\delta$, for all $t, s_t, a_t \in [H], \mathcal{S}, \mathcal{A}$: if $P_h(s_h|s_{h-1},a_{h-1}) \le \gamma$, then*

$$\left|\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1})\right| \le \sqrt{\frac{\gamma\log(HSA/\delta)}{2nd_m}} + \frac{2\log(HSA/\delta)}{3nd_m};$$

*if $P_h(s_h,|s_{h-1},a_{h-1}) > \gamma$, then*

$$\left|\frac{\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1})}{T_h(s_h|s_{h-1},a_{h-1})}\right| \le \sqrt{\frac{\log(HSA/\delta)}{2nd_m\gamma}} + \frac{2\log(HSA/\delta)}{3nd_m\gamma};$$

*Proof.* First consider the case where $P_h(s_h|s_{h-1},a_{h-1}) \le \gamma$.

$$\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1}) = \frac{1}{n_{s_{h-1},a_{h-1}}}\sum_{i=1}^{n_{s_{h-1},a_{h-1}}}\left(\mathbb{1}[s_h^{(i)}] - T_h(s_h|s_{h-1},a_{h-1})\right)\mathbb{1}(E_h),$$

since $\mathrm{Var}[\mathbb{1}[s_h^{(i)}]|s_{h-1},a_{h-1}] = P_h(s_h|s_{h-1},a_{h-1})(1-P_h(s_h|s_{h-1},a_{h-1})) \le P_h(s_h|s_{h-1},a_{h-1}) \le \gamma$, therefore by Lemma A.3,

$$\left|\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1})\right| \le \mathbb{1}(E_h)\left(\sqrt{\frac{\gamma\log(1/\delta)}{n_{s_{h-1},a_{h-1}}}} + \frac{2\log(1/\delta)}{n_{s_{h-1},a_{h-1}}}\right) \le \sqrt{\frac{\gamma\log(1/\delta)}{2nd_m}} + \frac{2\log(1/\delta)}{3nd_m};$$

Second, when $P_h(s_h|s_{h-1},a_{h-1}) > \gamma$.

$$\frac{\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1})}{T_h(s_h|s_{h-1},a_{h-1})} = \frac{1}{n_{s_{h-1},a_{h-1}}}\sum_{i=1}^{n_{s_{h-1},a_{h-1}}}\left(\frac{\mathbb{1}[s_h^{(i)}]}{T_h(s_h|s_{h-1},a_{h-1})} - 1\right)\mathbb{1}(E_h),$$

since

$$\mathrm{Var}\left[\frac{\mathbb{1}[s_h^{(i)}]}{T_h(s_h|s_{h-1},a_{h-1})}\middle|s_{h-1},a_{h-1}\right] \le \frac{1}{T_h(s_h|s_{h-1},a_{h-1})^2}\mathrm{Var}\left[\mathbb{1}[s_h^{(i)}]\middle|s_{h-1},a_{h-1}\right] \le \frac{1}{T_h(s_h|s_{h-1},a_{h-1})} \le \frac{1}{\gamma},$$

and since $\frac{\mathbb{1}[s_h^{(i)}]}{T_h(s_h|s_{h-1},a_{h-1})} \le 1/\gamma$, again be Bernstein inequality we have

$$\left|\frac{\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1})}{T_h(s_h|s_{h-1},a_{h-1})}\right| \le \sqrt{\frac{\log(1/\delta)}{2nd_m\gamma}} + \frac{2\log(1/\delta)}{3nd_m\gamma};$$

apply the union bound over $t, s_t, a_t$ we obtain the stated result. □

**Bounding $\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right|\right]$.** First note by Theorem C.5:

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right|\right] \le \sum_{h=2}^{H}\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle v_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right] + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_1^\pi(s), (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle\right|\right],$$

so it suffices to bound each $\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right]$. First of all,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right] \\
=&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\right|\right] \\
\le&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\right| \cdot \mathbb{1}[T_h(s_h|s_{h-1},a_{h-1}) > \gamma]\right] \\
+&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\right| \cdot \mathbb{1}[T_h(s_h|s_{h-1},a_{h-1}) \le \gamma]\right] \\
=&\underbrace{\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)T_h(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\frac{\widetilde{T}_h - T_h}{T_h}(s_h|s_{h-1},a_{h-1})\right| \cdot \mathbb{1}[T_h > \gamma]\right]}_{(a)} \\
+&\underbrace{\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\right| \cdot \mathbb{1}[T_h(s_h|s_{h-1},a_{h-1}) \le \gamma]\right]}_{(b)},
\end{aligned}$$

Apply Lemma D.2 with $\delta'/2$ where $\delta' = \delta/H$, then

$$\begin{aligned}
(a) \le& \sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)T_h(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\left(\sqrt{\frac{\log(2HSA/\delta')}{2nd_m\gamma}} + \frac{2\log(2HSA/\delta')}{3nd_m\gamma}\right)\right|(1-\frac{\delta'}{2}) \\
&+H\delta'/2 \\
\le& \sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)T_h(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\left(\sqrt{\frac{\log(2H^2SA/\delta)}{2nd_m\gamma}} + \frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)\right| \\
&+\delta/2 \\
\le& \sup_{\pi\in\Pi}\left|H\left(\sqrt{\frac{2\log(H^2SA/\delta)}{2nd_m\gamma}} + \frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)\right| + \delta/2 = H\left(\sqrt{\frac{\log(2H^2SA/\delta)}{2nd_m\gamma}} + \frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right) + \delta/2,
\end{aligned}$$

$$\begin{aligned}
(b) \le& \sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\left(\sqrt{\frac{\gamma\log(2HSA/\delta)}{2nd_m}} + \frac{2\log(2HSA/\delta)}{3nd_m}\right)\right|(1-\frac{\delta'}{2}) + H\frac{\delta'}{2} \\
\le& \sup_{\pi\in\Pi}\left|HS\left(\sqrt{\frac{\gamma\log(2H^2SA/\delta)}{2nd_m}} + \frac{2\log(2H^2SA/\delta)}{3nd_m}\right)\right| + \frac{\delta}{2} = HS\left(\sqrt{\frac{\gamma\log(2H^2SA/\delta)}{2nd_m}} + \frac{2\log(2H^2SA/\delta)}{3nd_m}\right) + \frac{\delta}{2},
\end{aligned}$$

Hence we have for any $\gamma$,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s),((\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right]\\ \leq& H\left(\sqrt{\frac{\log(2H^2SA/\delta)}{2nd_m\gamma}}+\frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)+HS\left(\sqrt{\frac{\gamma\log(2H^2SA/\delta)}{2nd_m}}+\frac{2\log(2H^2SA/\delta)}{3nd_m}\right)+\delta\end{aligned}$$

In particular, choose $\gamma=1/S<1$, then above becomes

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s),((\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right]\leq\sqrt{\frac{2H^2S\log(2H^2SA/\delta)}{nd_m}}+\frac{4HS\log(2H^2SA/\delta)}{3nd_m}+\delta$$

Critically, above holds for any $\forall 1>\delta>0$. Based on theorem condition $n>c\cdot 1/d_m\log(HSA/\theta)>c\cdot 1/d_m$[10], choose $\delta=\frac{c}{nd_m}$, then above is further less equal to

$$\sqrt{\frac{2H^2S\log(2nH^2SA)}{nd_m}}+\frac{4HS\log(2nH^2SA)}{3nd_m}+\frac{c}{nd_m}\leq\sqrt{\frac{2H^2S\log(2nH^2SA)}{nd_m}}+C\cdot\frac{HS\log(2nH^2SA)}{3nd_m}$$

where $C$ is a new constant absorbs $1/nd_m$. If we further reducing it to

Finally, summing over all $H$, and again using new constant $C'$ to absorb higher order term, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^H\langle\widetilde{d}_t^\pi-d_t^\pi,r_t\rangle\right|\right]\leq C'\sqrt{\frac{H^4S\log(nHSA)}{nd_m}}$$

Combing this with Theorem D.1 and Lemma C.4, we have proved Theorem 3.3.

**Remark D.3.** *The key for proving this uniform convergence bound is that applying concentration inequality only to terms that are independent of the policies, i.e. $\widetilde{T}_h(s_h|s_{h-1},a_{h-1})-T_h(s_h|s_{h-1},a_{h-1})$. Therefore when taking supremum over policies, high probability event holds with same probability without decay.*

# E Proof of uniform convergence in OPE with deterministic policies using martingale concentration inequalities: Theorem 3.5

A reminder that all results in this section use data $\mathcal{D}$ for OPEMA estimator $\widehat{v}^\pi$.

## E.1 Martingale concentration result on $\sum_{t=1}^H\langle\widetilde{d}_t^\pi-d_t^\pi,r_t\rangle$.

Let $X=\sum_{t=1}^H\langle\widetilde{d}_t^\pi-d_t^\pi,r_t\rangle$ and $\mathcal{D}_h:=\{s_t^{(i)},a_t^{(i)}:t=1,...,h\}_{i=1}^n$. Since $\mathcal{D}_h$ forms a filtration, then by law of total expectation we have $X_t=\mathbb{E}[X|\mathcal{D}_t]$ is martingale. Moreover, we have

**Lemma E.1.**

$$X_t:=\mathbb{E}[X|\mathcal{D}_t]=\sum_{h=2}^t\langle V_h^\pi,(\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi\rangle+\langle V_1^\pi,\widetilde{d}_1^\pi-d_1^\pi\rangle.$$

*Proof of Lemma E.1.* By martingale decomposition Theorem C.5 and note $\widetilde{T}_i,\widetilde{d}_i^\pi$ are measurable *w.r.t.* $\mathcal{D}_t$ for $i=1,...,t$, so we have

$$\mathbb{E}[X|\mathcal{D}_t]=\sum_{h=t+1}^H\mathbb{E}\left[\langle V_h^\pi,(\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi\rangle\middle|\mathcal{D}_t\right]+\sum_{h=2}^t\langle V_h^\pi,(\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi\rangle+\langle V_1^\pi,(\widetilde{d}_1^\pi-d_1^\pi)\rangle.$$

[10] Note the $\theta$ in $\log(HSA/\theta)$ is identical to the failure probability in Theorem D.1

Note for $h \geq t+1$, $\mathcal{D}_t \subset \mathcal{D}_{h-1}$, so by total law of expectation (tower property) we have

$$\begin{aligned}
&\mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle \Big| \mathcal{D}_t\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle \Big| \mathcal{D}_{h-1}\right] \Big| \mathcal{D}_t\right] \\
=&\mathbb{E}\left[\langle V_h^\pi, \mathbb{E}\left[(\widetilde{T}_h - T_h)\Big|\mathcal{D}_{h-1}\right]\widetilde{d}_{h-1}^\pi\rangle \Big| \mathcal{D}_t\right] = 0
\end{aligned}$$

where the last equality uses $\widetilde{T}_h$ is unbiased of $T_h$ given $\mathcal{D}_{h-1}$. This gives the desired result. □

Next we show martingale difference $|X_t - X_{t-1}|$ is bounded with high probability.

**Lemma E.2.** *With probability* $1-\delta$,

$$\sup_t |X_t - X_{t-1}| \leq O(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}).$$

*Proof.*

$$|X_t - X_{t-1}| = \langle V_t^\pi, (\widetilde{T}_t - T_t)\widetilde{d}_{t-1}^\pi\rangle \leq ||(\widetilde{T}_t - T_t)^T V_t^\pi||_\infty ||\widetilde{d}_{t-1}^\pi||_1 = ||(\widetilde{T}_t - T_t)^T V_t^\pi||_\infty.$$

For any fixed pair $(s_t, a_t)$, we have

$$\begin{aligned}
&((\widetilde{T}_t - T_t)^T V_t^\pi)(s_{t-1}, a_{t-1}) \\
=&\mathbb{1}(E_{t-1}) \cdot ((\widehat{T}_t - T_t)^T V_t^\pi)(s_{t-1}, a_{t-1}) \\
=&\mathbb{1}(E_{t-1}) \cdot \sum_{s_t} V_t^\pi(s_t)(\widehat{T}_t - T_t)(s_t|s_{t-1}, a_{t-1}) \\
=&\mathbb{1}(E_{t-1}) \cdot \left(\sum_{s_t} V_t^\pi(s_t)\widehat{T}_t(s_t|s_{t-1}, a_{t-1}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}]\right) \\
=&\mathbb{1}(E_{t-1}) \cdot \left(\sum_{s_t} V_t^\pi(s_t)\frac{1}{n_{s_{t-1},a_{t-1}}}\sum_{i=1}^n \mathbb{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}]\right) \\
=&\mathbb{1}(E_{t-1})\left(\frac{1}{n_{s_{t-1},a_{t-1}}}\sum_{i=1}^n V_t^\pi(s_t^{(i)})\mathbb{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}]\right) \\
=&\mathbb{1}(E_{t-1})\left(\frac{1}{n_{s_{t-1},a_{t-1}}}\sum_{i:s_{t-1}^{(i)}=s_{t-1}, a_{t-1}^{(i)}=a_{t-1}} V_t^\pi(s_t^{(i)}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}]\right),
\end{aligned}$$

where the fourth line uses the definition of $\widehat{T}_t$ and the fifth line uses the fact $\sum_{s_t} V_t^\pi(s_t)\mathbb{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1}) = V_t^\pi(s_t^{(i)})\mathbb{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1})$.

Note $||V_t^\pi(\cdot)||_\infty \leq H$ and also conditional on $E_t$, $n_{s_t,a_t} \geq n d_t^\mu(s_t, a_t)/2$, therefore by Hoeffding's inequality and a Union bound we obtain with probability $1-\delta$

$$\sup_t |X_t - X_{t-1}| \leq O(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot \min_{t,s_t,a_t} d_t^\mu(s_t, a_t)}}) = O(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}).$$

□

Next we calculate the conditional variance of $\mathrm{Var}[X_{t+1}|\mathcal{D}_t]$.

**Lemma E.3.** *We have the following decomposition of conditional variance:*

$$\mathrm{Var}[X_{t+1}|\mathcal{D}_t] = \sum_{s_t,a_t} \frac{\widetilde{d}_t^\pi(s_t, a_t)^2 \cdot \mathbb{1}(E_t)}{n_{s_t,a_t}} \cdot \mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)} = s_t, a_t^{(1)} = a_t]$$

*Proof.* Indeed,

$$\begin{aligned}
\mathrm{Var}[X_{t+1}|\mathcal{D}_t] &= \mathrm{Var}\left[\sum_{s_t,a_t}\sum_{s_{t+1}} V^\pi_{t+1}(s_{t+1})(\widetilde{T}-T)(s_{t+1}|s_t,a_t)\widetilde{d}^\pi_t(s_t,a_t)\middle|\mathcal{D}_t\right] \\
&= \sum_{s_t,a_t}\mathrm{Var}\left[\sum_{s_{t+1}} V^\pi_{t+1}(s_{t+1})(\widetilde{T}-T)(s_{t+1}|s_t,a_t)\middle|\mathcal{D}_t\right]\widetilde{d}^\pi_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\mathbb{1}(E_t)\cdot\mathrm{Var}\left[\sum_{s_{t+1}} V^\pi_{t+1}(s_{t+1})\widehat{T}(s_{t+1}|s_t,a_t)\middle|\mathcal{D}_t\right]\widetilde{d}^\pi_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\mathbb{1}(E_t)\cdot\mathrm{Var}\left[\sum_{s_{t+1}} V^\pi_{t+1}(s_{t+1})\frac{1}{n_{s_t,a_t}}\sum_{i=1}^n \mathbb{1}(s^{(i)}_{t+1}=s_{t+1},s^{(i)}_t=s_t,a^{(i)}_t=a_t)\middle|\mathcal{D}_t\right]\widetilde{d}^\pi_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\frac{\mathbb{1}(E_t)}{n^2_{s_t,a_t}}\cdot\mathrm{Var}\left[\sum_{i:s^{(i)}_t=s_t,a^{(i)}_t=a_t} V^\pi_{t+1}(s^{(i)}_{t+1})\middle|\mathcal{D}_t\right]\widetilde{d}^\pi_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\frac{\widetilde{d}^\pi_t(s_t,a_t)^2\cdot\mathbb{1}(E_t)}{n_{s_t,a_t}}\cdot\mathrm{Var}[V^\pi_{t+1}(s^{(1)}_{t+1})|s^{(1)}_t=s_t,a^{(1)}_t=a_t]
\end{aligned} \tag{9}$$

where the second equal sign comes from the fact that when conditional on $\mathcal{D}_t$, we can separate $n$ episodes into $SA$ groups and episodes from different groups are independent of each other. The third uses $\mathbb{1}(E_t)$ is measurable w.r.t $\mathcal{D}_t$. Similarly, the last equal sign again uses $n_{s_t,a_t}$ episodes are independent given $\mathcal{D}_t$. □

**Lemma E.4** (Yin and Wang (2020) Lemma 3.4)**.** *For any policy $\pi$ and any MDP.*

$$\begin{aligned}
\mathrm{Var}_\pi\left[\sum_{t=1}^H r^{(1)}_t\right] = \sum_{t=1}^H\Big(&\mathbb{E}_\pi\left[\mathrm{Var}\left[r^{(1)}_t + V^\pi_{t+1}(s^{(1)}_{t+1})\middle|s^{(1)}_t,a^{(1)}_t\right]\right] \\
&+\mathbb{E}_\pi\left[\mathrm{Var}\left[\mathbb{E}[r^{(1)}_t + V^\pi_{t+1}(s^{(1)}_{t+1})|s^{(1)}_t,a^{(1)}_t]\middle|s^{(1)}_t\right]\right]\Big).
\end{aligned}$$

This Lemma suggests if we can bound $\widetilde{d}^\pi_t$ by $O(d^\pi_t)$ with high probability, then by Lemma E.3 we have w.h.p

$$\sum_{t=1}^H \mathrm{Var}[X_{t+1}|\mathcal{D}_t] \le O(\frac{1}{nd_m}\cdot\sum_{t=1}^H\mathbb{E}[\mathrm{Var}[V^\pi_{t+1}(s^{(1)}_{t+1})|s^{(1)}_t,a^{(1)}_t]]) \le O(\frac{H^2}{nd_m})$$

Note this gives only $H^2$ dependence for $\sum_{t=1}^H \mathrm{Var}[X_{t+1}|\mathcal{D}_t]$ instead of a naive bound with $H^3$ and helps us to save a $H$ factor.

Next we show how to bound $\widetilde{d}^\pi_t$.

### E.2 Bounding $\widetilde{d}^\pi_t(s_t,a_t) - d^\pi_t(s_t,a_t)$

Our analysis is based on using martingale structure to derive bound on $\widetilde{d}^\pi_t(s_t,a_t) - d^\pi_t(s_t,a_t)$ for fixed $t,s_t,a_t$ with probability $1-\delta/HSA$, then use a union bound to get a bound for all $\widetilde{d}^\pi_t(s_t,a_t) - d^\pi_t(s_t,a_t)$ with probability $1-\delta$.

Concretely, in (7) if we only extract the specific $(s_t,a_t)$, then we have

$$\widetilde{d}^\pi_t(s_t,a_t) - d^\pi_t(s_t,a_t) = \sum_{h=2}^t(\Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}^\pi_{h-1})(s_t,a_t) + (\Gamma_{1:t}(\widetilde{d}^\pi_1 - d^\pi_1))(s_t,a_t),$$

here $\widetilde{d}^\pi_t(s_t,a_t) - d^\pi_t(s_t,a_t)$ already forms a martingale with filtration $\mathcal{F}_t = \sigma(\mathcal{D}_t)$ and $(\Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}^\pi_{h-1})(s_t,a_t)$ is the corresponding martingale difference since

$$\mathbb{E}[(\Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}^\pi_{h-1})(s_t,a_t)|\mathcal{F}_{h-1}] = (\Gamma_{h+1:t}\pi_h\mathbb{E}[(\widetilde{T}_h - T_h)|\mathcal{F}_{h-1}]\widetilde{d}^\pi_{h-1})(s_t,a_t) = 0.$$

Now we fix specific $(s_t, a_t)$. Then denote $(\Gamma_{h+1:t}\pi_h)(s_t, a_t) := \Gamma'_{h:t} \in \mathbb{R}^{1\times S}$, then we have

$$|(\Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}^{\pi}_{h-1})(s_t, a_t)| = |\Gamma'_{h:t}(\widetilde{T}_h - T_h)\widetilde{d}^{\pi}_{h-1}| = |\langle(\widetilde{T}_h - T_h)^T\Gamma'^T_{h:t}, \widetilde{d}^{\pi}_{h-1}\rangle| \le ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty \cdot 1.$$

Note here $\Gamma'_{h:t}(\widetilde{T}_h - T_h)$ is a row vector with dimension $SA$.

**Bounding** $||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty$ .

In fact, for any given $(s_{h-1}, a_{h-1})$, we have

$$\begin{aligned}
&\Gamma'_{h:t}(\widetilde{T}_h - T_h)(s_{h-1}, s_{h-1}) = \mathbb{1}(E_t)\cdot\Gamma'_{h:t}(\widehat{T}_h - T_h)(s_{h-1}, a_{h-1})\\
=&\mathbb{1}(E_t)\cdot\Gamma'_{h:t}\left(\frac{1}{n_{s_{t-1},a_{t-1}}}\sum_{i:s^{(i)}_{h-1}=s_{h-1},a^{(i)}_{h-1}=a_{h-1}}\mathbf{e}_{s^{(i)}_h} - \mathbb{E}[\mathbf{e}_{s^{(1)}_h}|s^{(1)}_{h-1}=s_{h-1}, a^{(1)}_{h-1}=a_{h-1}]\right)\\
=&\mathbb{1}(E_t)\left(\frac{1}{n_{s_{t-1},a_{t-1}}}\sum_{i:s^{(i)}_{h-1}=s_{h-1},a^{(i)}_{h-1}=a_{h-1}}\Gamma'_{h:t}(s^{(i)}_h) - \mathbb{E}[\Gamma'_{h:t}(s^{(1)}_h)|s^{(1)}_{h-1}=s_{h-1}, a^{(1)}_{h-1}=a_{h-1}]\right)
\end{aligned}$$

Note by definition $\Gamma'_{h:t}(s^{(i)}_h) \le 1$, since $(\Gamma_{h+1:t}\pi_h)(s_t, a_t) := \Gamma'_{h:t} \in \mathbb{R}^{1\times S}$ and $\Gamma_{h+1:t}, \pi_h$ are just probability transitions. Therefore by Hoeffding's inequality and law of total expectation, we have

$$\begin{aligned}
&\mathbb{P}\left(|\Gamma'_{h:t}(\widetilde{T}_h - T_h)(s_{h-1}, a_{h-1})| > \epsilon\right) = \mathbb{P}\left(|\Gamma'_{h:t}(\widehat{T}_h - T_h)(s_{h-1}, a_{h-1})| > \epsilon\middle|E_t\right)\\
&\le \mathbb{E}\left[\exp(-\frac{2n_{s_{h-1},a_{h-1}}\cdot\epsilon^2}{1})\middle|E_t\right] \le \exp(-\frac{nd^{\mu}_{h-1}(s_{h-1}, a_{h-1})\cdot\epsilon^2}{1})
\end{aligned}$$

and apply a union bound to get

$$\begin{aligned}
&P(\sup_h||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty > \epsilon) \le H\cdot\sup_h P(||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty > \epsilon)\\
\le& HSA\cdot\sup_{h,s_{h-1},a_{h-1}}\mathbb{P}\left(|\Gamma'_{h:t}(\widetilde{T}_h - T_h)(s_{h-1}, a_{h-1})| > \epsilon\right)\\
\le& HSA\cdot\exp(-\frac{n\min d^{\mu}_{h-1}(s_{h-1}, a_{h-1})\cdot\epsilon^2}{1}) := \frac{\delta}{HSA}
\end{aligned} \tag{10}$$

Let the right hand side of (10) to be $\delta/HSA$, then we have w.p. $1-\delta/HSA$,

$$\sup_h||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty \le O(\sqrt{\frac{1}{n\cdot d_m}\log\frac{H^2S^2A^2}{\delta}}). \tag{11}$$

**Go back to bounding** $\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)$**.** By Azuma-Hoeffding's inequality (Lemma A.5), we have[11]

$$\mathbb{P}(|\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)| > \epsilon) \le \exp(-\frac{\epsilon^2}{\sum_{i=1}^t(\sup_h||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty)^2}) := \delta/HSA,$$

where $\sum_{i=1}^t(\sup_h||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty)^2$ is the sum of bounded square differences in Azuma-Hoeffding's inequality. Therefore we have w.p. $1-\delta/HSA$,

$$|\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)| \le O(\sqrt{t\cdot(\sup_h||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty)^2\log\frac{HSA}{\delta}}), \tag{12}$$

[11] To be more precise here we actually use a weaker version of Azuma-Hoeffding's inequality, see Remark E.7.

combining (11) with above we further have that w.p. $1-2\delta/HSA$,

$$|\widetilde{d}_t^\pi(s_t,a_t) - d_t^\pi(s_t,a_t)| \le O(\sqrt{\frac{t}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}})$$

**Lastly**, by a union bound and simple scaling (from $2\delta$ to $\delta$) we have w.p. $1-\delta$

$$\sup_t ||\widetilde{d}_t^\pi - d_t^\pi||_\infty \le O(\sqrt{\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}}).$$

This implies that w.p. $1-\delta$, $\forall t, s_t, a_t$,

$$\widetilde{d}_t^\pi(s_t,a_t)^2 \le 2d_t^\pi(s_t,a_t)^2 + O(\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}). \tag{13}$$

Combining (13) with Lemma E.4 and Lemma E.3, we obtain:

**Lemma E.5.** *With probability* $1-\delta$,

$$\sum_{t=1}^H \mathrm{Var}[X_{t+1}|\mathcal{D}_t] \le O(\frac{H^2}{nd_m}) + O(\frac{H^4SA}{n^2d_m^2}\cdot\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}) \tag{14}$$

*Proof of Lemma E.5.* By (13) and Lemma E.3, we have $\forall t$, with probability ay least $1-\delta$,

$$\begin{aligned}
&\mathrm{Var}[X_{t+1}|\mathcal{D}_t] \le \sum_{s_t,a_t} O(\frac{\widetilde{d}_t^\pi(s_t,a_t)^2\cdot}{nd_m})\cdot\mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)}=s_t,a_t^{(1)}=a_t]\\
&\le \sum_{s_t,a_t} O(\frac{1}{nd_m})\left(2d_t^\pi(s_t,a_t)^2 + O(\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})\right)\cdot\mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)}=s_t,a_t^{(1)}=a_t]\\
&\le \sum_{s_t,a_t} O(\frac{1}{nd_m})\left(2d_t^\pi(s_t,a_t) + O(\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})\right)\cdot\mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)}=s_t,a_t^{(1)}=a_t]\\
&\le O(\frac{1}{nd_m})\mathbb{E}\left[\mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)},a_t^{(1)}]\right] + O(\frac{1}{nd_m}\cdot\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}\cdot H^2SA)\\
&= O(\frac{1}{nd_m})\mathbb{E}\left[\mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)},a_t^{(1)}]\right] + O(\frac{H^3SA}{n^2d_m^2}\cdot\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})
\end{aligned}$$

then sum over $t$ and apply Lemma E.4 gives the stated result. □

Combining all the results, we are able to prove:

**Theorem E.6.** *With probability* $1-\delta$, *we have*

$$\left|\sum_{t=1}^H\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right| \le O(\sqrt{\frac{H^2\log(HSA/\delta)}{nd_m}} + \sqrt{\frac{H^4SA\cdot\log(H^2S^2A^2/\delta)\log(HSA/\delta)}{n^2d_m^2}})$$

*where* $O(\cdot)$ *absorbs only the absolute constants.*

*Proof of Theorem E.6.* Recall $X = \sum_{t=1}^H\langle\widetilde{d}_t^\pi - d_t^\pi, r_t\rangle$ and by law of total expectation it is easy to show $E[X]=0$. Next denote $\sigma^2 = O(\frac{H^2}{nd_m}) + O(\frac{H^4SA}{n^2d_m^2}\cdot\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})$ as in Lemma E.5 and also let $M = \sup_t|X_t - X_{t-1}|$. Then by Freedman inequality (Lemma A.6), we have with probability $1-\delta/3$,

$$|X - \mathbb{E}[X]| \le \sqrt{8\sigma^2\cdot\log(3/\delta)} + \frac{2M}{3}\cdot\log(3/\delta), \quad \text{Or} \quad W \ge \sigma^2.$$

where $W = \sum_{t=1}^{H} \mathrm{Var}[X_{t+1}|\mathcal{D}_t]$. Next by Lemma E.5, we have $\mathbb{P}(W \geq \sigma^2) \leq 1/3\delta$, this implies with probability $1 - 2\delta/3$,

$$|X - \mathbb{E}[X]| \leq \sqrt{8\sigma^2 \cdot \log(3/\delta)} + \frac{2M}{3} \cdot \log(3/\delta).$$

Finally, by Lemma E.2, we have $\mathbb{P}(M \geq O(\sqrt{\frac{H^2 \log(HSA/\delta)}{nd_m}})) \leq \delta/3$. Also use $\mathbb{E}[X] = 0$, we have with probability $1 - \delta$,

$$|X| \leq \sqrt{8\sigma^2 \cdot \log(3/\delta)} + O(\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{nd_m}} \log(3/\delta)).$$

Plugging back the expression of $\sigma^2 = O(\frac{H^2}{nd_m}) + O(\frac{H^4 SA}{n^2 d_m^2} \cdot \log \frac{H^2 S^2 A^2}{\delta} \log \frac{HSA}{\delta})$ and assimilating the same order terms give the desired result. □

**Remark E.7.** *Rigorously, standard Azuma-Hoeffding's inequality Lemma A.5 does not apply to (12) since* $\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty$ *is not a deterministic upper bound, we only have the difference bound with high probability sense, see (11). Therefore, strictly speaking, we need to apply Theorem 32 in Chung and Lu (2006) which is a weaker Azuma-Hoeffding's inequality allowing bounded difference with high probability. The same logic applies for a weaker freedman's inequality consisting of Theorem 34 and Theorem 37 in Chung and Lu (2006) since our martingale difference* $M = \sup_t |X_t - X_{t-1}|$ *in the proof of Theorem E.6 is bounded with high probability. We avoid explicitly using them in order to make our proofs more readable for our readers.*

We end this section by giving the proofs of Theorem 3.4 and Theorem 3.5.

*Proof of Lemma 3.4 and Theorem 3.5.* The proof of Lemma 3.4 comes from Lemma C.3, Lemma C.4 and Theorem E.6. The proof of Theorem 3.5 relies on applying a union bound over $\Pi$ in Theorem 3.4 (recall all non-stationary deterministic policies have $|\Pi| = A^{HS}$), then extra dependence of $\sqrt{\log(|\Pi|)} = \sqrt{HS \log(A)}$ pops out. Note that the higher order term has two trailing log terms (see the right hand side of (14)), so when replacing $\delta$ by $\delta/|\Pi|$ with a union bound, both terms will give extra $\sqrt{HS}$ dependence so in higher order term we have extra $HS$ dependence but not just $\sqrt{HS}$.

□

## F Proof of uniform convergence problem with local policy class.

In this section, we consider using OPEMA estimator with data $\mathcal{D}'$. Also, WLOG we only consider deterministic reward (as implied by Lemma C.4 random reward only causes lower order dependence). Also, we fix $N > 0$ for the moment. First recall for all $t = 1, ..., H$

$$V_t^\pi(s_t) = \mathbb{E}_\pi \left[ \sum_{t'=t}^{H} r_{t'}(s_{t'}^{(1)}, a_t^{(1)}) \middle| s_t^{(1)} = s_t \right]$$
$$Q_t^\pi(s_t, a_t) = \mathbb{E}_\pi \left[ \sum_{t'=t}^{H} r_{t'}(s_{t'}^{(1)}, a_t^{(1)}) \middle| s_t^{(1)} = s_t, a_t^{(1)} = a_t \right]$$

where $r_t(s, a)$ are deterministic rewards and $s_t^{(1)}, a_t^{(1)}$ are random variables. Consider $V_t^\pi, Q_t^\pi$ as vectors, then by standard Bellman equations we have for all $t = 1, ..., H$ (define $V_{H+1} = Q_{H+1} = 0$)

$$Q_t^\pi = r_t + P_{t+1}^\pi Q_{t+1}^\pi = r_t + P_{t+1} V_{t+1}^\pi, \tag{15}$$

where $P_t^\pi \in \mathbb{R}^{(SA)\times(SA)}$ is the state-action transition and $P_t(\cdot|\cdot,\cdot) \in \mathbb{R}^{(SA)\times S}$ is the transition probabilities defined in Section 2. Also, we have bellman optimality equations:

$$Q_t^\star = r_t + P_{t+1} V_{t+1}^\star, \qquad V_t^\star(s_t) := \max_{a_t} Q_t^\star(s_t, a_t), \quad \pi_t^\star(s_t) := \operatorname*{argmax}_{a_t} Q_t^\star(s_t, a_t) \ \ \forall s_t \tag{16}$$

where $\pi^\star$ is one optimal deterministic policy. The corresponding Bellman equations and Bellman optimality equations for empirical MDP $\widehat{M}$ are defined similarly. Since we consider deterministic rewards, by Bellman equations we have

$$\widehat{Q}_t^\pi - Q_t^\pi = \widehat{P}_{t+1}^\pi \widehat{Q}_{t+1}^\pi - P_{t+1}^\pi Q_{t+1}^\pi = (\widehat{P}_{t+1}^\pi - P_{t+1}^\pi)\widehat{Q}_{t+1}^\pi + P_{t+1}^\pi(\widehat{Q}_{t+1}^\pi - Q_{t+1}^\pi)$$

for $t = 1, ..., H$. By writing it recursively, we have $\forall t = 1, ..., H-1$

$$\begin{aligned}\widehat{Q}_t^\pi - Q_t^\pi &= \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi (\widehat{P}_h^\pi - P_h^\pi)\widehat{Q}_h^\pi \\ &= \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi (\widehat{P}_h - P_h)\widehat{V}_h^\pi\end{aligned}$$

where $\Gamma_{t:h}^\pi = \prod_{i=t}^h P_i^\pi$ is the multi-step state-action transition and $\Gamma_{t+1:t}^\pi := I$.

Note $\widehat{\pi}^*$ to be the empirical optimal policy over $\widehat{M}$, we are interested in how to obtain uniform convergence for any policy $\pi$ that is close to $\widehat{\pi}^*$. More precisely, in this section we consider the policy class $\Pi_1$ to be:

$$\Pi_1 := \{\pi : s.t.\ ||\widehat{V}_t^\pi - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \le \epsilon_{\text{opt}},\ \forall t = 1, ..., H\}$$

where $\epsilon_{\text{opt}} \ge 0$ is a parameter decides how large the policy class is. We now assume $\widehat{\pi}$ to be any policy within $\Pi_1$ throughout this section. **Also, $\widehat{\pi}$ may be a policy learned from a learning algorithm using the data $\mathcal{D}$. In this case, $\widehat{\pi}$ may not be independent of $\widehat{P}$.**

We start with the following simple calculation:[12]

$$\begin{aligned}\left|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right| &\le \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi \left|(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}}\right| \\ &\le \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi \left|(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}^\star}\right|}_{(***)} + \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|}_{(****)}\end{aligned} \tag{17}$$

We now analyze $(***)$ and $(****)$.

### F.1 Analyzing $\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|$

First, by vector induced matrix norm[13] we have

$$\begin{aligned}\left\|\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\widehat{\pi}} \cdot \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|\right\|_\infty &\le H \cdot \sup_h \left\|\Gamma_{t+1:h-1}^{\widehat{\pi}}\right\|_\infty \left\||(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})|\right\|_\infty \\ &\le H \cdot \sup_h \left\||(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})|\right\|_\infty\end{aligned}$$

where the last equal sign uses multi-step transition $\Gamma_{t+1:h-1}^\pi$ is row-stochastic. Note given $N$, $\widehat{P}_t(\cdot|\cdot,\cdot)$ all have $N$ in the denominator. Therefore, by Hoeffding inequality and a union bound we have with probability $1-\delta$,

$$\sup_{t,s_t,s_{t-1},a_{t-1}} |\widehat{P}_t(s_t|s_{t-1},a_{t-1}) - P_t(s_t|s_{t-1},a_{t-1})| \le O(\sqrt{\frac{\log(HSA/\delta)}{N}}),$$

this indicates

$$\sup_h \left\||(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})|\right\|_\infty \le \epsilon_{\text{opt}} \cdot \sup_h \left\||\widehat{P}_h - P_h| \cdot \mathbf{1}\right\|_\infty \le \epsilon_{\text{opt}} \cdot O(S\sqrt{\frac{\log(HSA/\delta)}{N}}),$$

where $\mathbf{1} \in \mathbb{R}^S$ is all-one vector. To sum up, we have

[12] Since all quantities in the calculation are vectors, so the absolute value $|\cdot|$ used is point-wise operator.

[13] For $A$ a matrix and $x$ a vector we have $\|Ax\|_\infty \le \|A\|_\infty \|x\|_\infty$.

**Lemma F.1.** *Fix $N > 0$, we have with probability $1-\delta$, for all $t = 1, ..., H-1$*

$$\sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1} \left|(\widehat{P}_h - P_h)(\widehat{V}^{\widehat{\pi}^\star}_h - \widehat{V}^{\widehat{\pi}}_h)\right| \le \epsilon_{\text{opt}} \cdot O\left(\sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}} \cdot \mathbf{1}\right)$$

Now we consider $(***)$.

## F.2 Analyzing $\sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1} \left|(\widehat{P}_h - P_h)\widehat{V}^{\widehat{\pi}^\star}_h\right|$.

**Lemma F.2.** *Given $N$, we have with probability $1-\delta$, $\forall t = 1, ..., H-1$*

$$\sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1} \left|(\widehat{P}_h - P_h)\widehat{V}^{\widehat{\pi}^\star}_h\right| \le \sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1} \left(4\sqrt{\frac{\log(HSA/\delta)}{N}}\sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}^\star}_h)} + \frac{4(H-t)}{3N}\log(\frac{HSA}{\delta}) \cdot \mathbf{1}\right)$$

*where $\text{Var}(v^\pi_t) \in \mathbb{R}^{SA}$ and $\text{Var}(V^\pi_t)(s_{t-1}, a_{t-1}) = \text{Var}_{s_t}[V^\pi_t(\cdot)|s_{t-1}, a_{t-1}]$ and $|\cdot|, \sqrt{\cdot}$ are point-wise operator.*

*Proof of Lemma F.2.* The key point is to guarantee $\widehat{P}_h$ is independent of $\widehat{V}^{\widehat{\pi}^\star}_h$ so that we can apply Bernstein inequality w.r.t the randomness in $\widehat{P}_h$. In fact, note given $N$ all data pairs in $\mathcal{D}'$ are independent of each other, and $\widehat{P}_h$ only uses data from $h-1$ to $h$. Moreover, $\widehat{V}^{\widehat{\pi}^\star}_h$ only uses data from time $h$ to $H$ since $\widehat{V}^{\pi}_h$ uses data from $h$ to $H$ by bellman equation (15) for any $\pi$ and optimal policy $\widehat{\pi}^\star_{h:H}$ also only uses data from $h$ to $H$ by bellman optimality equation (16).

Then by Bernstein inequality (Lemma A.3), with probability $1-\delta$

$$\left|(\widehat{P}_h - P_h)\widehat{V}^{\widehat{\pi}^\star}_h\right|(s_{t-1}, a_{t-1}) \le 4\sqrt{\frac{\log(1/\delta)}{N}}\sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}^\star}_h)(s_{t-1}, a_{t-1})} + \frac{4(H-t)}{3N}\log(\frac{1}{\delta})$$

apply a union bound and take the sum we get the stated result. □

Now combine Lemma F.1 and Lemma F.2 we obtain with probability $1-\delta$, for all $t = 1, ..., H-1$

$$\begin{aligned}
\left|\widehat{Q}^{\widehat{\pi}}_t - Q^{\widehat{\pi}}_t\right| &\le \sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1} \left(4\sqrt{\frac{\log(HSA/\delta)}{N}}\sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}^\star}_h)} + \frac{4(H-t)}{3N}\log(\frac{HSA}{\delta}) \cdot \mathbf{1}\right) \\
&\quad + c_1\epsilon_{\text{opt}} \cdot \sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}} \cdot \mathbf{1} \\
&\le 4\sqrt{\frac{\log(HSA/\delta)}{N}} \sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1}\sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}^\star}_h)} + \frac{4H^2}{3N}\log(\frac{HSA}{\delta}) \cdot \mathbf{1} \\
&\quad + c_1\epsilon_{\text{opt}} \cdot \sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}} \cdot \mathbf{1},
\end{aligned} \tag{18}$$

Next note $\sqrt{\text{Var}(\cdot)}$ is a norm, therefore by norm triangle inequality we have

$$\begin{aligned}
\sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}^\star}_h)} &\le \sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}^\star}_h - \widehat{V}^{\widehat{\pi}}_h)} + \sqrt{\text{Var}(\widehat{V}^{\widehat{\pi}}_h - V^{\widehat{\pi}}_h)} + \sqrt{\text{Var}(V^{\widehat{\pi}}_h)} \\
&\le \left\|\widehat{V}^{\widehat{\pi}^\star}_h - \widehat{V}^{\widehat{\pi}}_h\right\|_\infty \cdot \mathbf{1} + \left\|\widehat{V}^{\widehat{\pi}}_h - V^{\widehat{\pi}}_h\right\|_\infty \cdot \mathbf{1} + \sqrt{\text{Var}(V^{\widehat{\pi}}_h)} \\
&\le \epsilon_{\text{opt}} \cdot \mathbf{1} + \left\|\widehat{Q}^{\widehat{\pi}}_h - Q^{\widehat{\pi}}_h\right\|_\infty \cdot \mathbf{1} + \sqrt{\text{Var}(V^{\widehat{\pi}}_h)}
\end{aligned}$$

Plug this into (18) to obtain

$$\begin{aligned}\left|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right| \le 4\sqrt{\frac{\log(HSA/\delta)}{N}} \sum_{h=t+1}^{H} \left(\Gamma_{t+1:h-1}^{\widehat{\pi}} \sqrt{\mathrm{Var}(V_h^{\widehat{\pi}})} + \left\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}\right\|_\infty \cdot \mathbf{1}\right) + \frac{4H^2}{3N}\log(\frac{HSA}{\delta}) \cdot \mathbf{1} \\ + c_2\epsilon_{\mathrm{opt}} \cdot \sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}} \cdot \mathbf{1}.\end{aligned} \tag{19}$$

Next lemma helps us to bound $\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\widehat{\pi}} \sqrt{\mathrm{Var}(V_h^{\widehat{\pi}})}$.

**Lemma F.3.** *A conditional version of Lemma E.4 holds:*

$$\begin{aligned}\mathrm{Var}_\pi\left[\sum_{t=h}^{H} r_t^{(1)} \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] = \sum_{t=h}^{H} \Big(\mathbb{E}_\pi\left[\mathrm{Var}\left[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)}) \middle| s_t^{(1)}, a_t^{(1)}\right] \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] \\ + \mathbb{E}_\pi\left[\mathrm{Var}\left[\mathbb{E}[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)}) | s_t^{(1)}, a_t^{(1)}] \middle| s_t^{(1)}\right] \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right]\Big).\end{aligned} \tag{20}$$

*and by using (20) we can show*

$$\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\widehat{\pi}} \sqrt{\mathrm{Var}(V_h^{\widehat{\pi}})} \le \sqrt{(H-t)^3} \cdot \mathbf{1}.$$

*Proof.* The proof of (20) uses the identical trick as Lemma E.4 except the total law of variance is replaced by the total law of conditional variance.

Moreover, recall $\Gamma_{t+1:h-1}^{\widehat{\pi}} = \prod_{i=t+1}^{h-1} P_i^{\widehat{\pi}}$ is the multi-step transition, so for any pair $(s_t, a_t)$,

$$\begin{aligned}&\sum_{h=t+1}^{H} \left(\Gamma_{t+1:h-1}^{\widehat{\pi}} \sqrt{\mathrm{Var}(V_h^{\widehat{\pi}})}\right)(s_t, a_t) \\ =& \sum_{h=t+1}^{H} \sum_{s_{h-1}, a_{h-1}} \sqrt{\mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}, a_{h-1}]} d_t^{\widehat{\pi}}(s_{h-1}, a_{h-1}|s_t, a_t) \\ =& \sum_{h=t+1}^{H} \sum_{s_{h-1}, a_{h-1}} \sqrt{\mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}, a_{h-1}] d_t^{\widehat{\pi}}(s_{h-1}, a_{h-1}|s_t, a_t)} \cdot \sqrt{d_t^{\widehat{\pi}}(s_{h-1}, a_{h-1}|s_t, a_t)} \\ \le& \sum_{h=t+1}^{H} \sqrt{\sum_{s_{h-1}, a_{h-1}} \mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}, a_{h-1}] d_t^{\widehat{\pi}}(s_{h-1}, a_{h-1}|s_t, a_t) \cdot \sum_{s_{h-1}, a_{h-1}} d_t^{\widehat{\pi}}(s_{h-1}, a_{h-1}|s_t, a_t)} \\ =& \sum_{h=t+1}^{H} \sqrt{\sum_{s_{h-1}, a_{h-1}} \mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}, a_{h-1}] d_t^{\widehat{\pi}}(s_{h-1}, a_{h-1}|s_t, a_t)} \\ =& \sum_{h=t+1}^{H} \sqrt{\mathbb{E}_{\widehat{\pi}}\left[\mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}^{(1)}, a_{h-1}^{(1)}] \middle| s_t, a_t\right]} \\ =& \sum_{h=t+1}^{H} \sqrt{1} \cdot \sqrt{\mathbb{E}_{\widehat{\pi}}\left[\mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}^{(1)}, a_{h-1}^{(1)}] \middle| s_t, a_t\right]} \\ \le& \sqrt{(H-t)\sum_{h=t+1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\mathrm{Var}[V_h^{\widehat{\pi}}|s_{h-1}^{(1)}, a_{h-1}^{(1)}] \middle| s_t, a_t\right]} \\ \le& \sqrt{(H-t) \cdot \mathrm{Var}_{\widehat{\pi}}\left[\sum_{h=t+1}^{H} r_h^{(1)} \middle| s_t^{(1)} = s_t, a_t^{(1)} = a_t\right]} \le \sqrt{(H-t)^3}\end{aligned}$$

where all the inequalities are Cauchy-Schwarz inequalities. □

Apply Lemma F.3 to bound (19), and use $\infty$ norm on both sides, we obtain

**Theorem F.4.** *Conditional on $N > 0$, then with probability $1-\delta$, we have for all $t = 1, ..., H-1$*

$$\left\|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right\|_\infty \le 4\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\sum_{h=t+1}^{H}\left\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}\right\|_\infty + \frac{4H^2}{3N}\log(\frac{HSA}{\delta}) + c_2\epsilon_{opt}\cdot\sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}}.$$

Then by using backward induction and Theorem F.4, we have the following:

**Theorem F.5.** *Suppose $N \ge 64H^2\cdot\log(HSA/\delta)$ and $\epsilon_{opt} \le \sqrt{H}/S$, then we have with probability $1-\delta$,*

$$\left\|\widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}}\right\|_\infty \le 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}$$

*where $c_2$ is the same constant in Theorem F.4.*

*Proof.* Under the condition, by Theorem F.4 it is easy to check for all $t = 1, ..., H-1$ with probability $1-\delta$,

$$\left\|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right\|_\infty \le (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\sum_{h=t+1}^{H}\left\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}\right\|_\infty,$$

which we conditional on.

For $t = H-1$, we have

$$\begin{aligned}\left\|\widehat{Q}_{H-1}^{\widehat{\pi}} - Q_{H-1}^{\widehat{\pi}}\right\|_\infty \le& (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\left\|\widehat{Q}_H^{\widehat{\pi}} - Q_H^{\widehat{\pi}}\right\|_\infty \\ \le& (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{H^2\log(HSA/\delta)}{N}} \\ \le& (9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}\end{aligned}$$

Suppose $\left\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}\right\|_\infty \le 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}$ holds for all $h = t+1, ..., H$, then for $h = t$, we have

$$\begin{aligned}\left\|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right\|_\infty &\le (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\sum_{h=t+1}^{H}\left\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}\right\|_\infty \\ &\le (9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{(H-1)^2\log(HSA/\delta)}{N}}\cdot 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} \\ &\le 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}\end{aligned}$$

where the last line uses the condition $N \ge 64H^2\cdot\log(HSA/\delta)$. By induction, we have the result. □

*Proof of Theorem 3.7.* By Theorem F.5 we have for $N \ge c\cdot H^2\cdot\log(HSA/\delta)$,

$$\mathbb{P}\left(\left\|\widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}}\right\|_\infty \ge 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}\middle| N\right) \le \delta$$

The only thing left is to use Lemma C.1 to bound the event that $\{N < nd_m/2\}$ has small probability.

Last but not least, the condition $n > c_1H^2\log(HSA/\delta)/d_m$ is sufficient for applying Lemma C.1 and it also implies $N \ge c\cdot H^2\cdot\log(HSA/\delta)$ (the condition of Theorem F.5) when $N \ge nd_m/2$ since:

$$n > c_1H^2\log(HSA/\delta)/d_m \Rightarrow nd_m/2 \ge c_2H^2\log(HSA/\delta)$$

which implies $N \ge c_2\cdot H^2\cdot\log(HSA/\delta)$ when $N \ge nd_m/2$. □

# G Proof of uniform convergence lower bound.

In this section we prove a uniform convergence OPE lower bound of $\Omega(H^3/d_m\epsilon^2)$. Conceptually, uniform convergence lower bound can be derived by a reduction to the lower bound of identifying the $\epsilon$-optimal policy. There are quite a few literature that provide information theoretical lower bounds in different setting, *e.g.* Dann and Brunskill (2015); Jiang et al. (2017); Krishnamurthy et al. (2016); Jin et al. (2018); Sidford et al. (2018). However, to the best of our knowledge, there is no result proven for the non-stationary transition finite horizon episodic setting with bounded rewards. For example, Sidford et al. (2018) prove the result sample complexity lower bound of $\Omega(H^3SA/\epsilon^2)$ with stationary MDP and their proof cannot be directly applied to non-stationary setting as they reduce the problem to infinite horizon discounted setting which always has stationary transitions. Dann and Brunskill (2015) prove the episode complexity of $\widetilde{\Omega}(H^2SA/\epsilon^2)$ for the stationary transition setting. Jin et al. (2018) prove the $\Omega(\sqrt{H^2SAT})$ regret lower bound for non-stationary finite horizon online setting but it is not clear how to translate the regret to PAC-learning setting by keeping the same sample complexity optimality. Jiang et al. (2017) prove the $\Omega(HSA/\epsilon^2)$ lower bound for the non-stationary finite horizon offline episodic setting where they assume $\sum_{i=1}^H r_i \leq 1$ and this is also different from our setting since we have $0 \leq r_t \leq 1$ for each time step.

Our proof consists of three steps. **1.** We will first show a minimax lower bound (**over all MDP instances**) for learning $\epsilon$-optimal policy is $\Omega(H^3SA/\epsilon^2)$; **2.** Based on 1, we can further show a minimax lower bound (**over problem class** $\mathcal{M}_{d_m}$) for learning $\epsilon$-optimal policy is $\Omega(H^3/d_m\epsilon^2)$; **3.** prove the uniform convergence OPE lower bound of the same rate.

## G.1 Information theoretical lower sample complexity bound over all MDP instances for identifying $\epsilon$-optimal policy.

In fact, a modified construction of Theorem 5 in Jiang et al. (2017) is our tool for obtaining $\Omega(H^3SA/\epsilon^2)$ lower bound. We can get the additional $H^2$ factor by using $\sum_{i=1}^H r_i$ can be of order $O(H)$.

**Theorem G.1.** *Given $H \geq 2$, $A \geq 2$, $0 < \epsilon < \frac{1}{48\sqrt{8}}$ and $S \geq c_1$ where $c_1$ is a universal constant. Then there exists another universal constant $c$ such that for any algorithm and any $n \leq cH^3SA/\epsilon^2$, there exists a non-stationary $H$ horizon MDP with probability at least $1/12$, the algorithm outputs a policy $\widehat{\pi}$ with $v^\star - v^{\widehat{\pi}} \geq \epsilon$.*

Like in Jiang et al. (2017), the proof relies on embedding $\Theta(HS)$ independent multi-arm bandit problems into a hard-to-learn MDP so that any algorithm that wants to output a near-optimal policy needs to identify the best action in $\Omega(HS)$ problems. However, in our construction we make a further modification of Jiang et al. (2017) so that there is **no** waiting states, which is crucial for the reduction from offline family. We also double the length of the hard-to-learn MDP instance so that the latter half uses a "naive" copy construction which is uninformative. The uninformative extension will help to produce the additional $H^2$ factor.

*Proof of Theorem G.2.* We construct a non-stationary MDP with $S$ states per level, $A$ actions per state and has horizon $2H$. At each time step, states are categorized into four types with two special states $g_h$, $b_h$ and the remaining $S-2$ "bandit" states denoted by $s_{h,i}$, $i \in [S-2]$. Each bandit state has an unknown best action $a^\star_{h,i}$ that provides the highest expected reward comparing to other actions.

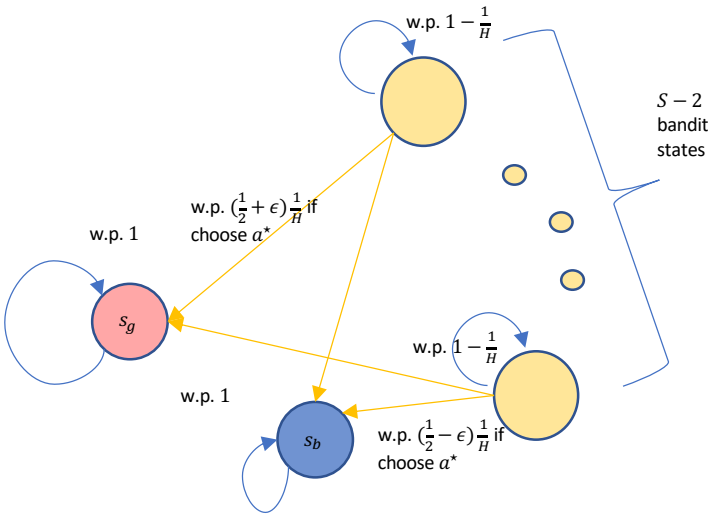


Figure 2: An illustration of the state-space transition diagram from our lower bound construction in Theorem G.2. The new construction eliminates the waiting states, thus making it possible to obtain a lower bound that explicitly depends on parameter $d_m$ in Theorem G.2.

The transition dynamics are defined as follows:

- for $h = 1, ..., H-1$,
  - For bandit states $b_{h,i}$, there is probability $1 - \frac{1}{H}$ to transition to $b_{h+1,i}$ regardless of the action chosen. For the rest of $\frac{1}{H}$ probability, optimal action $a^\star_{h,i}$ will have probability $\frac{1}{2} + \tau$ or $\frac{1}{2} - \tau$ transition to $g_{h+1}$ or $b_{h+1}$ and all other actions $a$ will have equal probability $\frac{1}{2}$ for either $g_{h+1}$ or $b_{h+1}$, where $\tau$ is a parameter will be decided later. Or equivalently,

$$\mathbb{P}(\cdot|s_{h,i}, a^\star_{h,i}) = \begin{cases} 1 - \frac{1}{H} & \text{if } \cdot = s_{h+1,i} \\ (\frac{1}{2} + \tau) \cdot \frac{1}{H} & \text{if } \cdot = g_{h+1} \\ (\frac{1}{2} - \tau) \cdot \frac{1}{H} & \text{if } \cdot = b_{h+1} \end{cases} \quad \mathbb{P}(\cdot|s_{h,i}, a) = \begin{cases} 1 - \frac{1}{H} & \text{if } \cdot = s_{h+1,i} \\ \frac{1}{2} \cdot \frac{1}{H} & \text{if } \cdot = g_{h+1} \\ \frac{1}{2} \cdot \frac{1}{H} & \text{if } \cdot = b_{h+1} \end{cases}$$

  - $g_h$ always transitions to $g_{h+1}$ and $b_h$ always transitions to $b_{h+1}$, *i.e.* for all $a \in \mathcal{A}$, we have

$$\mathbb{P}(g_{h+1}|g_h, a) = 1, \quad \mathbb{P}(b_{h+1}|b_h, a) = 1.$$

    We will determine parameter $\tau$ at the end of the proof.

- for $h = H, ..., 2H-1$, all states will always transition to the same type of states for the next step, *i.e.* $\forall a \in \mathcal{A}$,

$$\mathbb{P}(g_{h+1}|g_h, a) = \mathbb{P}(b_{h+1}|b_h, a) = \mathbb{P}(s_{h+1,i}|s_{h,i}, a) = 1, \ \forall i \in [S-2]. \tag{21}$$

- The initial distribution is decided by:

$$\mathbb{P}(s_{1,i}) = \frac{1}{S}, \ \forall i \in [S-2], \ \mathbb{P}(g_1) = \frac{1}{S}, \ \ \mathbb{P}(b_1) = \frac{1}{S} \tag{22}$$

- State $s$ will receives reward 1 if and only if $s = g_h$ and $h \geq H$. The reward at all other states is zero.

By this construction the optimal policy must take $a^\star_{h,i}$ for each bandit state $s_{h,i}$ for at least the first half of the MDP, *i.e.* need to take $a^\star_{h,i}$ for $h \leq H$. In other words, this construction embeds at least $H(S-2)$ independent best arm identification problems that are identical to the stochastic multi-arm bandit problem in Lemma A.7

into the MDP. **Note the key innovation here is that we can remove the waiting states used in Jiang et al. (2017) but still keep the multi-arm bandit problem independent!**[14]

Notice in our construction, for any bandit state $s_{h,i}$ with $h \le H$, the difference of the expected reward between optimal action $a^\star_{h,i}$ and other actions is:

$$\begin{aligned}
&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}]+(\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}]+(1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):2H}|s_{h+1,i}]\\
&-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}]-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}]-(1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):2H}|s_{h+1,i}]\\
=&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}]+(\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}] \qquad (23)\\
&-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}]-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}]\\
=&(\frac{1}{2}+\tau)\frac{1}{H}\cdot H+(\frac{1}{2}-\tau)\frac{1}{H}\cdot 0-\frac{1}{2H}\cdot H+\frac{1}{2H}\cdot 0=\tau
\end{aligned}$$

so it seems by Lemma A.7 one suffices to use the least possible $\frac{A}{72(\tau)^2}$ samples to identify the best action $a^\star_{h,i}$. However, note the construction of the latter half of the MDP (21) uses mindless reproduction of previous steps and therefore provides no additional information about the best action once the state at time $H$ is known. In other words, observing $\sum_{t=1}^{2H} r_t = H$ is equivalent as observing $\sum_{t=1}^{H} r_t = 1$. Therefore, for the bandit states in the first half the samples that provide information for identifying the best arm is up to time $H$. As a result, the difference of the expected reward between optimal action $a^\star_{h,i}$ and other action for identifying the best arm should be corrected as:

$$\begin{aligned}
&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]+(\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]+(1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):H}|s_{h+1,i}]\\
&-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]-(1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):H}|s_{h+1,i}]\\
=&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]+(\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]\\
&-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]\\
=&(\frac{1}{2}+\tau)\frac{1}{H}\cdot 1+(\frac{1}{2}-\tau)\frac{1}{H}\cdot 0-\frac{1}{2H}\cdot 1+\frac{1}{2H}\cdot 0=\frac{\tau}{H}
\end{aligned}$$

Now by Lemma A.7, for each bandit state $s_{h,i}$ satisfying $h \le H$, unless $\frac{A}{72(\tau/H)^2}$ samples are collected from that state, the learning algorithm fails to identify the optimal action $a^\star_{h,i}$ with probability at least $1/3$.

After running any algorithm, let $C$ be the set of $(h,s)$ pairs for which the algorithm identifies the correct action. Let $D$ be the set of $(h,s)$ pairs for which the algorithm collects fewer than $\frac{A}{72(\tau/H)^2}$ samples. Then by Lemma A.7 we have

$$\begin{aligned}
\mathbb{E}[|C|] &= \mathbb{E}\left[\sum_{(h,s)}\mathbb{1}[a_{h,s}=a^\star_{h,s}]\right] \le ((S-2)H-|D|)+\mathbb{E}\left[\sum_{(h,s)\in D}\mathbb{1}[a_{h,s}=a^\star_{h,s}]\right]\\
&\le ((S-2)H-|D|)+\frac{2}{3}|D| = (S-2)H-\frac{1}{3}|D|.
\end{aligned}$$

If we have $n \le \frac{H(S-2)}{2}\times\frac{A}{72(\tau/H)^2}$, by pigeonhole principle the algorithm can collect $\frac{A}{72(\tau/H)^2}$ samples for at most half of the bandit problems, *i.e.* $|D| \ge H(S-2)/2$. Therefore we have

$$\mathbb{E}[|C|] \le (S-2)H-\frac{1}{3}|D| \le \frac{5}{6}(S-2)H.$$

Then by Markov inequality

$$\mathbb{P}\left[|C| \ge \frac{11}{12}H(S-2)\right] \le \frac{5/6}{11/12} = \frac{10}{11}$$

[14] Here independence means solving one bandit problem provides no information on other bandit problems.

so the algorithm failed to identify the optimal action on 1/12 fraction of the bandit problems with probability at least 1/11. Note for each failure in identification, the reward is differ by $\tau$ (see (23)), therefore under the event $\{|C'| \geq \frac{1}{12}H(S-2)\}$, following the similar calculation of Jiang et al. (2017) the suboptimality of the policy produced by the algorithm is

$$\begin{aligned}
\epsilon := v^\star - v^{\widehat{\pi}} &= \mathbb{P}[\text{visit } C'] \times \tau + \mathbb{P}[\text{visit } C] \times 0 = \mathbb{P}[\bigcup_{(h,i)\in C'} \text{visit}(h,i)] \times \tau \\
&= \sum_{(h,i)\in C'} \mathbb{P}[\text{visit}(h,i)] \times \tau = \sum_{(h,i)\in C'} \frac{1}{HS}(1-1/H)^{h-1}\tau \\
&\geq \sum_{(h,i)\in C'} \frac{1}{HS}(1-1/H)^{H}\tau \geq \sum_{(h,i)\in C'} \frac{1}{HS}\frac{1}{4}\tau \\
&\geq \frac{H(S-2)}{12}\frac{1}{HS}\frac{1}{4}\tau = c_1\frac{\tau}{48}.
\end{aligned}$$

where the third equal sign uses all best arm identification problems are independent. Now we set $\tau = \min(\sqrt{1/8}, 48\epsilon/c_1)$ and under condition $n \leq cH^3SA/\epsilon^2$, we have

$$n \leq cH^3SA/\epsilon^2 \leq c48^2H^3SA/\tau^2 = c48^2 \cdot 72HS \cdot \frac{A}{72(\tau/H)^2} := c'HS \cdot \frac{A}{72\tau^2} \leq \frac{H(S-2)}{2} \cdot \frac{A}{72\tau^2},$$

the last inequality holds as long as $S \geq 2/(1-2c')$. Therefore in this situation, with probability at least 1/11, $v^\star - v^{\widehat{\pi}} \geq \epsilon$. Finally, we can use scaling to reduce the horizon from $2H$ to $H$.

□

### G.2 Information theoretical lower sample complexity bound over problems in $\mathcal{M}_{d_m}$ for identifying $\epsilon$-optimal policy.

For all $0 < d_m \leq \frac{1}{SA}$, let the class of problems be

$$\mathcal{M}_{d_m} := \left\{(\mu, M) \;\middle|\; \min_{t,s_t,a_t} d_t^\mu(s_t,a_t) \geq d_m\right\},$$

now we consider deriving minimax lower bound over this class.

**Theorem G.2.** *Under the same condition of Theorem G.1. In addition assume $0 < d_m \leq \frac{1}{SA}$. There exists another universal constant $c$ such that when $n \leq cH^3/d_m\epsilon^2$, we always have*

$$\inf_{v^{\pi_{alg}}} \sup_{(\mu,M)\in\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(v^* - v^{\pi_{alg}} \geq \epsilon\right) \geq p.$$

*Proof.* The hard instance $(\mu, M)$ we used is based on Theorem G.1, which is described as follows.

- for the MDP $M = (\mathcal{S}, \mathcal{A}, r, P, d_1, 2H+2)$,
  - Initial distribution $d_1$ will always enter state $s_0$, and there are two actions with action $a_1$ always transitions to $s_{\text{yes}}$ and action $a_2$ always transitions to $s_{\text{no}}$. The reward at the first time $r_1(s,a) = 0$ for any $s, a$.
  - For state $s_{\text{no}}$, it will always transition back to itself regardless of the action and receive reward 0, *i.e.*

$$P_t(s_{\text{no}}|s_{\text{no}},a) = 1,\ r_t(s_{\text{no}},a) = 0,\ \forall t,\ \forall a.$$

  - For state $s_{\text{yes}}$, it will transition to the MDP construction in Theorem G.1 with horizon $2H$ and $s_{\text{yes}}$ always receives reward zero.
  - For $t = 1$, choose $\mu(a_1|s_0) = \frac{1}{2}d_mSA$ and $\mu(a_2|s_0) = 1 - \frac{1}{2}d_mSA$. For $t \geq 2$, choose $\mu$ to be uniform policy, *i.e.* $\mu(a_t|s_t) = 1/A$.

Based on this construction, the optimal policy has the form $\pi^\star = (a_1, \ldots)$ and therefore the MDP branch that enters $s_{\text{no}}$ is uninformative. Hence, data collected by that part is uninformed about the optimal policy and there is only $\frac{1}{2}d_m SA$ proportion of data from $s_{\text{yes}}$ are useful. Moreover, by Theorem G.1 the rest of Markov chain succeeded from $s_{\text{yes}}$ requires $\Omega(H^3SA/\epsilon^2)$ episodes (regardless of the exploration strategy/logging policy), so the actual data complexity needed for the whole construction $(\mu, M)$ is $\frac{\Omega(H^3SA/\epsilon^2)}{d_m SA} = \Omega(H^3/d_m\epsilon^2)$.

It remains to check this construction $\mu, M$ stays within $\mathcal{M}_{d_m}$.

- For $t = 1$, we have $d_1(s_0, a_1) = \frac{1}{2}d_m SA \geq d_m$ (since $S \geq 2$) and $d_1(s_0, a_2) = 1 - \frac{1}{2}d_m SA \geq d_m$ (this is since $d_m \leq \frac{1}{SA} \leq \frac{2}{2+SA}$);
- For $t = 2$, $d_2(s_{\text{yes}}, a) = \frac{1}{2}d_m SA \cdot \frac{1}{A} = \frac{1}{2}d_m S \geq d_m$ (since $S \geq 2$) and similar for $s_{\text{no}}$;
- For $t \geq 3$, for $g_h$ and $b_h$ in the sub-chain inherited from $s_{\text{yes}}$, note $d_h(g_h) \leq d_{h+1}(g_{h+1})$ (since $g_h$ and $b_h$ are absorbing states regardless of actions), therefore $d_h(g_h) \geq d_1(g_1) = d_1(s_{\text{yes}}) \cdot \mathbb{P}(g_1|s_{\text{yes}}) = \frac{1}{2}d_m SA \cdot \frac{1}{S} = \frac{1}{2}d_m A$, since $\mu$ is uniform so $d_h(g_h, a) \geq \Omega(d_m A) \cdot \frac{1}{A} = \Omega(d_m)$ forall $a$. Similar result can be derived for $b_h$ in identical way.

  For bandit state, we have for all $i \in [S-2]$,

$$\begin{aligned} d^\mu_{t+1}(s_{t+1,i}) &\geq \mathbb{P}^\mu(s_{t+1,i}, s_{t,i}, s_{t-1,i}, \ldots, s_{2,i}, s_{1,i}, s_{\text{yes}}, s_0) \\ &= \prod_{u=1}^{t} \mathbb{P}^\mu(s_{u+1,i}|s_u)\mathbb{P}^\mu(s_{1,i}|s_{\text{yes}})\mathbb{P}^\mu(s_{\text{yes}}|s_0) \\ &= (1 - \frac{1}{H})^t \left(\frac{1}{S}\right)\left(\frac{1}{2}d_m SA\right) \geq c d_m A, \end{aligned}$$

  now by $\mu$ is uniform we have $d^\mu_{t+1}(s_{t+1,i}, a) \geq \Omega(d_m A) \cdot \frac{1}{A} = \Omega(d_m)$ for all $a$. This concludes the proof.

□

**Remark G.3.** *A directly corollary is that the sample complexity in Theorem 4.1 part 3. is optimal. Indeed, for the case $\epsilon_{opt} = 0$, Theorem 4.1 implies $\widehat{\pi}$ is the $\epsilon$-optimal policy learned with sample complexity $O(H^3 \log(HSA/\delta)/d_m\epsilon^2)$. Theorem G.2 implies this sample complexity cannot be further reduced up to the logarithmic factor.*

### G.3 Information theoretical lower sample complexity bound for uniform convergence in OPE.

By applying Theorem G.2, we can now prove Theorem 3.8.

*Proof of Theorem 3.8.* We prove it by contradiction. Suppose there is one off-policy evaluation method $\widehat{v}^\pi$ such that

$$\sup_{\pi \in \Pi} |\widehat{v}^\pi - v^\pi| \leq o\left(\sqrt{\frac{H^3}{d_m n}}\right),$$

where $o(\cdot)$ represents the standard small $o$-notation. Then by

$$\begin{aligned} 0 \leq v^{\pi^\star} - v^{\widehat{\pi}^\star} &= v^{\pi^\star} - \widehat{v}^{\widehat{\pi}^\star} + \widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star} \\ &\leq |v^{\pi^\star} - \widehat{v}^{\pi^\star}| + |\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}| \leq 2\sup_\pi |v^\pi - \widehat{v}^\pi|. \end{aligned}$$

this OPE method implies a $\epsilon$-optimal policy learning algorithm with sample complexity $o(H^3/d_m\epsilon^2)$ which is smaller than the information theoretical lower bound obtained in Theorem G.2. Contradiction! □

## H Proofs of Theorem 4.1

*Proof of Theorem 4.1.* Part 1. and Part 2. are just direct corollaries. We only prove Part 3. here. Indeed, by definition of empirical optimal policy we have $\widehat{Q}^{\pi^\star} \leq \widehat{Q}^{\widehat{\pi}^\star}$, so we have the following:

$$\begin{aligned} Q_1^{\pi^\star} - Q_1^{\widehat{\pi}} &= Q_1^{\pi^\star} - \widehat{Q}_1^{\widehat{\pi}^\star} + \widehat{Q}_1^{\widehat{\pi}^\star} - \widehat{Q}_1^{\widehat{\pi}} + \widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}} \\ &\leq Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star} + \widehat{Q}_1^{\widehat{\pi}^\star} - \widehat{Q}_1^{\widehat{\pi}} + \widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}} \\ &\leq Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star} + \epsilon_{\text{opt}} \cdot \mathbf{1} + \widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}} \end{aligned}$$

and $\widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}}$ can be bounded by Theorem 3.7 using local uniform convergence. $Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star}$ can be bounded by $O(\sqrt{\frac{H^3 \log(HSA/\delta)}{nd_m}})$ using the similar technique in Section F even without introducing $\epsilon_{\text{opt}}$ since $\pi^\star$ is a fixed policy. All these implies:

$$Q_1^{\pi^\star} - Q_1^{\widehat{\pi}} \leq \left( O(\sqrt{\frac{H^3 \log(HSA/\delta)}{nd_m}}) + \epsilon_{\text{opt}} \right) \cdot \mathbf{1}.$$

Especially when $\epsilon_{\text{opt}} = 0$ then this is slightly stronger than the stated result since:

$$v_1^{\pi^\star} - v_1^{\widehat{\pi}^\star} = Q_1^{\pi^\star}(\cdot, \pi^\star(\cdot)) - Q_1^{\widehat{\pi}^\star}(\cdot, \widehat{\pi}^\star(\cdot)) \leq Q_1^{\pi^\star}(\cdot, \pi^\star(\cdot)) - Q_1^{\widehat{\pi}^\star}(\cdot, \pi^\star(\cdot)) \leq \left\| Q_1^{\pi^\star} - Q_1^{\widehat{\pi}^\star} \right\|_\infty \leq O(\sqrt{\frac{H^3 \log(HSA/\delta)}{nd_m}}) \cdot \mathbf{1}$$

□

## I Simulation details

The non-stationary MDP with used for the experiments have 2 states $s_0, s_1$ and 2 actions $a_1, a_2$ where action $a_1$ has probability 1 always going back the current state and for action $a_2$, there is one state s.t. after choosing $a_2$ the dynamic transitions to both states with equal probability $\frac{1}{2}$ and the other one has asymmetric probability assignment ($\frac{1}{4}$ and $\frac{3}{4}$). The transition after choosing $a_2$ is changing over different time steps therefore the MDP is non-stationary and the change is decided by a sequence of pseudo-random numbers. More formally, $P_t$ can be either

$$\mathbb{P}(s_0|s_0, a_1) = 1; \mathbb{P}(s_1|s_1, a_1) = 1; \mathbb{P}(\cdot|s_0, a_2) = \begin{cases} \frac{1}{2}, \text{ if } \cdot = s_1 \\ \frac{1}{2}, \text{ if } \cdot = s_0 \end{cases} ; \mathbb{P}(\cdot|s_1, a_2) = \begin{cases} \frac{3}{4}, \text{ if } \cdot = s_1 \\ \frac{1}{4}, \text{ if } \cdot = s_0 \end{cases}$$

or

$$\mathbb{P}(s_0|s_0, a_1) = 1; \mathbb{P}(s_1|s_1, a_1) = 1; \mathbb{P}(\cdot|s_0, a_2) = \begin{cases} \frac{1}{4}, \text{ if } \cdot = s_1 \\ \frac{3}{4}, \text{ if } \cdot = s_0 \end{cases} ; \mathbb{P}(\cdot|s_1, a_2) = \begin{cases} \frac{1}{2}, \text{ if } \cdot = s_1 \\ \frac{1}{2}, \text{ if } \cdot = s_0 \end{cases}$$

Moreover, to make the learning problem non-trivial we use non-stationary rewards with 4 categories, *i.e.* $r_t(s, a) \in \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$ and assignment of $r_t(s, a)$ for each value is changing over time. That means, one possible assignment can be

$$r_t(s_0, a_1) = 1/4,\ r_t(s_0, a_2) = 2/4,\ r_t(s_1, a_1) = 3/4,\ r_t(s_1, a_2) = 1/4.$$

Moreover, the logging policy in Figure 1(a) is uniform with $\mu_t(a_1|s) = \mu_t(a_2|s) = \frac{1}{2}$ for both states. We implement the non-stationary MDP in the `Python` environment and pseudo-random numbers $p_t, r_t$'s are generated by keeping `numpy.random.seed(100)`.

We fix episodes $n = 2048$ and run each algorithm 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 for fixed policy is computed as:

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

and RMSE for suboptimality gap is computed as

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

and RMSE for empirical optimal policy gap is computed as

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

where $v^{\pi}_{\text{true}}$ 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 $v^{\pi}_{\text{true}} = \sum_{t=1}^{H}\sum_{s_t} d^{\pi}_t(s_t)r^{\pi}_t(s_t)$. $v^{\pi^{\star}}_{\text{true}}$ is obtained by running Value Iteration exhaustively until the error converges to 0. The average relative error for suboptimality (average of $|v^{\widehat{\pi}^{\star}_{[k]}} - v^{\pi^{\star}}_{\text{true}}|/v^{\pi^{\star}}_{\text{true}}$) at $H = 1000$ is 0.0011. Lastly, we also show the scaling of $|\widehat{v}^{\widehat{\pi}^{\star}} - v^{\widehat{\pi}^{\star}}|$ in Figure 3, which shares a similar pattern as the suboptimality plot as a whole. [15]

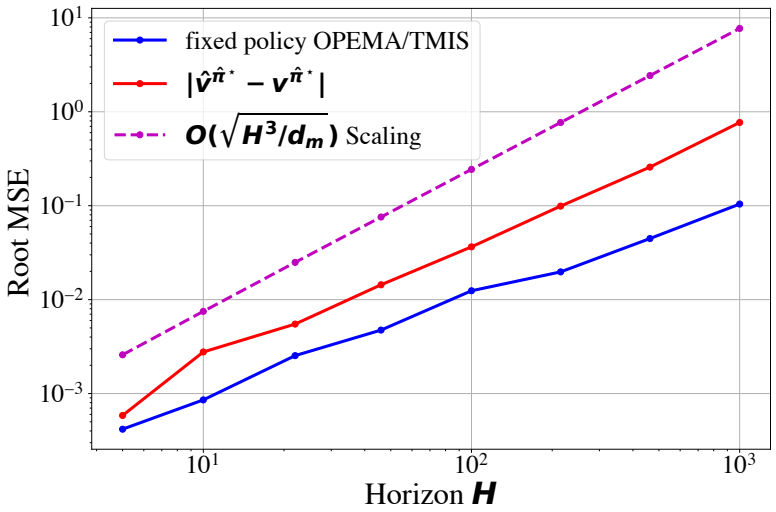


Figure 3: Log-log plot showing the dependence on horizon of uniform OPE and pointwise OPE via learning ($|\hat{v}^{\widehat{\pi}^{\star}} - v^{\widehat{\pi}^{\star}}|$) over a non-stationary MDP example.

## J On improvement over vanilla simulation lemma for fixed policy evaluation

**Vanilla simulation lemma, Lemma 1 of Jiang (2018).** Without loss of generality, assuming reward is determinsitic function over state-action. By definition of Bellman equation, we have the following:

$$\widehat{V}^{\pi}_t = r + \widehat{P}^{\pi}_{t+1}\widehat{V}^{\pi}_{t+1}, \quad V^{\pi}_t = r + P^{\pi}_{t+1}V^{\pi}_{t+1},$$

[15] Here we do point out the empirical dependence on $H$ for $|\widehat{v}^{\widehat{\pi}^{\star}} - v^{\widehat{\pi}^{\star}}|$ in the Figure 3 is actually less than $H^{1.5}$, this comes from that the MDP example we choose is not the "hardest" example for quantity $|\widehat{v}^{\widehat{\pi}^{\star}} - v^{\widehat{\pi}^{\star}}|$, as opposed to quantity $|v^{\star} - v^{\widehat{\pi}^{\star}}|$ in Figure 1.

define $\epsilon_P = \sup_{t,s_t,a_t} ||\widehat{P}_t(\cdot|s_t,a_t) - P_t(\cdot|s_t,a_t)||_1$, then by Hoeffding's inequality and union bound, with probability $1-\delta$,

$$\epsilon_P \leq S \cdot \sup_{t,s_t,a_t} ||\widehat{P}_t(\cdot|s_t,a_t) - P_t(\cdot|s_t,a_t)||_\infty \leq S \cdot \sup_{t,s_t,a_t} O\left(\sqrt{\frac{\log(HSA/\delta)}{n_{s_t,a_t}}}\mathbf{1}(E_t)\right) = O\left(\sqrt{\frac{S^2\log(HSA/\delta)}{n\cdot d_m}}\right)$$

then

$$\begin{aligned}
\widehat{V}_t^\pi - V_t^\pi =& \widehat{P}_{t+1}^\pi \widehat{V}_{t+1}^\pi - P_{t+1}^\pi V_{t+1}^\pi \\
\leq& \left(\left\|\widehat{P}_{t+1}^\pi - P_{t+1}^\pi\right\|_1 \left\|\widehat{V}_{t+1}^\pi\right\|_\infty + \left\|P_{t+1}^\pi\right\|_1 \left\|\widehat{V}_{t+1}^\pi - V_{t+1}^\pi\right\|_\infty\right)\cdot \mathbf{1} \\
\leq& \left(H\epsilon_P + \left\|\widehat{V}_{t+1}^\pi - V_{t+1}^\pi\right\|_\infty\right)\cdot \mathbf{1},
\end{aligned}$$

solving recursively, we have

$$\left\|\widehat{V}_1^\pi - V_1^\pi\right\|_\infty \leq H^2\epsilon_P \leq O\left(\sqrt{\frac{H^4S^2\log(HSA/\delta)}{n\cdot d_m}}\right).$$

This verifies SL has complexity $\widetilde{O}(H^4S^2/d_m\epsilon^2)$. We do point out above standard analysis can be improved (*e.g.* Jiang (2018) Section 2.2) to $\widetilde{O}(H^4S/d_m\epsilon^2)$, then in this case our analysis (Lemma 3.4) has an improvement of $H^2S$ with respect to the modified result.

# K Algorithms

**Algorithm 1** OPEMA

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

1: Calculate the on-policy estimation of initial distribution $d_1(\cdot)$ by $\widehat{d}_1(s) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_1^{(i)} = s)$, and set $\widehat{d}_1^\mu(\cdot) := \widehat{d}_1(\cdot)$, $\widehat{d}_1^\pi(s) := \widehat{d}_1(\cdot)$.
2: **for** $t = 2, 3, \ldots, H$ **do**
3: Choose all transition data at time step $t$, $\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{i=1}^n$.
4: Calculate the on-policy estimation of $d_t^\mu(\cdot)$ by $\widehat{d}_t^\mu(s) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s)$.
5: Set the off-policy estimation of $\widehat{P}_t(s_t|s_{t-1},a_{t-1})$:

$$\widehat{P}_t(s_t|s_{t-1},a_{t-1}) := \frac{\sum_{i=1}^n \mathbf{1}[(s_t^{(i)}, a_{t-1}^{(i)}, s_{t-1}^{(i)}) = (s_t, s_{t-1}, a_{t-1})]}{n_{s_{t-1},a_{t-1}}}$$

when $n_{s_{t-1},a_{t-1}} > 0$. Otherwise set it to be zero.
6: Estimate the reward function

$$\widehat{r}_t(s_t,a_t) := \frac{\sum_{i=1}^n r_t^{(i)}\mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}{\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}.$$

when $n_{s_t,a_t} > 0$. Otherwise set it to be zero.
7: Set $\widehat{d}_t^\pi(\cdot,\cdot)$ according to $\widehat{d}_t^\pi = \widehat{P}_t^\pi \widehat{d}_{t-1}^\pi$, where $\widehat{d}_t^\pi(\cdot,\cdot)$ is the estimated state-action distribution.
8: **end for**
9: Substitute the all estimated values above into $\widehat{v}^\pi = \sum_{t=1}^H \langle \widehat{d}_t^\pi, \widehat{r}_t\rangle$ to obtain $\widehat{v}^\pi$, the estimated value of $\pi$.

**Remark K.1.** *In short, we can see Algorithm 2 requires the splitting data size $M$ which is undecided by Yin and Wang (2020) and that makes the hyper-parameter requiring additional concrete specifications to make the data*

**Algorithm 2** Data Splitting TMIS in Yin and Wang (2020)

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

1: Randomly splitting the data $\mathcal{D}$ evenly into $N$ folds, with each fold $|\mathcal{D}^{(i)}| = M$, *i.e.* $n = M \cdot N$.
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 $v^{\pi}$.

*splitting estimator sample efficient. In contrast, OPEMA in Algorithm 1 is defined without ambiguity and can be implemented without extra work.*

*Their results require number of episodes in each splitted data $M$ to satisfy $\widetilde{O}(\sqrt{nSA}) > M > O(HSA)$. To achieve data efficiency, they need $n \approx \Theta(H^2SA/\epsilon^2)$ and by that condition $M$ has to satisfy $M \approx C \cdot HSA$. In this case, data-splitting version needs to create $N = n/M$ empirical transition dynamics and each dynamics use $H^3/N \approx C \cdot H^2SA/\epsilon^2$ episodes which is less than the lower bound ($O(H^3)$) required for learning. Most critically, due to data-splitting it has $N$ empirical transitions hence it is not clear which transition to plan over. Therefore in this sense their result does not enables efficient offline learning. Our Analysis for unsplitted version (OPEMA) addresses all these issues.*